US 20220284582A1

(19) United States

(12) Patent Application Publication

Yang et al.

(10) Pub. No.: US 2022/0284582 A1

(43) Pub. Date: Sep. 8, 2022

(54) SELECTING A NEURAL NETWORK BASED ON AN AMOUNT OF MEMORY

(71) Applicant: **NVIDIA Corporation**, Santa Clara, CA (US)

(72) Inventors: **Dong Yang**, Pocatello, ID (US); **Yufan He**, Philadelphia, PA (US); **Holger Reinhard Roth**, Rockville, MD (US); **Can Zhao**, North Potomac, MD (US); **Daguang Xu**, Potomac, MD (US)

(21) Appl. No.: **17/190,724**

(22) Filed: **Mar. 3, 2021**

Publication Classification

(51) **Int. Cl.**

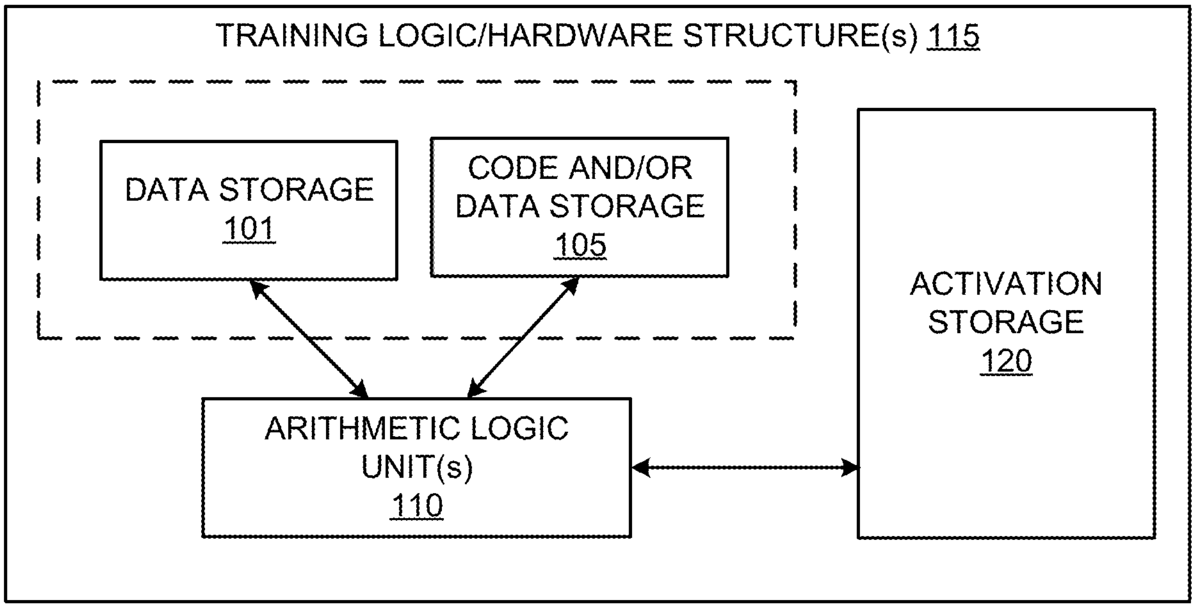
G06T 7/10	(2006.01)
G06T 7/00	(2006.01)
G06N 3/04	(2006.01)
G06N 3/10	(2006.01)

(52) **U.S. Cl.**

CPC ***G06T 7/10*** (2017.01); ***G06T 7/0012*** (2013.01); ***G06N 3/0454*** (2013.01); ***G06N 3/10*** (2013.01); *G06T 2207/20084* (2013.01)

(57) ABSTRACT

Apparatuses, systems, and techniques to select a neural network using an amount of memory to be used. In at least one embodiment, a processor includes one or more circuits to cause one or more neural networks to be selected from a plurality of neural networks based, at least in part, on an amount of memory to be used by the one oe more neural networks.


TRAINING LOGIC/HARDWARE STRUCTURE(s) 115
DATA STORAGE 101
CODE AND/OR DATA STORAGE 105
ACTIVATION STORAGE 120
ARITHMETIC LOGIC UNIT(s) 110

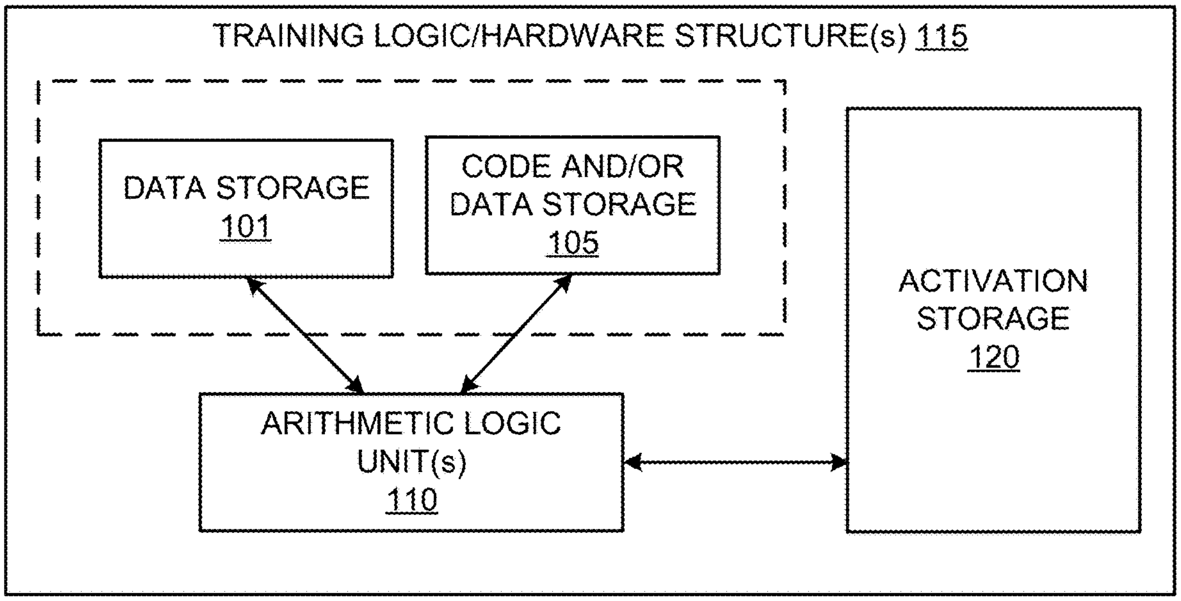

TRAINING LOGIC/HARDWARE STRUCTURE(s) 115
DATA STORAGE 101
CODE AND/OR DATA STORAGE 105
ACTIVATION STORAGE 120
ARITHMETIC LOGIC UNIT(s) 110

FIG. 1A

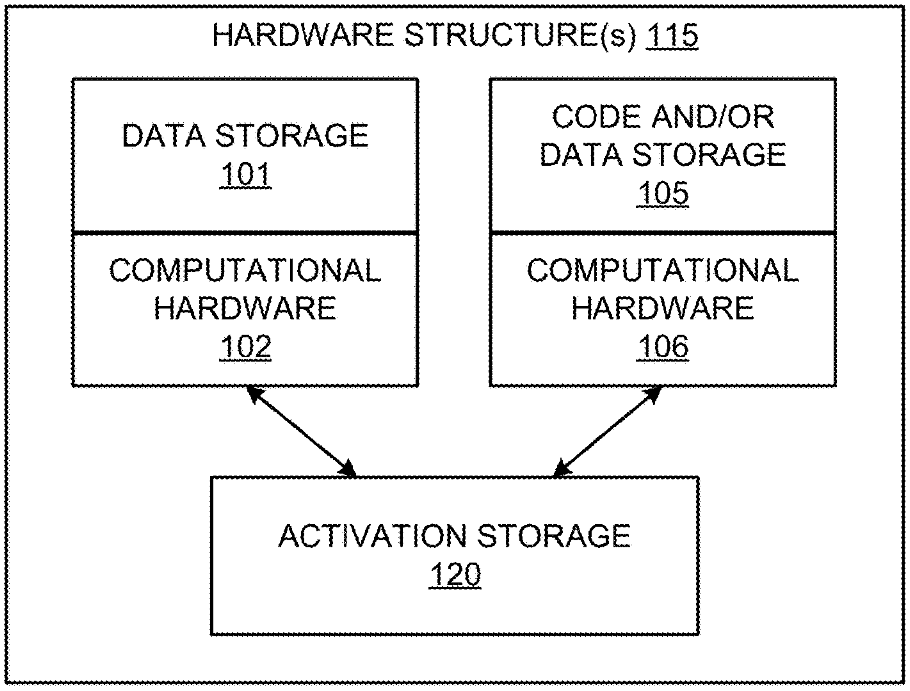

HARDWARE STRUCTURE(s) 115
DATA STORAGE 101
CODE AND/OR DATA STORAGE 105
COMPUTATIONAL HARDWARE 102
COMPUTATIONAL HARDWARE 106
ACTIVATION STORAGE 120

FIG. 1B

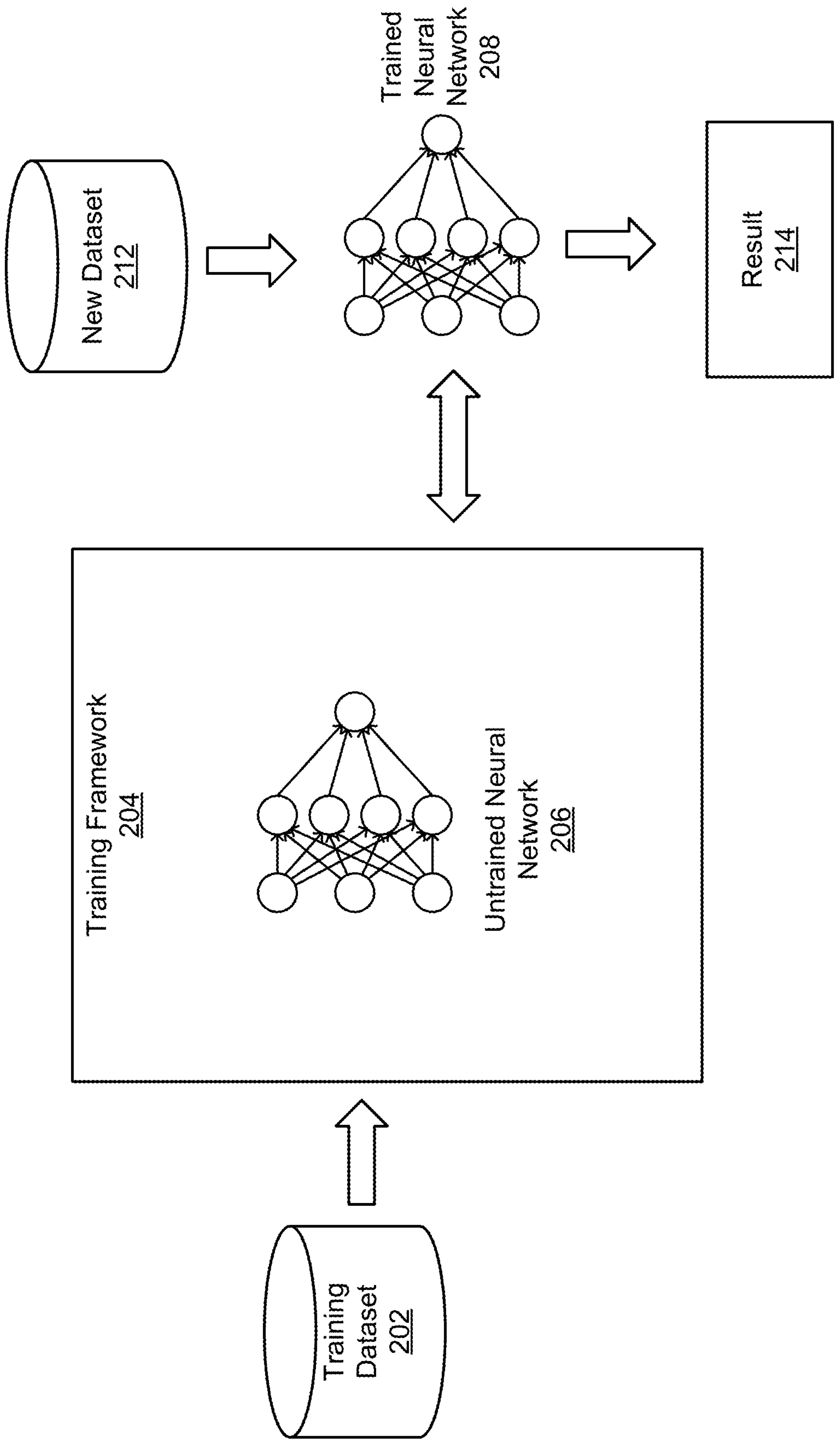

Training Dataset 202
Training Framework 204
Untrained Neural Network 206
Trained Neural Network 208
New Dataset 212
Result 214

FIG. 2

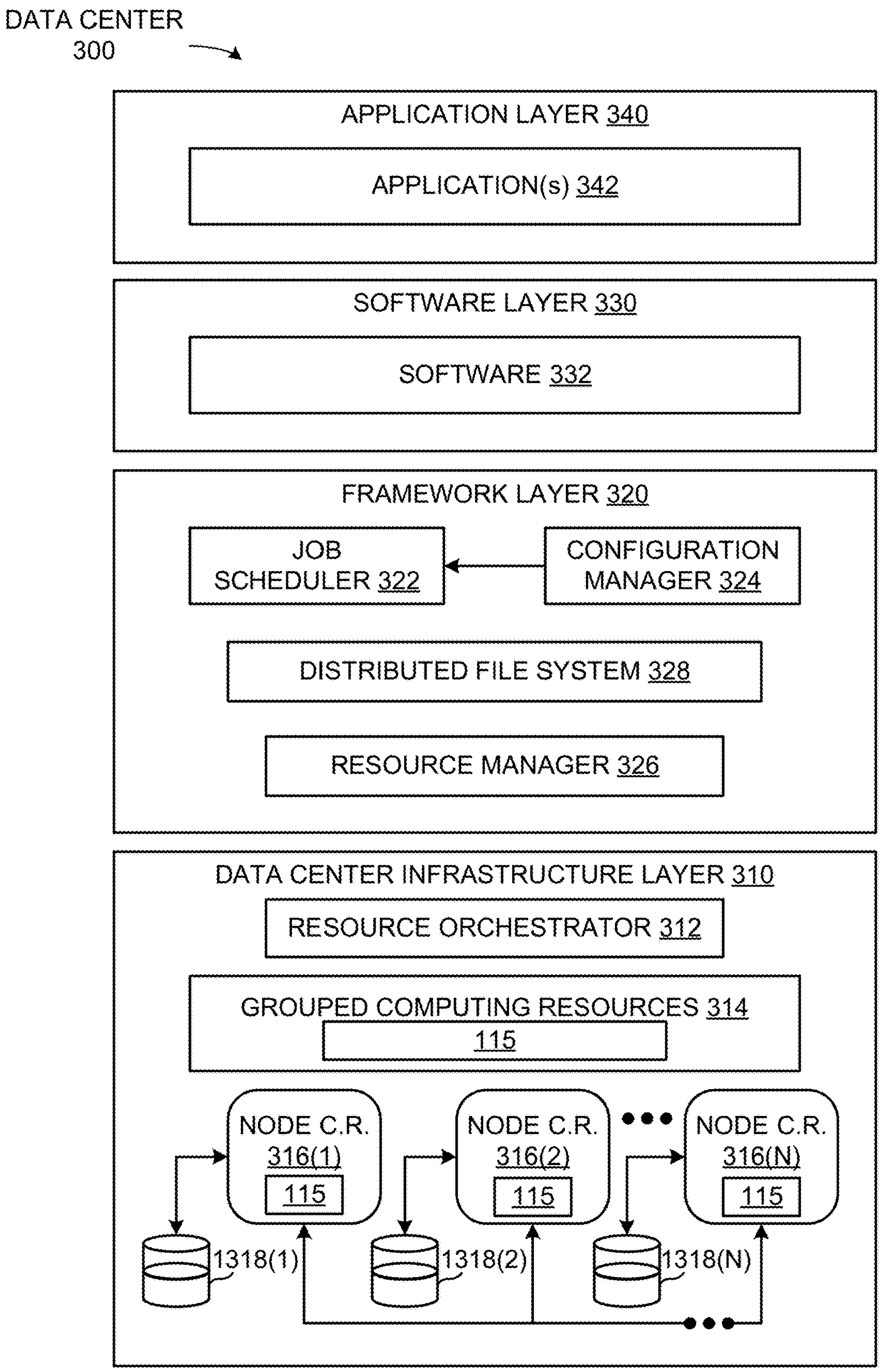

DATA CENTER
300
APPLICATION LAYER 340
APPLICATION(s) 342
SOFTWARE LAYER 330
SOFTWARE 332
FRAMEWORK LAYER 320
JOB SCHEDULER 322
CONFIGURATION MANAGER 324
DISTRIBUTED FILE SYSTEM 328
RESOURCE MANAGER 326
DATA CENTER INFRASTRUCTURE LAYER 310
RESOURCE ORCHESTRATOR 312
GROUPED COMPUTING RESOURCES 314
115
NODE C.R. 316(1)
115
NODE C.R. 316(2)
115
NODE C.R. 316(N)
115
1318(1)
1318(2)
1318(N)

FIG. 3

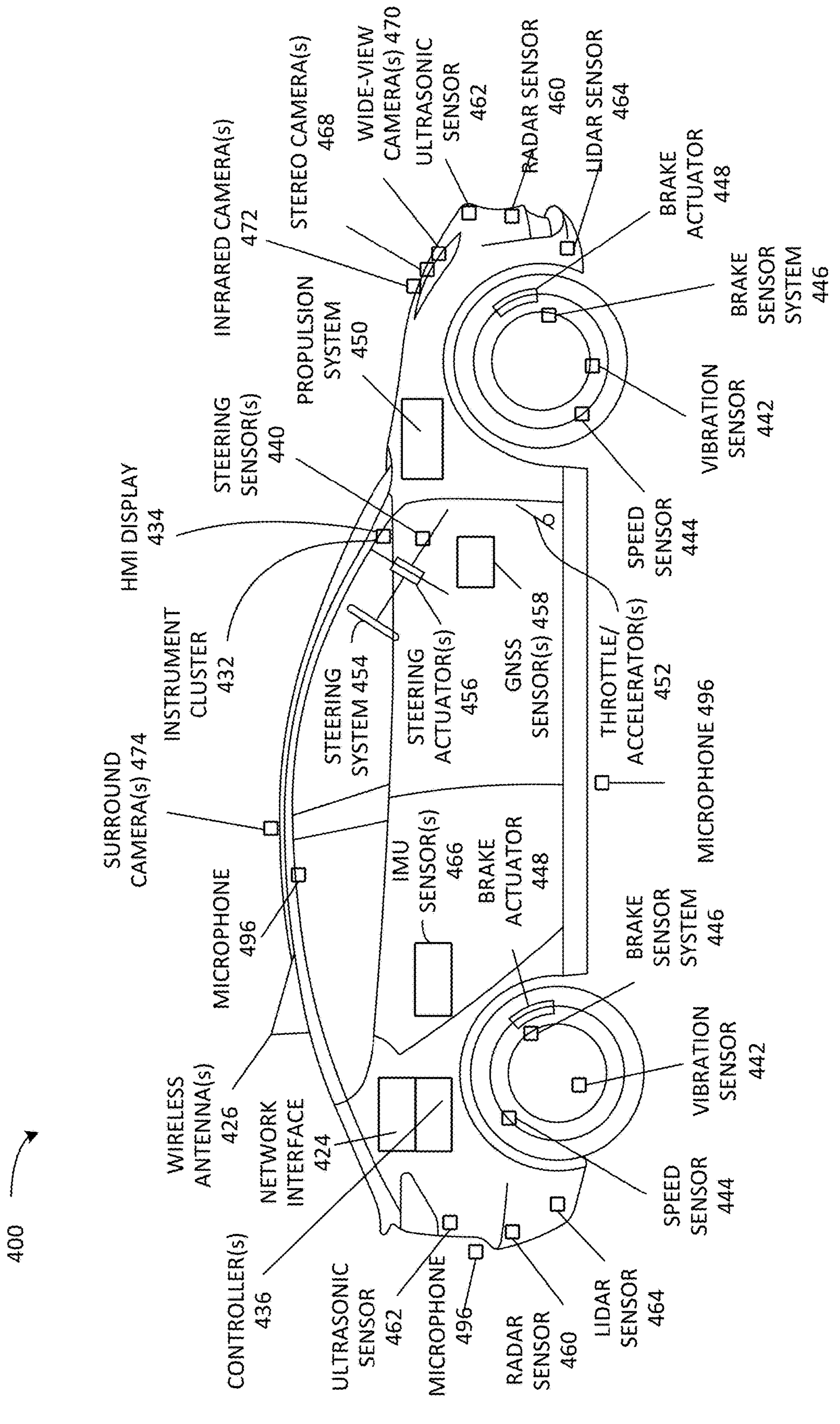

400
CONTROLLER(s) 436
ULTRASONIC SENSOR 462
MICROPHONE 496
RADAR SENSOR 460
LIDAR SENSOR 464
NETWORK INTERFACE 424
WIRELESS ANTENNA(s) 426
MICROPHONE 496
SURROUND CAMERA(s) 474
INSTRUMENT CLUSTER 432
HMI DISPLAY 434
STEERING SYSTEM 454
STEERING SENSOR(s) 440
STEERING ACTUATOR(s) 456
PROPULSION SYSTEM 450
INFRARED CAMERA(s) 472
STEREO CAMERA(s) 468
WIDE-VIEW CAMERA(s) 470
ULTRASONIC SENSOR 462
RADAR SENSOR 460
LIDAR SENSOR 464
BRAKE ACTUATOR 448
BRAKE SENSOR SYSTEM 446
VIBRATION SENSOR 442
SPEED SENSOR 444
GNSS SENSOR(s) 458
THROTTLE/ ACCELERATOR(s) 452
MICROPHONE 496
IMU SENSOR(s) 466
BRAKE ACTUATOR 448
BRAKE SENSOR SYSTEM 446
VIBRATION SENSOR 442
SPEED SENSOR 444

FIG. 4 A

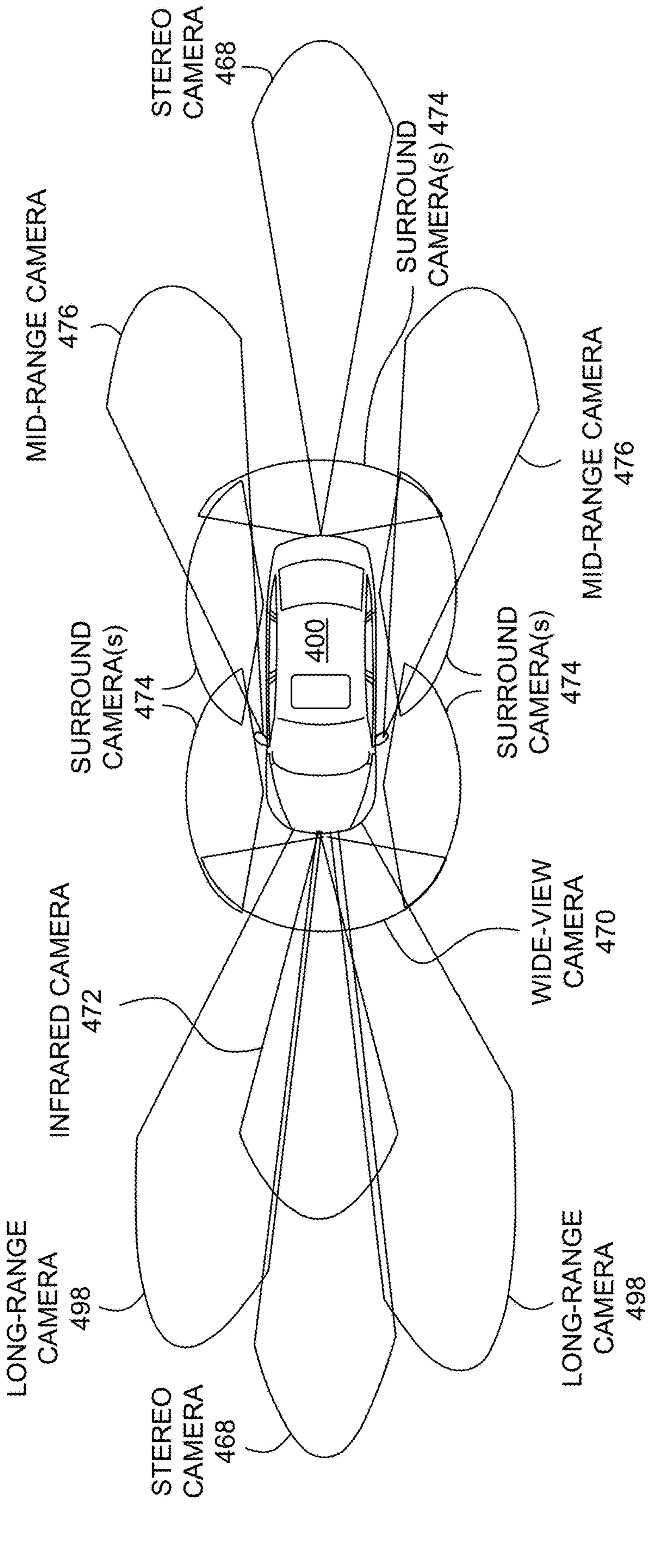

STEREO CAMERA 468
MID-RANGE CAMERA 476
SURROUND CAMERA(s) 474
MID-RANGE CAMERA 476
400
SURROUND CAMERA(s) 474
SURROUND CAMERA(s) 474
INFRARED CAMERA 472
WIDE-VIEW CAMERA 470
LONG-RANGE CAMERA 498
LONG-RANGE CAMERA 498
STEREO CAMERA 468

FIG. 4 B

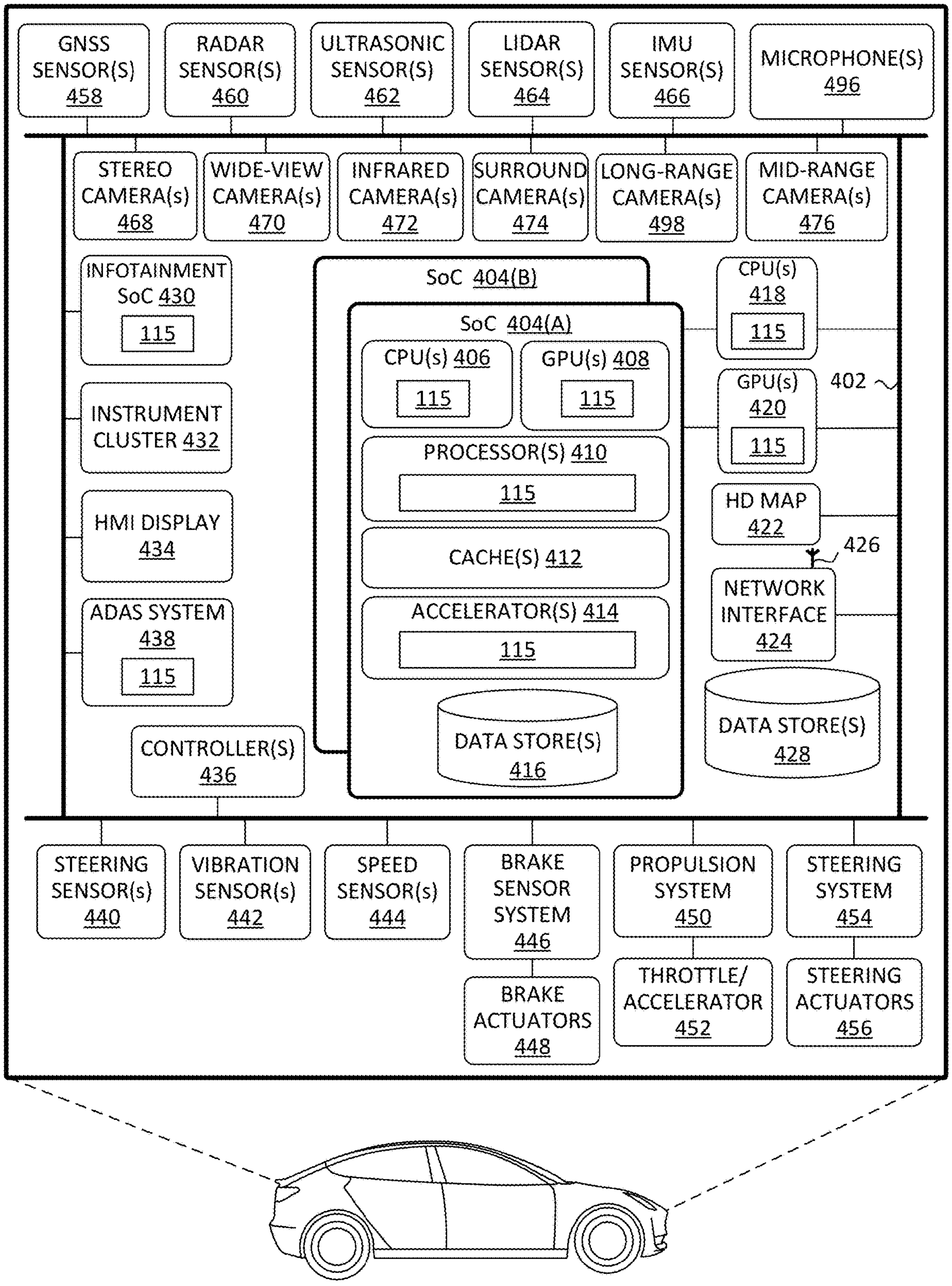

GNSS SENSOR(S) 458
RADAR SENSOR(S) 460
ULTRASONIC SENSOR(S) 462
LIDAR SENSOR(S) 464
IMU SENSOR(S) 466
MICROPHONE(S) 496
STEREO CAMERA(s) 468
WIDE-VIEW CAMERA(s) 470
INFRARED CAMERA(s) 472
SURROUND CAMERA(s) 474
LONG-RANGE CAMERA(s) 498
MID-RANGE CAMERA(s) 476
INFOTAINMENT SoC 430
115
SoC 404(B)
SoC 404(A)
CPU(s) 406
115
GPU(s) 408
115
CPU(s) 418
115
GPU(s) 420
115
402
INSTRUMENT CLUSTER 432
PROCESSOR(S) 410
115
HMI DISPLAY 434
CACHE(S) 412
HD MAP 422
426
NETWORK INTERFACE 424
ADAS SYSTEM 438
115
ACCELERATOR(S) 414
115
DATA STORE(S) 428
DATA STORE(S) 416
CONTROLLER(S) 436
STEERING SENSOR(s) 440
VIBRATION SENSOR(s) 442
SPEED SENSOR(s) 444
BRAKE SENSOR SYSTEM 446
PROPULSION SYSTEM 450
STEERING SYSTEM 454
BRAKE ACTUATORS 448
THROTTLE/ ACCELERATOR 452
STEERING ACTUATORS 456

FIG. 4 C

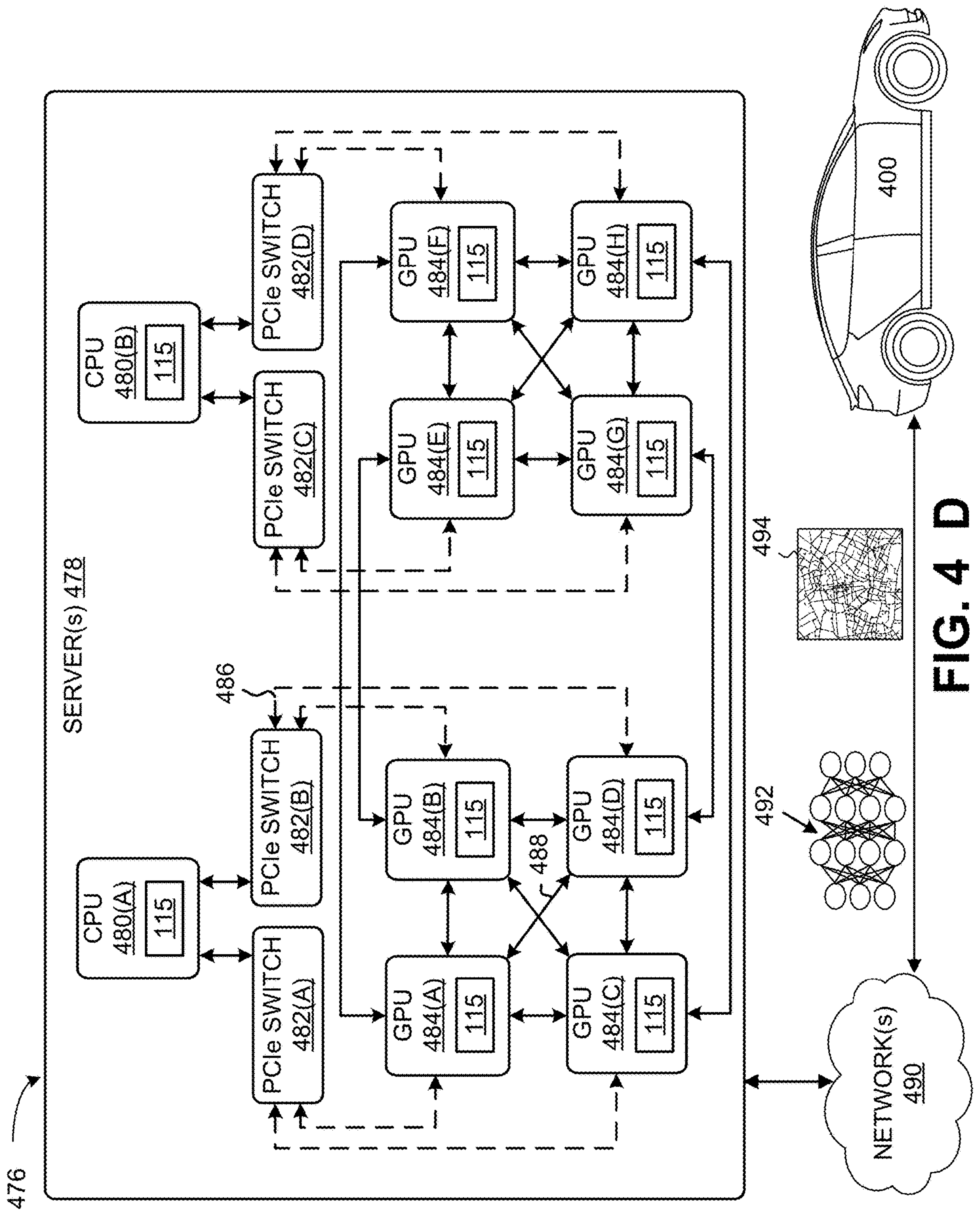

476
SERVER(s) 478
CPU 480(A)
115
CPU 480(B)
115
PCIe SWITCH 482(A)
PCIe SWITCH 482(B)
PCIe SWITCH 482(C)
PCIe SWITCH 482(D)
486
GPU 484(A)
GPU 484(B)
GPU 484(C)
GPU 484(D)
GPU 484(E)
GPU 484(F)
GPU 484(G)
GPU 484(H)
115
488
NETWORK(s) 490
492
494
400

FIG. 4 D

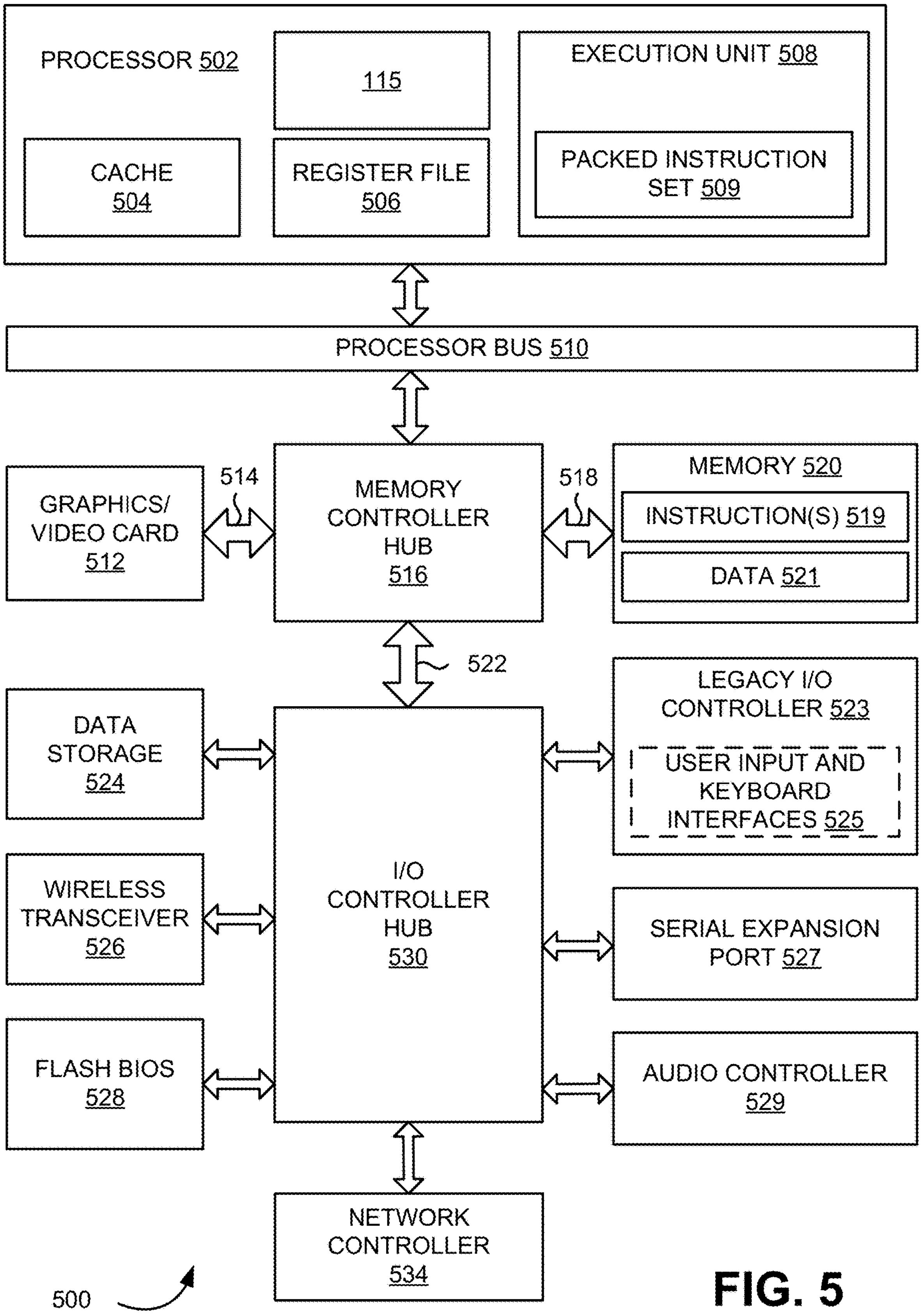

PROCESSOR 502
115
EXECUTION UNIT 508
CACHE 504
REGISTER FILE 506
PACKED INSTRUCTION SET 509
PROCESSOR BUS 510
GRAPHICS/ VIDEO CARD 512
514
MEMORY CONTROLLER HUB 516
518
MEMORY 520
INSTRUCTION(S) 519
DATA 521
522
DATA STORAGE 524
LEGACY I/O CONTROLLER 523
USER INPUT AND KEYBOARD INTERFACES 525
WIRELESS TRANSCEIVER 526
I/O CONTROLLER HUB 530
SERIAL EXPANSION PORT 527
FLASH BIOS 528
AUDIO CONTROLLER 529
NETWORK CONTROLLER 534
500

FIG. 5

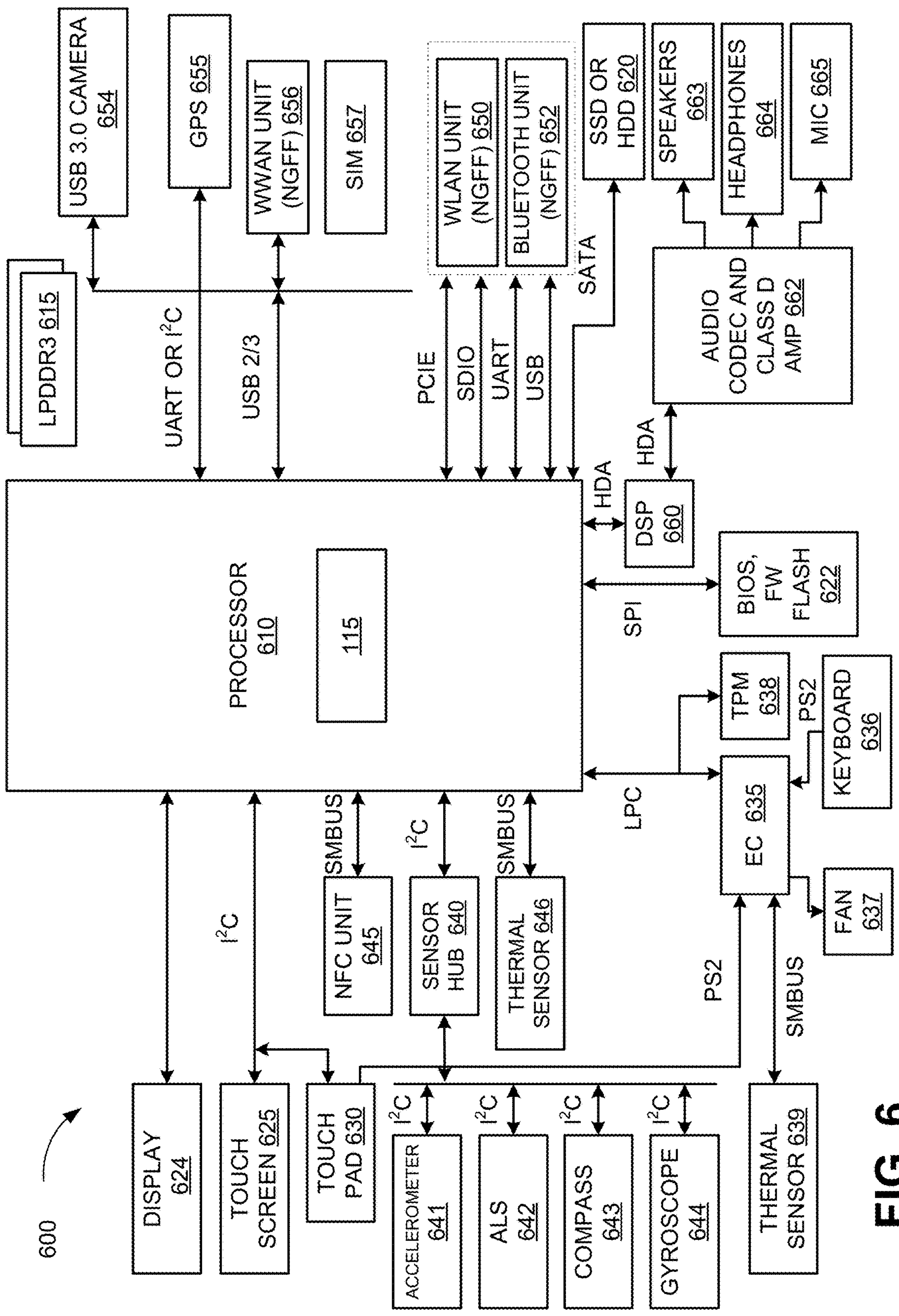

600
PROCESSOR 610
115
LPDDR3 615
USB 3.0 CAMERA 654
GPS 655
WWAN UNIT (NGFF) 656
SIM 657
WLAN UNIT (NGFF) 650
BLUETOOTH UNIT (NGFF) 652
SSD OR HDD 620
SPEAKERS 663
HEADPHONES 664
MIC 665
AUDIO CODEC AND CLASS D AMP 662
UART OR I2C
USB 2/3
PCIE
SDIO
UART
USB
SATA
HDA
HDA
DSP 660
SPI
BIOS, FW FLASH 622
TPM 638
PS2
KEYBOARD 636
LPC
EC 635
FAN 637
SMBUS
PS2
THERMAL SENSOR 639
SMBUS
NFC UNIT 645
I2C
SENSOR HUB 640
SMBUS
THERMAL SENSOR 646
I2C
DISPLAY 624
TOUCH SCREEN 625
TOUCH PAD 630
I2C
ACCELEROMETER 641
I2C
ALS 642
I2C
COMPASS 643
I2C
GYROSCOPE 644

FIG. 6

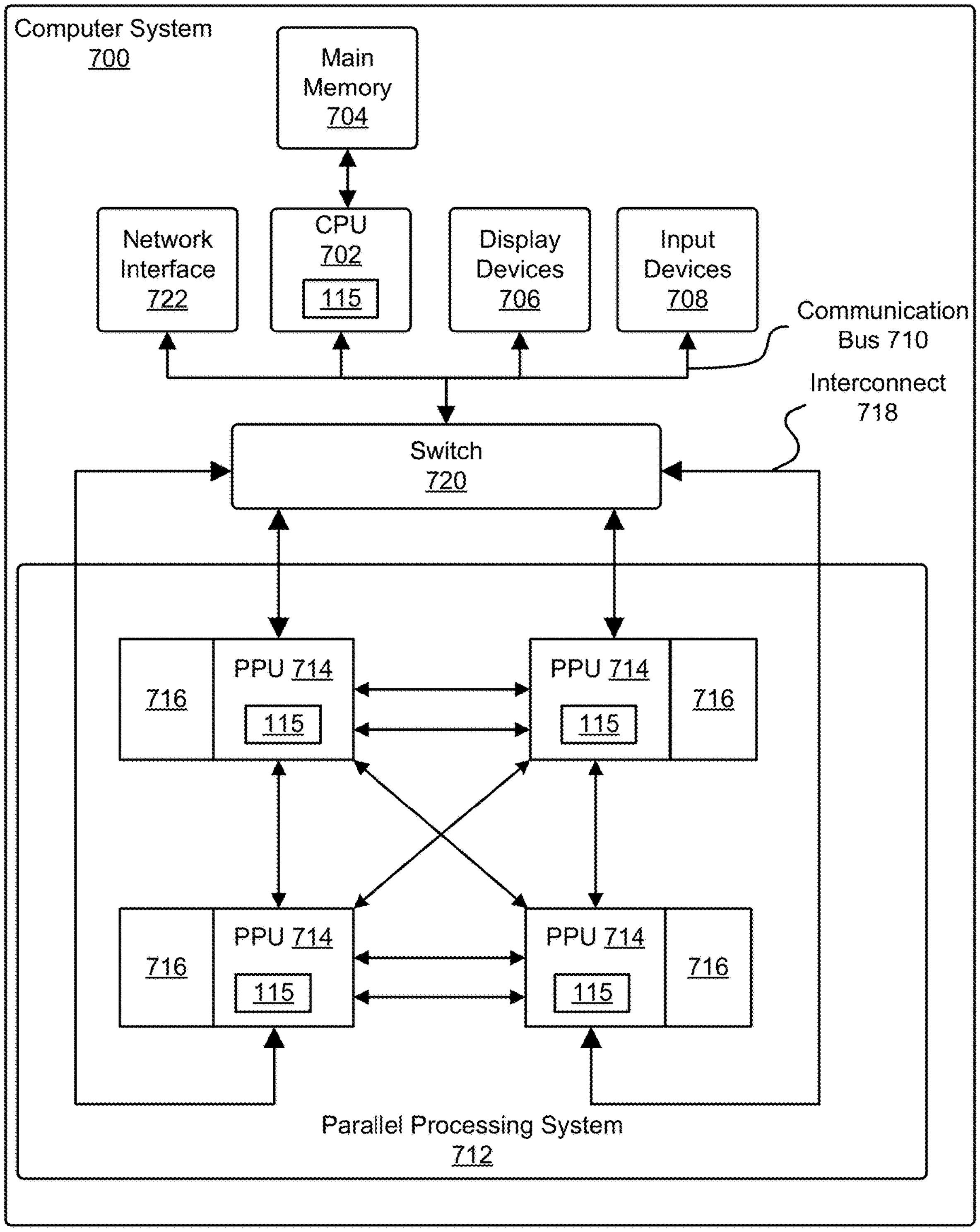

Computer System 700
Main Memory 704
Network Interface 722
CPU 702
115
Display Devices 706
Input Devices 708
Communication Bus 710
Interconnect 718
Switch 720
716
PPU 714
115
PPU 714
115
716
716
PPU 714
115
PPU 714
115
716
Parallel Processing System 712

FIG. 7

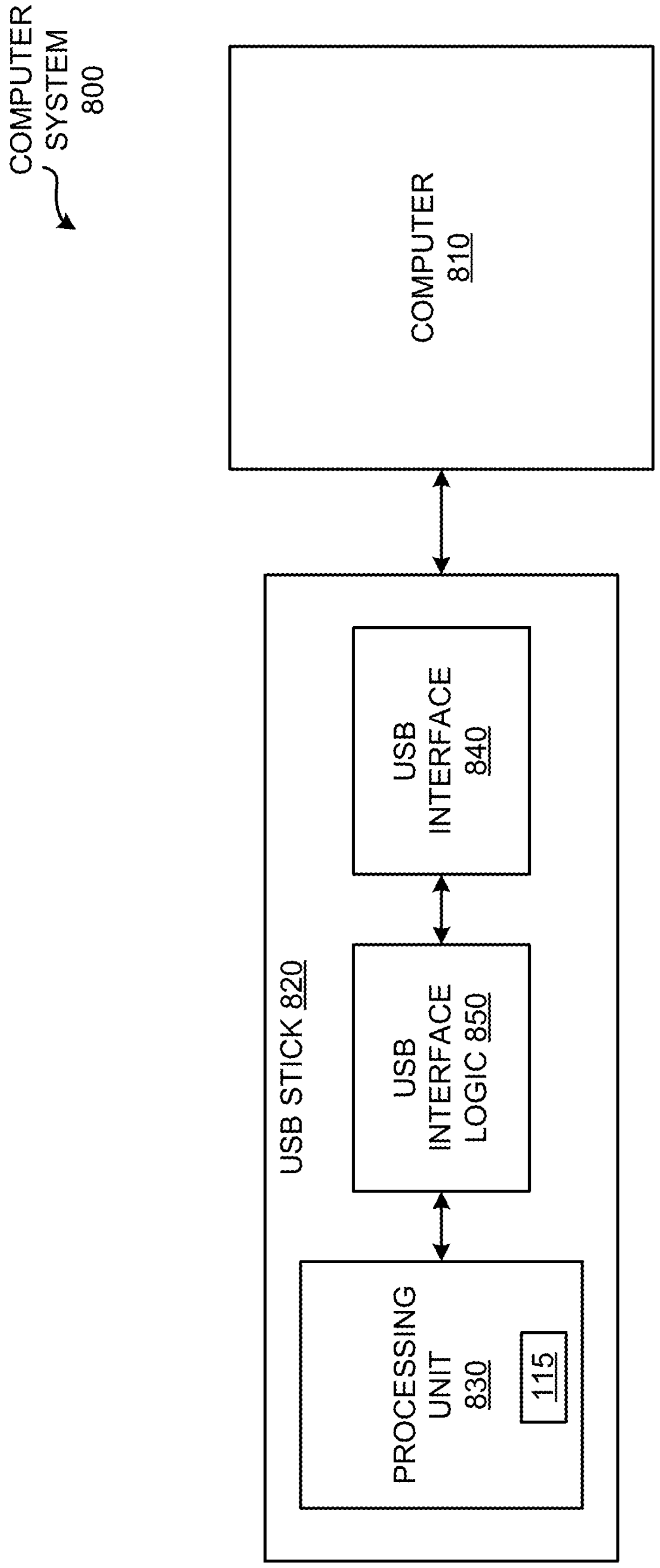

COMPUTER SYSTEM 800
COMPUTER 810
USB STICK 820
USB INTERFACE 840
USB INTERFACE LOGIC 850
PROCESSING UNIT 830
115

FIG. 8

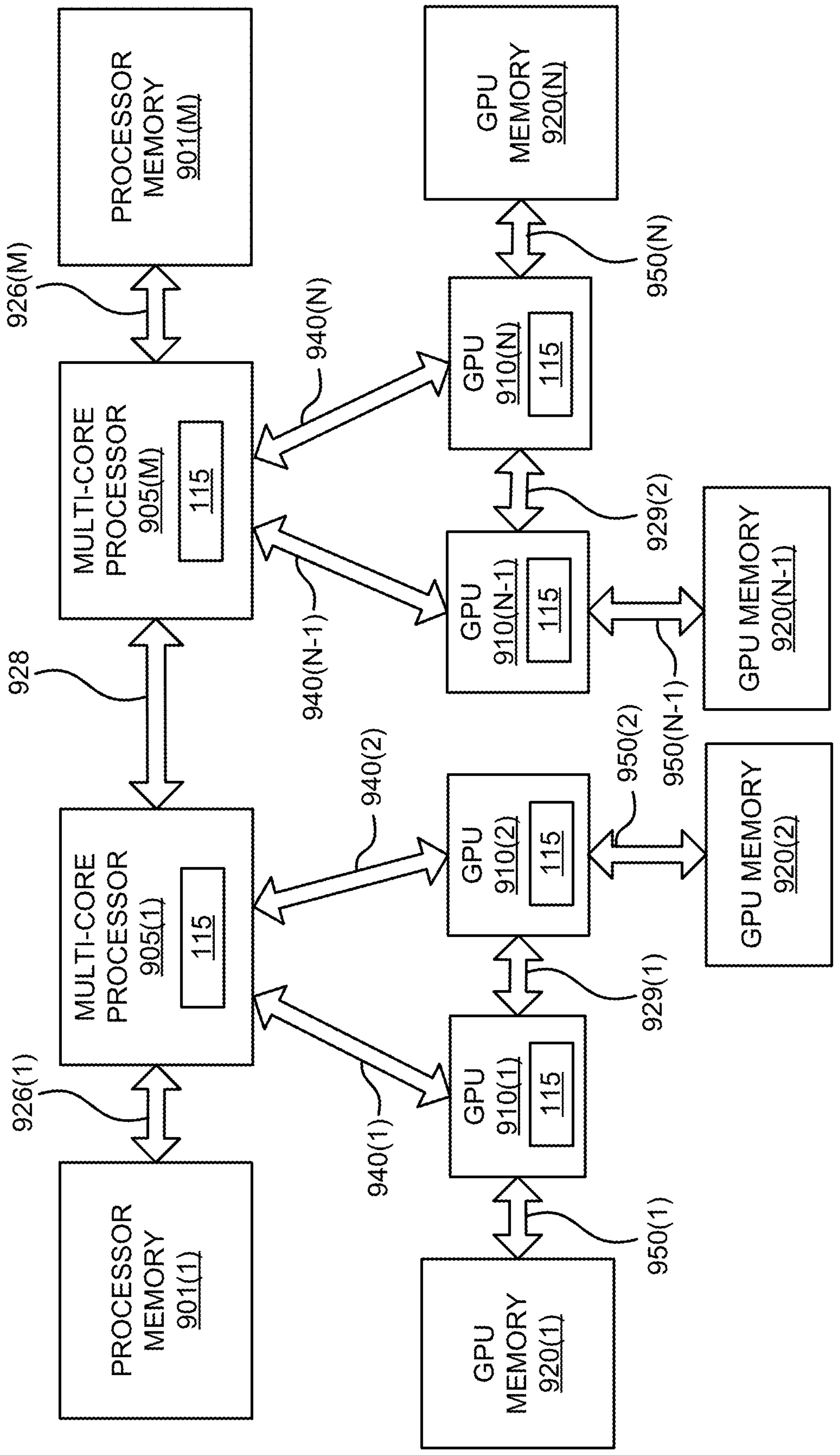

PROCESSOR MEMORY 901(M)
926(M)
MULTI-CORE PROCESSOR 905(M)
115
928
MULTI-CORE PROCESSOR 905(1)
115
926(1)
PROCESSOR MEMORY 901(1)
GPU MEMORY 920(N)
950(N)
940(N)
GPU 910(N)
115
929(2)
GPU 910(N-1)
115
940(N-1)
950(N-1)
GPU MEMORY 920(N-1)
950(2)
940(2)
GPU 910(2)
115
GPU MEMORY 920(2)
929(1)
GPU 910(1)
115
940(1)
950(1)
GPU MEMORY 920(1)

FIG. 9 A

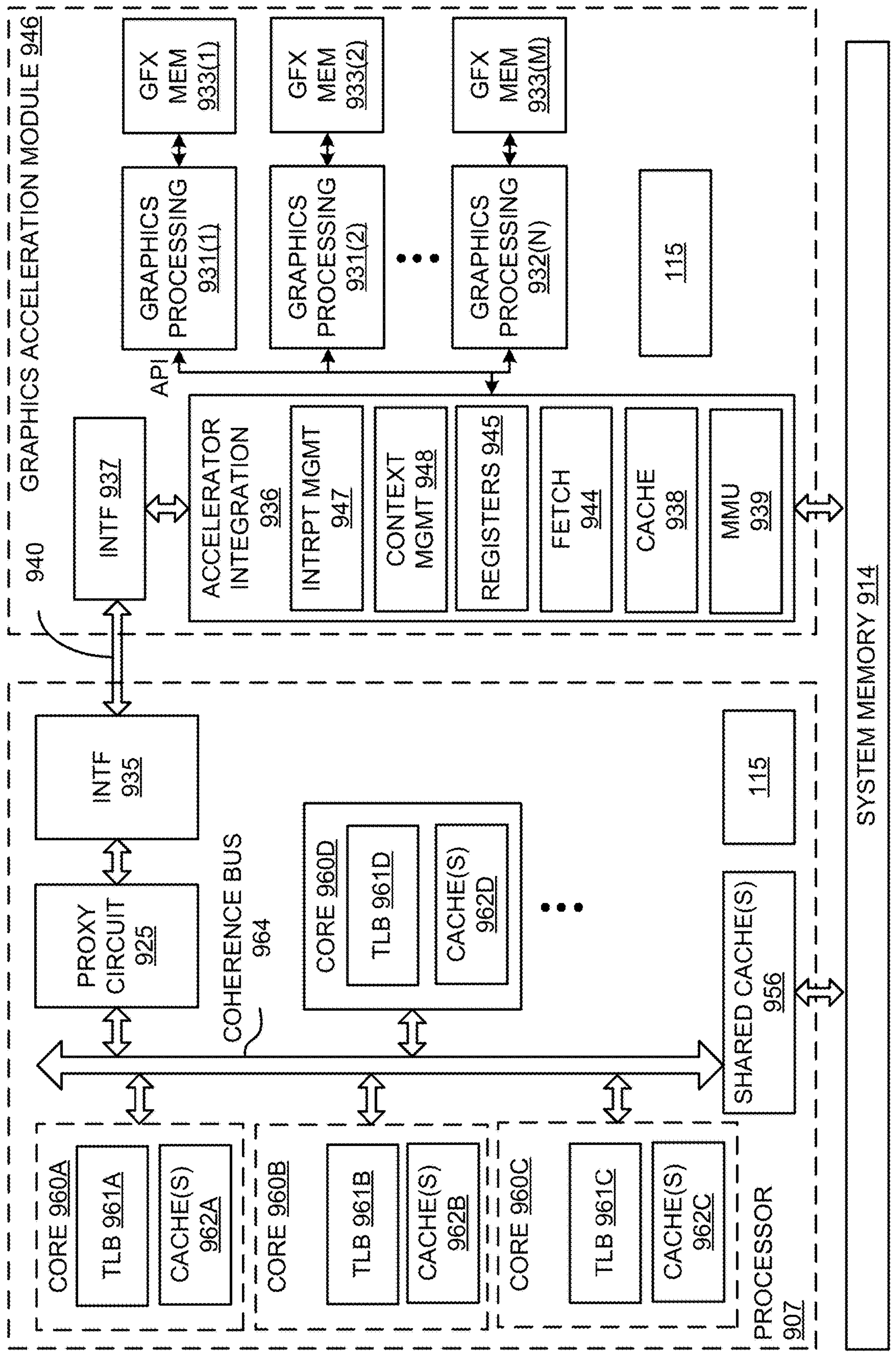

GRAPHICS ACCELERATION MODULE 946
940
INTF 937
API
ACCELERATOR INTEGRATION 936
INTRPT MGMT 947
CONTEXT MGMT 948
REGISTERS 945
FETCH 944
CACHE 938
MMU 939
GRAPHICS PROCESSING 931(1)
GRAPHICS PROCESSING 931(2)
GRAPHICS PROCESSING 932(N)
GFX MEM 933(1)
GFX MEM 933(2)
GFX MEM 933(M)
115
INTF 935
PROXY CIRCUIT 925
COHERENCE BUS 964
CORE 960D
TLB 961D
CACHE(S) 962D
115
SHARED CACHE(S) 956
CORE 960A
TLB 961A
CACHE(S) 962A
CORE 960B
TLB 961B
CACHE(S) 962B
CORE 960C
TLB 961C
CACHE(S) 962C
PROCESSOR 907
SYSTEM MEMORY 914

FIG. 9 B

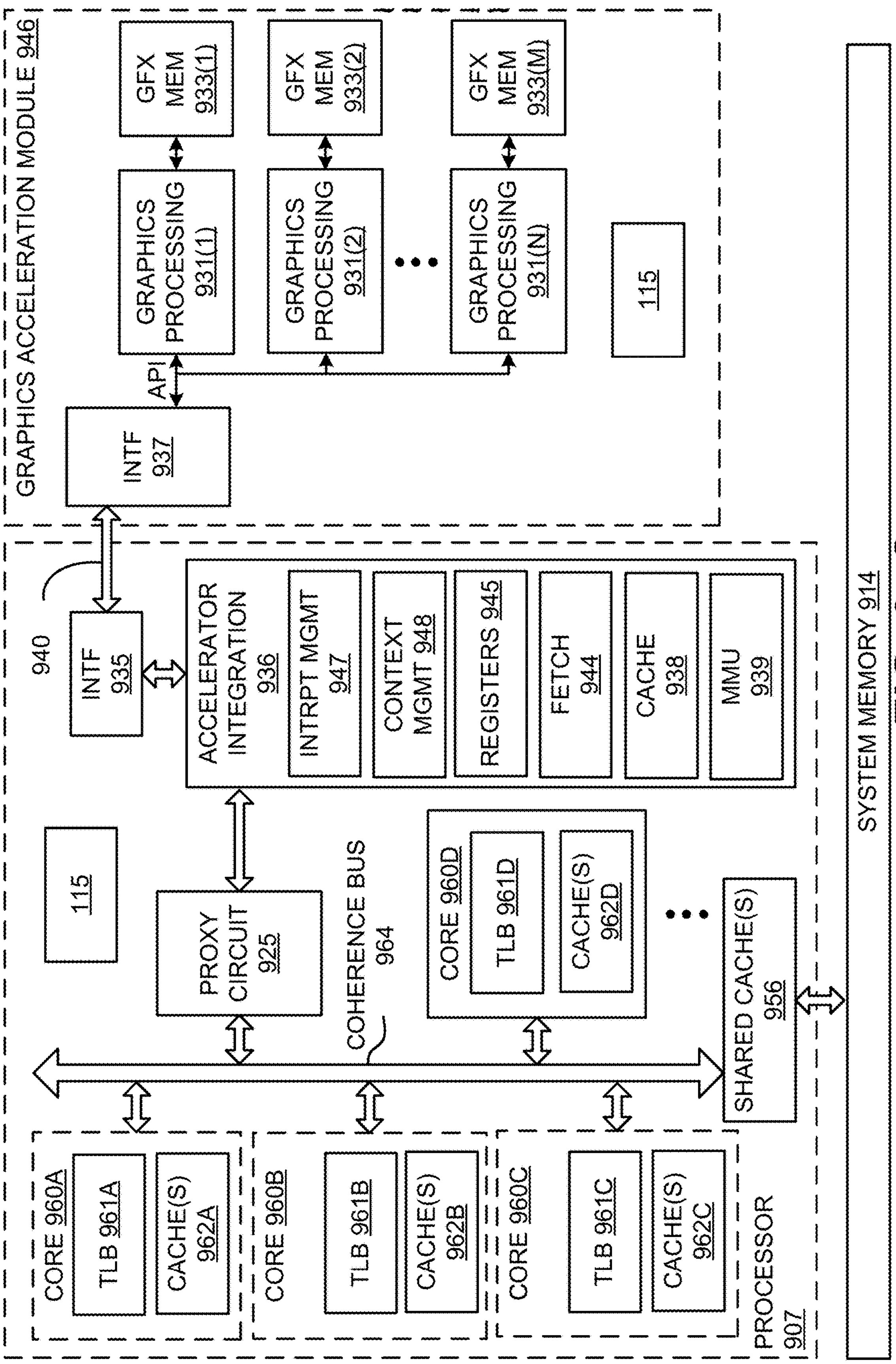

GRAPHICS ACCELERATION MODULE 946
GFX MEM 933(1)
GFX MEM 933(2)
GFX MEM 933(M)
GRAPHICS PROCESSING 931(1)
GRAPHICS PROCESSING 931(2)
GRAPHICS PROCESSING 931(N)
115
API
INTF 937
940
INTF 935
ACCELERATOR INTEGRATION 936
INTRPT MGMT 947
CONTEXT MGMT 948
REGISTERS 945
FETCH 944
CACHE 938
MMU 939
115
PROXY CIRCUIT 925
COHERENCE BUS 964
CORE 960D
TLB 961D
CACHE(S) 962D
SHARED CACHE(S) 956
CORE 960A
TLB 961A
CACHE(S) 962A
CORE 960B
TLB 961B
CACHE(S) 962B
CORE 960C
TLB 961C
CACHE(S) 962C
PROCESSOR 907
SYSTEM MEMORY 914

FIG. 9 C

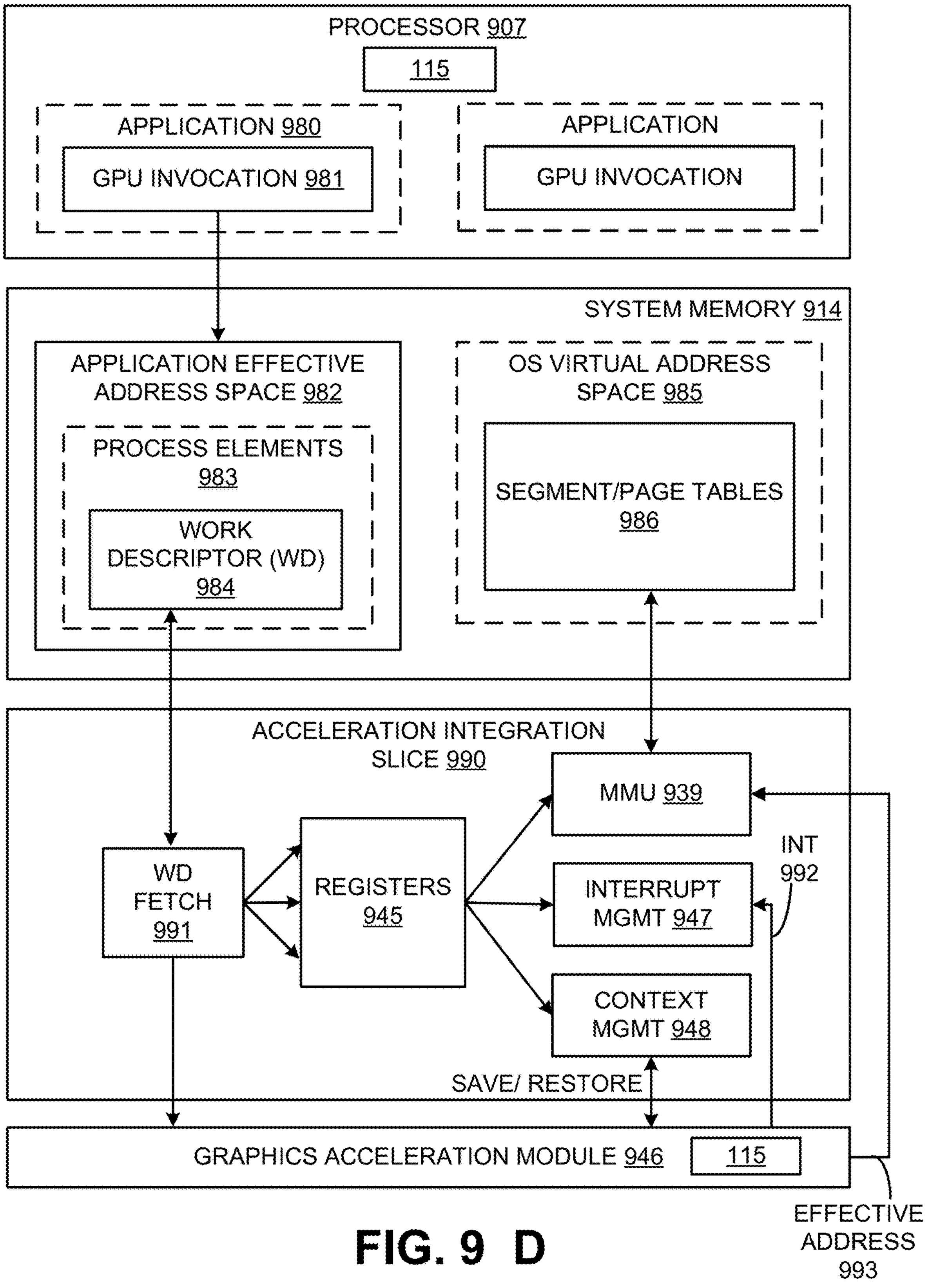

PROCESSOR 907
115
APPLICATION 980
GPU INVOCATION 981
APPLICATION
GPU INVOCATION
SYSTEM MEMORY 914
APPLICATION EFFECTIVE ADDRESS SPACE 982
PROCESS ELEMENTS 983
WORK DESCRIPTOR (WD) 984
OS VIRTUAL ADDRESS SPACE 985
SEGMENT/PAGE TABLES 986
ACCELERATION INTEGRATION SLICE 990
MMU 939
INT 992
WD FETCH 991
REGISTERS 945
INTERRUPT MGMT 947
CONTEXT MGMT 948
SAVE/ RESTORE
GRAPHICS ACCELERATION MODULE 946
115
EFFECTIVE ADDRESS 993

FIG. 9 D

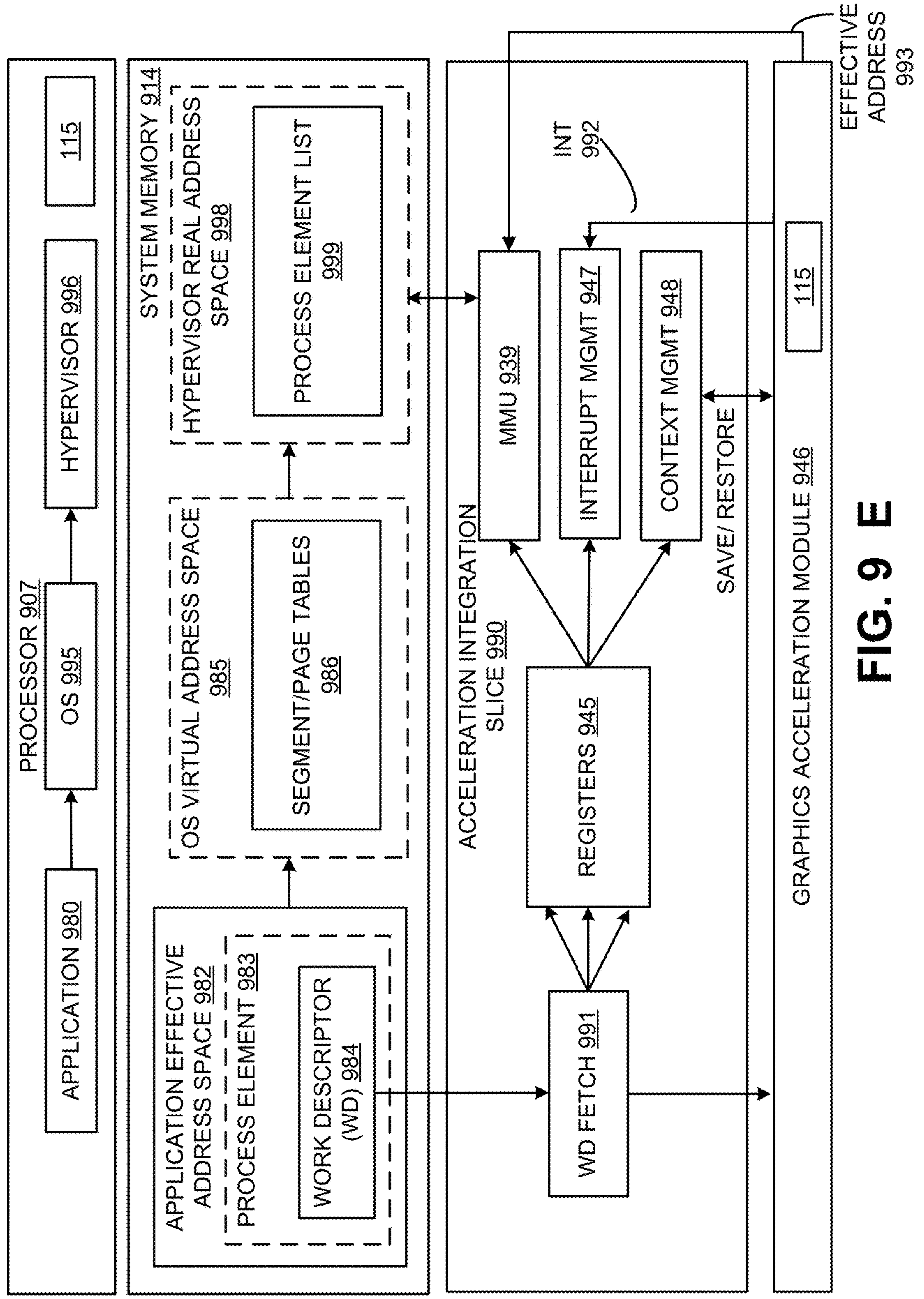

PROCESSOR 907
APPLICATION 980
OS 995
HYPERVISOR 996
115
SYSTEM MEMORY 914
APPLICATION EFFECTIVE ADDRESS SPACE 982
PROCESS ELEMENT 983
WORK DESCRIPTOR (WD) 984
OS VIRTUAL ADDRESS SPACE 985
SEGMENT/PAGE TABLES 986
HYPERVISOR REAL ADDRESS SPACE 998
PROCESS ELEMENT LIST 999
ACCELERATION INTEGRATION SLICE 990
WD FETCH 991
REGISTERS 945
MMU 939
INTERRUPT MGMT 947
CONTEXT MGMT 948
INT 992
SAVE/ RESTORE
GRAPHICS ACCELERATION MODULE 946
115
EFFECTIVE ADDRESS 993

FIG. 9 E

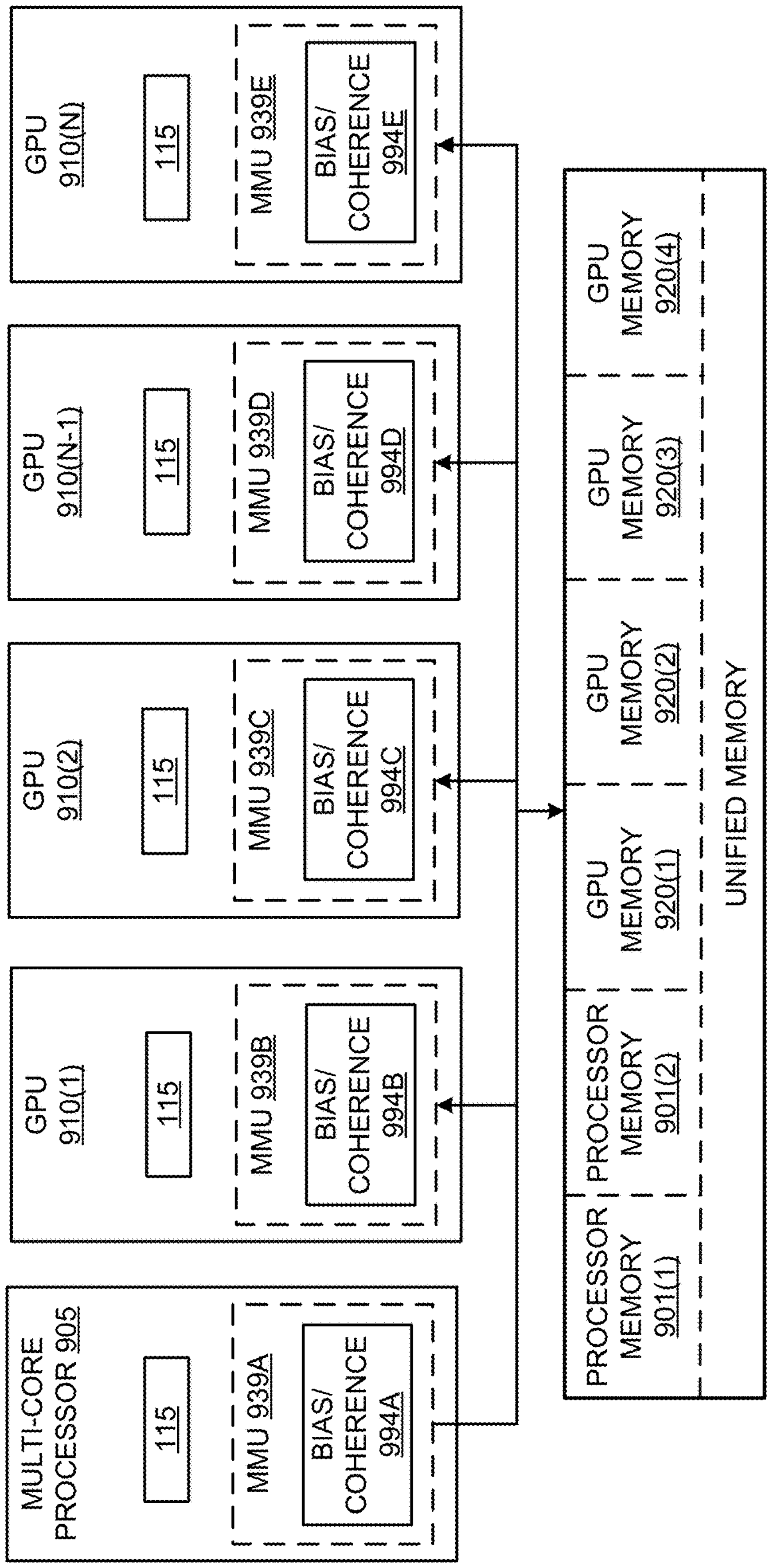

MULTI-CORE PROCESSOR 905
115
MMU 939A
BIAS/ COHERENCE 994A
GPU 910(1)
115
MMU 939B
BIAS/ COHERENCE 994B
GPU 910(2)
115
MMU 939C
BIAS/ COHERENCE 994C
GPU 910(N-1)
115
MMU 939D
BIAS/ COHERENCE 994D
GPU 910(N)
115
MMU 939E
BIAS/ COHERENCE 994E
PROCESSOR MEMORY 901(1)
PROCESSOR MEMORY 901(2)
GPU MEMORY 920(1)
GPU MEMORY 920(2)
GPU MEMORY 920(3)
GPU MEMORY 920(4)
UNIFIED MEMORY

FIG. 9 F

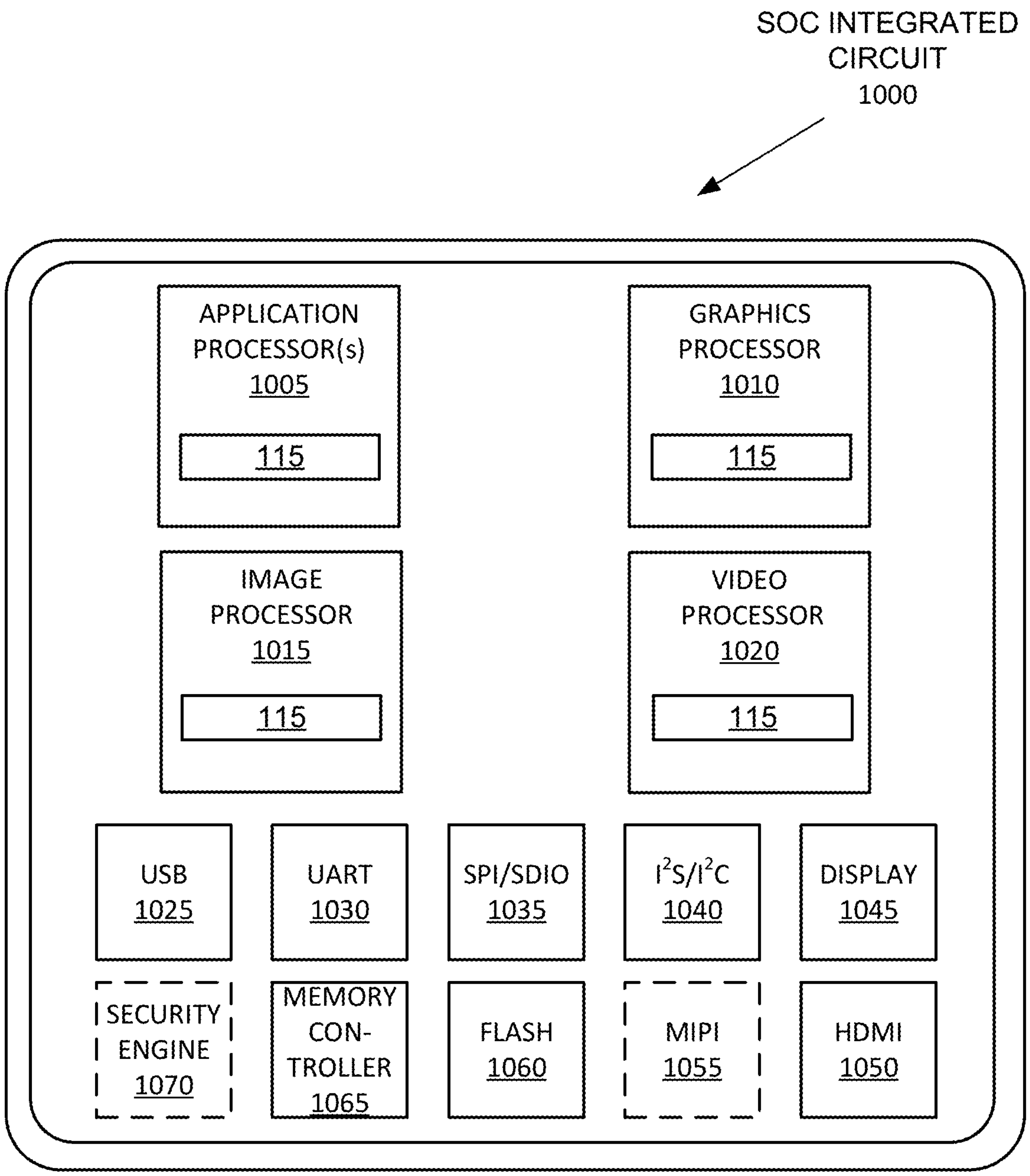

SOC INTEGRATED CIRCUIT 1000
APPLICATION PROCESSOR(s) 1005
115
GRAPHICS PROCESSOR 1010
115
IMAGE PROCESSOR 1015
115
VIDEO PROCESSOR 1020
115
USB 1025
UART 1030
SPI/SDIO 1035
I2S/I2C 1040
DISPLAY 1045
SECURITY ENGINE 1070
MEMORY CON-TROLLER 1065
FLASH 1060
MIPI 1055
HDMI 1050

FIG. 10

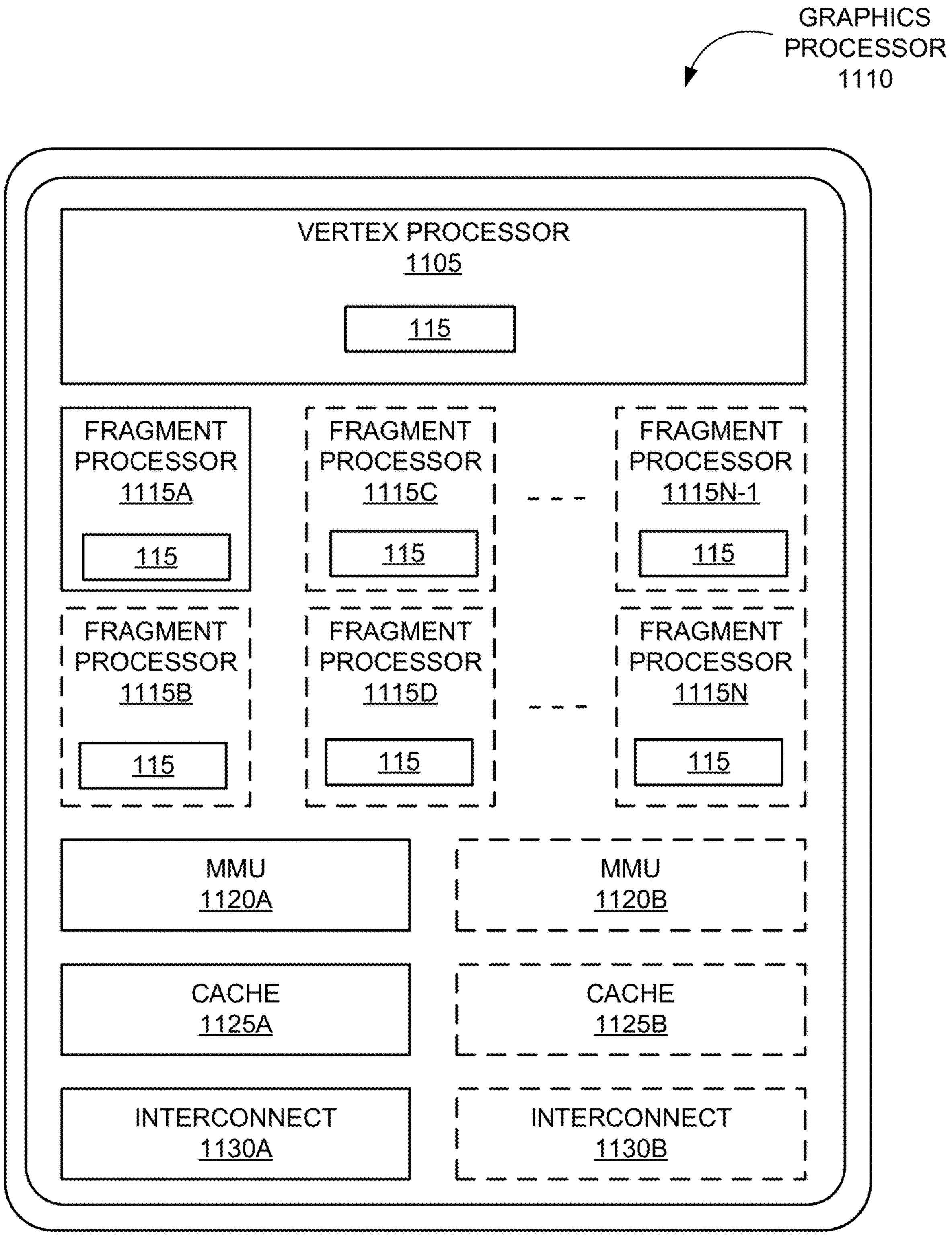

GRAPHICS PROCESSOR 1110
VERTEX PROCESSOR 1105
115
FRAGMENT PROCESSOR 1115A
115
FRAGMENT PROCESSOR 1115C
115
FRAGMENT PROCESSOR 1115N-1
115
FRAGMENT PROCESSOR 1115B
115
FRAGMENT PROCESSOR 1115D
115
FRAGMENT PROCESSOR 1115N
115
MMU 1120A
MMU 1120B
CACHE 1125A
CACHE 1125B
INTERCONNECT 1130A
INTERCONNECT 1130B

FIG. 11A

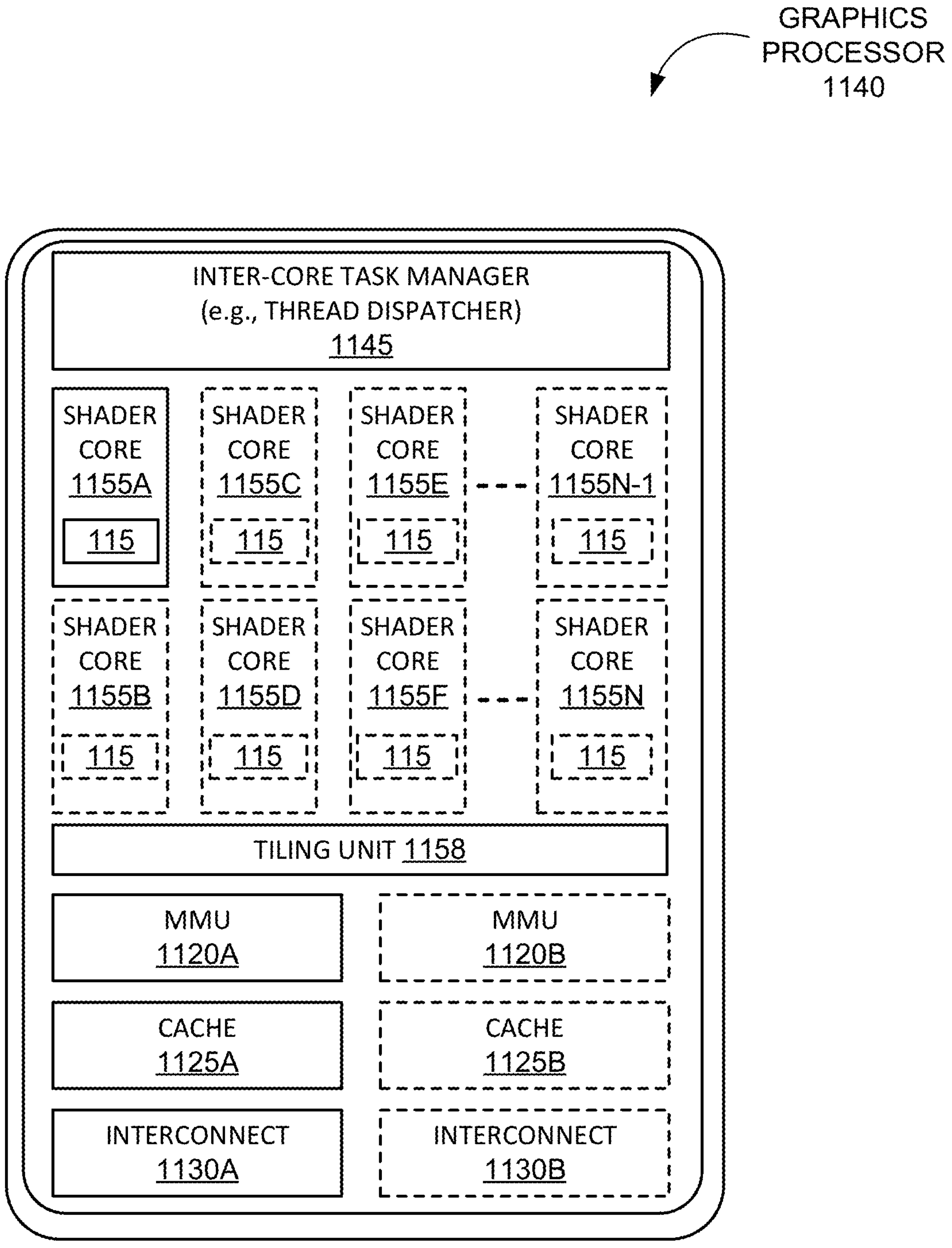

GRAPHICS PROCESSOR 1140
INTER-CORE TASK MANAGER (e.g., THREAD DISPATCHER) 1145
SHADER CORE 1155A
115
SHADER CORE 1155C
115
SHADER CORE 1155E
115
SHADER CORE 1155N-1
115
SHADER CORE 1155B
115
SHADER CORE 1155D
115
SHADER CORE 1155F
115
SHADER CORE 1155N
115
TILING UNIT 1158
MMU 1120A
MMU 1120B
CACHE 1125A
CACHE 1125B
INTERCONNECT 1130A
INTERCONNECT 1130B

FIG. 11B

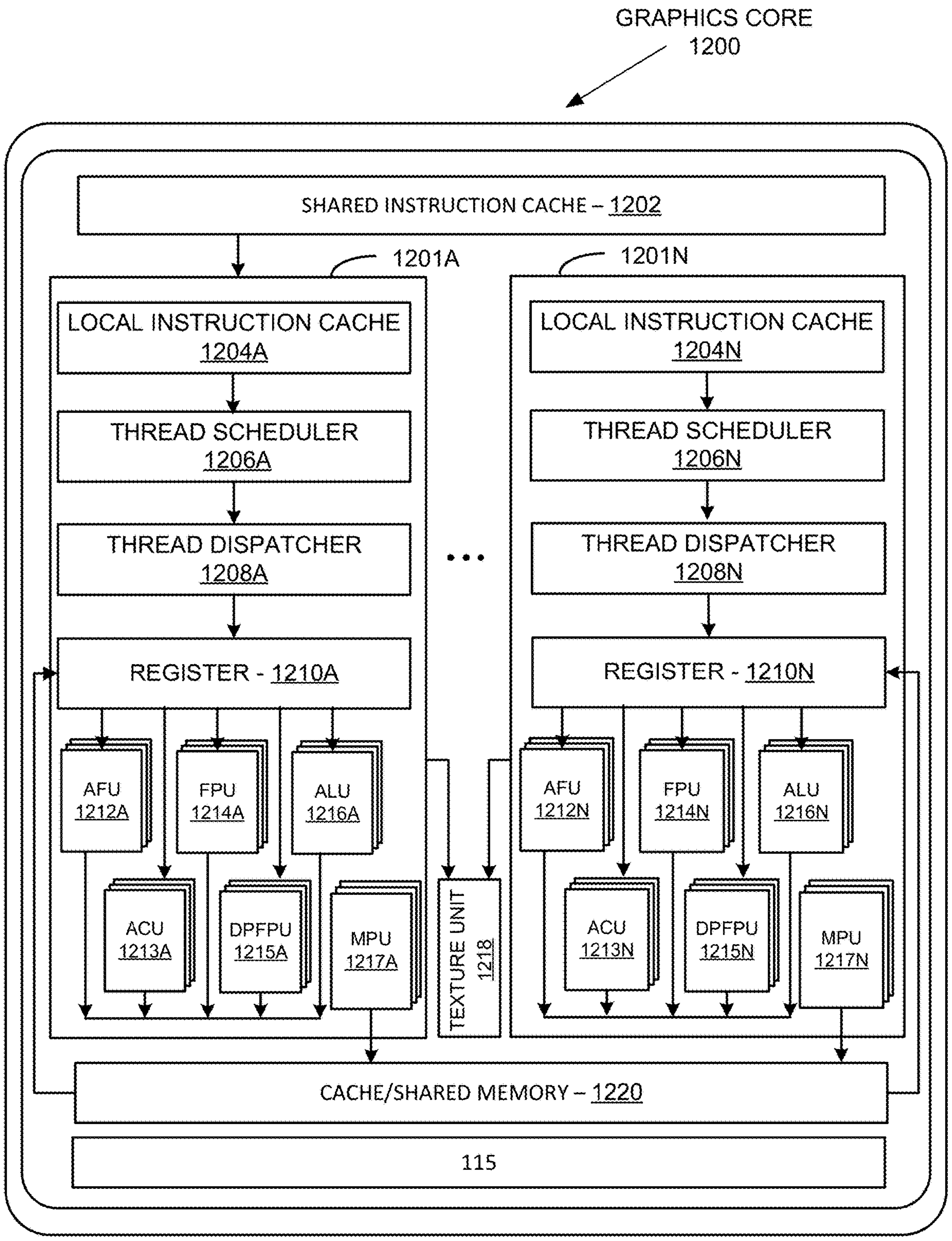

GRAPHICS CORE
1200
SHARED INSTRUCTION CACHE – 1202
1201A
1201N
LOCAL INSTRUCTION CACHE
1204A
THREAD SCHEDULER
1206A
THREAD DISPATCHER
1208A
REGISTER - 1210A
AFU
1212A
FPU
1214A
ALU
1216A
ACU
1213A
DPFPU
1215A
MPU
1217A
LOCAL INSTRUCTION CACHE
1204N
THREAD SCHEDULER
1206N
THREAD DISPATCHER
1208N
REGISTER - 1210N
AFU
1212N
FPU
1214N
ALU
1216N
ACU
1213N
DPFPU
1215N
MPU
1217N
TEXTURE UNIT
1218
CACHE/SHARED MEMORY – 1220
115

FIG. 12A

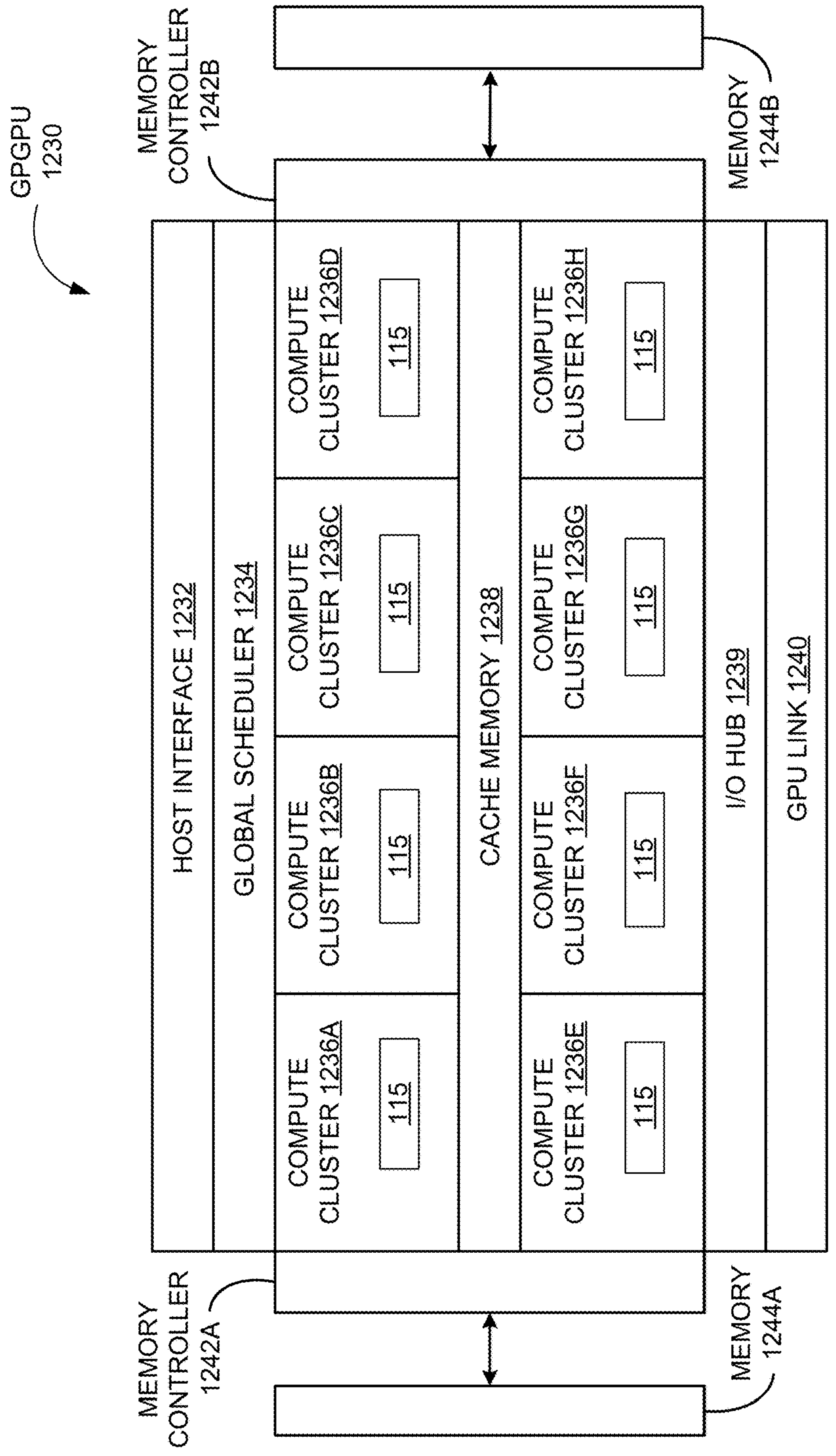

GPGPU 1230
MEMORY CONTROLLER 1242B
MEMORY 1244B
HOST INTERFACE 1232
GLOBAL SCHEDULER 1234
COMPUTE CLUSTER 1236A
COMPUTE CLUSTER 1236B
COMPUTE CLUSTER 1236C
COMPUTE CLUSTER 1236D
115
CACHE MEMORY 1238
COMPUTE CLUSTER 1236E
COMPUTE CLUSTER 1236F
COMPUTE CLUSTER 1236G
COMPUTE CLUSTER 1236H
I/O HUB 1239
GPU LINK 1240
MEMORY CONTROLLER 1242A
MEMORY 1244A

FIG. 12B

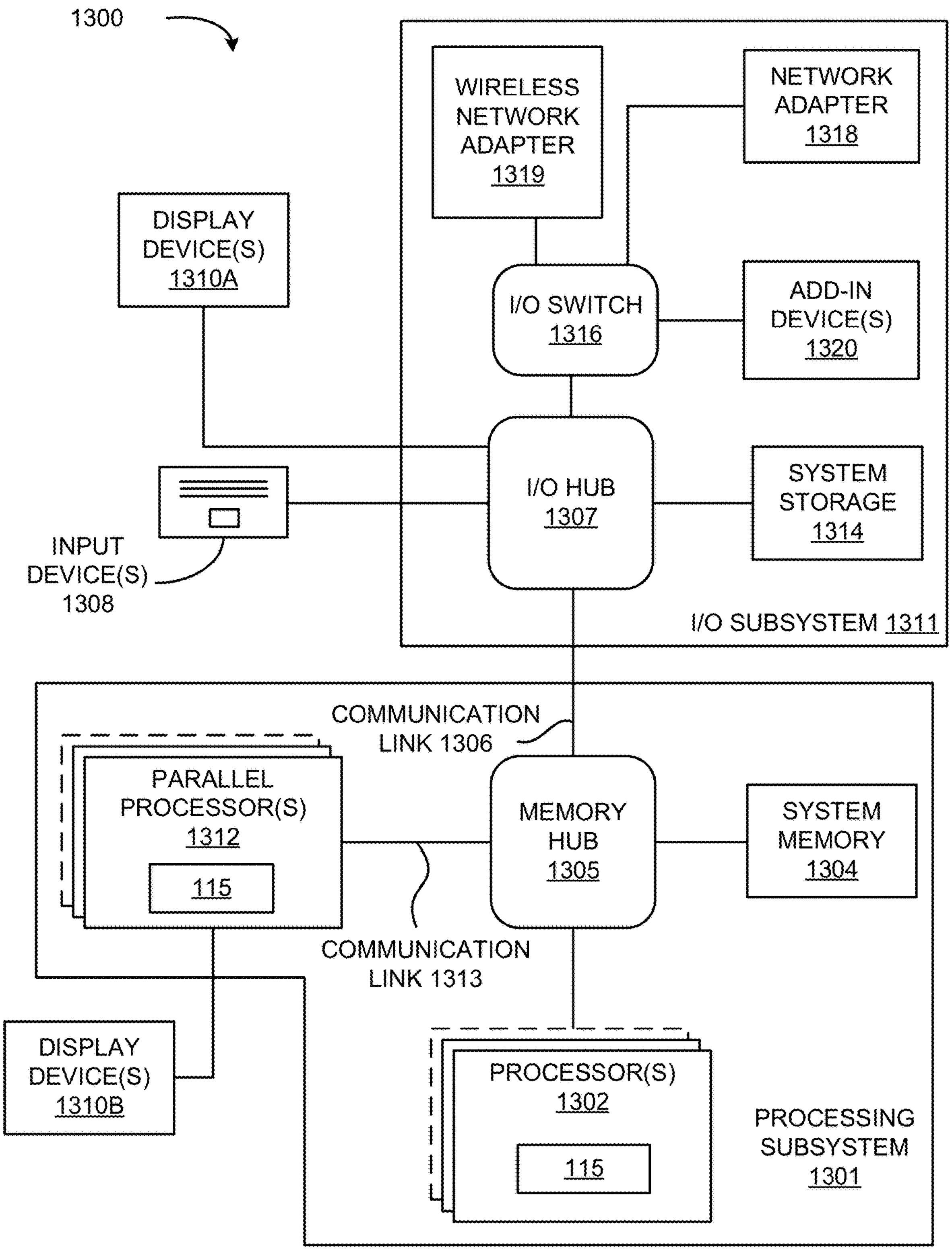

1300
WIRELESS NETWORK ADAPTER 1319
NETWORK ADAPTER 1318
DISPLAY DEVICE(S) 1310A
I/O SWITCH 1316
ADD-IN DEVICE(S) 1320
I/O HUB 1307
SYSTEM STORAGE 1314
INPUT DEVICE(S) 1308
I/O SUBSYSTEM 1311
COMMUNICATION LINK 1306
PARALLEL PROCESSOR(S) 1312
115
MEMORY HUB 1305
SYSTEM MEMORY 1304
COMMUNICATION LINK 1313
DISPLAY DEVICE(S) 1310B
PROCESSOR(S) 1302
115
PROCESSING SUBSYSTEM 1301

FIG. 13

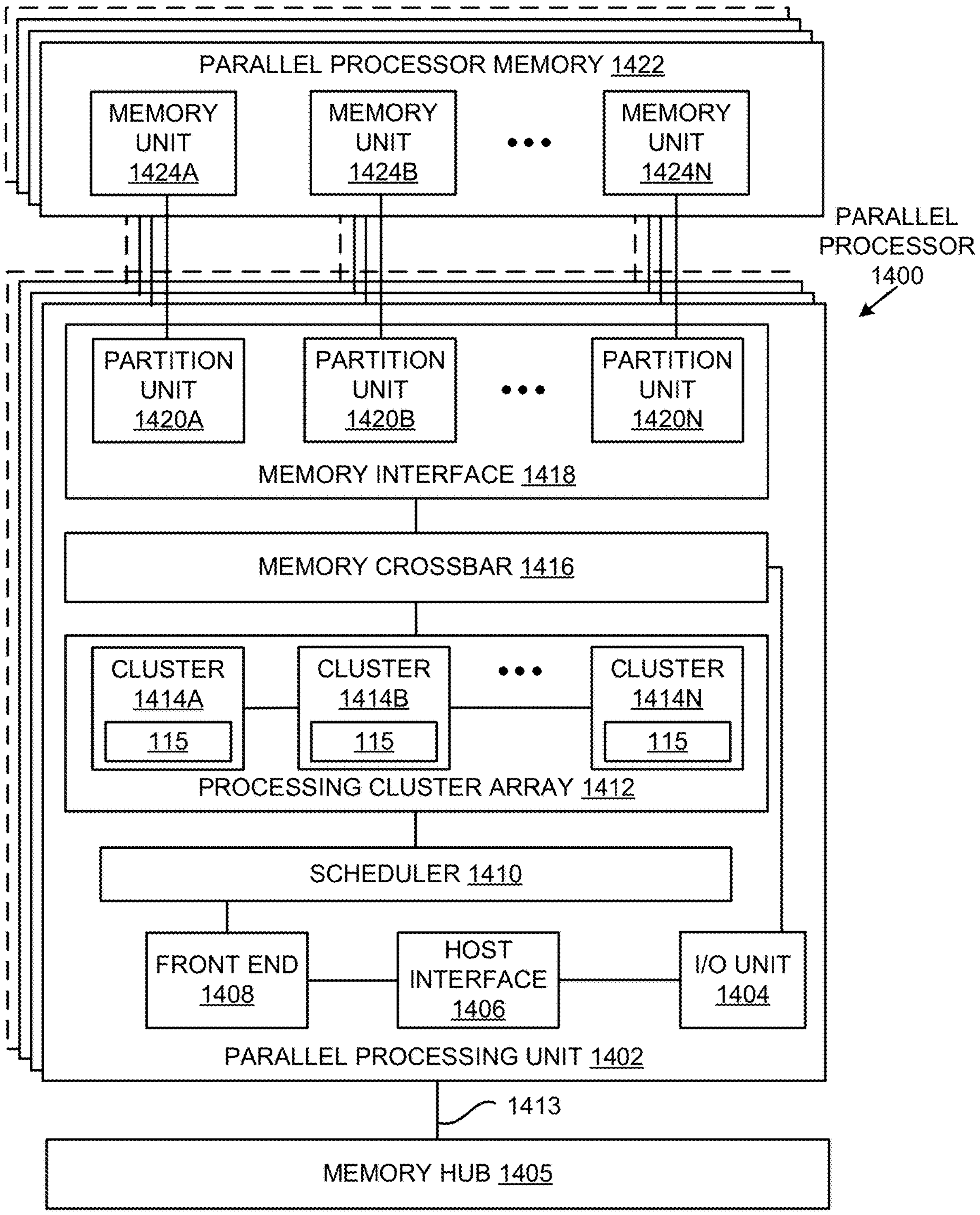

PARALLEL PROCESSOR MEMORY 1422
MEMORY UNIT 1424A
MEMORY UNIT 1424B
MEMORY UNIT 1424N
PARALLEL PROCESSOR 1400
PARTITION UNIT 1420A
PARTITION UNIT 1420B
PARTITION UNIT 1420N
MEMORY INTERFACE 1418
MEMORY CROSSBAR 1416
CLUSTER 1414A
115
CLUSTER 1414B
115
CLUSTER 1414N
115
PROCESSING CLUSTER ARRAY 1412
SCHEDULER 1410
FRONT END 1408
HOST INTERFACE 1406
I/O UNIT 1404
PARALLEL PROCESSING UNIT 1402
1413
MEMORY HUB 1405

FIG. 14 A

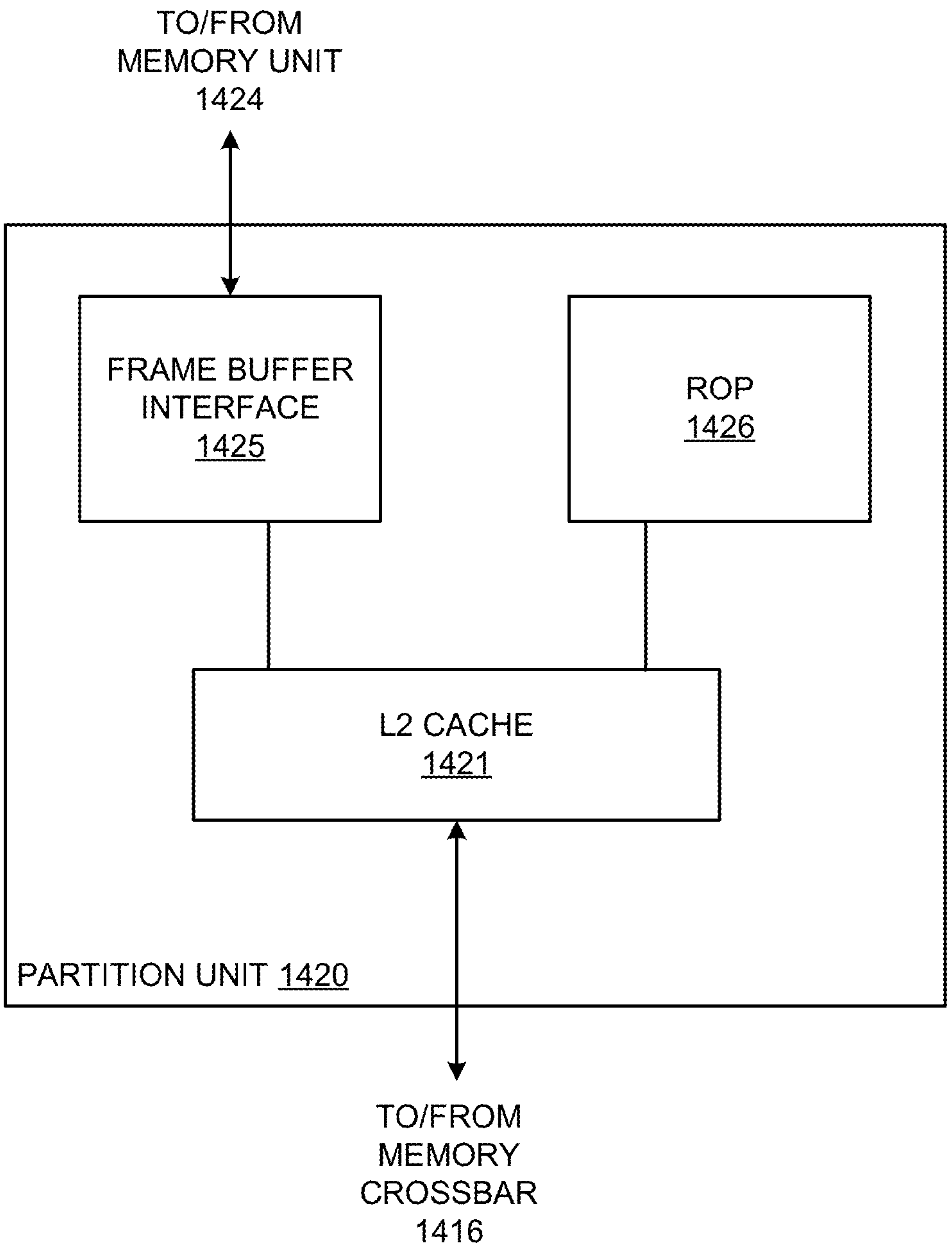

TO/FROM MEMORY UNIT 1424
FRAME BUFFER INTERFACE 1425
ROP 1426
L2 CACHE 1421
PARTITION UNIT 1420
TO/FROM MEMORY CROSSBAR 1416

FIG. 14 B

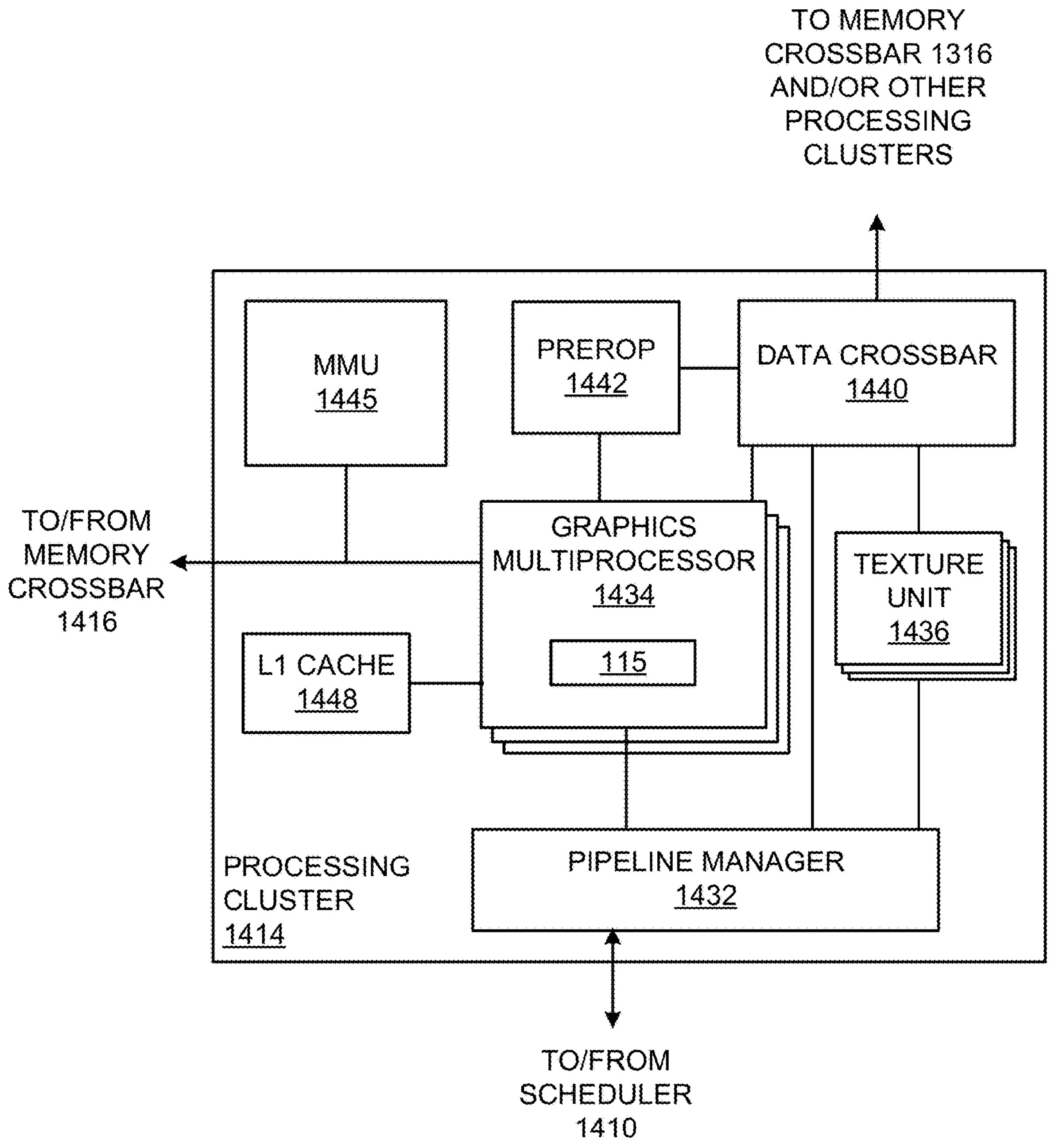
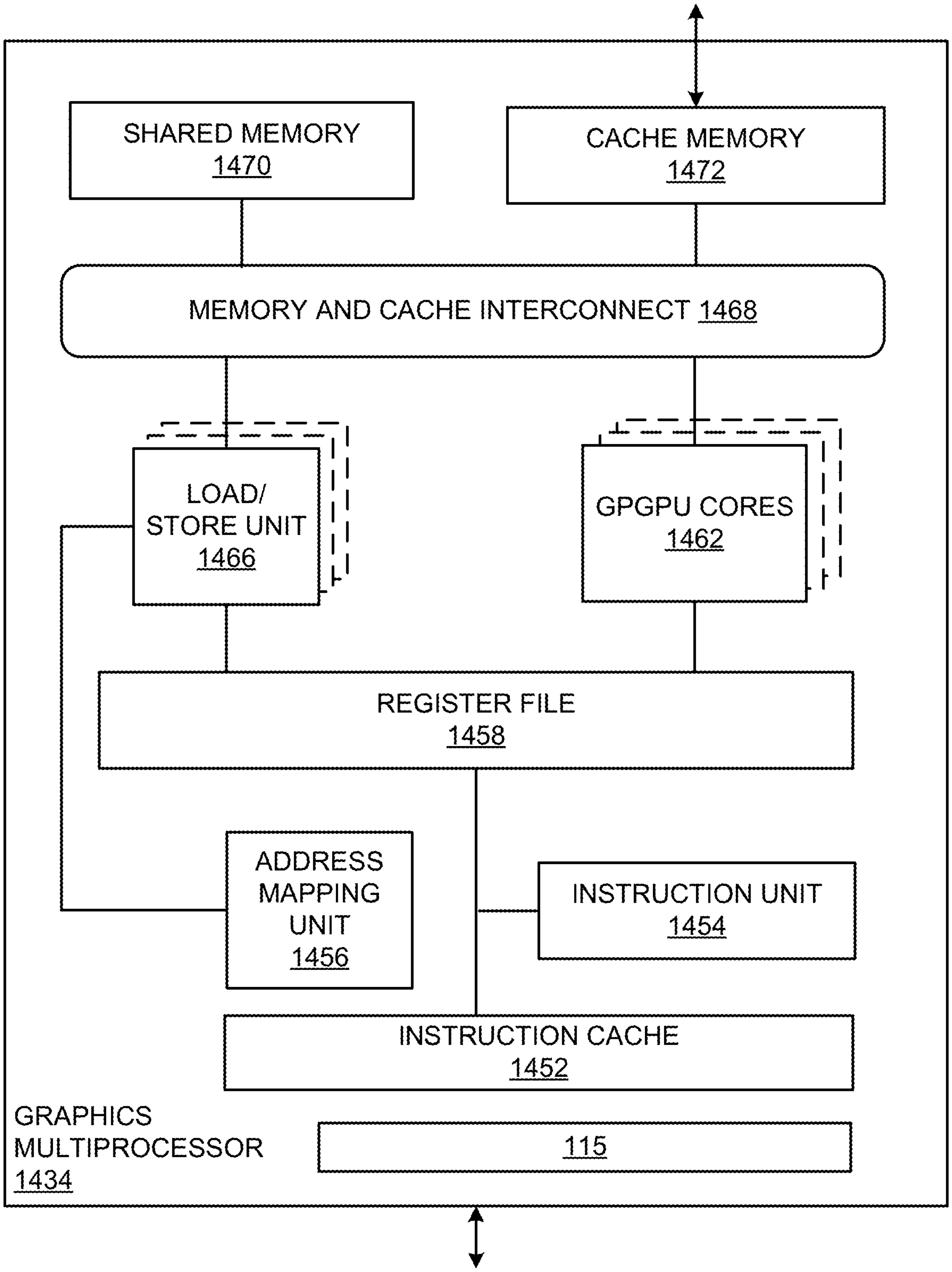

TO MEMORY CROSSBAR 1316 AND/OR OTHER PROCESSING CLUSTERS
MMU 1445
PREROP 1442
DATA CROSSBAR 1440
TO/FROM MEMORY CROSSBAR 1416
GRAPHICS MULTIPROCESSOR 1434
115
TEXTURE UNIT 1436
L1 CACHE 1448
PIPELINE MANAGER 1432
PROCESSING CLUSTER 1414
TO/FROM SCHEDULER 1410

FIG. 14 C



FROM PIPELINE MANAGER 1432

FIG. 14 D

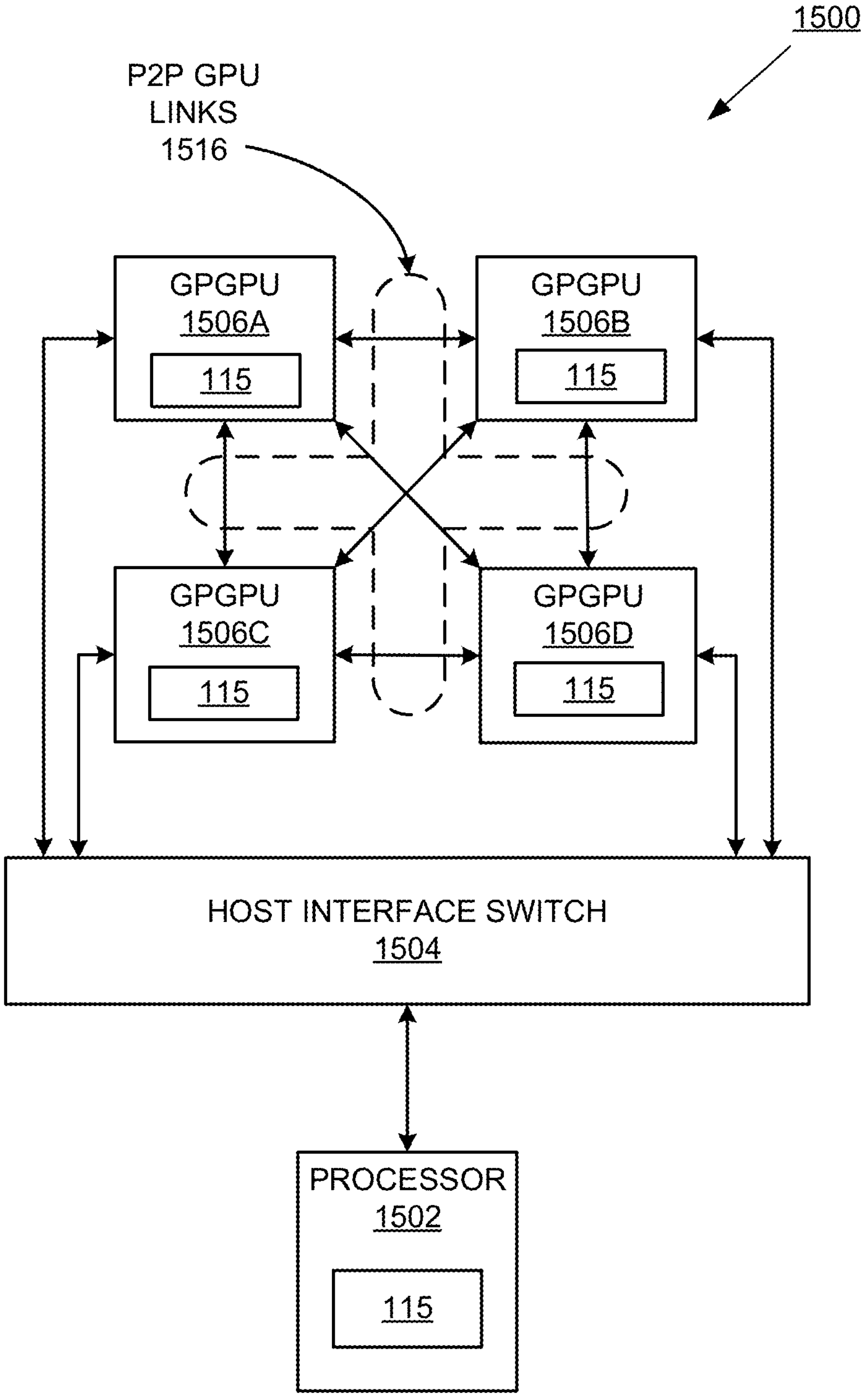

1500
P2P GPU LINKS 1516
GPGPU 1506A
115
GPGPU 1506B
115
GPGPU 1506C
115
GPGPU 1506D
115
HOST INTERFACE SWITCH 1504
PROCESSOR 1502
115

FIG. 15

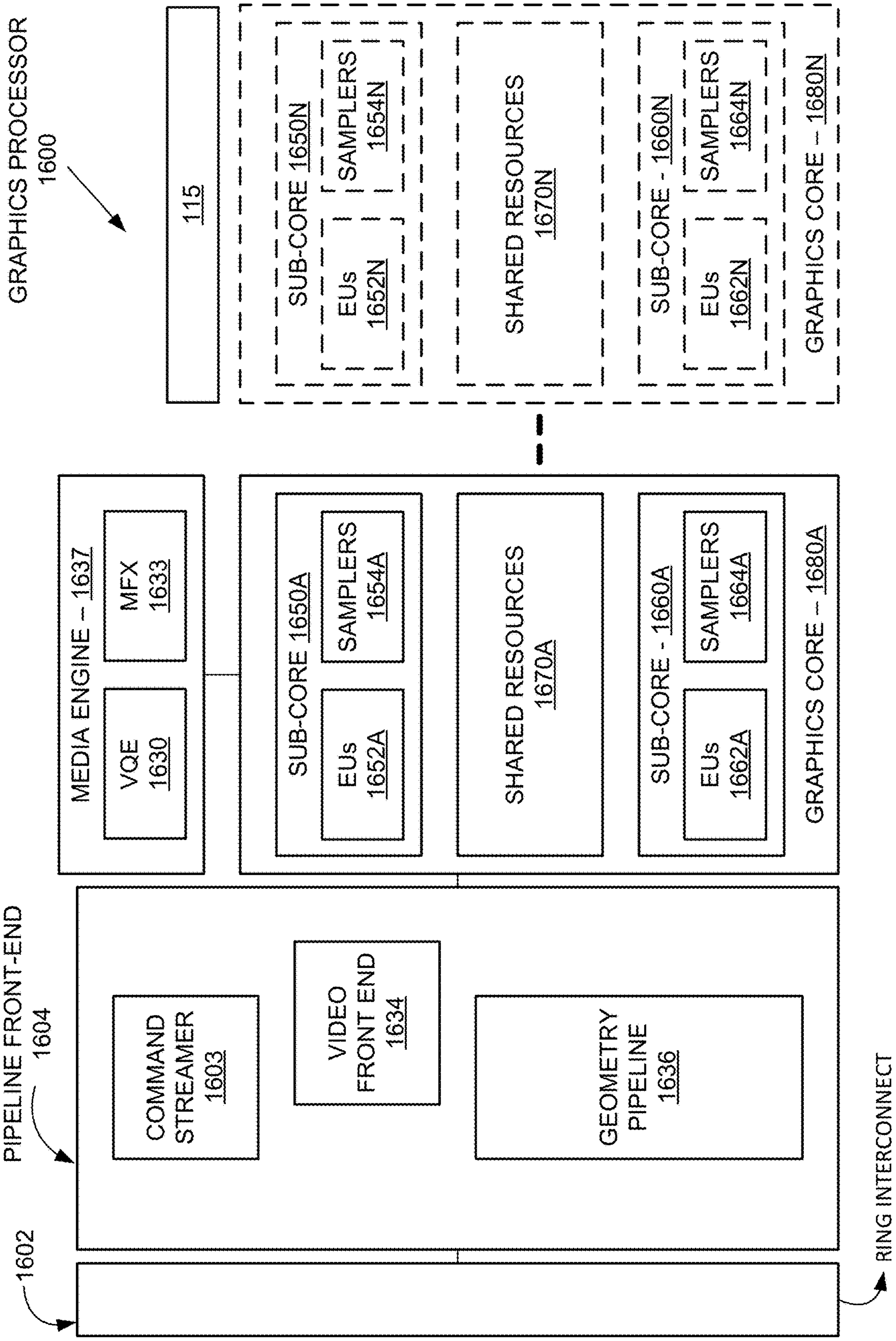

GRAPHICS PROCESSOR
1600
115
SUB-CORE 1650N
EUs
1652N
SAMPLERS
1654N
SHARED RESOURCES
1670N
SUB-CORE - 1660N
EUs
1662N
SAMPLERS
1664N
GRAPHICS CORE – 1680N
MEDIA ENGINE – 1637
VQE
1630
MFX
1633
SUB-CORE 1650A
EUs
1652A
SAMPLERS
1654A
SHARED RESOURCES
1670A
SUB-CORE - 1660A
EUs
1662A
SAMPLERS
1664A
GRAPHICS CORE – 1680A
PIPELINE FRONT-END
1604
COMMAND
STREAMER
1603
VIDEO
FRONT END
1634
GEOMETRY
PIPELINE
1636
1602
RING INTERCONNECT

FIG. 16

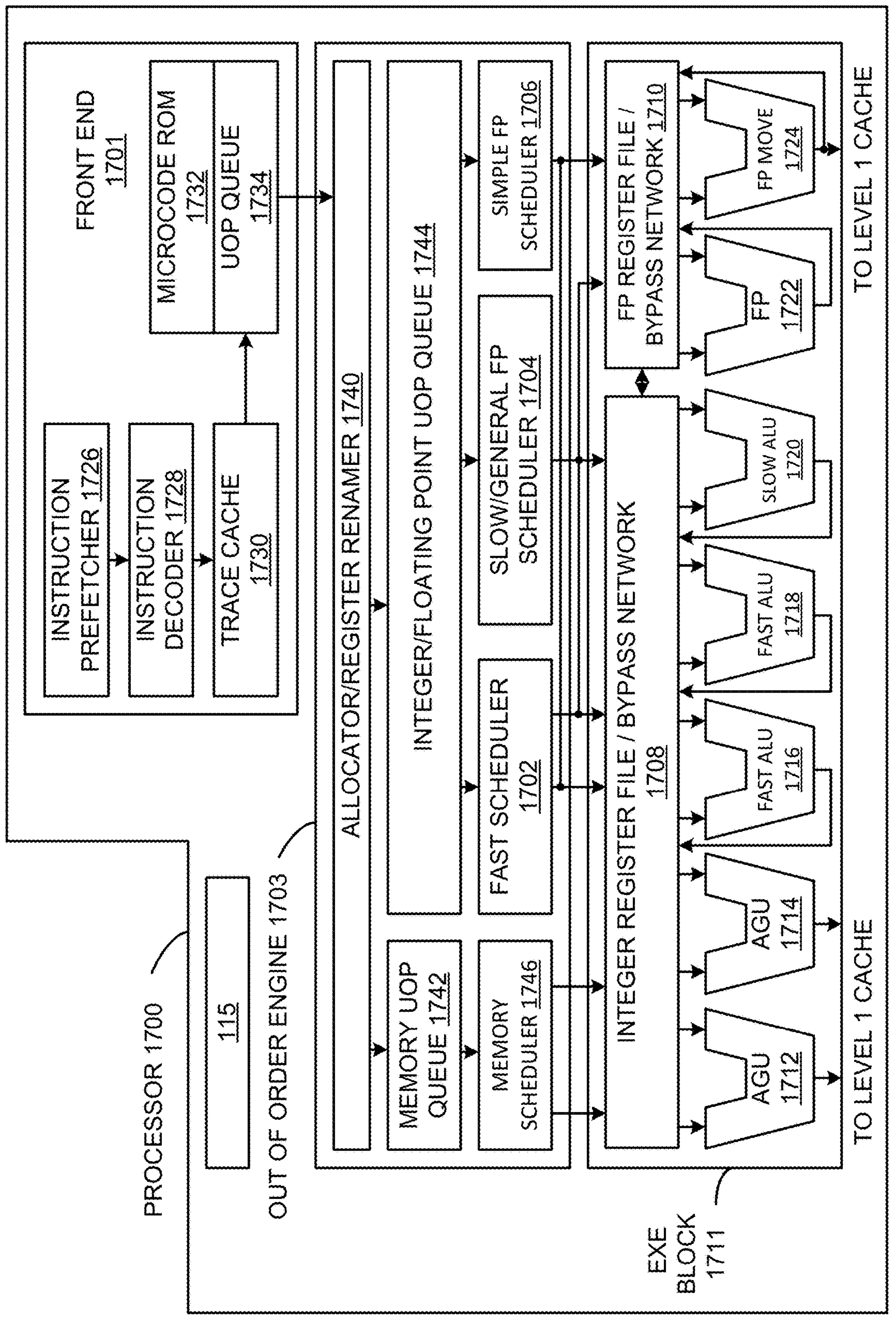

PROCESSOR 1700
115
FRONT END 1701
INSTRUCTION PREFETCHER 1726
INSTRUCTION DECODER 1728
TRACE CACHE 1730
MICROCODE ROM 1732
UOP QUEUE 1734
OUT OF ORDER ENGINE 1703
ALLOCATOR/REGISTER RENAMER 1740
MEMORY UOP QUEUE 1742
INTEGER/FLOATING POINT UOP QUEUE 1744
MEMORY SCHEDULER 1746
FAST SCHEDULER 1702
SLOW/GENERAL FP SCHEDULER 1704
SIMPLE FP SCHEDULER 1706
INTEGER REGISTER FILE / BYPASS NETWORK 1708
FP REGISTER FILE / BYPASS NETWORK 1710
AGU 1712
AGU 1714
FAST ALU 1716
FAST ALU 1718
SLOW ALU 1720
FP 1722
FP MOVE 1724
TO LEVEL 1 CACHE
TO LEVEL 1 CACHE
EXE BLOCK 1711

FIG. 17

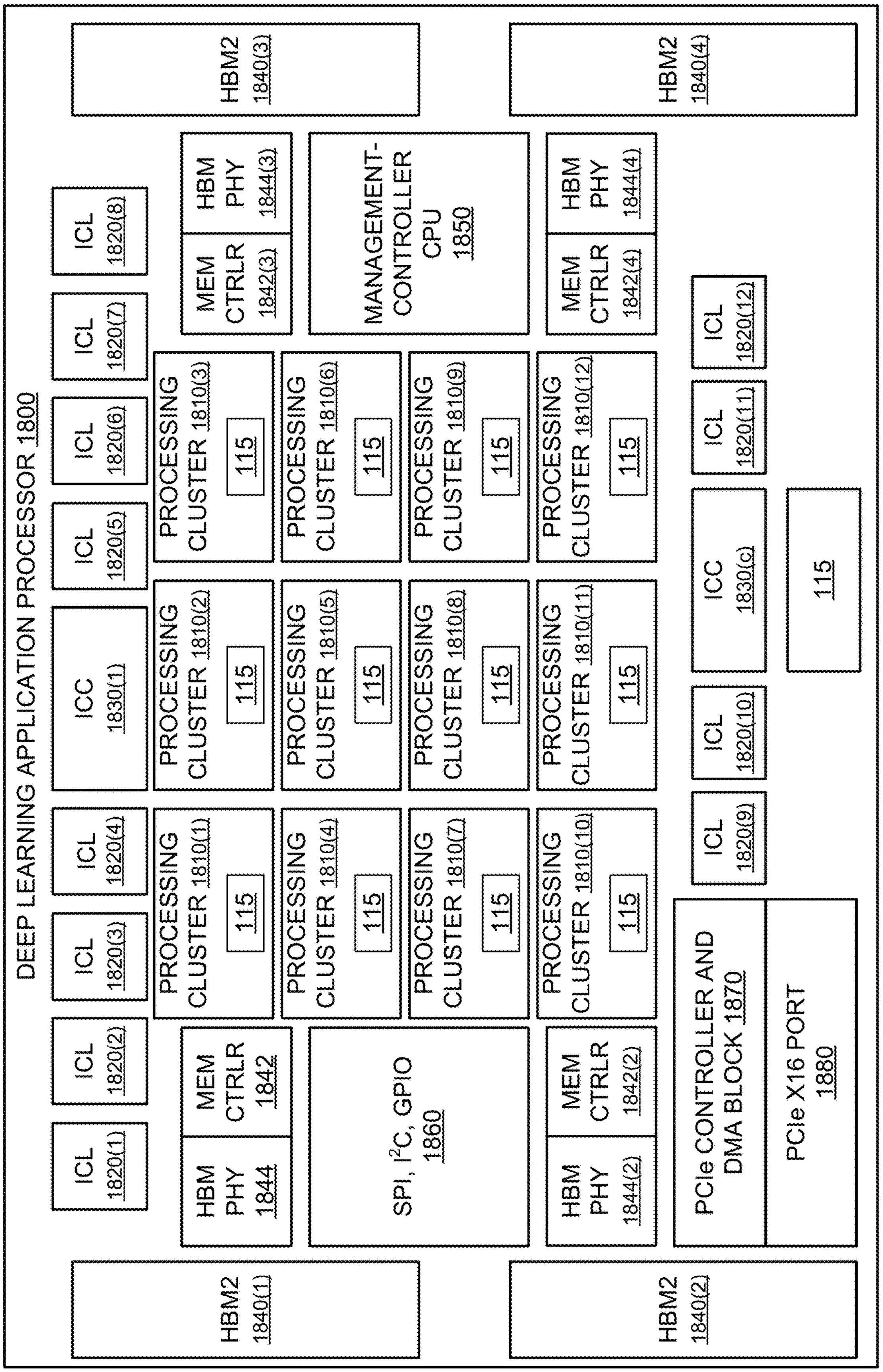

DEEP LEARNING APPLICATION PROCESSOR 1800
HBM2 1840(1)
HBM2 1840(2)
HBM2 1840(3)
HBM2 1840(4)
ICL 1820(1)
ICL 1820(2)
ICL 1820(3)
ICL 1820(4)
ICC 1830(1)
ICL 1820(5)
ICL 1820(6)
ICL 1820(7)
ICL 1820(8)
HBM PHY 1844
MEM CTRLR 1842
HBM PHY 1844(2)
MEM CTRLR 1842(2)
MEM CTRLR 1842(3)
HBM PHY 1844(3)
MEM CTRLR 1842(4)
HBM PHY 1844(4)
SPI, I2C, GPIO 1860
MANAGEMENT-CONTROLLER CPU 1850
PROCESSING CLUSTER 1810(1) 115
PROCESSING CLUSTER 1810(2) 115
PROCESSING CLUSTER 1810(3) 115
PROCESSING CLUSTER 1810(4) 115
PROCESSING CLUSTER 1810(5) 115
PROCESSING CLUSTER 1810(6) 115
PROCESSING CLUSTER 1810(7) 115
PROCESSING CLUSTER 1810(8) 115
PROCESSING CLUSTER 1810(9) 115
PROCESSING CLUSTER 1810(10) 115
PROCESSING CLUSTER 1810(11) 115
PROCESSING CLUSTER 1810(12) 115
PCIe CONTROLLER AND DMA BLOCK 1870
PCIe X16 PORT 1880
ICL 1820(9)
ICL 1820(10)
ICC 1830(c)
115
ICL 1820(11)
ICL 1820(12)

FIG. 18

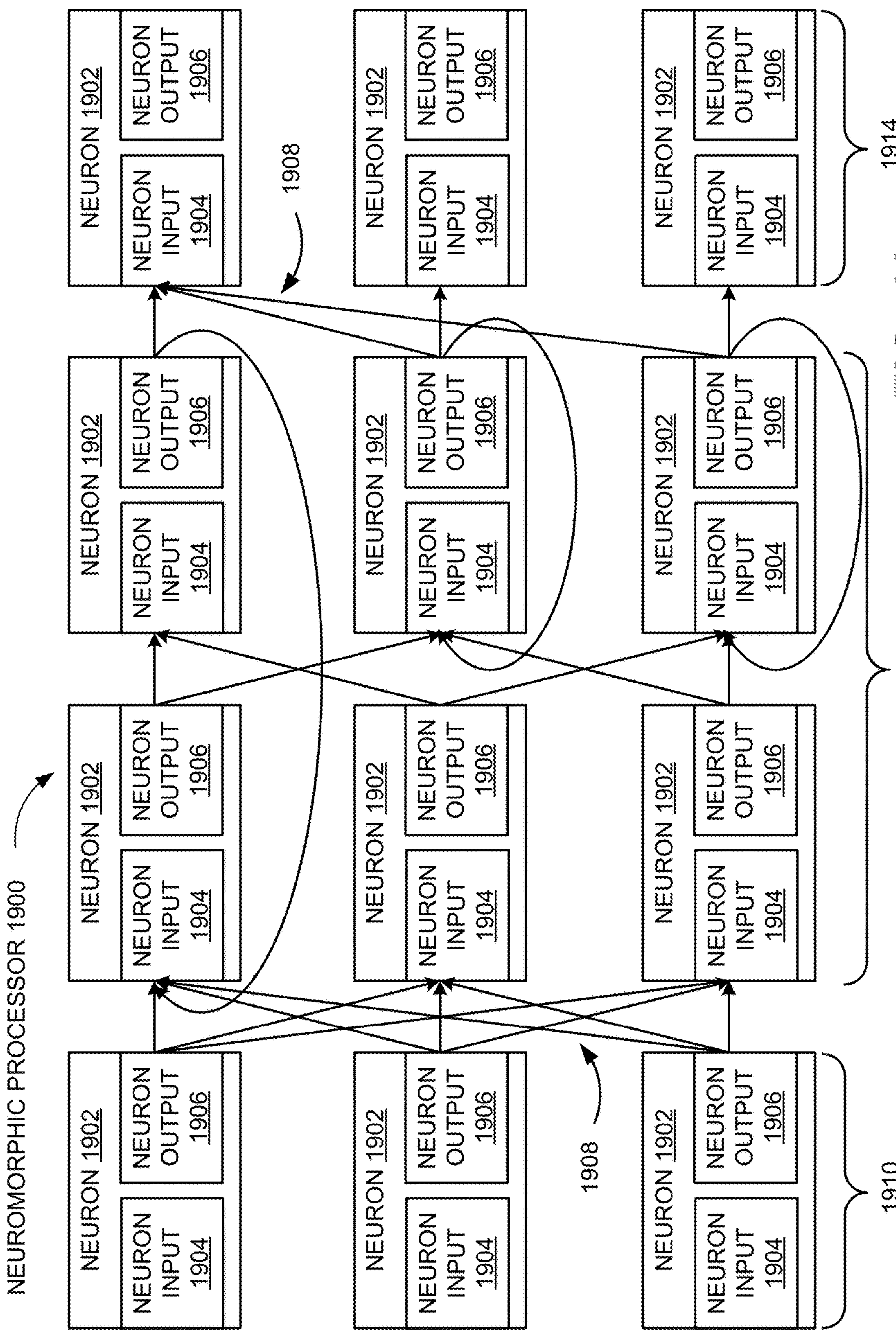

NEUROMORPHIC PROCESSOR 1900
NEURON 1902
NEURON INPUT 1904
NEURON OUTPUT 1906
1908
1910
1912
1914

FIG. 19

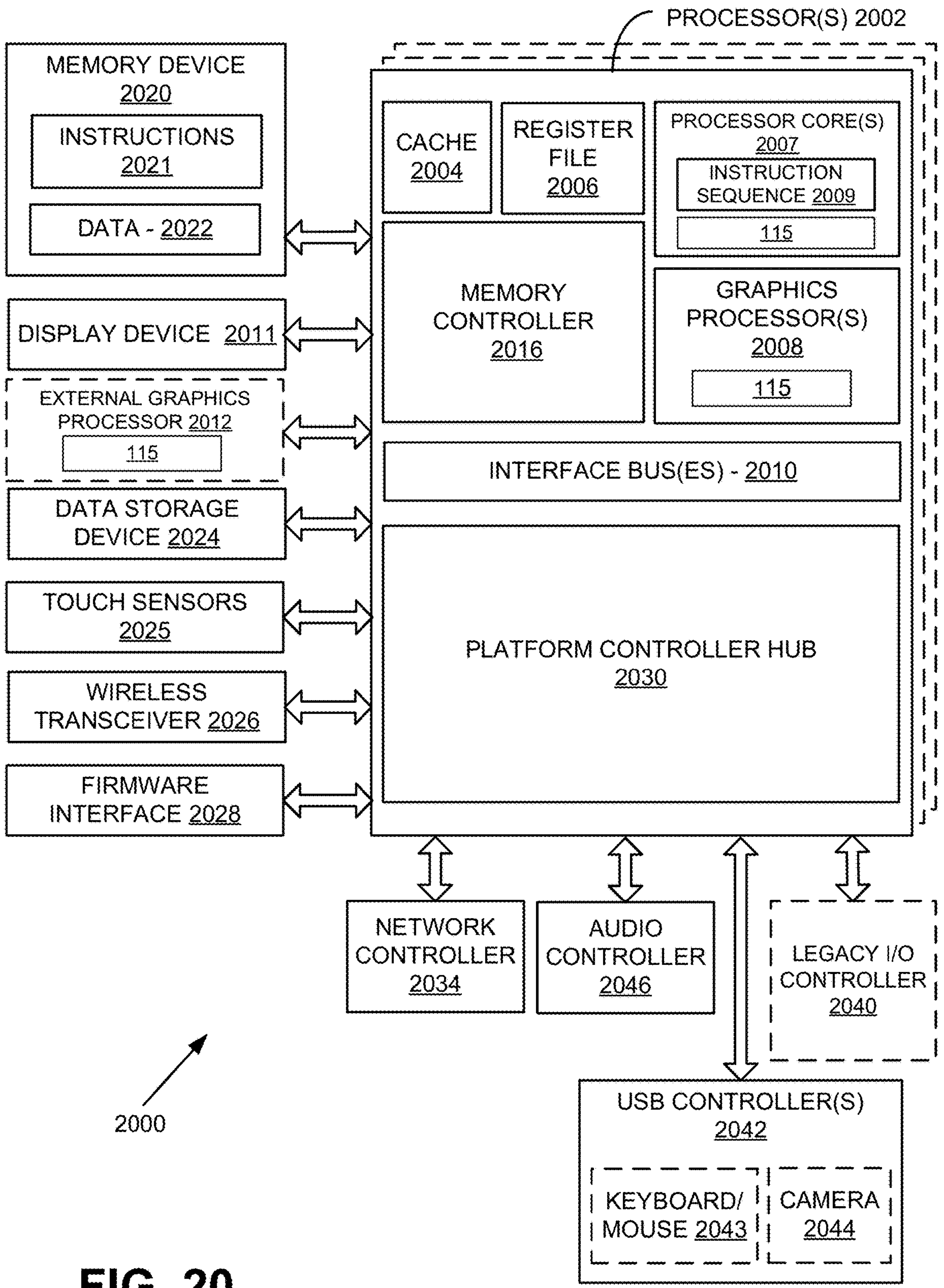

PROCESSOR(S) 2002
MEMORY DEVICE 2020
INSTRUCTIONS 2021
DATA - 2022
DISPLAY DEVICE 2011
EXTERNAL GRAPHICS PROCESSOR 2012
115
DATA STORAGE DEVICE 2024
TOUCH SENSORS 2025
WIRELESS TRANSCEIVER 2026
FIRMWARE INTERFACE 2028
CACHE 2004
REGISTER FILE 2006
PROCESSOR CORE(S) 2007
INSTRUCTION SEQUENCE 2009
115
MEMORY CONTROLLER 2016
GRAPHICS PROCESSOR(S) 2008
115
INTERFACE BUS(ES) - 2010
PLATFORM CONTROLLER HUB 2030
NETWORK CONTROLLER 2034
AUDIO CONTROLLER 2046
LEGACY I/O CONTROLLER 2040
USB CONTROLLER(S) 2042
KEYBOARD/ MOUSE 2043
CAMERA 2044
2000

FIG. 20

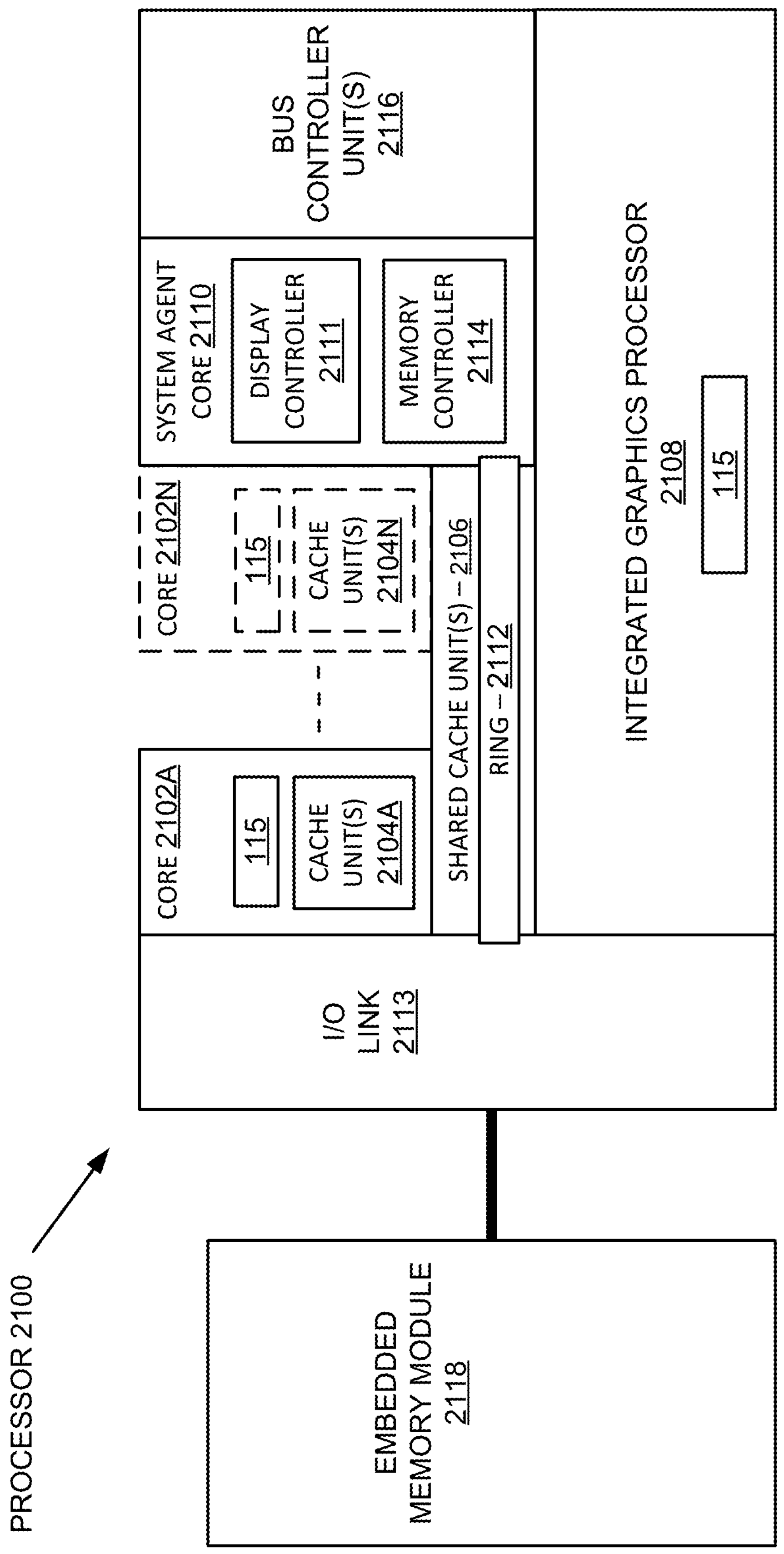

PROCESSOR 2100
EMBEDDED MEMORY MODULE 2118
I/O LINK 2113
CORE 2102A
115
CACHE UNIT(S) 2104A
CORE 2102N
115
CACHE UNIT(S) 2104N
SHARED CACHE UNIT(S) – 2106
RING – 2112
SYSTEM AGENT CORE 2110
DISPLAY CONTROLLER 2111
MEMORY CONTROLLER 2114
BUS CONTROLLER UNIT(S) 2116
INTEGRATED GRAPHICS PROCESSOR 2108
115

FIG. 21

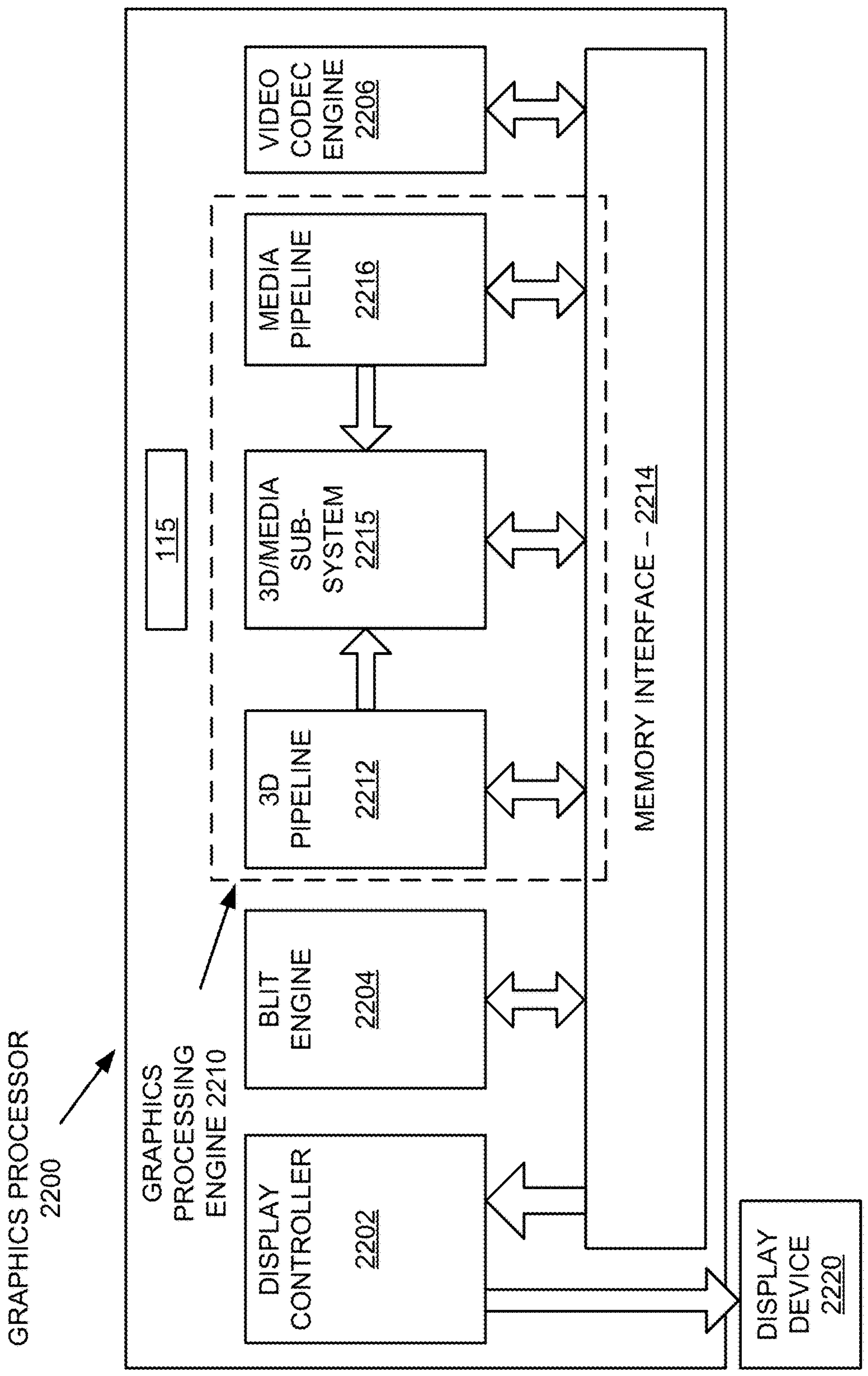

GRAPHICS PROCESSOR 2200
GRAPHICS PROCESSING ENGINE 2210
115
DISPLAY CONTROLLER 2202
BLIT ENGINE 2204
3D PIPELINE 2212
3D/MEDIA SUB-SYSTEM 2215
MEDIA PIPELINE 2216
VIDEO CODEC ENGINE 2206
MEMORY INTERFACE – 2214
DISPLAY DEVICE 2220

FIG. 22

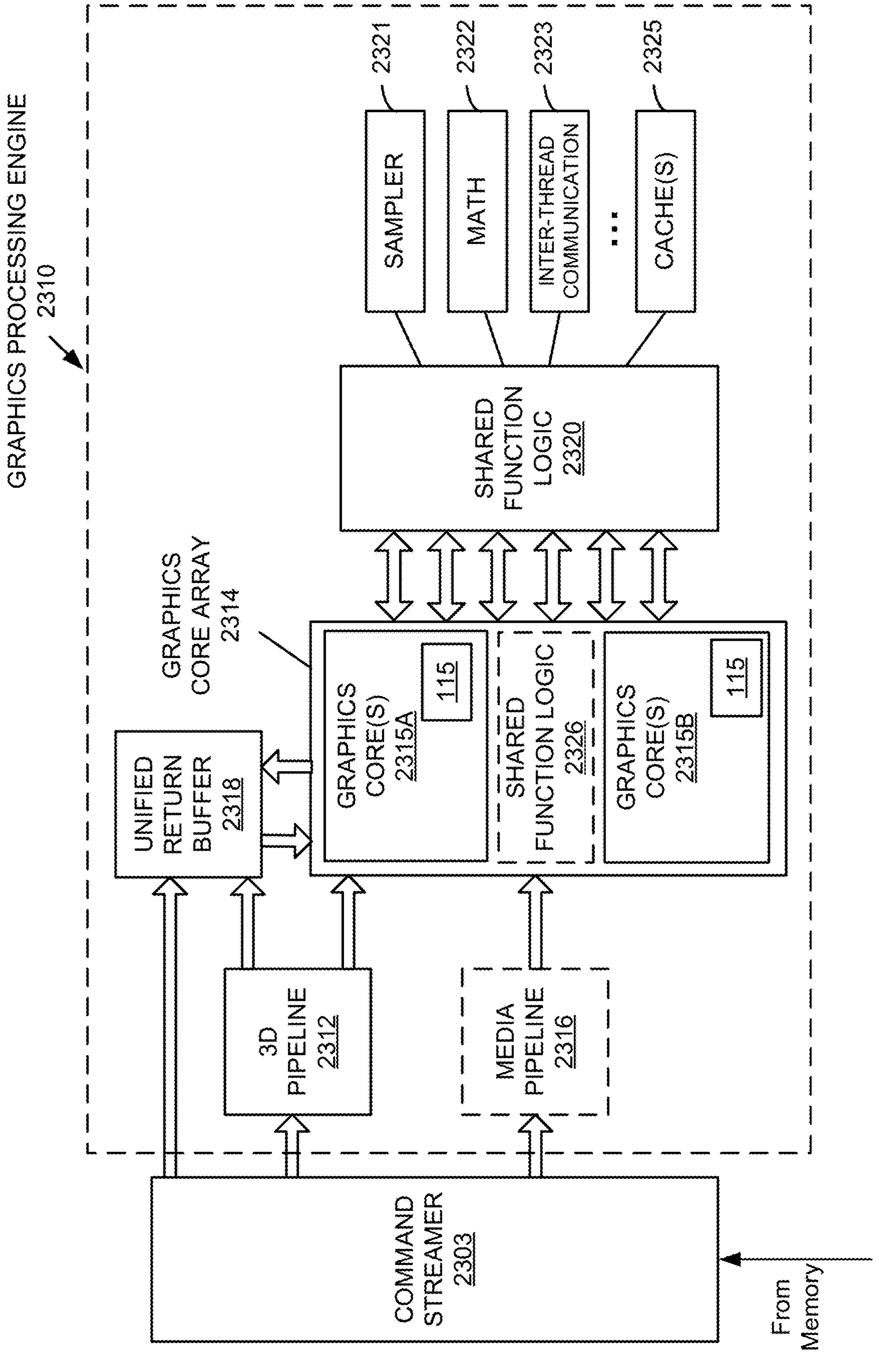

GRAPHICS PROCESSING ENGINE
2310
SAMPLER
MATH
INTER-THREAD COMMUNICATION
CACHE(S)
2321
2322
2323
2325
SHARED FUNCTION LOGIC
2320
GRAPHICS CORE ARRAY
2314
UNIFIED RETURN BUFFER
2318
GRAPHICS CORE(S)
2315A
115
SHARED FUNCTION LOGIC
2326
GRAPHICS CORE(S)
2315B
115
3D PIPELINE
2312
MEDIA PIPELINE
2316
COMMAND STREAMER
2303
From Memory

FIG. 23

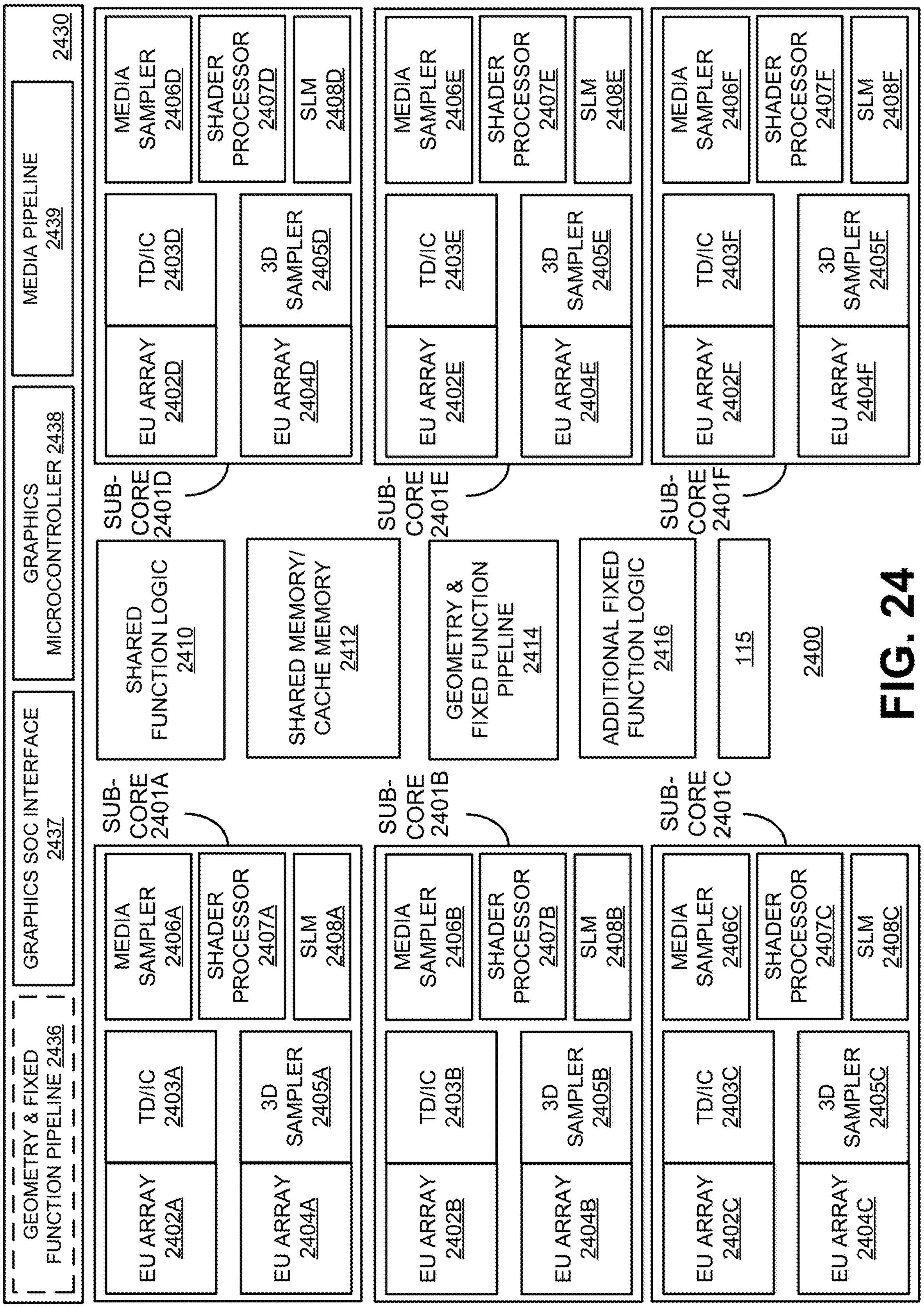

GEOMETRY & FIXED FUNCTION PIPELINE 2436
GRAPHICS SOC INTERFACE 2437
GRAPHICS MICROCONTROLLER 2438
MEDIA PIPELINE 2439
2430
SUB-CORE 2401A
SUB-CORE 2401B
SUB-CORE 2401C
SUB-CORE 2401D
SUB-CORE 2401E
SUB-CORE 2401F
EU ARRAY 2402A
TD/IC 2403A
MEDIA SAMPLER 2406A
EU ARRAY 2404A
3D SAMPLER 2405A
SHADER PROCESSOR 2407A
SLM 2408A
EU ARRAY 2402B
TD/IC 2403B
MEDIA SAMPLER 2406B
EU ARRAY 2404B
3D SAMPLER 2405B
SHADER PROCESSOR 2407B
SLM 2408B
EU ARRAY 2402C
TD/IC 2403C
MEDIA SAMPLER 2406C
EU ARRAY 2404C
3D SAMPLER 2405C
SHADER PROCESSOR 2407C
SLM 2408C
EU ARRAY 2402D
TD/IC 2403D
MEDIA SAMPLER 2406D
EU ARRAY 2404D
3D SAMPLER 2405D
SHADER PROCESSOR 2407D
SLM 2408D
EU ARRAY 2402E
TD/IC 2403E
MEDIA SAMPLER 2406E
EU ARRAY 2404E
3D SAMPLER 2405E
SHADER PROCESSOR 2407E
SLM 2408E
EU ARRAY 2402F
TD/IC 2403F
MEDIA SAMPLER 2406F
EU ARRAY 2404F
3D SAMPLER 2405F
SHADER PROCESSOR 2407F
SLM 2408F
SHARED FUNCTION LOGIC 2410
SHARED MEMORY/ CACHE MEMORY 2412
GEOMETRY & FIXED FUNCTION PIPELINE 2414
ADDITIONAL FIXED FUNCTION LOGIC 2416
115
2400

FIG. 24

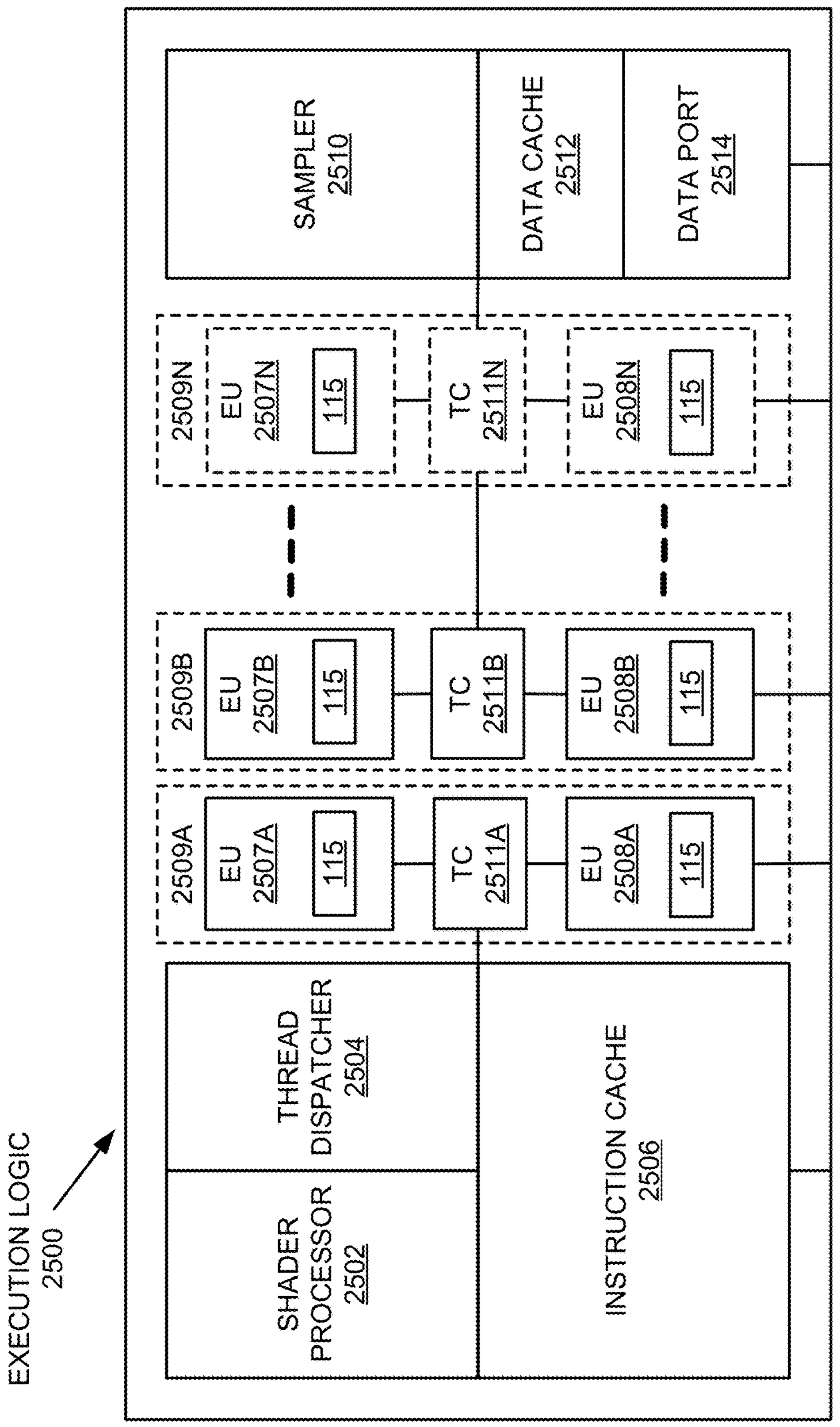

EXECUTION LOGIC
2500
SHADER PROCESSOR
2502
THREAD DISPATCHER
2504
INSTRUCTION CACHE
2506
2509A
EU
2507A
115
TC
2511A
EU
2508A
115
2509B
EU
2507B
115
TC
2511B
EU
2508B
115
2509N
EU
2507N
115
TC
2511N
EU
2508N
115
SAMPLER
2510
DATA CACHE
2512
DATA PORT
2514

FIG. 25 A

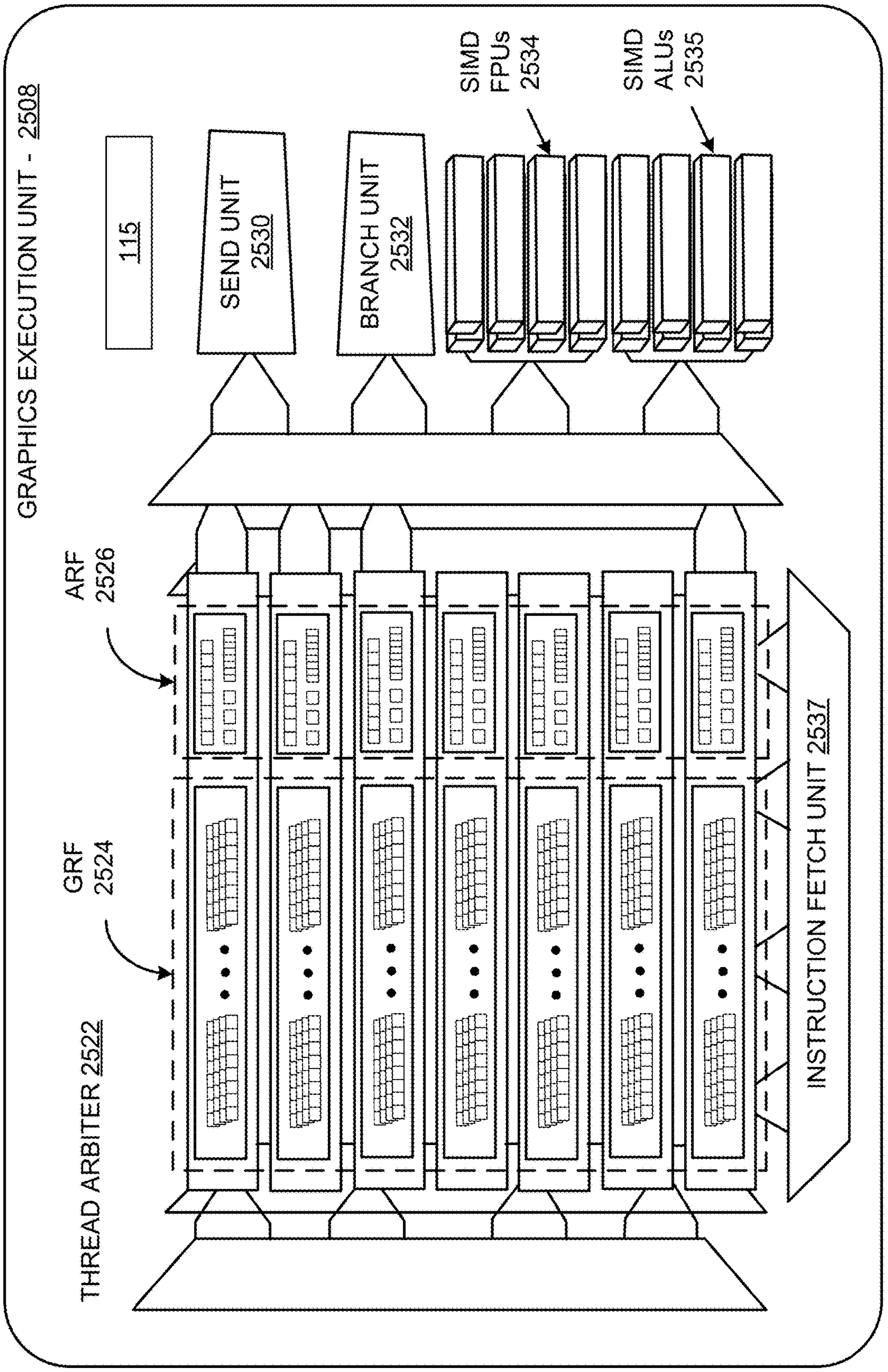

GRAPHICS EXECUTION UNIT - 2508
115
SEND UNIT 2530
BRANCH UNIT 2532
SIMD FPUs 2534
SIMD ALUs 2535
ARF 2526
GRF 2524
THREAD ARBITER 2522
INSTRUCTION FETCH UNIT 2537

FIG. 25 B

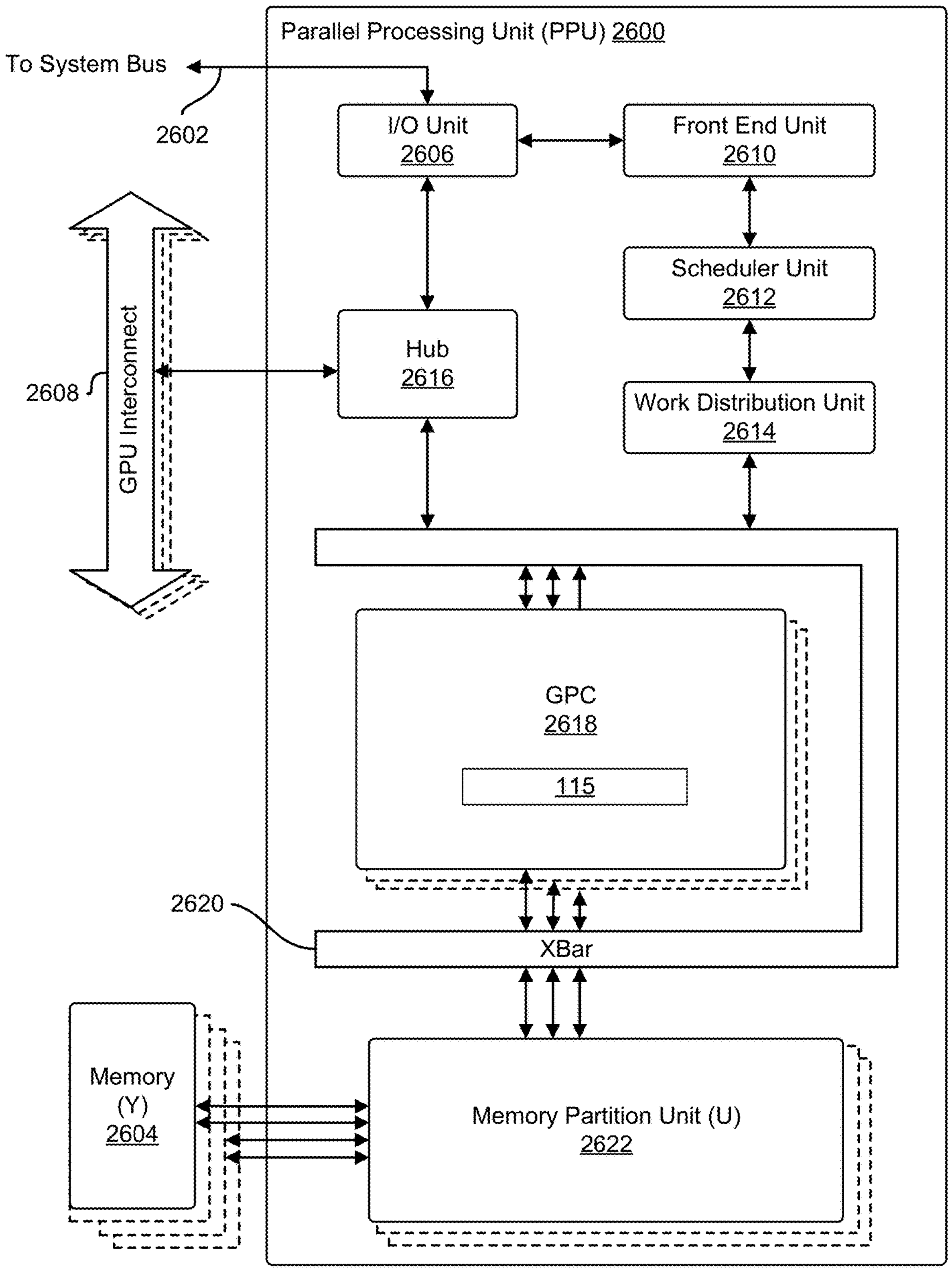

Parallel Processing Unit (PPU) 2600
To System Bus
2602
I/O Unit 2606
Front End Unit 2610
Scheduler Unit 2612
Work Distribution Unit 2614
Hub 2616
GPU Interconnect
2608
GPC 2618
115
2620
XBar
Memory (Y) 2604
Memory Partition Unit (U) 2622

FIG. 26

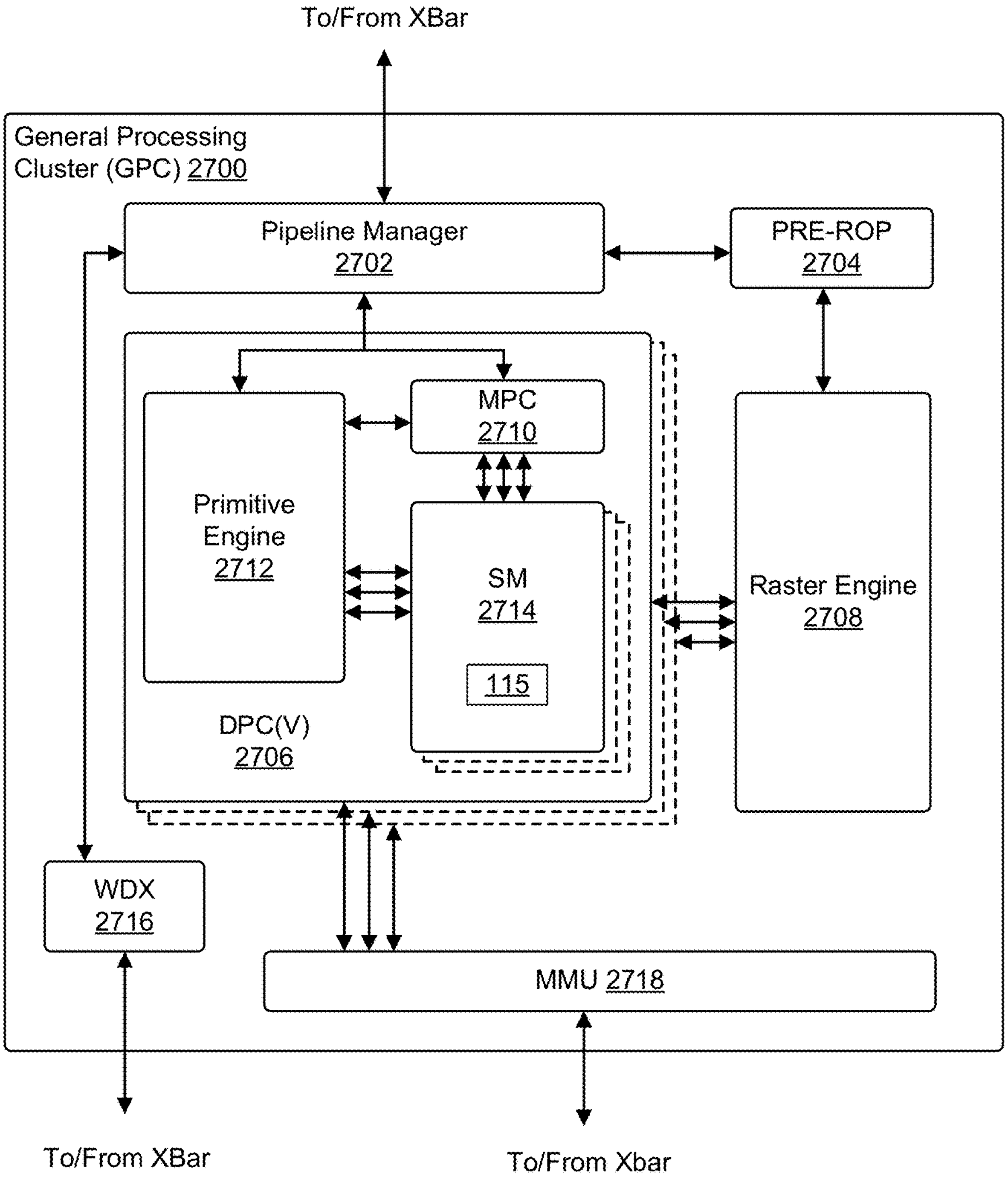

To/From XBar
General Processing Cluster (GPC) 2700
Pipeline Manager 2702
PRE-ROP 2704
MPC 2710
Primitive Engine 2712
SM 2714
115
DPC(V) 2706
Raster Engine 2708
WDX 2716
MMU 2718
To/From XBar
To/From Xbar

FIG. 27

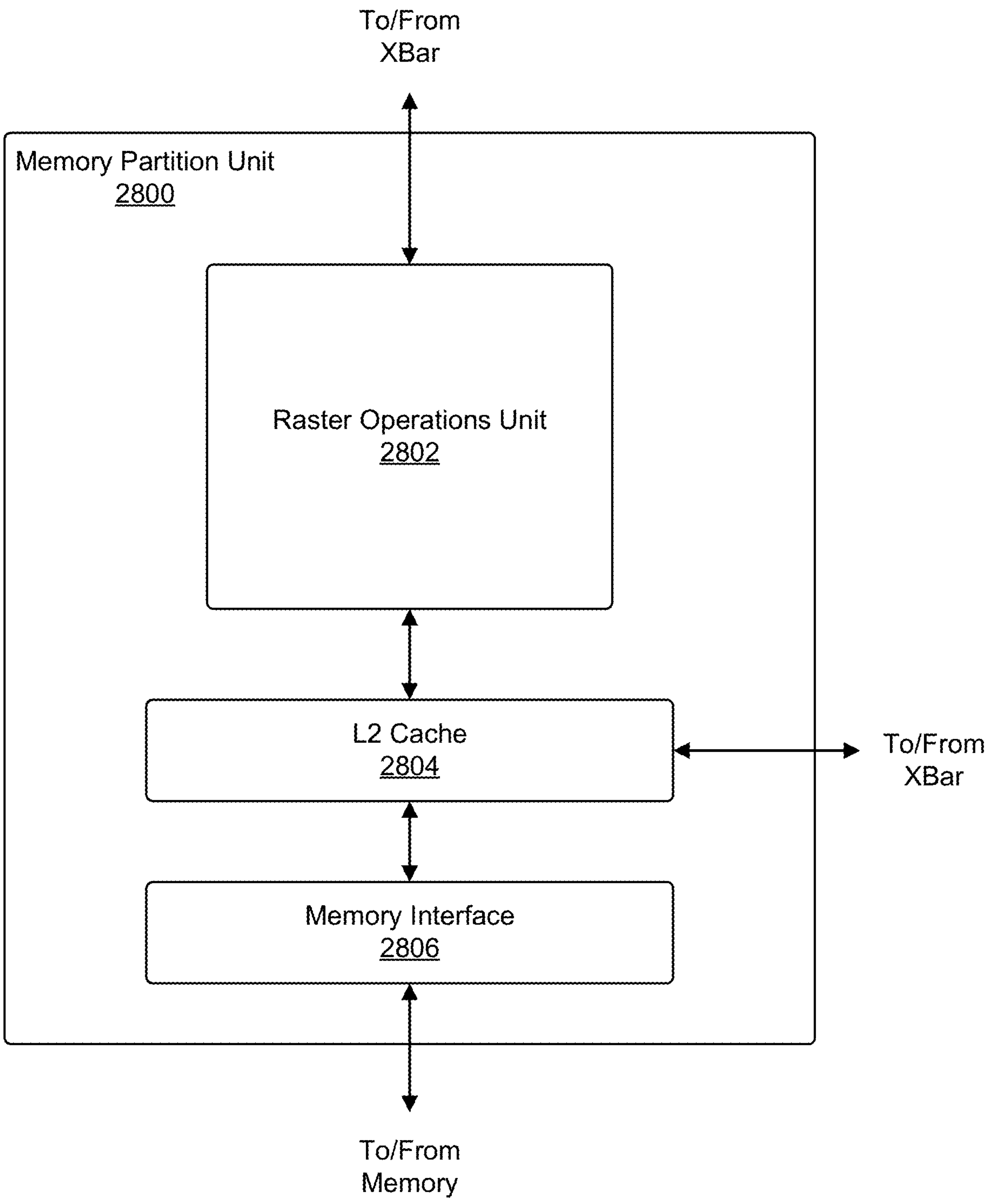

To/From XBar
Memory Partition Unit 2800
Raster Operations Unit 2802
L2 Cache 2804
To/From XBar
Memory Interface 2806
To/From Memory

FIG. 28

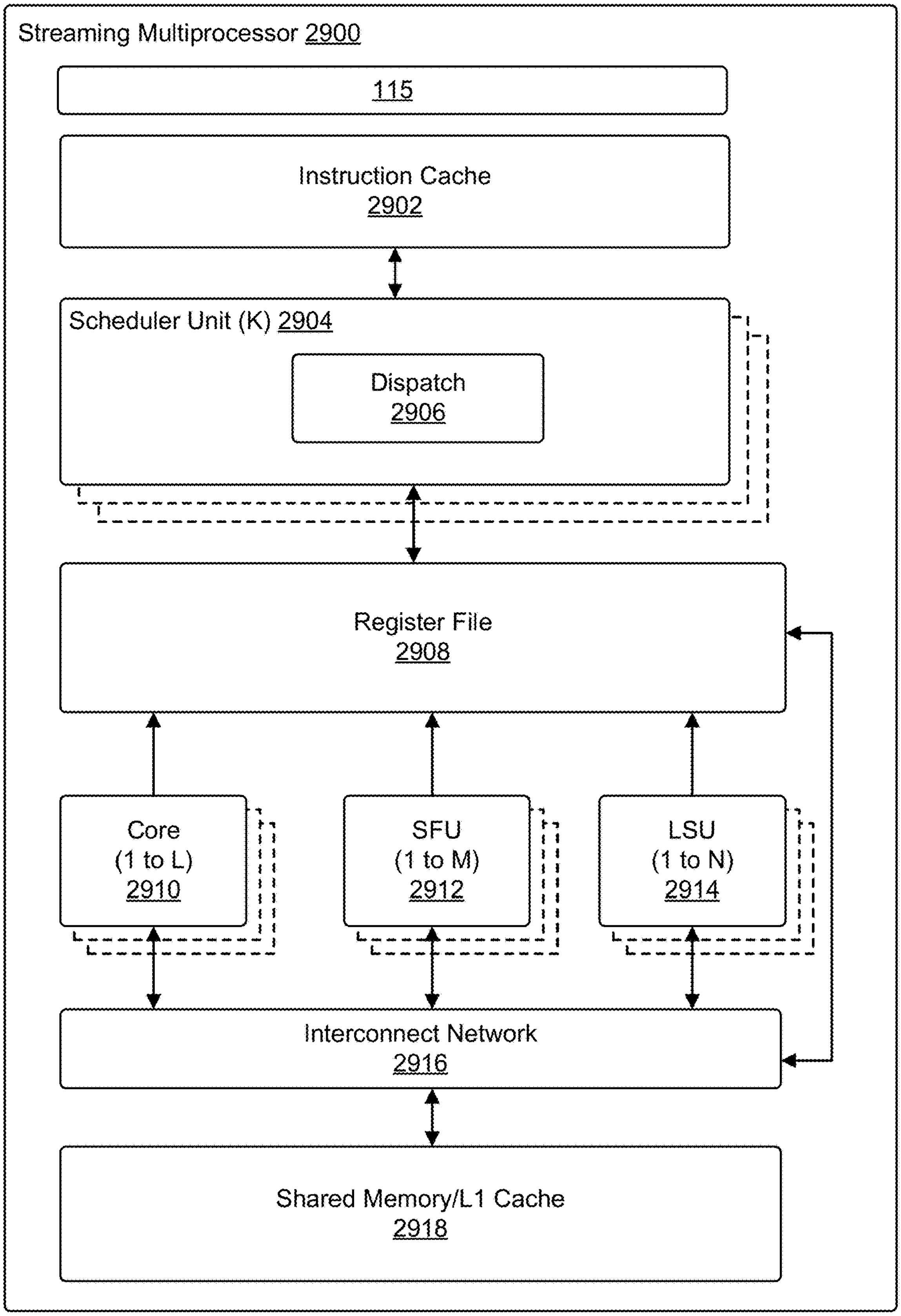

Streaming Multiprocessor 2900
115
Instruction Cache 2902
Scheduler Unit (K) 2904
Dispatch 2906
Register File 2908
Core (1 to L) 2910
SFU (1 to M) 2912
LSU (1 to N) 2914
Interconnect Network 2916
Shared Memory/L1 Cache 2918

FIG. 29

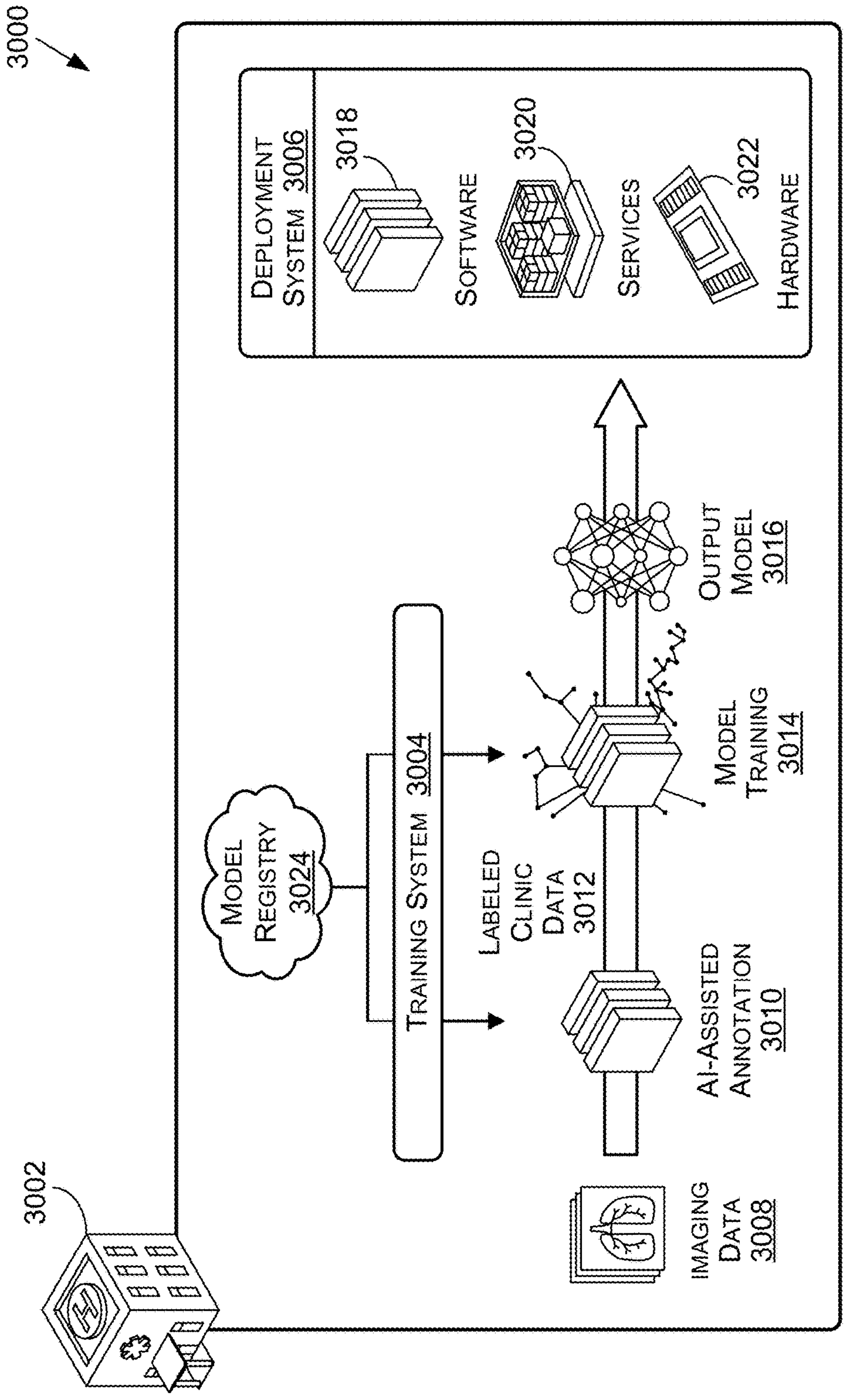

3000
3002
Imaging Data 3008
Training System 3004
Model Registry 3024
AI-Assisted Annotation 3010
Labeled Clinic Data 3012
Model Training 3014
Output Model 3016
Deployment System 3006
3018
Software
3020
Services
3022
Hardware

FIG. 30

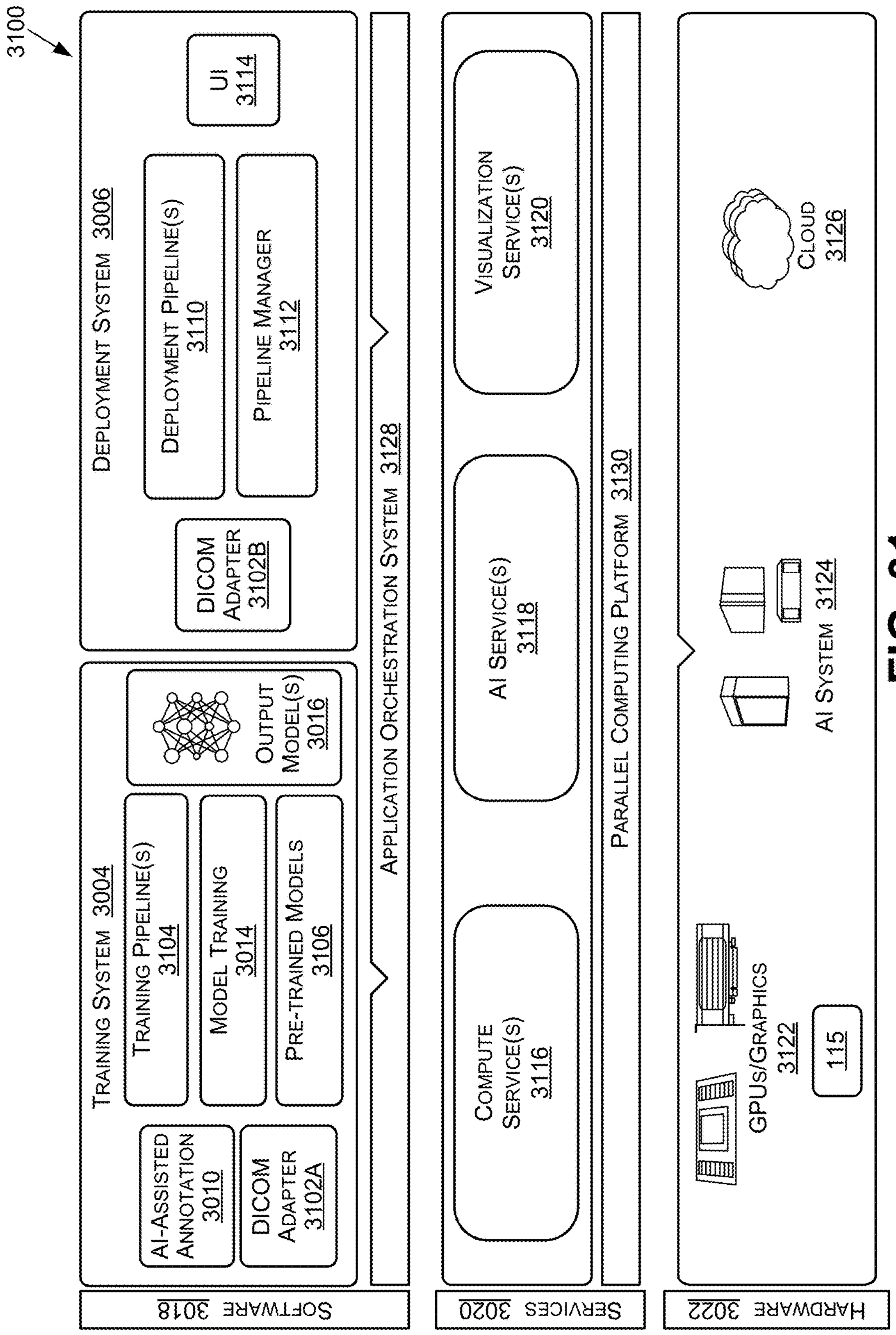

3100
SOFTWARE 3018
TRAINING SYSTEM 3004
AI-ASSISTED ANNOTATION 3010
DICOM ADAPTER 3102A
TRAINING PIPELINE(S) 3104
MODEL TRAINING 3014
PRE-TRAINED MODELS 3106
OUTPUT MODEL(S) 3016
DEPLOYMENT SYSTEM 3006
DICOM ADAPTER 3102B
DEPLOYMENT PIPELINE(S) 3110
PIPELINE MANAGER 3112
UI 3114
APPLICATION ORCHESTRATION SYSTEM 3128
SERVICES 3020
COMPUTE SERVICE(S) 3116
AI SERVICE(S) 3118
VISUALIZATION SERVICE(S) 3120
PARALLEL COMPUTING PLATFORM 3130
HARDWARE 3022
GPUs/GRAPHICS 3122
115
AI SYSTEM 3124
CLOUD 3126

FIG. 31

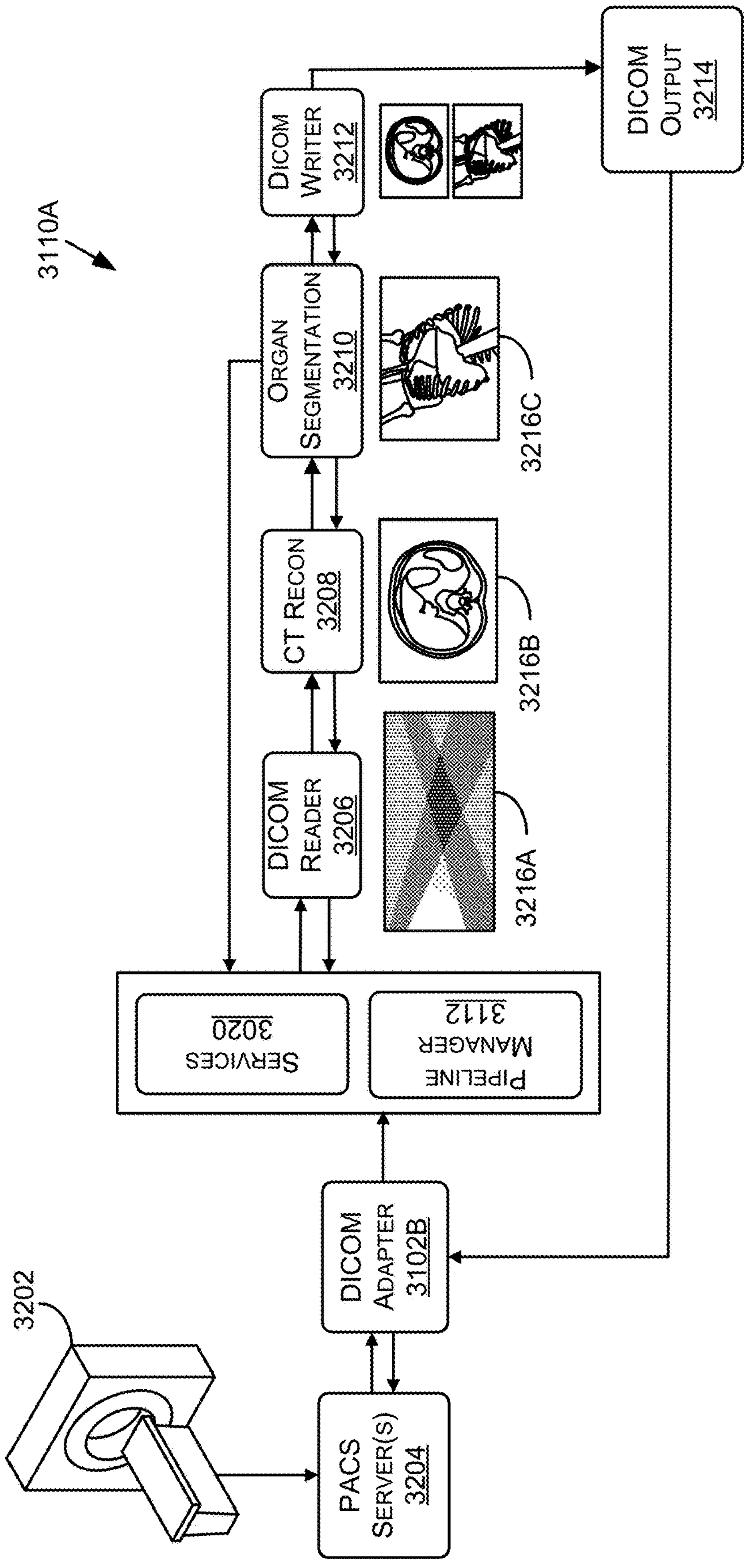

3110A
3202
PACS SERVER(S) 3204
DICOM ADAPTER 3102B
PIPELINE MANAGER 3112
SERVICES 3020
DICOM READER 3206
CT RECON 3208
ORGAN SEGMENTATION 3210
DICOM WRITER 3212
DICOM OUTPUT 3214
3216A
3216B
3216C

FIG. 32

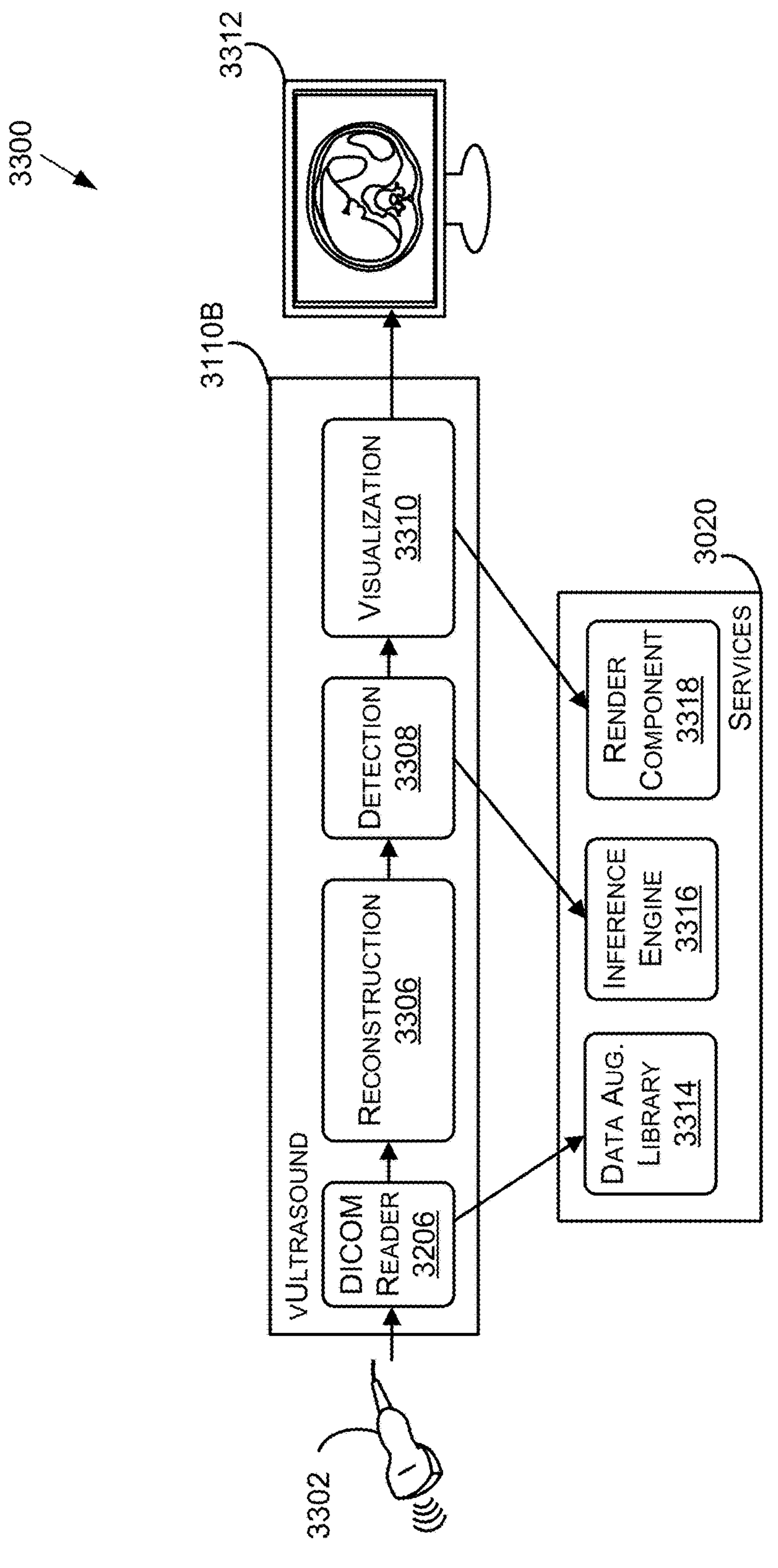

3300
3312
3110B
vUltrasound
DICOM Reader 3206
Reconstruction 3306
Detection 3308
Visualization 3310
3020
Services
Data Aug. Library 3314
Inference Engine 3316
Render Component 3318
3302

FIG. 33A

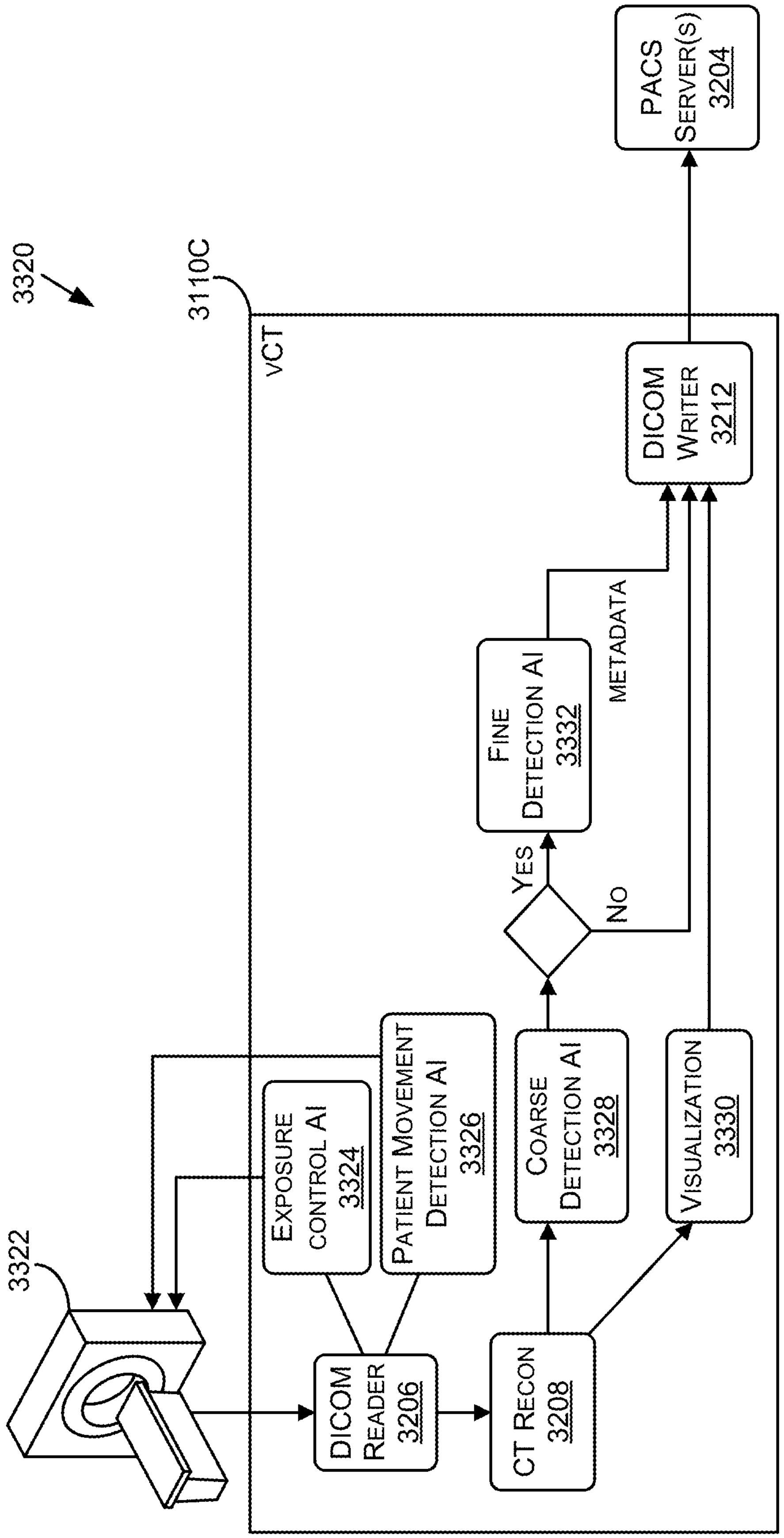

3320
3110C
vCT
3322
DICOM Reader 3206
CT Recon 3208
Exposure Control AI 3324
Patient Movement Detection AI 3326
Coarse Detection AI 3328
Visualization 3330
Yes
No
Fine Detection AI 3332
Metadata
DICOM Writer 3212
PACS Server(s) 3204

FIG. 33B

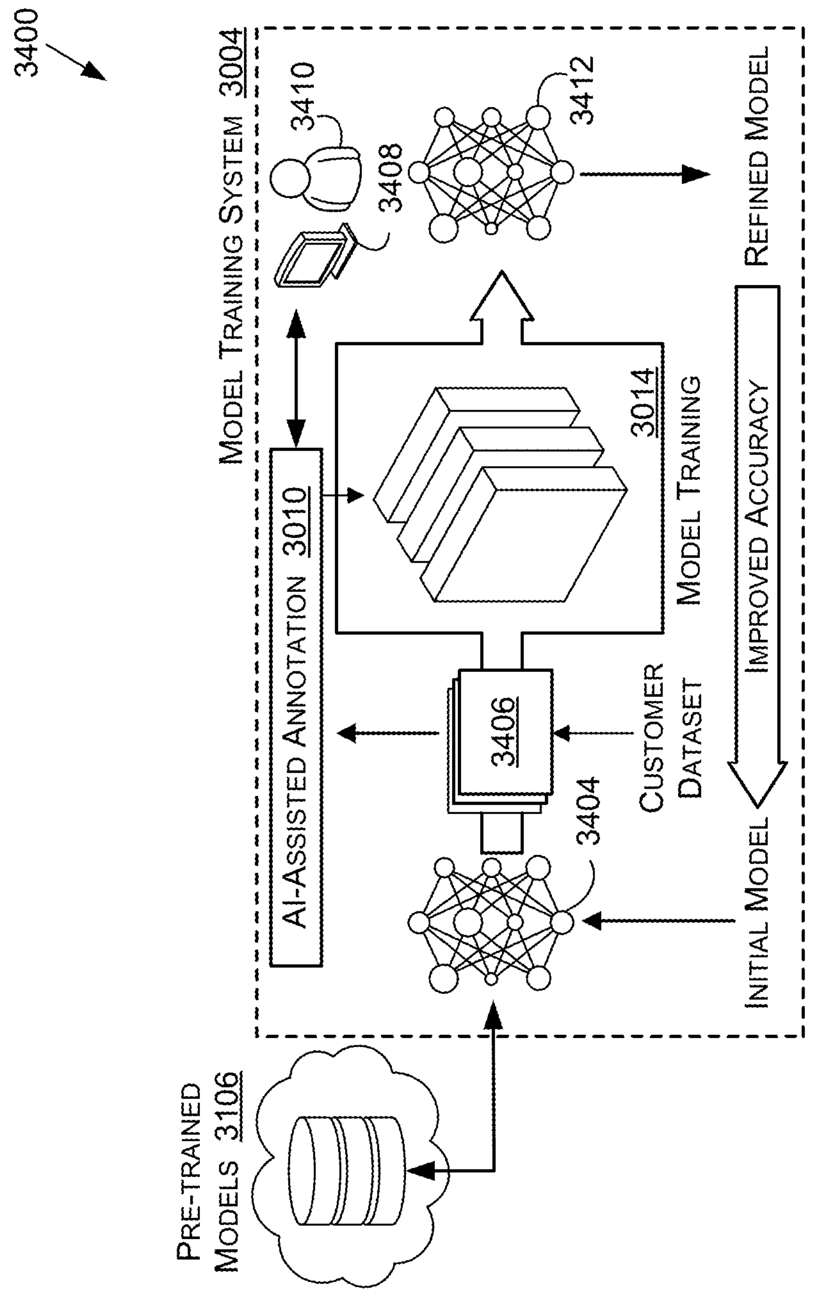

3400
Model Training System 3004
Pre-trained Models 3106
AI-Assisted Annotation 3010
3410
3408
3412
3014
Model Training
3404
3406
Customer Dataset
Initial Model
Improved Accuracy
Refined Model

FIG. 34A

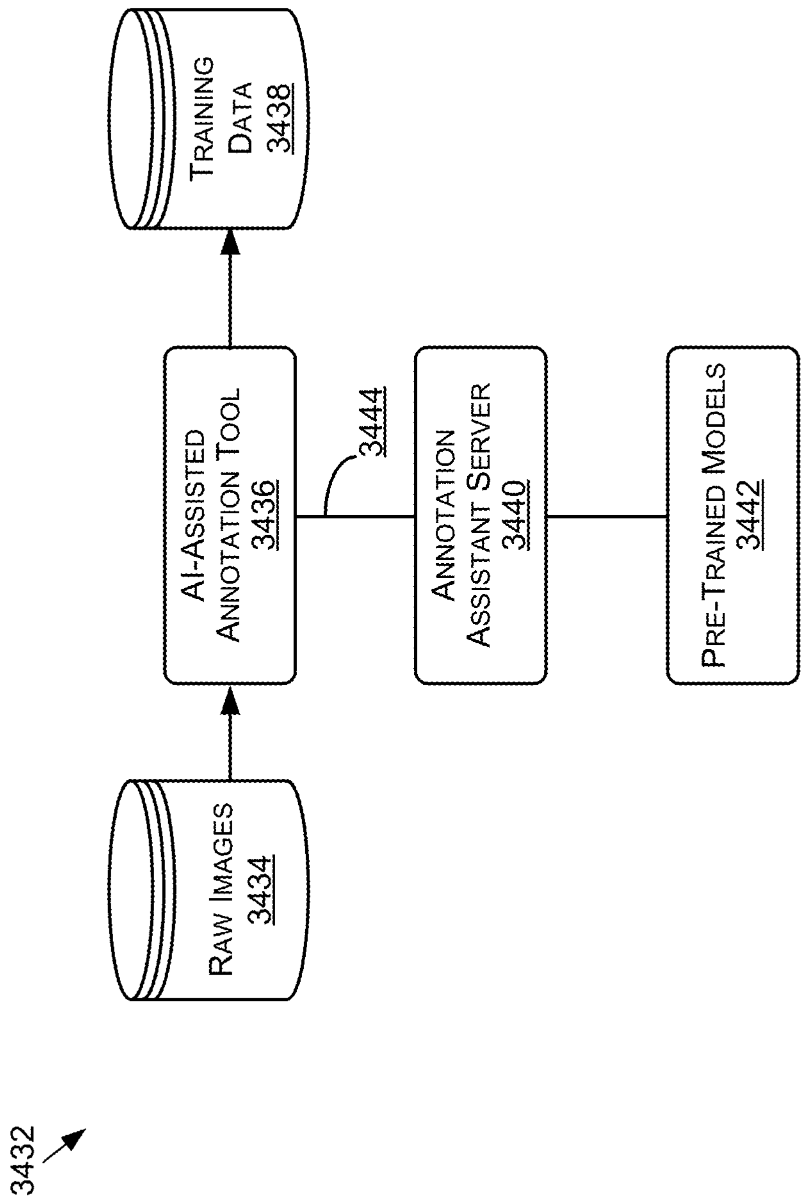

3432
Raw Images 3434
AI-Assisted Annotation Tool 3436
Training Data 3438
3444
Annotation Assistant Server 3440
Pre-Trained Models 3442

FIG. 34B

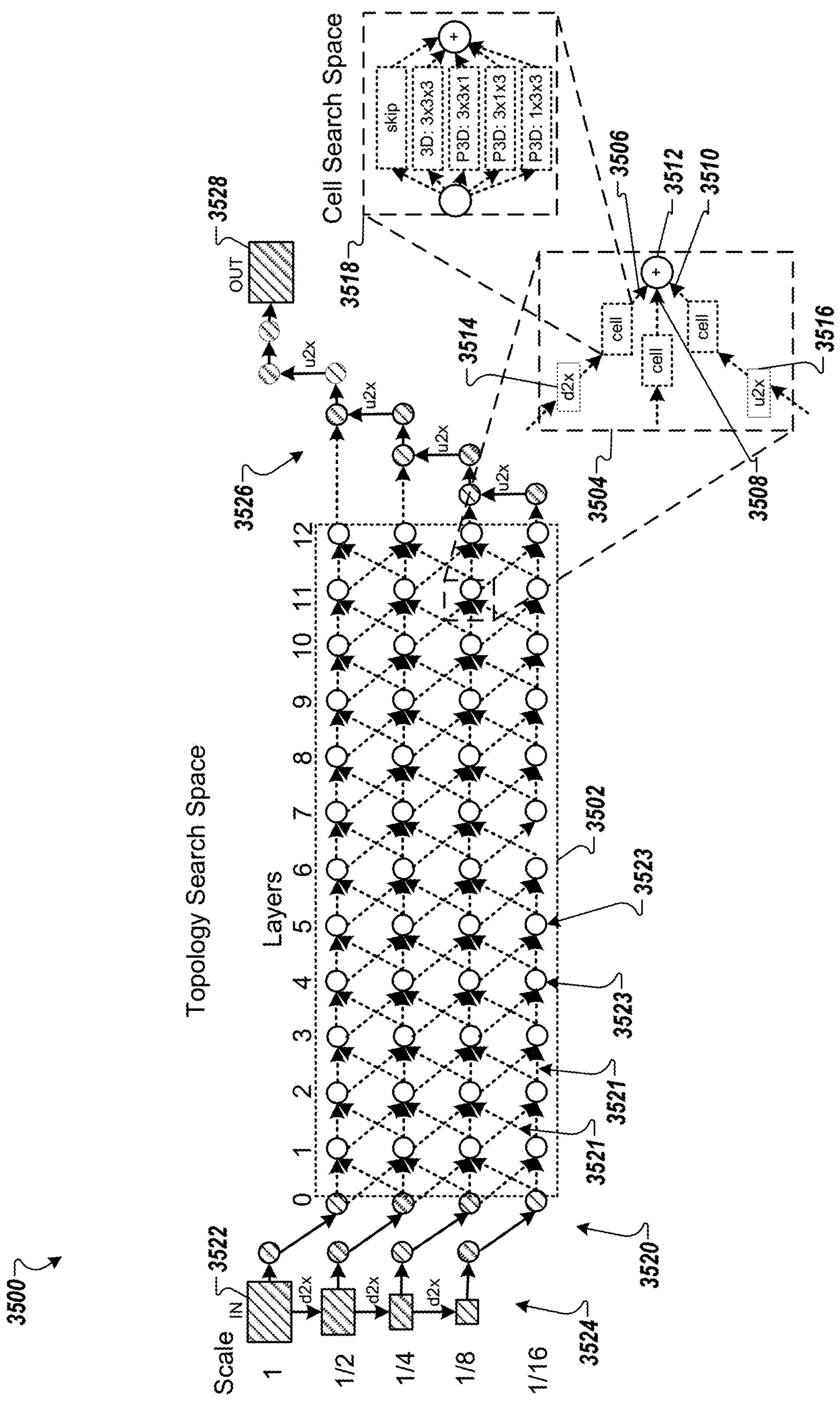

3500
Scale
1
1/2
1/4
1/8
1/16
IN
d2x
d2x
d2x
3522
3524
3520
3521
3521
3523
3523
3502
Topology Search Space
Layers
0
1
2
3
4
5
6
7
8
9
10
11
12
3526
u2x
u2x
u2x
u2x
OUT
3528
3518
Cell Search Space
skip
3D: 3x3x3
P3D: 3x3x1
P3D: 3x1x3
P3D: 1x3x3
+
3506
3512
3510
3514
d2x
cell
cell
cell
+
u2x
3516
3504
3508

FIG. 35

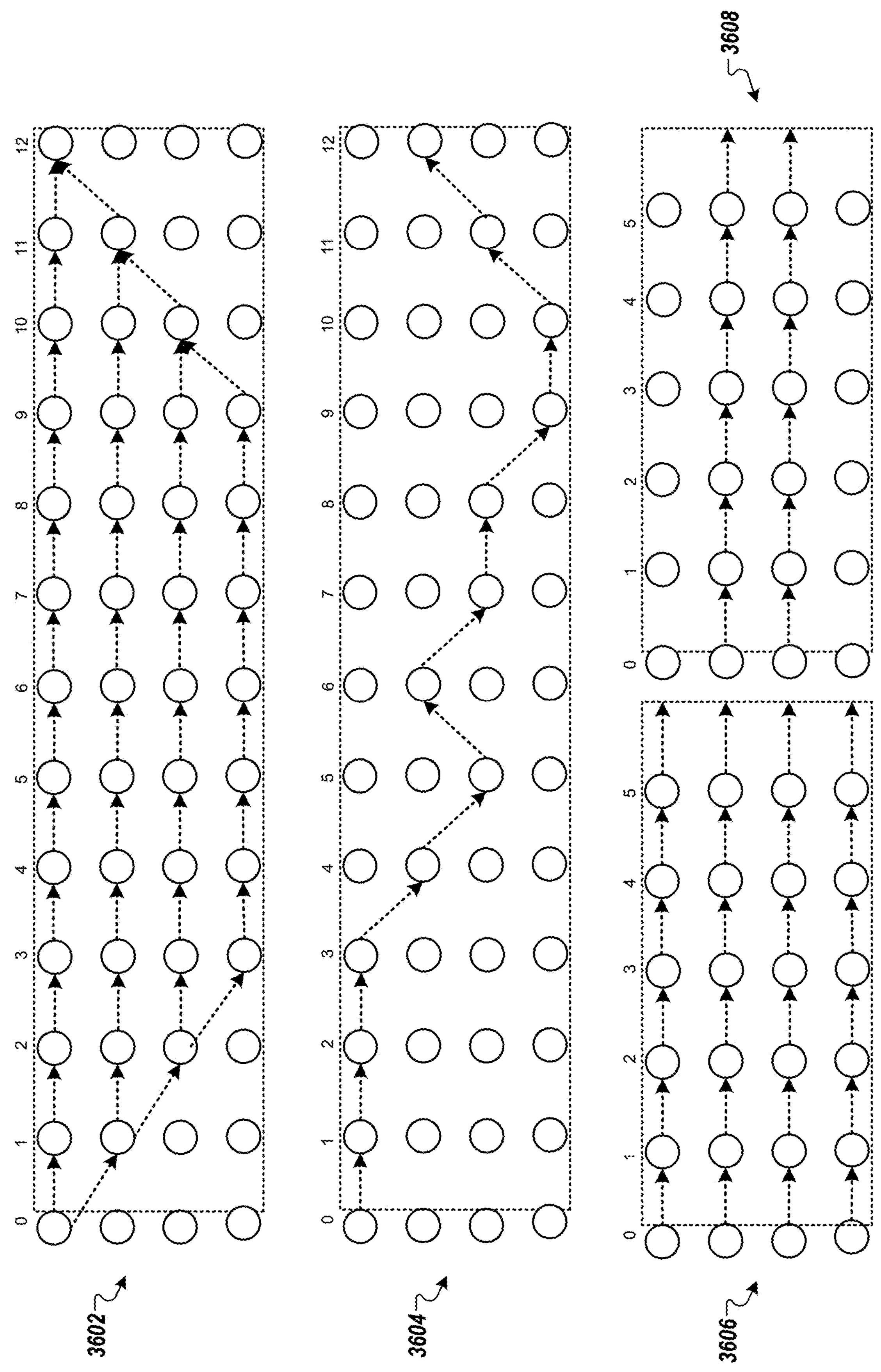


FIG. 36

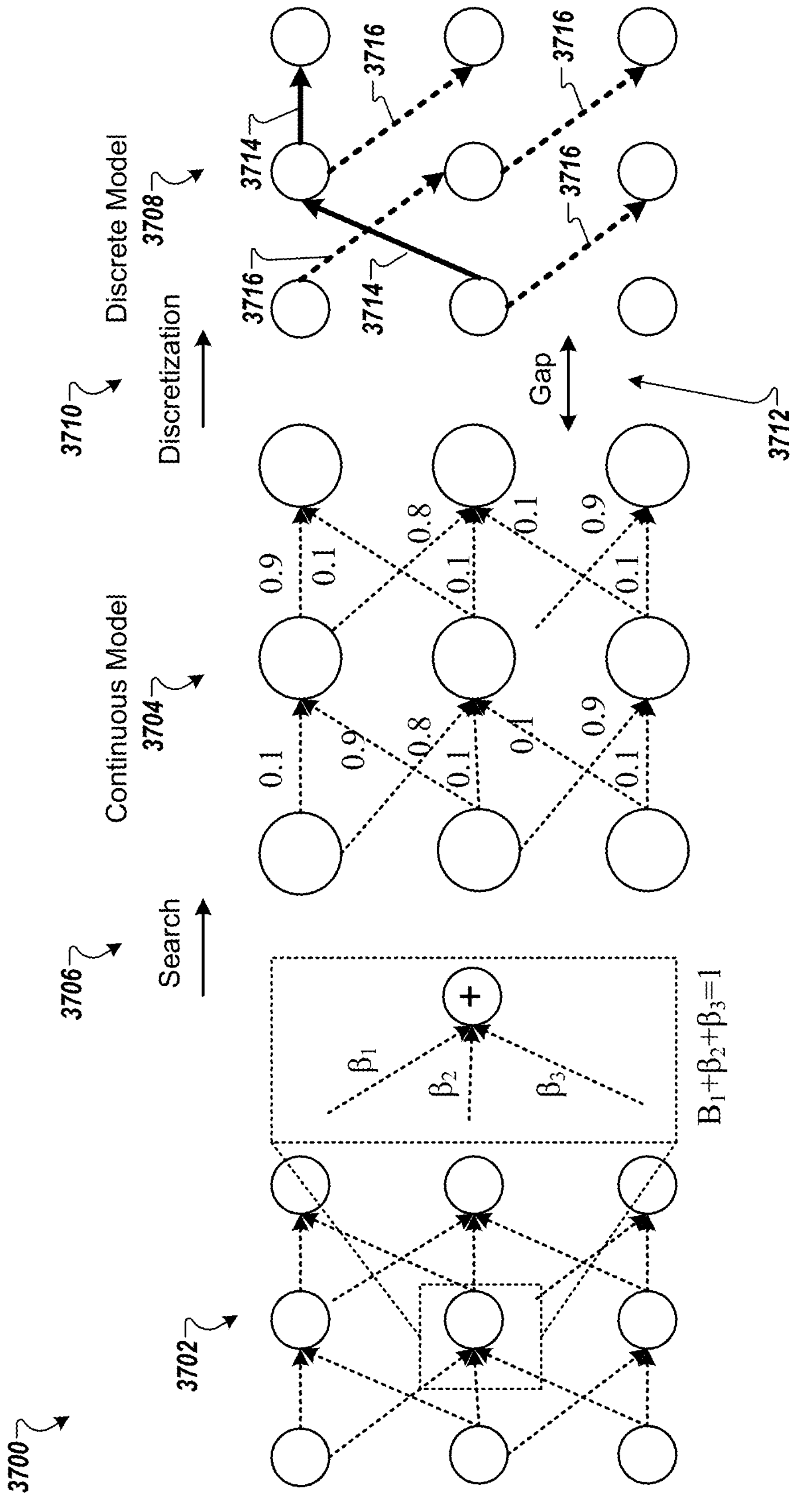

3700
3702
3706
Search
$B_1+\beta_2+\beta_3=1$
β_1
β_2
β_3
+
Continuous Model
3704
0.1
0.9
0.8
0.1
0.1
0.9
0.1
0.9
0.1
0.8
0.1
0.1
0.9
0.1
3710
Discretization
Gap
3712
Discrete Model
3708
3714
3714
3716
3716
3716
3716

FIG. 37

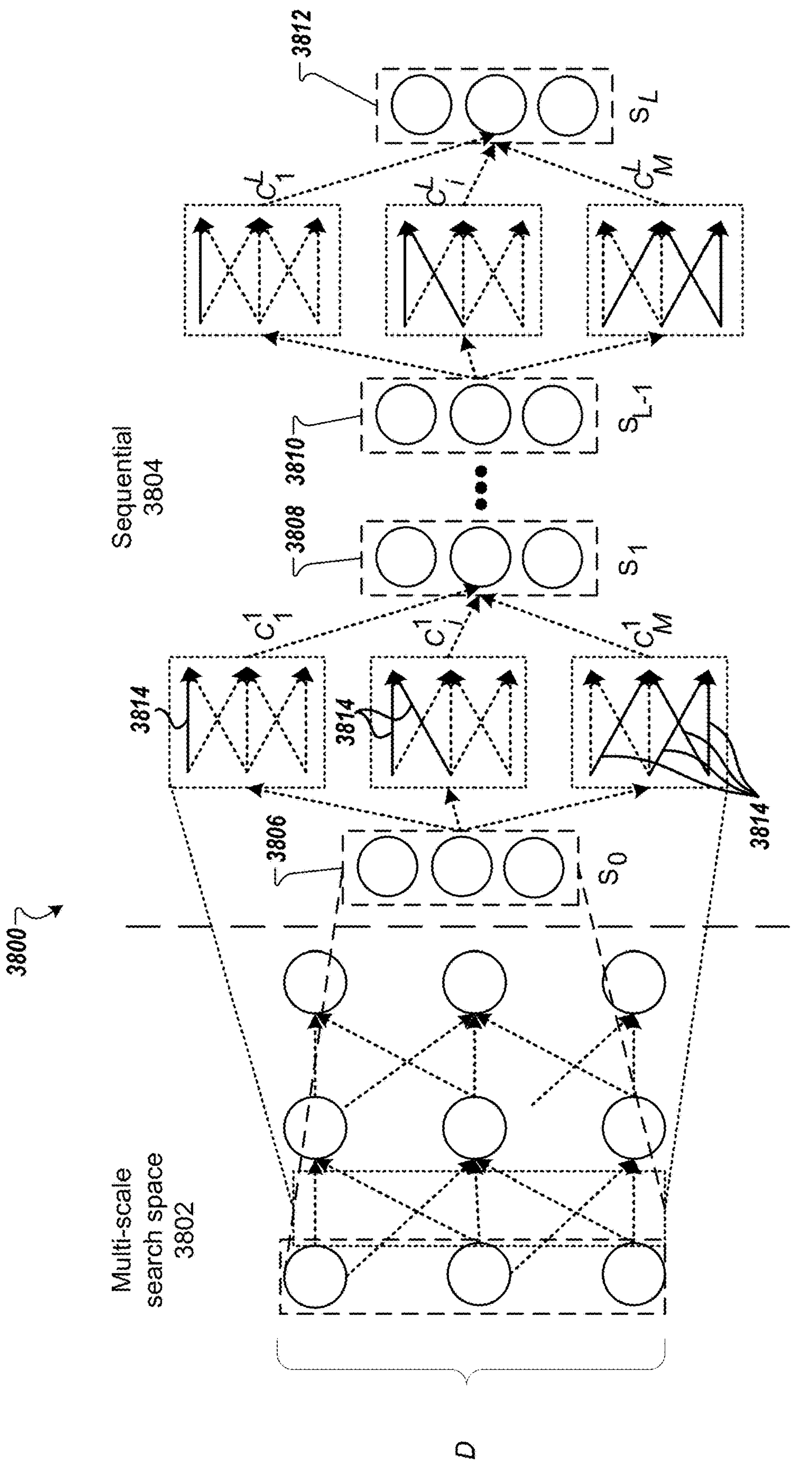

3800
Multi-scale search space 3802
Sequential 3804
3806
3808
3810
3812
3814
S_0
S_1
S_{L-1}
S_L
C^1_1
C^1_i
C^1_M
C^L_1
C^L_i
C^L_M
D

FIG. 38

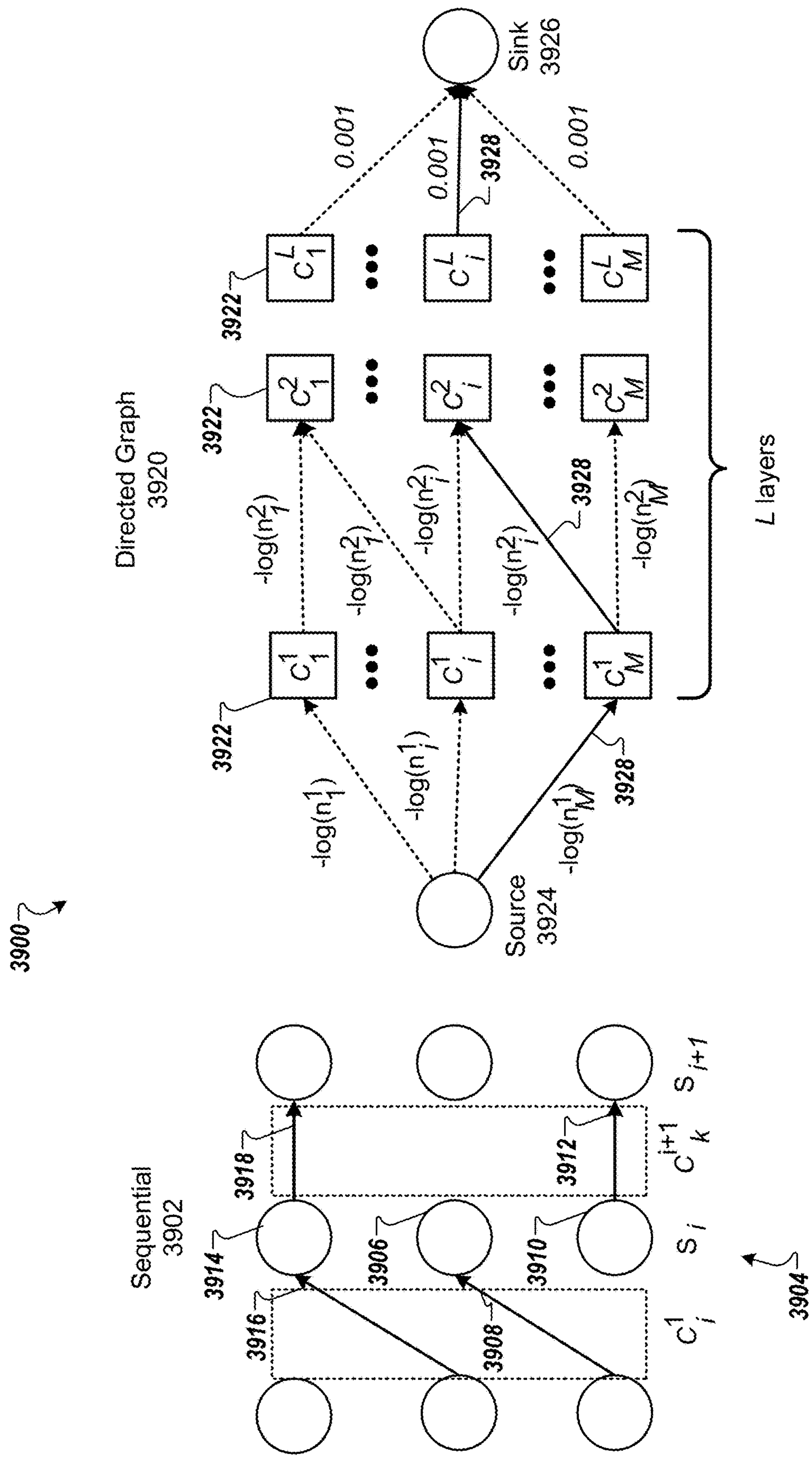

3900
Sequential
3902
3914
3916
3918
3906
3908
3910
3912
3904
C_i^1
S_i
C_k^{i+1}
S_{i+1}
Directed Graph
3920
Source
3924
Sink
3926
3922
3928
$-\log(n_1^1)$
$-\log(n_i^1)$
$-\log(n_M^1)$
$-\log(n_1^2)$
$-\log(n_i^2)$
$-\log(n_M^2)$
C_1^1
C_i^1
C_M^1
C_1^2
C_i^2
C_M^2
C_1^L
C_i^L
C_M^L
0.001
L layers

FIG. 39

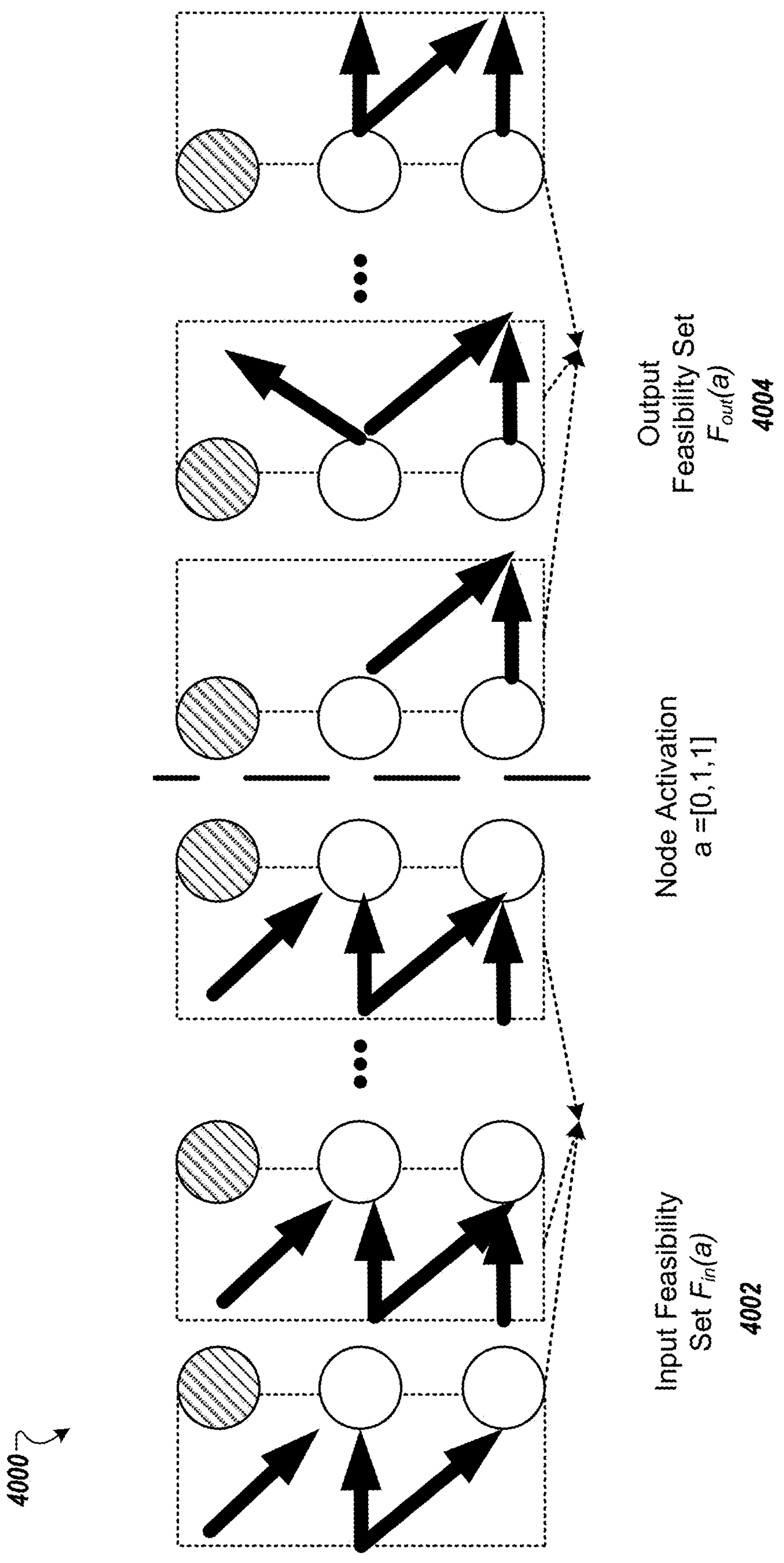

4000
Input Feasibility Set $F_{in}(a)$
4002
Node Activation a =[0,1,1]
Output Feasibility Set $F_{out}(a)$
4004

FIG. 40

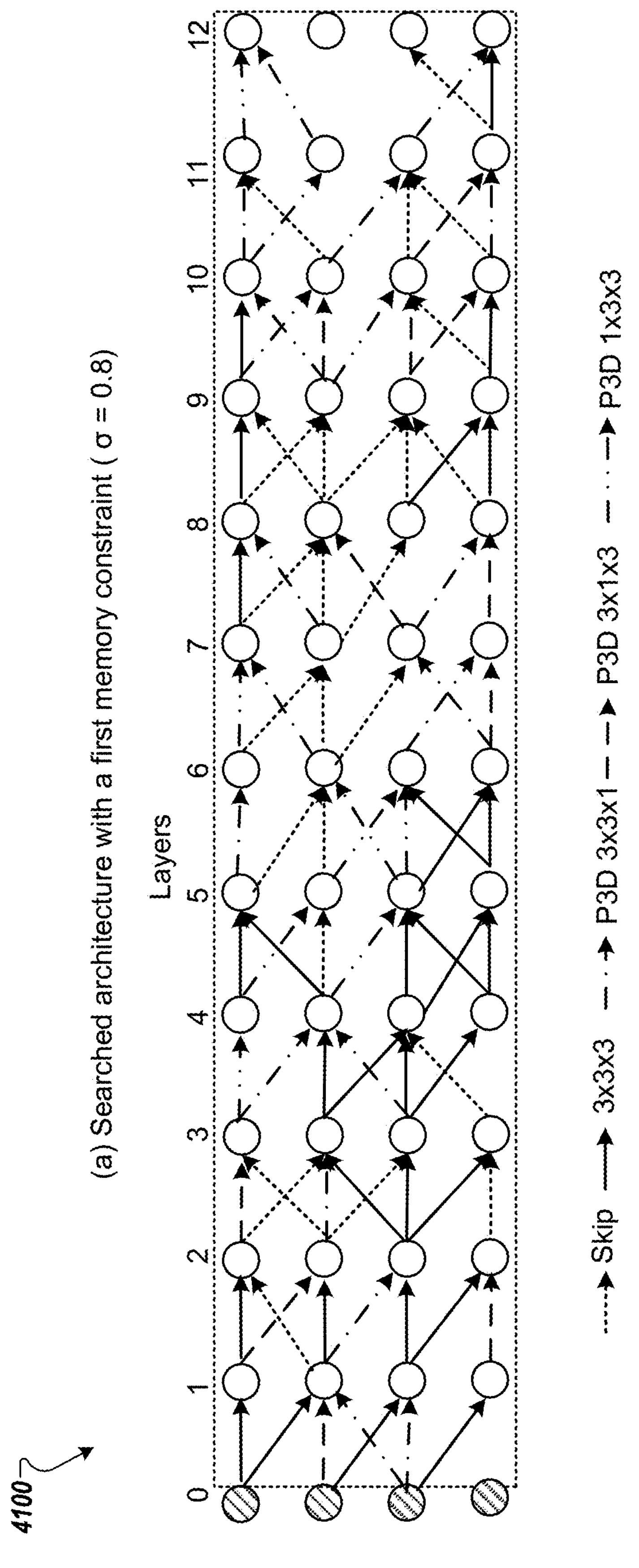

4100
(a) Searched architecture with a first memory constraint (σ = 0.8)
Layers
0
1
2
3
4
5
6
7
8
9
10
11
12
Skip
3x3x3
P3D 3x3x1
P3D 3x1x3
P3D 1x3x3

FIG. 41A

4110

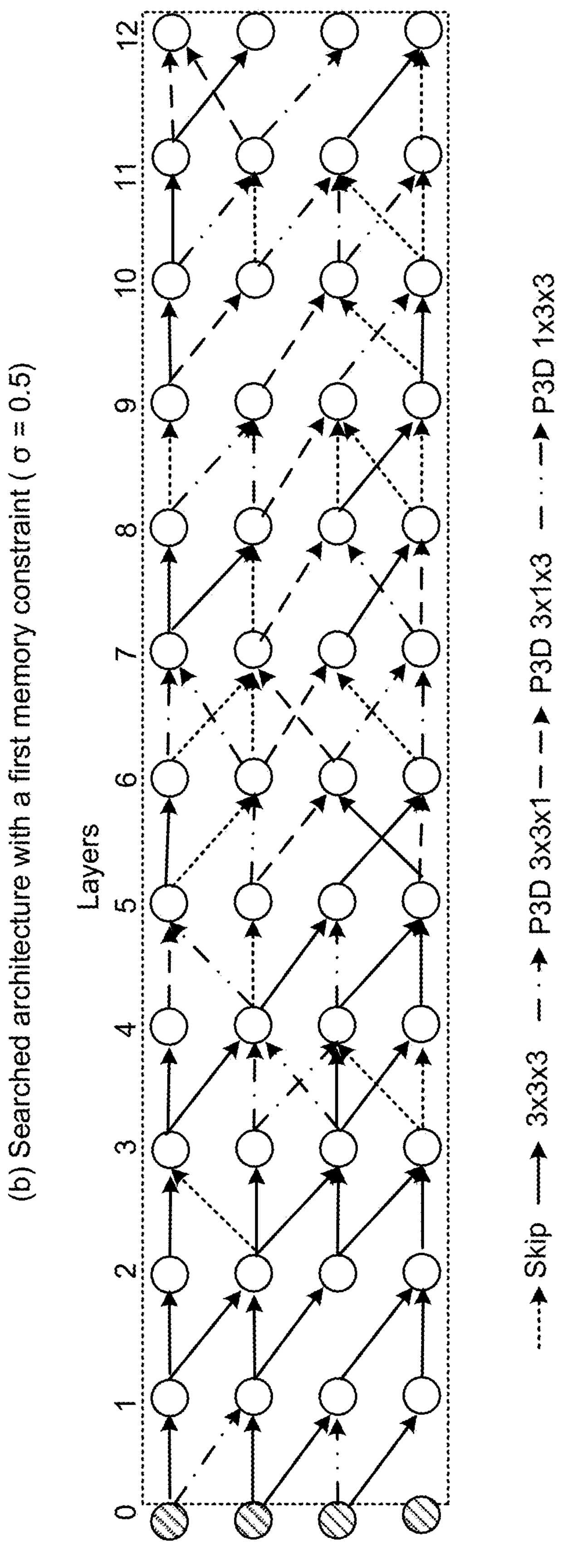

(b) Searched architecture with a first memory constraint (σ = 0.5)
Layers
0
1
2
3
4
5
6
7
8
9
10
11
12
Skip
3x3x3
P3D 3x3x1
P3D 3x1x3
P3D 1x3x3

FIG. 41B

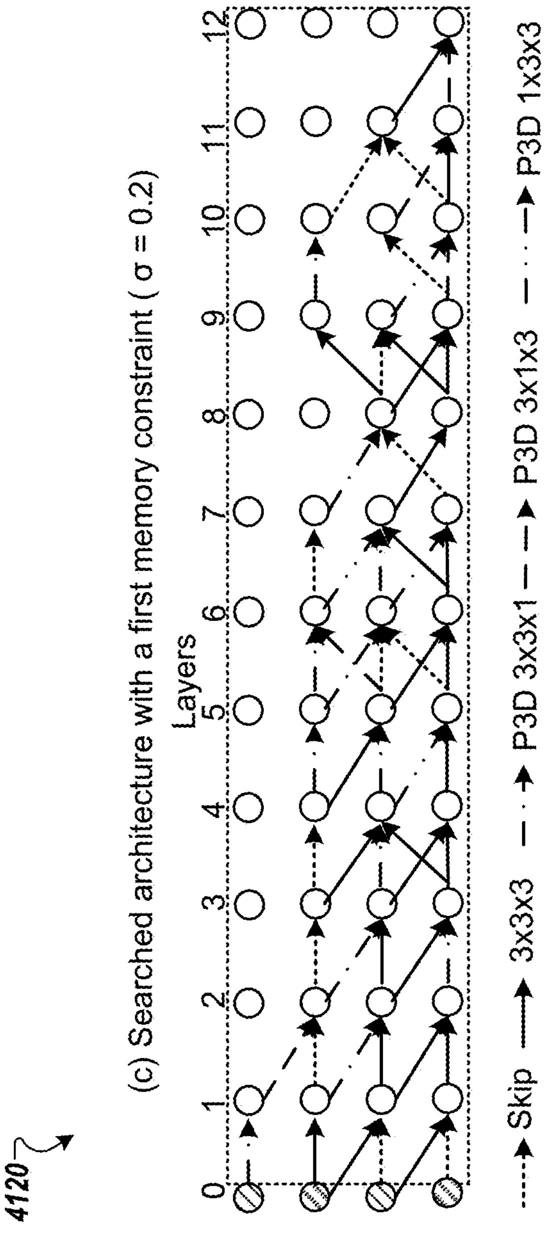

4120
(c) Searched architecture with a first memory constraint (σ = 0.2)
Layers
0
1
2
3
4
5
6
7
8
9
10
11
12
Skip
3x3x3
P3D 3x3x1
P3D 3x1x3
P3D 1x3x3

FIG. 41C

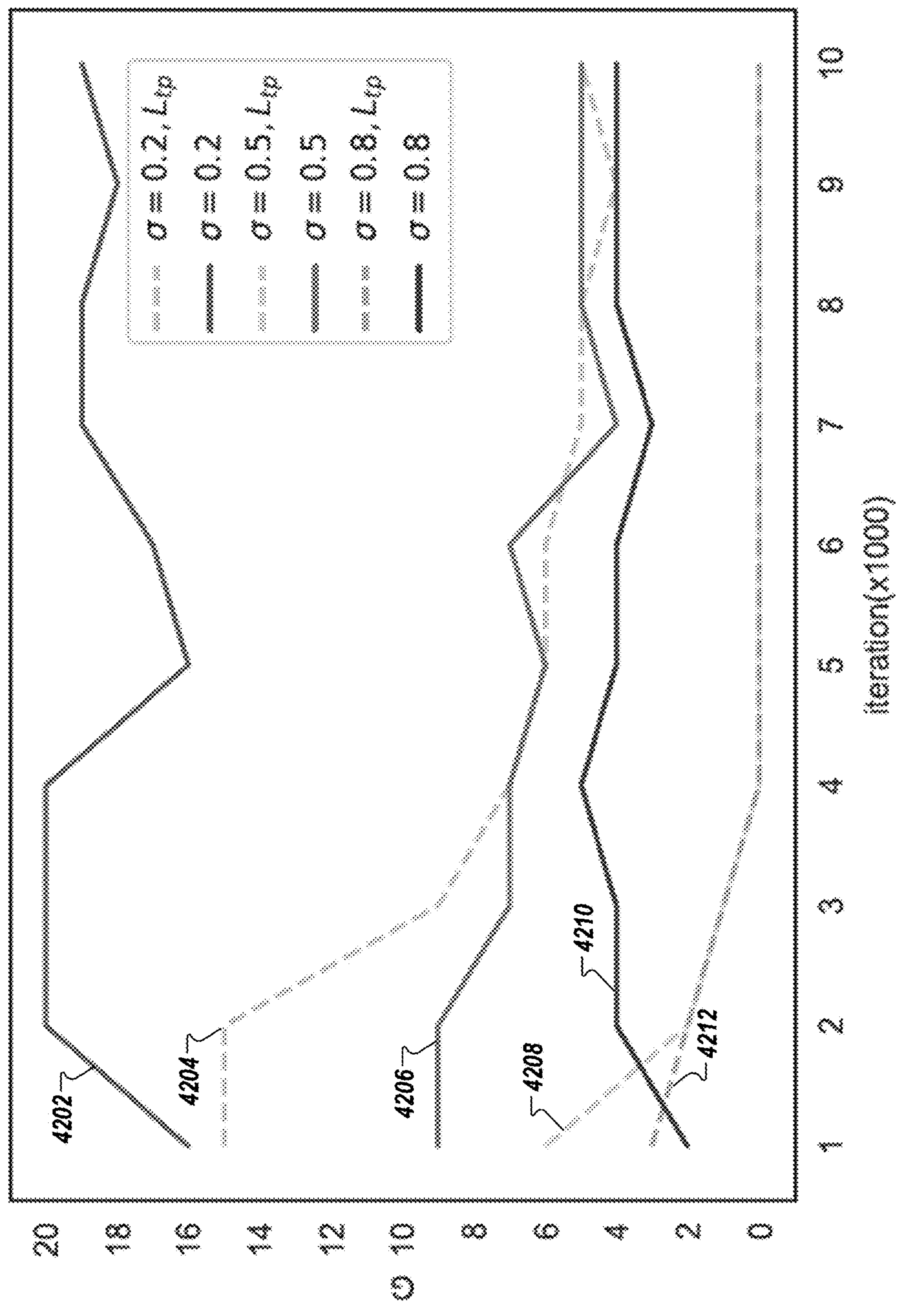

G
0
2
4
6
8
10
12
14
16
18
20
iteration(x1000)
1
2
3
4
5
6
7
8
9
10
$\sigma = 0.2, L_{tp}$
$\sigma = 0.2$
$\sigma = 0.5, L_{tp}$
$\sigma = 0.5$
$\sigma = 0.8, L_{tp}$
$\sigma = 0.8$
4202
4204
4206
4208
4210
4212

FIG. 42

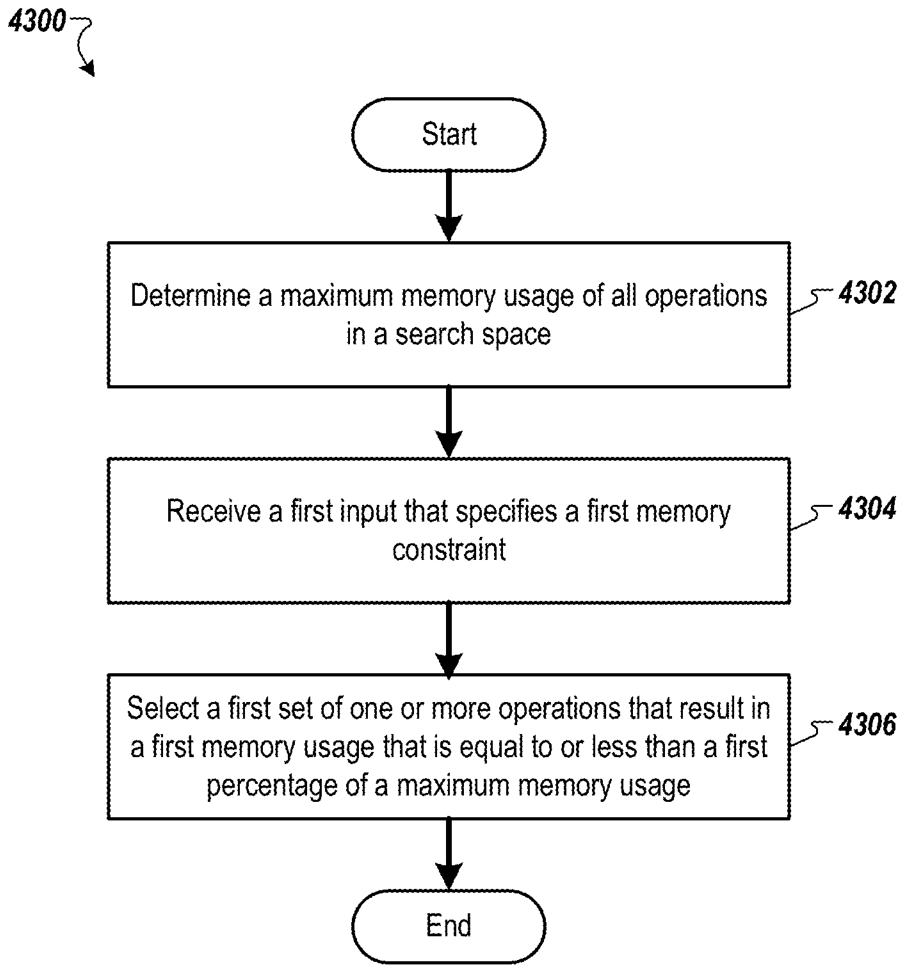

4300
Start
Determine a maximum memory usage of all operations in a search space
4302
Receive a first input that specifies a first memory constraint
4304
Select a first set of one or more operations that result in a first memory usage that is equal to or less than a first percentage of a maximum memory usage
4306
End

FIG. 43

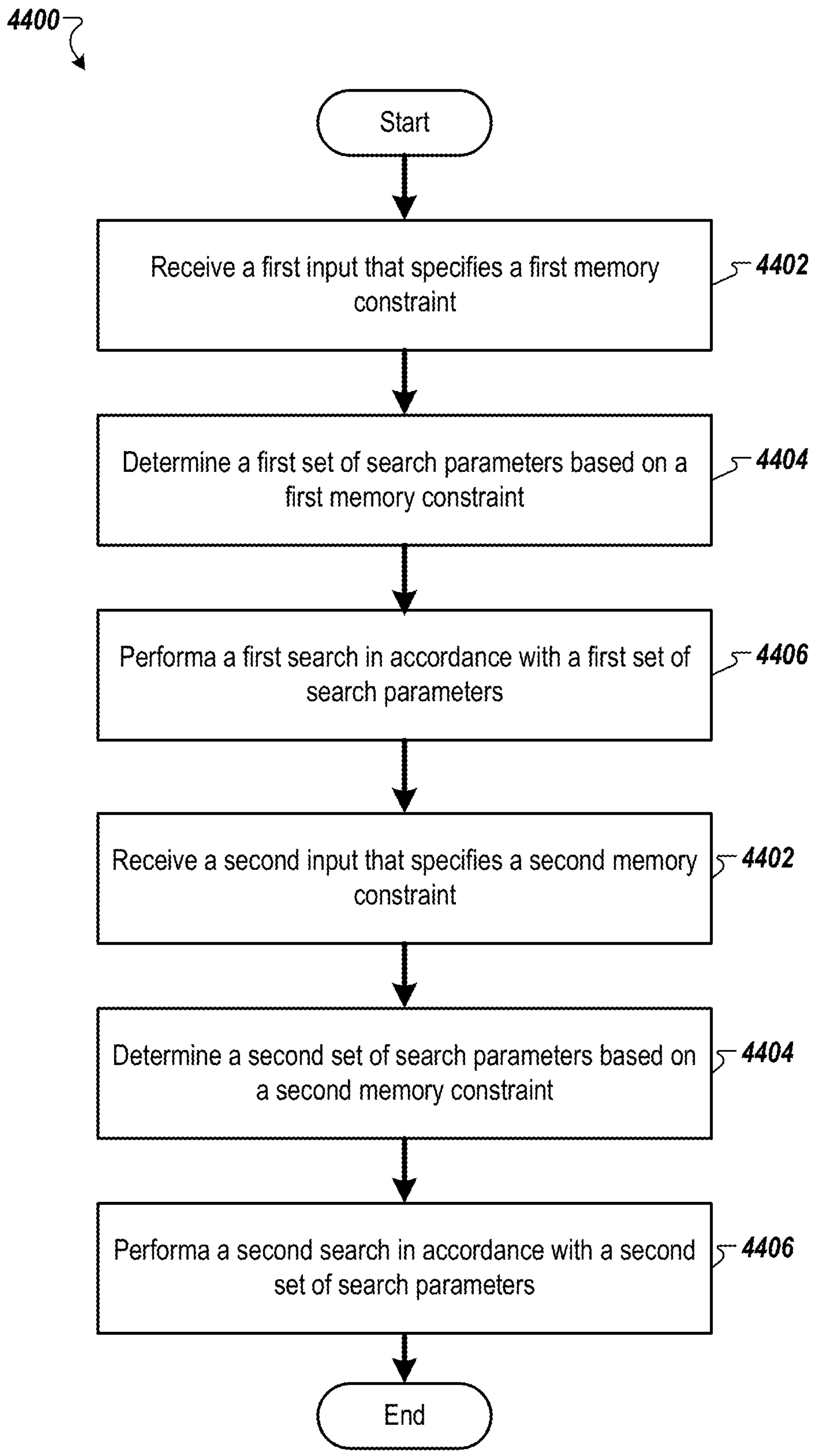

4400
Start
Receive a first input that specifies a first memory constraint
4402
Determine a first set of search parameters based on a first memory constraint
4404
Performa a first search in accordance with a first set of search parameters
4406
Receive a second input that specifies a second memory constraint
4402
Determine a second set of search parameters based on a second memory constraint
4404
Performa a second search in accordance with a second set of search parameters
4406
End

FIG. 44

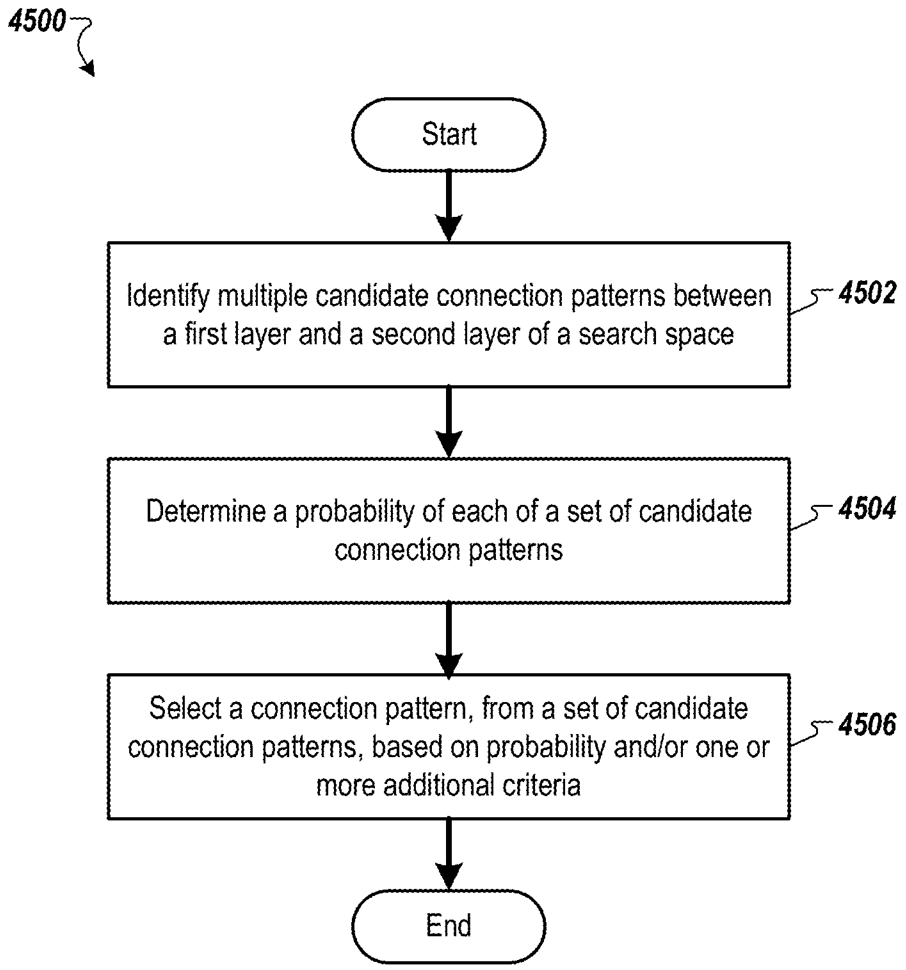

4500
Start
Identify multiple candidate connection patterns between a first layer and a second layer of a search space
4502
Determine a probability of each of a set of candidate connection patterns
4504
Select a connection pattern, from a set of candidate connection patterns, based on probability and/or one or more additional criteria
4506
End

FIG. 45

SELECTING A NEURAL NETWORK BASED ON AN AMOUNT OF MEMORY

TECHNICAL FIELD

[0001] At least one embodiment pertains to processing resources used to perform and facilitate artificial intelligence. For example, at least one embodiment pertains to processors or computing systems used to train and use neural networks according to various novel techniques described herein.

BACKGROUND

[0002] Neural networks can be used for image segmentation tasks. Image segmentation is a process of partitioning an image into multiple sets of pixels and labeling every pixel such that pixels with a same label share certain characteristics. Image segmentation can classify all pixels of an image into classes of objects, allowing identification of parts of an image and understand what objects these parts belong to.

BRIEF DESCRIPTION OF DRAWINGS

[0003] FIG. **1**A illustrates inference and/or training logic, according to at least one embodiment;

[0004] FIG. **1**B illustrates inference and/or training logic, according to at least one embodiment;

[0005] FIG. **2** illustrates training and deployment of a neural network, according to at least one embodiment;

[0006] FIG. **3** illustrates an example data center system, according to at least one embodiment;

[0007] FIG. **4**A illustrates an example of an autonomous vehicle, according to at least one embodiment;

[0008] FIG. **4**B illustrates an example of camera locations and fields of view for an autonomous vehicle of FIG. **4**A, according to at least one embodiment;

[0009] FIG. **4**C is a block diagram illustrating an example system architecture for an autonomous vehicle of FIG. **4**A, according to at least one embodiment;

[0010] FIG. **4**D is a diagram illustrating a system for communication between cloud-based server(s) and an autonomous vehicle of FIG. **4**A, according to at least one embodiment;

[0011] FIG. **5** is a block diagram illustrating a computer system, according to at least one embodiment;

[0012] FIG. **6** is a block diagram illustrating a computer system, according to at least one embodiment;

[0013] FIG. **7** illustrates a computer system, according to at least one embodiment;

[0014] FIG. **8** illustrates a computer system, according to at least one embodiment;

[0015] FIG. **9**A illustrates a computer system, according to at least one embodiment;

[0016] FIG. **9**B illustrates a computer system, according to at least one embodiment;

[0017] FIG. **9**C illustrates a computer system, according to at least one embodiment;

[0018] FIG. **9**D illustrates a computer system, according to at least one embodiment;

[0019] FIGS. **9**E and **9**F illustrate a shared programming model, according to at least one embodiment;

[0020] FIG. **10** illustrates exemplary integrated circuits and associated graphics processors, according to at least one embodiment;

[0021] FIGS. **11**A-**11**B illustrate exemplary integrated circuits and associated graphics processors, according to at least one embodiment;

[0022] FIGS. **12**A-**12**B illustrate additional exemplary graphics processor logic according to at least one embodiment;

[0023] FIG. **13** illustrates a computer system, according to at least one embodiment;

[0024] FIG. **14**A illustrates a parallel processor, according to at least one embodiment;

[0025] FIG. **14**B illustrates a partition unit, according to at least one embodiment;

[0026] FIG. **14**C illustrates a processing cluster, according to at least one embodiment;

[0027] FIG. **14**D illustrates a graphics multiprocessor, according to at least one embodiment;

[0028] FIG. **15** illustrates a multi-graphics processing unit (GPU) system, according to at least one embodiment;

[0029] FIG. **16** illustrates a graphics processor, according to at least one embodiment;

[0030] FIG. **17** is a block diagram illustrating a processor micro-architecture for a processor, according to at least one embodiment;

[0031] FIG. **18** illustrates a deep learning application processor, according to at least one embodiment;

[0032] FIG. **19** is a block diagram illustrating an example neuromorphic processor, according to at least one embodiment;

[0033] FIG. **20** illustrates at least portions of a graphics processor, according to one or more embodiments;

[0034] FIG. **21** illustrates at least portions of a graphics processor, according to one or more embodiments;

[0035] FIG. **22** illustrates at least portions of a graphics processor, according to one or more embodiments;

[0036] FIG. **23** is a block diagram of a graphics processing engine of a graphics processor in accordance with at least one embodiment;

[0037] FIG. **24** is a block diagram of at least portions of a graphics processor core, according to at least one embodiment;

[0038] FIGS. **25**A-**25**B illustrate thread execution logic including an array of processing elements of a graphics processor core according to at least one embodiment;

[0039] FIG. **26** illustrates a parallel processing unit ("PPU"), according to at least one embodiment;

[0040] FIG. **27** illustrates a general processing cluster ("GPC"), according to at least one embodiment;

[0041] FIG. **28** illustrates a memory partition unit of a parallel processing unit ("PPU"), according to at least one embodiment;

[0042] FIG. **29** illustrates a streaming multi-processor, according to at least one embodiment;

[0043] FIG. **30** is an example data flow diagram for an advanced computing pipeline, in accordance with at least one embodiment;

[0044] FIG. **31** is a system diagram for an example system for training, adapting, instantiating, and deploying machine learning models in an advanced computing pipeline, in accordance with at least one embodiment;

[0045] FIG. **32** includes an example illustration of an advanced computing pipeline **3110**A for processing imaging data, in accordance with at least one embodiment;

[0046] FIG. **33**A includes an example data flow diagram of a virtual instrument supporting an ultrasound device, in accordance with at least one embodiment;

[0047] FIG. **33**B includes an example data flow diagram of a virtual instrument supporting a CT scanner, in accordance with at least one embodiment;

[0048] FIG. **34**A illustrates a data flow diagram for a process to train a machine learning model, in accordance with at least one embodiment;

[0049] FIG. **34**B is an example illustration of a client-server architecture to enhance annotation tools with pre-trained annotation models, in accordance with at least one embodiment;

[0050] FIG. **35** is a visual representation of a search space with fully connected edges between adjacent layers for a differentiable NAS method, according to at least one embodiment;

[0051] FIG. **36** illustrates a multi-path topology, a single-path topology, a multi-path topology with four input resolutions, and a multi-path topology with two input resolutions, according to at least one embodiment;

[0052] FIG. **37** illustrates a search stage and a discretization stage of a differentiable NAS method, according to at least one embodiment;

[0053] FIG. **38** illustrates a differentiable NAS method with a sequential model with super nodes, according to at least one embodiment;

[0054] FIG. **39** illustrates a differentiable NAS method that discretizes a sequential model with topology feasibility constraints, according to at least one embodiment;

[0055] FIG. **40** illustrates an input feasibility set and an output feasibility set that represent feasible input and output connection pattern indexes for a super node with a node activation, according to at least one embodiment;

[0056] FIGS. **41**A-**41**C illustrates three searched architectures with different memory constraints, according to at least one embodiment;

[0057] FIG. **42** is a graph illustrating indications in discretization gaps with and without topology loss in a discretization algorithm under different memory constraints, according to at least one embodiment;

[0058] FIG. **43** is a flow diagram of a method of performing a search of a search space, according to at least one embodiment;

[0059] FIG. **44** is a flow diagram of a method of performing a search of a search space, according to at least one embodiment; and

[0060] FIG. **45** is a flow diagram of a method of performing a search of a search space, according to at least one embodiment.

DETAILED DESCRIPTION

Inference and Training Logic

[0061] FIG. **1**A illustrates inference and/or training logic **115** used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided below in conjunction with FIGS. **1**A and/or **1**B.

[0062] In at least one embodiment, inference and/or training logic **115** may include, without limitation, code and/or data storage **101** to store forward and/or output weight and/or input/output data, and/or other parameters to configure neurons or layers of a neural network trained and/or used for inferencing in aspects of one or more embodiments. In at least one embodiment, training logic **115** may include, or be coupled to code and/or data storage **101** to store graph code or other software to control timing and/or order, in which weight and/or other parameter information is to be loaded to configure, logic, including integer and/or floating point units (collectively, arithmetic logic units (ALUs). In at least one embodiment, code, such as graph code, loads weight or other parameter information into processor ALUs based on an architecture of a neural network to which such code corresponds. In at least one embodiment, code and/or data storage **101** stores weight parameters and/or input/output data of each layer of a neural network trained or used in conjunction with one or more embodiments during forward propagation of input/output data and/or weight parameters during training and/or inferencing using aspects of one or more embodiments. In at least one embodiment, any portion of code and/or data storage **101** may be included with other on-chip or off-chip data storage, including a processor's L1, L2, or L3 cache or system memory.

[0063] In at least one embodiment, any portion of code and/or data storage **101** may be internal or external to one or more processors or other hardware logic devices or circuits. In at least one embodiment, code and/or code and/or data storage **101** may be cache memory, dynamic randomly addressable memory ("DRAM"), static randomly addressable memory ("SRAM"), non-volatile memory (e.g., flash memory), or other storage. In at least one embodiment, a choice of whether code and/or code and/or data storage **101** is internal or external to a processor, for example, or comprising DRAM, SRAM, flash, or some other storage type may depend on available storage on-chip versus off-chip, latency requirements of training and/or inferencing functions being performed, batch size of data used in inferencing and/or training of a neural network, or some combination of these factors.

[0064] In at least one embodiment, inference and/or training logic **115** may include, without limitation, a code and/or data storage **105** to store backward and/or output weight and/or input/output data corresponding to neurons or layers of a neural network trained and/or used for inferencing in aspects of one or more embodiments. In at least one embodiment, code and/or data storage **105** stores weight parameters and/or input/output data of each layer of a neural network trained or used in conjunction with one or more embodiments during backward propagation of input/output data and/or weight parameters during training and/or inferencing using aspects of one or more embodiments. In at least one embodiment, training logic **115** may include, or be coupled to code and/or data storage **105** to store graph code or other software to control timing and/or order, in which weight and/or other parameter information is to be loaded to configure, logic, including integer and/or floating point units (collectively, arithmetic logic units (ALUs).

[0065] In at least one embodiment, code, such as graph code, causes loading of weight or other parameter information into processor ALUs based on an architecture of a neural network to which such code corresponds. In at least one embodiment, any portion of code and/or data storage **105** may be included with other on-chip or off-chip data storage, including a processor's L1, L2, or L3 cache or system memory. In at least one embodiment, any portion of code and/or data storage **105** may be internal or external to one or more processors or other hardware logic devices or circuits. In at least one embodiment, code and/or data storage **105** may be cache memory, DRAM, SRAM, non-volatile memory (e.g., flash memory), or other storage. In at least one embodiment, a choice of whether code and/or data storage **105** is internal or external to a processor, for example, or comprising DRAM, SRAM, flash memory or some other storage type may depend on available storage on-chip versus off-chip, latency requirements of training and/or inferencing functions being performed, batch size of data used in inferencing and/or training of a neural network, or some combination of these factors.

[0066] In at least one embodiment, code and/or data storage **101** and code and/or data storage **105** may be separate storage structures. In at least one embodiment, code and/or data storage **101** and code and/or data storage **105** may be a combined storage structure. In at least one embodiment, code and/or data storage **101** and code and/or data storage **105** may be partially combined and partially separate. In at least one embodiment, any portion of code and/or data storage **101** and code and/or data storage **105** may be included with other on-chip or off-chip data storage, including a processor's L1, L2, or L3 cache or system memory.

[0067] In at least one embodiment, inference and/or training logic **115** may include, without limitation, one or more arithmetic logic unit(s) ("ALU(s)") **110**, including integer and/or floating point units, to perform logical and/or mathematical operations based, at least in part on, or indicated by, training and/or inference code (e.g., graph code), a result of which may produce activations (e.g., output values from layers or neurons within a neural network) stored in an activation storage **120** that are functions of input/output and/or weight parameter data stored in code and/or data storage **101** and/or code and/or data storage **105**. In at least one embodiment, activations stored in activation storage **120** are generated according to linear algebraic and or matrix-based mathematics performed by ALU(s) **110** in response to performing instructions or other code, wherein weight values stored in code and/or data storage **105** and/or data storage **101** are used as operands along with other values, such as bias values, gradient information, momentum values, or other parameters or hyperparameters, any or all of which may be stored in code and/or data storage **105** or code and/or data storage **101** or another storage on or off-chip.

[0068] In at least one embodiment, ALU(s) **110** are included within one or more processors or other hardware logic devices or circuits, whereas in another embodiment, ALU(s) **110** may be external to a processor or other hardware logic device or circuit that uses them (e.g., a co-processor). In at least one embodiment, ALUs **110** may be included within a processor's execution units or otherwise within a bank of ALUs accessible by a processor's execution units either within same processor or distributed between different processors of different types (e.g., central processing units, graphics processing units, fixed function units, etc.). In at least one embodiment, code and/or data storage **101**, code and/or data storage **105**, and activation storage **120** may share a processor or other hardware logic device or circuit, whereas in another embodiment, they may be in different processors or other hardware logic devices or circuits, or some combination of same and different processors or other hardware logic devices or circuits. In at least one embodiment, any portion of activation storage **120** may be included with other on-chip or off-chip data storage, including a processor's L1, L2, or L3 cache or system memory. Furthermore, inferencing and/or training code may be stored with other code accessible to a processor or other hardware logic or circuit and fetched and/or processed using a processor's fetch, decode, scheduling, execution, retirement, and/or other logical circuits.

[0069] In at least one embodiment, activation storage **120** may be cache memory, DRAM, SRAM, non-volatile memory (e.g., flash memory), or other storage. In at least one embodiment, activation storage **120** may be completely or partially within or external to one or more processors or other logical circuits. In at least one embodiment, a choice of whether activation storage **120** is internal or external to a processor, for example, or comprising DRAM, SRAM, flash memory or some other storage type may depend on available storage on-chip versus off-chip, latency requirements of training and/or inferencing functions being performed, batch size of data used in inferencing and/or training of a neural network, or some combination of these factors.

[0070] In at least one embodiment, inference and/or training logic **115** illustrated in FIG. **1**A may be used in conjunction with an application-specific integrated circuit ("ASIC"), such as a TensorFlow® Processing Unit from Google, an inference processing unit (IPU) from Graphcore™, or a Nervana® (e.g., "Lake Crest") processor from Intel Corp. In at least one embodiment, inference and/or training logic **115** illustrated in FIG. **1**A may be used in conjunction with central processing unit ("CPU") hardware, graphics processing unit ("GPU") hardware, or other hardware, such as field programmable gate arrays ("FPGAs").

[0071] FIG. **1**B illustrates inference and/or training logic **115**, according to at least one embodiment. In at least one embodiment, inference and/or training logic **115** may include, without limitation, hardware logic in which computational resources are dedicated or otherwise exclusively used in conjunction with weight values or other information corresponding to one or more layers of neurons within a neural network. In at least one embodiment, inference and/or training logic **115** illustrated in FIG. **1**B may be used in conjunction with an application-specific integrated circuit (ASIC), such as TensorFlow® Processing Unit from Google, an inference processing unit (IPU) from Graphcore™, or a Nervana® (e.g., "Lake Crest") processor from Intel Corp. In at least one embodiment, inference and/or training logic **115** illustrated in FIG. **1**B may be used in conjunction with central processing unit (CPU) hardware, graphics processing unit (GPU) hardware, or other hardware, such as field programmable gate arrays (FPGAs). In at least one embodiment, inference and/or training logic **115** includes, without limitation, code and/or data storage **101** and code and/or data storage **105**, which may be used to store code (e.g., graph code), weight values and/or other information, including bias values, gradient information, momentum values, and/or other parameter or hyperparameter information. In at least one embodiment illustrated in FIG. **1**B, each of code and/or data storage **101** and code and/or data storage **105** is associated with a dedicated computational resource, such as computational hardware **102** and computational hardware **106**, respectively. In at least one embodiment, each of computational hardware **102** and computational hardware **106** comprises one or more ALUs that perform mathematical functions, such as linear algebraic functions, only on information stored in code and/or data storage **101** and code and/or data storage **105**, respectively, result of which is stored in activation storage **120**.

[0072] In at least one embodiment, each of code and/or data storage **101** and **105** and corresponding computational hardware **102** and **106**, respectively, correspond to different layers of a neural network, such that resulting activation from one storage/computational pair **101/102** of code and/or data storage **101** and computational hardware **102** is provided as an input to a next storage/computational pair **105/106** of code and/or data storage **105** and computational hardware **106**, in order to mirror a conceptual organization of a neural network. In at least one embodiment, each of storage/computational pairs **101/102** and **105/106** may correspond to more than one neural network layer. In at least one embodiment, additional storage/computation pairs (not shown) subsequent to or in parallel with storage/computation pairs **101/102** and **105/106** may be included in inference and/or training logic **115**.

Neural Network Training and Deployment

[0073] FIG. **2** illustrates training and deployment of a deep neural network, according to at least one embodiment. In at least one embodiment, untrained neural network **206** is trained using a training dataset **202**. In at least one embodiment, training framework **204** is a PyTorch framework, whereas in other embodiments, training framework **204** is a TensorFlow, Boost, Caffe, Microsoft Cognitive Toolkit/CNTK, MXNet, Chainer, Keras, Deeplearning4j, or other training framework. In at least one embodiment, training framework **204** trains an untrained neural network **206** and enables it to be trained using processing resources described herein to generate a trained neural network **208**. In at least one embodiment, weights may be chosen randomly or by pre-training using a deep belief network. In at least one embodiment, training may be performed in either a supervised, partially supervised, or unsupervised manner.

[0074] In at least one embodiment, untrained neural network **206** is trained using supervised learning, wherein training dataset **202** includes an input paired with a desired output for an input, or where training dataset **202** includes input having a known output and an output of neural network **206** is manually graded. In at least one embodiment, untrained neural network **206** is trained in a supervised manner and processes inputs from training dataset **202** and compares resulting outputs against a set of expected or desired outputs. In at least one embodiment, errors are then propagated back through untrained neural network **206**. In at least one embodiment, training framework **204** adjusts weights that control untrained neural network **206**. In at least one embodiment, training framework **204** includes tools to monitor how well untrained neural network **206** is converging towards a model, such as trained neural network **208**, suitable to generating correct answers, such as in result **214**, based on input data such as a new dataset **212**. In at least one embodiment, training framework **204** trains untrained neural network **206** repeatedly while adjusting weights to refine an output of untrained neural network **206** using a loss function and adjustment algorithm, such as stochastic gradient descent. In at least one embodiment, training framework **204** trains untrained neural network **206** until untrained neural network **206** achieves a desired accuracy. In at least one embodiment, trained neural network **208** can then be deployed to implement any number of machine learning operations.

[0075] In at least one embodiment, untrained neural network **206** is trained using unsupervised learning, wherein untrained neural network **206** attempts to train itself using unlabeled data. In at least one embodiment, unsupervised learning training dataset **202** will include input data without any associated output data or "ground truth" data. In at least one embodiment, untrained neural network **206** can learn groupings within training dataset **202** and can determine how individual inputs are related to untrained dataset **202**. In at least one embodiment, unsupervised training can be used to generate a self-organizing map in trained neural network **208** capable of performing operations useful in reducing dimensionality of new dataset **212**. In at least one embodiment, unsupervised training can also be used to perform anomaly detection, which allows identification of data points in new dataset **212** that deviate from normal patterns of new dataset **212**.

[0076] In at least one embodiment, semi-supervised learning may be used, which is a technique in which in training dataset **202** includes a mix of labeled and unlabeled data. In at least one embodiment, training framework **204** may be used to perform incremental learning, such as through transferred learning techniques. In at least one embodiment, incremental learning enables trained neural network **208** to adapt to new dataset **212** without forgetting knowledge instilled within trained neural network **208** during initial training.

Data Center

[0077] FIG. **3** illustrates an example data center **300**, in which at least one embodiment may be used. In at least one embodiment, data center **300** includes a data center infrastructure layer **310**, a framework layer **320**, a software layer **330**, and an application layer **340**.

[0078] In at least one embodiment, as shown in FIG. **3**, data center infrastructure layer **310** may include a resource orchestrator **312**, grouped computing resources **314**, and node computing resources ("node C.R.s") **316(1)**-**316**(N), where "N" represents a positive integer (which may be a different integer "N" than used in other figures). In at least one embodiment, node C.R.s **316(1)**-**316**(N) may include, but are not limited to, any number of central processing units ("CPUs") or other processors (including accelerators, field programmable gate arrays (FPGAs), graphics processors, etc.), memory storage devices **318(1)**-**318**(N) (e.g., dynamic read-only memory, solid state storage or disk drives), network input/output ("NW I/O") devices, network switches, virtual machines ("VMs"), power modules, and cooling modules, etc. In at least one embodiment, one or more node C.R.s from among node C.R.s **316**(1)-**316**(N) may be a server having one or more of above-mentioned computing resources.

[0079] In at least one embodiment, grouped computing resources **314** may include separate groupings of node C.R.s housed within one or more racks (not shown), or many racks housed in data centers at various geographical locations (also not shown). In at least one embodiment, separate groupings of node C.R.s within grouped computing resources **314** may include grouped compute, network, memory, or storage resources that may be configured or allocated to support one or more workloads. In at least one embodiment, several node C.R.s including CPUs or processors may be grouped within one or more racks to provide compute resources to support one or more workloads. In at least one embodiment, one or more racks may also include any number of power modules, cooling modules, and network switches, in any combination.

[0080] In at least one embodiment, resource orchestrator **312** may configure or otherwise control one or more node C.R.s **316**(**1**)-**316**(N) and/or grouped computing resources **314**. In at least one embodiment, resource orchestrator **312** may include a software design infrastructure ("SDI") management entity for data center **300**. In at least one embodiment, resource orchestrator **112** may include hardware, software, or some combination thereof.

[0081] In at least one embodiment, as shown in FIG. **3**, framework layer **320** includes a job scheduler **322**, a configuration manager **324**, a resource manager **326**, and a distributed file system **328**. In at least one embodiment, framework layer **320** may include a framework to support software **332** of software layer **330** and/or one or more application(s) **342** of application layer **340**. In at least one embodiment, software **332** or application(s) **342** may respectively include web-based service software or applications, such as those provided by Amazon Web Services, Google Cloud, and Microsoft Azure. In at least one embodiment, framework layer **320** may be, but is not limited to, a type of free and open-source software web application framework such as Apache Spark™ (hereinafter "Spark") that may utilize distributed file system **328** for large-scale data processing (e.g., "big data"). In at least one embodiment, job scheduler **332** may include a Spark driver to facilitate scheduling of workloads supported by various layers of data center **300**. In at least one embodiment, configuration manager **324** may be capable of configuring different layers such as software layer **330** and framework layer **320** including Spark and distributed file system **328** for supporting large-scale data processing. In at least one embodiment, resource manager **326** may be capable of managing clustered or grouped computing resources mapped to or allocated for support of distributed file system **328** and job scheduler **322**. In at least one embodiment, clustered or grouped computing resources may include grouped computing resources **314** at data center infrastructure layer **310**. In at least one embodiment, resource manager **326** may coordinate with resource orchestrator **312** to manage these mapped or allocated computing resources.

[0082] In at least one embodiment, software **332** included in software layer **330** may include software used by at least portions of node C.R.s **316**(**1**)-**316**(N), grouped computing resources **314**, and/or distributed file system **328** of framework layer **320**. In at least one embodiment, one or more types of software may include, but are not limited to, Internet web page search software, e-mail virus scan software, database software, and streaming video content software.

[0083] In at least one embodiment, application(s) **342** included in application layer **340** may include one or more types of applications used by at least portions of node C.R.s **316**(**1**)-**316**(N), grouped computing resources **314**, and/or distributed file system **328** of framework layer **320**. In at least one embodiment, one or more types of applications may include, but are not limited to, any number of a genomics application, a cognitive compute application, and a machine learning application, including training or inferencing software, machine learning framework software (e.g., PyTorch, TensorFlow, Caffe, etc.) or other machine learning applications used in conjunction with one or more embodiments.

[0084] In at least one embodiment, any of configuration manager **324**, resource manager **326**, and resource orchestrator **312** may implement any number and type of self-modifying actions based on any amount and type of data acquired in any technically feasible fashion. In at least one embodiment, self-modifying actions may relieve a data center operator of data center **300** from making possibly bad configuration decisions and possibly avoiding underutilized and/or poor performing portions of a data center.

[0085] In at least one embodiment, data center **300** may include tools, services, software, or other resources to train one or more machine learning models or predict or infer information using one or more machine learning models according to one or more embodiments described herein. For example, in at least one embodiment, a machine learning model may be trained by calculating weight parameters according to a neural network architecture using software and computing resources described above with respect to data center **300**. In at least one embodiment, trained machine learning models corresponding to one or more neural networks may be used to infer or predict information using resources described above with respect to data center **300** by using weight parameters calculated through one or more training techniques described herein.

[0086] In at least one embodiment, data center may use CPUs, application-specific integrated circuits (ASICs), GPUs, FPGAs, or other hardware to perform training and/or inferencing using above-described resources. Moreover, one or more software and/or hardware resources described above may be configured as a service to allow users to train or performing inferencing of information, such as image recognition, speech recognition, or other artificial intelligence services.

[0087] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, inference and/or training logic **115** may be used in system FIG. **3** for inferencing or predicting operations based, at least in part, on weight parameters calculated using neural network training operations, neural network functions and/or architectures, or neural network use cases described herein.

Autonomous Vehicle

[0088] FIG. **4**A illustrates an example of an autonomous vehicle **400**, according to at least one embodiment. In at least one embodiment, autonomous vehicle **400** (alternatively referred to herein as "vehicle **400**") may be, without limitation, a passenger vehicle, such as a car, a truck, a bus, and/or another type of vehicle that accommodates one or more passengers. In at least one embodiment, vehicle **400** may be a semi-tractor-trailer truck used for hauling cargo. In at least one embodiment, vehicle **400** may be an airplane, robotic vehicle, or other kinds of vehicles.

[0089] Autonomous vehicles may be described in terms of automation levels, defined by National Highway Traffic Safety Administration ("NHTSA"), a division of US Department of Transportation, and Society of Automotive Engineers ("SAE") "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles" (e.g., Standard No. J3016-201806, published on Jun. 15, 2018, Standard No. J3016-201609, published on Sep. 30, 2016, and previous and future versions of this standard). In at least one embodiment, vehicle **400** may be capable of functioning in accordance with one or more of Level 1 through Level 5 of autonomous driving levels. For example, in at least one embodiment, vehicle **400** may be capable of conditional automation (Level 3), high automation (Level 4), and/or full automation (Level 5), depending on embodiment.

[0090] In at least one embodiment, vehicle **400** may include, without limitation, components such as a chassis, a vehicle body, wheels (e.g., 2, 4, 6, 8, 18, etc.), tires, axles, and other components of a vehicle. In at least one embodiment, vehicle **400** may include, without limitation, a propulsion system **450**, such as an internal combustion engine, hybrid electric power plant, an all-electric engine, and/or another propulsion system type. In at least one embodiment, propulsion system **450** may be connected to a drive train of vehicle **400**, which may include, without limitation, a transmission, to enable propulsion of vehicle **400**. In at least one embodiment, propulsion system **450** may be controlled in response to receiving signals from a throttle/accelerator(s) **452**.

[0091] In at least one embodiment, a steering system **454**, which may include, without limitation, a steering wheel, is used to steer vehicle **400** (e.g., along a desired path or route) when propulsion system **450** is operating (e.g., when vehicle **400** is in motion). In at least one embodiment, steering system **454** may receive signals from steering actuator(s) **456**. In at least one embodiment, a steering wheel may be optional for full automation (Level 5) functionality. In at least one embodiment, a brake sensor system **446** may be used to operate vehicle brakes in response to receiving signals from brake actuator(s) **448** and/or brake sensors.

[0092] In at least one embodiment, controller(s) **436**, which may include, without limitation, one or more system on chips ("SoCs") (not shown in FIG. **4**A) and/or graphics processing unit(s) ("GPU(s)"), provide signals (e.g., representative of commands) to one or more components and/or systems of vehicle **400**. For instance, in at least one embodiment, controller(s) **436** may send signals to operate vehicle brakes via brake actuator(s) **448**, to operate steering system **454** via steering actuator(s) **456**, to operate propulsion system **450** via throttle/accelerator(s) **452**. In at least one embodiment, controller(s) **436** may include one or more onboard (e.g., integrated) computing devices that process sensor signals, and output operation commands (e.g., signals representing commands) to enable autonomous driving and/or to assist a human driver in driving vehicle **400**. In at least one embodiment, controller(s) **436** may include a first controller for autonomous driving functions, a second controller for functional safety functions, a third controller for artificial intelligence functionality (e.g., computer vision), a fourth controller for infotainment functionality, a fifth controller for redundancy in emergency conditions, and/or other controllers. In at least one embodiment, a single controller may handle two or more of above functionalities, two or more controllers may handle a single functionality, and/or any combination thereof.

[0093] In at least one embodiment, controller(s) **436** provide signals for controlling one or more components and/or systems of vehicle **400** in response to sensor data received from one or more sensors (e.g., sensor inputs). In at least one embodiment, sensor data may be received from, for example and without limitation, global navigation satellite systems ("GNSS") sensor(s) **458** (e.g., Global Positioning System sensor(s)), RADAR sensor(s) **460**, ultrasonic sensor(s) **462**, LIDAR sensor(s) **464**, inertial measurement unit ("IMU") sensor(s) **466** (e.g., accelerometer(s), gyroscope(s), a magnetic compass or magnetic compasses, magnetometer(s), etc.), microphone(s) **496**, stereo camera(s) **468**, wide-view camera(s) **470** (e.g., fisheye cameras), infrared camera(s) **472**, surround camera(s) **474** (e.g., 360 degree cameras), long-range cameras (not shown in FIG. **4**A), mid-range camera(s) (not shown in FIG. **4**A), speed sensor(s) **444** (e.g., for measuring speed of vehicle **400**), vibration sensor(s) **442**, steering sensor(s) **440**, brake sensor(s) (e.g., as part of brake sensor system **446**), and/or other sensor types.

[0094] In at least one embodiment, one or more of controller(s) **436** may receive inputs (e.g., represented by input data) from an instrument cluster **432** of vehicle **400** and provide outputs (e.g., represented by output data, display data, etc.) via a human-machine interface ("HMI") display **434**, an audible annunciator, a loudspeaker, and/or via other components of vehicle **400**. In at least one embodiment, outputs may include information such as vehicle velocity, speed, time, map data (e.g., a High Definition map (not shown in FIG. **4**A), location data (e.g., vehicle's **400** location, such as on a map), direction, location of other vehicles (e.g., an occupancy grid), information about objects and status of objects as perceived by controller(s) **436**, etc. For example, in at least one embodiment, HMI display **434** may display information about presence of one or more objects (e.g., a street sign, caution sign, traffic light changing, etc.), and/or information about driving maneuvers vehicle has made, is making, or will make (e.g., changing lanes now, taking exit **34**B in two miles, etc.).

[0095] In at least one embodiment, vehicle **400** further includes a network interface **424** which may use wireless antenna(s) **426** and/or modem(s) to communicate over one or more networks. For example, in at least one embodiment, network interface **424** may be capable of communication over Long-Term Evolution ("LTE"), Wideband Code Division Multiple Access ("WCDMA"), Universal Mobile Telecommunications System ("UMTS"), Global System for Mobile communication ("GSM"), IMT-CDMA Multi-Carrier ("CDMA2000") networks, etc. In at least one embodiment, wireless antenna(s) **426** may also enable communication between objects in environment (e.g., vehicles, mobile devices, etc.), using local area network(s), such as Bluetooth, Bluetooth Low Energy ("LE"), Z-Wave, ZigBee, etc., and/or low power wide-area network(s) ("LPWANs"), such as LoRaWAN, SigFox, etc. protocols.

[0096] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, inference and/or training logic **115** may be used in system FIG. **4**A for inferencing or predicting operations based, at least in part, on weight parameters calculated using neural network training operations, neural network functions and/or architectures, or neural network use cases described herein.

[0097] FIG. **4**B illustrates an example of camera locations and fields of view for autonomous vehicle **400** of FIG. **4**A, according to at least one embodiment. In at least one embodiment, cameras and respective fields of view are one example embodiment and are not intended to be limiting. For instance, in at least one embodiment, additional and/or alternative cameras may be included, and/or cameras may be located at different locations on vehicle **400**.

[0098] In at least one embodiment, camera types for cameras may include, but are not limited to, digital cameras that may be adapted for use with components and/or systems of vehicle **400**. In at least one embodiment, camera(s) may operate at automotive safety integrity level ("ASIL") B and/or at another ASIL. In at least one embodiment, camera types may be capable of any image capture rate, such as 60 frames per second (fps), 1220 fps, 240 fps, etc., depending on embodiment. In at least one embodiment, cameras may be capable of using rolling shutters, global shutters, another type of shutter, or a combination thereof. In at least one embodiment, color filter array may include a red clear clear clear ("RCCC") color filter array, a red clear clear blue ("RCCB") color filter array, a red blue green clear ("RBGC") color filter array, a Foveon X3 color filter array, a Bayer sensors ("RGGB") color filter array, a monochrome sensor color filter array, and/or another type of color filter array. In at least one embodiment, clear pixel cameras, such as cameras with an RCCC, an RCCB, and/or an RBGC color filter array, may be used in an effort to increase light sensitivity.

[0099] In at least one embodiment, one or more camera(s) may be used to perform advanced driver assistance systems ("ADAS") functions (e.g., as part of a redundant or fail-safe design). For example, in at least one embodiment, a Multi-Function Mono Camera may be installed to provide functions including lane departure warning, traffic sign assist, and intelligent headlamp control. In at least one embodiment, one or more camera(s) (e.g., all cameras) may record and provide image data (e.g., video) simultaneously.

[0100] In at least one embodiment, one or more camera may be mounted in a mounting assembly, such as a custom designed (three-dimensional ("3D") printed) assembly, in order to cut out stray light and reflections from within vehicle **400** (e.g., reflections from dashboard reflected in windshield mirrors) which may interfere with camera image data capture abilities. With reference to wing-mirror mounting assemblies, in at least one embodiment, wing-mirror assemblies may be custom 3D printed so that a camera mounting plate matches a shape of a wing-mirror. In at least one embodiment, camera(s) may be integrated into wing-mirrors. In at least one embodiment, for side-view cameras, camera(s) may also be integrated within four pillars at each corner of a cabin.

[0101] In at least one embodiment, cameras with a field of view that include portions of an environment in front of vehicle **400** (e.g., front-facing cameras) may be used for surround view, to help identify forward facing paths and obstacles, as well as aid in, with help of one or more of controller(s) **436** and/or control SoCs, providing information critical to generating an occupancy grid and/or determining preferred vehicle paths. In at least one embodiment, front-facing cameras may be used to perform many similar ADAS functions as LIDAR, including, without limitation, emergency braking, pedestrian detection, and collision avoidance. In at least one embodiment, front-facing cameras may also be used for ADAS functions and systems including, without limitation, Lane Departure Warnings ("LDW"), Autonomous Cruise Control ("ACC"), and/or other functions such as traffic sign recognition.

[0102] In at least one embodiment, a variety of cameras may be used in a front-facing configuration, including, for example, a monocular camera platform that includes a CMOS ("complementary metal oxide semiconductor") color imager. In at least one embodiment, a wide-view camera **470** may be used to perceive objects coming into view from a periphery (e.g., pedestrians, crossing traffic, or bicycles). Although only one wide-view camera **470** is illustrated in FIG. **4**B, in other embodiments, there may be any number (including zero) wide-view cameras on vehicle **400**. In at least one embodiment, any number of long-range camera(s) **498** (e.g., a long-view stereo camera pair) may be used for depth-based object detection, especially for objects for which a neural network has not yet been trained. In at least one embodiment, long-range camera(s) **498** may also be used for object detection and classification, as well as basic object tracking.

[0103] In at least one embodiment, any number of stereo camera(s) **468** may also be included in a front-facing configuration. In at least one embodiment, one or more of stereo camera(s) **468** may include an integrated control unit comprising a scalable processing unit, which may provide a programmable logic ("FPGA") and a multi-core microprocessor with an integrated Controller Area Network ("CAN") or Ethernet interface on a single chip. In at least one embodiment, such a unit may be used to generate a 3D map of an environment of vehicle **400**, including a distance estimate for all points in an image. In at least one embodiment, one or more of stereo camera(s) **468** may include, without limitation, compact stereo vision sensor(s) that may include, without limitation, two camera lenses (one each on left and right) and an image processing chip that may measure distance from vehicle **400** to target object and use generated information (e.g., metadata) to activate autonomous emergency braking and lane departure warning functions. In at least one embodiment, other types of stereo camera(s) **468** may be used in addition to, or alternatively from, those described herein.

[0104] In at least one embodiment, cameras with a field of view that include portions of environment to sides of vehicle **400** (e.g., side-view cameras) may be used for surround view, providing information used to create and update an occupancy grid, as well as to generate side impact collision warnings. For example, in at least one embodiment, surround camera(s) **474** (e.g., four surround cameras as illustrated in FIG. **4**B) could be positioned on vehicle **400**. In at least one embodiment, surround camera(s) **474** may include, without limitation, any number and combination of wide-view cameras, fisheye camera(s), 360 degree camera(s), and/or similar cameras. For instance, in at least one embodiment, four fisheye cameras may be positioned on a front, a rear, and sides of vehicle **400**. In at least one embodiment, vehicle **400** may use three surround camera(s) **474** (e.g., left, right, and rear), and may leverage one or more other camera(s) (e.g., a forward-facing camera) as a fourth surround-view camera.

[0105] In at least one embodiment, cameras with a field of view that include portions of an environment behind vehicle **400** (e.g., rear-view cameras) may be used for parking assistance, surround view, rear collision warnings, and creating and updating an occupancy grid. In at least one embodiment, a wide variety of cameras may be used including, but not limited to, cameras that are also suitable as a front-facing camera(s) (e.g., long-range cameras **498** and/or mid-range camera(s) **476**, stereo camera(s) **468**), infrared camera(s) **472**, etc.), as described herein.

[0106] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, inference and/or training logic **115** may be used in system FIG. **4**B for inferencing or predicting operations based, at least in part, on weight parameters calculated using neural network training operations, neural network functions and/or architectures, or neural network use cases described herein.

[0107] FIG. **4**C is a block diagram illustrating an example system architecture for autonomous vehicle **400** of FIG. **4**A, according to at least one embodiment. In at least one embodiment, each of components, features, and systems of vehicle **400** in FIG. **4**C is illustrated as being connected via a bus **402**. In at least one embodiment, bus **402** may include, without limitation, a CAN data interface (alternatively referred to herein as a "CAN bus"). In at least one embodiment, a CAN may be a network inside vehicle **400** used to aid in control of various features and functionality of vehicle **400**, such as actuation of brakes, acceleration, braking, steering, windshield wipers, etc. In at least one embodiment, bus **402** may be configured to have dozens or even hundreds of nodes, each with its own unique identifier (e.g., a CAN ID). In at least one embodiment, bus **402** may be read to find steering wheel angle, ground speed, engine revolutions per minute ("RPMs"), button positions, and/or other vehicle status indicators. In at least one embodiment, bus **402** may be a CAN bus that is ASIL B compliant.

[0108] In at least one embodiment, in addition to, or alternatively from CAN, FlexRay and/or Ethernet protocols may be used. In at least one embodiment, there may be any number of busses forming bus **402**, which may include, without limitation, zero or more CAN busses, zero or more FlexRay busses, zero or more Ethernet busses, and/or zero or more other types of busses using different protocols. In at least one embodiment, two or more busses may be used to perform different functions, and/or may be used for redundancy. For example, a first bus may be used for collision avoidance functionality and a second bus may be used for actuation control. In at least one embodiment, each bus of bus **402** may communicate with any of components of vehicle **400**, and two or more busses of bus **402** may communicate with corresponding components. In at least one embodiment, each of any number of system(s) on chip(s) ("SoC(s)") **404** (such as SoC **404**(A) and SoC **404**(B), each of controller(s) **436**, and/or each computer within a vehicle may have access to same input data (e.g., inputs from sensors of vehicle **400**), and may be connected to a common bus, such CAN bus.

[0109] In at least one embodiment, vehicle **400** may include one or more controller(s) **436**, such as those described herein with respect to FIG. **4**A. In at least one embodiment, controller(s) **436** may be used for a variety of functions. In at least one embodiment, controller(s) **436** may be coupled to any of various other components and systems of vehicle **400**, and may be used for control of vehicle **400**, artificial intelligence of vehicle **400**, infotainment for vehicle **400**, and/or other functions.

[0110] In at least one embodiment, vehicle **400** may include any number of SoCs **404**. In at least one embodiment, each of SoCs **404** may include, without limitation, central processing units ("CPU(s)") **406**, graphics processing units ("GPU(s)") **408**, processor(s) **410**, cache(s) **412**, accelerator(s) **414**, data store(s) **416**, and/or other components and features not illustrated. In at least one embodiment, SoC(s) **404** may be used to control vehicle **400** in a variety of platforms and systems. For example, in at least one embodiment, SoC(s) **404** may be combined in a system (e.g., system of vehicle **400**) with a High Definition ("HD") map **422** which may obtain map refreshes and/or updates via network interface **424** from one or more servers (not shown in FIG. **4**C).

[0111] In at least one embodiment, CPU(s) **406** may include a CPU cluster or CPU complex (alternatively referred to herein as a "CCPLEX"). In at least one embodiment, CPU(s) **406** may include multiple cores and/or level two ("L2") caches. For instance, in at least one embodiment, CPU(s) **406** may include eight cores in a coherent multiprocessor configuration. In at least one embodiment, CPU(s) **406** may include four dual-core clusters where each cluster has a dedicated L2 cache (e.g., a 2 megabyte (MB) L2 cache). In at least one embodiment, CPU(s) **406** (e.g., CCPLEX) may be configured to support simultaneous cluster operations enabling any combination of clusters of CPU(s) **406** to be active at any given time.

[0112] In at least one embodiment, one or more of CPU(s) **406** may implement power management capabilities that include, without limitation, one or more of following features: individual hardware blocks may be clock-gated automatically when idle to save dynamic power; each core clock may be gated when such core is not actively executing instructions due to execution of Wait for Interrupt ("WFI")/Wait for Event ("WFE") instructions; each core may be independently power-gated; each core cluster may be independently clock-gated when all cores are clock-gated or power-gated; and/or each core cluster may be independently power-gated when all cores are power-gated. In at least one embodiment, CPU(s) **406** may further implement an enhanced algorithm for managing power states, where allowed power states and expected wakeup times are specified, and hardware/microcode determines which best power state to enter for core, cluster, and CCPLEX. In at least one embodiment, processing cores may support simplified power state entry sequences in software with work offloaded to microcode.

[0113] In at least one embodiment, GPU(s) **408** may include an integrated GPU (alternatively referred to herein as an "iGPU"). In at least one embodiment, GPU(s) **408** may be programmable and may be efficient for parallel workloads. In at least one embodiment, GPU(s) **408** may use an enhanced tensor instruction set. In at least one embodiment, GPU(s) **408** may include one or more streaming microprocessors, where each streaming microprocessor may include a level one ("L1") cache (e.g., an L1 cache with at least 96 KB storage capacity), and two or more streaming microprocessors may share an L2 cache (e.g., an L2 cache with a 512 KB storage capacity). In at least one embodiment, GPU(s) **408** may include at least eight streaming microprocessors. In at least one embodiment, GPU(s) **408** may use compute application programming interface(s) (API(s)). In at least one embodiment, GPU(s) **408** may use one or more parallel computing platforms and/or programming models (e.g., NVIDIA's CUDA model).

[0114] In at least one embodiment, one or more of GPU(s) **408** may be power-optimized for best performance in automotive and embedded use cases. For example, in at least one embodiment, GPU(s) **408** could be fabricated on Fin field-effect transistor ("FinFET") circuitry. In at least one embodiment, each streaming microprocessor may incorporate a number of mixed-precision processing cores partitioned into multiple blocks. For example, and without limitation, 64 PF32 cores and 32 PF64 cores could be partitioned into four processing blocks. In at least one embodiment, each processing block could be allocated 16 FP32 cores, 8 FP64 cores, 16 INT32 cores, two mixed-precision NVIDIA Tensor cores for deep learning matrix arithmetic, a level zero ("L0") instruction cache, a warp scheduler, a dispatch unit, and/or a 64 KB register file. In at least one embodiment, streaming microprocessors may include independent parallel integer and floating-point data paths to provide for efficient execution of workloads with a mix of computation and addressing calculations. In at least one embodiment, streaming microprocessors may include independent thread scheduling capability to enable finer-grain synchronization and cooperation between parallel threads. In at least one embodiment, streaming microprocessors may include a combined L1 data cache and shared memory unit in order to improve performance while simplifying programming.

[0115] In at least one embodiment, one or more of GPU(s) **408** may include a high bandwidth memory ("HBM) and/or a 16 GB HBM2 memory subsystem to provide, in some examples, about 900 GB/second peak memory bandwidth. In at least one embodiment, in addition to, or alternatively from, HBM memory, a synchronous graphics random-access memory ("SGRAM") may be used, such as a graphics double data rate type five synchronous random-access memory ("GDDR5").

[0116] In at least one embodiment, GPU(s) **408** may include unified memory technology. In at least one embodiment, address translation services ("ATS") support may be used to allow GPU(s) **408** to access CPU(s) **406** page tables directly. In at least one embodiment, embodiment, when a GPU of GPU(s) **408** memory management unit ("MMU") experiences a miss, an address translation request may be transmitted to CPU(s) **406**. In response, 2 CPU of CPU(s) **406** may look in its page tables for a virtual-to-physical mapping for an address and transmit translation back to GPU(s) **408**, in at least one embodiment. In at least one embodiment, unified memory technology may allow a single unified virtual address space for memory of both CPU(s) **406** and GPU(s) **408**, thereby simplifying GPU(s) **408** programming and porting of applications to GPU(s) **408**.

[0117] In at least one embodiment, GPU(s) **408** may include any number of access counters that may keep track of frequency of access of GPU(s) **408** to memory of other processors. In at least one embodiment, access counter(s) may help ensure that memory pages are moved to physical memory of a processor that is accessing pages most frequently, thereby improving efficiency for memory ranges shared between processors.

[0118] In at least one embodiment, one or more of SoC(s) **404** may include any number of cache(s) **412**, including those described herein. For example, in at least one embodiment, cache(s) **412** could include a level three ("L3") cache that is available to both CPU(s) **406** and GPU(s) **408** (e.g., that is connected to CPU(s) **406** and GPU(s) **408**). In at least one embodiment, cache(s) **412** may include a write-back cache that may keep track of states of lines, such as by using a cache coherence protocol (e.g., MEI, MESI, MSI, etc.). In at least one embodiment, a L3 cache may include 4 MB of memory or more, depending on embodiment, although smaller cache sizes may be used.

[0119] In at least one embodiment, one or more of SoC(s) **404** may include one or more accelerator(s) **414** (e.g., hardware accelerators, software accelerators, or a combination thereof). In at least one embodiment, SoC(s) **404** may include a hardware acceleration cluster that may include optimized hardware accelerators and/or large on-chip memory. In at least one embodiment, large on-chip memory (e.g., 4 MB of SRAM), may enable a hardware acceleration cluster to accelerate neural networks and other calculations. In at least one embodiment, a hardware acceleration cluster may be used to complement GPU(s) **408** and to off-load some of tasks of GPU(s) **408** (e.g., to free up more cycles of GPU(s) **408** for performing other tasks). In at least one embodiment, accelerator(s) **414** could be used for targeted workloads (e.g., perception, convolutional neural networks ("CNNs"), recurrent neural networks ("RNNs"), etc.) that are stable enough to be amenable to acceleration. In at least one embodiment, a CNN may include a region-based or regional convolutional neural networks ("RCNNs") and Fast RCNNs (e.g., as used for object detection) or other type of CNN.

[0120] In at least one embodiment, accelerator(s) **414** (e.g., hardware acceleration cluster) may include one or more deep learning accelerator ("DLA"). In at least one embodiment, DLA(s) may include, without limitation, one or more Tensor processing units ("TPUs") that may be configured to provide an additional ten trillion operations per second for deep learning applications and inferencing. In at least one embodiment, TPUs may be accelerators configured to, and optimized for, performing image processing functions (e.g., for CNNs, RCNNs, etc.). In at least one embodiment, DLA(s) may further be optimized for a specific set of neural network types and floating point operations, as well as inferencing. In at least one embodiment, design of DLA(s) may provide more performance per millimeter than a typical general-purpose GPU, and typically vastly exceeds performance of a CPU. In at least one embodiment, TPU(s) may perform several functions, including a single-instance convolution function, supporting, for example, INT8, INT16, and FP16 data types for both features and weights, as well as post-processor functions. In at least one embodiment, DLA(s) may quickly and efficiently execute neural networks, especially CNNs, on processed or unprocessed data for any of a variety of functions, including, for example and without limitation: a CNN for object identification and detection using data from camera sensors; a CNN for distance estimation using data from camera sensors; a CNN for emergency vehicle detection and identification and detection using data from microphones; a CNN for facial recognition and vehicle owner identification using data from camera sensors; and/or a CNN for security and/or safety related events.

[0121] In at least one embodiment, DLA(s) may perform any function of GPU(s) **408**, and by using an inference accelerator, for example, a designer may target either DLA (s) or GPU(s) **408** for any function. For example, in at least one embodiment, a designer may focus processing of CNNs and floating point operations on DLA(s) and leave other functions to GPU(s) **408** and/or accelerator(s) **414**.

[0122] In at least one embodiment, accelerator(s) **414** may include programmable vision accelerator ("PVA"), which may alternatively be referred to herein as a computer vision accelerator. In at least one embodiment, PVA may be designed and configured to accelerate computer vision algorithms for advanced driver assistance system ("ADAS") **438**, autonomous driving, augmented reality ("AR") applications, and/or virtual reality ("VR") applications. In at least one embodiment, PVA may provide a balance between performance and flexibility. For example, in at least one embodiment, each PVA may include, for example and without limitation, any number of reduced instruction set computer ("RISC") cores, direct memory access ("DMA"), and/or any number of vector processors.

[0123] In at least one embodiment, RISC cores may interact with image sensors (e.g., image sensors of any cameras described herein), image signal processor(s), etc. In at least one embodiment, each RISC core may include any amount of memory. In at least one embodiment, RISC cores may use any of a number of protocols, depending on embodiment. In at least one embodiment, RISC cores may execute a real-time operating system ("RTOS"). In at least one embodiment, RISC cores may be implemented using one or more integrated circuit devices, application specific integrated circuits ("ASICs"), and/or memory devices. For example, in at least one embodiment, RISC cores could include an instruction cache and/or a tightly coupled RAM.

[0124] In at least one embodiment, DMA may enable components of PVA to access system memory independently of CPU(s) **406**. In at least one embodiment, DMA may support any number of features used to provide optimization to a PVA including, but not limited to, supporting multi-dimensional addressing and/or circular addressing. In at least one embodiment, DMA may support up to six or more dimensions of addressing, which may include, without limitation, block width, block height, block depth, horizontal block stepping, vertical block stepping, and/or depth stepping.

[0125] In at least one embodiment, vector processors may be programmable processors that may be designed to efficiently and flexibly execute programming for computer vision algorithms and provide signal processing capabilities. In at least one embodiment, a PVA may include a PVA core and two vector processing subsystem partitions. In at least one embodiment, a PVA core may include a processor subsystem, DMA engine(s) (e.g., two DMA engines), and/or other peripherals. In at least one embodiment, a vector processing subsystem may operate as a primary processing engine of a PVA, and may include a vector processing unit ("VPU"), an instruction cache, and/or vector memory (e.g., "VMEM"). In at least one embodiment, VPU core may include a digital signal processor such as, for example, a single instruction, multiple data ("SIMD"), very long instruction word ("VLIW") digital signal processor. In at least one embodiment, a combination of SIMD and VLIW may enhance throughput and speed.

[0126] In at least one embodiment, each of vector processors may include an instruction cache and may be coupled to dedicated memory. As a result, in at least one embodiment, each of vector processors may be configured to execute independently of other vector processors. In at least one embodiment, vector processors that are included in a particular PVA may be configured to employ data parallelism. For instance, in at least one embodiment, plurality of vector processors included in a single PVA may execute a common computer vision algorithm, but on different regions of an image. In at least one embodiment, vector processors included in a particular PVA may simultaneously execute different computer vision algorithms, on one image, or even execute different algorithms on sequential images or portions of an image. In at least one embodiment, among other things, any number of PVAs may be included in hardware acceleration cluster and any number of vector processors may be included in each PVA. In at least one embodiment, PVA may include additional error correcting code ("ECC") memory, to enhance overall system safety.

[0127] In at least one embodiment, accelerator(s) **414** may include a computer vision network on-chip and static random-access memory ("SRAM"), for providing a high-bandwidth, low latency SRAM for accelerator(s) **414**. In at least one embodiment, on-chip memory may include at least 4 MB SRAM, comprising, for example and without limitation, eight field-configurable memory blocks, that may be accessible by both a PVA and a DLA. In at least one embodiment, each pair of memory blocks may include an advanced peripheral bus ("APB") interface, configuration circuitry, a controller, and a multiplexer. In at least one embodiment, any type of memory may be used. In at least one embodiment, a PVA and a DLA may access memory via a backbone that provides a PVA and a DLA with high-speed access to memory. In at least one embodiment, a backbone may include a computer vision network on-chip that interconnects a PVA and a DLA to memory (e.g., using APB).

[0128] In at least one embodiment, a computer vision network on-chip may include an interface that determines, before transmission of any control signal/address/data, that both a PVA and a DLA provide ready and valid signals. In at least one embodiment, an interface may provide for separate phases and separate channels for transmitting control signals/addresses/data, as well as burst-type communications for continuous data transfer. In at least one embodiment, an interface may comply with International Organization for Standardization ("ISO") 26262 or International Electrotechnical Commission ("IEC") 61508 standards, although other standards and protocols may be used.

[0129] In at least one embodiment, one or more of SoC(s) **404** may include a real-time ray-tracing hardware accelerator. In at least one embodiment, real-time ray-tracing hardware accelerator may be used to quickly and efficiently determine positions and extents of objects (e.g., within a world model), to generate real-time visualization simulations, for RADAR signal interpretation, for sound propagation synthesis and/or analysis, for simulation of SONAR systems, for general wave propagation simulation, for comparison to LIDAR data for purposes of localization and/or other functions, and/or for other uses.

[0130] In at least one embodiment, accelerator(s) **414** can have a wide array of uses for autonomous driving. In at least one embodiment, a PVA may be used for key processing stages in ADAS and autonomous vehicles. In at least one embodiment, a PVA's capabilities are a good match for algorithmic domains needing predictable processing, at low power and low latency. In other words, a PVA performs well on semi-dense or dense regular computation, even on small data sets, which might require predictable run-times with low latency and low power. In at least one embodiment, such as in vehicle **400**, PVAs might be designed to run classic computer vision algorithms, as they can be efficient at object detection and operating on integer math.

[0131] For example, according to at least one embodiment of technology, a PVA is used to perform computer stereo vision. In at least one embodiment, a semi-global matching-based algorithm may be used in some examples, although this is not intended to be limiting. In at least one embodiment, applications for Level 3-5 autonomous driving use motion estimation/stereo matching on-the-fly (e.g., structure from motion, pedestrian recognition, lane detection, etc.). In at least one embodiment, a PVA may perform computer stereo vision functions on inputs from two monocular cameras.

[0132] In at least one embodiment, a PVA may be used to perform dense optical flow. For example, in at least one embodiment, a PVA could process raw RADAR data (e.g., using a 4D Fast Fourier Transform) to provide processed RADAR data. In at least one embodiment, a PVA is used for time of flight depth processing, by processing raw time of flight data to provide processed time of flight data, for example.

[0133] In at least one embodiment, a DLA may be used to run any type of network to enhance control and driving safety, including for example and without limitation, a neural network that outputs a measure of confidence for each object detection. In at least one embodiment, confidence may be represented or interpreted as a probability, or as providing a relative "weight" of each detection compared to other detections. In at least one embodiment, a confidence measure enables a system to make further decisions regarding which detections should be considered as true positive detections rather than false positive detections. In at least one embodiment, a system may set a threshold value for confidence and consider only detections exceeding threshold value as true positive detections. In an embodiment in which an automatic emergency braking ("AEB") system is used, false positive detections would cause vehicle to automatically perform emergency braking, which is obviously undesirable. In at least one embodiment, highly confident detections may be considered as triggers for AEB In at least one embodiment, a DLA may run a neural network for regressing confidence value. In at least one embodiment, neural network may take as its input at least some subset of parameters, such as bounding box dimensions, ground plane estimate obtained (e.g., from another subsystem), output from IMU sensor(s) **466** that correlates with vehicle **400** orientation, distance, 3D location estimates of object obtained from neural network and/or other sensors (e.g., LIDAR sensor(s) **464** or RADAR sensor(s) **460**), among others.

[0134] In at least one embodiment, one or more of SoC(s) **404** may include data store(s) **416** (e.g., memory). In at least one embodiment, data store(s) **416** may be on-chip memory of SoC(s) **404**, which may store neural networks to be executed on GPU(s) **408** and/or a DLA. In at least one embodiment, data store(s) **416** may be large enough in capacity to store multiple instances of neural networks for redundancy and safety. In at least one embodiment, data store(s) **416** may comprise L2 or L3 cache(s).

[0135] In at least one embodiment, one or more of SoC(s) **404** may include any number of processor(s) **410** (e.g., embedded processors). In at least one embodiment, processor(s) **410** may include a boot and power management processor that may be a dedicated processor and subsystem to handle boot power and management functions and related security enforcement. In at least one embodiment, a boot and power management processor may be a part of a boot sequence of SoC(s) **404** and may provide runtime power management services. In at least one embodiment, a boot power and management processor may provide clock and voltage programming, assistance in system low power state transitions, management of SoC(s) **404** thermals and temperature sensors, and/or management of SoC(s) **404** power states. In at least one embodiment, each temperature sensor may be implemented as a ring-oscillator whose output frequency is proportional to temperature, and SoC(s) **404** may use ring-oscillators to detect temperatures of CPU(s) **406**, GPU(s) **408**, and/or accelerator(s) **414**. In at least one embodiment, if temperatures are determined to exceed a threshold, then a boot and power management processor may enter a temperature fault routine and put SoC(s) **404** into a lower power state and/or put vehicle **400** into a chauffeur to safe stop mode (e.g., bring vehicle **400** to a safe stop).

[0136] In at least one embodiment, processor(s) **410** may further include a set of embedded processors that may serve as an audio processing engine which may be an audio subsystem that enables full hardware support for multi-channel audio over multiple interfaces, and a broad and flexible range of audio I/O interfaces. In at least one embodiment, an audio processing engine is a dedicated processor core with a digital signal processor with dedicated RAM.

[0137] In at least one embodiment, processor(s) **410** may further include an always-on processor engine that may provide necessary hardware features to support low power sensor management and wake use cases. In at least one embodiment, an always-on processor engine may include, without limitation, a processor core, a tightly coupled RAM, supporting peripherals (e.g., timers and interrupt controllers), various I/O controller peripherals, and routing logic.

[0138] In at least one embodiment, processor(s) **410** may further include a safety cluster engine that includes, without limitation, a dedicated processor subsystem to handle safety management for automotive applications. In at least one embodiment, a safety cluster engine may include, without limitation, two or more processor cores, a tightly coupled RAM, support peripherals (e.g., timers, an interrupt controller, etc.), and/or routing logic. In a safety mode, two or more cores may operate, in at least one embodiment, in a lockstep mode and function as a single core with comparison logic to detect any differences between their operations. In at least one embodiment, processor(s) **410** may further include a real-time camera engine that may include, without limitation, a dedicated processor subsystem for handling real-time camera management. In at least one embodiment, processor (s) **410** may further include a high-dynamic range signal processor that may include, without limitation, an image signal processor that is a hardware engine that is part of a camera processing pipeline.

[0139] In at least one embodiment, processor(s) **410** may include a video image compositor that may be a processing block (e.g., implemented on a microprocessor) that implements video post-processing functions needed by a video playback application to produce a final image for a player window. In at least one embodiment, a video image compositor may perform lens distortion correction on wide-view camera(s) **470**, surround camera(s) **474**, and/or on in-cabin monitoring camera sensor(s). In at least one embodiment, in-cabin monitoring camera sensor(s) are preferably monitored by a neural network running on another instance of SoC **404**, configured to identify in cabin events and respond accordingly. In at least one embodiment, an in-cabin system may perform, without limitation, lip reading to activate cellular service and place a phone call, dictate emails, change a vehicle's destination, activate or change a vehicle's infotainment system and settings, or provide voice-activated web surfing. In at least one embodiment, certain functions are available to a driver when a vehicle is operating in an autonomous mode and are disabled otherwise.

[0140] In at least one embodiment, a video image compositor may include enhanced temporal noise reduction for both spatial and temporal noise reduction. For example, in at least one embodiment, where motion occurs in a video, noise reduction weights spatial information appropriately, decreasing weights of information provided by adjacent frames. In at least one embodiment, where an image or portion of an image does not include motion, temporal noise reduction performed by video image compositor may use information from a previous image to reduce noise in a current image.

[0141] In at least one embodiment, a video image compositor may also be configured to perform stereo rectification on input stereo lens frames. In at least one embodiment, a video image compositor may further be used for user interface composition when an operating system desktop is in use, and GPU(s) **408** are not required to continuously render new surfaces. In at least one embodiment, when GPU(s) **408** are powered on and active doing 3D rendering, a video image compositor may be used to offload GPU(s) **408** to improve performance and responsiveness.

[0142] In at least one embodiment, one or more SoC of SoC(s) **404** may further include a mobile industry processor interface ("MIPI") camera serial interface for receiving video and input from cameras, a high-speed interface, and/or a video input block that may be used for a camera and related pixel input functions. In at least one embodiment, one or more of SoC(s) **404** may further include an input/output controller(s) that may be controlled by software and may be used for receiving I/O signals that are uncommitted to a specific role.

[0143] In at least one embodiment, one or more Soc of SoC(s) **404** may further include a broad range of peripheral interfaces to enable communication with peripherals, audio encoders/decoders ("codecs"), power management, and/or other devices. In at least one embodiment, SoC(s) **404** may be used to process data from cameras (e.g., connected over Gigabit Multimedia Serial Link and Ethernet channels), sensors (e.g., LIDAR sensor(s) **464**, RADAR sensor(s) **460**, etc. that may be connected over Ethernet channels), data from bus **402** (e.g., speed of vehicle **400**, steering wheel position, etc.), data from GNSS sensor(s) **458** (e.g., connected over an Ethernet bus or a CAN bus), etc. In at least one embodiment, one or more SoC of SoC(s) **404** may further include dedicated high-performance mass storage controllers that may include their own DMA engines, and that may be used to free CPU(s) **406** from routine data management tasks.

[0144] In at least one embodiment, SoC(s) **404** may be an end-to-end platform with a flexible architecture that spans automation Levels 3-5, thereby providing a comprehensive functional safety architecture that leverages and makes efficient use of computer vision and ADAS techniques for diversity and redundancy, and provides a platform for a flexible, reliable driving software stack, along with deep learning tools. In at least one embodiment, SoC(s) **404** may be faster, more reliable, and even more energy-efficient and space-efficient than conventional systems. For example, in at least one embodiment, accelerator(s) **414**, when combined with CPU(s) **406**, GPU(s) **408**, and data store(s) **416**, may provide for a fast, efficient platform for Level 3-5 autonomous vehicles.

[0145] In at least one embodiment, computer vision algorithms may be executed on CPUs, which may be configured using a high-level programming language, such as C, to execute a wide variety of processing algorithms across a wide variety of visual data. However, in at least one embodiment, CPUs are oftentimes unable to meet performance requirements of many computer vision applications, such as those related to execution time and power consumption, for example. In at least one embodiment, many CPUs are unable to execute complex object detection algorithms in real-time, which is used in in-vehicle ADAS applications and in practical Level 3-5 autonomous vehicles.

[0146] Embodiments described herein allow for multiple neural networks to be performed simultaneously and/or sequentially, and for results to be combined together to enable Level 3-5 autonomous driving functionality. For example, in at least one embodiment, a CNN executing on a DLA or a discrete GPU (e.g., GPU(s) **420**) may include text and word recognition, allowing reading and understanding of traffic signs, including signs for which a neural network has not been specifically trained. In at least one embodiment, a DLA may further include a neural network that is able to identify, interpret, and provide semantic understanding of a sign, and to pass that semantic understanding to path planning modules running on a CPU Complex.

[0147] In at least one embodiment, multiple neural networks may be run simultaneously, as for Level 3, 4, or 5 driving. For example, in at least one embodiment, a warning sign stating "Caution: flashing lights indicate icy conditions," along with an electric light, may be independently or collectively interpreted by several neural networks. In at least one embodiment, such warning sign itself may be identified as a traffic sign by a first deployed neural network (e.g., a neural network that has been trained), text "flashing lights indicate icy conditions" may be interpreted by a second deployed neural network, which informs a vehicle's path planning software (preferably executing on a CPU Complex) that when flashing lights are detected, icy conditions exist. In at least one embodiment, a flashing light may be identified by operating a third deployed neural network over multiple frames, informing a vehicle's path-planning software of a presence (or an absence) of flashing lights. In at least one embodiment, all three neural networks may run simultaneously, such as within a DLA and/or on GPU(s) **408**.

[0148] In at least one embodiment, a CNN for facial recognition and vehicle owner identification may use data from camera sensors to identify presence of an authorized driver and/or owner of vehicle **400**. In at least one embodiment, an always-on sensor processing engine may be used to unlock a vehicle when an owner approaches a driver door and turns on lights, and, in a security mode, to disable such vehicle when an owner leaves such vehicle. In this way, SoC(s) **404** provides for security against theft and/or car-jacking.

[0149] In at least one embodiment, a CNN for emergency vehicle detection and identification may use data from microphones **496** to detect and identify emergency vehicle sirens. In at least one embodiment, SoC(s) **404** uses a CNN for classifying environmental and urban sounds, as well as classifying visual data. In at least one embodiment, a CNN running on a DLA is trained to identify a relative closing speed of an emergency vehicle (e.g., by using a Doppler effect). In at least one embodiment, a CNN may also be trained to identify emergency vehicles specific to a local area in which a vehicle is operating, as identified by GNSS sensor(s) **458**. In at least one embodiment, when operating in Europe, a CNN will seek to detect European sirens, and when in North America, a CNN will seek to identify only North American sirens. In at least one embodiment, once an emergency vehicle is detected, a control program may be used to execute an emergency vehicle safety routine, slowing a vehicle, pulling over to a side of a road, parking a vehicle, and/or idling a vehicle, with assistance of ultrasonic sensor(s) **462**, until emergency vehicles pass.

[0150] In at least one embodiment, vehicle **400** may include CPU(s) **418** (e.g., discrete CPU(s), or dCPU(s)), that may be coupled to SoC(s) **404** via a high-speed interconnect (e.g., PCIe). In at least one embodiment, CPU(s) **418** may include an X86 processor, for example. CPU(s) **418** may be used to perform any of a variety of functions, including arbitrating potentially inconsistent results between ADAS sensors and SoC(s) **404**, and/or monitoring status and health of controller(s) **436** and/or an infotainment system on a chip ("infotainment SoC") **430**, for example.

[0151] In at least one embodiment, vehicle **400** may include GPU(s) **420** (e.g., discrete GPU(s), or dGPU(s)), that may be coupled to SoC(s) **404** via a high-speed interconnect (e.g., NVIDIA's NVLINK channel). In at least one embodiment, GPU(s) **420** may provide additional artificial intelligence functionality, such as by executing redundant and/or different neural networks, and may be used to train and/or update neural networks based at least in part on input (e.g., sensor data) from sensors of a vehicle **400**.

[0152] In at least one embodiment, vehicle **400** may further include network interface **424** which may include, without limitation, wireless antenna(s) **426** (e.g., one or more wireless antennas for different communication protocols, such as a cellular antenna, a Bluetooth antenna, etc.). In at least one embodiment, network interface **424** may be used to enable wireless connectivity to Internet cloud services (e.g., with server(s) and/or other network devices), with other vehicles, and/or with computing devices (e.g., client devices of passengers). In at least one embodiment, to communicate with other vehicles, a direct link may be established between vehicle **40** and another vehicle, and/or an indirect link may be established (e.g., across networks and over the Internet). In at least one embodiment, direct links may be provided using a vehicle-to-vehicle communication link. In at least one embodiment, a vehicle-to-vehicle communication link may provide vehicle **400** information about vehicles in proximity to vehicle **400** (e.g., vehicles in front of, on a side of, and/or behind vehicle **400**). In at least one embodiment, such aforementioned functionality may be part of a cooperative adaptive cruise control functionality of vehicle **400**.

[0153] In at least one embodiment, network interface **424** may include an SoC that provides modulation and demodulation functionality and enables controller(s) **436** to communicate over wireless networks. In at least one embodiment, network interface **424** may include a radio frequency front-end for up-conversion from baseband to radio frequency, and down conversion from radio frequency to baseband. In at least one embodiment, frequency conversions may be performed in any technically feasible fashion. For example, frequency conversions could be performed through well-known processes, and/or using super-heterodyne processes. In at least one embodiment, radio frequency front end functionality may be provided by a separate chip. In at least one embodiment, network interfaces may include wireless functionality for communicating over LTE, WCDMA, UMTS, GSM, CDMA2000, Bluetooth, Bluetooth LE, Wi-Fi, Z-Wave, ZigBee, LoRaWAN, and/or other wireless protocols.

[0154] In at least one embodiment, vehicle **400** may further include data store(s) **428** which may include, without limitation, off-chip (e.g., off SoC(s) **404**) storage. In at least one embodiment, data store(s) **428** may include, without limitation, one or more storage elements including RAM, SRAM, dynamic random-access memory ("DRAM"), video random-access memory ("VRAM"), flash memory, hard disks, and/or other components and/or devices that may store at least one bit of data.

[0155] In at least one embodiment, vehicle **400** may further include GNSS sensor(s) **458** (e.g., GPS and/or assisted GPS sensors), to assist in mapping, perception, occupancy grid generation, and/or path planning functions. In at least one embodiment, any number of GNSS sensor(s) **458** may be used, including, for example and without limitation, a GPS using a USB connector with an Ethernet-to-Serial (e.g., RS-232) bridge.

[0156] In at least one embodiment, vehicle **400** may further include RADAR sensor(s) **460**. In at least one embodiment, RADAR sensor(s) **460** may be used by vehicle **400** for long-range vehicle detection, even in darkness and/or severe weather conditions. In at least one embodiment, RADAR functional safety levels may be ASIL B. In at least one embodiment, RADAR sensor(s) **460** may use a CAN bus and/or bus **402** (e.g., to transmit data generated by RADAR sensor(s) **460**) for control and to access object tracking data, with access to Ethernet channels to access raw data in some examples. In at least one embodiment, a wide variety of RADAR sensor types may be used. For example, and without limitation, RADAR sensor(s) **460** may be suitable for front, rear, and side RADAR use. In at least one embodiment, one or more sensor of RADAR sensors(s) **460** is a Pulse Doppler RADAR sensor.

[0157] In at least one embodiment, RADAR sensor(s) **460** may include different configurations, such as long-range with narrow field of view, short-range with wide field of view, short-range side coverage, etc. In at least one embodiment, long-range RADAR may be used for adaptive cruise control functionality. In at least one embodiment, long-range RADAR systems may provide a broad field of view realized by two or more independent scans, such as within a 250 m (meter) range. In at least one embodiment, RADAR sensor (s) **460** may help in distinguishing between static and moving objects, and may be used by ADAS system **438** for emergency brake assist and forward collision warning. In at least one embodiment, sensors **460**(s) included in a long-range RADAR system may include, without limitation, monostatic multimodal RADAR with multiple (e.g., six or more) fixed RADAR antennae and a high-speed CAN and FlexRay interface. In at least one embodiment, with six antennae, a central four antennae may create a focused beam pattern, designed to record vehicle's **400** surroundings at higher speeds with minimal interference from traffic in adjacent lanes. In at least one embodiment, another two antennae may expand field of view, making it possible to quickly detect vehicles entering or leaving a lane of vehicle **400**.

[0158] In at least one embodiment, mid-range RADAR systems may include, as an example, a range of up to 160 m (front) or 80 m (rear), and a field of view of up to 42 degrees (front) or 150 degrees (rear). In at least one embodiment, short-range RADAR systems may include, without limitation, any number of RADAR sensor(s) **460** designed to be installed at both ends of a rear bumper. When installed at both ends of a rear bumper, in at least one embodiment, a RADAR sensor system may create two beams that constantly monitor blind spots in a rear direction and next to a vehicle. In at least one embodiment, short-range RADAR systems may be used in ADAS system **438** for blind spot detection and/or lane change assist.

[0159] In at least one embodiment, vehicle **400** may further include ultrasonic sensor(s) **462**. In at least one embodiment, ultrasonic sensor(s) **462**, which may be positioned at a front, a back, and/or side location of vehicle **400**, may be used for parking assist and/or to create and update an occupancy grid. In at least one embodiment, a wide variety of ultrasonic sensor(s) **462** may be used, and different ultrasonic sensor(s) **462** may be used for different ranges of detection (e.g., 2.5 m, 4 m). In at least one embodiment, ultrasonic sensor(s) **462** may operate at functional safety levels of ASIL B.

[0160] In at least one embodiment, vehicle **400** may include LIDAR sensor(s) **464**. In at least one embodiment, LIDAR sensor(s) **464** may be used for object and pedestrian detection, emergency braking, collision avoidance, and/or other functions. In at least one embodiment, LIDAR sensor (s) **464** may operate at functional safety level ASIL B. In at least one embodiment, vehicle **400** may include multiple LIDAR sensors **464** (e.g., two, four, six, etc.) that may use an Ethernet channel (e.g., to provide data to a Gigabit Ethernet switch).

[0161] In at least one embodiment, LIDAR sensor(s) **464** may be capable of providing a list of objects and their distances for a 360-degree field of view. In at least one embodiment, commercially available LIDAR sensor(s) **464** may have an advertised range of approximately 100 m, with an accuracy of 2 cm to 3 cm, and with support for a 100 Mbps Ethernet connection, for example. In at least one embodiment, one or more non-protruding LIDAR sensors may be used. In such an embodiment, LIDAR sensor(s) **464** may include a small device that may be embedded into a front, a rear, a side, and/or a corner location of vehicle **400**. In at least one embodiment, LIDAR sensor(s) **464**, in such an embodiment, may provide up to a 120-degree horizontal and 35-degree vertical field-of-view, with a 200 m range even for low-reflectivity objects. In at least one embodiment, front-mounted LIDAR sensor(s) **464** may be configured for a horizontal field of view between 45 degrees and 135 degrees.

[0162] In at least one embodiment, LIDAR technologies, such as 3D flash LIDAR, may also be used. In at least one embodiment, 3D flash LIDAR uses a flash of a laser as a transmission source, to illuminate surroundings of vehicle **400** up to approximately 200 m. In at least one embodiment, a flash LIDAR unit includes, without limitation, a receptor, which records laser pulse transit time and reflected light on each pixel, which in turn corresponds to a range from vehicle **400** to objects. In at least one embodiment, flash LIDAR may allow for highly accurate and distortion-free images of surroundings to be generated with every laser flash. In at least one embodiment, four flash LIDAR sensors may be deployed, one at each side of vehicle **400**. In at least one embodiment, 3D flash LIDAR systems include, without limitation, a solid-state 3D staring array LIDAR camera with no moving parts other than a fan (e.g., a non-scanning LIDAR device). In at least one embodiment, flash LIDAR device may use a 5 nanosecond class I (eye-safe) laser pulse per frame and may capture reflected laser light as a 3D range point cloud and co-registered intensity data.

[0163] In at least one embodiment, vehicle **400** may further include IMU sensor(s) **466**. In at least one embodiment, IMU sensor(s) **466** may be located at a center of a rear axle of vehicle **400**. In at least one embodiment, IMU sensor(s) **466** may include, for example and without limitation, accelerometer(s), magnetometer(s), gyroscope(s), a magnetic compass, magnetic compasses, and/or other sensor types. In at least one embodiment, such as in six-axis applications, IMU sensor(s) **466** may include, without limitation, accelerometers and gyroscopes. In at least one embodiment, such as in nine-axis applications, IMU sensor (s) **466** may include, without limitation, accelerometers, gyroscopes, and magnetometers.

[0164] In at least one embodiment, IMU sensor(s) **466** may be implemented as a miniature, high performance GPS-Aided Inertial Navigation System ("GPS/INS") that combines micro-electro-mechanical systems ("MEMS") inertial sensors, a high-sensitivity GPS receiver, and advanced Kalman filtering algorithms to provide estimates of position, velocity, and attitude. In at least one embodiment, IMU sensor(s) **466** may enable vehicle **400** to estimate its heading without requiring input from a magnetic sensor by directly observing and correlating changes in velocity from a GPS to IMU sensor(s) **466**. In at least one embodiment, IMU sensor(s) **466** and GNSS sensor(s) **458** may be combined in a single integrated unit.

[0165] In at least one embodiment, vehicle **400** may include microphone(s) **496** placed in and/or around vehicle **400**. In at least one embodiment, microphone(s) **496** may be used for emergency vehicle detection and identification, among other things.

[0166] In at least one embodiment, vehicle **400** may further include any number of camera types, including stereo camera(s) **468**, wide-view camera(s) **470**, infrared camera(s) **472**, surround camera(s) **474**, long-range camera(s) **498**, mid-range camera(s) **476**, and/or other camera types. In at least one embodiment, cameras may be used to capture image data around an entire periphery of vehicle **400**. In at least one embodiment, which types of cameras used depends on vehicle **400**. In at least one embodiment, any combination of camera types may be used to provide necessary coverage around vehicle **400**. In at least one embodiment, a number of cameras deployed may differ depending on embodiment. For example, in at least one embodiment, vehicle **400** could include six cameras, seven cameras, ten cameras, twelve cameras, or another number of cameras. In at least one embodiment, cameras may support, as an example and without limitation, Gigabit Multimedia Serial Link ("GMSL") and/or Gigabit Ethernet communications. In at least one embodiment, each camera might be as described with more detail previously herein with respect to FIG. **4**A and FIG. **4**B.

[0167] In at least one embodiment, vehicle **400** may further include vibration sensor(s) **442**. In at least one embodiment, vibration sensor(s) **442** may measure vibrations of components of vehicle **400**, such as axle(s). For example, in at least one embodiment, changes in vibrations may indicate a change in road surfaces. In at least one embodiment, when two or more vibration sensors **442** are used, differences between vibrations may be used to determine friction or slippage of road surface (e.g., when a difference in vibration is between a power-driven axle and a freely rotating axle).

[0168] In at least one embodiment, vehicle **400** may include ADAS system **438**. In at least one embodiment, ADAS system **438** may include, without limitation, an SoC, in some examples. In at least one embodiment, ADAS system **438** may include, without limitation, any number and combination of an autonomous/adaptive/automatic cruise control ("ACC") system, a cooperative adaptive cruise control ("CACC") system, a forward crash warning ("FCW") system, an automatic emergency braking ("AEB") system, a lane departure warning ("LDW)" system, a lane keep assist ("LKA") system, a blind spot warning ("BSW") system, a rear cross-traffic warning ("RCTW") system, a collision warning ("CW") system, a lane centering ("LC") system, and/or other systems, features, and/or functionality.

[0169] In at least one embodiment, ACC system may use RADAR sensor(s) **460**, LIDAR sensor(s) **464**, and/or any number of camera(s). In at least one embodiment, ACC system may include a longitudinal ACC system and/or a lateral ACC system. In at least one embodiment, a longitudinal ACC system monitors and controls distance to another vehicle immediately ahead of vehicle **400** and automatically adjusts speed of vehicle **400** to maintain a safe distance from vehicles ahead. In at least one embodiment, a lateral ACC system performs distance keeping, and advises vehicle **400** to change lanes when necessary. In at least one embodiment, a lateral ACC is related to other ADAS applications, such as LC and CW.

[0170] In at least one embodiment, a CACC system uses information from other vehicles that may be received via network interface **424** and/or wireless antenna(s) **426** from other vehicles via a wireless link, or indirectly, over a network connection (e.g., over the Internet). In at least one embodiment, direct links may be provided by a vehicle-to-vehicle ("V2V") communication link, while indirect links may be provided by an infrastructure-to-vehicle ("I2V") communication link. In general, V2V communication provides information about immediately preceding vehicles (e.g., vehicles immediately ahead of and in same lane as vehicle **400**), while I2V communication provides information about traffic further ahead. In at least one embodiment, a CACC system may include either or both I2V and V2V information sources. In at least one embodiment, given information of vehicles ahead of vehicle **400**, a CACC system may be more reliable and it has potential to improve traffic flow smoothness and reduce congestion on road.

[0171] In at least one embodiment, an FCW system is designed to alert a driver to a hazard, so that such driver may take corrective action. In at least one embodiment, an FCW system uses a front-facing camera and/or RADAR sensor(s) **460**, coupled to a dedicated processor, DSP, FPGA, and/or ASIC, that is electrically coupled to provide driver feedback, such as a display, speaker, and/or vibrating component. In at least one embodiment, an FCW system may provide a warning, such as in form of a sound, visual warning, vibration, and/or a quick brake pulse.

[0172] In at least one embodiment, an AEB system detects an impending forward collision with another vehicle or other object, and may automatically apply brakes if a driver does not take corrective action within a specified time or distance parameter. In at least one embodiment, AEB system may use front-facing camera(s) and/or RADAR sensor(s) **460**, coupled to a dedicated processor, DSP, FPGA, and/or ASIC. In at least one embodiment, when an AEB system detects a hazard, it will typically first alert a driver to take corrective action to avoid collision and, if that driver does not take corrective action, that AEB system may automatically apply brakes in an effort to prevent, or at least mitigate, an impact of a predicted collision. In at least one embodiment, an AEB system may include techniques such as dynamic brake support and/or crash imminent braking.

[0173] In at least one embodiment, an LDW system provides visual, audible, and/or tactile warnings, such as steering wheel or seat vibrations, to alert driver when vehicle **400** crosses lane markings. In at least one embodiment, an LDW system does not activate when a driver indicates an intentional lane departure, such as by activating a turn signal. In at least one embodiment, an LDW system may use front-side facing cameras, coupled to a dedicated processor, DSP, FPGA, and/or ASIC, that is electrically coupled to provide driver feedback, such as a display, speaker, and/or vibrating component. In at least one embodiment, an LKA system is a variation of an LDW system. In at least one embodiment, an LKA system provides steering input or braking to correct vehicle **400** if vehicle **400** starts to exit its lane.

[0174] In at least one embodiment, a BSW system detects and warns a driver of vehicles in an automobile's blind spot. In at least one embodiment, a BSW system may provide a visual, audible, and/or tactile alert to indicate that merging or changing lanes is unsafe. In at least one embodiment, a BSW system may provide an additional warning when a driver uses a turn signal. In at least one embodiment, a BSW system may use rear-side facing camera(s) and/or RADAR sensor(s) **460**, coupled to a dedicated processor, DSP, FPGA, and/or ASIC, that is electrically coupled to driver feedback, such as a display, speaker, and/or vibrating component.

[0175] In at least one embodiment, an RCTW system may provide visual, audible, and/or tactile notification when an object is detected outside a rear-camera range when vehicle **400** is backing up. In at least one embodiment, an RCTW system includes an AEB system to ensure that vehicle brakes are applied to avoid a crash. In at least one embodiment, an RCTW system may use one or more rear-facing RADAR sensor(s) **460**, coupled to a dedicated processor, DSP, FPGA, and/or ASIC, that is electrically coupled to provide driver feedback, such as a display, speaker, and/or vibrating component.

[0176] In at least one embodiment, conventional ADAS systems may be prone to false positive results which may be annoying and distracting to a driver, but typically are not catastrophic, because conventional ADAS systems alert a driver and allow that driver to decide whether a safety condition truly exists and act accordingly. In at least one embodiment, vehicle **400** itself decides, in case of conflicting results, whether to heed result from a primary computer or a secondary computer (e.g., a first controller or a second controller of controllers **436**). For example, in at least one embodiment, ADAS system **438** may be a backup and/or secondary computer for providing perception information to a backup computer rationality module. In at least one embodiment, a backup computer rationality monitor may run redundant diverse software on hardware components to detect faults in perception and dynamic driving tasks. In at least one embodiment, outputs from ADAS system **438** may be provided to a supervisory MCU. In at least one embodiment, if outputs from a primary computer and outputs from a secondary computer conflict, a supervisory MCU determines how to reconcile conflict to ensure safe operation.

[0177] In at least one embodiment, a primary computer may be configured to provide a supervisory MCU with a confidence score, indicating that primary computer's confidence in a chosen result. In at least one embodiment, if that confidence score exceeds a threshold, that supervisory MCU may follow that primary computer's direction, regardless of whether that secondary computer provides a conflicting or inconsistent result. In at least one embodiment, where a confidence score does not meet a threshold, and where primary and secondary computers indicate different results (e.g., a conflict), a supervisory MCU may arbitrate between computers to determine an appropriate outcome.

[0178] In at least one embodiment, a supervisory MCU may be configured to run a neural network(s) that is trained and configured to determine, based at least in part on outputs from a primary computer and outputs from a secondary computer, conditions under which that secondary computer provides false alarms. In at least one embodiment, neural network(s) in a supervisory MCU may learn when a secondary computer's output may be trusted, and when it cannot. For example, in at least one embodiment, when that secondary computer is a RADAR-based FCW system, a neural network(s) in that supervisory MCU may learn when an FCW system is identifying metallic objects that are not, in fact, hazards, such as a drainage grate or manhole cover that triggers an alarm. In at least one embodiment, when a secondary computer is a camera-based LDW system, a neural network in a supervisory MCU may learn to override LDW when bicyclists or pedestrians are present and a lane departure is, in fact, a safest maneuver. In at least one embodiment, a supervisory MCU may include at least one of a DLA or a GPU suitable for running neural network(s) with associated memory. In at least one embodiment, a supervisory MCU may comprise and/or be included as a component of SoC(s) **404**.

[0179] In at least one embodiment, ADAS system **438** may include a secondary computer that performs ADAS functionality using traditional rules of computer vision. In at least one embodiment, that secondary computer may use classic computer vision rules (if-then), and presence of a neural network(s) in a supervisory MCU may improve reliability, safety, and performance. For example, in at least one embodiment, diverse implementation and intentional non-identity makes an overall system more fault-tolerant, especially to faults caused by software (or software-hardware interface) functionality. For example, in at least one embodiment, if there is a software bug or error in software running on a primary computer, and non-identical software code running on a secondary computer provides a consistent overall result, then a supervisory MCU may have greater confidence that an overall result is correct, and a bug in software or hardware on that primary computer is not causing a material error.

[0180] In at least one embodiment, an output of ADAS system **438** may be fed into a primary computer's perception block and/or a primary computer's dynamic driving task block. For example, in at least one embodiment, if ADAS system **438** indicates a forward crash warning due to an object immediately ahead, a perception block may use this information when identifying objects. In at least one embodiment, a secondary computer may have its own neural network that is trained and thus reduces a risk of false positives, as described herein.

[0181] In at least one embodiment, vehicle **400** may further include infotainment SoC **430** (e.g., an in-vehicle infotainment system (IVI)). Although illustrated and described as an SoC, infotainment system SoC **430**, in at least one embodiment, may not be an SoC, and may include, without limitation, two or more discrete components. In at least one embodiment, infotainment SoC **430** may include, without limitation, a combination of hardware and software that may be used to provide audio (e.g., music, a personal digital assistant, navigational instructions, news, radio, etc.), video (e.g., TV, movies, streaming, etc.), phone (e.g., hands-free calling), network connectivity (e.g., LTE, WiFi, etc.), and/or information services (e.g., navigation systems, rear-parking assistance, a radio data system, vehicle related information such as fuel level, total distance covered, brake fuel level, oil level, door open/close, air filter information, etc.) to vehicle **400**. For example, infotainment SoC **430** could include radios, disk players, navigation systems, video players, USB and Bluetooth connectivity, carputers, in-car entertainment, WiFi, steering wheel audio controls, hands free voice control, a heads-up display ("HUD"), HMI display **434**, a telematics device, a control panel (e.g., for controlling and/or interacting with various components, features, and/or systems), and/or other components. In at least one embodiment, infotainment SoC **430** may further be used to provide information (e.g., visual and/or audible) to user(s) of vehicle **400**, such as information from ADAS system **438**, autonomous driving information such as planned vehicle maneuvers, trajectories, surrounding environment information (e.g., intersection information, vehicle information, road information, etc.), and/or other information.

[0182] In at least one embodiment, infotainment SoC **430** may include any amount and type of GPU functionality. In at least one embodiment, infotainment SoC **430** may communicate over bus **402** with other devices, systems, and/or components of vehicle **400**. In at least one embodiment, infotainment SoC **430** may be coupled to a supervisory MCU such that a GPU of an infotainment system may perform some self-driving functions in event that primary controller(s) **436** (e.g., primary and/or backup computers of vehicle **400**) fail. In at least one embodiment, infotainment SoC **430** may put vehicle **400** into a chauffeur to safe stop mode, as described herein.

[0183] In at least one embodiment, vehicle **400** may further include instrument cluster **432** (e.g., a digital dash, an electronic instrument cluster, a digital instrument panel, etc.). In at least one embodiment, instrument cluster **432** may include, without limitation, a controller and/or supercomputer (e.g., a discrete controller or supercomputer). In at least one embodiment, instrument cluster **432** may include, without limitation, any number and combination of a set of instrumentation such as a speedometer, fuel level, oil pressure, tachometer, odometer, turn indicators, gearshift position indicator, seat belt warning light(s), parking-brake warning light(s), engine-malfunction light(s), supplemental restraint system (e.g., airbag) information, lighting controls, safety system controls, navigation information, etc. In some examples, information may be displayed and/or shared among infotainment SoC **430** and instrument cluster **432**. In at least one embodiment, instrument cluster **432** may be included as part of infotainment SoC **430**, or vice versa.

[0184] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, inference and/or training logic **115** may be used in system FIG. **4**C for inferencing or predicting operations based, at least in part, on weight parameters calculated using neural network training operations, neural network functions and/or architectures, or neural network use cases described herein.

[0185] FIG. **4**D is a diagram of a system **476** for communication between cloud-based server(s) and autonomous vehicle **400** of FIG. **4**A, according to at least one embodiment. In at least one embodiment, system **476** may include, without limitation, server(s) **478**, network(s) **490**, and any number and type of vehicles, including vehicle **400**. In at least one embodiment, server(s) **478** may include, without limitation, a plurality of GPUs **484**(A)-**484**(H) (collectively referred to herein as GPUs **484**), PCIe switches **482**(A)-**482**(D) (collectively referred to herein as PCIe switches **482**), and/or CPUs **480**(A)-**480**(B) (collectively referred to herein as CPUs **480**). In at least one embodiment, GPUs **484**, CPUs **480**, and PCIe switches **482** may be interconnected with high-speed interconnects such as, for example and without limitation, NVLink interfaces **488** developed by NVIDIA and/or PCIe connections **486**. In at least one embodiment, GPUs **484** are connected via an NVLink and/or NVSwitch SoC and GPUs **484** and PCIe switches **482** are connected via PCIe interconnects. Although eight GPUs **484**, two CPUs **480**, and four PCIe switches **482** are illustrated, this is not intended to be limiting. In at least one embodiment, each of server(s) **478** may include, without limitation, any number of GPUs **484**, CPUs **480**, and/or PCIe switches **482**, in any combination. For example, in at least one embodiment, server(s) **478** could each include eight, sixteen, thirty-two, and/or more GPUs **484**.

[0186] In at least one embodiment, server(s) **478** may receive, over network(s) **490** and from vehicles, image data representative of images showing unexpected or changed road conditions, such as recently commenced road-work. In at least one embodiment, server(s) **478** may transmit, over network(s) **490** and to vehicles, neural networks **492**, updated or otherwise, and/or map information **494**, including, without limitation, information regarding traffic and road conditions. In at least one embodiment, updates to map information **494** may include, without limitation, updates for HD map **422**, such as information regarding construction sites, potholes, detours, flooding, and/or other obstructions. In at least one embodiment, neural networks **492**, and/or map information **494** may have resulted from new training and/or experiences represented in data received from any number of vehicles in an environment, and/or based at least in part on training performed at a data center (e.g., using server(s) **478** and/or other servers).

[0187] In at least one embodiment, server(s) **478** may be used to train machine learning models (e.g., neural networks) based at least in part on training data. In at least one embodiment, training data may be generated by vehicles, and/or may be generated in a simulation (e.g., using a game engine). In at least one embodiment, any amount of training data is tagged (e.g., where associated neural network benefits from supervised learning) and/or undergoes other preprocessing. In at least one embodiment, any amount of training data is not tagged and/or pre-processed (e.g., where associated neural network does not require supervised learning). In at least one embodiment, once machine learning models are trained, machine learning models may be used by vehicles (e.g., transmitted to vehicles over network(s) **490**), and/or machine learning models may be used by server(s) **478** to remotely monitor vehicles.

[0188] In at least one embodiment, server(s) **478** may receive data from vehicles and apply data to up-to-date real-time neural networks for real-time intelligent inferencing. In at least one embodiment, server(s) **478** may include deep-learning supercomputers and/or dedicated AI computers powered by GPU(s) **484**, such as a DGX and DGX Station machines developed by NVIDIA. However, in at least one embodiment, server(s) **478** may include deep learning infrastructure that uses CPU-powered data centers.

[0189] In at least one embodiment, deep-learning infrastructure of server(s) **478** may be capable of fast, real-time inferencing, and may use that capability to evaluate and verify health of processors, software, and/or associated hardware in vehicle **400**. For example, in at least one embodiment, deep-learning infrastructure may receive periodic updates from vehicle **400**, such as a sequence of images and/or objects that vehicle **400** has located in that sequence of images (e.g., via computer vision and/or other machine learning object classification techniques). In at least one embodiment, deep-learning infrastructure may run its own neural network to identify objects and compare them with objects identified by vehicle **400** and, if results do not match and deep-learning infrastructure concludes that AI in vehicle **400** is malfunctioning, then server(s) **478** may transmit a signal to vehicle **400** instructing a fail-safe computer of vehicle **400** to assume control, notify passengers, and complete a safe parking maneuver.

[0190] In at least one embodiment, server(s) **478** may include GPU(s) **484** and one or more programmable inference accelerators (e.g., NVIDIA's TensorRT 3 devices). In at least one embodiment, a combination of GPU-powered servers and inference acceleration may make real-time responsiveness possible. In at least one embodiment, such as where performance is less critical, servers powered by CPUs, FPGAs, and other processors may be used for inferencing. In at least one embodiment, hardware structure(s) **115** are used to perform one or more embodiments. Details regarding hardware structure(x) **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B.

Computer Systems

[0191] FIG. **5** is a block diagram illustrating an exemplary computer system, which may be a system with interconnected devices and components, a system-on-a-chip (SOC) or some combination thereof formed with a processor that may include execution units to execute an instruction, according to at least one embodiment. In at least one embodiment, a computer system **500** may include, without limitation, a component, such as a processor **502** to employ execution units including logic to perform algorithms for process data, in accordance with present disclosure, such as in embodiment described herein. In at least one embodiment, computer system **500** may include processors, such as PENTIUM® Processor family, Xeon™ Itanium®, XScale™ and/or StrongARM™, Intel® Core™, or Intel® Nervana™ microprocessors available from Intel Corporation of Santa Clara, Calif., although other systems (including PCs having other microprocessors, engineering workstations, set-top boxes and like) may also be used. In at least one embodiment, computer system **500** may execute a version of WINDOWS operating system available from Microsoft Corporation of Redmond, Wash., although other operating systems (UNIX and Linux, for example), embedded software, and/or graphical user interfaces, may also be used.

[0192] Embodiments may be used in other devices such as handheld devices and embedded applications. Some examples of handheld devices include cellular phones, Internet Protocol devices, digital cameras, personal digital assistants ("PDAs"), and handheld PCs. In at least one embodiment, embedded applications may include a microcontroller, a digital signal processor ("DSP"), system on a chip, network computers ("NetPCs"), set-top boxes, network hubs, wide area network ("WAN") switches, or any other system that may perform one or more instructions in accordance with at least one embodiment.

[0193] In at least one embodiment, computer system **500** may include, without limitation, processor **502** that may include, without limitation, one or more execution units **508** to perform machine learning model training and/or inferencing according to techniques described herein. In at least one embodiment, computer system **500** is a single processor desktop or server system, but in another embodiment, computer system **500** may be a multiprocessor system. In at least one embodiment, processor **502** may include, without limitation, a complex instruction set computer ("CISC") microprocessor, a reduced instruction set computing ("RISC") microprocessor, a very long instruction word ("VLIW") microprocessor, a processor implementing a combination of instruction sets, or any other processor device, such as a digital signal processor, for example. In at least one embodiment, processor **502** may be coupled to a processor bus **510** that may transmit data signals between processor **502** and other components in computer system **500**.

[0194] In at least one embodiment, processor **502** may include, without limitation, a Level 1 ("L1") internal cache memory ("cache") **504**. In at least one embodiment, processor **502** may have a single internal cache or multiple levels of internal cache. In at least one embodiment, cache memory may reside external to processor **502**. Other embodiments may also include a combination of both internal and external caches depending on particular implementation and needs. In at least one embodiment, a register file **506** may store different types of data in various registers including, without limitation, integer registers, floating point registers, status registers, and an instruction pointer register.

[0195] In at least one embodiment, execution unit **508**, including, without limitation, logic to perform integer and floating point operations, also resides in processor **502**. In at least one embodiment, processor **502** may also include a microcode ("ucode") read only memory ("ROM") that stores microcode for certain macro instructions. In at least one embodiment, execution unit **508** may include logic to handle a packed instruction set **509**. In at least one embodiment, by including packed instruction set **509** in an instruction set of a general-purpose processor, along with associated circuitry to execute instructions, operations used by many multimedia applications may be performed using packed data in processor **502**. In at least one embodiment, many multimedia applications may be accelerated and executed more efficiently by using a full width of a processor's data bus for performing operations on packed data, which may eliminate a need to transfer smaller units of data across that processor's data bus to perform one or more operations one data element at a time.

[0196] In at least one embodiment, execution unit **508** may also be used in microcontrollers, embedded processors, graphics devices, DSPs, and other types of logic circuits. In at least one embodiment, computer system **500** may include, without limitation, a memory **520**. In at least one embodiment, memory **520** may be a Dynamic Random Access Memory ("DRAM") device, a Static Random Access Memory ("SRAM") device, a flash memory device, or another memory device. In at least one embodiment, memory **520** may store instruction(s) **519** and/or data **521** represented by data signals that may be executed by processor **502**.

[0197] In at least one embodiment, a system logic chip may be coupled to processor bus **510** and memory **520**. In at least one embodiment, a system logic chip may include, without limitation, a memory controller hub ("MCH") **516**, and processor **502** may communicate with MCH **516** via processor bus **510**. In at least one embodiment, MCH **516** may provide a high bandwidth memory path **518** to memory **520** for instruction and data storage and for storage of graphics commands, data and textures. In at least one embodiment, MCH **516** may direct data signals between processor **502**, memory **520**, and other components in computer system **500** and to bridge data signals between processor bus **510**, memory **520**, and a system I/O interface **522**. In at least one embodiment, a system logic chip may provide a graphics port for coupling to a graphics controller. In at least one embodiment, MCH **516** may be coupled to memory **520** through high bandwidth memory path **518** and a graphics/video card **512** may be coupled to MCH **516** through an Accelerated Graphics Port ("AGP") interconnect **514**.

[0198] In at least one embodiment, computer system **500** may use system I/O interface **522** as a proprietary hub interface bus to couple MCH **516** to an I/O controller hub ("ICH") **530**. In at least one embodiment, ICH **530** may provide direct connections to some I/O devices via a local I/O bus. In at least one embodiment, a local I/O bus may include, without limitation, a high-speed I/O bus for connecting peripherals to memory **520**, a chipset, and processor **502**. Examples may include, without limitation, an audio controller **529**, a firmware hub ("flash BIOS") **528**, a wireless transceiver **526**, a data storage **524**, a legacy I/O controller **523** containing user input and keyboard interfaces **525**, a serial expansion port **527**, such as a Universal Serial Bus ("USB") port, and a network controller **534**. In at least one embodiment, data storage **524** may comprise a hard disk drive, a floppy disk drive, a CD-ROM device, a flash memory device, or other mass storage device.

[0199] In at least one embodiment, FIG. **5** illustrates a system, which includes interconnected hardware devices or "chips", whereas in other embodiments, FIG. **5** may illustrate an exemplary SoC. In at least one embodiment, devices illustrated in FIG. **5** may be interconnected with proprietary interconnects, standardized interconnects (e.g., PCIe) or some combination thereof. In at least one embodiment, one or more components of computer system **500** are interconnected using compute express link (CXL) interconnects.

[0200] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, inference and/or training logic **115** may be used in system FIG. **5** for inferencing or predicting operations based, at least in part, on weight parameters calculated using neural network training operations, neural network functions and/or architectures, or neural network use cases described herein.

[0201] FIG. **6** is a block diagram illustrating an electronic device **600** for utilizing a processor **610**, according to at least one embodiment. In at least one embodiment, electronic device **600** may be, for example and without limitation, a notebook, a tower server, a rack server, a blade server, a laptop, a desktop, a tablet, a mobile device, a phone, an embedded computer, or any other suitable electronic device.

[0202] In at least one embodiment, electronic device **600** may include, without limitation, processor **610** communicatively coupled to any suitable number or kind of components, peripherals, modules, or devices. In at least one embodiment, processor **610** is coupled using a bus or interface, such as a I^2C bus, a System Management Bus ("SMBus"), a Low Pin Count (LPC) bus, a Serial Peripheral Interface ("SPI"), a High Definition Audio ("HDA") bus, a Serial Advance Technology Attachment ("SATA") bus, a Universal Serial Bus ("USB") (versions 1, 2, 3, etc.), or a Universal Asynchronous Receiver/Transmitter ("UART") bus. In at least one embodiment, FIG. **6** illustrates a system, which includes interconnected hardware devices or "chips", whereas in other embodiments, FIG. **6** may illustrate an exemplary SoC. In at least one embodiment, devices illustrated in FIG. **6** may be interconnected with proprietary interconnects, standardized interconnects (e.g., PCIe) or some combination thereof. In at least one embodiment, one or more components of FIG. **6** are interconnected using compute express link (CXL) interconnects.

[0203] In at least one embodiment, FIG. **6** may include a display **624**, a touch screen **625**, a touch pad **630**, a Near Field Communications unit ("NEC") **645**, a sensor hub **640**, a thermal sensor **646**, an Express Chipset ("EC") **635**, a Trusted Platform Module ("TPM") **638**, BIOS/firmware/flash memory ("BIOS, FW Flash") **622**, a DSP **660**, a drive **620** such as a Solid State Disk ("SSD") or a Hard Disk Drive ("HDD"), a wireless local area network unit ("WLAN") **650**, a Bluetooth unit **652**, a Wireless Wide Area Network unit ("WWAN") **656**, a Global Positioning System (GPS) unit **655**, a camera ("USB 3.0 camera") **654** such as a USB 3.0 camera, and/or a Low Power Double Data Rate ("LPDDR") memory unit ("LPDDR3") **615** implemented in, for example, an LPDDR3 standard. These components may each be implemented in any suitable manner.

[0204] In at least one embodiment, other components may be communicatively coupled to processor **610** through components described herein. In at least one embodiment, an accelerometer **641**, an ambient light sensor ("ALS") **642**, a compass **643**, and a gyroscope **644** may be communicatively coupled to sensor hub **640**. In at least one embodiment, a thermal sensor **639**, a fan **637**, a keyboard **636**, and touch pad **630** may be communicatively coupled to EC **635**. In at least one embodiment, speakers **663**, headphones **664**, and a microphone ("mic") **665** may be communicatively coupled to an audio unit ("audio codec and class D amp") **662**, which may in turn be communicatively coupled to DSP **660**. In at least one embodiment, audio unit **662** may include, for example and without limitation, an audio coder/decoder ("codec") and a class D amplifier. In at least one embodiment, a SIM card ("SIM") **657** may be communicatively coupled to WWAN unit **656**. In at least one embodiment, components such as WLAN unit **650** and Bluetooth unit **652**, as well as WWAN unit **656** may be implemented in a Next Generation Form Factor ("NGFF").

[0205] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, inference and/or training logic **115** may be used in system FIG. **6** for inferencing or predicting operations based, at least in part, on weight parameters calculated using neural network training operations, neural network functions and/or architectures, or neural network use cases described herein.

[0206] FIG. **7** illustrates a computer system **700**, according to at least one embodiment. In at least one embodiment, computer system **700** is configured to implement various processes and methods described throughout this disclosure.

[0207] In at least one embodiment, computer system **700** comprises, without limitation, at least one central processing unit ("CPU") **702** that is connected to a communication bus **710** implemented using any suitable protocol, such as PCI ("Peripheral Component Interconnect"), peripheral component interconnect express ("PCI-Express"), AGP ("Accelerated Graphics Port"), HyperTransport, or any other bus or point-to-point communication protocol(s). In at least one embodiment, computer system **700** includes, without limitation, a main memory **704** and control logic (e.g., implemented as hardware, software, or a combination thereof) and data are stored in main memory **704**, which may take form of random access memory ("RAM"). In at least one embodiment, a network interface subsystem ("network interface") **722** provides an interface to other computing devices and networks for receiving data from and transmitting data to other systems with computer system **700**.

[0208] In at least one embodiment, computer system **700**, in at least one embodiment, includes, without limitation, input devices **708**, a parallel processing system **712**, and display devices **706** that can be implemented using a conventional cathode ray tube ("CRT"), a liquid crystal display ("LCD"), a light emitting diode ("LED") display, a plasma display, or other suitable display technologies. In at least one embodiment, user input is received from input devices **708**

such as keyboard, mouse, touchpad, microphone, etc. In at least one embodiment, each module described herein can be situated on a single semiconductor platform to form a processing system.

[0209] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, inference and/or training logic **115** may be used in system FIG. **7** for inferencing or predicting operations based, at least in part, on weight parameters calculated using neural network training operations, neural network functions and/or architectures, or neural network use cases described herein.

[0210] FIG. **8** illustrates a computer system **800**, according to at least one embodiment. In at least one embodiment, computer system **800** includes, without limitation, a computer **810** and a USB stick **820**. In at least one embodiment, computer **810** may include, without limitation, any number and type of processor(s) (not shown) and a memory (not shown). In at least one embodiment, computer **810** includes, without limitation, a server, a cloud instance, a laptop, and a desktop computer.

[0211] In at least one embodiment, USB stick **820** includes, without limitation, a processing unit **830**, a USB interface **840**, and USB interface logic **850**. In at least one embodiment, processing unit **830** may be any instruction execution system, apparatus, or device capable of executing instructions. In at least one embodiment, processing unit **830** may include, without limitation, any number and type of processing cores (not shown). In at least one embodiment, processing unit **830** comprises an application specific integrated circuit ("ASIC") that is optimized to perform any amount and type of operations associated with machine learning. For instance, in at least one embodiment, processing unit **830** is a tensor processing unit ("TPC") that is optimized to perform machine learning inference operations. In at least one embodiment, processing unit **830** is a vision processing unit ("VPU") that is optimized to perform machine vision and machine learning inference operations.

[0212] In at least one embodiment, USB interface **840** may be any type of USB connector or USB socket. For instance, in at least one embodiment, USB interface **840** is a USB 3.0 Type-C socket for data and power. In at least one embodiment, USB interface **840** is a USB 3.0 Type-A connector. In at least one embodiment, USB interface logic **850** may include any amount and type of logic that enables processing unit **830** to interface with devices (e.g., computer **810**) via USB connector **840**.

[0213] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, inference and/or training logic **115** may be used in system FIG. **8** for inferencing or predicting operations based, at least in part, on weight parameters calculated using neural network training operations, neural network functions and/or architectures, or neural network use cases described herein.

[0214] FIG. **9**A illustrates an exemplary architecture in which a plurality of GPUs **910**(**1**)-**910**(N) is communicatively coupled to a plurality of multi-core processors **905**(**1**)-**905**(M) over high-speed links **940**(**1**)-**940**(N) (e.g., buses, point-to-point interconnects, etc.). In at least one embodiment, high-speed links **940**(**1**)-**940**(N) support a communication throughput of 4 GB/s, 30 GB/s, 80 GB/s or higher. In at least one embodiment, various interconnect protocols may be used including, but not limited to, PCIe 4.0 or 5.0 and NVLink 2.0. In various figures, "N" and "M" represent positive integers, values of which may be different from figure to figure.

[0215] In addition, and in at least one embodiment, two or more of GPUs **910** are interconnected over high-speed links **929**(**1**)-**929**(**2**), which may be implemented using similar or different protocols/links than those used for high-speed links **940**(**1**)-**940**(N). Similarly, two or more of multi-core processors **905** may be connected over a high-speed link **928** which may be symmetric multi-processor (SMP) buses operating at 20 GB/s, 30 GB/s, 120 GB/s or higher. Alternatively, all communication between various system components shown in FIG. **9**A may be accomplished using similar protocols/links (e.g., over a common interconnection fabric).

[0216] In at least one embodiment, each multi-core processor **905** is communicatively coupled to a processor memory **901**(**1**)-**901**(M), via memory interconnects **926**(**1**)-**926**(M), respectively, and each GPU **910**(**1**)-**910**(N) is communicatively coupled to GPU memory **920**(**1**)-**920**(N) over GPU memory interconnects **950**(**1**)-**950**(N), respectively. In at least one embodiment, memory interconnects **926** and **950** may utilize similar or different memory access technologies. By way of example, and not limitation, processor memories **901**(**1**)-**901**(M) and GPU memories **920** may be volatile memories such as dynamic random access memories (DRAMs) (including stacked DRAMs), Graphics DDR SDRAM (GDDR) (e.g., GDDRS, GDDR6), or High Bandwidth Memory (HBM) and/or may be non-volatile memories such as 3D XPoint or Nano-Ram. In at least one embodiment, some portion of processor memories **901** may be volatile memory and another portion may be non-volatile memory (e.g., using a two-level memory (2LM) hierarchy).

[0217] As described herein, although various multi-core processors **905** and GPUs **910** may be physically coupled to a particular memory **901**, **920**, respectively, and/or a unified memory architecture may be implemented in which a virtual system address space (also referred to as "effective address" space) is distributed among various physical memories. For example, processor memories **901**(**1**)-**901**(M) may each comprise 64 GB of system memory address space and GPU memories **920**(**1**)-**920**(N) may each comprise 32 GB of system memory address space resulting in a total of 256 GB addressable memory when M=2 and N=4. Other values for N and M are possible.

[0218] FIG. **9**B illustrates additional details for an interconnection between a multi-core processor **907** and a graphics acceleration module **946** in accordance with one exemplary embodiment. In at least one embodiment, graphics acceleration module **946** may include one or more GPU chips integrated on a line card which is coupled to processor **907** via high-speed link **940** (e.g., a PCIe bus, NVLink, etc.). In at least one embodiment, graphics acceleration module **946** may alternatively be integrated on a package or chip with processor **907**.

[0219] In at least one embodiment, processor **907** includes a plurality of cores **960**A-**960**D, each with a translation lookaside buffer ("TLB") **961**A-**961**D and one or more caches **962**A-**962**D. In at least one embodiment, cores **960**A-**960**D may include various other components for executing instructions and processing data that are not illustrated. In at least one embodiment, caches **962**A-**962**D may comprise Level 1 (L1) and Level 2 (L2) caches. In addition, one or more shared caches **956** may be included in caches **962**A-**962**D and shared by sets of cores **960**A-**960**D. For example, one embodiment of processor **907** includes 24 cores, each with its own L1 cache, twelve shared L2 caches, and twelve shared L3 caches. In this embodiment, one or more L2 and L3 caches are shared by two adjacent cores. In at least one embodiment, processor **907** and graphics acceleration module **946** connect with system memory **914**, which may include processor memories **901**(**1**)-**901**(M) of FIG. **9**A.

[0220] In at least one embodiment, coherency is maintained for data and instructions stored in various caches **962**A-**962**D, **956** and system memory **914** via inter-core communication over a coherence bus **964**. In at least one embodiment, for example, each cache may have cache coherency logic/circuitry associated therewith to communicate to over coherence bus **964** in response to detected reads or writes to particular cache lines. In at least one embodiment, a cache snooping protocol is implemented over coherence bus **964** to snoop cache accesses.

[0221] In at least one embodiment, a proxy circuit **925** communicatively couples graphics acceleration module **946** to coherence bus **964**, allowing graphics acceleration module **946** to participate in a cache coherence protocol as a peer of cores **960**A-**960**D. In particular, in at least one embodiment, an interface **935** provides connectivity to proxy circuit **925** over high-speed link **940** and an interface **937** connects graphics acceleration module **946** to high-speed link **940**.

[0222] In at least one embodiment, an accelerator integration circuit **936** provides cache management, memory access, context management, and interrupt management services on behalf of a plurality of graphics processing engines **931**(**1**)-**931**(N) of graphics acceleration module **946**. In at least one embodiment, graphics processing engines **931**(**1**)-**931**(N) may each comprise a separate graphics processing unit (GPU). In at least one embodiment, graphics processing engines **931**(**1**)-**931**(N) alternatively may comprise different types of graphics processing engines within a GPU, such as graphics execution units, media processing engines (e.g., video encoders/decoders), samplers, and blit engines. In at least one embodiment, graphics acceleration module **946** may be a GPU with a plurality of graphics processing engines **931**(**1**)-**931**(N) or graphics processing engines **931**(**1**)-**931**(N) may be individual GPUs integrated on a common package, line card, or chip.

[0223] In at least one embodiment, accelerator integration circuit **936** includes a memory management unit (MMU) **939** for performing various memory management functions such as virtual-to-physical memory translations (also referred to as effective-to-real memory translations) and memory access protocols for accessing system memory **914**. In at least one embodiment, MMU **939** may also include a translation lookaside buffer (TLB) (not shown) for caching virtual/effective to physical/real address translations. In at least one embodiment, a cache **938** can store commands and data for efficient access by graphics processing engines **931**(**1**)-**931**(N). In at least one embodiment, data stored in cache **938** and graphics memories **933**(**1**)-**933**(M) is kept coherent with core caches **962**A-**962**D, **956** and system memory **914**, possibly using a fetch unit **944**. As mentioned, this may be accomplished via proxy circuit **925** on behalf of cache **938** and memories **933**(**1**)-**933**(M) (e.g., sending updates to cache **938** related to modifications/accesses of cache lines on processor caches **962**A-**962**D, **956** and receiving updates from cache **938**).

[0224] In at least one embodiment, a set of registers **945** store context data for threads executed by graphics processing engines **931**(**1**)-**931**(N) and a context management circuit **948** manages thread contexts. For example, context management circuit **948** may perform save and restore operations to save and restore contexts of various threads during contexts switches (e.g., where a first thread is saved and a second thread is stored so that a second thread can be execute by a graphics processing engine). For example, on a context switch, context management circuit **948** may store current register values to a designated region in memory (e.g., identified by a context pointer). It may then restore register values when returning to a context. In at least one embodiment, an interrupt management circuit **947** receives and processes interrupts received from system devices.

[0225] In at least one embodiment, virtual/effective addresses from a graphics processing engine **931** are translated to real/physical addresses in system memory **914** by MMU **939**. In at least one embodiment, accelerator integration circuit **936** supports multiple (e.g., 4, 8, 16) graphics accelerator modules **946** and/or other accelerator devices. In at least one embodiment, graphics accelerator module **946** may be dedicated to a single application executed on processor **907** or may be shared between multiple applications. In at least one embodiment, a virtualized graphics execution environment is presented in which resources of graphics processing engines **931**(**1**)-**931**(N) are shared with multiple applications or virtual machines (VMs). In at least one embodiment, resources may be subdivided into "slices" which are allocated to different VMs and/or applications based on processing requirements and priorities associated with VMs and/or applications.

[0226] In at least one embodiment, accelerator integration circuit **936** performs as a bridge to a system for graphics acceleration module **946** and provides address translation and system memory cache services. In addition, in at least one embodiment, accelerator integration circuit **936** may provide virtualization facilities for a host processor to manage virtualization of graphics processing engines **931**(**1**)-**931**(N), interrupts, and memory management.

[0227] In at least one embodiment, because hardware resources of graphics processing engines **931**(**1**)-**931**(N) are mapped explicitly to a real address space seen by host processor **907**, any host processor can address these resources directly using an effective address value. In at least one embodiment, one function of accelerator integration circuit **936** is physical separation of graphics processing engines **931**(**1**)-**931**(N) so that they appear to a system as independent units.

[0228] In at least one embodiment, one or more graphics memories **933**(**1**)-**933**(M) are coupled to each of graphics processing engines **931**(**1**)-**931**(N), respectively and N=M. In at least one embodiment, graphics memories **933**(**1**)-**933**(M) store instructions and data being processed by each of graphics processing engines **931**(**1**)-**931**(N). In at least one embodiment, graphics memories **933**(**1**)-**933**(M) may be volatile memories such as DRAMs (including stacked

DRAMs), GDDR memory (e.g., GDDR5, GDDR6), or HBM, and/or may be non-volatile memories such as 3D XPoint or Nano-Ram.

[0229] In at least one embodiment, to reduce data traffic over high-speed link **940**, biasing techniques can be used to ensure that data stored in graphics memories **933**(**1**)-**933**(M) is data that will be used most frequently by graphics processing engines **931**(**1**)-**931**(N) and preferably not used by cores **960**A-**960**D (at least not frequently). Similarly, in at least one embodiment, a biasing mechanism attempts to keep data needed by cores (and preferably not graphics processing engines **931**(**1**)-**931**(N)) within caches **962**A-**962**D, **956** and system memory **914**.

[0230] FIG. **9**C illustrates another exemplary embodiment in which accelerator integration circuit **936** is integrated within processor **907**. In this embodiment, graphics processing engines **931**(**1**)-**931**(N) communicate directly over high-speed link **940** to accelerator integration circuit **936** via interface **937** and interface **935** (which, again, may be any form of bus or interface protocol). In at least one embodiment, accelerator integration circuit **936** may perform similar operations as those described with respect to FIG. **9**B, but potentially at a higher throughput given its close proximity to coherence bus **964** and caches **962**A-**962**D, **956**. In at least one embodiment, an accelerator integration circuit supports different programming models including a dedicated-process programming model (no graphics acceleration module virtualization) and shared programming models (with virtualization), which may include programming models which are controlled by accelerator integration circuit **936** and programming models which are controlled by graphics acceleration module **946**.

[0231] In at least one embodiment, graphics processing engines **931**(**1**)-**931**(N) are dedicated to a single application or process under a single operating system. In at least one embodiment, a single application can funnel other application requests to graphics processing engines **931**(**1**)-**931**(N), providing virtualization within a VM/partition.

[0232] In at least one embodiment, graphics processing engines **931**(**1**)-**931**(N), may be shared by multiple VM/application partitions. In at least one embodiment, shared models may use a system hypervisor to virtualize graphics processing engines **931**(**1**)-**931**(N) to allow access by each operating system. In at least one embodiment, for single-partition systems without a hypervisor, graphics processing engines **931**(**1**)-**931**(N) are owned by an operating system. In at least one embodiment, an operating system can virtualize graphics processing engines **931**(**1**)-**931**(N) to provide access to each process or application.

[0233] In at least one embodiment, graphics acceleration module **946** or an individual graphics processing engine **931**(**1**)-**931**(N) selects a process element using a process handle. In at least one embodiment, process elements are stored in system memory **914** and are addressable using an effective address to real address translation technique described herein. In at least one embodiment, a process handle may be an implementation-specific value provided to a host process when registering its context with graphics processing engine **931**(**1**)-**931**(N) (that is, calling system software to add a process element to a process element linked list). In at least one embodiment, a lower 16-bits of a process handle may be an offset of a process element within a process element linked list.

[0234] FIG. **9**D illustrates an exemplary accelerator integration slice **990**. In at least one embodiment, a "slice" comprises a specified portion of processing resources of accelerator integration circuit **936**. In at least one embodiment, an application is effective address space **982** within system memory **914** stores process elements **983**. In at least one embodiment, process elements **983** are stored in response to GPU invocations **981** from applications **980** executed on processor **907**. In at least one embodiment, a process element **983** contains process state for corresponding application **980**. In at least one embodiment, a work descriptor (WD) **984** contained in process element **983** can be a single job requested by an application or may contain a pointer to a queue of jobs. In at least one embodiment, WD **984** is a pointer to a job request queue in an application's effective address space **982**.

[0235] In at least one embodiment, graphics acceleration module **946** and/or individual graphics processing engines **931**(**1**)-**931**(N) can be shared by all or a subset of processes in a system. In at least one embodiment, an infrastructure for setting up process states and sending a WD **984** to a graphics acceleration module **946** to start a job in a virtualized environment may be included.

[0236] In at least one embodiment, a dedicated-process programming model is implementation-specific. In at least one embodiment, in this model, a single process owns graphics acceleration module **946** or an individual graphics processing engine **931**. In at least one embodiment, when graphics acceleration module **946** is owned by a single process, a hypervisor initializes accelerator integration circuit **936** for an owning partition and an operating system initializes accelerator integration circuit **936** for an owning process when graphics acceleration module **946** is assigned.

[0237] In at least one embodiment, in operation, a WD fetch unit **991** in accelerator integration slice **990** fetches next WD **984**, which includes an indication of work to be done by one or more graphics processing engines of graphics acceleration module **946**. In at least one embodiment, data from WD **984** may be stored in registers **945** and used by MMU **939**, interrupt management circuit **947** and/or context management circuit **948** as illustrated. For example, one embodiment of MMU **939** includes segment/page walk circuitry for accessing segment/page tables **986** within an OS virtual address space **985**. In at least one embodiment, interrupt management circuit **947** may process interrupt events **992** received from graphics acceleration module **946**. In at least one embodiment, when performing graphics operations, an effective address **993** generated by a graphics processing engine **931**(**1**)-**931**(N) is translated to a real address by MMU **939**.

[0238] In at least one embodiment, registers **945** are duplicated for each graphics processing engine **931**(**1**)-**931**(N) and/or graphics acceleration module **946** and may be initialized by a hypervisor or an operating system. In at least one embodiment, each of these duplicated registers may be included in an accelerator integration slice **990**. Exemplary registers that may be initialized by a hypervisor are shown in Table 1.

TABLE 1

Hypervisor Initialized Registers

Register #	Description
1	Slice Control Register
2	Real Address (RA) Scheduled Processes Area Pointer
3	Authority Mask Override Register
4	Interrupt Vector Table Entry Offset
5	Interrupt Vector Table Entry Limit
6	State Register
7	Logical Partition ID
8	Real address (RA) Hypervisor Accelerator Utilization Record Pointer
9	Storage Description Register

[0239] Exemplary registers that may be initialized by an operating system are shown in Table 2.

TABLE 2

Operating System Initialized Registers

Register #	Description
1	Process and Thread Identification
2	Effective Address (EA) Context Save/Restore Pointer
3	Virtual Address (VA) Accelerator Utilization Record Pointer
4	Virtual Address (VA) Storage Segment Table Pointer
5	Authority Mask
6	Work descriptor

[0240] In at least one embodiment, each WD **984** is specific to a particular graphics acceleration module **946** and/or graphics processing engines **931**(**1**)-**931**(N). In at least one embodiment, it contains all information required by a graphics processing engine **931**(**1**)-**931**(N) to do work, or it can be a pointer to a memory location where an application has set up a command queue of work to be completed.

[0241] FIG. **9**E illustrates additional details for one exemplary embodiment of a shared model. This embodiment includes a hypervisor real address space **998** in which a process element list **999** is stored. In at least one embodiment, hypervisor real address space **998** is accessible via a hypervisor **996** which virtualizes graphics acceleration module engines for operating system **995**.

[0242] In at least one embodiment, shared programming models allow for all or a subset of processes from all or a subset of partitions in a system to use a graphics acceleration module **946**. In at least one embodiment, there are two programming models where graphics acceleration module **946** is shared by multiple processes and partitions, namely time-sliced shared and graphics directed shared.

[0243] In at least one embodiment, in this model, system hypervisor **996** owns graphics acceleration module **946** and makes its function available to all operating systems **995**. In at least one embodiment, for a graphics acceleration module **946** to support virtualization by system hypervisor **996**, graphics acceleration module **946** may adhere to certain requirements, such as (1) an application's job request must be autonomous (that is, state does not need to be maintained between jobs), or graphics acceleration module **946** must provide a context save and restore mechanism, (2) an application's job request is guaranteed by graphics acceleration module **946** to complete in a specified amount of time, including any translation faults, or graphics acceleration module **946** provides an ability to preempt processing of a job, and (3) graphics acceleration module **946** must be guaranteed fairness between processes when operating in a directed shared programming model.

[0244] In at least one embodiment, application **980** is required to make an operating system **995** system call with a graphics acceleration module type, a work descriptor (WD), an authority mask register (AMR) value, and a context save/restore area pointer (CSRP). In at least one embodiment, graphics acceleration module type describes a targeted acceleration function for a system call. In at least one embodiment, graphics acceleration module type may be a system-specific value. In at least one embodiment, WD is formatted specifically for graphics acceleration module **946** and can be in a form of a graphics acceleration module **946** command, an effective address pointer to a user-defined structure, an effective address pointer to a queue of commands, or any other data structure to describe work to be done by graphics acceleration module **946**.

[0245] In at least one embodiment, an AMR value is an AMR state to use for a current process. In at least one embodiment, a value passed to an operating system is similar to an application setting an AMR. In at least one embodiment, if accelerator integration circuit **936** (not shown) and graphics acceleration module **946** implementations do not support a User Authority Mask Override Register (UAMOR), an operating system may apply a current UAMOR value to an AMR value before passing an AMR in a hypervisor call. In at least one embodiment, hypervisor **996** may optionally apply a current Authority Mask Override Register (AMOR) value before placing an AMR into process element **983**. In at least one embodiment, CSRP is one of registers **945** containing an effective address of an area in an application's effective address space **982** for graphics acceleration module **946** to save and restore context state. In at least one embodiment, this pointer is optional if no state is required to be saved between jobs or when a job is preempted. In at least one embodiment, context save/restore area may be pinned system memory.

[0246] Upon receiving a system call, operating system **995** may verify that application **980** has registered and been given authority to use graphics acceleration module **946**. In at least one embodiment, operating system **995** then calls hypervisor **996** with information shown in Table 3.

TABLE 3

OS to Hypervisor Call Parameters

Parameter #	Description
1	A work descriptor (WD)
2	An Authority Mask Register (AMR) value (potentially masked)
3	An effective address (EA) Context Save/Restore Area Pointer (CSRP)
4	A process ID (PID) and optional thread ID (TID)
5	A virtual address (VA) accelerator utilization record pointer (AURP)
6	Virtual address of storage segment table pointer (SSTP)
7	A logical interrupt service number (LISN)

[0247] In at least one embodiment, upon receiving a hypervisor call, hypervisor **996** verifies that operating system **995** has registered and been given authority to use graphics acceleration module **946**. In at least one embodiment, hypervisor **996** then puts process element **983** into a process element linked list for a corresponding graphics acceleration module **946** type. In at least one embodiment, a process element may include information shown in Table 4.

TABLE 4

Process Element Information

Element #	Description
1	A work descriptor (WD)
2	An Authority Mask Register (AMR) value (potentially masked).
3	An effective address (EA) Context Save/Restore Area Pointer (CSRP)
4	A process ID (PID) and optional thread ID (TID)
5	A virtual address (VA) accelerator utilization record pointer (AURP)
6	Virtual address of storage segment table pointer (SSTP)
7	A logical interrupt service number (LISN)
8	Interrupt vector table, derived from hypervisor call parameters
9	A state register (SR) value
10	A logical partition ID (LPID)
11	A real address (RA) hypervisor accelerator utilization record pointer
12	Storage Descriptor Register (SDR)

[0248] In at least one embodiment, hypervisor initializes a plurality of accelerator integration slice **990** registers **945**.

[0249] As illustrated in FIG. **9**F, in at least one embodiment, a unified memory is used, addressable via a common virtual memory address space used to access physical processor memories **901**(**1**)-**901**(N) and GPU memories **920**(**1**)-**920**(N). In this implementation, operations executed on GPUs **910**(**1**)-**910**(N) utilize a same virtual/effective memory address space to access processor memories **901**(**1**)-**901**(M) and vice versa, thereby simplifying programmability. In at least one embodiment, a first portion of a virtual/effective address space is allocated to processor memory **901**(**1**), a second portion to second processor memory **901**(N), a third portion to GPU memory **920**(**1**), and so on. In at least one embodiment, an entire virtual/effective memory space (sometimes referred to as an effective address space) is thereby distributed across each of processor memories **901** and GPU memories **920**, allowing any processor or GPU to access any physical memory with a virtual address mapped to that memory.

[0250] In at least one embodiment, bias/coherence management circuitry **994**A-**994**E within one or more of MMUs **939**A-**939**E ensures cache coherence between caches of one or more host processors (e.g., **905**) and GPUs **910** and implements biasing techniques indicating physical memories in which certain types of data should be stored. In at least one embodiment, while multiple instances of bias/coherence management circuitry **994**A-**994**E are illustrated in FIG. **9**F, bias/coherence circuitry may be implemented within an MMU of one or more host processors **905** and/or within accelerator integration circuit **936**.

[0251] One embodiment allows GPU memories **920** to be mapped as part of system memory, and accessed using shared virtual memory (SVM) technology, but without suffering performance drawbacks associated with full system cache coherence. In at least one embodiment, an ability for GPU memories **920** to be accessed as system memory without onerous cache coherence overhead provides a beneficial operating environment for GPU offload. In at least one embodiment, this arrangement allows software of host processor **905** to setup operands and access computation results, without overhead of tradition I/O DMA data copies. In at least one embodiment, such traditional copies involve driver calls, interrupts and memory mapped I/O (MMIO) accesses that are all inefficient relative to simple memory accesses. In at least one embodiment, an ability to access GPU memories **920** without cache coherence overheads can be critical to execution time of an offloaded computation. In at least one embodiment, in cases with substantial streaming write memory traffic, for example, cache coherence overhead can significantly reduce an effective write bandwidth seen by a GPU **910**. In at least one embodiment, efficiency of operand setup, efficiency of results access, and efficiency of GPU computation may play a role in determining effectiveness of a GPU offload.

[0252] In at least one embodiment, selection of GPU bias and host processor bias is driven by a bias tracker data structure. In at least one embodiment, a bias table may be used, for example, which may be a page-granular structure (e.g., controlled at a granularity of a memory page) that includes 1 or 2 bits per GPU-attached memory page. In at least one embodiment, a bias table may be implemented in a stolen memory range of one or more GPU memories **920**, with or without a bias cache in a GPU **910** (e.g., to cache frequently/recently used entries of a bias table). Alternatively, in at least one embodiment, an entire bias table may be maintained within a GPU.

[0253] In at least one embodiment, a bias table entry associated with each access to a GPU attached memory **920** is accessed prior to actual access to a GPU memory, causing following operations. In at least one embodiment, local requests from a GPU **910** that find their page in GPU bias are forwarded directly to a corresponding GPU memory **920**. In at least one embodiment, local requests from a GPU that find their page in host bias are forwarded to processor **905** (e.g., over a high-speed link as described herein). In at least one embodiment, requests from processor **905** that find a requested page in host processor bias complete a request like a normal memory read. Alternatively, requests directed to a GPU-biased page may be forwarded to a GPU **910**. In at least one embodiment, a GPU may then transition a page to a host processor bias if it is not currently using a page. In at least one embodiment, a bias state of a page can be changed either by a software-based mechanism, a hardware-assisted software-based mechanism, or, for a limited set of cases, a purely hardware-based mechanism.

[0254] In at least one embodiment, one mechanism for changing bias state employs an API call (e.g., OpenCL), which, in turn, calls a GPU's device driver which, in turn, sends a message (or enqueues a command descriptor) to a GPU directing it to change a bias state and, for some transitions, perform a cache flushing operation in a host. In at least one embodiment, a cache flushing operation is used for a transition from host processor **905** bias to GPU bias, but is not for an opposite transition.

[0255] In at least one embodiment, cache coherency is maintained by temporarily rendering GPU-biased pages uncacheable by host processor **905**. In at least one embodiment, to access these pages, processor **905** may request access from GPU **910**, which may or may not grant access right away. In at least one embodiment, thus, to reduce communication between processor **905** and GPU **910** it is beneficial to ensure that GPU-biased pages are those which are required by a GPU but not host processor **905** and vice versa.

[0256] Hardware structure(s) **115** are used to perform one or more embodiments. Details regarding a hardware structure(s) **115** may be provided herein in conjunction with FIGS. **1**A and/or **1**B.

[0257] FIG. **10** illustrates exemplary integrated circuits and associated graphics processors that may be fabricated using one or more IP cores, according to various embodiments described herein. In addition to what is illustrated, other logic and circuits may be included in at least one embodiment, including additional graphics processors/cores, peripheral interface controllers, or general-purpose processor cores.

[0258] FIG. **10** is a block diagram illustrating an exemplary system on a chip integrated circuit **1000** that may be fabricated using one or more IP cores, according to at least one embodiment. In at least one embodiment, integrated circuit **1000** includes one or more application processor(s) **1005** (e.g., CPUs), at least one graphics processor **1010**, and may additionally include an image processor **1015** and/or a video processor **1020**, any of which may be a modular IP core. In at least one embodiment, integrated circuit **1000** includes peripheral or bus logic including a USB controller **1025**, a UART controller **1030**, an SPI/SDIO controller **1035**, and an I^2S/I^2C controller **1040**. In at least one embodiment, integrated circuit **1000** can include a display device **1045** coupled to one or more of a high-definition multimedia interface (HDMI) controller **1050** and a mobile industry processor interface (MIPI) display interface **1055**. In at least one embodiment, storage may be provided by a flash memory subsystem **1060** including flash memory and a flash memory controller. In at least one embodiment, a memory interface may be provided via a memory controller **1065** for access to SDRAM or SRAM memory devices. In at least one embodiment, some integrated circuits additionally include an embedded security engine **1070**.

[0259] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, inference and/or training logic **115** may be used in integrated circuit **1000** for inferencing or predicting operations based, at least in part, on weight parameters calculated using neural network training operations, neural network functions and/or architectures, or neural network use cases described herein.

[0260] FIGS. **11**A-**11**B illustrate exemplary integrated circuits and associated graphics processors that may be fabricated using one or more IP cores, according to various embodiments described herein. In addition to what is illustrated, other logic and circuits may be included in at least one embodiment, including additional graphics processors/cores, peripheral interface controllers, or general-purpose processor cores.

[0261] FIGS. **11**A-**11**B are block diagrams illustrating exemplary graphics processors for use within an SoC, according to embodiments described herein. FIG. **11**A illustrates an exemplary graphics processor **1110** of a system on a chip integrated circuit that may be fabricated using one or more IP cores, according to at least one embodiment. FIG. **11**B illustrates an additional exemplary graphics processor **1140** of a system on a chip integrated circuit that may be fabricated using one or more IP cores, according to at least one embodiment. In at least one embodiment, graphics processor **1110** of FIG. **11**A is a low power graphics processor core. In at least one embodiment, graphics processor **1140** of FIG. **11**B is a higher performance graphics processor core. In at least one embodiment, each of graphics processors **1110**, **1140** can be variants of graphics processor **1010** of FIG. **10**.

[0262] In at least one embodiment, graphics processor **1110** includes a vertex processor **1105** and one or more fragment processor(s) **1115**A-**1115**N (e.g., **1115**A, **1115**B, **1115**C, **1115**D, through **1115**N−1, and **1115**N). In at least one embodiment, graphics processor **1110** can execute different shader programs via separate logic, such that vertex processor **1105** is optimized to execute operations for vertex shader programs, while one or more fragment processor(s) **1115**A-**1115**N execute fragment (e.g., pixel) shading operations for fragment or pixel shader programs. In at least one embodiment, vertex processor **1105** performs a vertex processing stage of a 3D graphics pipeline and generates primitives and vertex data. In at least one embodiment, fragment processor (s) **1115**A-**1115**N use primitive and vertex data generated by vertex processor **1105** to produce a framebuffer that is displayed on a display device. In at least one embodiment, fragment processor(s) **1115**A-**1115**N are optimized to execute fragment shader programs as provided for in an OpenGL API, which may be used to perform similar operations as a pixel shader program as provided for in a Direct 3D API.

[0263] In at least one embodiment, graphics processor **1110** additionally includes one or more memory management units (MMUs) **1120**A-**1120**B, cache(s) **1125**A-**1125**B, and circuit interconnect(s) **1130**A-**1130**B. In at least one embodiment, one or more MMU(s) **1120**A-**1120**B provide for virtual to physical address mapping for graphics processor **1110**, including for vertex processor **1105** and/or fragment processor(s) **1115**A-**1115**N, which may reference vertex or image/texture data stored in memory, in addition to vertex or image/texture data stored in one or more cache(s) **1125**A-**1125**B. In at least one embodiment, one or more MMU(s) **1120**A-**1120**B may be synchronized with other MMUs within a system, including one or more MMUs associated with one or more application processor(s) **1005**, image processors **1015**, and/or video processors **1020** of FIG. **10**, such that each processor **1005**-**1020** can participate in a shared or unified virtual memory system. In at least one embodiment, one or more circuit interconnect(s) **1130**A-**1130**B enable graphics processor **1110** to interface with other IP cores within SoC, either via an internal bus of SoC or via a direct connection.

[0264] In at least one embodiment, graphics processor **1140** includes one or more shader core(s) **1155**A-**1155**N (e.g., **1155**A, **1155**B, **1155**C, **1155**D, **1155**E, **1155**F, through **1155**N−1, and **1155**N) as shown in FIG. **11**B, which provides for a unified shader core architecture in which a single core or type or core can execute all types of programmable shader code, including shader program code to implement vertex shaders, fragment shaders, and/or compute shaders. In at least one embodiment, a number of shader cores can vary. In at least one embodiment, graphics processor **1140** includes an inter-core task manager **1145**, which acts as a thread dispatcher to dispatch execution threads to one or more shader cores **1155**A-**1155**N and a tiling unit **1158** to accelerate tiling operations for tile-based rendering, in which rendering operations for a scene are subdivided in image space, for example to exploit local spatial coherence within a scene or to optimize use of internal caches.

[0265] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, inference and/or training logic **115** may be used in integrated circuit **11**A and/or **11**B for inferencing or predicting operations based, at least in part, on weight parameters calculated using neural network training operations, neural network functions and/or architectures, or neural network use cases described herein.

[0266] FIGS. **12**A-**12**B illustrate additional exemplary graphics processor logic according to embodiments described herein. FIG. **12**A illustrates a graphics core **1200** that may be included within graphics processor **1010** of FIG. **10**, in at least one embodiment, and may be a unified shader core **1155**A-**1155**N as in FIG. **11**B in at least one embodiment. FIG. **12**B illustrates a highly-parallel general-purpose graphics processing unit ("GPGPU") **1230** suitable for deployment on a multi-chip module in at least one embodiment.

[0267] In at least one embodiment, graphics core **1200** includes a shared instruction cache **1202**, a texture unit **1218**, and a cache/shared memory **1220** that are common to execution resources within graphics core **1200**. In at least one embodiment, graphics core **1200** can include multiple slices **1201**A-**1201**N or a partition for each core, and a graphics processor can include multiple instances of graphics core **1200**. In at least one embodiment, slices **1201**A-**1201**N can include support logic including a local instruction cache **1204**A-**1204**N, a thread scheduler **1206**A-**1206**N, a thread dispatcher **1208**A-**1208**N, and a set of registers **1210**A-**1210**N. In at least one embodiment, slices **1201**A-**1201**N can include a set of additional function units (AFUs **1212**A-**1212**N), floating-point units (FPUs **1214**A-**1214**N), integer arithmetic logic units (ALUs **1216**A-**1216**N), address computational units (ACUs **1213**A-**1213**N), double-precision floating-point units (DPFPUs **1215**A-**1215**N), and matrix processing units (MPUs **1217**A-**1217**N).

[0268] In at least one embodiment, FPUs **1214**A-**1214**N can perform single-precision (32-bit) and half-precision (16-bit) floating point operations, while DPFPUs **1215**A-**1215**N perform double precision (64-bit) floating point operations. In at least one embodiment, ALUs **1216**A-**1216**N can perform variable precision integer operations at 8-bit, 16-bit, and 32-bit precision, and can be configured for mixed precision operations. In at least one embodiment, MPUs **1217**A-**1217**N can also be configured for mixed precision matrix operations, including half-precision floating point and 8-bit integer operations. In at least one embodiment, MPUs **1217**-**1217**N can perform a variety of matrix operations to accelerate machine learning application frameworks, including enabling support for accelerated general matrix to matrix multiplication (GEMM). In at least one embodiment, AFUs **1212**A-**1212**N can perform additional logic operations not supported by floating-point or integer units, including trigonometric operations (e.g., sine, cosine, etc.).

[0269] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, inference and/or training logic **115** may be used in graphics core **1200** for inferencing or predicting operations based, at least in part, on weight parameters calculated using neural network training operations, neural network functions and/or architectures, or neural network use cases described herein.

[0270] FIG. **12**B illustrates a general-purpose processing unit (GPGPU) **1230** that can be configured to enable highly-parallel compute operations to be performed by an array of graphics processing units, in at least one embodiment. In at least one embodiment, GPGPU **1230** can be linked directly to other instances of GPGPU **1230** to create a multi-GPU cluster to improve training speed for deep neural networks. In at least one embodiment, GPGPU **1230** includes a host interface **1232** to enable a connection with a host processor. In at least one embodiment, host interface **1232** is a PCI Express interface. In at least one embodiment, host interface **1232** can be a vendor-specific communications interface or communications fabric. In at least one embodiment, GPGPU **1230** receives commands from a host processor and uses a global scheduler **1234** to distribute execution threads associated with those commands to a set of compute clusters **1236**A-**1236**H. In at least one embodiment, compute clusters **1236**A-**1236**H share a cache memory **1238**. In at least one embodiment, cache memory **1238** can serve as a higher-level cache for cache memories within compute clusters **1236**A-**1236**H.

[0271] In at least one embodiment, GPGPU **1230** includes memory **1244**A-**1244**B coupled with compute clusters **1236**A-**1236**H via a set of memory controllers **1242**A-**1242**B. In at least one embodiment, memory **1244**A-**1244**B can include various types of memory devices including dynamic random access memory (DRAM) or graphics random access memory, such as synchronous graphics random access memory (SGRAM), including graphics double data rate (GDDR) memory.

[0272] In at least one embodiment, compute clusters **1236**A-**1236**H each include a set of graphics cores, such as graphics core **1200** of FIG. **12**A, which can include multiple types of integer and floating point logic units that can perform computational operations at a range of precisions including suited for machine learning computations. For example, in at least one embodiment, at least a subset of floating point units in each of compute clusters **1236**A-**1236**H can be configured to perform 16-bit or 32-bit floating point operations, while a different subset of floating point units can be configured to perform 64-bit floating point operations.

[0273] In at least one embodiment, multiple instances of GPGPU **1230** can be configured to operate as a compute cluster. In at least one embodiment, communication used by compute clusters **1236**A-**1236**H for synchronization and data exchange varies across embodiments. In at least one embodiment, multiple instances of GPGPU **1230** communicate over host interface **1232**. In at least one embodiment, GPGPU **1230** includes an I/O hub **1239** that couples GPGPU **1230** with a GPU link **1240** that enables a direct connection to other instances of GPGPU **1230**. In at least one embodiment, GPU link **1240** is coupled to a dedicated GPU-to-GPU bridge that enables communication and synchronization between multiple instances of GPGPU **1230**. In at least one embodiment, GPU link **1240** couples with a high-speed interconnect to transmit and receive data to other GPGPUs or parallel processors. In at least one embodiment, multiple instances of GPGPU **1230** are located in separate data processing systems and communicate via a network device that is accessible via host interface **1232**. In at least one embodiment GPU link **1240** can be configured to enable a connection to a host processor in addition to or as an alternative to host interface **1232**.

[0274] In at least one embodiment, GPGPU **1230** can be configured to train neural networks. In at least one embodiment, GPGPU **1230** can be used within an inferencing platform. In at least one embodiment, in which GPGPU **1230** is used for inferencing, GPGPU **1230** may include fewer compute clusters **1236**A-**1236**H relative to when GPGPU **1230** is used for training a neural network. In at least one embodiment, memory technology associated with memory **1244**A-**1244**B may differ between inferencing and training configurations, with higher bandwidth memory technologies devoted to training configurations. In at least one embodiment, an inferencing configuration of GPGPU **1230** can support inferencing specific instructions. For example, in at least one embodiment, an inferencing configuration can provide support for one or more 8-bit integer dot product instructions, which may be used during inferencing operations for deployed neural networks.

[0275] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, inference and/or training logic **115** may be used in GPGPU **1230** for inferencing or predicting operations based, at least in part, on weight parameters calculated using neural network training operations, neural network functions and/or architectures, or neural network use cases described herein.

[0276] FIG. **13** is a block diagram illustrating a computing system **1300** according to at least one embodiment. In at least one embodiment, computing system **1300** includes a processing subsystem **1301** having one or more processor(s) **1302** and a system memory **1304** communicating via an interconnection path that may include a memory hub **1305**. In at least one embodiment, memory hub **1305** may be a separate component within a chipset component or may be integrated within one or more processor(s) **1302**. In at least one embodiment, memory hub **1305** couples with an I/O subsystem **1311** via a communication link **1306**. In at least one embodiment, I/O subsystem **1311** includes an I/O hub **1307** that can enable computing system **1300** to receive input from one or more input device(s) **1308**. In at least one embodiment, I/O hub **1307** can enable a display controller, which may be included in one or more processor(s) **1302**, to provide outputs to one or more display device(s) **1310**A. In at least one embodiment, one or more display device(s) **1310**A coupled with I/O hub **1307** can include a local, internal, or embedded display device.

[0277] In at least one embodiment, processing subsystem **1301** includes one or more parallel processor(s) **1312** coupled to memory hub **1305** via a bus or other communication link **1313**. In at least one embodiment, communication link **1313** may use one of any number of standards based communication link technologies or protocols, such as, but not limited to PCI Express, or may be a vendor-specific communications interface or communications fabric. In at least one embodiment, one or more parallel processor(s) **1312** form a computationally focused parallel or vector processing system that can include a large number of processing cores and/or processing clusters, such as a many-integrated core (MIC) processor. In at least one embodiment, some or all of parallel processor(s) **1312** form a graphics processing subsystem that can output pixels to one of one or more display device(s) **1310**A coupled via I/O Hub **1307**. In at least one embodiment, parallel processor(s) **1312** can also include a display controller and display interface (not shown) to enable a direct connection to one or more display device(s) **1310**B.

[0278] In at least one embodiment, a system storage unit **1314** can connect to I/O hub **1307** to provide a storage mechanism for computing system **1300**. In at least one embodiment, an I/O switch **1316** can be used to provide an interface mechanism to enable connections between I/O hub **1307** and other components, such as a network adapter **1318** and/or a wireless network adapter **1319** that may be integrated into platform, and various other devices that can be added via one or more add-in device(s) **1320**. In at least one embodiment, network adapter **1318** can be an Ethernet adapter or another wired network adapter. In at least one embodiment, wireless network adapter **1319** can include one or more of a Wi-Fi, Bluetooth, near field communication (NFC), or other network device that includes one or more wireless radios.

[0279] In at least one embodiment, computing system **1300** can include other components not explicitly shown, including USB or other port connections, optical storage drives, video capture devices, and like, may also be connected to I/O hub **1307**. In at least one embodiment, communication paths interconnecting various components in FIG. **13** may be implemented using any suitable protocols, such as PCI (Peripheral Component Interconnect) based protocols (e.g., PCI-Express), or other bus or point-to-point communication interfaces and/or protocol(s), such as NV-Link high-speed interconnect, or interconnect protocols.

[0280] In at least one embodiment, parallel processor(s) **1312** incorporate circuitry optimized for graphics and video processing, including, for example, video output circuitry, and constitutes a graphics processing unit (GPU). In at least one embodiment, parallel processor(s) **1312** incorporate circuitry optimized for general purpose processing. In at least embodiment, components of computing system **1300** may be integrated with one or more other system elements on a single integrated circuit. For example, in at least one embodiment, parallel processor(s) **1312**, memory hub **1305**, processor(s) **1302**, and I/O hub **1307** can be integrated into a system on chip (SoC) integrated circuit. In at least one embodiment, components of computing system **1300** can be integrated into a single package to form a system in package (SIP) configuration. In at least one embodiment, at least a portion of components of computing system **1300** can be integrated into a multi-chip module (MCM), which can be interconnected with other multi-chip modules into a modular computing system.

[0281] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, inference and/or training logic **115** may be used in system FIG. **1300** for inferencing or predicting operations based, at least in part, on weight parameters calculated using neural network training operations, neural network functions and/or architectures, or neural network use cases described herein.

Processors

[0282] FIG. **14**A illustrates a parallel processor **1400** according to at least one embodiment. In at least one embodiment, various components of parallel processor **1400** may be implemented using one or more integrated circuit devices, such as programmable processors, application specific integrated circuits (ASICs), or field programmable gate arrays (FPGA). In at least one embodiment, illustrated parallel processor **1400** is a variant of one or more parallel processor(s) **1312** shown in FIG. **13** according to an exemplary embodiment.

[0283] In at least one embodiment, parallel processor **1400** includes a parallel processing unit **1402**. In at least one embodiment, parallel processing unit **1402** includes an I/O unit **1404** that enables communication with other devices, including other instances of parallel processing unit **1402**. In at least one embodiment, I/O unit **1404** may be directly connected to other devices. In at least one embodiment, I/O unit **1404** connects with other devices via use of a hub or switch interface, such as a memory hub **1405**. In at least one embodiment, connections between memory hub **1405** and I/O unit **1404** form a communication link **1413**. In at least one embodiment, I/O unit **1404** connects with a host interface **1406** and a memory crossbar **1416**, where host interface **1406** receives commands directed to performing processing operations and memory crossbar **1416** receives commands directed to performing memory operations.

[0284] In at least one embodiment, when host interface **1406** receives a command buffer via I/O unit **1404**, host interface **1406** can direct work operations to perform those commands to a front end **1408**. In at least one embodiment, front end **1408** couples with a scheduler **1410**, which is configured to distribute commands or other work items to a processing cluster array **1412**. In at least one embodiment, scheduler **1410** ensures that processing cluster array **1412** is properly configured and in a valid state before tasks are distributed to a cluster of processing cluster array **1412**. In at least one embodiment, scheduler **1410** is implemented via firmware logic executing on a microcontroller. In at least one embodiment, microcontroller implemented scheduler **1410** is configurable to perform complex scheduling and work distribution operations at coarse and fine granularity, enabling rapid preemption and context switching of threads executing on processing array **1412**. In at least one embodiment, host software can prove workloads for scheduling on processing cluster array **1412** via one of multiple graphics processing paths. In at least one embodiment, workloads can then be automatically distributed across processing array cluster **1412** by scheduler **1410** logic within a microcontroller including scheduler **1410**.

[0285] In at least one embodiment, processing cluster array **1412** can include up to "N" processing clusters (e.g., cluster **1414**A, cluster **1414**B, through cluster **1414**N), where "N" represents a positive integer (which may be a different integer "N" than used in other figures). In at least one embodiment, each cluster **1414**A-**1414**N of processing cluster array **1412** can execute a large number of concurrent threads. In at least one embodiment, scheduler **1410** can allocate work to clusters **1414**A-**1414**N of processing cluster array **1412** using various scheduling and/or work distribution algorithms, which may vary depending on workload arising for each type of program or computation. In at least one embodiment, scheduling can be handled dynamically by scheduler **1410**, or can be assisted in part by compiler logic during compilation of program logic configured for execution by processing cluster array **1412**. In at least one embodiment, different clusters **1414**A-**1414**N of processing cluster array **1412** can be allocated for processing different types of programs or for performing different types of computations.

[0286] In at least one embodiment, processing cluster array **1412** can be configured to perform various types of parallel processing operations. In at least one embodiment, processing cluster array **1412** is configured to perform general-purpose parallel compute operations. For example, in at least one embodiment, processing cluster array **1412** can include logic to execute processing tasks including filtering of video and/or audio data, performing modeling operations, including physics operations, and performing data transformations.

[0287] In at least one embodiment, processing cluster array **1412** is configured to perform parallel graphics processing operations. In at least one embodiment, processing cluster array **1412** can include additional logic to support execution of such graphics processing operations, including but not limited to, texture sampling logic to perform texture operations, as well as tessellation logic and other vertex processing logic. In at least one embodiment, processing cluster array **1412** can be configured to execute graphics processing related shader programs such as, but not limited to, vertex shaders, tessellation shaders, geometry shaders, and pixel shaders. In at least one embodiment, parallel processing unit **1402** can transfer data from system memory via I/O unit **1404** for processing. In at least one embodiment, during processing, transferred data can be stored to on-chip memory (e.g., parallel processor memory **1422**) during processing, then written back to system memory.

[0288] In at least one embodiment, when parallel processing unit **1402** is used to perform graphics processing, scheduler **1410** can be configured to divide a processing workload into approximately equal sized tasks, to better enable distribution of graphics processing operations to multiple clusters **1414**A-**1414**N of processing cluster array **1412**. In at least one embodiment, portions of processing cluster array **1412** can be configured to perform different types of processing. For example, in at least one embodiment, a first portion may be configured to perform vertex shading and topology generation, a second portion may be configured to perform tessellation and geometry shading, and a third portion may be configured to perform pixel shading or other screen space operations, to produce a rendered image for display. In at least one embodiment, intermediate data produced by one or more of clusters **1414**A-**1414**N may be stored in buffers to allow intermediate data to be transmitted between clusters **1414**A-**1414**N for further processing.

[0289] In at least one embodiment, processing cluster array **1412** can receive processing tasks to be executed via scheduler **1410**, which receives commands defining processing tasks from front end **1408**. In at least one embodiment, processing tasks can include indices of data to be processed, e.g., surface (patch) data, primitive data, vertex data, and/or pixel data, as well as state parameters and commands defining how data is to be processed (e.g., what program is to be executed). In at least one embodiment, scheduler **1410** may be configured to fetch indices corresponding to tasks or may receive indices from front end **1408**. In at least one embodiment, front end **1408** can be configured to ensure processing cluster array **1412** is configured to a valid state before a workload specified by incoming command buffers (e.g., batch-buffers, push buffers, etc.) is initiated.

[0290] In at least one embodiment, each of one or more instances of parallel processing unit **1402** can couple with a parallel processor memory **1422**. In at least one embodiment, parallel processor memory **1422** can be accessed via memory crossbar **1416**, which can receive memory requests from processing cluster array **1412** as well as I/O unit **1404**. In at least one embodiment, memory crossbar **1416** can access parallel processor memory **1422** via a memory interface **1418**. In at least one embodiment, memory interface **1418** can include multiple partition units (e.g., partition unit **1420**A, partition unit **1420**B, through partition unit **1420**N) that can each couple to a portion (e.g., memory unit) of parallel processor memory **1422**. In at least one embodiment, a number of partition units **1420**A-**1420**N is configured to be equal to a number of memory units, such that a first partition unit **1420**A has a corresponding first memory unit **1424**A, a second partition unit **1420**B has a corresponding memory unit **1424**B, and an N-th partition unit **1420**N has a corresponding N-th memory unit **1424**N. In at least one embodiment, a number of partition units **1420**A-**1420**N may not be equal to a number of memory units.

[0291] In at least one embodiment, memory units **1424**A-**1424**N can include various types of memory devices, including dynamic random access memory (DRAM) or graphics random access memory, such as synchronous graphics random access memory (SGRAM), including graphics double data rate (GDDR) memory. In at least one embodiment, memory units **1424**A-**1424**N may also include 3D stacked memory, including but not limited to high bandwidth memory (HBM). In at least one embodiment, render targets, such as frame buffers or texture maps may be stored across memory units **1424**A-**1424**N, allowing partition units **1420**A-**1420**N to write portions of each render target in parallel to efficiently use available bandwidth of parallel processor memory **1422**. In at least one embodiment, a local instance of parallel processor memory **1422** may be excluded in favor of a unified memory design that utilizes system memory in conjunction with local cache memory.

[0292] In at least one embodiment, any one of clusters **1414**A-**1414**N of processing cluster array **1412** can process data that will be written to any of memory units **1424**A-**1424**N within parallel processor memory **1422**. In at least one embodiment, memory crossbar **1416** can be configured to transfer an output of each cluster **1414**A-**1414**N to any partition unit **1420**A-**1420**N or to another cluster **1414**A-**1414**N, which can perform additional processing operations on an output. In at least one embodiment, each cluster **1414**A-**1414**N can communicate with memory interface **1418** through memory crossbar **1416** to read from or write to various external memory devices. In at least one embodiment, memory crossbar **1416** has a connection to memory interface **1418** to communicate with I/O unit **1404**, as well as a connection to a local instance of parallel processor memory **1422**, enabling processing units within different processing clusters **1414**A-**1414**N to communicate with system memory or other memory that is not local to parallel processing unit **1402**. In at least one embodiment, memory crossbar **1416** can use virtual channels to separate traffic streams between clusters **1414**A-**1414**N and partition units **1420**A-**1420**N.

[0293] In at least one embodiment, multiple instances of parallel processing unit **1402** can be provided on a single add-in card, or multiple add-in cards can be interconnected. In at least one embodiment, different instances of parallel processing unit **1402** can be configured to interoperate even if different instances have different numbers of processing cores, different amounts of local parallel processor memory, and/or other configuration differences. For example, in at least one embodiment, some instances of parallel processing unit **1402** can include higher precision floating point units relative to other instances. In at least one embodiment, systems incorporating one or more instances of parallel processing unit **1402** or parallel processor **1400** can be implemented in a variety of configurations and form factors, including but not limited to desktop, laptop, or handheld personal computers, servers, workstations, game consoles, and/or embedded systems.

[0294] FIG. **14**B is a block diagram of a partition unit **1420** according to at least one embodiment. In at least one embodiment, partition unit **1420** is an instance of one of partition units **1420**A-**1420**N of FIG. **14**A. In at least one embodiment, partition unit **1420** includes an L2 cache **1421**, a frame buffer interface **1425**, and a ROP **1426** (raster operations unit). In at least one embodiment, L2 cache **1421** is a read/write cache that is configured to perform load and store operations received from memory crossbar **1416** and ROP **1426**. In at least one embodiment, read misses and urgent write-back requests are output by L2 cache **1421** to frame buffer interface **1425** for processing. In at least one embodiment, updates can also be sent to a frame buffer via frame buffer interface **1425** for processing. In at least one embodiment, frame buffer interface **1425** interfaces with one of memory units in parallel processor memory, such as memory units **1424**A-**1424**N of FIG. **14** (e.g., within parallel processor memory **1422**).

[0295] In at least one embodiment, ROP **1426** is a processing unit that performs raster operations such as stencil, z test, blending, etc. In at least one embodiment, ROP **1426** then outputs processed graphics data that is stored in graphics memory. In at least one embodiment, ROP **1426** includes compression logic to compress depth or color data that is written to memory and decompress depth or color data that is read from memory. In at least one embodiment, compression logic can be lossless compression logic that makes use of one or more of multiple compression algorithms. In at least one embodiment, a type of compression that is performed by ROP **1426** can vary based on statistical characteristics of data to be compressed. For example, in at least one embodiment, delta color compression is performed on depth and color data on a per-tile basis.

[0296] In at least one embodiment, ROP **1426** is included within each processing cluster (e.g., cluster **1414**A-**1414**N of FIG. **14**A) instead of within partition unit **1420**. In at least one embodiment, read and write requests for pixel data are transmitted over memory crossbar **1416** instead of pixel fragment data. In at least one embodiment, processed graphics data may be displayed on a display device, such as one of one or more display device(s) **1310** of FIG. **13**, routed for further processing by processor(s) **1302**, or routed for further processing by one of processing entities within parallel processor **1400** of FIG. **14**A.

[0297] FIG. **14**C is a block diagram of a processing cluster **1414** within a parallel processing unit according to at least one embodiment. In at least one embodiment, a processing cluster is an instance of one of processing clusters **1414**A-**1414**N of FIG. **14**A. In at least one embodiment, processing cluster **1414** can be configured to execute many threads in parallel, where "thread" refers to an instance of a particular program executing on a particular set of input data. In at least one embodiment, single-instruction, multiple-data (SIMD) instruction issue techniques are used to support parallel execution of a large number of threads without providing multiple independent instruction units. In at least one embodiment, single-instruction, multiple-thread (SIMT) techniques are used to support parallel execution of a large number of generally synchronized threads, using a common instruction unit configured to issue instructions to a set of processing engines within each one of processing clusters.

[0298] In at least one embodiment, operation of processing cluster **1414** can be controlled via a pipeline manager **1432** that distributes processing tasks to SIMT parallel processors. In at least one embodiment, pipeline manager **1432** receives instructions from scheduler **1410** of FIG. **14**A and manages execution of those instructions via a graphics multiprocessor **1434** and/or a texture unit **1436**. In at least one embodiment, graphics multiprocessor **1434** is an exemplary instance of a SIMT parallel processor. However, in at least one embodiment, various types of SIMT parallel processors of differing architectures may be included within processing cluster **1414**. In at least one embodiment, one or more instances of graphics multiprocessor **1434** can be included within a processing cluster **1414**. In at least one embodiment, graphics multiprocessor **1434** can process data and a data crossbar **1440** can be used to distribute processed data to one of multiple possible destinations, including other shader units. In at least one embodiment, pipeline manager **1432** can facilitate distribution of processed data by specifying destinations for processed data to be distributed via data crossbar **1440**.

[0299] In at least one embodiment, each graphics multiprocessor **1434** within processing cluster **1414** can include an identical set of functional execution logic (e.g., arithmetic logic units, load-store units, etc.). In at least one embodiment, functional execution logic can be configured in a pipelined manner in which new instructions can be issued before previous instructions are complete. In at least one embodiment, functional execution logic supports a variety of operations including integer and floating point arithmetic, comparison operations, Boolean operations, bit-shifting, and computation of various algebraic functions. In at least one embodiment, same functional-unit hardware can be leveraged to perform different operations and any combination of functional units may be present.

[0300] In at least one embodiment, instructions transmitted to processing cluster **1414** constitute a thread. In at least one embodiment, a set of threads executing across a set of parallel processing engines is a thread group. In at least one embodiment, a thread group executes a common program on different input data. In at least one embodiment, each thread within a thread group can be assigned to a different processing engine within a graphics multiprocessor **1434**. In at least one embodiment, a thread group may include fewer threads than a number of processing engines within graphics multiprocessor **1434**. In at least one embodiment, when a thread group includes fewer threads than a number of processing engines, one or more of processing engines may be idle during cycles in which that thread group is being processed. In at least one embodiment, a thread group may also include more threads than a number of processing engines within graphics multiprocessor **1434**. In at least one embodiment, when a thread group includes more threads than number of processing engines within graphics multiprocessor **1434**, processing can be performed over consecutive clock cycles. In at least one embodiment, multiple thread groups can be executed concurrently on a graphics multiprocessor **1434**.

[0301] In at least one embodiment, graphics multiprocessor **1434** includes an internal cache memory to perform load and store operations. In at least one embodiment, graphics multiprocessor **1434** can forego an internal cache and use a cache memory (e.g., L1 cache **1448**) within processing cluster **1414**. In at least one embodiment, each graphics multiprocessor **1434** also has access to L2 caches within partition units (e.g., partition units **1420**A-**1420**N of FIG. **14**A) that are shared among all processing clusters **1414** and may be used to transfer data between threads. In at least one embodiment, graphics multiprocessor **1434** may also access off-chip global memory, which can include one or more of local parallel processor memory and/or system memory. In at least one embodiment, any memory external to parallel processing unit **1402** may be used as global memory. In at least one embodiment, processing cluster **1414** includes multiple instances of graphics multiprocessor **1434** and can share common instructions and data, which may be stored in L1 cache **1448**.

[0302] In at least one embodiment, each processing cluster **1414** may include an MMU **1445** (memory management unit) that is configured to map virtual addresses into physical addresses. In at least one embodiment, one or more instances of MMU **1445** may reside within memory interface **1418** of FIG. **14**A. In at least one embodiment, MMU **1445** includes a set of page table entries (PTEs) used to map a virtual address to a physical address of a tile and optionally a cache line index. In at least one embodiment, MMU **1445** may include address translation lookaside buffers (TLB) or caches that may reside within graphics multiprocessor **1434** or L1 **1448** cache or processing cluster **1414**. In at least one embodiment, a physical address is processed to distribute surface data access locally to allow for efficient request interleaving among partition units. In at least one embodiment, a cache line index may be used to determine whether a request for a cache line is a hit or miss.

[0303] In at least one embodiment, a processing cluster **1414** may be configured such that each graphics multiprocessor **1434** is coupled to a texture unit **1436** for performing texture mapping operations, e.g., determining texture sample positions, reading texture data, and filtering texture data. In at least one embodiment, texture data is read from an internal texture L1 cache (not shown) or from an L1 cache within graphics multiprocessor **1434** and is fetched from an L2 cache, local parallel processor memory, or system memory, as needed. In at least one embodiment, each graphics multiprocessor **1434** outputs processed tasks to data crossbar **1440** to provide processed task to another processing cluster **1414** for further processing or to store processed task in an L2 cache, local parallel processor memory, or system memory via memory crossbar **1416**. In at least one embodiment, a preROP **1442** (pre-raster operations unit) is configured to receive data from graphics multiprocessor **1434**, and direct data to ROP units, which may be located with partition units as described herein (e.g., partition units **1420**A-**1420**N of FIG. **14**A). In at least one embodiment, preROP **1442** unit can perform optimizations for color blending, organizing pixel color data, and performing address translations.

[0304] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, inference and/or training logic **115** may be used in graphics processing cluster **1414** for inferencing or predicting operations based, at least in part, on weight parameters calculated using neural network training operations, neural network functions and/or architectures, or neural network use cases described herein.

[0305] FIG. **14**D shows a graphics multiprocessor **1434** according to at least one embodiment. In at least one embodiment, graphics multiprocessor **1434** couples with pipeline manager **1432** of processing cluster **1414**. In at least one embodiment, graphics multiprocessor **1434** has an execution pipeline including but not limited to an instruction cache **1452**, an instruction unit **1454**, an address mapping unit **1456**, a register file **1458**, one or more general purpose graphics processing unit (GPGPU) cores **1462**, and one or more load/store units **1466**. In at least one embodiment, GPGPU cores **1462** and load/store units **1466** are coupled with cache memory **1472** and shared memory **1470** via a memory and cache interconnect **1468**.

[0306] In at least one embodiment, instruction cache **1452** receives a stream of instructions to execute from pipeline manager **1432**. In at least one embodiment, instructions are cached in instruction cache **1452** and dispatched for execution by an instruction unit **1454**. In at least one embodiment, instruction unit **1454** can dispatch instructions as thread groups (e.g., warps), with each thread of thread group assigned to a different execution unit within GPGPU cores **1462**. In at least one embodiment, an instruction can access any of a local, shared, or global address space by specifying an address within a unified address space. In at least one embodiment, address mapping unit **1456** can be used to translate addresses in a unified address space into a distinct memory address that can be accessed by load/store units **1466**.

[0307] In at least one embodiment, register file **1458** provides a set of registers for functional units of graphics multiprocessor **1434**. In at least one embodiment, register file **1458** provides temporary storage for operands connected to data paths of functional units (e.g., GPGPU cores **1462**, load/store units **1466**) of graphics multiprocessor **1434**. In at least one embodiment, register file **1458** is divided between each of functional units such that each functional unit is allocated a dedicated portion of register file **1458**. In at least one embodiment, register file **1458** is divided between different warps being executed by graphics multiprocessor **1434**.

[0308] In at least one embodiment, GPGPU cores **1462** can each include floating point units (FPUs) and/or integer arithmetic logic units (ALUs) that are used to execute instructions of graphics multiprocessor **1434**. In at least one embodiment, GPGPU cores **1462** can be similar in architecture or can differ in architecture. In at least one embodiment, a first portion of GPGPU cores **1462** include a single precision FPU and an integer ALU while a second portion of GPGPU cores include a double precision FPU. In at least one embodiment, FPUs can implement IEEE 754-2008 standard floating point arithmetic or enable variable precision floating point arithmetic. In at least one embodiment, graphics multiprocessor **1434** can additionally include one or more fixed function or special function units to perform specific functions such as copy rectangle or pixel blending operations. In at least one embodiment, one or more of GPGPU cores **1462** can also include fixed or special function logic.

[0309] In at least one embodiment, GPGPU cores **1462** include SIMD logic capable of performing a single instruction on multiple sets of data. In at least one embodiment, GPGPU cores **1462** can physically execute SIMD4, SIMD8, and SIMD16 instructions and logically execute SIMD1, SIMD2, and SIMD32 instructions. In at least one embodiment, SIMD instructions for GPGPU cores can be generated at compile time by a shader compiler or automatically generated when executing programs written and compiled for single program multiple data (SPMD) or SIMT architectures. In at least one embodiment, multiple threads of a program configured for an SIMT execution model can executed via a single SIMD instruction. For example, in at least one embodiment, eight SIMT threads that perform same or similar operations can be executed in parallel via a single SIMD8 logic unit.

[0310] In at least one embodiment, memory and cache interconnect **1468** is an interconnect network that connects each functional unit of graphics multiprocessor **1434** to register file **1458** and to shared memory **1470**. In at least one embodiment, memory and cache interconnect **1468** is a crossbar interconnect that allows load/store unit **1466** to implement load and store operations between shared memory **1470** and register file **1458**. In at least one embodiment, register file **1458** can operate at a same frequency as GPGPU cores **1462**, thus data transfer between GPGPU cores **1462** and register file **1458** can have very low latency. In at least one embodiment, shared memory **1470** can be used to enable communication between threads that execute on functional units within graphics multiprocessor **1434**. In at least one embodiment, cache memory **1472** can be used as a data cache for example, to cache texture data communicated between functional units and texture unit **1436**. In at least one embodiment, shared memory **1470** can also be used as a program managed cache. In at least one embodiment, threads executing on GPGPU cores **1462** can programmatically store data within shared memory in addition to automatically cached data that is stored within cache memory **1472**.

[0311] In at least one embodiment, a parallel processor or GPGPU as described herein is communicatively coupled to host/processor cores to accelerate graphics operations, machine-learning operations, pattern analysis operations, and various general purpose GPU (GPGPU) functions. In at least one embodiment, a GPU may be communicatively coupled to host processor/cores over a bus or other interconnect (e.g., a high-speed interconnect such as PCIe or NVLink). In at least one embodiment, a GPU may be integrated on a package or chip as cores and communicatively coupled to cores over an internal processor bus/

interconnect internal to a package or chip. In at least one embodiment, regardless a manner in which a GPU is connected, processor cores may allocate work to such GPU in a form of sequences of commands/instructions contained in a work descriptor. In at least one embodiment, that GPU then uses dedicated circuitry/logic for efficiently processing these commands/instructions.

[0312] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, inference and/or training logic **115** may be used in graphics multiprocessor **1434** for inferencing or predicting operations based, at least in part, on weight parameters calculated using neural network training operations, neural network functions and/or architectures, or neural network use cases described herein.

[0313] FIG. **15** illustrates a multi-GPU computing system **1500**, according to at least one embodiment. In at least one embodiment, multi-GPU computing system **1500** can include a processor **1502** coupled to multiple general purpose graphics processing units (GPGPUs) **1506**A-D via a host interface switch **1504**. In at least one embodiment, host interface switch **1504** is a PCI express switch device that couples processor **1502** to a PCI express bus over which processor **1502** can communicate with GPGPUs **1506**A-D. In at least one embodiment, GPGPUs **1506**A-D can interconnect via a set of high-speed point-to-point GPU-to-GPU links **1516**. In at least one embodiment, GPU-to-GPU links **1516** connect to each of GPGPUs **1506**A-D via a dedicated GPU link. In at least one embodiment, P2P GPU links **1516** enable direct communication between each of GPGPUs **1506**A-D without requiring communication over host interface bus **1504** to which processor **1502** is connected. In at least one embodiment, with GPU-to-GPU traffic directed to P2P GPU links **1516**, host interface bus **1504** remains available for system memory access or to communicate with other instances of multi-GPU computing system **1500**, for example, via one or more network devices. While in at least one embodiment GPGPUs **1506**A-D connect to processor **1502** via host interface switch **1504**, in at least one embodiment processor **1502** includes direct support for P2P GPU links **1516** and can connect directly to GPGPUs **1506**A-D.

[0314] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, inference and/or training logic **115** may be used in multi-GPU computing system **1500** for inferencing or predicting operations based, at least in part, on weight parameters calculated using neural network training operations, neural network functions and/or architectures, or neural network use cases described herein.

[0315] FIG. **16** is a block diagram of a graphics processor **1600**, according to at least one embodiment. In at least one embodiment, graphics processor **1600** includes a ring interconnect **1602**, a pipeline front-end **1604**, a media engine **1637**, and graphics cores **1680**A-**1680**N. In at least one embodiment, ring interconnect **1602** couples graphics processor **1600** to other processing units, including other graphics processors or one or more general-purpose processor cores. In at least one embodiment, graphics processor **1600** is one of many processors integrated within a multi-core processing system.

[0316] In at least one embodiment, graphics processor **1600** receives batches of commands via ring interconnect **1602**. In at least one embodiment, incoming commands are interpreted by a command streamer **1603** in pipeline front-end **1604**. In at least one embodiment, graphics processor **1600** includes scalable execution logic to perform 3D geometry processing and media processing via graphics core(s) **1680**A-**1680**N. In at least one embodiment, for 3D geometry processing commands, command streamer **1603** supplies commands to geometry pipeline **1636**. In at least one embodiment, for at least some media processing commands, command streamer **1603** supplies commands to a video front end **1634**, which couples with media engine **1637**. In at least one embodiment, media engine **1637** includes a Video Quality Engine (VQE) **1630** for video and image post-processing and a multi-format encode/decode (MFX) **1633** engine to provide hardware-accelerated media data encoding and decoding. In at least one embodiment, geometry pipeline **1636** and media engine **1637** each generate execution threads for thread execution resources provided by at least one graphics core **1680**.

[0317] In at least one embodiment, graphics processor **1600** includes scalable thread execution resources featuring graphics cores **1680**A-**1680**N (which can be modular and are sometimes referred to as core slices), each having multiple sub-cores **1650**A-**50**N, **1660**A-**1660**N (sometimes referred to as core sub-slices). In at least one embodiment, graphics processor **1600** can have any number of graphics cores **1680**A. In at least one embodiment, graphics processor **1600** includes a graphics core **1680**A having at least a first sub-core **1650**A and a second sub-core **1660**A. In at least one embodiment, graphics processor **1600** is a low power processor with a single sub-core (e.g., **1650**A). In at least one embodiment, graphics processor **1600** includes multiple graphics cores **1680**A-**1680**N, each including a set of first sub-cores **1650**A-**1650**N and a set of second sub-cores **1660**A-**1660**N. In at least one embodiment, each sub-core in first sub-cores **1650**A-**1650**N includes at least a first set of execution units **1652**A-**1652**N and media/texture samplers **1654**A-**1654**N. In at least one embodiment, each sub-core in second sub-cores **1660**A-**1660**N includes at least a second set of execution units **1662**A-**1662**N and samplers **1664**A-**1664**N. In at least one embodiment, each sub-core **1650**A-**1650**N, **1660**A-**1660**N shares a set of shared resources **1670**A-**1670**N. In at least one embodiment, shared resources include shared cache memory and pixel operation logic.

[0318] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, inference and/or training logic **115** may be used in graphics processor **1600** for inferencing or predicting operations based, at least in part, on weight parameters calculated using neural network training operations, neural network functions and/or architectures, or neural network use cases described herein.

[0319] FIG. **17** is a block diagram illustrating micro-architecture for a processor **1700** that may include logic circuits to perform instructions, according to at least one embodiment. In at least one embodiment, processor **1700**

may perform instructions, including x86 instructions, ARM instructions, specialized instructions for application-specific integrated circuits (ASICs), etc. In at least one embodiment, processor **1700** may include registers to store packed data, such as 64-bit wide MMX™ registers in microprocessors enabled with MMX technology from Intel Corporation of Santa Clara, Calif. In at least one embodiment, MMX registers, available in both integer and floating point forms, may operate with packed data elements that accompany single instruction, multiple data ("SIND") and streaming SIMD extensions ("SSE") instructions. In at least one embodiment, 128-bit wide XMM registers relating to SSE2, SSE3, SSE4, AVX, or beyond (referred to generically as "SSEx") technology may hold such packed data operands. In at least one embodiment, processor **1700** may perform instructions to accelerate machine learning or deep learning algorithms, training, or inferencing.

[0320] In at least one embodiment, processor **1700** includes an in-order front end ("front end") **1701** to fetch instructions to be executed and prepare instructions to be used later in a processor pipeline. In at least one embodiment, front end **1701** may include several units. In at least one embodiment, an instruction prefetcher **1726** fetches instructions from memory and feeds instructions to an instruction decoder **1728** which in turn decodes or interprets instructions. For example, in at least one embodiment, instruction decoder **1728** decodes a received instruction into one or more operations called "micro-instructions" or "micro-operations" (also called "micro ops" or "uops") that a machine may execute. In at least one embodiment, instruction decoder **1728** parses an instruction into an opcode and corresponding data and control fields that may be used by micro-architecture to perform operations in accordance with at least one embodiment. In at least one embodiment, a trace cache **1730** may assemble decoded uops into program ordered sequences or traces in a uop queue **1734** for execution. In at least one embodiment, when trace cache **1730** encounters a complex instruction, a microcode ROM **1732** provides uops needed to complete an operation.

[0321] In at least one embodiment, some instructions may be converted into a single micro-op, whereas others need several micro-ops to complete full operation. In at least one embodiment, if more than four micro-ops are needed to complete an instruction, instruction decoder **1728** may access microcode ROM **1732** to perform that instruction. In at least one embodiment, an instruction may be decoded into a small number of micro-ops for processing at instruction decoder **1728**. In at least one embodiment, an instruction may be stored within microcode ROM **1732** should a number of micro-ops be needed to accomplish such operation. In at least one embodiment, trace cache **1730** refers to an entry point programmable logic array ("PLA") to determine a correct micro-instruction pointer for reading microcode sequences to complete one or more instructions from microcode ROM **1732** in accordance with at least one embodiment. In at least one embodiment, after microcode ROM **1732** finishes sequencing micro-ops for an instruction, front end **1701** of a machine may resume fetching micro-ops from trace cache **1730**.

[0322] In at least one embodiment, out-of-order execution engine ("out of order engine") **1703** may prepare instructions for execution. In at least one embodiment, out-of-order execution logic has a number of buffers to smooth out and re-order flow of instructions to optimize performance as they go down a pipeline and get scheduled for execution. In at least one embodiment, out-of-order execution engine **1703** includes, without limitation, an allocator/register renamer **1740**, a memory uop queue **1742**, an integer/floating point uop queue **1744**, a memory scheduler **1746**, a fast scheduler **1702**, a slow/general floating point scheduler ("slow/general FP scheduler") **1704**, and a simple floating point scheduler ("simple FP scheduler") **1706**. In at least one embodiment, fast schedule **1702**, slow/general floating point scheduler **1704**, and simple floating point scheduler **1706** are also collectively referred to herein as "uop schedulers **1702**, **1704**, **1706**." In at least one embodiment, allocator/register renamer **1740** allocates machine buffers and resources that each uop needs in order to execute. In at least one embodiment, allocator/register renamer **1740** renames logic registers onto entries in a register file. In at least one embodiment, allocator/register renamer **1740** also allocates an entry for each uop in one of two uop queues, memory uop queue **1742** for memory operations and integer/floating point uop queue **1744** for non-memory operations, in front of memory scheduler **1746** and uop schedulers **1702**, **1704**, **1706**. In at least one embodiment, uop schedulers **1702**, **1704**, **1706**, determine when a uop is ready to execute based on readiness of their dependent input register operand sources and availability of execution resources uops need to complete their operation. In at least one embodiment, fast scheduler **1702** may schedule on each half of a main clock cycle while slow/general floating point scheduler **1704** and simple floating point scheduler **1706** may schedule once per main processor clock cycle. In at least one embodiment, uop schedulers **1702**, **1704**, **1706** arbitrate for dispatch ports to schedule uops for execution.

[0323] In at least one embodiment, execution block **1711** includes, without limitation, an integer register file/bypass network **1708**, a floating point register file/bypass network ("FP register file/bypass network") **1710**, address generation units ("AGUs") **1712** and **1714**, fast Arithmetic Logic Units (ALUs) ("fast ALUs") **1716** and **1718**, a slow Arithmetic Logic Unit ("slow ALU") **1720**, a floating point ALU ("FP") **1722**, and a floating point move unit ("FP move") **1724**. In at least one embodiment, integer register file/bypass network **1708** and floating point register file/bypass network **1710** are also referred to herein as "register files **1708**, **1710**." In at least one embodiment, AGUSs **1712** and **1714**, fast ALUs **1716** and **1718**, slow ALU **1720**, floating point ALU **1722**, and floating point move unit **1724** are also referred to herein as "execution units **1712**, **1714**, **1716**, **1718**, **1720**, **1722**, and **1724**." In at least one embodiment, execution block **1711** may include, without limitation, any number (including zero) and type of register files, bypass networks, address generation units, and execution units, in any combination.

[0324] In at least one embodiment, register networks **1708**, **1710** may be arranged between uop schedulers **1702**, **1704**, **1706**, and execution units **1712**, **1714**, **1716**, **1718**, **1720**, **1722**, and **1724**. In at least one embodiment, integer register file/bypass network **1708** performs integer operations. In at least one embodiment, floating point register file/bypass network **1710** performs floating point operations. In at least one embodiment, each of register networks **1708**, **1710** may include, without limitation, a bypass network that may bypass or forward just completed results that have not yet been written into a register file to new dependent uops. In at least one embodiment, register networks **1708**, **1710** may communicate data with each other. In at least one embodiment, integer register file/bypass network **1708** may include, without limitation, two separate register files, one register file for a low-order thirty-two bits of data and a second register file for a high order thirty-two bits of data. In at least one embodiment, floating point register file/bypass network **1710** may include, without limitation, 128-bit wide entries because floating point instructions typically have operands from 64 to 128 bits in width.

[0325] In at least one embodiment, execution units **1712**, **1714**, **1716**, **1718**, **1720**, **1722**, **1724** may execute instructions. In at least one embodiment, register networks **1708**, **1710** store integer and floating point data operand values that micro-instructions need to execute. In at least one embodiment, processor **1700** may include, without limitation, any number and combination of execution units **1712**, **1714**, **1716**, **1718**, **1720**, **1722**, **1724**. In at least one embodiment, floating point ALU **1722** and floating point move unit **1724**, may execute floating point, MMX, SIMD, AVX and SSE, or other operations, including specialized machine learning instructions. In at least one embodiment, floating point ALU **1722** may include, without limitation, a 64-bit by 64-bit floating point divider to execute divide, square root, and remainder micro ops. In at least one embodiment, instructions involving a floating point value may be handled with floating point hardware. In at least one embodiment, ALU operations may be passed to fast ALUs **1716**, **1718**. In at least one embodiment, fast ALUS **1716**, **1718** may execute fast operations with an effective latency of half a clock cycle. In at least one embodiment, most complex integer operations go to slow ALU **1720** as slow ALU **1720** may include, without limitation, integer execution hardware for long-latency type of operations, such as a multiplier, shifts, flag logic, and branch processing. In at least one embodiment, memory load/store operations may be executed by AGUs **1712**, **1714**. In at least one embodiment, fast ALU **1716**, fast ALU **1718**, and slow ALU **1720** may perform integer operations on 64-bit data operands. In at least one embodiment, fast ALU **1716**, fast ALU **1718**, and slow ALU **1720** may be implemented to support a variety of data bit sizes including sixteen, thirty-two, 128, 256, etc. In at least one embodiment, floating point ALU **1722** and floating point move unit **1724** may be implemented to support a range of operands having bits of various widths, such as 128-bit wide packed data operands in conjunction with SIMD and multimedia instructions.

[0326] In at least one embodiment, uop schedulers **1702**, **1704**, **1706** dispatch dependent operations before a parent load has finished executing. In at least one embodiment, as uops may be speculatively scheduled and executed in processor **1700**, processor **1700** may also include logic to handle memory misses. In at least one embodiment, if a data load misses in a data cache, there may be dependent operations in flight in a pipeline that have left a scheduler with temporarily incorrect data. In at least one embodiment, a replay mechanism tracks and re-executes instructions that use incorrect data. In at least one embodiment, dependent operations might need to be replayed and independent ones may be allowed to complete. In at least one embodiment, schedulers and a replay mechanism of at least one embodiment of a processor may also be designed to catch instruction sequences for text string comparison operations.

[0327] In at least one embodiment, "registers" may refer to on-board processor storage locations that may be used as part of instructions to identify operands. In at least one embodiment, registers may be those that may be usable from outside of a processor (from a programmer's perspective). In at least one embodiment, registers might not be limited to a particular type of circuit. Rather, in at least one embodiment, a register may store data, provide data, and perform functions described herein. In at least one embodiment, registers described herein may be implemented by circuitry within a processor using any number of different techniques, such as dedicated physical registers, dynamically allocated physical registers using register renaming, combinations of dedicated and dynamically allocated physical registers, etc. In at least one embodiment, integer registers store 32-bit integer data. A register file of at least one embodiment also contains eight multimedia SIND registers for packed data.

[0328] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1A** and/or **1B**. In at least one embodiment portions or all of inference and/or training logic **115** may be incorporated into execution block **1711** and other memory or registers shown or not shown. For example, in at least one embodiment, training and/or inferencing techniques described herein may use one or more of ALUs illustrated in execution block **1711**. Moreover, weight parameters may be stored in on-chip or off-chip memory and/or registers (shown or not shown) that configure ALUs of execution block **1711** to perform one or more machine learning algorithms, neural network architectures, use cases, or training techniques described herein.

[0329] FIG. **18** illustrates a deep learning application processor **1800**, according to at least one embodiment. In at least one embodiment, deep learning application processor **1800** uses instructions that, if executed by deep learning application processor **1800**, cause deep learning application processor **1800** to perform some or all of processes and techniques described throughout this disclosure. In at least one embodiment, deep learning application processor **1800** is an application-specific integrated circuit (ASIC). In at least one embodiment, application processor **1800** performs matrix multiply operations either "hard-wired" into hardware as a result of performing one or more instructions or both. In at least one embodiment, deep learning application processor **1800** includes, without limitation, processing clusters **1810(1)**-**1810(12)**, Inter-Chip Links ("ICLs") **1820(1)**-**1820(12)**, Inter-Chip Controllers ("ICCs") **1830(1)**-**1830(2)**, high-bandwidth memory second generation ("HBM2") **1840(1)**-**1840(4)**, memory controllers ("Mem Ctrlrs") **1842(1)**-**1842(4)**, high bandwidth memory physical layer ("HBM PHY") **1844(1)**-**1844(4)**, a management-controller central processing unit ("management-controller CPU") **1850**, a Serial Peripheral Interface, Inter-Integrated Circuit, and General Purpose Input/Output block ("SPI, I^{2}C, GPIO") **1860**, a peripheral component interconnect express controller and direct memory access block ("PCIe Controller and DMA") **1870**, and a sixteen-lane peripheral component interconnect express port ("PCI Express x 16") **1880**.

[0330] In at least one embodiment, processing clusters **1810** may perform deep learning operations, including inference or prediction operations based on weight parameters calculated one or more training techniques, including those described herein. In at least one embodiment, each processing cluster **1810** may include, without limitation, any number and type of processors. In at least one embodiment, deep learning application processor **1800** may include any number and type of processing clusters **1800**. In at least one embodiment, Inter-Chip Links **1820** are bi-directional. In at least one embodiment, Inter-Chip Links **1820** and Inter-Chip Controllers **1830** enable multiple deep learning application processors **1800** to exchange information, including activation information resulting from performing one or more machine learning algorithms embodied in one or more neural networks. In at least one embodiment, deep learning application processor **1800** may include any number (including zero) and type of ICLs **1820** and ICCs **1830**.

[0331] In at least one embodiment, HBM2s **1840** provide a total of 32 Gigabytes (GB) of memory. In at least one embodiment, HBM2 **1840**(i) is associated with both memory controller **1842**(*i*) and HBM PHY **1844**(*i*) where "i" is an arbitrary integer. In at least one embodiment, any number of HBM2s **1840** may provide any type and total amount of high bandwidth memory and may be associated with any number (including zero) and type of memory controllers **1842** and HBM PHYs **1844**. In at least one embodiment, SPI, I^2C, GPIO **1860**, PCIe Controller and DMA **1870**, and/or PCIe **1880** may be replaced with any number and type of blocks that enable any number and type of communication standards in any technically feasible fashion.

[0332] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, deep learning application processor is used to train a machine learning model, such as a neural network, to predict or infer information provided to deep learning application processor **1800**. In at least one embodiment, deep learning application processor **1800** is used to infer or predict information based on a trained machine learning model (e.g., neural network) that has been trained by another processor or system or by deep learning application processor **1800**. In at least one embodiment, processor **1800** may be used to perform one or more neural network use cases described herein.

[0333] FIG. **19** is a block diagram of a neuromorphic processor **1900**, according to at least one embodiment. In at least one embodiment, neuromorphic processor **1900** may receive one or more inputs from sources external to neuromorphic processor **1900**. In at least one embodiment, these inputs may be transmitted to one or more neurons **1902** within neuromorphic processor **1900**. In at least one embodiment, neurons **1902** and components thereof may be implemented using circuitry or logic, including one or more arithmetic logic units (ALUs). In at least one embodiment, neuromorphic processor **1900** may include, without limitation, thousands or millions of instances of neurons **1902**, but any suitable number of neurons **1902** may be used. In at least one embodiment, each instance of neuron **1902** may include a neuron input **1904** and a neuron output **1906**. In at least one embodiment, neurons **1902** may generate outputs that may be transmitted to inputs of other instances of neurons **1902**. For example, in at least one embodiment, neuron inputs **1904** and neuron outputs **1906** may be interconnected via synapses **1908**.

[0334] In at least one embodiment, neurons **1902** and synapses **1908** may be interconnected such that neuromorphic processor **1900** operates to process or analyze information received by neuromorphic processor **1900**. In at least one embodiment, neurons **1902** may transmit an output pulse (or "fire" or "spike") when inputs received through neuron input **1904** exceed a threshold. In at least one embodiment, neurons **1902** may sum or integrate signals received at neuron inputs **1904**. For example, in at least one embodiment, neurons **1902** may be implemented as leaky integrate-and-fire neurons, wherein if a sum (referred to as a "membrane potential") exceeds a threshold value, neuron **1902** may generate an output (or "fire") using a transfer function such as a sigmoid or threshold function. In at least one embodiment, a leaky integrate-and-fire neuron may sum signals received at neuron inputs **1904** into a membrane potential and may also apply a decay factor (or leak) to reduce a membrane potential. In at least one embodiment, a leaky integrate-and-fire neuron may fire if multiple input signals are received at neuron inputs **1904** rapidly enough to exceed a threshold value (i.e., before a membrane potential decays too low to fire). In at least one embodiment, neurons **1902** may be implemented using circuits or logic that receive inputs, integrate inputs into a membrane potential, and decay a membrane potential. In at least one embodiment, inputs may be averaged, or any other suitable transfer function may be used. Furthermore, in at least one embodiment, neurons **1902** may include, without limitation, comparator circuits or logic that generate an output spike at neuron output **1906** when result of applying a transfer function to neuron input **1904** exceeds a threshold. In at least one embodiment, once neuron **1902** fires, it may disregard previously received input information by, for example, resetting a membrane potential to 0 or another suitable default value. In at least one embodiment, once membrane potential is reset to 0, neuron **1902** may resume normal operation after a suitable period of time (or refractory period).

[0335] In at least one embodiment, neurons **1902** may be interconnected through synapses **1908**. In at least one embodiment, synapses **1908** may operate to transmit signals from an output of a first neuron **1902** to an input of a second neuron **1902**. In at least one embodiment, neurons **1902** may transmit information over more than one instance of synapse **1908**. In at least one embodiment, one or more instances of neuron output **1906** may be connected, via an instance of synapse **1908**, to an instance of neuron input **1904** in same neuron **1902**. In at least one embodiment, an instance of neuron **1902** generating an output to be transmitted over an instance of synapse **1908** may be referred to as a "presynaptic neuron" with respect to that instance of synapse **1908**. In at least one embodiment, an instance of neuron **1902** receiving an input transmitted over an instance of synapse **1908** may be referred to as a "post-synaptic neuron" with respect to that instance of synapse **1908**. Because an instance of neuron **1902** may receive inputs from one or more instances of synapse **1908**, and may also transmit outputs over one or more instances of synapse **1908**, a single instance of neuron **1902** may therefore be both a "presynaptic neuron" and "post-synaptic neuron," with respect to various instances of synapses **1908**, in at least one embodiment.

[0336] In at least one embodiment, neurons **1902** may be organized into one or more layers. In at least one embodiment, each instance of neuron **1902** may have one neuron output **1906** that may fan out through one or more synapses **1908** to one or more neuron inputs **1904**. In at least one embodiment, neuron outputs **1906** of neurons **1902** in a first layer **1910** may be connected to neuron inputs **1904** of neurons **1902** in a second layer **1912**. In at least one embodiment, layer **1910** may be referred to as a "feed-forward layer." In at least one embodiment, each instance of neuron **1902** in an instance of first layer **1910** may fan out to each instance of neuron **1902** in second layer **1912**. In at least one embodiment, first layer **1910** may be referred to as a "fully connected feed-forward layer." In at least one embodiment, each instance of neuron **1902** in an instance of second layer **1912** may fan out to fewer than all instances of neuron **1902** in a third layer **1914**. In at least one embodiment, second layer **1912** may be referred to as a "sparsely connected feed-forward layer." In at least one embodiment, neurons **1902** in second layer **1912** may fan out to neurons **1902** in multiple other layers, including to neurons **1902** also in second layer **1912**. In at least one embodiment, second layer **1912** may be referred to as a "recurrent layer." In at least one embodiment, neuromorphic processor **1900** may include, without limitation, any suitable combination of recurrent layers and feed-forward layers, including, without limitation, both sparsely connected feed-forward layers and fully connected feed-forward layers.

[0337] In at least one embodiment, neuromorphic processor **1900** may include, without limitation, a reconfigurable interconnect architecture or dedicated hard-wired interconnects to connect synapse **1908** to neurons **1902**. In at least one embodiment, neuromorphic processor **1900** may include, without limitation, circuitry or logic that allows synapses to be allocated to different neurons **1902** as needed based on neural network topology and neuron fan-in/out. For example, in at least one embodiment, synapses **1908** may be connected to neurons **1902** using an interconnect fabric, such as network-on-chip, or with dedicated connections. In at least one embodiment, synapse interconnections and components thereof may be implemented using circuitry or logic.

[0338] FIG. **20** is a block diagram of a processing system, according to at least one embodiment. In at least one embodiment, system **2000** includes one or more processors **2002** and one or more graphics processors **2008**, and may be a single processor desktop system, a multiprocessor workstation system, or a server system having a large number of processors **2002** or processor cores **2007**. In at least one embodiment, system **2000** is a processing platform incorporated within a system-on-a-chip (SoC) integrated circuit for use in mobile, handheld, or embedded devices.

[0339] In at least one embodiment, system **2000** can include, or be incorporated within a server-based gaming platform, a game console, including a game and media console, a mobile gaming console, a handheld game console, or an online game console. In at least one embodiment, system **2000** is a mobile phone, a smart phone, a tablet computing device or a mobile Internet device. In at least one embodiment, processing system **2000** can also include, couple with, or be integrated within a wearable device, such as a smart watch wearable device, a smart eyewear device, an augmented reality device, or a virtual reality device. In at least one embodiment, processing system **2000** is a television or set top box device having one or more processors **2002** and a graphical interface generated by one or more graphics processors **2008**.

[0340] In at least one embodiment, one or more processors **2002** each include one or more processor cores **2007** to process instructions which, when executed, perform operations for system and user software. In at least one embodiment, each of one or more processor cores **2007** is configured to process a specific instruction sequence **2009**. In at least one embodiment, instruction sequence **2009** may facilitate Complex Instruction Set Computing (CISC), Reduced Instruction Set Computing (RISC), or computing via a Very Long Instruction Word (VLIW). In at least one embodiment, processor cores **2007** may each process a different instruction sequence **2009**, which may include instructions to facilitate emulation of other instruction sequences. In at least one embodiment, processor core **2007** may also include other processing devices, such a Digital Signal Processor (DSP).

[0341] In at least one embodiment, processor **2002** includes a cache memory **2004**. In at least one embodiment, processor **2002** can have a single internal cache or multiple levels of internal cache. In at least one embodiment, cache memory is shared among various components of processor **2002**. In at least one embodiment, processor **2002** also uses an external cache (e.g., a Level-3 (L3) cache or Last Level Cache (LLC)) (not shown), which may be shared among processor cores **2007** using known cache coherency techniques. In at least one embodiment, a register file **2006** is additionally included in processor **2002**, which may include different types of registers for storing different types of data (e.g., integer registers, floating point registers, status registers, and an instruction pointer register). In at least one embodiment, register file **2006** may include general-purpose registers or other registers.

[0342] In at least one embodiment, one or more processor(s) **2002** are coupled with one or more interface bus(es) **2010** to transmit communication signals such as address, data, or control signals between processor **2002** and other components in system **2000**. In at least one embodiment, interface bus **2010** can be a processor bus, such as a version of a Direct Media Interface (DMI) bus. In at least one embodiment, interface bus **2010** is not limited to a DMI bus, and may include one or more Peripheral Component Interconnect buses (e.g., PCI, PCI Express), memory busses, or other types of interface busses. In at least one embodiment processor(s) **2002** include an integrated memory controller **2016** and a platform controller hub **2030**. In at least one embodiment, memory controller **2016** facilitates communication between a memory device and other components of system **2000**, while platform controller hub (PCH) **2030** provides connections to I/O devices via a local I/O bus.

[0343] In at least one embodiment, a memory device **2020** can be a dynamic random access memory (DRAM) device, a static random access memory (SRAM) device, flash memory device, phase-change memory device, or some other memory device having suitable performance to serve as process memory. In at least one embodiment, memory device **2020** can operate as system memory for system **2000**, to store data **2022** and instructions **2021** for use when one or more processors **2002** executes an application or process. In at least one embodiment, memory controller **2016** also couples with an optional external graphics processor **2012**, which may communicate with one or more graphics processors **2008** in processors **2002** to perform graphics and media operations. In at least one embodiment, a display device **2011** can connect to processor(s) **2002**. In at least one embodiment, display device **2011** can include one or more of an internal display device, as in a mobile electronic device or a laptop device, or an external display device attached via a display interface (e.g., DisplayPort, etc.). In at least one embodiment, display device **2011** can include a head mounted display (HMD) such as a stereoscopic display device for use in virtual reality (VR) applications or augmented reality (AR) applications.

[0344] In at least one embodiment, platform controller hub **2030** enables peripherals to connect to memory device **2020** and processor **2002** via a high-speed I/O bus. In at least one embodiment, I/O peripherals include, but are not limited to, an audio controller **2046**, a network controller **2034**, a firmware interface **2028**, a wireless transceiver **2026**, touch sensors **2025**, a data storage device **2024** (e.g., hard disk drive, flash memory, etc.). In at least one embodiment, data storage device **2024** can connect via a storage interface (e.g., SATA) or via a peripheral bus, such as a Peripheral Component Interconnect bus (e.g., PCI, PCI Express). In at least one embodiment, touch sensors **2025** can include touch screen sensors, pressure sensors, or fingerprint sensors. In at least one embodiment, wireless transceiver **2026** can be a Wi-Fi transceiver, a Bluetooth transceiver, or a mobile network transceiver such as a 3G, 4G, or Long Term Evolution (LTE) transceiver. In at least one embodiment, firmware interface **2028** enables communication with system firmware, and can be, for example, a unified extensible firmware interface (UEFI). In at least one embodiment, network controller **2034** can enable a network connection to a wired network. In at least one embodiment, a high-performance network controller (not shown) couples with interface bus **2010**. In at least one embodiment, audio controller **2046** is a multi-channel high definition audio controller. In at least one embodiment, system **2000** includes an optional legacy I/O controller **2040** for coupling legacy (e.g., Personal System 2 (PS/2)) devices to system **2000**. In at least one embodiment, platform controller hub **2030** can also connect to one or more Universal Serial Bus (USB) controllers **2042** connect input devices, such as keyboard and mouse **2043** combinations, a camera **2044**, or other USB input devices.

[0345] In at least one embodiment, an instance of memory controller **2016** and platform controller hub **2030** may be integrated into a discreet external graphics processor, such as external graphics processor **2012**. In at least one embodiment, platform controller hub **2030** and/or memory controller **2016** may be external to one or more processor(s) **2002**. For example, in at least one embodiment, system **2000** can include an external memory controller **2016** and platform controller hub **2030**, which may be configured as a memory controller hub and peripheral controller hub within a system chipset that is in communication with processor(s) **2002**.

[0346] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment portions or all of inference and/or training logic **115** may be incorporated into graphics processor **2000**. For example, in at least one embodiment, training and/or inferencing techniques described herein may use one or more of ALUs embodied in a 3D pipeline. Moreover, in at least one embodiment, inferencing and/or training operations described herein may be done using logic other than logic illustrated in FIG. **1**A or **1**B. In at least one embodiment, weight parameters may be stored in on-chip or off-chip memory and/or registers (shown or not shown) that configure ALUs of graphics processor **2000** to perform one or more machine learning algorithms, neural network architectures, use cases, or training techniques described herein.

[0347] FIG. **21** is a block diagram of a processor **2100** having one or more processor cores **2102**A-**2102**N, an integrated memory controller **2114**, and an integrated graphics processor **2108**, according to at least one embodiment. In at least one embodiment, processor **2100** can include additional cores up to and including additional core **2102**N represented by dashed lined boxes. In at least one embodiment, each of processor cores **2102**A-**2102**N includes one or more internal cache units **2104**A-**2104**N. In at least one embodiment, each processor core also has access to one or more shared cached units **2106**.

[0348] In at least one embodiment, internal cache units **2104**A-**2104**N and shared cache units **2106** represent a cache memory hierarchy within processor **2100**. In at least one embodiment, cache memory units **2104**A-**2104**N may include at least one level of instruction and data cache within each processor core and one or more levels of shared mid-level cache, such as a Level 2 (L2), Level 3 (L3), Level 4 (L4), or other levels of cache, where a highest level of cache before external memory is classified as an LLC. In at least one embodiment, cache coherency logic maintains coherency between various cache units **2106** and **2104**A-**2104**N.

[0349] In at least one embodiment, processor **2100** may also include a set of one or more bus controller units **2116** and a system agent core **2110**. In at least one embodiment, bus controller units **2116** manage a set of peripheral buses, such as one or more PCI or PCI express busses. In at least one embodiment, system agent core **2110** provides management functionality for various processor components. In at least one embodiment, system agent core **2110** includes one or more integrated memory controllers **2114** to manage access to various external memory devices (not shown).

[0350] In at least one embodiment, one or more of processor cores **2102**A-**2102**N include support for simultaneous multi-threading. In at least one embodiment, system agent core **2110** includes components for coordinating and operating cores **2102**A-**2102**N during multi-threaded processing. In at least one embodiment, system agent core **2110** may additionally include a power control unit (PCU), which includes logic and components to regulate one or more power states of processor cores **2102**A-**2102**N and graphics processor **2108**.

[0351] In at least one embodiment, processor **2100** additionally includes graphics processor **2108** to execute graphics processing operations. In at least one embodiment, graphics processor **2108** couples with shared cache units **2106**, and system agent core **2110**, including one or more integrated memory controllers **2114**. In at least one embodiment, system agent core **2110** also includes a display controller **2111** to drive graphics processor output to one or more coupled displays. In at least one embodiment, display controller **2111** may also be a separate module coupled with graphics processor **2108** via at least one interconnect, or may be integrated within graphics processor **2108**.

[0352] In at least one embodiment, a ring-based interconnect unit **2112** is used to couple internal components of processor **2100**. In at least one embodiment, an alternative interconnect unit may be used, such as a point-to-point interconnect, a switched interconnect, or other techniques. In at least one embodiment, graphics processor **2108** couples with ring interconnect **2112** via an I/O link **2113**.

[0353] In at least one embodiment, I/O link **2113** represents at least one of multiple varieties of I/O interconnects, including an on package I/O interconnect which facilitates communication between various processor components and a high-performance embedded memory module **2118**, such as an eDRAM module. In at least one embodiment, each of processor cores **2102**A-**2102**N and graphics processor **2108** use embedded memory module **2118** as a shared Last Level Cache.

[0354] In at least one embodiment, processor cores **2102**A-**2102**N are homogeneous cores executing a common instruction set architecture. In at least one embodiment, processor cores **2102**A-**2102**N are heterogeneous in terms of instruction set architecture (ISA), where one or more of processor cores **2102**A-**2102**N execute a common instruction set, while one or more other cores of processor cores **2102**A-**2102**N executes a subset of a common instruction set or a different instruction set. In at least one embodiment, processor cores **2102**A-**2102**N are heterogeneous in terms of microarchitecture, where one or more cores having a relatively higher power consumption couple with one or more power cores having a lower power consumption. In at least one embodiment, processor **2100** can be implemented on one or more chips or as an SoC integrated circuit.

[0355] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment portions or all of inference and/or training logic **115** may be incorporated into graphics processor **2110**. For example, in at least one embodiment, training and/or inferencing techniques described herein may use one or more of ALUs embodied in a 3D pipeline, graphics core(s) **2102**, shared function logic, or other logic in FIG. **21**. Moreover, in at least one embodiment, inferencing and/or training operations described herein may be done using logic other than logic illustrated in FIG. **1**A or **1**B. In at least one embodiment, weight parameters may be stored in on-chip or off-chip memory and/or registers (shown or not shown) that configure ALUs of processor **2100** to perform one or more machine learning algorithms, neural network architectures, use cases, or training techniques described herein.

[0356] FIG. **22** is a block diagram of a graphics processor **2200**, which may be a discrete graphics processing unit, or may be a graphics processor integrated with a plurality of processing cores. In at least one embodiment, graphics processor **2200** communicates via a memory mapped I/O interface to registers on graphics processor **2200** and with commands placed into memory. In at least one embodiment, graphics processor **2200** includes a memory interface **2214** to access memory. In at least one embodiment, memory interface **2214** is an interface to local memory, one or more internal caches, one or more shared external caches, and/or to system memory.

[0357] In at least one embodiment, graphics processor **2200** also includes a display controller **2202** to drive display output data to a display device **2220**. In at least one embodiment, display controller **2202** includes hardware for one or more overlay planes for display device **2220** and composition of multiple layers of video or user interface elements. In at least one embodiment, display device **2220** can be an internal or external display device. In at least one embodiment, display device **2220** is a head mounted display device, such as a virtual reality (VR) display device or an augmented reality (AR) display device. In at least one embodiment, graphics processor **2200** includes a video codec engine **2206** to encode, decode, or transcode media to, from, or between one or more media encoding formats, including, but not limited to Moving Picture Experts Group (MPEG) formats such as MPEG-2, Advanced Video Coding (AVC) formats such as H.264/MPEG-4 AVC, as well as the Society of Motion Picture & Television Engineers (SMPTE) 421M/VC-1, and Joint Photographic Experts Group (JPEG) formats such as JPEG, and Motion JPEG (MJPEG) formats.

[0358] In at least one embodiment, graphics processor **2200** includes a block image transfer (BLIT) engine **2204** to perform two-dimensional (2D) rasterizer operations including, for example, bit-boundary block transfers. However, in at least one embodiment, 2D graphics operations are performed using one or more components of a graphics processing engine (GPE) **2210**. In at least one embodiment, GPE **2210** is a compute engine for performing graphics operations, including three-dimensional (3D) graphics operations and media operations.

[0359] In at least one embodiment, GPE **2210** includes a 3D pipeline **2212** for performing 3D operations, such as rendering three-dimensional images and scenes using processing functions that act upon 3D primitive shapes (e.g., rectangle, triangle, etc.). In at least one embodiment, 3D pipeline **2212** includes programmable and fixed function elements that perform various tasks and/or spawn execution threads to a 3D/Media sub-system **2215**. While 3D pipeline **2212** can be used to perform media operations, in at least one embodiment, GPE **2210** also includes a media pipeline **2216** that is used to perform media operations, such as video post-processing and image enhancement.

[0360] In at least one embodiment, media pipeline **2216** includes fixed function or programmable logic units to perform one or more specialized media operations, such as video decode acceleration, video de-interlacing, and video encode acceleration in place of, or on behalf of, video codec engine **2206**. In at least one embodiment, media pipeline **2216** additionally includes a thread spawning unit to spawn threads for execution on 3D/Media sub-system **2215**. In at least one embodiment, spawned threads perform computations for media operations on one or more graphics execution units included in 3D/Media sub-system **2215**.

[0361] In at least one embodiment, 3D/Media subsystem **2215** includes logic for executing threads spawned by 3D pipeline **2212** and media pipeline **2216**. In at least one embodiment, 3D pipeline **2212** and media pipeline **2216** send thread execution requests to 3D/Media subsystem **2215**, which includes thread dispatch logic for arbitrating and dispatching various requests to available thread execution resources. In at least one embodiment, execution resources include an array of graphics execution units to process 3D and media threads. In at least one embodiment, 3D/Media subsystem **2215** includes one or more internal caches for thread instructions and data. In at least one embodiment, subsystem **2215** also includes shared memory, including registers and addressable memory, to share data between threads and to store output data.

[0362] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment portions or all of inference and/or training logic **115** may be incorporated into graphics processor **2200**. For example, in at least one embodiment, training and/or inferencing techniques described herein may use one or more of ALUs embodied in 3D pipeline **2212**. Moreover, in at least one embodiment, inferencing and/or training operations described herein may be done using logic other than logic illustrated in FIG. **1**A or **1**B. In at least one embodiment, weight parameters may be stored in on-chip or off-chip memory and/or registers (shown or not shown) that configure ALUs of graphics processor **2200** to perform one or more machine learning algorithms, neural network architectures, use cases, or training techniques described herein.

[0363] FIG. **23** is a block diagram of a graphics processing engine **2310** of a graphics processor in accordance with at least one embodiment. In at least one embodiment, graphics processing engine (GPE) **2310** is a version of GPE **2210** shown in FIG. **22**. In at least one embodiment, a media pipeline **2316** is optional and may not be explicitly included within GPE **2310**. In at least one embodiment, a separate media and/or image processor is coupled to GPE **2310**.

[0364] In at least one embodiment, GPE **2310** is coupled to or includes a command streamer **2303**, which provides a command stream to a 3D pipeline **2312** and/or media pipeline **2316**. In at least one embodiment, command streamer **2303** is coupled to memory, which can be system memory, or one or more of internal cache memory and shared cache memory. In at least one embodiment, command streamer **2303** receives commands from memory and sends commands to 3D pipeline **2312** and/or media pipeline **2316**. In at least one embodiment, commands are instructions, primitives, or micro-operations fetched from a ring buffer, which stores commands for 3D pipeline **2312** and media pipeline **2316**. In at least one embodiment, a ring buffer can additionally include batch command buffers storing batches of multiple commands. In at least one embodiment, commands for 3D pipeline **2312** can also include references to data stored in memory, such as, but not limited to, vertex and geometry data for 3D pipeline **2312** and/or image data and memory objects for media pipeline **2316**. In at least one embodiment, 3D pipeline **2312** and media pipeline **2316** process commands and data by performing operations or by dispatching one or more execution threads to a graphics core array **2314**. In at least one embodiment, graphics core array **2314** includes one or more blocks of graphics cores (e.g., graphics core(s) **2315**A, graphics core (s) **2315**B), each block including one or more graphics cores. In at least one embodiment, each graphics core includes a set of graphics execution resources that includes general-purpose and graphics specific execution logic to perform graphics and compute operations, as well as fixed function texture processing and/or machine learning and artificial intelligence acceleration logic, including inference and/or training logic **115** in FIG. **1**A and FIG. **1**B.

[0365] In at least one embodiment, 3D pipeline **2312** includes fixed function and programmable logic to process one or more shader programs, such as vertex shaders, geometry shaders, pixel shaders, fragment shaders, compute shaders, or other shader programs, by processing instructions and dispatching execution threads to graphics core array **2314**. In at least one embodiment, graphics core array **2314** provides a unified block of execution resources for use in processing shader programs. In at least one embodiment, a multi-purpose execution logic (e.g., execution units) within graphics core(s) **2315**A-**2315**B of graphic core array **2314** includes support for various 3D API shader languages and can execute multiple simultaneous execution threads associated with multiple shaders.

[0366] In at least one embodiment, graphics core array **2314** also includes execution logic to perform media functions, such as video and/or image processing. In at least one embodiment, execution units additionally include general-purpose logic that is programmable to perform parallel general-purpose computational operations, in addition to graphics processing operations.

[0367] In at least one embodiment, output data generated by threads executing on graphics core array **2314** can output data to memory in a unified return buffer (URB) **2318**. In at least one embodiment, URB **2318** can store data for multiple threads. In at least one embodiment, URB **2318** may be used to send data between different threads executing on graphics core array **2314**. In at least one embodiment, URB **2318** may additionally be used for synchronization between threads on graphics core array **2314** and fixed function logic within shared function logic **2320**.

[0368] In at least one embodiment, graphics core array **2314** is scalable, such that graphics core array **2314** includes a variable number of graphics cores, each having a variable number of execution units based on a target power and performance level of GPE **2310**. In at least one embodiment, execution resources are dynamically scalable, such that execution resources may be enabled or disabled as needed.

[0369] In at least one embodiment, graphics core array **2314** is coupled to shared function logic **2320** that includes multiple resources that are shared between graphics cores in graphics core array **2314**. In at least one embodiment, shared functions performed by shared function logic **2320** are embodied in hardware logic units that provide specialized supplemental functionality to graphics core array **2314**. In at least one embodiment, shared function logic **2320** includes but is not limited to a sampler unit **2321**, a math unit **2322**, and inter-thread communication (ITC) logic **2323**. In at least one embodiment, one or more cache(s) **2325** are included in, or coupled to, shared function logic **2320**.

[0370] In at least one embodiment, a shared function is used if demand for a specialized function is insufficient for inclusion within graphics core array **2314**. In at least one embodiment, a single instantiation of a specialized function is used in shared function logic **2320** and shared among other execution resources within graphics core array **2314**. In at least one embodiment, specific shared functions within shared function logic **2320** that are used extensively by graphics core array **2314** may be included within shared function logic **2616** within graphics core array **2314**. In at least one embodiment, shared function logic **2616** within graphics core array **2314** can include some or all logic within shared function logic **2320**. In at least one embodiment, all logic elements within shared function logic **2320** may be duplicated within shared function logic **2326** of graphics core array **2314**. In at least one embodiment, shared function logic **2320** is excluded in favor of shared function logic **2326** within graphics core array **2314**.

[0371] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment portions or all of inference and/or training logic **115** may be incorporated into graphics processor **2310**. For example, in at least one embodiment, training and/or inferencing techniques described herein may use one or more of ALUs embodied in 3D pipeline **2312**, graphics core(s) **2315**, shared function logic **2326**, shared function logic **2320**, or other logic in FIG. **23**. Moreover, in at least one embodiment, inferencing and/or training operations described herein may be done using logic other than logic illustrated in FIG. **1**A or **1**B. In at least one embodiment, weight parameters may be stored in on-chip or off-chip memory and/or registers (shown or not shown) that configure ALUs of graphics processor **2310** to perform one or more machine learning algorithms, neural network architectures, use cases, or training techniques described herein.

[0372] FIG. **24** is a block diagram of hardware logic of a graphics processor core **2400**, according to at least one embodiment described herein. In at least one embodiment, graphics processor core **2400** is included within a graphics core array. In at least one embodiment, graphics processor core **2400**, sometimes referred to as a core slice, can be one or multiple graphics cores within a modular graphics processor. In at least one embodiment, graphics processor core **2400** is exemplary of one graphics core slice, and a graphics processor as described herein may include multiple graphics core slices based on target power and performance envelopes. In at least one embodiment, each graphics core **2400** can include a fixed function block **2430** coupled with multiple sub-cores **2401**A-**2401**F, also referred to as sub-slices, that include modular blocks of general-purpose and fixed function logic.

[0373] In at least one embodiment, fixed function block **2430** includes a geometry and fixed function pipeline **2436** that can be shared by all sub-cores in graphics processor **2400**, for example, in lower performance and/or lower power graphics processor implementations. In at least one embodiment, geometry and fixed function pipeline **2436** includes a 3D fixed function pipeline, a video front-end unit, a thread spawner and thread dispatcher, and a unified return buffer manager, which manages unified return buffers.

[0374] In at least one embodiment, fixed function block **2430** also includes a graphics SoC interface **2437**, a graphics microcontroller **2438**, and a media pipeline **2439**. In at least one embodiment, graphics SoC interface **2437** provides an interface between graphics core **2400** and other processor cores within a system on a chip integrated circuit. In at least one embodiment, graphics microcontroller **2438** is a programmable sub-processor that is configurable to manage various functions of graphics processor **2400**, including thread dispatch, scheduling, and pre-emption. In at least one embodiment, media pipeline **2439** includes logic to facilitate decoding, encoding, pre-processing, and/or post-processing of multimedia data, including image and video data. In at least one embodiment, media pipeline **2439** implements media operations via requests to compute or sampling logic within sub-cores **2401**A-**2401**F.

[0375] In at least one embodiment, SoC interface **2437** enables graphics core **2400** to communicate with general-purpose application processor cores (e.g., CPUs) and/or other components within an SoC, including memory hierarchy elements such as a shared last level cache memory, system RAM, and/or embedded on-chip or on-package DRAM. In at least one embodiment, SoC interface **2437** can also enable communication with fixed function devices within an SoC, such as camera imaging pipelines, and enables use of and/or implements global memory atomics that may be shared between graphics core **2400** and CPUs within an SoC. In at least one embodiment, graphics SoC interface **2437** can also implement power management controls for graphics processor core **2400** and enable an interface between a clock domain of graphics processor core **2400** and other clock domains within an SoC. In at least one embodiment, SoC interface **2437** enables receipt of command buffers from a command streamer and global thread dispatcher that are configured to provide commands and instructions to each of one or more graphics cores within a graphics processor. In at least one embodiment, commands and instructions can be dispatched to media pipeline **2439**, when media operations are to be performed, or a geometry and fixed function pipeline (e.g., geometry and fixed function pipeline **2436**, and/or a geometry and fixed function pipeline **2414**) when graphics processing operations are to be performed.

[0376] In at least one embodiment, graphics microcontroller **2438** can be configured to perform various scheduling and management tasks for graphics core **2400**. In at least one embodiment, graphics microcontroller **2438** can perform graphics and/or compute workload scheduling on various graphics parallel engines within execution unit (EU) arrays **2402**A-**2402**F, **2404**A-**2404**F within sub-cores **2401**A-**2401**F. In at least one embodiment, host software executing on a CPU core of an SoC including graphics core **2400** can submit workloads to one of multiple graphic processor paths, which invokes a scheduling operation on an appropriate graphics engine. In at least one embodiment, scheduling operations include determining which workload to run next, submitting a workload to a command streamer, pre-empting existing workloads running on an engine, monitoring progress of a workload, and notifying host software when a workload is complete. In at least one embodiment, graphics microcontroller **2438** can also facilitate low-power or idle states for graphics core **2400**, providing graphics core **2400** with an ability to save and restore registers within graphics core **2400** across low-power state transitions independently from an operating system and/or graphics driver software on a system.

[0377] In at least one embodiment, graphics core **2400** may have greater than or fewer than illustrated sub-cores **2401**A-**2401**F, up to N modular sub-cores. For each set of N sub-cores, in at least one embodiment, graphics core **2400** can also include shared function logic **2410**, shared and/or cache memory **2412**, geometry/fixed function pipeline **2414**, as well as additional fixed function logic **2416** to accelerate various graphics and compute processing operations. In at least one embodiment, shared function logic **2410** can include logic units (e.g., sampler, math, and/or inter-thread communication logic) that can be shared by each N sub-cores within graphics core **2400**. In at least one embodiment, shared and/or cache memory **2412** can be a last-level cache for N sub-cores **2401**A-**2401**F within graphics core **2400** and can also serve as shared memory that is accessible by multiple sub-cores. In at least one embodiment, geometry/fixed function pipeline **2414** can be included instead of geometry/fixed function pipeline **2436** within fixed function block **2430** and can include similar logic units.

[0378] In at least one embodiment, graphics core **2400** includes additional fixed function logic **2416** that can include various fixed function acceleration logic for use by graphics core **2400**. In at least one embodiment, additional fixed function logic **2416** includes an additional geometry pipeline for use in position-only shading. In position-only shading, at least two geometry pipelines exist, whereas in a full geometry pipeline within geometry and fixed function pipelines **2414**, **2436**, and a cull pipeline, which is an additional geometry pipeline that may be included within additional fixed function logic **2416**. In at least one embodiment, a cull pipeline is a trimmed down version of a full geometry pipeline. In at least one embodiment, a full pipeline and a cull pipeline can execute different instances of an application, each instance having a separate context. In at least one embodiment, position only shading can hide long cull runs of discarded triangles, enabling shading to be completed earlier in some instances. For example, in at least one embodiment, cull pipeline logic within additional fixed function logic **2416** can execute position shaders in parallel with a main application and generally generates critical results faster than a full pipeline, as a cull pipeline fetches and shades position attributes of vertices, without performing rasterization and rendering of pixels to a frame buffer. In at least one embodiment, a cull pipeline can use generated critical results to compute visibility information for all triangles without regard to whether those triangles are culled. In at least one embodiment, a full pipeline (which in this instance may be referred to as a replay pipeline) can consume visibility information to skip culled triangles to shade only visible triangles that are finally passed to a rasterization phase.

[0379] In at least one embodiment, additional fixed function logic **2416** can also include machine-learning acceleration logic, such as fixed function matrix multiplication logic, for implementations including optimizations for machine learning training or inferencing.

[0380] In at least one embodiment, within each graphics sub-core **2401**A-**2401**F includes a set of execution resources that may be used to perform graphics, media, and compute operations in response to requests by graphics pipeline, media pipeline, or shader programs. In at least one embodiment, graphics sub-cores **2401**A-**2401**F include multiple EU arrays **2402**A-**2402**F, **2404**A-**2404**F, thread dispatch and inter-thread communication (TD/IC) logic **2403**A-**2403**F, a 3D (e.g., texture) sampler **2405**A-**2405**F, a media sampler **2406**A-**2406**F, a shader processor **2407**A-**2407**F, and shared local memory (SLM) **2408**A-**2408**F. In at least one embodiment, EU arrays **2402**A-**2402**F, **2404**A-**2404**F each include multiple execution units, which are general-purpose graphics processing units capable of performing floating-point and integer/fixed-point logic operations in service of a graphics, media, or compute operation, including graphics, media, or compute shader programs. In at least one embodiment, TD/IC logic **2403**A-**2403**F performs local thread dispatch and thread control operations for execution units within a sub-core and facilitates communication between threads executing on execution units of a sub-core. In at least one embodiment, 3D samplers **2405**A-**2405**F can read texture or other 3D graphics related data into memory. In at least one embodiment, 3D samplers can read texture data differently based on a configured sample state and texture format associated with a given texture. In at least one embodiment, media samplers **2406**A-**2406**F can perform similar read operations based on a type and format associated with media data. In at least one embodiment, each graphics sub-core **2401**A-**2401**F can alternately include a unified 3D and media sampler. In at least one embodiment, threads executing on execution units within each of sub-cores **2401**A-**2401**F can make use of shared local memory **2408**A-**2408**F within each sub-core, to enable threads executing within a thread group to execute using a common pool of on-chip memory.

[0381] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, portions or all of inference and/or training logic **115** may be incorporated into graphics processor **2410**. For example, in at least one embodiment, training and/or inferencing techniques described herein may use one or more of ALUs embodied in a 3D pipeline, graphics microcontroller **2438**, geometry and fixed function pipeline **2414** and **2436**, or other logic in FIG. **24**. Moreover, in at least one embodiment, inferencing and/or training operations described herein may be done using logic other than logic illustrated in FIG. **1**A or **1**B. In at least one embodiment, weight parameters may be stored in on-chip or off-chip memory and/or registers (shown or not shown) that configure ALUs of graphics processor **2400** to perform one or more machine learning algorithms, neural network architectures, use cases, or training techniques described herein.

[0382] FIGS. **25**A-**25**B illustrate thread execution logic **2500** including an array of processing elements of a graphics processor core according to at least one embodiment. FIG. **25**A illustrates at least one embodiment, in which thread execution logic **2500** is used. FIG. **25**B illustrates exemplary internal details of a graphics execution unit **2508**, according to at least one embodiment.

[0383] As illustrated in FIG. **25**A, in at least one embodiment, thread execution logic **2500** includes a shader processor **2502**, a thread dispatcher **2504**, an instruction cache **2506**, a scalable execution unit array including a plurality of execution units **2507**A-**2507**N and **2508**A-**2508**N, a sampler **2510**, a data cache **2512**, and a data port **2514**. In at least one embodiment, a scalable execution unit array can dynamically scale by enabling or disabling one or more execution units (e.g., any of execution unit **2508**A-N or **2507**A-N) based on computational requirements of a workload, for example. In at least one embodiment, scalable execution units are interconnected via an interconnect fabric that links to each execution unit. In at least one embodiment, thread execution logic **2500** includes one or more connections to memory, such as system memory or cache memory, through one or more of instruction cache **2506**, data port **2514**, sampler **2510**, and execution units **2507** or **2508**. In at least one embodiment, each execution unit (e.g., **2507**A) is a stand-alone programmable general-purpose computational unit that is capable of executing multiple simultaneous hardware threads while processing multiple data elements in parallel for each thread. In at least one embodiment, array of execution units **2507** and/or **2508** is scalable to include any number individual execution units.

[0384] In at least one embodiment, execution units **2507** and/or **2508** are primarily used to execute shader programs. In at least one embodiment, shader processor **2502** can process various shader programs and dispatch execution threads associated with shader programs via a thread dispatcher **2504**. In at least one embodiment, thread dispatcher **2504** includes logic to arbitrate thread initiation requests from graphics and media pipelines and instantiate requested threads on one or more execution units in execution units **2507** and/or **2508**. For example, in at least one embodiment, a geometry pipeline can dispatch vertex, tessellation, or geometry shaders to thread execution logic for processing. In at least one embodiment, thread dispatcher **2504** can also process runtime thread spawning requests from executing shader programs.

[0385] In at least one embodiment, execution units **2507** and/or **2508** support an instruction set that includes native support for many standard 3D graphics shader instructions, such that shader programs from graphics libraries (e.g., Direct 3D and OpenGL) are executed with a minimal translation. In at least one embodiment, execution units support vertex and geometry processing (e.g., vertex programs, geometry programs, and/or vertex shaders), pixel processing (e.g., pixel shaders, fragment shaders) and general-purpose processing (e.g., compute and media shaders). In at least one embodiment, each of execution units **2507** and/or **2508**, which include one or more arithmetic logic units (ALUs), is capable of multi-issue single instruction multiple data (SIMD) execution and multi-threaded operation enables an efficient execution environment despite higher latency memory accesses. In at least one embodiment, each hardware thread within each execution unit has a dedicated high-bandwidth register file and associated independent thread-state. In at least one embodiment, execution is multi-issue per clock to pipelines capable of integer, single and double precision floating point operations, SIMD branch capability, logical operations, transcendental operations, and other miscellaneous operations. In at least one embodiment, while waiting for data from memory or one of shared functions, dependency logic within execution units **2507** and/or **2508** causes a waiting thread to sleep until requested data has been returned. In at least one embodiment, while an awaiting thread is sleeping, hardware resources may be devoted to processing other threads. For example, in at least one embodiment, during a delay associated with a vertex shader operation, an execution unit can perform operations for a pixel shader, fragment shader, or another type of shader program, including a different vertex shader.

[0386] In at least one embodiment, each execution unit in execution units **2507** and/or **2508** operates on arrays of data elements. In at least one embodiment, a number of data elements is an "execution size," or number of channels for an instruction. In at least one embodiment, an execution channel is a logical unit of execution for data element access, masking, and flow control within instructions. In at least one embodiment, a number of channels may be independent of a number of physical arithmetic logic units (ALUs) or floating point units (FPUs) for a particular graphics processor. In at least one embodiment, execution units **2507** and/or **2508** support integer and floating-point data types.

[0387] In at least one embodiment, an execution unit instruction set includes SIMD instructions. In at least one embodiment, various data elements can be stored as a packed data type in a register and execution unit will process various elements based on data size of elements. For example, in at least one embodiment, when operating on a 256-bit wide vector, 256 bits of a vector are stored in a register and an execution unit operates on a vector as four separate 64-bit packed data elements (Quad-Word (QW) size data elements), eight separate 32-bit packed data elements (Double Word (DW) size data elements), sixteen separate 16-bit packed data elements (Word (W) size data elements), or thirty-two separate 8-bit data elements (byte (B) size data elements). However, in at least one embodiment, different vector widths and register sizes are possible.

[0388] In at least one embodiment, one or more execution units can be combined into a fused execution unit **2509**A-**2509**N having thread control logic (**2511**A-**2511**N) that is common to fused EUs such as execution unit **2507**A fused with execution unit **2508**A into fused execution unit **2509**A. In at least one embodiment, multiple EUs can be fused into an EU group. In at least one embodiment, each EU in a fused EU group can be configured to execute a separate SIMD hardware thread, with a number of EUs in a fused EU group possibly varying according to various embodiments. In at least one embodiment, various SIMD widths can be performed per-EU, including but not limited to SIMD8, SIMD16, and SIMD32. In at least one embodiment, each fused graphics execution unit **2509**A-**2509**N includes at least two execution units. For example, in at least one embodiment, fused execution unit **2509**A includes a first EU **2507**A, second EU **2508**A, and thread control logic **2511**A that is common to first EU **2507**A and second EU **2508**A. In at least one embodiment, thread control logic **2511**A controls threads executed on fused graphics execution unit **2509**A, allowing each EU within fused execution units **2509**A-**2509**N to execute using a common instruction pointer register.

[0389] In at least one embodiment, one or more internal instruction caches (e.g., **2506**) are included in thread execution logic **2500** to cache thread instructions for execution units. In at least one embodiment, one or more data caches (e.g., **2512**) are included to cache thread data during thread execution. In at least one embodiment, sampler **2510** is included to provide texture sampling for 3D operations and media sampling for media operations. In at least one embodiment, sampler **2510** includes specialized texture or media sampling functionality to process texture or media data during sampling process before providing sampled data to an execution unit.

[0390] During execution, in at least one embodiment, graphics and media pipelines send thread initiation requests to thread execution logic **2500** via thread spawning and dispatch logic. In at least one embodiment, once a group of geometric objects has been processed and rasterized into pixel data, pixel processor logic (e.g., pixel shader logic, fragment shader logic, etc.) within shader processor **2502** is invoked to further compute output information and cause results to be written to output surfaces (e.g., color buffers, depth buffers, stencil buffers, etc.). In at least one embodiment, a pixel shader or a fragment shader calculates values of various vertex attributes that are to be interpolated across a rasterized object. In at least one embodiment, pixel processor logic within shader processor **2502** then executes an application programming interface (API)-supplied pixel or fragment shader program. In at least one embodiment, to execute a shader program, shader processor **2502** dispatches threads to an execution unit (e.g., **2508**A) via thread dispatcher **2504**. In at least one embodiment, shader processor **2502** uses texture sampling logic in sampler **2510** to access texture data in texture maps stored in memory. In at least one embodiment, arithmetic operations on texture data and input geometry data compute pixel color data for each geometric fragment, or discards one or more pixels from further processing.

[0391] In at least one embodiment, data port **2514** provides a memory access mechanism for thread execution logic **2500** to output processed data to memory for further processing on a graphics processor output pipeline. In at least one embodiment, data port **2514** includes or couples to one or more cache memories (e.g., data cache **2512**) to cache data for memory access via a data port.

[0392] As illustrated in FIG. **25**B, in at least one embodiment, a graphics execution unit **2508** can include an instruction fetch unit **2537**, a general register file array (GRF) **2524**, an architectural register file array (ARF) **2526**, a thread arbiter **2522**, a send unit **2530**, a branch unit **2532**, a set of SIMD floating point units (FPUs) **2534**, and a set of dedicated integer SIMD ALUs **2535**. In at least one embodiment, GRF **2524** and ARF **2526** includes a set of general register files and architecture register files associated with each simultaneous hardware thread that may be active in graphics execution unit **2508**. In at least one embodiment, per thread architectural state is maintained in ARF **2526**, while data used during thread execution is stored in GRF **2524**. In at least one embodiment, execution state of each thread, including instruction pointers for each thread, can be held in thread-specific registers in ARF **2526**.

[0393] In at least one embodiment, graphics execution unit **2508** has an architecture that is a combination of Simultaneous Multi-Threading (SMT) and fine-grained Interleaved Multi-Threading (IMT). In at least one embodiment, architecture has a modular configuration that can be fine-tuned at design time based on a target number of simultaneous threads and number of registers per execution unit, where execution unit resources are divided across logic used to execute multiple simultaneous threads.

[0394] In at least one embodiment, graphics execution unit **2508** can co-issue multiple instructions, which may each be different instructions. In at least one embodiment, thread arbiter **2522** of graphics execution unit thread **2508** can dispatch instructions to one of send unit **2530**, branch unit **2532**, or SIMD FPU(s) **2534** for execution. In at least one embodiment, each execution thread can access **128** general-purpose registers within GRF **2524**, where each register can store 32 bytes, accessible as a SIMD 8-element vector of 32-bit data elements. In at least one embodiment, each execution unit thread has access to 4 kilobytes within GRF **2524**, although embodiments are not so limited, and greater or fewer register resources may be provided in other embodiments. In at least one embodiment, up to seven threads can execute simultaneously, although a number of threads per execution unit can also vary according to embodiments. In at least one embodiment, in which seven threads may access 4 kilobytes, GRF **2524** can store a total of 28 kilobytes. In at least one embodiment, flexible addressing modes can permit registers to be addressed together to build effectively wider registers or to represent strided rectangular block data structures.

[0395] In at least one embodiment, memory operations, sampler operations, and other longer-latency system communications are dispatched via "send" instructions that are executed by message passing to send unit **2530**. In at least one embodiment, branch instructions are dispatched to branch unit **2532** to facilitate SIMD divergence and eventual convergence.

[0396] In at least one embodiment, graphics execution unit **2508** includes one or more SIMD floating point units (FPU(s)) **2534** to perform floating-point operations. In at least one embodiment, FPU(s) **2534** also support integer computation. In at least one embodiment, FPU(s) **2534** can SIMD execute up to M number of 32-bit floating-point (or integer) operations, or SIMD execute up to 2M 16-bit integer or 16-bit floating-point operations. In at least one embodiment, at least one FPU provides extended math capability to support high-throughput transcendental math functions and double precision 64-bit floating-point. In at least one embodiment, a set of 8-bit integer SIMD ALUs **2535** are also present, and may be specifically optimized to perform operations associated with machine learning computations.

[0397] In at least one embodiment, arrays of multiple instances of graphics execution unit **2508** can be instantiated in a graphics sub-core grouping (e.g., a sub-slice). In at least one embodiment, execution unit **2508** can execute instructions across a plurality of execution channels. In at least one embodiment, each thread executed on graphics execution unit **2508** is executed on a different channel.

[0398] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, portions or all of inference and/or training logic **115** may be incorporated into thread execution logic **2500**. Moreover, in at least one embodiment, inferencing and/or training operations described herein may be done using logic other than logic illustrated in FIG. **1**A or **1**B. In at least one embodiment, weight parameters may be stored in on-chip or off-chip memory and/or registers (shown or not shown) that configure ALUs thread of execution logic **2500** to perform one or more machine learning algorithms, neural network architectures, use cases, or training techniques described herein.

[0399] FIG. **26** illustrates a parallel processing unit ("PPU") **2600**, according to at least one embodiment. In at least one embodiment, PPU **2600** is configured with machine-readable code that, if executed by PPU **2600**, causes PPU **2600** to perform some or all of processes and techniques described throughout this disclosure. In at least one embodiment, PPU **2600** is a multi-threaded processor that is implemented on one or more integrated circuit devices and that utilizes multithreading as a latency-hiding technique designed to process computer-readable instructions (also referred to as machine-readable instructions or simply instructions) on multiple threads in parallel. In at least one embodiment, a thread refers to a thread of execution and is an instantiation of a set of instructions configured to be executed by PPU **2600**. In at least one embodiment, PPU **2600** is a graphics processing unit ("GPU") configured to implement a graphics rendering pipeline for processing three-dimensional ("3D") graphics data in order to generate two-dimensional ("2D") image data for display on a display device such as a liquid crystal display ("LCD") device. In at least one embodiment, PPU **2600** is utilized to perform computations such as linear algebra operations and machine-learning operations. FIG. **26** illustrates an example parallel processor for illustrative purposes only and should be construed as a non-limiting example of processor architectures contemplated within scope of this disclosure and that any suitable processor may be employed to supplement and/or substitute for same.

[0400] In at least one embodiment, one or more PPUs **2600** are configured to accelerate High Performance Computing ("HPC"), data center, and machine learning applications. In at least one embodiment, PPU **2600** is configured to accelerate deep learning systems and applications including following non-limiting examples: autonomous vehicle platforms, deep learning, high-accuracy speech, image, text recognition systems, intelligent video analytics, molecular simulations, drug discovery, disease diagnosis, weather forecasting, big data analytics, astronomy, molecular dynamics simulation, financial modeling, robotics, factory automation, real-time language translation, online search optimizations, and personalized user recommendations, and more.

[0401] In at least one embodiment, PPU **2600** includes, without limitation, an Input/Output ("I/O") unit **2606**, a front-end unit **2610**, a scheduler unit **2612**, a work distribution unit **2614**, a hub **2616**, a crossbar ("XBar") **2620**, one or more general processing clusters ("GPCs") **2618**, and one or more partition units ("memory partition units") **2622**. In at least one embodiment, PPU **2600** is connected to a host processor or other PPUs **2600** via one or more high-speed GPU interconnects ("GPU interconnects") **2608**. In at least one embodiment, PPU **2600** is connected to a host processor or other peripheral devices via a system bus **2602**. In at least one embodiment, PPU **2600** is connected to a local memory comprising one or more memory devices ("memory") **2604**. In at least one embodiment, memory devices **2604** include, without limitation, one or more dynamic random access memory ("DRAM") devices. In at least one embodiment, one or more DRAM devices are configured and/or configurable as high-bandwidth memory ("HBM") subsystems, with multiple DRAM dies stacked within each device.

[0402] In at least one embodiment, high-speed GPU interconnect **2608** may refer to a wire-based multi-lane communications link that is used by systems to scale and include one or more PPUs **2600** combined with one or more central processing units ("CPUs"), supports cache coherence between PPUs **2600** and CPUs, and CPU mastering. In at least one embodiment, data and/or commands are transmitted by high-speed GPU interconnect **2608** through hub **2616** to/from other units of PPU **2600** such as one or more copy engines, video encoders, video decoders, power management units, and other components which may not be explicitly illustrated in FIG. **26**.

[0403] In at least one embodiment, I/O unit **2606** is configured to transmit and receive communications (e.g., commands, data) from a host processor (not illustrated in FIG. **26**) over system bus **2602**. In at least one embodiment, I/O unit **2606** communicates with host processor directly via system bus **2602** or through one or more intermediate devices such as a memory bridge. In at least one embodiment, I/O unit **2606** may communicate with one or more other processors, such as one or more of PPUs **2600** via system bus **2602**. In at least one embodiment, I/O unit **2606** implements a Peripheral Component Interconnect Express ("PCIe") interface for communications over a PCIe bus. In at least one embodiment, I/O unit **2606** implements interfaces for communicating with external devices.

[0404] In at least one embodiment, I/O unit **2606** decodes packets received via system bus **2602**. In at least one embodiment, at least some packets represent commands configured to cause PPU **2600** to perform various operations. In at least one embodiment, I/O unit **2606** transmits decoded commands to various other units of PPU **2600** as specified by commands. In at least one embodiment, commands are transmitted to front-end unit **2610** and/or transmitted to hub **2616** or other units of PPU **2600** such as one or more copy engines, a video encoder, a video decoder, a power management unit, etc. (not explicitly illustrated in FIG. **26**). In at least one embodiment, I/O unit **2606** is configured to route communications between and among various logical units of PPU **2600**.

[0405] In at least one embodiment, a program executed by host processor encodes a command stream in a buffer that provides workloads to PPU **2600** for processing. In at least one embodiment, a workload comprises instructions and data to be processed by those instructions. In at least one embodiment, a buffer is a region in a memory that is accessible (e.g., read/write) by both a host processor and PPU **2600**—a host interface unit may be configured to access that buffer in a system memory connected to system bus **2602** via memory requests transmitted over system bus **2602** by I/O unit **2606**. In at least one embodiment, a host processor writes a command stream to a buffer and then transmits a pointer to a start of a command stream to PPU **2600** such that front-end unit **2610** receives pointers to one or more command streams and manages one or more command streams, reading commands from command streams and forwarding commands to various units of PPU **2600**.

[0406] In at least one embodiment, front-end unit **2610** is coupled to scheduler unit **2612** that configures various GPCs **2618** to process tasks defined by one or more command streams. In at least one embodiment, scheduler unit **2612** is configured to track state information related to various tasks managed by scheduler unit **2612** where state information may indicate which of GPCs **2618** a task is assigned to, whether task is active or inactive, a priority level associated with task, and so forth. In at least one embodiment, scheduler unit **2612** manages execution of a plurality of tasks on one or more of GPCs **2618**.

[0407] In at least one embodiment, scheduler unit **2612** is coupled to work distribution unit **2614** that is configured to dispatch tasks for execution on GPCs **2618**. In at least one embodiment, work distribution unit **2614** tracks a number of scheduled tasks received from scheduler unit **2612** and work distribution unit **2614** manages a pending task pool and an active task pool for each of GPCs **2618**. In at least one embodiment, pending task pool comprises a number of slots (e.g., 32 slots) that contain tasks assigned to be processed by a particular GPC **2618**; an active task pool may comprise a number of slots (e.g., 4 slots) for tasks that are actively being processed by GPCs **2618** such that as one of GPCs **2618** completes execution of a task, that task is evicted from that active task pool for GPC **2618** and another task from a pending task pool is selected and scheduled for execution on GPC **2618**. In at least one embodiment, if an active task is idle on GPC **2618**, such as while waiting for a data dependency to be resolved, then that active task is evicted from GPC **2618** and returned to that pending task pool while another task in that pending task pool is selected and scheduled for execution on GPC **2618**.

[0408] In at least one embodiment, work distribution unit **2614** communicates with one or more GPCs **2618** via XBar **2620**. In at least one embodiment, XBar **2620** is an interconnect network that couples many of units of PPU **2600** to other units of PPU **2600** and can be configured to couple work distribution unit **2614** to a particular GPC **2618**. In at least one embodiment, one or more other units of PPU **2600** may also be connected to XBar **2620** via hub **2616**.

[0409] In at least one embodiment, tasks are managed by scheduler unit **2612** and dispatched to one of GPCs **2618** by work distribution unit **2614**. In at least one embodiment, GPC **2618** is configured to process task and generate results. In at least one embodiment, results may be consumed by other tasks within GPC **2618**, routed to a different GPC **2618** via XBar **2620**, or stored in memory **2604**. In at least one embodiment, results can be written to memory **2604** via partition units **2622**, which implement a memory interface for reading and writing data to/from memory **2604**. In at least one embodiment, results can be transmitted to another PPU **2604** or CPU via high-speed GPU interconnect **2608**. In at least one embodiment, PPU **2600** includes, without limitation, a number U of partition units **2622** that is equal to a number of separate and distinct memory devices **2604** coupled to PPU **2600**, as described in more detail herein in conjunction with FIG. **28**.

[0410] In at least one embodiment, a host processor executes a driver kernel that implements an application programming interface ("API") that enables one or more applications executing on a host processor to schedule operations for execution on PPU **2600**. In at least one embodiment, multiple compute applications are simultaneously executed by PPU **2600** and PPU **2600** provides isolation, quality of service ("QoS"), and independent address spaces for multiple compute applications. In at least one embodiment, an application generates instructions (e.g., in form of API calls) that cause a driver kernel to generate one or more tasks for execution by PPU **2600** and that driver kernel outputs tasks to one or more streams being processed by PPU **2600**. In at least one embodiment, each task comprises one or more groups of related threads, which may be referred to as a warp. In at least one embodiment, a warp comprises a plurality of related threads (e.g., 32 threads) that can be executed in parallel. In at least one embodiment, cooperating threads can refer to a plurality of threads including instructions to perform task and that exchange data through shared memory. In at least one embodiment, threads and cooperating threads are described in more detail in conjunction with FIG. **28**.

[0411] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, deep learning application processor is used to train a machine learning model, such as a neural network, to predict or infer information provided to PPU **2600**. In at least one embodiment, deep learning application processor **2600** is used to infer or predict information based on a trained machine learning model (e.g., neural network) that has been trained by another processor or system or by PPU **2600**. In at least one embodiment, PPU **2600** may be used to perform one or more neural network use cases described herein.

[0412] FIG. **27** illustrates a general processing cluster ("GPC") **2700**, according to at least one embodiment. In at least one embodiment, GPC **2700** is GPC **2618** of FIG. **26**. In at least one embodiment, each GPC **2700** includes, without limitation, a number of hardware units for processing tasks and each GPC **2700** includes, without limitation, a pipeline manager **2702**, a pre-raster operations unit ("preROP") **2704**, a raster engine **2708**, a work distribution crossbar ("WDX") **2716**, a memory management unit ("MMU") **2718**, one or more Data Processing Clusters ("DPCs") **2706**, and any suitable combination of parts.

[0413] In at least one embodiment, operation of GPC **2700** is controlled by pipeline manager **2702**. In at least one embodiment, pipeline manager **2702** manages configuration of one or more DPCs **2706** for processing tasks allocated to GPC **2700**. In at least one embodiment, pipeline manager **2702** configures at least one of one or more DPCs **2706** to implement at least a portion of a graphics rendering pipeline. In at least one embodiment, DPC **2706** is configured to execute a vertex shader program on a programmable streaming multi-processor ("SM") **2714**. In at least one embodiment, pipeline manager **2702** is configured to route packets received from a work distribution unit to appropriate logical units within GPC **2700**, in at least one embodiment, and some packets may be routed to fixed function hardware units in preROP **2704** and/or raster engine **2708** while other packets may be routed to DPCs **2706** for processing by a primitive engine **2712** or SM **2714**. In at least one embodiment, pipeline manager **2702** configures at least one of DPCs **2706** to implement a neural network model and/or a computing pipeline.

[0414] In at least one embodiment, preROP unit **2704** is configured, in at least one embodiment, to route data generated by raster engine **2708** and DPCs **2706** to a Raster Operations ("ROP") unit in partition unit **2622**, described in more detail above in conjunction with FIG. **26**. In at least one embodiment, preROP unit **2704** is configured to perform optimizations for color blending, organize pixel data, perform address translations, and more. In at least one embodiment, raster engine **2708** includes, without limitation, a number of fixed function hardware units configured to perform various raster operations, in at least one embodiment, and raster engine **2708** includes, without limitation, a setup engine, a coarse raster engine, a culling engine, a clipping engine, a fine raster engine, a tile coalescing engine, and any suitable combination thereof. In at least one embodiment, setup engine receives transformed vertices and generates plane equations associated with geometric primitive defined by vertices; plane equations are transmitted to a coarse raster engine to generate coverage information (e.g., an x, y coverage mask for a tile) for primitive; output of a coarse raster engine is transmitted to a culling engine where fragments associated with a primitive that fail a z-test are culled, and transmitted to a clipping engine where fragments lying outside a viewing frustum are clipped. In at least one embodiment, fragments that survive clipping and culling are passed to a fine raster engine to generate attributes for pixel fragments based on plane equations generated by a setup engine. In at least one embodiment, an output of raster engine **2708** comprises fragments to be processed by any suitable entity, such as by a fragment shader implemented within DPC **2706**.

[0415] In at least one embodiment, each DPC **2706** included in GPC **2700** comprises, without limitation, an M-Pipe Controller ("MPC") **2710**; primitive engine **2712**; one or more SMs **2714**; and any suitable combination thereof. In at least one embodiment, MPC **2710** controls operation of DPC **2706**, routing packets received from pipeline manager **2702** to appropriate units in DPC **2706**. In at least one embodiment, packets associated with a vertex are routed to primitive engine **2712**, which is configured to fetch vertex attributes associated with a vertex from memory; in contrast, packets associated with a shader program may be transmitted to SM **2714**.

[0416] In at least one embodiment, SM **2714** comprises, without limitation, a programmable streaming processor that is configured to process tasks represented by a number of threads. In at least one embodiment, SM **2714** is multi-threaded and configured to execute a plurality of threads (e.g., 32 threads) from a particular group of threads concurrently and implements a Single-Instruction, Multiple-Data ("SIMD") architecture where each thread in a group of threads (e.g., a warp) is configured to process a different set of data based on same set of instructions. In at least one embodiment, all threads in group of threads execute a common set of instructions. In at least one embodiment, SM **2714** implements a Single-Instruction, Multiple Thread ("SIMT") architecture wherein each thread in a group of threads is configured to process a different set of data based on that common set of instructions, but where individual threads in a group of threads are allowed to diverge during execution. In at least one embodiment, a program counter, call stack, and execution state is maintained for each warp, enabling concurrency between warps and serial execution within warps when threads within a warp diverge. In another embodiment, a program counter, call stack, and execution state is maintained for each individual thread, enabling equal concurrency between all threads, within and between warps. In at least one embodiment, execution state is maintained for each individual thread and threads executing common instructions may be converged and executed in parallel for better efficiency. At least one embodiment of SM **2714** is described in more detail herein.

[0417] In at least one embodiment, MMU **2718** provides an interface between GPC **2700** and a memory partition unit (e.g., partition unit **2622** of FIG. **26**) and MMU **2718** provides translation of virtual addresses into physical addresses, memory protection, and arbitration of memory requests. In at least one embodiment, MMU **2718** provides one or more translation lookaside buffers ("TLBs") for performing translation of virtual addresses into physical addresses in memory.

[0418] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, deep learning application processor is used to train a machine learning model, such as a neural network, to predict or infer information provided to GPC **2700**. In at least one embodiment, GPC **2700** is used to infer or predict information based on a trained machine learning model (e.g., neural network) that has been trained by another processor or system or by GPC **2700**. In at least one embodiment, GPC **2700** may be used to perform one or more neural network use cases described herein.

[0419] FIG. **28** illustrates a memory partition unit **2800** of a parallel processing unit ("PPU"), in accordance with at least one embodiment. In at least one embodiment, memory partition unit **2800** includes, without limitation, a Raster Operations ("ROP") unit **2802**, a level two ("L2") cache **2804**, a memory interface **2806**, and any suitable combination thereof. In at least one embodiment, memory interface **2806** is coupled to memory. In at least one embodiment, memory interface **2806** may implement 32, 64, 128, 1024-bit data buses, or like, for high-speed data transfer. In at least one embodiment, PPU incorporates U memory interfaces **2806** where U is a positive integer, with one memory interface **2806** per pair of partition units **2800**, where each pair of partition units **2800** is connected to a corresponding memory device. For example, in at least one embodiment, PPU may be connected to up to Y memory devices, such as high bandwidth memory stacks or graphics double-data-rate, version 5, synchronous dynamic random access memory ("GDDR5 SDRAM").

[0420] In at least one embodiment, memory interface **2806** implements a high bandwidth memory second generation ("HBM2") memory interface and Y equals half of U. In at least one embodiment, HBM2 memory stacks are located on a physical package with a PPU, providing substantial power and area savings compared with conventional GDDR5 SDRAM systems. In at least one embodiment, each HBM2 stack includes, without limitation, four memory dies with Y=4, with each HBM2 stack including two 128-bit channels per die for a total of 8 channels and a data bus width of 1024 bits. In at least one embodiment, that memory supports Single-Error Correcting Double-Error Detecting ("SECDED") Error Correction Code ("ECC") to protect data. In at least one embodiment, ECC can provide higher reliability for compute applications that are sensitive to data corruption.

[0421] In at least one embodiment, PPU implements a multi-level memory hierarchy. In at least one embodiment, memory partition unit **2800** supports a unified memory to provide a single unified virtual address space for central processing unit ("CPU") and PPU memory, enabling data sharing between virtual memory systems. In at least one embodiment frequency of accesses by a PPU to a memory located on other processors is traced to ensure that memory pages are moved to physical memory of PPU that is accessing pages more frequently. In at least one embodiment, high-speed GPU interconnect **2608** supports address translation services allowing PPU to directly access a CPU's page tables and providing full access to CPU memory by a PPU.

[0422] In at least one embodiment, copy engines transfer data between multiple PPUs or between PPUs and CPUs. In at least one embodiment, copy engines can generate page faults for addresses that are not mapped into page tables and memory partition unit **2800** then services page faults, mapping addresses into page table, after which copy engine performs a transfer. In at least one embodiment, memory is pinned (i.e., non-pageable) for multiple copy engine operations between multiple processors, substantially reducing available memory. In at least one embodiment, with hardware page faulting, addresses can be passed to copy engines without regard as to whether memory pages are resident, and a copy process is transparent.

[0423] Data from memory **2604** of FIG. **26** or other system memory is fetched by memory partition unit **2800** and stored in L2 cache **2804**, which is located on-chip and is shared between various GPCs, in accordance with at least one embodiment. Each memory partition unit **2800**, in at least one embodiment, includes, without limitation, at least a portion of L2 cache associated with a corresponding memory device. In at least one embodiment, lower level caches are implemented in various units within GPCs. In at least one embodiment, each of SMs **2714** in FIG. **27** may implement a Level 1 ("L1") cache wherein that L1 cache is private memory that is dedicated to a particular SM **2714** and data from L2 cache **2804** is fetched and stored in each L1 cache for processing in functional units of SMs **2714**. In at least one embodiment, L2 cache **2804** is coupled to memory interface **2806** and XBar **2620** shown in FIG. **26**.

[0424] ROP unit **2802** performs graphics raster operations related to pixel color, such as color compression, pixel blending, and more, in at least one embodiment. ROP unit **2802**, in at least one embodiment, implements depth testing in conjunction with raster engine **2708**, receiving a depth for a sample location associated with a pixel fragment from a culling engine of raster engine **2708**. In at least one embodiment, depth is tested against a corresponding depth in a depth buffer for a sample location associated with a fragment. In at least one embodiment, if that fragment passes that depth test for that sample location, then ROP unit **2802** updates depth buffer and transmits a result of that depth test to raster engine **2708**. It will be appreciated that a number of partition units **2800** may be different than a number of GPCs and, therefore, each ROP unit **2802** can, in at least one embodiment, be coupled to each GPC. In at least one embodiment, ROP unit **2802** tracks packets received from different GPCs and determines whether a result generated by ROP unit **2802** is to be routed to through XBar **2620**.

[0425] FIG. **29** illustrates a streaming multi-processor ("SM") **2900**, according to at least one embodiment. In at least one embodiment, SM **2900** is SM of FIG. **27**. In at least one embodiment, SM **2900** includes, without limitation, an instruction cache **2902**, one or more scheduler units **2904**, a register file **2908**, one or more processing cores ("cores") **2910**, one or more special function units ("SFUs") **2912**, one or more load/store units ("LSUs") **2914**, an interconnect network **2916**, a shared memory/level one ("L1") cache **2918**, and/or any suitable combination thereof.

[0426] In at least one embodiment, a work distribution unit dispatches tasks for execution on general processing clusters ("GPCs") of parallel processing units ("PPUs") and each task is allocated to a particular Data Processing Cluster ("DPC") within a GPC and, if a task is associated with a shader program, that task is allocated to one of SMs **2900**. In at least one embodiment, scheduler unit **2904** receives tasks from a work distribution unit and manages instruction scheduling for one or more thread blocks assigned to SM **2900**. In at least one embodiment, scheduler unit **2904** schedules thread blocks for execution as warps of parallel threads, wherein each thread block is allocated at least one warp. In at least one embodiment, each warp executes threads. In at least one embodiment, scheduler unit **2904** manages a plurality of different thread blocks, allocating warps to different thread blocks and then dispatching instructions from plurality of different cooperative groups to various functional units (e.g., processing cores **2910**, SFUs **2912**, and LSUs **2914**) during each clock cycle.

[0427] In at least one embodiment, Cooperative Groups may refer to a programming model for organizing groups of communicating threads that allows developers to express granularity at which threads are communicating, enabling expression of richer, more efficient parallel decompositions. In at least one embodiment, cooperative launch APIs support synchronization amongst thread blocks for execution of parallel algorithms. In at least one embodiment, applications of conventional programming models provide a single, simple construct for synchronizing cooperating threads: a barrier across all threads of a thread block (e.g., syncthreads( ) function). However, in at least one embodiment, programmers may define groups of threads at smaller than thread block granularities and synchronize within defined groups to enable greater performance, design flexibility, and software reuse in form of collective group-wide function interfaces. In at least one embodiment, Cooperative Groups enables programmers to define groups of threads explicitly at sub-block (i.e., as small as a single thread) and multi-block granularities, and to perform collective operations such as synchronization on threads in a cooperative group. In at least one embodiment, that programming model supports clean composition across software boundaries, so that libraries and utility functions can synchronize safely within their local context without having to make assumptions about convergence. In at least one embodiment, Cooperative Groups primitives enable new patterns of cooperative parallelism, including, without limitation, producer-consumer parallelism, opportunistic parallelism, and global synchronization across an entire grid of thread blocks.

[0428] In at least one embodiment, a dispatch unit **2906** is configured to transmit instructions to one or more functional units and scheduler unit **2904** and includes, without limitation, two dispatch units **2906** that enable two different instructions from a common warp to be dispatched during each clock cycle. In at least one embodiment, each scheduler unit **2904** includes a single dispatch unit **2906** or additional dispatch units **2906**.

[0429] In at least one embodiment, each SM **2900**, in at least one embodiment, includes, without limitation, register file **2908** that provides a set of registers for functional units of SM **2900**. In at least one embodiment, register file **2908** is divided between each functional unit such that each functional unit is allocated a dedicated portion of register file **2908**. In at least one embodiment, register file **2908** is divided between different warps being executed by SM **2900** and register file **2908** provides temporary storage for operands connected to data paths of functional units. In at least one embodiment, each SM **2900** comprises, without limitation, a plurality of L processing cores **2910**, where L is a positive integer. In at least one embodiment, SM **2900** includes, without limitation, a large number (e.g., 128 or more) of distinct processing cores **2910**. In at least one embodiment, each processing core **2910** includes, without limitation, a fully-pipelined, single-precision, double-precision, and/or mixed precision processing unit that includes, without limitation, a floating point arithmetic logic unit and an integer arithmetic logic unit. In at least one embodiment, floating point arithmetic logic units implement IEEE 754-2008 standard for floating point arithmetic. In at least one embodiment, processing cores **2910** include, without limitation, 64 single-precision (32-bit) floating point cores, 64 integer cores, 32 double-precision (64-bit) floating point cores, and 8 tensor cores.

[0430] Tensor cores are configured to perform matrix operations in accordance with at least one embodiment. In at least one embodiment, one or more tensor cores are included in processing cores **2910**. In at least one embodiment, tensor cores are configured to perform deep learning matrix arithmetic, such as convolution operations for neural network training and inferencing. In at least one embodiment, each tensor core operates on a 4×4 matrix and performs a matrix multiply and accumulate operation, $D=A\times B+C$, where A, B, C, and D are 4×4 matrices.

[0431] In at least one embodiment, matrix multiply inputs A and B are 16-bit floating point matrices and accumulation matrices C and D are16-bit floating point or 32-bit floating point matrices. In at least one embodiment, tensor cores operate on 16-bit floating point input data with 32-bit floating point accumulation. In at least one embodiment, 16-bit floating point multiply uses 64 operations and results in a full precision product that is then accumulated using 32-bit floating point addition with other intermediate products for a 4×4×4 matrix multiply. Tensor cores are used to perform much larger two-dimensional or higher dimensional matrix operations, built up from these smaller elements, in at least one embodiment. In at least one embodiment, an API, such as a CUDA 9 C++ API, exposes specialized matrix load, matrix multiply and accumulate, and matrix store operations to efficiently use tensor cores from a CUDA-C++ program. In at least one embodiment, at a CUDA level, a warp-level interface assumes 16×16 size matrices spanning all 32 threads of warp.

[0432] In at least one embodiment, each SM **2900** comprises, without limitation, M SFUs **2912** that perform special functions (e.g., attribute evaluation, reciprocal square root, and like). In at least one embodiment, SFUs **2912** include, without limitation, a tree traversal unit configured to traverse a hierarchical tree data structure. In at least one embodiment, SFUs **2912** include, without limitation, a texture unit configured to perform texture map filtering operations. In at least one embodiment, texture units are configured to load texture maps (e.g., a 2D array of texels) from memory and sample texture maps to produce sampled texture values for use in shader programs executed by SM **2900**. In at least one embodiment, texture maps are stored in shared memory/L1 cache **2918**. In at least one embodiment, texture units implement texture operations such as filtering operations using mip-maps (e.g., texture maps of varying levels of detail), in accordance with at least one embodiment. In at least one embodiment, each SM **2900** includes, without limitation, two texture units.

[0433] Each SM **2900** comprises, without limitation, N LSUs **2914** that implement load and store operations between shared memory/L1 cache **2918** and register file **2908**, in at least one embodiment. Interconnect network **2916** connects each functional unit to register file **2908** and LSU **2914** to register file **2908** and shared memory/L1 cache **2918** in at least one embodiment. In at least one embodiment, interconnect network **2916** is a crossbar that can be configured to connect any functional units to any registers in register file **2908** and connect LSUs **2914** to register file **2908** and memory locations in shared memory/L1 cache **2918**.

[0434] In at least one embodiment, shared memory/L1 cache **2918** is an array of on-chip memory that allows for data storage and communication between SM **2900** and primitive engine and between threads in SM **2900**, in at least one embodiment. In at least one embodiment, shared memory/L1 cache **2918** comprises, without limitation, 128 KB of storage capacity and is in a path from SM **2900** to a partition unit. In at least one embodiment, shared memory/L1 cache **2918**, in at least one embodiment, is used to cache reads and writes. In at least one embodiment, one or more of shared memory/L1 cache **2918**, L2 cache, and memory are backing stores.

[0435] Combining data cache and shared memory functionality into a single memory block provides improved performance for both types of memory accesses, in at least one embodiment. In at least one embodiment, capacity is used or is usable as a cache by programs that do not use shared memory, such as if shared memory is configured to use half of a capacity, and texture and load/store operations can use remaining capacity. Integration within shared memory/L1 cache **2918** enables shared memory/L1 cache **2918** to function as a high-throughput conduit for streaming data while simultaneously providing high-bandwidth and low-latency access to frequently reused data, in accordance with at least one embodiment. In at least one embodiment, when configured for general purpose parallel computation, a simpler configuration can be used compared with graphics processing. In at least one embodiment, fixed function graphics processing units are bypassed, creating a much simpler programming model. In a general purpose parallel computation configuration, a work distribution unit assigns and distributes blocks of threads directly to DPCs, in at least one embodiment. In at least one embodiment, threads in a block execute a common program, using a unique thread ID in calculation to ensure each thread generates unique results, using SM **2900** to execute program and perform calculations, shared memory/L1 cache **2918** to communicate between threads, and LSU **2914** to read and write global memory through shared memory/L1 cache **2918** and memory partition unit. In at least one embodiment, when configured for general purpose parallel computation, SM **2900** writes commands that scheduler unit **2904** can use to launch new work on DPCs.

[0436] In at least one embodiment, a PPU is included in or coupled to a desktop computer, a laptop computer, a tablet computer, servers, supercomputers, a smart-phone (e.g., a wireless, hand-held device), personal digital assistant ("PDA"), a digital camera, a vehicle, a head mounted display, a hand-held electronic device, and more. In at least one embodiment, a PPU is embodied on a single semiconductor substrate. In at least one embodiment, a PPU is included in a system-on-a-chip ("SoC") along with one or more other devices such as additional PPUs, memory, a reduced instruction set computer ("RISC") CPU, a memory management unit ("MMU"), a digital-to-analog converter ("DAC"), and like.

[0437] In at least one embodiment, a PPU may be included on a graphics card that includes one or more memory devices. In at least one embodiment, that graphics card may be configured to interface with a PCIe slot on a motherboard of a desktop computer. In at least one embodiment, that PPU may be an integrated graphics processing unit ("iGPU") included in chipset of a motherboard.

[0438] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B. In at least one embodiment, deep learning application processor is used to train a machine learning model, such as a neural network, to predict or infer information provided to SM **2900**. In at least one embodiment, SM **2900** is used to infer or predict information based on a trained machine learning model (e.g., neural network) that has been trained by another processor or system or by SM **2900**. In at least one embodiment, SM **2900** may be used to perform one or more neural network use cases described herein.

[0439] Embodiments are disclosed related a virtualized computing platform for advanced computing, such as image inferencing and image processing in medical applications. Without limitation, embodiments may include radiography, magnetic resonance imaging (MRI), nuclear medicine, ultrasound, sonography, elastography, photoacoustic imaging, tomography, echocardiography, functional near-infrared spectroscopy, and magnetic particle imaging, or a combination thereof. In at least one embodiment, a virtualized computing platform and associated processes described herein may additionally or alternatively be used, without limitation, in forensic science analysis, sub-surface detection and imaging (e.g., oil exploration, archaeology, paleontology, etc.), topography, oceanography, geology, osteology, meteorology, intelligent area or object tracking and monitoring, sensor data processing (e.g., RADAR, SONAR, LIDAR, etc.), and/or genomics and gene sequencing.

[0440] With reference to FIG. **30**, FIG. **30** is an example data flow diagram for a process **3000** of generating and deploying an image processing and inferencing pipeline, in accordance with at least one embodiment. In at least one embodiment, process **3000** may be deployed for use with imaging devices, processing devices, genomics devices, gene sequencing devices, radiology devices, and/or other device types at one or more facilities **3002**, such as medical facilities, hospitals, healthcare institutes, clinics, research or diagnostic labs, etc. In at least one embodiment, process **3000** may be deployed to perform genomics analysis and inferencing on sequencing data. Examples of genomic analyses that may be performed using systems and processes described herein include, without limitation, variant calling, mutation detection, and gene expression quantification.

[0441] In at least one embodiment, process **3000** may be executed within a training system **3004** and/or a deployment system **3006**. In at least one embodiment, training system **3004** may be used to perform training, deployment, and implementation of machine learning models (e.g., neural networks, object detection algorithms, computer vision algorithms, etc.) for use in deployment system **3006**. In at least one embodiment, deployment system **3006** may be configured to offload processing and compute resources among a distributed computing environment to reduce infrastructure requirements at facility **3002**. In at least one embodiment, deployment system **3006** may provide a streamlined platform for selecting, customizing, and implementing virtual instruments for use with imaging devices (e.g., MRI, CT Scan, X-Ray, Ultrasound, etc.) or sequencing devices at facility **3002**. In at least one embodiment, virtual instruments may include software-defined applications for performing one or more processing operations with respect to imaging data generated by imaging devices, sequencing devices, radiology devices, and/or other device types. In at least one embodiment, one or more applications in a pipeline may use or call upon services (e.g., inference, visualization, compute, AI, etc.) of deployment system **3006** during execution of applications.

[0442] In at least one embodiment, some of applications used in advanced processing and inferencing pipelines may use machine learning models or other AI to perform one or more processing steps. In at least one embodiment, machine learning models may be trained at facility **3002** using data **3008** (such as imaging data) generated at facility **3002** (and stored on one or more picture archiving and communication system (PACS) servers at facility **3002**), may be trained using imaging or sequencing data **3008** from another facility or facilities (e.g., a different hospital, lab, clinic, etc.), or a combination thereof. In at least one embodiment, training system **3004** may be used to provide applications, services, and/or other resources for generating working, deployable machine learning models for deployment system **3006**.

[0443] In at least one embodiment, a model registry **3024** may be backed by object storage that may support versioning and object metadata. In at least one embodiment, object storage may be accessible through, for example, a cloud storage (e.g., a cloud **3126** of FIG. **31**) compatible application programming interface (API) from within a cloud platform. In at least one embodiment, machine learning models within model registry **3024** may uploaded, listed, modified, or deleted by developers or partners of a system interacting with an API. In at least one embodiment, an API may provide access to methods that allow users with appropriate credentials to associate models with applications, such that models may be executed as part of execution of containerized instantiations of applications.

[0444] In at least one embodiment, a training pipeline **3104** (FIG. **31**) may include a scenario where facility **3002** is training their own machine learning model, or has an existing machine learning model that needs to be optimized or updated. In at least one embodiment, imaging data **3008** generated by imaging device(s), sequencing devices, and/or other device types may be received. In at least one embodiment, once imaging data **3008** is received, AI-assisted annotation **3010** may be used to aid in generating annotations corresponding to imaging data **3008** to be used as ground truth data for a machine learning model. In at least one embodiment, AI-assisted annotation **3010** may include one or more machine learning models (e.g., convolutional neural networks (CNNs)) that may be trained to generate annotations corresponding to certain types of imaging data **3008** (e.g., from certain devices) and/or certain types of anomalies in imaging data **3008**. In at least one embodiment, AI-assisted annotations **3010** may then be used directly, or may be adjusted or fine-tuned using an annotation tool (e.g., by a researcher, a clinician, a doctor, a scientist, etc.), to generate ground truth data. In at least one embodiment, in some examples, labeled clinic data **3012** (e.g., annotations provided by a clinician, doctor, scientist, technician, etc.) may be used as ground truth data for training a machine learning model. In at least one embodiment, AI-assisted annotations **3010**, labeled clinic data **3012**, or a combination thereof may be used as ground truth data for training a machine learning model. In at least one embodiment, a trained machine learning model may be referred to as an output model **3016**, and may be used by deployment system **3006**, as described herein.

[0445] In at least one embodiment, training pipeline **3104** (FIG. **31**) may include a scenario where facility **3002** needs a machine learning model for use in performing one or more processing tasks for one or more applications in deployment system **3006**, but facility **3002** may not currently have such a machine learning model (or may not have a model that is optimized, efficient, or effective for such purposes). In at least one embodiment, an existing machine learning model may be selected from model registry **3024**. In at least one embodiment, model registry **3024** may include machine learning models trained to perform a variety of different inference tasks on imaging data. In at least one embodiment, machine learning models in model registry **3024** may have been trained on imaging data from different facilities than facility **3002** (e.g., facilities remotely located). In at least one embodiment, machine learning models may have been trained on imaging data from one location, two locations, or any number of locations. In at least one embodiment, when being trained on imaging data from a specific location, training may take place at that location, or at least in a manner that protects confidentiality of imaging data or restricts imaging data from being transferred off-premises (e.g., to comply with HIPAA regulations, privacy regulations, etc.). In at least one embodiment, once a model is trained—or partially trained—at one location, a machine learning model may be added to model registry **3024**. In at least one embodiment, a machine learning model may then be retrained, or updated, at any number of other facilities, and a retrained or updated model may be made available in model registry **3024**. In at least one embodiment, a machine learning model may then be selected from model registry **3024**—and referred to as output model **3016**—and may be used in deployment system **3006** to perform one or more processing tasks for one or more applications of a deployment system.

[0446] In at least one embodiment, training pipeline **3104** (FIG. **31**) may be used in a scenario that includes facility **3002** requiring a machine learning model for use in performing one or more processing tasks for one or more applications in deployment system **3006**, but facility **3002** may not currently have such a machine learning model (or may not have a model that is optimized, efficient, or effective for such purposes). In at least one embodiment, a machine learning model selected from model registry **3024** might not be fine-tuned or optimized for imaging data **3008** generated at facility **3002** because of differences in populations, genetic variations, robustness of training data used to train a machine learning model, diversity in anomalies of training data, and/or other issues with training data. In at least one embodiment, AI-assisted annotation **3010** may be used to aid in generating annotations corresponding to imaging data **3008** to be used as ground truth data for retraining or updating a machine learning model. In at least one embodiment, labeled clinic data **3012** (e.g., annotations provided by a clinician, doctor, scientist, etc.) may be used as ground truth data for training a machine learning model. In at least one embodiment, retraining or updating a machine learning model may be referred to as model training **3014**. In at least one embodiment, model training **3014**—e.g., AI-assisted annotations **3010**, labeled clinic data **3012**, or a combination thereof—may be used as ground truth data for retraining or updating a machine learning model.

[0447] In at least one embodiment, deployment system **3006** may include software **3018**, services **3020**, hardware **3022**, and/or other components, features, and functionality. In at least one embodiment, deployment system **3006** may include a software "stack," such that software **3018** may be built on top of services **3020** and may use services **3020** to perform some or all of processing tasks, and services **3020** and software **3018** may be built on top of hardware **3022** and use hardware **3022** to execute processing, storage, and/or other compute tasks of deployment system **3006**.

[0448] In at least one embodiment, software **3018** may include any number of different containers, where each container may execute an instantiation of an application. In at least one embodiment, each application may perform one or more processing tasks in an advanced processing and inferencing pipeline (e.g., inferencing, object detection, feature detection, segmentation, image enhancement, calibration, etc.). In at least one embodiment, for each type of imaging device (e.g., CT, MRI, X-Ray, ultrasound, sonography, echocardiography, etc.), sequencing device, radiology device, genomics device, etc., there may be any number of containers that may perform a data processing task with respect to imaging data **3008** (or other data types, such as those described herein) generated by a device. In at least one embodiment, an advanced processing and inferencing pipeline may be defined based on selections of different containers that are desired or required for processing imaging data **3008**, in addition to containers that receive and configure imaging data for use by each container and/or for use by facility **3002** after processing through a pipeline (e.g., to convert outputs back to a usable data type, such as digital imaging and communications in medicine (DICOM) data, radiology information system (RIS) data, clinical information system (CIS) data, remote procedure call (RPC) data, data substantially compliant with a representation state transfer (REST) interface, data substantially compliant with a file-based interface, and/or raw data, for storage and display at facility **3002**). In at least one embodiment, a combination of containers within software **3018** (e.g., that make up a pipeline) may be referred to as a virtual instrument (as described in more detail herein), and a virtual instrument may leverage services **3020** and hardware **3022** to execute some or all processing tasks of applications instantiated in containers.

[0449] In at least one embodiment, a data processing pipeline may receive input data (e.g., imaging data **3008**) in a DICOM, RIS, CIS, REST compliant, RPC, raw, and/or other format in response to an inference request (e.g., a request from a user of deployment system **3006**, such as a clinician, a doctor, a radiologist, etc.). In at least one embodiment, input data may be representative of one or more images, video, and/or other data representations generated by one or more imaging devices, sequencing devices, radiology devices, genomics devices, and/or other device types. In at least one embodiment, data may undergo preprocessing as part of data processing pipeline to prepare data for processing by one or more applications. In at least one embodiment, post-processing may be performed on an output of one or more inferencing tasks or other processing tasks of a pipeline to prepare an output data for a next application and/or to prepare output data for transmission and/or use by a user (e.g., as a response to an inference request). In at least one embodiment, inferencing tasks may be performed by one or more machine learning models, such as trained or deployed neural networks, which may include output models **3016** of training system **3004**.

[0450] In at least one embodiment, tasks of data processing pipeline may be encapsulated in a container(s) that each represent a discrete, fully functional instantiation of an application and virtualized computing environment that is able to reference machine learning models. In at least one embodiment, containers or applications may be published into a private (e.g., limited access) area of a container registry (described in more detail herein), and trained or deployed models may be stored in model registry **3024** and associated with one or more applications. In at least one embodiment, images of applications (e.g., container images) may be available in a container registry, and once selected by a user from a container registry for deployment in a pipeline, an image may be used to generate a container for an instantiation of an application for use by a user's system.

[0451] In at least one embodiment, developers (e.g., software developers, clinicians, doctors, etc.) may develop, publish, and store applications (e.g., as containers) for performing image processing and/or inferencing on supplied data. In at least one embodiment, development, publishing, and/or storing may be performed using a software development kit (SDK) associated with a system (e.g., to ensure that an application and/or container developed is compliant with or compatible with a system). In at least one embodiment, an application that is developed may be tested locally (e.g., at a first facility, on data from a first facility) with an SDK which may support at least some of services **3020** as a system (e.g., system **3100** of FIG. **31**). In at least one embodiment, because DICOM objects may contain anywhere from one to hundreds of images or other data types, and due to a variation in data, a developer may be responsible for managing (e.g., setting constructs for, building pre-processing into an application, etc.) extraction and preparation of incoming DICOM data. In at least one embodiment, once validated by system **3100** (e.g., for accuracy, safety, patient privacy, etc.), an application may be available in a container registry for selection and/or implementation by a user (e.g., a hospital, clinic, lab, healthcare provider, etc.) to perform one or more processing tasks with respect to data at a facility (e.g., a second facility) of a user.

[0452] In at least one embodiment, developers may then share applications or containers through a network for access and use by users of a system (e.g., system **3100** of FIG. **31**). In at least one embodiment, completed and validated applications or containers may be stored in a container registry and associated machine learning models may be stored in model registry **3024**. In at least one embodiment, a requesting entity (e.g., a user at a medical facility)—who provides an inference or image processing request—may browse a container registry and/or model registry **3024** for an application, container, dataset, machine learning model, etc., select a desired combination of elements for inclusion in data processing pipeline, and submit an imaging processing request. In at least one embodiment, a request may include input data (and associated patient data, in some examples) that is necessary to perform a request, and/or may include a selection of application(s) and/or machine learning models to be executed in processing a request. In at least one embodiment, a request may then be passed to one or more components of deployment system **3006** (e.g., a cloud) to perform processing of data processing pipeline. In at least one embodiment, processing by deployment system **3006** may include referencing selected elements (e.g., applications, containers, models, etc.) from a container registry and/or model registry **3024**. In at least one embodiment, once results are generated by a pipeline, results may be returned to a user for reference (e.g., for viewing in a viewing application suite executing on a local, on-premises workstation or terminal). In at least one embodiment, a radiologist may receive results from an data processing pipeline including any number of application and/or containers, where results may include anomaly detection in X-rays, CT scans, MRIs, etc.

[0453] In at least one embodiment, to aid in processing or execution of applications or containers in pipelines, services **3020** may be leveraged. In at least one embodiment, services **3020** may include compute services, artificial intelligence (AI) services, visualization services, and/or other service types. In at least one embodiment, services **3020** may provide functionality that is common to one or more applications in software **3018**, so functionality may be abstracted to a service that may be called upon or leveraged by applications. In at least one embodiment, functionality provided by services **3020** may run dynamically and more efficiently, while also scaling well by allowing applications to process data in parallel (e.g., using a parallel computing platform **3130** (FIG. **31**)). In at least one embodiment, rather than each application that shares a same functionality offered by a service **3020** being required to have a respective instance of service **3020**, service **3020** may be shared between and among various applications. In at least one embodiment, services may include an inference server or engine that may be used for executing detection or segmentation tasks, as non-limiting examples. In at least one embodiment, a model training service may be included that may provide machine learning model training and/or retraining capabilities. In at least one embodiment, a data augmentation service may further be included that may provide GPU accelerated data (e.g., DICOM, RIS, CIS, REST compliant, RPC, raw, etc.) extraction, resizing, scaling, and/or other augmentation. In at least one embodiment, a visualization service may be used that may add image rendering effects—such as ray-tracing, rasterization, denoising, sharpening, etc.—to add realism to two-dimensional (2D) and/or three-dimensional (3D) models. In at least one embodiment, virtual instrument services may be included that provide for beam-forming, segmentation, inferencing, imaging, and/or support for other applications within pipelines of virtual instruments.

[0454] In at least one embodiment, where a service **3020** includes an AI service (e.g., an inference service), one or more machine learning models associated with an application for anomaly detection (e.g., tumors, growth abnormalities, scarring, etc.) may be executed by calling upon (e.g., as an API call) an inference service (e.g., an inference server) to execute machine learning model(s), or processing thereof, as part of application execution. In at least one embodiment, where another application includes one or more machine learning models for segmentation tasks, an application may call upon an inference service to execute machine learning models for performing one or more of processing operations associated with segmentation tasks. In at least one embodiment, software **3018** implementing advanced processing and inferencing pipeline that includes segmentation application and anomaly detection application may be streamlined because each application may call upon a same inference service to perform one or more inferencing tasks.

[0455] In at least one embodiment, hardware **3022** may include GPUs, CPUs, graphics cards, an AI/deep learning system (e.g., an AI supercomputer, such as NVIDIA's DGX supercomputer system), a cloud platform, or a combination thereof. In at least one embodiment, different types of hardware **3022** may be used to provide efficient, purpose-built support for software **3018** and services **3020** in deployment system **3006**. In at least one embodiment, use of GPU processing may be implemented for processing locally (e.g., at facility **3002**), within an AI/deep learning system, in a cloud system, and/or in other processing components of deployment system **3006** to improve efficiency, accuracy, and efficacy of image processing, image reconstruction, segmentation, MRI exams, stroke or heart attack detection (e.g., in real-time), image quality in rendering, etc. In at least one embodiment, a facility may include imaging devices, genomics devices, sequencing devices, and/or other device types on-premises that may leverage GPUs to generate imaging data representative of a subject's anatomy.

[0456] In at least one embodiment, software **3018** and/or services **3020** may be optimized for GPU processing with respect to deep learning, machine learning, and/or high-performance computing, as non-limiting examples. In at least one embodiment, at least some of computing environment of deployment system **3006** and/or training system **3004** may be executed in a datacenter one or more supercomputers or high performance computing systems, with GPU optimized software (e.g., hardware and software combination of NVIDIA's DGX system). In at least one embodiment, datacenters may be compliant with provisions of HIPAA, such that receipt, processing, and transmission of imaging data and/or other patient data is securely handled with respect to privacy of patient data. In at least one embodiment, hardware **3022** may include any number of GPUs that may be called upon to perform processing of data in parallel, as described herein. In at least one embodiment, cloud platform may further include GPU processing for GPU-optimized execution of deep learning tasks, machine learning tasks, or other computing tasks. In at least one embodiment, cloud platform (e.g., NVIDIA's NGC) may be executed using an AI/deep learning supercomputer(s) and/or GPU-optimized software (e.g., as provided on NVIDIA's DGX systems) as a hardware abstraction and scaling platform. In at least one embodiment, cloud platform may integrate an application container clustering system or orchestration system (e.g., KUBERNETES) on multiple GPUs to enable seamless scaling and load balancing.

[0457] FIG. **31** is a system diagram for an example system **3100** for generating and deploying an imaging deployment pipeline, in accordance with at least one embodiment. In at least one embodiment, system **3100** may be used to implement process **3000** of FIG. **30** and/or other processes including advanced processing and inferencing pipelines. In at least one embodiment, system **3100** may include training system **3004** and deployment system **3006**. In at least one embodiment, training system **3004** and deployment system **3006** may be implemented using software **3018**, services **3020**, and/or hardware **3022**, as described herein.

[0458] In at least one embodiment, system **3100** (e.g., training system **3004** and/or deployment system **3006**) may implemented in a cloud computing environment (e.g., using cloud **3126**). In at least one embodiment, system **3100** may be implemented locally with respect to a healthcare services facility, or as a combination of both cloud and local computing resources. In at least one embodiment, in embodiments where cloud computing is implemented, patient data may be separated from, or unprocessed by, by one or more components of system **3100** that would render processing non-compliant with HIPAA and/or other data handling and privacy regulations or laws. In at least one embodiment, access to APIs in cloud **3126** may be restricted to authorized users through enacted security measures or protocols. In at least one embodiment, a security protocol may include web tokens that may be signed by an authentication (e.g., AuthN, AuthZ, Gluecon, etc.) service and may carry appropriate authorization. In at least one embodiment, APIs of virtual instruments (described herein), or other instantiations of system **3100**, may be restricted to a set of public IPs that have been vetted or authorized for interaction.

[0459] In at least one embodiment, various components of system **3100** may communicate between and among one another using any of a variety of different network types, including but not limited to local area networks (LANs) and/or wide area networks (WANs) via wired and/or wireless communication protocols. In at least one embodiment, communication between facilities and components of system **3100** (e.g., for transmitting inference requests, for receiving results of inference requests, etc.) may be communicated over a data bus or data busses, wireless data protocols (Wi-Fi), wired data protocols (e.g., Ethernet), etc.

[0460] In at least one embodiment, training system **3004** may execute training pipelines **3104**, similar to those described herein with respect to FIG. **30**. In at least one embodiment, where one or more machine learning models are to be used in deployment pipelines **3110** by deployment system **3006**, training pipelines **3104** may be used to train or retrain one or more (e.g., pre-trained) models, and/or implement one or more of pre-trained models **3106** (e.g., without a need for retraining or updating). In at least one embodiment, as a result of training pipelines **3104**, output model(s) **3016** may be generated. In at least one embodiment, training pipelines **3104** may include any number of processing steps, such as but not limited to imaging data (or other input data) conversion or adaption (e.g., using DICOM adapter **3102**A to convert DICOM images to another format suitable for processing by respective machine learning models, such as Neuroimaging Informatics Technology Initiative (NIfTI) format), AI-assisted annotation **3010**, labeling or annotating of imaging data **3008** to generate labeled clinic data **3012**, model selection from a model registry, model training **3014**, training, retraining, or updating models, and/or other processing steps. In at least one embodiment, for different machine learning models used by deployment system **3006**, different training pipelines **3104** may be used. In at least one embodiment, training pipeline **3104** similar to a first example described with respect to FIG. **30** may be used for a first machine learning model, training pipeline **3104** similar to a second example described with respect to FIG. **30** may be used for a second machine learning model, and training pipeline **3104** similar to a third example described with respect to FIG. **30** may be used for a third machine learning model. In at least one embodiment, any combination of tasks within training system **3004** may be used depending on what is required for each respective machine learning model. In at least one embodiment, one or more of machine learning models may already be trained and ready for deployment so machine learning models may not undergo any processing by training system **3004**, and may be implemented by deployment system **3006**.

[0461] In at least one embodiment, output model(s) **3016** and/or pre-trained model(s) **3106** may include any types of machine learning models depending on implementation or embodiment. In at least one embodiment, and without limitation, machine learning models used by system **3100** may include machine learning model(s) using linear regression, logistic regression, decision trees, support vector machines (SVM), Naïve Bayes, k-nearest neighbor (Knn), K means clustering, random forest, dimensionality reduction algorithms, gradient boosting algorithms, neural networks (e.g., auto-encoders, convolutional, recurrent, perceptrons, Long/Short Term Memory (LSTM), Hopfield, Boltzmann, deep belief, deconvolutional, generative adversarial, liquid state machine, etc.), and/or other types of machine learning models.

[0462] In at least one embodiment, training pipelines **3104** may include AI-assisted annotation, as described in more detail herein with respect to at least FIG. **34**B. In at least one embodiment, labeled clinic data **3012** (e.g., traditional annotation) may be generated by any number of techniques. In at least one embodiment, labels or other annotations may be generated within a drawing program (e.g., an annotation program), a computer aided design (CAD) program, a labeling program, another type of program suitable for generating annotations or labels for ground truth, and/or may be hand drawn, in some examples. In at least one embodiment, ground truth data may be synthetically produced (e.g., generated from computer models or renderings), real produced (e.g., designed and produced from real-world data), machine-automated (e.g., using feature analysis and learning to extract features from data and then generate labels), human annotated (e.g., labeler, or annotation expert, defines location of labels), and/or a combination thereof. In at least one embodiment, for each instance of imaging data **3008** (or other data type used by machine learning models), there may be corresponding ground truth data generated by training system **3004**. In at least one embodiment, AI-assisted annotation may be performed as part of deployment pipelines **3110**; either in addition to, or in lieu of AI-assisted annotation included in training pipelines **3104**. In at least one embodiment, system **3100** may include a multi-layer platform that may include a software layer (e.g., software **3018**) of diagnostic applications (or other application types) that may perform one or more medical imaging and diagnostic functions. In at least one embodiment, system **3100** may be communicatively coupled to (e.g., via encrypted links) PACS server networks of one or more facilities. In at least one embodiment, system **3100** may be configured to access and referenced data (e.g., DICOM data, RIS data, raw data, CIS data, REST compliant data, RPC data, raw data, etc.) from PACS servers (e.g., via a DICOM adapter **3102**, or another data type adapter such as RIS, CIS, REST compliant, RPC, raw, etc.) to perform operations, such as training machine learning models, deploying machine learning models, image processing, inferencing, and/or other operations.

[0463] In at least one embodiment, a software layer may be implemented as a secure, encrypted, and/or authenticated API through which applications or containers may be invoked (e.g., called) from an external environment(s) (e.g., facility **3002**). In at least one embodiment, applications may then call or execute one or more services **3020** for performing compute, AI, or visualization tasks associated with respective applications, and software **3018** and/or services **3020** may leverage hardware **3022** to perform processing tasks in an effective and efficient manner.

[0464] In at least one embodiment, deployment system **3006** may execute deployment pipelines **3110**. In at least one embodiment, deployment pipelines **3110** may include any number of applications that may be sequentially, non-sequentially, or otherwise applied to imaging data (and/or other data types) generated by imaging devices, sequencing devices, genomics devices, etc.—including AI-assisted annotation, as described above. In at least one embodiment, as described herein, a deployment pipeline **3110** for an individual device may be referred to as a virtual instrument for a device (e.g., a virtual ultrasound instrument, a virtual CT scan instrument, a virtual sequencing instrument, etc.). In at least one embodiment, for a single device, there may be more than one deployment pipeline **3110** depending on information desired from data generated by a device. In at least one embodiment, where detections of anomalies are desired from an MRI machine, there may be a first deployment pipeline **3110**, and where image enhancement is desired from output of an MRI machine, there may be a second deployment pipeline **3110**.

[0465] In at least one embodiment, applications available for deployment pipelines **3110** may include any application that may be used for performing processing tasks on imaging data or other data from devices. In at least one embodiment, different applications may be responsible for image enhancement, segmentation, reconstruction, anomaly detection, object detection, feature detection, treatment planning, dosimetry, beam planning (or other radiation treatment procedures), and/or other analysis, image processing, or inferencing tasks. In at least one embodiment, deployment system **3006** may define constructs for each of applications, such that users of deployment system **3006** (e.g., medical facilities, labs, clinics, etc.) may understand constructs and adapt applications for implementation within their respective facility. In at least one embodiment, an application for image reconstruction may be selected for inclusion in deployment pipeline **3110**, but data type generated by an imaging device may be different from a data type used within an application. In at least one embodiment, DICOM adapter **3102**B (and/or a DICOM reader) or another data type adapter or reader (e.g., RIS, CIS, REST compliant, RPC, raw, etc.) may be used within deployment pipeline **3110** to convert data to a form useable by an application within deployment system **3006**. In at least one embodiment, access to DICOM, RIS, CIS, REST compliant, RPC, raw, and/or other data type libraries may be accumulated and pre-processed, including decoding, extracting, and/or performing any convolutions, color corrections, sharpness, gamma, and/or other augmentations to data. In at least one embodiment, DICOM, RIS, CIS, REST compliant, RPC, and/or raw data may be unordered and a pre-pass may be executed to organize or sort collected data. In at least one embodiment, because various applications may share common image operations, in some embodiments, a data augmentation library (e.g., as one of services **3020**) may be used to accelerate these operations. In at least one embodiment, to avoid bottlenecks of conventional processing approaches that rely on CPU processing, parallel computing platform **3130** may be used for GPU acceleration of these processing tasks.

[0466] In at least one embodiment, an image reconstruction application may include a processing task that includes use of a machine learning model. In at least one embodiment, a user may desire to use their own machine learning model, or to select a machine learning model from model registry **3024**. In at least one embodiment, a user may implement their own machine learning model or select a machine learning model for inclusion in an application for performing a processing task. In at least one embodiment, applications may be selectable and customizable, and by defining constructs of applications, deployment and implementation of applications for a particular user are presented as a more seamless user experience. In at least one embodiment, by leveraging other features of system **3100**—such as services **3020** and hardware **3022**—deployment pipelines **3110** may be even more user friendly, provide for easier integration, and produce more accurate, efficient, and timely results.

[0467] In at least one embodiment, deployment system **3006** may include a user interface **3114** (e.g., a graphical user interface, a web interface, etc.) that may be used to select applications for inclusion in deployment pipeline(s) **3110**, arrange applications, modify or change applications or parameters or constructs thereof, use and interact with deployment pipeline(s) **3110** during set-up and/or deployment, and/or to otherwise interact with deployment system **3006**. In at least one embodiment, although not illustrated with respect to training system **3004**, user interface **3114** (or a different user interface) may be used for selecting models for use in deployment system **3006**, for selecting models for training, or retraining, in training system **3004**, and/or for otherwise interacting with training system **3004**.

[0468] In at least one embodiment, pipeline manager **3112** may be used, in addition to an application orchestration system **3128**, to manage interaction between applications or containers of deployment pipeline(s) **3110** and services **3020** and/or hardware **3022**. In at least one embodiment, pipeline manager **3112** may be configured to facilitate interactions from application to application, from application to service **3020**, and/or from application or service to hardware **3022**. In at least one embodiment, although illustrated as included in software **3018**, this is not intended to be limiting, and in some examples (e.g., as illustrated in FIG. **32**) pipeline manager **3112** may be included in services **3020**. In at least one embodiment, application orchestration system **3128** (e.g., Kubernetes, DOCKER, etc.) may include a container orchestration system that may group applications into containers as logical units for coordination, management, scaling, and deployment. In at least one embodiment, by associating applications from deployment pipeline(s) **3110** (e.g., a reconstruction application, a segmentation application, etc.) with individual containers, each application may execute in a self-contained environment (e.g., at a kernel level) to increase speed and efficiency.

[0469] In at least one embodiment, each application and/or container (or image thereof) may be individually developed, modified, and deployed (e.g., a first user or developer may develop, modify, and deploy a first application and a second user or developer may develop, modify, and deploy a second application separate from a first user or developer), which may allow for focus on, and attention to, a task of a single application and/or container(s) without being hindered by tasks of another application(s) or container(s). In at least one embodiment, communication, and cooperation between different containers or applications may be aided by pipeline manager **3112** and application orchestration system **3128**. In at least one embodiment, so long as an expected input and/or output of each container or application is known by a system (e.g., based on constructs of applications or containers), application orchestration system **3128** and/or pipeline manager **3112** may facilitate communication among and between, and sharing of resources among and between, each of applications or containers. In at least one embodiment, because one or more of applications or containers in deployment pipeline(s) **3110** may share same services and resources, application orchestration system **3128** may orchestrate, load balance, and determine sharing of services or resources between and among various applications or containers. In at least one embodiment, a scheduler may be used to track resource requirements of applications or containers, current usage or planned usage of these resources, and resource availability. In at least one embodiment, a scheduler may thus allocate resources to different applications and distribute resources between and among applications in view of requirements and availability of a system. In some examples, a scheduler (and/or other component of application orchestration system **3128**) may determine resource availability and distribution based on constraints imposed on a system (e.g., user constraints), such as quality of service (QoS), urgency of need for data outputs (e.g., to determine whether to execute real-time processing or delayed processing), etc.

[0470] In at least one embodiment, services **3020** leveraged by and shared by applications or containers in deployment system **3006** may include compute services **3116**, AI services **3118**, visualization services **3120**, and/or other service types. In at least one embodiment, applications may call (e.g., execute) one or more of services **3020** to perform processing operations for an application. In at least one embodiment, compute services **3116** may be leveraged by applications to perform super-computing or other high-performance computing (HPC) tasks. In at least one embodiment, compute service(s) **3116** may be leveraged to perform parallel processing (e.g., using a parallel computing platform **3130**) for processing data through one or more of applications and/or one or more tasks of a single application, substantially simultaneously. In at least one embodiment, parallel computing platform **3130** (e.g., NVIDIA's CUDA) may enable general purpose computing on GPUs (GPGPU) (e.g., GPUs **3122**). In at least one embodiment, a software layer of parallel computing platform **3130** may provide access to virtual instruction sets and parallel computational elements of GPUs, for execution of compute kernels. In at least one embodiment, parallel computing platform **3130** may include memory and, in some embodiments, a memory may be shared between and among multiple containers, and/or between and among different processing tasks within a single container. In at least one embodiment, inter-process communication (IPC) calls may be generated for multiple containers and/or for multiple processes within a container to use same data from a shared segment of memory of parallel computing platform **3130** (e.g., where multiple different stages of an application or multiple applications are processing same information). In at least one embodiment, rather than making a copy of data and moving data to different locations in memory (e.g., a read/write operation), same data in same location of a memory may be used for any number of processing tasks (e.g., at a same time, at different times, etc.). In at least one embodiment, as data is used to generate new data as a result of processing, this information of a new location of data may be stored and shared between various applications. In at least one embodiment, location of data and a location of updated or modified data may be part of a definition of how a payload is understood within containers.

[0471] In at least one embodiment, AI services **3118** may be leveraged to perform inferencing services for executing machine learning model(s) associated with applications (e.g., tasked with performing one or more processing tasks of an application). In at least one embodiment, AI services **3118** may leverage AI system **3124** to execute machine learning model(s) (e.g., neural networks, such as CNNs) for segmentation, reconstruction, object detection, feature detection, classification, and/or other inferencing tasks. In at least one embodiment, applications of deployment pipeline (s) **3110** may use one or more of output models **3016** from training system **3004** and/or other models of applications to perform inference on imaging data (e.g., DICOM data, RIS data, CIS data, REST compliant data, RPC data, raw data, etc.). In at least one embodiment, two or more examples of inferencing using application orchestration system **3128** (e.g., a scheduler) may be available. In at least one embodiment, a first category may include a high priority/low latency path that may achieve higher service level agreements, such as for performing inference on urgent requests during an emergency, or for a radiologist during diagnosis. In at least one embodiment, a second category may include a standard priority path that may be used for requests that may be non-urgent or where analysis may be performed at a later time. In at least one embodiment, application orchestration system **3128** may distribute resources (e.g., services **3020** and/or hardware **3022**) based on priority paths for different inferencing tasks of AI services **3118**.

[0472] In at least one embodiment, shared storage may be mounted to AI services **3118** within system **3100**. In at least one embodiment, shared storage may operate as a cache (or other storage device type) and may be used to process inference requests from applications. In at least one embodiment, when an inference request is submitted, a request may be received by a set of API instances of deployment system **3006**, and one or more instances may be selected (e.g., for best fit, for load balancing, etc.) to process a request. In at least one embodiment, to process a request, a request may be entered into a database, a machine learning model may be located from model registry **3024** if not already in a cache, a validation step may ensure appropriate machine learning model is loaded into a cache (e.g., shared storage), and/or a copy of a model may be saved to a cache. In at least one embodiment, a scheduler (e.g., of pipeline manager **3112**) may be used to launch an application that is referenced in a request if an application is not already running or if there are not enough instances of an application. In at least one embodiment, if an inference server is not already launched to execute a model, an inference server may be launched. In at least one embodiment, any number of inference servers may be launched per model. In at least one embodiment, in a pull model, in which inference servers are clustered, models may be cached whenever load balancing is advantageous. In at least one embodiment, inference servers may be statically loaded in corresponding, distributed servers.

[0473] In at least one embodiment, inferencing may be performed using an inference server that runs in a container. In at least one embodiment, an instance of an inference server may be associated with a model (and optionally a plurality of versions of a model). In at least one embodiment, if an instance of an inference server does not exist when a request to perform inference on a model is received, a new instance may be loaded. In at least one embodiment, when starting an inference server, a model may be passed to an inference server such that a same container may be used to serve different models so long as inference server is running as a different instance.

[0474] In at least one embodiment, during application execution, an inference request for a given application may be received, and a container (e.g., hosting an instance of an inference server) may be loaded (if not already), and a start procedure may be called. In at least one embodiment, pre-processing logic in a container may load, decode, and/or perform any additional pre-processing on incoming data (e.g., using a CPU(s) and/or GPU(s)). In at least one embodiment, once data is prepared for inference, a container may perform inference as necessary on data. In at least one embodiment, this may include a single inference call on one image (e.g., a hand X-ray), or may require inference on hundreds of images (e.g., a chest CT). In at least one embodiment, an application may summarize results before completing, which may include, without limitation, a single confidence score, pixel level-segmentation, voxel-level segmentation, generating a visualization, or generating text to summarize findings. In at least one embodiment, different models or applications may be assigned different priorities. For example, some models may have a real-time (TAT less than one minute) priority while others may have lower priority (e.g., TAT less than 10 minutes). In at least one embodiment, model execution times may be measured from requesting institution or entity and may include partner network traversal time, as well as execution on an inference service.

[0475] In at least one embodiment, transfer of requests between services **3020** and inference applications may be hidden behind a software development kit (SDK), and robust transport may be provide through a queue. In at least one embodiment, a request will be placed in a queue via an API for an individual application/tenant ID combination and an SDK will pull a request from a queue and give a request to an application. In at least one embodiment, a name of a queue may be provided in an environment from where an SDK will pick it up. In at least one embodiment, asynchronous communication through a queue may be useful as it may allow any instance of an application to pick up work as it becomes available. In at least one embodiment, results may be transferred back through a queue, to ensure no data is lost. In at least one embodiment, queues may also provide an ability to segment work, as highest priority work may go to a queue with most instances of an application connected to it, while lowest priority work may go to a queue with a single instance connected to it that processes tasks in an order received. In at least one embodiment, an application may run on a GPU-accelerated instance generated in cloud **3126**, and an inference service may perform inferencing on a GPU.

[0476] In at least one embodiment, visualization services **3120** may be leveraged to generate visualizations for viewing outputs of applications and/or deployment pipeline(s) **3110**. In at least one embodiment, GPUs **3122** may be leveraged by visualization services **3120** to generate visualizations. In at least one embodiment, rendering effects, such as ray-tracing, may be implemented by visualization services **3120** to generate higher quality visualizations. In at least one embodiment, visualizations may include, without limitation, 2D image renderings, 3D volume renderings, 3D volume reconstruction, 2D tomographic slices, virtual reality displays, augmented reality displays, etc. In at least one embodiment, virtualized environments may be used to generate a virtual interactive display or environment (e.g., a virtual environment) for interaction by users of a system (e.g., doctors, nurses, radiologists, etc.). In at least one embodiment, visualization services **3120** may include an internal visualizer, cinematics, and/or other rendering or image processing capabilities or functionality (e.g., ray tracing, rasterization, internal optics, etc.).

[0477] In at least one embodiment, hardware **3022** may include GPUs **3122**, AI system **3124**, cloud **3126**, and/or any other hardware used for executing training system **3004** and/or deployment system **3006**. In at least one embodiment, GPUs **3122** (e.g., NVIDIA's TESLA and/or QUADRO GPUs) may include any number of GPUs that may be used for executing processing tasks of compute services **3116**, AI services **3118**, visualization services **3120**, other services, and/or any of features or functionality of software **3018**. For example, with respect to AI services **3118**, GPUs **3122** may be used to perform pre-processing on imaging data (or other data types used by machine learning models), post-processing on outputs of machine learning models, and/or to perform inferencing (e.g., to execute machine learning models). In at least one embodiment, cloud **3126**, AI system **3124**, and/or other components of system **3100** may use GPUs **3122**. In at least one embodiment, cloud **3126** may include a GPU-optimized platform for deep learning tasks. In at least one embodiment, AI system **3124** may use GPUs, and cloud **3126**—or at least a portion tasked with deep learning or inferencing—may be executed using one or more AI systems **3124**. As such, although hardware **3022** is illustrated as discrete components, this is not intended to be limiting, and any components of hardware **3022** may be combined with, or leveraged by, any other components of hardware **3022**.

[0478] In at least one embodiment, AI system **3124** may include a purpose-built computing system (e.g., a supercomputer or an HPC) configured for inferencing, deep learning, machine learning, and/or other artificial intelligence tasks. In at least one embodiment, AI system **3124** (e.g., NVIDIA's DGX) may include GPU-optimized software (e.g., a software stack) that may be executed using a plurality of GPUs **3122**, in addition to CPUs, RAM, storage, and/or other components, features, or functionality. In at least one embodiment, one or more AI systems **3124** may be implemented in cloud **3126** (e.g., in a data center) for performing some or all of AI-based processing tasks of system **3100**.

[0479] In at least one embodiment, cloud **3126** may include a GPU-accelerated infrastructure (e.g., NVIDIA's NGC) that may provide a GPU-optimized platform for executing processing tasks of system **3100**. In at least one embodiment, cloud **3126** may include an AI system(s) **3124** for performing one or more of AI-based tasks of system **3100** (e.g., as a hardware abstraction and scaling platform). In at least one embodiment, cloud **3126** may integrate with application orchestration system **3128** leveraging multiple GPUs to enable seamless scaling and load balancing between and among applications and services **3020**. In at least one embodiment, cloud **3126** may tasked with executing at least some of services **3020** of system **3100**, including compute services **3116**, AI services **3118**, and/or visualization services **3120**, as described herein. In at least one embodiment, cloud **3126** may perform small and large batch inference (e.g., executing NVIDIA's TENSOR RT), provide an accelerated parallel computing API and platform **3130** (e.g., NVIDIA's CUDA), execute application orchestration system **3128** (e.g., KUBERNETES), provide a graphics rendering API and platform (e.g., for ray-tracing, 2D graphics, 3D graphics, and/or other rendering techniques to produce higher quality cinematics), and/or may provide other functionality for system **3100**.

[0480] In at least one embodiment, in an effort to preserve patient confidentiality (e.g., where patient data or records are to be used off-premises), cloud **3126** may include a registry—such as a deep learning container registry. In at least one embodiment, a registry may store containers for instantiations of applications that may perform pre-processing, post-processing, or other processing tasks on patient data. In at least one embodiment, cloud **3126** may receive data that includes patient data as well as sensor data in containers, perform requested processing for just sensor data in those containers, and then forward a resultant output and/or visualizations to appropriate parties and/or devices (e.g., on-premises medical devices used for visualization or diagnoses), all without having to extract, store, or otherwise access patient data. In at least one embodiment, confidentiality of patient data is preserved in compliance with HIPAA and/or other data regulations.

[0481] FIG. **32** includes an example illustration of a deployment pipeline **3110**A for processing imaging data, in accordance with at least one embodiment. In at least one embodiment, system **3100**—and specifically deployment system **3006**—may be used to customize, update, and/or integrate deployment pipeline(s) **3110**A into one or more production environments. In at least one embodiment, deployment pipeline **3110**A of FIG. **32** includes a non-limiting example of a deployment pipeline **3110**A that may be custom defined by a particular user (or team of users) at a facility (e.g., at a hospital, clinic, lab, research environment, etc.). In at least one embodiment, to define deployment pipelines **3110**A for a CT scanner **3202**, a user may select—from a container registry, for example—one or more applications that perform specific functions or tasks with respect to imaging data generated by CT scanner **3202**. In at least one embodiment, applications may be applied to deployment pipeline **3110**A as containers that may leverage services **3020** and/or hardware **3022** of system **3100**. In addition, deployment pipeline **3110**A may include additional processing tasks or applications that may be implemented to prepare data for use by applications (e.g., DICOM adapter **3102**B and DICOM reader **3206** may be used in deployment pipeline **3110**A to prepare data for use by CT reconstruction **3208**, organ segmentation **3210**, etc.). In at least one embodiment, deployment pipeline **3110**A may be customized or selected for consistent deployment, one time use, or for another frequency or interval. In at least one embodiment, a user may desire to have CT reconstruction **3208** and organ segmentation **3210** for several subjects over a specific interval, and thus may deploy pipeline **3110**A for that period of time. In at least one embodiment, a user may select, for each request from system **3100**, applications that a user wants to perform processing on that data for that request. In at least one embodiment, deployment pipeline **3110**A may be adjusted at any interval and, because of adaptability and scalability of a container structure within system **3100**, this may be a seamless process.

[0482] In at least one embodiment, deployment pipeline **3110**A of FIG. **32** may include CT scanner **3202** generating imaging data of a patient or subject. In at least one embodiment, imaging data from CT scanner **3202** may be stored on a PACS server(s) **3204** associated with a facility housing CT scanner **3202**. In at least one embodiment, PACS server(s) **3204** may include software and/or hardware components that may directly interface with imaging modalities (e.g., CT scanner **3202**) at a facility. In at least one embodiment, DICOM adapter **3102**B may enable sending and receipt of DICOM objects using DICOM protocols. In at least one embodiment, DICOM adapter **3102**B may aid in preparation or configuration of DICOM data from PACS server(s) **3204** for use by deployment pipeline **3110**A. In at least one embodiment, once DICOM data is processed through DICOM adapter **3102**B, pipeline manager **3112** may route data through to deployment pipeline **3110**A. In at least one embodiment, DICOM reader **3206** may extract image files and any associated metadata from DICOM data (e.g., raw sinogram data, as illustrated in visualization **3216**A). In at least one embodiment, working files that are extracted may be stored in a cache for faster processing by other applications in deployment pipeline **3110**A. In at least one embodiment, once DICOM reader **3206** has finished extracting and/or storing data, a signal of completion may be communicated to pipeline manager **3112**. In at least one embodiment, pipeline manager **3112** may then initiate or call upon one or more other applications or containers in deployment pipeline **3110**A.

[0483] In at least one embodiment, CT reconstruction **3208** application and/or container may be executed once data (e.g., raw sinogram data) is available for processing by CT reconstruction **3208** application. In at least one embodiment, CT reconstruction **3208** may read raw sinogram data from a cache, reconstruct an image file out of raw sinogram data (e.g., as illustrated in visualization **3216**B), and store resulting image file in a cache. In at least one embodiment, at completion of reconstruction, pipeline manager **3112** may be signaled that reconstruction task is complete. In at least one embodiment, once reconstruction is complete, and a reconstructed image file may be stored in a cache (or other storage device), organ segmentation **3210** application and/or container may be triggered by pipeline manager **3112**. In at least one embodiment, organ segmentation **3210** application and/or container may read an image file from a cache, normalize or convert an image file to format suitable for inference (e.g., convert an image file to an input resolution of a machine learning model), and run inference against a normalized image. In at least one embodiment, to run inference on a normalized image, organ segmentation **3210** application and/or container may rely on services **3020**, and pipeline manager **3112** and/or application orchestration system **3128** may facilitate use of services **3020** by organ segmentation **3210** application and/or container. In at least one embodiment, for example, organ segmentation **3210** application and/or container may leverage AI services **3118** to perform inference on a normalized image, and AI services **3118** may leverage hardware **3022** (e.g., AI system **3124**) to execute AI services **3118**. In at least one embodiment, a result of an inference may be a mask file (e.g., as illustrated in visualization **3216**C) that may be stored in a cache (or other storage device).

[0484] In at least one embodiment, once applications that process DICOM data and/or data extracted from DICOM data have completed processing, a signal may be generated for pipeline manager **3112**. In at least one embodiment, pipeline manager **3112** may then execute DICOM writer **3212** to read results from a cache (or other storage device), package results into a DICOM format (e.g., as DICOM output **3214**) for use by users at a facility who generated a request. In at least one embodiment, DICOM output **3214** may then be transmitted to DICOM adapter **3102**B to prepare DICOM output **3214** for storage on PACS server(s) **3204** (e.g., for viewing by a DICOM viewer at a facility). In at least one embodiment, in response to a request for reconstruction and segmentation, visualizations **3216**B and **3216**C may be generated and available to a user for diagnoses, research, and/or for other purposes.

[0485] Although illustrated as consecutive application in deployment pipeline **3110**A, CT reconstruction **3208** and organ segmentation **3210** applications may be processed in parallel in at least one embodiment. In at least one embodiment, where applications do not have dependencies on one another, and data is available for each application (e.g., after DICOM reader **3206** extracts data), applications may be executed at a same time, substantially at a same time, or with some overlap. In at least one embodiment, where two or more applications require similar services **3020**, a scheduler of system **3100** may be used to load balance and distribute compute or processing resources between and among various applications. In at least one embodiment, in some embodiments, parallel computing platform **3130** may be used to perform parallel processing for applications to decrease run-time of deployment pipeline **3110**A to provide real-time results.

[0486] In at least one embodiment, and with reference to FIGS. **33**A-**33**B, deployment system **3006** may be implemented as one or more virtual instruments to perform different functionalities—such as image processing, segmentation, enhancement, AI, visualization, and inferencing—with imaging devices (e.g., CT scanners, X-ray machines, MRI machines, etc.), sequencing devices, genomics devices, and/or other device types. In at least one embodiment, system **3100** may allow for creation and provision of virtual instruments that may include a software-defined deployment pipeline **3110** that may receive raw/unprocessed input data generated by a device(s) and output processed/reconstructed data. In at least one embodiment, deployment pipelines **3110** (e.g., **3110**A and **3110**B) that represent virtual instruments may implement intelligence into a pipeline, such as by leveraging machine learning models, to provide containerized inference support to a system. In at least one embodiment, virtual instruments may execute any number of containers each including instantiations of applications. In at least one embodiment, such as where real-time processing is desired, deployment pipelines **3110** representing virtual instruments may be static (e.g., containers and/or applications may be set), while in other examples, container and/or applications for virtual instruments may be selected (e.g., on a per-request basis) from a pool of applications or resources (e.g., within a container registry).

[0487] In at least one embodiment, system **3100** may be instantiated or executed as one or more virtual instruments on-premise at a facility in, for example, a computing system deployed next to or otherwise in communication with a radiology machine, an imaging device, and/or another device type at a facility. In at least one embodiment, however, an on-premise installation may be instantiated or executed within a computing system of a device itself (e.g., a computing system integral to an imaging device), in a local datacenter (e.g., a datacenter on-premise), and/or in a cloud-environment (e.g., in cloud **3126**). In at least one embodiment, deployment system **3006**, operating as a virtual instrument, may be instantiated by a supercomputer or other HPC system in some examples. In at least one embodiment, on-premise installation may allow for high-bandwidth uses (via, for example, higher throughput local communication interfaces, such as RF over Ethernet) for real-time processing. In at least one embodiment, real-time or near real-time processing may be particularly useful where a virtual instrument supports an ultrasound device or other imaging modality where immediate visualizations are expected or required for accurate diagnoses and analyses. In at least one embodiment, a cloud-computing architecture may be capable of dynamic bursting to a cloud computing service provider, or other compute cluster, when local demand exceeds on-premise capacity or capability. In at least one embodiment, a cloud architecture, when implemented, may be tuned for training neural networks or other machine learning models, as described herein with respect to training system **3004**. In at least one embodiment, with training pipelines in place, machine learning models may be continuously learn and improve as they process additional data from devices they support. In at least one embodiment, virtual instruments may be continually improved using additional data, new data, existing machine learning models, and/or new or updated machine learning models.

[0488] In at least one embodiment, a computing system may include some or all of hardware **3022** described herein, and hardware **3022** may be distributed in any of a number of ways including within a device, as part of a computing device coupled to and located proximate a device, in a local datacenter at a facility, and/or in cloud **3126**. In at least one embodiment, because deployment system **3006** and associated applications or containers are created in software (e.g., as discrete containerized instantiations of applications), behavior, operation, and configuration of virtual instruments, as well as outputs generated by virtual instruments, may be modified or customized as desired, without having to change or alter raw output of a device that a virtual instrument supports.

[0489] FIG. **33**A includes an example data flow diagram of a virtual instrument supporting an ultrasound device, in accordance with at least one embodiment. In at least one embodiment, deployment pipeline **3110**B may leverage one or more of services **3020** of system **3100**. In at least one embodiment, deployment pipeline **3110**B and services **3020** may leverage hardware **3022** of a system either locally or in cloud **3126**. In at least one embodiment, although not illustrated, process **3300** may be facilitated by pipeline manager **3112**, application orchestration system **3128**, and/or parallel computing platform **3130**.

[0490] In at least one embodiment, process **3300** may include receipt of imaging data from an ultrasound device **3302**. In at least one embodiment, imaging data may be stored on PACS server(s) in a DICOM format (or other format, such as RIS, CIS, REST compliant, RPC, raw, etc.), and may be received by system **3100** for processing through deployment pipeline **3110** selected or customized as a virtual instrument (e.g., a virtual ultrasound) for ultrasound device **3302**. In at least one embodiment, imaging data may be received directly from an imaging device (e.g., ultrasound device **3302**) and processed by a virtual instrument. In at least one embodiment, a transducer or other signal converter communicatively coupled between an imaging device and a virtual instrument may convert signal data generated by an imaging device to image data that may be processed by a virtual instrument. In at least one embodiment, raw data and/or image data may be applied to DICOM reader **3206** to extract data for use by applications or containers of deployment pipeline **3110**B. In at least one embodiment, DICOM reader **3206** may leverage data augmentation library **3314** (e.g., NVIDIA's DALI) as a service **3020** (e.g., as one of compute service(s) **3116**) for extracting, resizing, rescaling, and/or otherwise preparing data for use by applications or containers.

[0491] In at least one embodiment, once data is prepared, a reconstruction **3306** application and/or container may be executed to reconstruct data from ultrasound device **3302** into an image file. In at least one embodiment, after reconstruction **3306**, or at a same time as reconstruction **3306**, a detection **3308** application and/or container may be executed for anomaly detection, object detection, feature detection, and/or other detection tasks related to data. In at least one embodiment, an image file generated during reconstruction **3306** may be used during detection **3308** to identify anomalies, objects, features, etc. In at least one embodiment, detection **3308** application may leverage an inference engine **3316** (e.g., as one of AI service(s) **3118**) to perform inference on data to generate detections. In at least one embodiment, one or more machine learning models (e.g., from training system **3004**) may be executed or called by detection **3308** application.

[0492] In at least one embodiment, once reconstruction **3306** and/or detection **3308** is/are complete, data output from these application and/or containers may be used to generate visualizations **3310**, such as visualization **3312** (e.g., a grayscale output) displayed on a workstation or display terminal. In at least one embodiment, visualization may allow a technician or other user to visualize results of deployment pipeline **3110**B with respect to ultrasound device **3302**. In at least one embodiment, visualization **3310** may be executed by leveraging a render component **3318** of system **3100** (e.g., one of visualization service(s) **3120**). In at least one embodiment, render component **3318** may execute a 2D, OpenGL, or ray-tracing service to generate visualization **3312**.

[0493] FIG. **33**B includes an example data flow diagram of a virtual instrument supporting a CT scanner, in accordance with at least one embodiment. In at least one embodiment, deployment pipeline **3110**C may leverage one or more of services **3020** of system **3100**. In at least one embodiment, deployment pipeline **3110**C and services **3020** may leverage hardware **3022** of a system either locally or in cloud **3126**. In at least one embodiment, although not illustrated, process **3320** may be facilitated by pipeline manager **3112**, application orchestration system **3128**, and/or parallel computing platform **3130**.

[0494] In at least one embodiment, process **3320** may include CT scanner **3322** generating raw data that may be received by DICOM reader **3206** (e.g., directly, via a PACS server **3204**, after processing, etc.). In at least one embodiment, a Virtual CT (instantiated by deployment pipeline **3110**C) may include a first, real-time pipeline for monitoring a patient (e.g., patient movement detection AI **3326**) and/or for adjusting or optimizing exposure of CT scanner **3322** (e.g., using exposure control AI **3324**). In at least one embodiment, one or more of applications (e.g., **3324** and **3326**) may leverage a service **3020**, such as AI service(s) **3118**. In at least one embodiment, outputs of exposure control AI **3324** application (or container) and/or patient movement detection AI **3326** application (or container) may be used as feedback to CT scanner **3322** and/or a technician for adjusting exposure (or other settings of CT scanner **3322**) and/or informing a patient to move less.

[0495] In at least one embodiment, deployment pipeline **3110**C may include a non-real-time pipeline for analyzing data generated by CT scanner **3322**. In at least one embodiment, a second pipeline may include CT reconstruction **3208** application and/or container, a coarse detection AI **3328** application and/or container, a fine detection AI **3332** application and/or container (e.g., where certain results are detected by coarse detection AI **3328**), a visualization **3330** application and/or container, and a DICOM writer **3212** (and/or other data type writer, such as RIS, CIS, REST compliant, RPC, raw, etc.) application and/or container. In at least one embodiment, raw data generated by CT scanner **3322** may be passed through pipelines of deployment pipeline **3110**C (instantiated as a virtual CT instrument) to generate results. In at least one embodiment, results from DICOM writer **3212** may be transmitted for display and/or may be stored on PACS server(s) **3204** for later retrieval, analysis, or display by a technician, practitioner, or other user.

[0496] FIG. **34**A illustrates a data flow diagram for a process **3400** to train, retrain, or update a machine learning model, in accordance with at least one embodiment. In at least one embodiment, process **3400** may be executed using, as a non-limiting example, system **3100** of FIG. **31**. In at least one embodiment, process **3400** may leverage services **3020** and/or hardware **3022** of system **3100**, as described herein. In at least one embodiment, refined models **3412** generated by process **3400** may be executed by deployment system **3006** for one or more containerized applications in deployment pipelines **3110**.

[0497] In at least one embodiment, model training **3014** may include retraining or updating an initial model **3404** (e.g., a pre-trained model) using new training data (e.g., new input data, such as customer dataset **3406**, and/or new ground truth data associated with input data). In at least one embodiment, to retrain, or update, initial model **3404**, output or loss layer(s) of initial model **3404** may be reset, or deleted, and/or replaced with an updated or new output or loss layer(s). In at least one embodiment, initial model **3404** may have previously fine-tuned parameters (e.g., weights and/or biases) that remain from prior training, so training or retraining **3014** may not take as long or require as much processing as training a model from scratch. In at least one embodiment, during model training **3014**, by having reset or replaced output or loss layer(s) of initial model **3404**, parameters may be updated and re-tuned for a new data set based on loss calculations associated with accuracy of output or loss layer(s) at generating predictions on new, customer dataset **3406** (e.g., image data **3008** of FIG. **30**).

[0498] In at least one embodiment, pre-trained models **3106** may be stored in a data store, or registry (e.g., model registry **3024** of FIG. **30**). In at least one embodiment, pre-trained models **3106** may have been trained, at least in part, at one or more facilities other than a facility executing process **3400**. In at least one embodiment, to protect privacy and rights of patients, subjects, or clients of different facilities, pre-trained models **3106** may have been trained, on-premise, using customer or patient data generated on-premise. In at least one embodiment, pre-trained models **3106** may be trained using cloud **3126** and/or other hardware **3022**, but confidential, privacy protected patient data may not be transferred to, used by, or accessible to any components of cloud **3126** (or other off premise hardware). In at least one embodiment, where a pre-trained model **3106** is trained at using patient data from more than one facility, pre-trained model **3106** may have been individually trained for each facility prior to being trained on patient or customer data from another facility. In at least one embodiment, such as where a customer or patient data has been released of privacy concerns (e.g., by waiver, for experimental use, etc.), or where a customer or patient data is included in a public data set, a customer or patient data from any number of facilities may be used to train pre-trained model **3106** on-premise and/or off premise, such as in a datacenter or other cloud computing infrastructure.

[0499] In at least one embodiment, when selecting applications for use in deployment pipelines **3110**, a user may also select machine learning models to be used for specific applications. In at least one embodiment, a user may not have a model for use, so a user may select a pre-trained model **3106** to use with an application. In at least one embodiment, pre-trained model **3106** may not be optimized for generating accurate results on customer dataset **3406** of a facility of a user (e.g., based on patient diversity, demographics, types of medical imaging devices used, etc.). In at least one embodiment, prior to deploying pre-trained model **3106** into deployment pipeline **3110** for use with an application(s), pre-trained model **3106** may be updated, retrained, and/or fine-tuned for use at a respective facility.

[0500] In at least one embodiment, a user may select pre-trained model **3106** that is to be updated, retrained, and/or fine-tuned, and pre-trained model **3106** may be referred to as initial model **3404** for training system **3004** within process **3400**. In at least one embodiment, customer dataset **3406** (e.g., imaging data, genomics data, sequencing data, or other data types generated by devices at a facility) may be used to perform model training **3014** (which may include, without limitation, transfer learning) on initial model **3404** to generate refined model **3412**. In at least one embodiment, ground truth data corresponding to customer dataset **3406** may be generated by training system **3004**. In at least one embodiment, ground truth data may be generated, at least in part, by clinicians, scientists, doctors, practitioners, at a facility (e.g., as labeled clinic data **3012** of FIG. **30**).

[0501] In at least one embodiment, AI-assisted annotation **3010** may be used in some examples to generate ground truth data. In at least one embodiment, AI-assisted annotation **3010** (e.g., implemented using an AI-assisted annotation SDK) may leverage machine learning models (e.g., neural networks) to generate suggested or predicted ground truth data for a customer dataset. In at least one embodiment, user **3410** may use annotation tools within a user interface (a graphical user interface (GUI)) on computing device **3408**.

[0502] In at least one embodiment, user **3410** may interact with a GUI via computing device **3408** to edit or fine-tune annotations or auto-annotations. In at least one embodiment, a polygon editing feature may be used to move vertices of a polygon to more accurate or fine-tuned locations.

[0503] In at least one embodiment, once customer dataset **3406** has associated ground truth data, ground truth data (e.g., from AI-assisted annotation, manual labeling, etc.) may be used by during model training **3014** to generate refined model **3412**. In at least one embodiment, customer dataset **3406** may be applied to initial model **3404** any number of times, and ground truth data may be used to update parameters of initial model **3404** until an acceptable level of accuracy is attained for refined model **3412**. In at least one embodiment, once refined model **3412** is generated, refined model **3412** may be deployed within one or more deployment pipelines **3110** at a facility for performing one or more processing tasks with respect to medical imaging data.

[0504] In at least one embodiment, refined model **3412** may be uploaded to pre-trained models **3106** in model registry **3024** to be selected by another facility. In at least one embodiment, his process may be completed at any number of facilities such that refined model **3412** may be further refined on new datasets any number of times to generate a more universal model.

[0505] FIG. **34**B is an example illustration of a client-server architecture **3432** to enhance annotation tools with pre-trained annotation models, in accordance with at least one embodiment. In at least one embodiment, AI-assisted annotation tools **3436** may be instantiated based on a client-server architecture **3432**. In at least one embodiment, annotation tools **3436** in imaging applications may aid radiologists, for example, identify organs and abnormalities. In at least one embodiment, imaging applications may include software tools that help user **3410** to identify, as a non-limiting example, a few extreme points on a particular organ of interest in raw images **3434** (e.g., in a 3D MRI or CT scan) and receive auto-annotated results for all 2D slices of a particular organ. In at least one embodiment, results may be stored in a data store as training data **3438** and used as (for example and without limitation) ground truth data for training. In at least one embodiment, when computing device **3408** sends extreme points for AI-assisted annotation **3010**, a deep learning model, for example, may receive this data as input and return inference results of a segmented organ or abnormality. In at least one embodiment, pre-instantiated annotation tools, such as AI-Assisted Annotation Tool **3436**B in FIG. **34**B, may be enhanced by making API calls (e.g., API Call **3444**) to a server, such as an Annotation Assistant Server **3440** that may include a set of pre-trained models **3442** stored in an annotation model registry, for example. In at least one embodiment, an annotation model registry may store pre-trained models **3442** (e.g., machine learning models, such as deep learning models) that are pre-trained to perform AI-assisted annotation on a particular organ or abnormality. In at least one embodiment, these models may be further updated by using training pipelines **3104**. In at least one embodiment, pre-installed annotation tools may be improved over time as new labeled clinic data **3012** is added.

[0506] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B.

Differentiable Neural Network Topology Search (DINTS)

[0507] Medical image segmentation faces some unique challenges like lacking manual labels and vast memory usage for processing 3D high-resolution images. Manually designed networks (e.g., U-Net) have been used for medical image segmentation tasks, but such manually designed networks are often not optimal. Diversity of these segmentation tasks could be extremely high since image characteristics and appearances can be completely distinct for different modalities and presentations of diseases can vary considerably. Automated medical image segmentation is essential for many clinical applications like finding new biomarkers and monitoring disease progression. Similar challenges exist for other neural network-based image processing tasks as well as neural network-based tasks performed on data other than image data. Different machine learning tasks have different optimal neural network architectures, and design of neural network architectures to provide optimal solutions is challenging.

[0508] Neural architecture search (NAS) focuses on designing a neural network automatically. NAS can be categorized into three dimensions: a search space, a search method, and performance estimation. A search space defines what architectures (e.g., which neural networks) can be searched, which can be further divided into a network topology level and cell level. A more flexible search space has more potential to contain better performing neural network architectures. Search method and performance estimation focus on finding an optimal architecture from a search space. Some existing NAS methods use evolutionary or reinforcement learning-based algorithms, which are too slow to be practicable (e.g., it can take 333 GPU days to search just one 3D segmentation network). Other existing NAS methods that are faster are only capable of finding single-path network topologies. A path may include an operation, upscaling and/or downscaling that connects two layers of a neural network. For a multi-path topology, two different layers are connected by at least two different types of operations and/or two different scales of inputs, whereas for a single-path topology different layers are connected by only a single operation and/or a single scale of inputs. For example, a multi-path topology may include at a layer a first input downsampled from a higher scale and a second input from a current scale. Limited single-path network topologies that can be generated from existing NAS techniques make a direct application of a successful network to a new task less likely to be optimal. In addition, existing NAS methods do not include any capability to specify an amount of memory to be used for performing a NAS search or an amount of memory to be used by a neural network having an automatically generated neural network architecture.

[0509] In at least one embodiment, a process or comprises one or more circuits to perform a differentiable network topology search of a neural network architecture search space to identify a neural network with a multi-path topology for an image-based task. In at least one embodiment, a processor comprises one or more circuits to cause one or more neural networks to be selected from a plurality of neural networks based, at least in part, on an amount of memory to be used by one or more neural networks. In at least one embodiment, a processor comprises one or more circuits to select a neural network architecture based on a plurality of connection patterns, each comprising a distinct combination of paths between layers of a neural network.

[0510] FIG. **35** is a visual representation of a topology search space **3502** with fully connected edges between adjacent layers for a differentiable NAS method **3500**, according to at least one embodiment. In at least one embodiment, topology search space **3502** include a first number, L, of layers (e.g., 12 layers) and each layer includes multiple feature nodes from a second number, D, of resolutions or scales (e.g., 4 resolutions), resulting in multiple candidate input edges **3521** between adjacent layers

(E=3D−2 candidate input edges). In at least one embodiment, edges **3521** or paths are represented by dashed lines between feature nodes **3523**. In at least one embodiment, a candidate input edge contains a cell operation and, in some cases, an upsample operation or a downsample operation (factor 2) is used before a cell operation if a candidate input edge is an upsample edge (represented as an upward facing arrow) or a downsample edge (represented as a downward facing arrow). In at least one embodiment, a feature node is a summation of output features such as feature maps from each candidate input edge. In at least one embodiment, as illustrated in FIG. **35**, a first feature node **3504** includes a first candidate input edge **3506** with a first cell operation, a second candidate input edge **3508** with a second cell operation, a third candidate input edge **3510** with a third cell operation, and a summation **3512** of output features from each candidate input edge. In at least one embodiment, first candidate input edge **3506** is a downsample edge and performs a downsample operation **3514**, second candidate input edge **3508** is a direct pass edge, and third candidate input edge **3510** is a upsample edge that performs an upsample operation **3516**. In at least one embodiment, cell operations are defined on candidate input edges. In at least one embodiment, feature nodes include feature maps. In at least one embodiment, edges in topology search space **3502** are selected for features to flow from input to output from a candidate network topology (also referred to as a candidate neural network architecture).

[0511] In at least one embodiment, each edge in topology search space **3502** includes a cell search space **3518** which contains a set of operations to select from. In at least one embodiment, cell search space **3518** is a set of operations where input and output feature maps have a same spatial resolution. In at least one embodiment, cell search space **3518** includes multiple blocks and connects among blocks that can be searched. In at least one embodiment, searched cells can be repeated over all cells in a network topology level. In at least one embodiment, differentiable NAS method **3500** searches operations of each cell independently with one or more operation selected from a set of operations. In at least one embodiment, for a NAS search performed for 3D images such as 3D medical images or other input feature vectors with similar dimensionality, a set of operations of a cell can include one or more types of operations, including: i) a skip connection operation in which no operation is performed; ii) a 3×3×3 3D convolution operation; iii) a pseudo 3D (P3D) 3×3×1 operation, which is a 3×3×1 convolution operation followed by a 1×1×3 convolution operation; iv) a P3D 3×1×3 operation, which is a 3×1×3 convolution operation followed by a 1'3×1 convolution operation; v) a P3D 1×3×3 operation, which is a 1×3×3 convolution operation followed by a 3×1×1 convolution operation. In at least one embodiment, other operations can be part of a set of operations of a cell. In at least one embodiment, a cell can also include an activation function (e.g., rectified linear unit (ReLU) activation) and instance normalization, which are used before and after those other operations respectively, except for a skip-connection operation. In at least one embodiment, feature spatial changes can be performed by upsampling and downsampling operations in edges searched at a topology level, allowing a set of operations to exclude some operations such as multi-scale feature aggregation operations like atrous convolution and pooling operations.

[0512] In at least one embodiment, as illustrated in FIG. **35**, topology search space **3502** contains input pre-defined operations **3520** that can receive an input image **3522** (e.g., a 3D or 2D image) and generate multiple multi-resolution images **3524** by downsampling input image **3522** by different scales (e.g., ½, ¼, ⅛) along each axis of said input image. In at least one embodiment, as illustrated in FIG. **35**, topology search space **3502** contains output pre-defined operations **3526** that can generate an output image **3528** by upsampling features from multiple resolutions from topology search space **3502**. In at least one embodiment, output pre-defined operations **3526** perform upsampling on multiple multi-resolution features from topology search space **3502**. In at least one embodiment, input pre-defined operations **3520** and output pre-defined operations **3526** are referred to as a stem that includes cell operations at various scales. In at least one embodiment, input image **3522** is a same scale and/or resolution as output image **3528**. In at least one embodiment, a first layer (layer 0) can receive a stack of multi-resolution images (e.g., a stack of images all generated from an initial input image, where each image in a stack has a different resolution) and perform convolutions (e.g., 3×3×3 3D convolutions with stride 2) to generate multi-resolution features for topology search space **3502**. In at least one embodiment, topology search space **3502** includes twelve layers, pre-defined operations for a first layer, and pre-defined operations for a last layer.

[0513] In at least one embodiment, differentiable NAS method **3500** performs a joint two-level search in a network topology level of search space **3502** to find a flow of feature maps across different spatial scales between layers and a cell level of search space **3518** to find one or more operations to perform at layers. In at least one embodiment, differentiable NAS method **3500**, to identify a candidate neural network architecture, selects input candidate edges in topology search space **3502** for features to flow from input to output and selects an operation from a set of operations of each cell associated with each input candidate edge selected. In at least one embodiment, differentiable NAS method **3500** can identify a variety of topologies, including a single-path topology and a multi-path topology, using a flexible search space that includes topology search space **3502** and cell search space **3518** for each input candidate edge between adjacent layers.

[0514] In at least one embodiment, processing logic performs a joint two-level search of a topology search space and a cell search space to identify one or more neural networks for an image-based task. In at least one embodiment, processing logic performs a joint two-level search of a topology search space and a cell search space to identify one or more neural networks for an image segmentation task (e.g., 3D image segmentation task). In at least one embodiment, a topology search space comprises a set of layers, each layer comprising a set of candidate feature nodes each at a different image scale for a first selection of one or more candidate feature nodes and/or paths or edges such as upsampling paths and downsampling paths between layers for one or more neural networks. In at least one embodiment, each candidate feature node comprises a set of candidate edges or paths that connect to a feature node in a previous layer.

[0515] In at least one embodiment, a cell search space comprises a set of candidate operations for a second selection of one or more of a set of candidate operations at each of a set of candidate edges or paths. In at least one embodiment, each of a set of candidate operations of a candidate feature node is to receive an input feature map and provides an output feature map after performing a candidate operation. In at least one embodiment, an input feature map and an output feature map have a same spatial resolution. In at least one embodiment, each candidate feature node is a summation of output features from each of a set of candidate edges. In at least one embodiment, during a joint two-level search, processing logic searches a topology search space with a set of input features, each at a different image scale, and searches a topology search space for a connection pattern between each adjacent layer of a set of layers, a connection pattern comprising one or more of a set of candidate edges. In at least one embodiment, a first candidate feature node of a set of candidate feature nodes comprises: a first candidate edge comprising a downsample operation and a first cell search space comprising a first set of candidate operations; a second candidate edge comprising a second cell search space comprising a second set of candidate operations; and a third candidate edge comprising an upsample operation and a third cell search space comprising a third set of candidate operations.

[0516] In at least one embodiment, a topology search space is a multi-scale search space and processing logic converts a multi-scale search space into a sequential search space comprising a super node for each respective layer of a set of layers. In at least one embodiment, each super node comprises a set of candidate feature nodes at a respective layer. In at least one embodiment, processing logic identifies a set of candidate connection patterns between a first layer and a second layer. In at least one embodiment, each of a set of candidate connection patterns comprises a different combination of one or more of a set of candidate edges between a first layer and a second layer. In at least one embodiment, processing logic determines a probability of each of a set of candidate connection patterns and selects a connection pattern, from a set of candidate connection patterns, based on probability. In at least one embodiment, processing logic selects a connection pattern based on probability, such as a connection pattern having a highest probability or another probability as a highest probability violates another constraint. In at least one embodiment, processing logic selects a connection pattern with a highest probability that also satisfies one or more other selection criteria. In at least one embodiment, one or more other selection criteria include a valid edges criterion. In at least one embodiment, a valid edges criterion is satisfied if both an input and an output of a path or edge connect to feature nodes. In at least one embodiment, a connection pattern between two layers includes multiple candidate edges **3521**. In at least one embodiment, a connection pattern between two layers includes a single candidate edge **3521**.

[0517] In at least one embodiment, differentiable NAS method **3500** can support a variety of topologies and can select input resolutions, such as illustrated in FIG. **36**. In at least one embodiment, more or fewer than 12 layers are searched to design a neural network. In at least one embodiment, more or fewer than four scales of inputs are searched to design a neural network.

[0518] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments of differentiable NAS method **3500**. In at least one embodiment, once NSA method **3500** is performed, a neural network having an optimized neural network architecture is generated. In at least one embodiment, inference and training logic **115** are used to perform inference and/or training operations on a generated neural network having an optimized neural network architecture. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B.

[0519] FIG. **36** illustrates a multi-path topology **3602**, a single-path topology **3604**, a multi-path topology with four input resolutions **3606**, and a multi-path topology with two input resolutions **3608**, according to at least one embodiment. In at least one embodiment, as described herein, a differentiable NAS method can perform a joint two-level search in a search space to identify a neural network architecture having any one of multi-path topology **3602**, single-path topology **3604**, multi-path topology with four input resolutions **3606**, and multi-path topology with two input resolutions **3608**.

[0520] FIG. **37** illustrates a search stage and a discretization stage of a differentiable NAS method **3700**, according to at least one embodiment. In at least one embodiment, instead of searching a discrete set of candidate architectures in a search space, differentiable NAS method **3700** performs a differentiable search where a discrete set of candidate architectures, as represented in search space **3702**, are relaxed into continuous model representations **3704** in a search stage **706**. In at least one embodiment, search stage **3706** results in a set of probabilities for combinations of edges or paths between feature nodes at different scales in layers. In at least one embodiment, relaxed continuous model representations **3704** allow gradient-based searching (e.g., gradient-descent searching) of search space **3702** in search stage **3706**. In at least one embodiment, probabilities of connecting different feature nodes at different scales of layers are computed and maintained in continuous model **3704**.

[0521] In at least one embodiment, after gradient-based searching in search stage **3706**, continuous model representations **3704** are converted back to a discrete architecture **3708** having distinct paths or edges between layers during a discretization stage **3710**. In at least one embodiment, during discretization stage **3710** one or more operations are selected for feature nodes and one or more paths or edges are selected for connecting feature nodes of different layers. In at least one embodiment, during discretization stage **3710**, differentiable NAS method **3700** selects a candidate operation from N candidate operations (e.g., O_1, O_2, . . . O_N) for each computational node and each operation (e.g., O_i) is paired with a first trainable parameter, where $\Sigma_{i=1}^{N} \propto_i O_i(x_{in})$, where x_{in} is an input feature. In at least one embodiment, a discrete operation is relaxed by a continuous model representation, $\propto_i$, which can be optimized using gradient descent. In at least one embodiment, after optimization, an operation O_i with a larger trainable parameter, $\propto_i$ will be selected. In at least one embodiment, a smaller trainable parameter, $\propto_j$, can still make a significant difference on an output feature, x_{out}, and following layers, as long as $\propto_j$ is not zero.

[0522] In at least one embodiment, during edge selection in discretization stage **3710**, each candidate input edge is paired with a second trainable parameter, β (where 0≥β≥1), and second parameters paired with candidate input edges that point to a same feature node sum to one. In at least one embodiment, after discretization stage **3710**, instead of keeping a single path and discarding candidate input edges with smaller trainable parameters, differentiable NAS method **3700** maintains non-zero candidate input edges for edge selection using a discretization algorithm that converts a continuous model to a discrete model and guarantees a topology (also referred to as a "topology guaranteed discretization algorithm") and a topology loss as described herein.

[0523] In at least one embodiment, if a differentiable NAS search uses a single-path constraint which assumes that only one input edge is kept for each node, a final searched model only has a single path from input to output which limits its complexity and can cause a sub-optimal discrete final architecture with a large performance gap, called a "discretization gap." In at least one embodiment, during search stage **306**, each candidate input edge in continuous model representations **3704** is given a probability and probabilities of candidate input edge to a node sum to one ($\beta_1+\beta_2+\beta_3=1$). In at least one embodiment, as illustrated in FIG. **37**, a discretization gap **3712** results from discarding some edges with small but non-zero probabilities in a searched continuous model representations **3704** during discretization stage **3710** and if a discretization algorithm uses a single-path topology constraint because edges causing infeasible topologies are not allowed even if they have large probabilities in continuous model representations **3704**. In at least one embodiment, without using a single-path constraint, differentiable NAS method **3700** finds a single path **3714** between layers of discrete model representations **3708**. In at least one embodiment, using a topology guaranteed discretization algorithm and a topology loss, differentiable NAS method **3700** mitigates discretization gap **3712**. In at least one embodiment, instead of extracting a single-path discrete model from a searched continuous model, differentiable NAS method **3700** extracts connection patterns that include one or more paths between layers including edges **3716** with non-zero probabilities in continuous model representations **3704**. In at least one embodiment, extracting connection patterns to support flexible topologies makes a direct application of a successful neural network to a new task more likely to be optimal.

[0524] In at least one embodiment, a differentiable NAS method **3700** relaxes a discrete architecture into continuous representations and allows a direct gradient-based search. In at least one embodiment, converting continuous representations back to a discrete architecture causes a "discretization gap" and can have a large memory usage during a search stage. In at least one embodiment, a searched model can be retrained while increasing a filter number, a batch size, or a patch size to gain better performance. In at least one embodiment for 3D medical image segmentation, a change of retraining scheme, which can be done when transferring to a new task that requires a larger input size, can still cause out-of-memory problems. In at least one embodiment, to solve a "discretization gap," differentiable NAS method **3700** uses a topology loss to push continuous representations close to binary as described in more detail below.

[0525] In at least one embodiment, instead of determining a score for each candidate input edge at a node and selecting a candidate input edge having a highest score for a single-path topology, differentiable NAS method **3700** considers all candidate input edges in sets of connection patterns between two layers as possibilities and selects an input connection pattern between each set of two layers. In at least one embodiment, differentiable NAS method **3700** considers multiple candidate input edges (paths) and/or combinations of candidate input edges between two layers as being possible, where each candidate input edge represents a connection between a first feature node at an image scale (e.g., full scale, ½ scale, a ¼ scale, a ⅛ scale and a 1/16 scale) at a first layer and a second feature node at an image scale at a second layer. In at least one embodiment, a first feature node is at a first image scale and a second feature node is at a second image scale that is different from first image scale. In at least one embodiment, a first feature node and a second feature node are at a same image scale. In at least one embodiment, differentiable NAS method **3700** generates a set of connection patterns, where there is a separate connection pattern for each possible combination of candidate input edges (paths) that can connect two layers. In at least one embodiment, differentiable NAS method **3700** scores each connection pattern, as opposed to scoring individual candidate input edges, and this is repeated for each pair of adjacent layers. In at least one embodiment, differentiable NAS method **3700** selects a connection pattern that together has a highest combined score and that results in valid connections. In at least one embodiment, a valid connection occurs when a feature node has an input edge and an output edge. In at least one embodiment, differentiable NAS method **3700** tests multiple connection patterns during a search stage to automatically discover a neural network architecture for an image-based task, such as 2D or 3D image segmentation, image classification, or similar tasks.

[0526] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments of differentiable NAS method **3700**. In at least one embodiment, inference and/or training logic **115** are used to perform inferencing and/or training on a neural network generated according to differentiable NAS method **3700**. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B.

[0527] FIG. **38** illustrates a differentiable NAS method **3800** with a sequential model **3804** with super nodes, according to at least one embodiment. In at least one embodiment, differential NAS method **3800** converts a multi-scale search space **3802** into a sequential model **3804** using super nodes. In at least one embodiment, feature nodes at a same layer, i, are combined as a super node, s_i. In at least one embodiment, for multi-scale search space **3802** with L layers and D resolution levels (image scales), feature nodes in a same layer, i, are combined as super nodes and features flow sequentially from these L super nodes as shown with super nodes **3806** (s_0), **3808** (s_1), **3810** (s_{L-1}), **3812** (s_L). In at least one embodiment, there are a number of candidate input edges to each super node E (E=3D−2), and a network topology search by differentiable NAS method **3800** selects an optimal set of input edges for each super node. In at least one embodiment, a connection pattern is a set of selected edges and there can be a number of feasible candidate connection patterns, M($M=2^E-1$). In at least one embodiment, a j-th connection pattern, cp_j, is an indication vector of length E, where $cp_j(e)=1$, if e-th edge is selected in j-th pattern.

[0528] In at least one embodiment, differentiable NAS method **3800** performs a network topology search using an input connection operation. In at least one embodiment, an input connection operation is defined to s_i with connection pattern cp_j as c_j^i, where cp_j defines a network topology of that also includes cell operations on selected edges in cp_j and c_j^i and c_k^{i+1} identify input/output connection patterns for a super node, s_i, as cp_j, cp_k respectively. In at least one embodiment, using this formulation, differentiable NAS method **3800** performs a network topology search by selecting an input connection pattern for each super node and competition is among all M connection patterns, not among candidate input edges. In at least one embodiment, differentiable NAS method **3800** can associate a variable n_j^i to each connection operation c_j^i for every super node, s_i, and every input connection pattern j. In at least one embodiment, input features at layer **0** are denoted as s_0 and a sequential feature flow equation is expressed in equation (1) below.

$$s_i = \sum_{j=1}^{M} (\eta_j^i * c_i^i(s_{i-1}))i = 1, 2, \quad \ldots, L \tag{1}$$

$$\sum_{j=1}^{M} \eta_j^i *= 1, n_j \geq 0 \forall\, i,\ j$$

In at least one embodiment, differentiable NAS method **3800** parametrizes variable η_j^i with a set of edge probability parameters (p_e^i), e=1, . . . E, since M is growing exponentially with D, as expressed in equation (2) below.

$$\frac{\prod_{e=1}^{E} (1 - p_e^{i^{1-cp_j(e)}} (p_e^i)^{cp_j(e)}}{\sum_{j=1}^{M} \prod_{e=1}^{E} (1 - p_e^{i^{1-cp_j(e)}} (p_e^i)^{cp_j(e)}} \tag{2}$$

$$0 \leq p_e^i \leq 1 \forall\, i,\, e$$

In at least one embodiment, for an example with a search space with 12 layers (L=12) and four resolution levels (D=4), a network topology parameter number is reduced from M×L (1023×12) to E×L (10×12). In at least one embodiment, under this formulation, probabilities n of connections are highly correlated such that if an input edge, e, to a super node, si, has a lower probability, all candidate patterns to a super node, si, with a selected input edge, e, will have lower probabilities. In at least one embodiment, as illustrated in FIG. **38**, a set of selected edges **3814** (solid lines as opposed to dashed lines for non-selected edges) that connects a first super node, s_{i-1}, and a second super node, s_i is a selected candidate connection pattern. In at least one embodiment, a topology search selects one candidate connection pattern to connect adjacent super nodes sequentially.

[0529] In at least one embodiment, differentiable NAS method **3800** uses a cell operation relaxation as described above for a cell level search. In at least one embodiment, each cell on a candidate input edge, e, to a given super node, si, has its own cell architecture parameters, $\alpha_1^{i,e}$, $\alpha_2^{i,e}$, . . . , $\alpha_N^{i,e} \propto_i$ and will be optimized during a search stage. In at least one embodiment, a variable c_j^i in equation (1) above contains cell operations defined on selected edges and contains relaxed cell architecture parameters α, allowing a gradient-based search for topology and cell levels jointly.

[0530] In at least one embodiment, once a continuous model is optimized in a search stage, a final discrete architecture is derived from an optimized continuous architecture representation n, which is derived from a set of edge probability parameters (p_e^i) and cell architecture probability parameters, α. In at least one embodiment, a variable η_i^j represents a probability of using an input connection pattern cphd j^i for a super node s_i. In at least one embodiment, an easiest way for topology discretization is to select an input connection pattern cp_j^i with a maximum probability η_i^j, since a network topology search space is converted into a sequential space. In at least one embodiment, a topology infeasibility is considered since a particular topology may not be feasible. In at least one embodiment, topology infeasiability occurs when a feature node has an input edge but no output edge or has an output edge but no input edge, such as illustrated in FIG. **39**.

[0531] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments of differentiable NAS method **3800**. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B.

[0532] FIG. **39** illustrates a differentiable NAS method **3900** that discretizes a sequential model **3902** with topology feasibility constraints, according to at least one embodiment. In at least one embodiment, differential NAS method **3900** evaluates topology feasibility of a sequential model **3902**. In at least one embodiment, as illustrated in FIG. **39**, a super node **3904** includes two of three nodes that are infeasible topologies as a first feature node **3906** has an input edge **3908** but no output edge, and a second feature node **3910** has an output edge **3912** but no input edge. These edges or paths are therefore invalid. In at least one embodiment, as illustrated in FIG. **39**, a super node **3904** includes a third feature node **3914** that is a feasible topology since it has an input edge **3916** and an output edge **3918**. In at least one embodiment, for every connection pattern cp_j, a feasiable set, F(j) is generated. In at least one embodiment, if a super node with an input pattern, j, and output pattern, k, is topologically feasible, then k ∈ F(j). In at least one embodiment, differentiable NAS method **3900** can denote an array of selected input connection pattern indexes for L super nodes as I and can perform a topology discretization by sampling an index I from its probability distribution p(I) using a maximum likelihood (or minimizing a negative log-likelihood), as expressed in equations (3) and (4) below.

$$p(I) = \begin{cases} \prod_{i=1}^{L} n_i^{I(i)}, \forall\, i{:}I(i+1) \in F(I(i)) \\ 0, \text{ else} \end{cases} \tag{3}$$

$$I = \text{argmin} \sum_{i=1}^{L} = -\log(\eta_i^{I(i)}), \forall\, i{:}I(i+1) \in F(I(i)) \tag{4}$$

In at least one embodiment, differentiable NAS method **3900** builds a directed graph **3920**, G, using parameters η and a feasibility set F, where a node c_j^i is connected with an adjacent node c_k^{i+1} and c_m^{i-1} if adjacent nodes are feasible (if j ∈ F(M) and k ∈ F(j)), as illustrated in FIG. **39**. In at least one embodiment, directed graph **3920**, G, contains a number of nodes **3922** (L×M+2), which represent connection operations and an input edge cost to a node c_j^i in terms of a negative log-likelihood—log(n_j^i). In at least one embodiment, a source **3924** connects to all first layer nodes **3922** and all L-th layer nodes **3922** connect to a sink **3926** with an edge cost of a constant value (e.g., 0.001). In at least one embodiment, L nodes **3922** on a shortest path **3928** from source **3924** to sink **3926** in directed graph **3920**, G, represents an optimal final architecture with feasible connection operations. In at least one embodiment, differentiable NAS method **3900** uses equation (4) to determine a path with a minimum cost from source **3924** to sink **3924** and an optimal index, I, using Dijkstra algorithm. In at least one embodiment, differentiable NAS method **3900** selects an operation with a largest cell architecture parameters a for cell operations on selected edges from an optimal index, I.

[0533] In at least one embodiment, to minimize a discretization gap between a continuous model representation and a final discretized model representation (final architecture), entropy losses can be used by differentiable NAS method **3900** to encourage binarization of cell architecture parameters a and parameters η of input connections, as set forth in equations (5).

$$Ł_{\propto} = \frac{-1}{L*E*N}\sum_{l=1}^{L}\sum_{e=1}^{E}\sum_{n=1}^{N}\propto_n^{i,e}*\log(\propto_n^{i,e}) \tag{5}$$

$$Ł_{\eta} = \frac{-1}{L*M}\sum_{i=1}^{L}\sum_{j=1}^{M}\eta_j^i*\log(\eta_j^i)$$

[0534] In at least one embodiment, even with architecture parameters ∝ and η being almost binarized, there may still be a large gap due to topology constraints in a discretization algorithm, since topology feasibility indicates that an activated feature node with at least one input edge must have an output edge while an in-activated feature node cannot have an output edge. In at least one embodiment, each super node has D feature nodes, thus there are 2^D-1 node activation patterns. In at least one embodiment, a variable A is defined as a set of all node activation patterns and each element of a set (a ∈ A) is an indication function of length D, where a(i)=1 of an i-th node of a super node is activated. In at least one embodiment, differentiable NAS method **3900** defines two feasibility sets, $F_{in}(a)$ and $F_{out}(a)$, representing all feasible input and output connection pattern indexes for a super node with node activation, a, as shown in FIG. **40**.

[0535] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments of differentiable NAS method **3900**. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B.

[0536] FIG. **40** illustrates a differentiable NAS method **4000** that minimizes a discretization gap, according to at least one embodiment. In at least one embodiment, as illustrated in FIG. **40**, input feasibility set **4002** and output feasibility set **4004** represent feasible input and output connection pattern indexes for a super node with a node activation, according to at least one embodiment. In at least one embodiment, connection patterns in input feasibility set **4002** with activation pattern a (a=[0,1,1]) and all feasible output connection patterns are in output feasibility set **4004**. In at least one embodiment, a pattern a=[0,1,1] indicates that two of three nodes of a super node are activated. In at least one embodiment, differentiable NAS method **4000** uses a topology loss function that can be minimized to encourage all super nodes to be topologically feasible in a search stage that is aware of topology constraints, thereby reducing a discretization gap caused by topology constraints in a discretization stage. In at least one embodiment, differentiable NAS method **4000** uses a topology loss as expressed in equations (6) and (7) below.

$$p_{in}^i(a) = \sum_{j\in F_{in}(a)}\eta_j^i \tag{6}$$

$$p_{out}^i(a) = \sum_{j\in F_{out}(a)}\eta_j^{i+j}$$

$$Ł_{tp} = -\sum_{i=1}^{L-1}\sum_{a\in A}(p_{in}^i(a)\log(p_{out}^i(a)) + 1 - p_{in}^i(a))\log(1 - p_{out}^i(a)), \tag{7}$$

where $p_{in}{}^i(a)$ is a probability that an activation pattern for a super node, s_i, is activation pattern, a, and $p_{out}{}^i(a)$ is a probability that a super node, s_i, with activation pattern a is feasible. In at least one embodiment, by minimizing $Ł_{tp}$, a search stage of differentiable NAS method **4000** is aware of topology constraints and encourages all super nodes to be topologically feasible, thus reducing a discretization gap caused during a discretization step.

[0537] In at least one embodiment, a searched model is retrained under different training settings like patch size (e.g., 4× larger image patch), filter number (e.g., 6× more filters), or tasks, which can cause out of memory problems for 3D images during re-training. In at least one embodiment, a differentiable NAS method can use a memory budget mechanism in a search stage for a neural network architecture. In at least one embodiment, a cell's expected memory usage is estimated by $M^{i,e}=\Sigma_{n=1}^{N}\propto_n{}^{i,e}M_n$, where M_n is memory usage of operation O_n (e.g., an estimated tensor size). In at least one embodiment, an expected memory usage M_e of a searched model can be represented in equation (8) below.

$$M_e = \sum_{i=1}^{L}\sum_{j=1}^{M}\eta_j^i*(\sum_{e=1}^{E}\left(\sum_{n=1}^{N}M_n\right)*cp_j(e), \tag{8}$$

[0538] In at least one embodiment, differentiable NAS method **4000** uses a budget as a percentage σ of a maximum memory usage M_a, of which all parameters ∝ and η equal to one, as set forth in equations (9) and (10) below.

$$M_a = \sum_{i=1}^{L}\sum_{j=1}^{M}*(\sum_{e=1}^{E}\left(\sum_{n=1}^{N}M_n\right)*cp_j(e) \tag{9}$$

$$Ł_m = \left|\frac{M_e}{M_a} - \sigma\right|_1 \tag{10}$$

[0539] In at least one embodiment, a differentiable neural network topology search uses an optimization strategy in which a training set is partitioned into a first training subset, train**1**, and a second training subset, train**2**, and network weights, w (e.g., convolution kernels) are optimized using a segmentation loss optimization algorithm, L_{seg}, on trainl, and network architecture weights, ∝ and η using an archirecture optimization algorithm, L_{arch}, on train**2** alternatively. In at least one embodiment, a segmentation loss optimization, Lseg, for weight w is an even sum of dice and cross-entropy loss in segmentation, while t and tall are current and total iterations for architecture optimization such that searching is focused more on segmentation loss L_{seg} at a starting point, as set forth in equation (11) below.

$$L_{arch} = L_{seg} + \frac{t}{t_{all}} * (L_{\propto} + L_{\eta} + \lambda * L_{tp} + L_{m}) \tag{11}$$

[0540] In at least one embodiment, a topology loss, L_{tp} can be empirically scaled to a same range of other losses by using a multiplier, λ, such as 0.001 (e.g., λ=0.001).

[0541] In at least one embodiment, a differentiable neural network topology search, as described herein, can be performed for an image-based task, such as medical image segmentation tasks. In at least one embodiment, medical image segmentation tasks can cover different anatomies of interest, modalities, and imaging sources. In at least one embodiment, for example, a differentiable neural network topology search can be performed on an annotated medical image dataset and can deploy a searched model on one or more medical image segmentation tasks. In at least one embodiment, images are sampled to have a 1.0×1.0×1.0 mm^3 voxel resolution. In at least one embodiment, a differentiable neural network topology search, as described herein, can be performed for a medical image segmentation task for each of a pancreas, a liver, a lung, a brain, a colon, a heart, a prostate, a spleen, or other anatomies.

[0542] In at least one embodiment, a search space includes L=12 layers and D=4 resolution levels, such as illustrated in FIG. **35**. In at least one embodiment, a stem cell at a first scale (scale 1) has sixteen filters, and a number of filters are doubled for each decrease in spatial size (e.g., by half) in each axis, for example, a second scale (scale ½) has thirty-two filters. In at least one embodiment, a differentiable neural network topology search is performed on a training dataset following a 5-fold data split with 4 for training and a last set for validation. In at least one embodiment, a stochastic gradient descent (SGD) optimizer can be used, such as with momentum set to 0.9 and weight decay of $4e^{-5}$ set for network weights, w. In at least one embodiment, weight w is trained for a first set of warm-up iterations (e.g., 1K) and followed by a second set of iterations (e.g., 10K iterations) without updating architecture. In at least one embodiment, architecture parameters ∝ and η are initialized with Gaussian N(1, 0.01), N(0, 0.01) respectively before softmax and sigmoid. In at least one embodiment, in a following set of iterations (e.g., 10K iterations), weight w is jointly optimized with SGD optimizer and architecture parameters ∝ and η with an Adam optimizer (e.g., learning rate 0.008, weight decay 0). In at least one embodiment, a learning rate of SGD linearly increases from 0.025 to 0.2 in a first set of iterations and decays with a factor of 0.5 at following sets of iterations (e.g 8K, 16K). In at least one embodiment, a search is conducted on a set of GPUs (e.g., 8 GPUs) with a batch size (e.g., 8) (e.g., each GPU with one 96×96×96 patch). In at least one embodiment, patches are randomly augmented with 2D rotation by 90, 180, and 270 degrees in an x-y plane and flip in all three axes. In at least one embodiment, in a search stage, a discretized model is identified.

[0543] In at least one embodiment, after a search stage, a discretized model is retrained with doubled filter number and doubled batch size. In at least one embodiment, an SGD optimizer can be used with a first set of warm-up iterations (e.g, 1K) and a set of training iterations (e.g., 40K) and a decay learning rate by a factor of 0.5 at sets of iterations (e.g., 8K, 16K, 24K, 32K) after warm-up iterations. In at least one embodiment, a learning rate schedule can be similar to a start stage in a warm-up and a first set of iterations (e.g., 20K). In at least one embodiment, during a latter set of iterations (e.g., 20K) can be used for better convergence. In at least one embodiment, to test generalizability of a searched model, a searched model can be retrained on other image segmentation tasks. In at least one embodiment, a differentiable neural network topology search, as described herein, can be performed in less than 333 GPU days, such as 5.8 GPU days.

[0544] In at least one embodiment, a differentiable neural network topology search, as described herein, can be performed with varying memory constraints, allowing searched models to have highly flexible topologies given different memory constraints and allowing searched models to be searched jointly at a network topology level and a cell level. In at least one embodiment, increasing a memory constraint results in a searched model being denser in connection and can achieve better performance while requiring more GPU memory, such as illustrated in searched architectures with different memory constraints of FIGS. **41**A-**41**C. In at least one embodiment, a memory constraint for a neural network and/or a NAS search based on known properties of a device that will perform inferences using a neural network and/or on properties of a device that will perform an NAS search. Accordingly, in at least one embodiment, if an available memory of an edge device that will execute a neural network is known, then that available memory may be used to set a memory constraint used to search for a neural network. In at least one embodiment, memory, such as GPU memory, can be estimated using functions in training, as described above. In at least one embodiment, a differentiable neural network topology search, as described herein, can be performed with different memory constraints to evaluate performance, allowing a reduction in memory usage with an acceptable level of accuracy. In at least one embodiment, a memory-aware search of a differentiable neural network topology search, as described herein, can be used to further reduce memory usage from a NAS perspective. In at least one embodiment, this can result in a reduction in memory used to perform a NAS and/or a reduction in memory used by a trained neural network selected based on a result of a NAS.

[0545] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments of differentiable NAS method **4000**. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B.

[0546] FIGS. **41**A-**41**C illustrates three searched architectures **4100**, **4110**, **4120** with different memory constraints, according to at least one embodiment. In at least one embodiment, a differentiable neural network topology search can identify a first neural network architecture **4100** in FIG. **41**A using a first memory constraint (e.g., σ=0.8), a second neural network architecture **4110** in FIG. **41**B using a second memory constraint (e.g., σ=0.5), and a third neural network architecture **4120** in FIG. **41**C using a third memory constraint (e.g., σ=0.2). In at least one embodiment, as illustrated in FIGS. **41A-41C**, increasing a memory constraint can increase a number of connections between layers in a neural network, leading to better performance in an image-based task but also leading to increased memory usage.

[0547] In at least one embodiment, a discretization algorithm of a differentiable neural network topology search, as described herein, do not discard topologically infeasible edges (most importantly edges with large probabilities), reducing a discretization gap between feature flow in an optimized continuous model, such as set forth in equation (1), and a discrete model. In at least one embodiment, a differentiable neural network topology search, as described herein, includes a topology loss that encourages connections with large probabilities to be feasible, thus will not be discarded and causing a gap, such as illustrated in FIG. **42** below.

[0548] FIG. **42** is a graph illustrating indications in discretization gaps with and without topology loss in a discretization algorithm under different memory constraints, according to at least one embodiment. In at least one embodiment, a discretization algorithm without a topology loss results in a first indication of discretization gap **4202** during an architecture search with a first memory constraint (e.g., σ=0.8). In at least one embodiment, a discretization algorithm with a topology loss, as described herein, results in a second indication of discretization gap **4204** with a same first memory constraint (e.g., σ=0.8) that is lower than first indication of discretization gap **4202**. In at least one embodiment, a discretization algorithm without a topology loss results in a third indication of discretization gap **4206** during an architecture search with a second memory constraint (e.g., σ=0.5). In at least one embodiment, a discretization algorithm with a topology loss, as described herein, results in a fourth indication of discretization gap **4208** with a same second memory constraint (e.g., σ=0.5) that is lower than third indication of discretization gap **4206**. In at least one embodiment, a discretization algorithm without a topology loss results in a fifth indication of discretization gap **4210** during an architecture search with a third memory constraint (e.g., σ=0.2). In at least one embodiment, a discretization algorithm with a topology loss, as described herein, results in a sixth indication of discretization gap **4212** with a same third memory constraint (e.g., σ=0.2) that is lower than fifth indication of discretization gap **4210**. In at least one embodiment, as illustrated in FIG. **42**, a topology loss (dashed lines) is decreased compared to no topology loss (solid lines), demonstrating an importance of topology loss in a discretization algorithm as described herein.

[0549] In at least one embodiment, a differentiable neural network topology search, as described herein, can implement a topology loss by denoting i) C_{max} as a topology decoded by selecting a connection j with a largest probability η_j^i for each layer i, even if infeasible, and ii) C_{top} as a topology decoded by a discretization algorithm of a differentiable neural network topology search as described herein. Topologies C_{max} and C_{top} are indication matrices of size [L,E] representing whether an edge is selected and an indication of discretization gap, G, is expressed as $G=\Sigma_{i=1}^{L}\Sigma_{e=1}^{E}\lfloor C_{max}(i, e)-C_{top}(i, e)\rfloor$. In at least one embodiment, as illustrated in FIG. **42**, larger G represents a larger gap between feature flows before and after discretization. In at least one embodiment, as illustrated in FIG. **42**, with topology loss, a gap between Cmax and Ctop is reduced, which is more important for smaller memory constraints (e.g., smaller σ) where a searched architecture is more sparse and more likely to have topology infeasibility.

[0550] In at least one embodiment, a differentiable network topology search framework (DiNTS) can be used for 3D medical image segmentation tasks. In at least one embodiment, by converting feature nodes with varying spatial resolutions into super nodes, operations can focus on connection patterns rather than individual edges, enabling more flexible network topologies and a discretization-aware search framework. In at least one embodiment, a differentiable network topology search framework (DiNTS) can discover highly flexible, high-performance neural network architectures for medical image segmentation tasks, as well as other image-based tasks. In at least one embodiment, medical images can be 3D medical images and a differentiable network topology search framework (DiNTS) can use a memory constraint to identify different neural network architectures using different memory constraints during a search.

[0551] In at least one embodiment, one or more circuits perform a new NAS method for automatically searching a NAS search space to discover an optimal architecture for a neural network with a multi-path topology for an image-based task like medical image segmentation. In at least one embodiment, a new NAS method, called Differentiable Neural Network Topology Search (DiNTS) performs a differentiable network topology search that separately searches for an optimal network topology and for optimal cell operations to discover a neural network architecture for an image-based task. In at least one embodiment, a NAS search space for segmentation contains a network topology level and a cell level. In at least one embodiment, a network topology level controls connections among cells and a flow of feature maps across different spatial scales, and a cell level controls specific operations on feature maps within each cell. In at least one embodiment, a differentiable search scheme is capable of performing a joint two-level search in a network topology level and a cell level in a NAS search space and finding multi-path topologies with reduced use of processor resources. In at least one embodiment, resulting architectures with multi-path topologies can have better performance than an architecture with a single path from input to output. In at least one embodiment, a differentiable NAS search uses a topology guaranteed discretization algorithm and a discretization aware topology loss for a search stage to minimize a discretization gap. In at least one embodiment, a differentiable search scheme can use a memory-usage constraint during searching to discover neural networks for different memory requirements. In at least one embodiment, a memory usage criterion can be set and a search can be performed in a manner that does not exceed a memory usage criterion. In at least one embodiment, a memory usage aware search method is able to search 3D networks with different GPU memory requirements.

[0552] FIG. **43** is a flow diagram of a method **4300** of performing a search of a search space, according to at least one embodiment. In at least one embodiment, method **4300** is performed by processing logic comprising hardware, firmware, software, or any combination thereof. In at least one embodiment, method **4300** begins by processing logic determining a maximum memory usage of all operations associated with one or more candidate feature nodes of search space (block **4302**). In at least one embodiment, processing logic receives a first input that specifies a first memory constraint (block **4304**). In at least one embodiment, a first memory constraint is specified as a first percentage of a maximum memory usage. In at least one embodiment, processing logic selects a first set of one or more operations associated with one or more candidate feature modes for one or more neural networks to be selected (block **4306**). In at least one embodiment, a first memory usage of a first set of one or more operations is equal to or less than a first percentage of a maximum memory usage; and method **4300** ends.

[0553] In at least one embodiment, method **4300** includes performing a second search of a search space by receiving a second input that specifies a second memory constraint as a second percentage of a maximum memory usage. In at least one embodiment, processing logic selects a second set of one or more operations associated with one or more candidate feature modes for one or more neural networks to be selected. In at least one embodiment, a second memory usage of a second set of one or more operations is equal to or less than a second percentage of a maximum memory usage.

[0554] In at least one embodiment, method **4300** includes performing a third search of a search space by receiving a third input that specifies a third memory constraint as a third percentage of a maximum memory usage. In at least one embodiment, processing logic selects a third set of one or more operations associated with one or more candidate feature modes for one or more neural networks to be selected. In at least one embodiment, a second memory usage of a second set of one or more operations is equal to or less than a second percentage of a maximum memory usage.

[0555] In at least one embodiment, a first set of one or more operations, a second set of one or more operations, and a third set of one or more operations are different, such as illustrated in FIGS. **41**A-**41**C based on a first memory constraint, a second memory constraint, and a third memory constraint, respectively.

[0556] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments of method **4300**. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B.

[0557] FIG. **44** is a flow diagram of a method **4400** of performing multiple searches of a search space, according to at least one embodiment. In at least one embodiment, method **4300** is performed by processing logic comprising hardware, firmware, software, or any combination thereof. In at least one embodiment, method **4400** begins by processing logic receiving input data that specifies a first memory constraint (block **4402**). In at least one embodiment, processing logic determines a first set of one or more search parameters based on a first memory constraint. In at least one embodiment, processing logic performs a first search in accordance with a first set of one or more search parameters (block **4404**). In at least one embodiment, first results of a first search for one or more neural networks include a first set of one or more operations associated with one or more candidate feature nodes of a search space that satisfy a first memory constraint. In at least one embodiment, method **4400** receives input data that specifies a second memory constraint (block **4406**). In at least one embodiment, processing logic determines a second set of one or more search parameters based on a second memory constraint (block **4408**). In at least one embodiment, processing logic performs a second search in accordance with a second set of one or more search parameters. In at least one embodiment, second results of a second search for one or more neural networks include a second set of one or more operations associated with one or more candidate feature nodes of a search space that satisfy a second memory constraint. In at least one embodiment, processing logic performs a search of a search space using a topology-guaranteed discretization algorithm and a discretization-aware topology loss.

[0558] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments of method **4400**. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B.

[0559] FIG. **45** is a flow diagram of a method **4500** of performing a search of a search space, according to at least one embodiment. In at least one embodiment, method **4500** is performed by processing logic comprising hardware, firmware, software, or any combination thereof. In at least one embodiment, method **4500** begins by processing logic identifying multiple candidate connection patterns between a first layer and a second layer of a search space, a first layer and a second layer each including a set of candidate feature nodes at a different image scale (block **4502**). In at least one embodiment, each candidate feature node includes multiple candidate edges that connect to a feature node in a previous layer. In at least one embodiment, processing logic determines a probability of each of a set of candidate connection patterns (block **4504**) and selects a connection pattern, from a set of candidate connection patterns, based on probability (block **4506**); and method **4500** ends. In at least one embodiment, at block **4506** a candidate connection pattern having a highest probability is selected. In at least one embodiment, one or more criteria are used to select a candidate connection pattern for connecting two layers. In at least one embodiment, one or more candidate connection patterns are discarded based on invalid connections. For example, candidate connection patterns that connect to inactive feature nodes of either a first layer or a second layer may be discarded. In at least one embodiment, a remaining candidate connection pattern with a highest probability after discarding any candidate connection patterns with invalid connections is selected.

[0560] In at least one embodiment, processing logic performs a search by identifying a second set of candidate connection patterns between a second layer and a third layer of a search space, a third layer including a set of candidate feature nodes at a different image scale. In at least one embodiment, processing logic determines a feasible set of candidate connection patterns, from a set of candidate connection patterns based on whether a respective candidate feature node includes an input connection in a set of candidate connection patterns and an output connection in a second set of candidate connection patterns. In at least one embodiment, processing logic selects, from a feasible set of candidate connection patterns, a connection pattern having a highest probability.

[0561] In at least one embodiment, processing logic performs a joint two-level search of a search space by identifying a connection pattern, from a set of connection patterns, between each of a set of layers in a search space for one or more neural networks to be selected, wherein each of a set of layers includes a set of candidate feature nodes at a different image scale, wherein each candidate feature node includes a set of candidate edges that connect to a feature node in a previous layer. In at least one embodiment, each of a set of connection patterns includes a different combination of candidate edges. In at least one embodiment, processing logic identifies, for each of a respective candidate edges in a connection pattern, an operation from a set of candidate operations for one or more neural networks to be selected.

[0562] Inference and/or training logic **115** are used to perform inferencing and/or training operations associated with one or more embodiments of method **4500**. Details regarding inference and/or training logic **115** are provided herein in conjunction with FIGS. **1**A and/or **1**B.

[0563] At least one embodiment of the disclosure can be described in view of the following clauses:

[0564] In clause 1, a processor comprising: one or more circuits to cause one or more neural networks to be selected from a plurality of neural networks based, at least in part, on an amount of memory to be used by the one or more neural networks.

[0565] In clause 2, a processor of clause 1, wherein the one or more circuits are further to perform a search to select the one or more neural networks from the plurality of neural networks that satisfy a memory constraint.

[0566] In clause 3, a processor of any of clauses 1-2, wherein the one or more circuits are further to cause one or more additional neural networks to be selected from the plurality of neural networks or from a second plurality of neural networks, wherein the one or more additional neural networks are different from the one or more neural networks and satisfy a second memory constraint that is different from a first memory satisfied by the one or more neural networks.

[0567] In clause 4, a processor of any of clauses 1-3, wherein the one or more circuits are further to perform a search to select the one or more neural networks from the plurality of neural networks in accordance with a set of one or more search parameters determined at least in part on the amount of memory to be used by the one or more neural networks.

[0568] In clause 5, a processor of any of clauses 1-4, wherein a percentage of a maximum memory usage of operations associated with one or more candidate feature nodes of a search space comprising the plurality of neural networks is less than or equal to the amount of memory.

[0569] In clause 6, a processor of any of clauses 1-5, wherein the one or more circuits further cause the one or more neural networks to be selected by performing a joint two-level search of a topology search space and a cell search space to identify the one or more neural networks for an image-based task.

[0570] In clause 7, a system comprises: one or more processors to cause one or more neural networks to be selected from a plurality of neural networks based, at least in part, on an amount of memory to be used by the one or more neural networks; and one or more memories to store parameters corresponding to the one or more neural networks.

[0571] In clause 8, a system of clause 7, wherein the one or more processors are further to perform a first search to select the one or more neural networks from the plurality of neural networks that satisfy a first memory constraint.

[0572] In clause 9, a system of any of clauses 7-8, wherein the one or more processors are further to select a connection pattern, from a plurality of candidate connection patterns between a first layer and a second layer of the one or more neural networks, based at least in part on probabilities of each of the plurality of candidate connection patterns.

[0573] In clause 10, a system of any of clauses 7-9, wherein the one or more processors are further to select a feature node from a set of candidate features nodes for one or more layers of the one or more neural networks, wherein the set of candidate feature nodes comprises feature nodes at different image scales that comprise a plurality of candidate edges that connect to a feature node in a previous layer.

[0574] In clause 11, a system of any of clauses 7-10, wherein the one or more processors further cause the one or more neural networks to be selected by performing a joint two-level search of a topology search space and a cell search space to identify the one or more neural networks for an image-based task.

[0575] In clause 12, a system of any of clauses 7-11, wherein the one or more neural networks are to perform an image segmentation task.

[0576] In clause 13, a system of any of clauses 7-12, wherein the one or more processors further cause the one or more neural networks to be selected by performing a search of a topology search space comprising a plurality of candidate edges that connect candidate feature nodes of a plurality of layers and a cell search space comprising a plurality of candidate operations.

[0577] In clause 14, a system of any of clauses 7-13, wherein the one or more processors are further to select a connection pattern, from a plurality of candidate connection patterns between layers of the one or more neural networks, based at least in part on probabilities of the plurality of candidate connection patterns.

[0578] In clause 15, a datacenter comprising: one or more processors to cause one or more neural networks to be selected from a plurality of neural networks to perform a medical image segmentation task based, at least in part, on an amount of memory to be used by the one or more neural networks.

[0579] In clause 16, a datacenter of clause 15, wherein the one or more processors are further to perform a search to select the one or more neural networks from the plurality of neural networks that satisfy a memory constraint.

[0580] In clause 17, a datacenter of any of clauses 15-16, wherein the one or more processors are further to cause one or more additional neural networks to be selected from the plurality of neural networks or from a second plurality of neural networks, wherein the one or more additional neural networks are different from the one or more neural networks and satisfy a second memory constraint that is different from a first memory satisfied by the one or more neural networks.

[0581] In clause 18, a datacenter of any of clauses 15-17, wherein the one or more processors are to perform a joint two-level search of a topology search space and a cell search space to identify the one or more neural networks for the medical image segmentation task.

[0582] In clause 19, a datacenter of any of clauses 15-18, wherein the one or more processors are to perform a search of a search space to cause the one or more neural networks to be selected, wherein the search comprises selecting a connection pattern between layers of the one or more neural networks, from a plurality of candidate connection patterns, based at least in part on probabilities of the plurality of candidate connection patterns.

[0583] In clause 20, a datacenter of any of clauses 15-19, wherein the one or more processors are to perform a search of a search space to cause the one or more neural networks to be selected, wherein the search comprises selecting a connection pattern from a feasible set of candidate connection patterns between layers of the one or more neural networks, wherein each feasible connection pattern in the feasible set of candidate connection patterns comprises valid input connections and output connections between the layers.

[0584] In clause 21, a method comprising: selecting one or more neural networks from a plurality of neural networks based, at least in part, on an amount of memory to be used by the one or more neural networks.

[0585] In clause 22, a method of clause 21, further comprising selecting a first set of one or more operations for the one or more neural networks that satisfy a first memory constraint.

[0586] In clause 23, a method of any of clauses 21-22, further comprising performing a search of a search space to select the one or more neural networks, wherein performing the search comprises selecting a connection pattern, from a plurality of candidate connection patterns between layers of the one or more neural networks, based at least in part on probabilities of the plurality of candidate connection patterns.

[0587] In clause 24, a method of any of clauses 21-23, further comprising performing a joint two-level search of a topology search space and a cell search space to identify the one or more neural networks for an image-based task.

[0588] In clause 25, a method of any of clauses 21-24, wherein the one or more neural networks are to perform an image segmentation task.

[0589] In clause 26, a method of any of clauses 21-25, further comprising performing a search of a multi-scale topology search space by converting the multi-scale topology search space into a sequential search space comprising a super node for each respective layer of a plurality of layers, wherein each super node comprises a set of candidate feature nodes at the respective layer.

[0590] In clause 27, a processor comprising: one or more circuits to cause one or more neural networks to be selected from a plurality of neural networks based, at least in part, on an amount of memory to be used by the one or more neural networks.

[0591] In clause 28, a processor of clause 27, wherein the one or more circuits further cause the one or more neural networks to be selected by performing a first search of a search space, wherein the first search comprises: determine a maximum memory usage of all operations associated with one or more candidate feature nodes of the search space; receive a first input that specifies a first memory constraint as a first percentage of the maximum memory usage; and select a first set of one or more operations associated with one or more candidate feature nodes for the one or more neural networks to be selected, wherein a first memory usage of the first set of one or more operations is equal to or less than the first percentage of the maximum memory usage.

[0592] In clause 29, a processor of any of clauses 27-28, wherein the one or more circuits further cause the one or more neural networks to be selected by performing a second search of the search space, wherein the second search comprises: receive a second input that specifies a second memory constraint as a second percentage of the maximum memory usage; and select a second set of one or more operations associated with one or more candidate feature modes for the one or more neural networks to be selected, wherein a second memory usage of the second set of one or more operations is equal to or less than the second percentage of the maximum memory usage.

[0593] In clause 30, a processor of any of clauses 27-29, wherein the one or more circuits further cause the one or more neural networks to be selected by performing a first search of a search space using a first memory constraint, wherein the first search comprises: receive input data that specifies the first memory constraint; determine a first set of one or more search parameters based on the first memory constraint; and perform the first search in accordance with the first set of one or more search parameters, wherein first results of the first search for the one or more neural networks comprises a first set of one or more operations associated with one or more candidate feature nodes of the search space that satisfy the first memory constraint.

[0594] In clause 31, a processor of any of clauses 27-30, wherein the one or more circuits further cause the one or more neural networks to be selected by performing a second search of the search space using a second memory constraint that is different than the first memory constraint, wherein the second search comprises: receive input data that specifies the second memory constraint; determine a second set of one or more search parameters based on the second memory constraint; and perform the second search in accordance with the second set of one or more search parameters, wherein second results of the second search for the one or more neural networks comprises a second set of one or more operations associated with one or more candidate feature nodes of the search space that satisfy the second memory constraint.

[0595] In clause 32, a processor of any of clauses 27-31, wherein the one or more circuits further cause the one or more neural networks to be selected by performing a joint two-level search of a topology search space and a cell search space to identify the one or more neural networks for an image-based task.

[0596] In clause 33, a system comprising: one or more processors to cause one or more neural networks to be selected from a plurality of neural networks based, at least in part, on an amount of memory to be used by the one or more neural networks; and one or more memories to store parameters corresponding to the one or more neural networks.

[0597] In clause 34, a system of clause 33, wherein the one or more processors further cause the one or more neural networks to be selected by performing a first search of a search space, wherein the first search comprises: determine a maximum memory usage of all operations associated with one or more candidate feature nodes of the search space; receive a first input that specifies a first memory constraint as a first percentage of the maximum memory usage; and select a first set of one or more operations associated with one or more candidate feature modes for the one or more neural networks to be selected, wherein a first memory usage of the first set of one or more operations is equal to or less than the first percentage of the maximum memory usage.

[0598] In clause 35, a system of any of clauses 33-34, wherein the one or more processors further cause the one or more neural networks to be selected by performing a search of a search space, wherein the search comprises: identify a plurality of candidate connection patterns between a first layer and a second layer of the search space, the first layer and the second layer each comprising a set of candidate feature nodes at a different image scale, wherein each candidate feature node comprises a plurality of candidate edges that connect to a feature node in one or more previous layers; determine a probability of each of the plurality of candidate connection patterns; and select a connection pattern, from the plurality of candidate connection patterns, having a highest probability.

[0599] In clause 36, a system of any of clauses 33-35, wherein the one or more processors further cause the one or more neural networks to be selected by performing a joint two-level search of a topology search space and a cell search space to identify the one or more neural networks for an image-based task.

[0600] In clause 37, a system of any of clauses 33-36, wherein the image-based task is an image segmentation task.

[0601] In clause 38, a system of any of clauses 33-37, wherein: the topology search space comprises a plurality of layers, each layer comprising a set of candidate feature nodes each at a different image scale for a first selection of one or more candidate feature nodes for the one or more neural networks, wherein each candidate feature node comprises a plurality of candidate edges that connect to a feature node in one or more previous layers; and the cell search space comprises a plurality of candidate operations for a second selection of one of the plurality of candidate operations at each of the plurality of candidate edges.

[0602] In clause 39, a system of any of clauses 33-38, wherein each of the plurality of candidate operations is to receive an input feature map and provide an output feature map, the input feature map and the output feature map having a same spatial resolution, wherein each candidate feature node is a summation of output features from each of the plurality of candidate edges of the candidate feature node, wherein the one or more processors, during the joint two-level search, search the topology search space with a plurality of input features, each at a different image scale, and wherein the one or more processors, during the joint two-level search, search the topology search space for a connection pattern between each of the plurality of layers, the connection pattern comprising one or more of the plurality of candidate edges.

[0603] In clause 40, a system of any of clauses 33-39, wherein the topology search space is a multi-scale search space, and wherein the one or more processors further cause the one or more neural networks to: convert the multi-scale search space into a sequential search space comprising a super node for each respective layer of the plurality of layers, wherein each super node comprises the set of candidate feature nodes at the respective layer; identify a plurality of candidate connection patterns between a first layer and a second layer, each of the plurality of candidate connection patterns comprising a different combination of one or more of the plurality of candidate edges between the first layer and the second layer; determine a probability of each of the plurality of candidate connection patterns; and select a connection pattern, from the plurality of candidate connection patterns, based on the probability.

[0604] In clause 41, a datacenter comprising: or more processors to cause one or more neural networks to be selected from a plurality of neural networks to perform a medical image segmentation task based, at least in part, on an amount of memory to be used by the one or more neural networks.

[0605] In clause 42, a datacenter of clause 41, wherein the one or more processors further cause the one or more neural networks to be selected by performing a first search of a search space, wherein the first search comprises: determine a maximum memory usage of all operations associated with one or more candidate feature nodes of the search space; receive a first input that specifies a first memory constraint as a first percentage of the maximum memory usage; and select a first set of one or more operations associated with one or more candidate feature modes for the one or more neural networks to be selected, wherein a first memory usage of the first set of one or more operations is equal to or less than the first percentage of the maximum memory usage.

[0606] In clause 43, a datacenter of any of clauses 41-42, wherein the one or more processors further cause the one or more neural networks to be selected by performing a second search of the search space, wherein the second search comprises: receive a second input that specifies a second memory constraint as a second percentage of the maximum memory usage; and select a second set of one or more operations associated with one or more candidate feature modes for the one or more neural networks to be selected, wherein a second memory usage of the second set of one or more operations is equal to or less than the second percentage of the maximum memory usage.

[0607] In clause 44, a datacenter of any of clauses 41-43, wherein the one or more processors further cause the one or more neural networks to be selected by performing a joint two-level search of a topology search space and a cell search space to identify the one or more neural networks for the medical image segmentation task.

[0608] In clause 45, a datacenter of any of clauses 41-44, wherein the one or more processors further cause the one or more neural networks to be selected by performing a search of a search space, wherein the search comprises: identify a plurality of candidate connection patterns between a first layer and a second layer of the search space, the first layer and the second layer each comprising a set of candidate feature nodes at a different image scale, wherein each candidate feature node comprises a plurality of candidate edges that connect to a feature node in one or more previous layers; determine a probability of each of the plurality of candidate connection patterns; and select a connection pattern, from the plurality of candidate connection patterns, based on the probability.

[0609] In clause 46, a datacenter of any of clauses 41-45, wherein the search further comprises: identify a second plurality of candidate connection patterns between the second layer and a third layer of the search space, the third layer comprising a set of candidate feature nodes at a different image scale; determine a feasible set of candidate connection patterns, from the plurality of candidate connection patterns based on whether a respective candidate feature node comprises an input connection in the plurality of candidate connection patterns and an output connection in the second plurality of candidate connection patterns; and select, from the feasible set of candidate connection patterns, the connection pattern based on the probability.

[0610] In clause 47, a method comprising: receiving an input indicative of an amount of memory to be used; and selecting one or more neural networks from a plurality of neural networks based, at least in part, on the amount of memory to be used by the one or more neural networks.

[0611] In clause 48, a method of clause 47, further comprising performing a first search of a search space, wherein performing the first search comprises: determining a maximum memory usage of all operations associated with one or more candidate feature nodes of the search space; receiving a first input that specifies a first memory constraint as a first percentage of the maximum memory usage; and selecting a first set of one or more operations associated with one or more candidate feature modes for the one or more neural networks to be selected, wherein a first memory usage of the first set of one or more operations is equal to or less than the first percentage of the maximum memory usage.

[0612] In clause 49, a datacenter of any of clauses 47-48, further comprising performing a search of a search space, wherein the search comprises: identifying a plurality of candidate connection patterns between a first layer and a second layer of the search space, the first layer and the second layer each comprising a set of candidate feature nodes at a different image scale, wherein each candidate feature node comprises a plurality of candidate edges that connect to a feature node in one or more previous layers; determining a probability of each of the plurality of candidate connection patterns; and selecting a connection pattern, from the plurality of candidate connection patterns, having a highest probability.

[0613] In clause 50, a datacenter of any of clauses 47-49, further comprising performing a joint two-level search of a topology search space and a cell search space to identify the one or more neural networks for an image-based task.

[0614] In clause 51, a datacenter of any of clauses 47-50, wherein the image-based task is an image segmentation task.

[0615] In clause 52, a datacenter of any of clauses 47-51, wherein the topology search space is a multi-scale search space, and wherein the method further comprises: converting the multi-scale search space into a sequential search space comprising a super node for each respective layer of a plurality of layers, wherein each super node comprises a set of candidate feature nodes at the respective layer; identifying a plurality of candidate connection patterns between a first layer and a second layer, each of the plurality of candidate connection patterns comprising a different combination of one or more of a plurality of candidate edges between the first layer and the second layer; determining a probability of each of the plurality of candidate connection patterns; and selecting a connection pattern, from the plurality of candidate connection patterns, based on the probability.

[0616] In at least one embodiment, a single semiconductor platform may refer to a sole unitary semiconductor-based integrated circuit or chip. In at least one embodiment, multi-chip modules may be used with increased connectivity which simulates on-chip operation, and make substantial improvements over utilizing a conventional central processing unit ("CPU") and bus implementation. In at least one embodiment, various modules may also be situated separately or in various combinations of semiconductor platforms per desires of user.

[0617] In at least one embodiment, referring back to FIG. **7**, computer programs in form of machine-readable executable code or computer control logic algorithms are stored in main memory **704** and/or secondary storage. Computer programs, if executed by one or more processors, enable system **700** to perform various functions in accordance with at least one embodiment. In at least one embodiment, memory **704**, storage, and/or any other storage are possible examples of computer-readable media. In at least one embodiment, secondary storage may refer to any suitable storage device or system such as a hard disk drive and/or a removable storage drive, representing a floppy disk drive, a magnetic tape drive, a compact disk drive, digital versatile disk ("DVD") drive, recording device, universal serial bus ("USB") flash memory, etc. In at least one embodiment, architecture and/or functionality of various previous figures are implemented in context of CPU **702**, parallel processing system **712**, an integrated circuit capable of at least a portion of capabilities of both CPU **702**, parallel processing system **712**, a chipset (e.g., a group of integrated circuits designed to work and sold as a unit for performing related functions, etc.), and/or any suitable combination of integrated circuit (s).

[0618] In at least one embodiment, architecture and/or functionality of various previous figures are implemented in context of a general computer system, a circuit board system, a game console system dedicated for entertainment purposes, an application-specific system, and more. In at least one embodiment, computer system **700** may take form of a desktop computer, a laptop computer, a tablet computer, servers, supercomputers, a smart-phone (e.g., a wireless, hand-held device), personal digital assistant ("PDA"), a digital camera, a vehicle, a head mounted display, a hand-held electronic device, a mobile phone device, a television, workstation, game consoles, embedded system, and/or any other type of logic.

[0619] In at least one embodiment, parallel processing system **712** includes, without limitation, a plurality of parallel processing units ("PPUs") **714** and associated memories **716**. In at least one embodiment, PPUs **714** are connected to a host processor or other peripheral devices via an interconnect **718** and a switch **720** or multiplexer. In at least one embodiment, parallel processing system **712** distributes computational tasks across PPUs **714** which can be parallelizable—for example, as part of distribution of computational tasks across multiple graphics processing unit ("GPU") thread blocks. In at least one embodiment, memory is shared and accessible (e.g., for read and/or write access) across some or all of PPUs **714**, although such shared memory may incur performance penalties relative to use of local memory and registers resident to a PPU **714**. In at least one embodiment, operation of PPUs **714** is synchronized through use of a command such as _syncthreads( ), wherein all threads in a block (e.g., executed across multiple PPUs **714**) to reach a certain point of execution of code before proceeding.

[0620] Other variations are within spirit of present disclosure. Thus, while disclosed techniques are susceptible to various modifications and alternative constructions, certain illustrated embodiments thereof are shown in drawings and have been described above in detail. It should be understood, however, that there is no intention to limit disclosure to specific form or forms disclosed, but on contrary, intention is to cover all modifications, alternative constructions, and equivalents falling within spirit and scope of disclosure, as defined in appended claims.

[0621] Use of terms "a" and "an" and "the" and similar referents in context of describing disclosed embodiments (especially in context of following claims) are to be construed to cover both singular and plural, unless otherwise indicated herein or clearly contradicted by context, and not as a definition of a term. Terms "comprising," "having," "including," and "containing" are to be construed as open-ended terms (meaning "including, but not limited to,") unless otherwise noted. "Connected," when unmodified and referring to physical connections, is to be construed as partly or wholly contained within, attached to, or joined together, even if there is something intervening. Recitation of ranges of values herein are merely intended to serve as a shorthand method of referring individually to each separate value falling within range, unless otherwise indicated herein and each separate value is incorporated into specification as if it were individually recited herein. In at least one embodiment, use of term "set" (e.g., "a set of items") or "subset" unless otherwise noted or contradicted by context, is to be construed as a nonempty collection comprising one or more members. Further, unless otherwise noted or contradicted by context, term "subset" of a corresponding set does not necessarily denote a proper subset of corresponding set, but subset and corresponding set may be equal.

[0622] Conjunctive language, such as phrases of form "at least one of A, B, and C," or "at least one of A, B and C," unless specifically stated otherwise or otherwise clearly contradicted by context, is otherwise understood with context as used in general to present that an item, term, etc., may be either A or B or C, or any nonempty subset of set of A and B and C. For instance, in illustrative example of a set having three members, conjunctive phrases "at least one of A, B, and C" and "at least one of A, B and C" refer to any of following sets: {A}, {B}, {C}, {A, B}, {A, C}, {B, C}, {A, B, C}. Thus, such conjunctive language is not generally intended to imply that certain embodiments require at least one of A, at least one of B and at least one of C each to be present. In addition, unless otherwise noted or contradicted by context, term "plurality" indicates a state of being plural (e.g., "a plurality of items" indicates multiple items). In at least one embodiment, number of items in a plurality is at least two, but can be more when so indicated either explicitly or by context. Further, unless stated otherwise or otherwise clear from context, phrase "based on" means "based at least in part on" and not "based solely on."

[0623] Operations of processes described herein can be performed in any suitable order unless otherwise indicated herein or otherwise clearly contradicted by context. In at least one embodiment, a process such as those processes described herein (or variations and/or combinations thereof) is performed under control of one or more computer systems configured with executable instructions and is implemented as code (e.g., executable instructions, one or more computer programs or one or more applications) executing collectively on one or more processors, by hardware or combinations thereof. In at least one embodiment, code is stored on a computer-readable storage medium, for example, in form of a computer program comprising a plurality of instructions executable by one or more processors. In at least one embodiment, a computer-readable storage medium is a non-transitory computer-readable storage medium that excludes transitory signals (e.g., a propagating transient electric or electromagnetic transmission) but includes non-transitory data storage circuitry (e.g., buffers, cache, and queues) within transceivers of transitory signals. In at least one embodiment, code (e.g., executable code or source code) is stored on a set of one or more non-transitory computer-readable storage media having stored thereon executable instructions (or other memory to store executable instructions) that, when executed (i.e., as a result of being executed) by one or more processors of a computer system, cause computer system to perform operations described herein. In at least one embodiment, set of non-transitory computer-readable storage media comprises multiple non-transitory computer-readable storage media and one or more of individual non-transitory storage media of multiple non-transitory computer-readable storage media lack all of code while multiple non-transitory computer-readable storage media collectively store all of code. In at least one embodiment, executable instructions are executed such that different instructions are executed by different processors—for example, a non-transitory computer-readable storage medium store instructions and a main central processing unit ("CPU") executes some of instructions while a graphics processing unit ("GPU") executes other instructions. In at least one embodiment, different components of a computer system have separate processors and different processors execute different subsets of instructions.

[0624] Accordingly, in at least one embodiment, computer systems are configured to implement one or more services that singly or collectively perform operations of processes described herein and such computer systems are configured with applicable hardware and/or software that enable performance of operations. Further, a computer system that implements at least one embodiment of present disclosure is a single device and, in another embodiment, is a distributed computer system comprising multiple devices that operate differently such that distributed computer system performs operations described herein and such that a single device does not perform all operations.

[0625] Use of any and all examples, or exemplary language (e.g., "such as") provided herein, is intended merely to better illuminate embodiments of disclosure and does not pose a limitation on scope of disclosure unless otherwise claimed. No language in specification should be construed as indicating any non-claimed element as essential to practice of disclosure.

[0626] All references, including publications, patent applications, and patents, cited herein are hereby incorporated by reference to same extent as if each reference were individually and specifically indicated to be incorporated by reference and were set forth in its entirety herein.

[0627] In description and claims, terms "coupled" and "connected," along with their derivatives, may be used. It should be understood that these terms may be not intended as synonyms for each other. Rather, in particular examples, "connected" or "coupled" may be used to indicate that two or more elements are in direct or indirect physical or electrical contact with each other. "Coupled" may also mean that two or more elements are not in direct contact with each other, but yet still co-operate or interact with each other.

[0628] Unless specifically stated otherwise, it may be appreciated that throughout specification terms such as "processing," "computing," "calculating," "determining," or like, refer to action and/or processes of a computer or computing system, or similar electronic computing device, that manipulate and/or transform data represented as physical, such as electronic, quantities within computing system's registers and/or memories into other data similarly represented as physical quantities within computing system's memories, registers or other such information storage, transmission or display devices.

[0629] In a similar manner, term "processor" may refer to any device or portion of a device that processes electronic data from registers and/or memory and transform that electronic data into other electronic data that may be stored in registers and/or memory. As non-limiting examples, "processor" may be a CPU or a GPU. A "computing platform" may comprise one or more processors. As used herein, "software" processes may include, for example, software and/or hardware entities that perform work over time, such as tasks, threads, and intelligent agents. Also, each process may refer to multiple processes, for carrying out instructions in sequence or in parallel, continuously or intermittently. In at least one embodiment, terms "system" and "method" are used herein interchangeably insofar as system may embody one or more methods and methods may be considered a system.

[0630] In present document, references may be made to obtaining, acquiring, receiving, or inputting analog or digital data into a subsystem, computer system, or computer-implemented machine. In at least one embodiment, process of obtaining, acquiring, receiving, or inputting analog and digital data can be accomplished in a variety of ways such as by receiving data as a parameter of a function call or a call to an application programming interface. In at least one embodiment, processes of obtaining, acquiring, receiving, or inputting analog or digital data can be accomplished by transferring data via a serial or parallel interface. In at least one embodiment, processes of obtaining, acquiring, receiving, or inputting analog or digital data can be accomplished by transferring data via a computer network from providing entity to acquiring entity. In at least one embodiment, references may also be made to providing, outputting, transmitting, sending, or presenting analog or digital data. In various examples, processes of providing, outputting, transmitting, sending, or presenting analog or digital data can be accomplished by transferring data as an input or output parameter of a function call, a parameter of an application programming interface or interprocess communication mechanism.

[0631] Although descriptions herein set forth example implementations of described techniques, other architectures may be used to implement described functionality, and are intended to be within scope of this disclosure. Furthermore, although specific distributions of responsibilities may be defined above for purposes of description, various functions and responsibilities might be distributed and divided in different ways, depending on circumstances.

[0632] Furthermore, although subject matter has been described in language specific to structural features and/or methodological acts, it is to be understood that subject matter claimed in appended claims is not necessarily limited to specific features or acts described. Rather, specific features and acts are disclosed as exemplary forms of implementing the claims.

What is claimed is:

1. A processor comprising: one or more circuits to cause one or more neural networks to be selected from a plurality of neural networks based, at least in part, on an amount of memory to be used by the one or more neural networks.

2. The processor of claim **1**, wherein the one or more circuits are further to perform a search to select the one or more neural networks from the plurality of neural networks that satisfy a memory constraint.

3. The processor of claim **1**, wherein the one or more circuits are further to cause one or more additional neural networks to be selected from the plurality of neural networks or from a second plurality of neural networks, wherein the one or more additional neural networks are different from the one or more neural networks and satisfy a second memory constraint that is different from a first memory satisfied by the one or more neural networks.

4. The processor of claim **1**, wherein the one or more circuits are further to perform a search to select the one or more neural networks from the plurality of neural networks in accordance with a set of one or more search parameters determined at least in part on the amount of memory to be used by the one or more neural networks.

5. The processor of claim **1**, wherein a percentage of a maximum memory usage of operations associated with one or more candidate feature nodes of a search space comprising the plurality of neural networks is less than or equal to the amount of memory.

6. The processor of claim **1**, wherein the one or more circuits further cause the one or more neural networks to be selected by performing a joint two-level search of a topology search space and a cell search space to identify the one or more neural networks for an image-based task.

7. A system comprising:

one or more processors to cause one or more neural networks to be selected from a plurality of neural networks based, at least in part, on an amount of memory to be used by the one or more neural networks; and

one or more memories to store parameters corresponding to the one or more neural networks.

8. The system of claim **7**, wherein the one or more processors are further to perform a first search to select the one or more neural networks from the plurality of neural networks that satisfy a first memory constraint.

9. The system of claim **7**, wherein the one or more processors are further to select a connection pattern, from a plurality of candidate connection patterns between a first layer and a second layer of the one or more neural networks, based at least in part on probabilities of each of the plurality of candidate connection patterns.

10. The system of claim **7**, wherein the one or more processors are further to select a feature node from a set of candidate features nodes for one or more layers of the one or more neural networks, wherein the set of candidate feature nodes comprises feature nodes at different image scales that comprise a plurality of candidate edges that connect to a feature node in a previous layer.

11. The system of claim **7**, wherein the one or more processors further cause the one or more neural networks to be selected by performing a joint two-level search of a topology search space and a cell search space to identify the one or more neural networks for an image-based task.

12. The system of claim **7**, wherein the one or more neural networks are to perform an image segmentation task.

13. The system of claim **7**, wherein the one or more processors further cause the one or more neural networks to be selected by performing a search of a topology search space comprising a plurality of candidate edges that connect candidate feature nodes of a plurality of layers and a cell search space comprising a plurality of candidate operations.

14. The system of claim **7**, wherein the one or more processors are further to select a connection pattern, from a plurality of candidate connection patterns between layers of the one or more neural networks, based at least in part on probabilities of the plurality of candidate connection patterns.

15. A datacenter comprising: one or more processors to cause one or more neural networks to be selected from a plurality of neural networks to perform a medical image segmentation task based, at least in part, on an amount of memory to be used by the one or more neural networks.

16. The datacenter of claim **15**, wherein the one or more processors are further to perform a search to select the one or more neural networks from the plurality of neural networks that satisfy a memory constraint.

17. The datacenter of claim **15**, wherein the one or more processors are further to cause one or more additional neural networks to be selected from the plurality of neural networks or from a second plurality of neural networks, wherein the one or more additional neural networks are different from the one or more neural networks and satisfy a second memory constraint that is different from a first memory satisfied by the one or more neural networks.

18. The datacenter of claim **15**, wherein the one or more processors are to perform a joint two-level search of a topology search space and a cell search space to identify the one or more neural networks for the medical image segmentation task.

19. The datacenter of claim **15**, wherein the one or more processors are to perform a search of a search space to cause the one or more neural networks to be selected, wherein the search comprises selecting a connection pattern between layers of the one or more neural networks, from a plurality of candidate connection patterns, based at least in part on probabilities of the plurality of candidate connection patterns.

20. The datacenter of claim **15**, wherein the one or more processors are to perform a search of a search space to cause the one or more neural networks to be selected, wherein the search comprises selecting a connection pattern from a feasible set of candidate connection patterns between layers of the one or more neural networks, wherein each feasible connection pattern in the feasible set of candidate connection patterns comprises valid input connections and output connections between the layers.

21. A method comprising:

selecting one or more neural networks from a plurality of neural networks based, at least in part, on an amount of memory to be used by the one or more neural networks.

22. The method of claim **21**, further comprising selecting a first set of one or more operations for the one or more neural networks that satisfy a first memory constraint.

23. The method of claim **21**, further comprising performing a search of a search space to select the one or more neural networks, wherein performing the search comprises selecting a connection pattern, from a plurality of candidate connection patterns between layers of the one or more neural networks, based at least in part on probabilities of the plurality of candidate connection patterns.

24. The method of claim **21**, further comprising performing a joint two-level search of a topology search space and a cell search space to identify the one or more neural networks for an image-based task.

25. The method of claim **21**, wherein the one or more neural networks are to perform an image segmentation task.

26. The method of claim **21**, further comprising performing a search of a multi-scale topology search space by converting the multi-scale topology search space into a sequential search space comprising a super node for each respective layer of a plurality of layers, wherein each super node comprises a set of candidate feature nodes at the respective layer.

* * * * *