(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114731163 A
(43)申请公布日 2022.07.08

(21)申请号 202080080904.3
(22)申请日 2020.12.11
(30)优先权数据
16/716,350 2019.12.16 US
(85)PCT国际申请进入国家阶段日
2022.05.19
(86)PCT国际申请的申请数据
PCT/US2020/064705 2020.12.11
(87)PCT国际申请的公布数据
WO2021/126718 EN 2021.06.24
(71)申请人 辉达公司
地址 美国加利福尼亚州
(72)发明人 C·伊瓦尔斯·卡萨斯
H·D·巴努里·南叶·高达
J·D·富特克 J·H·德尔菲尔德
A·米勒 V·蒂雅
(74)专利代理机构 北京市磐华律师事务所 11336
专利代理师 高伟
(51)Int.Cl.
H03M 13/00(2006.01)

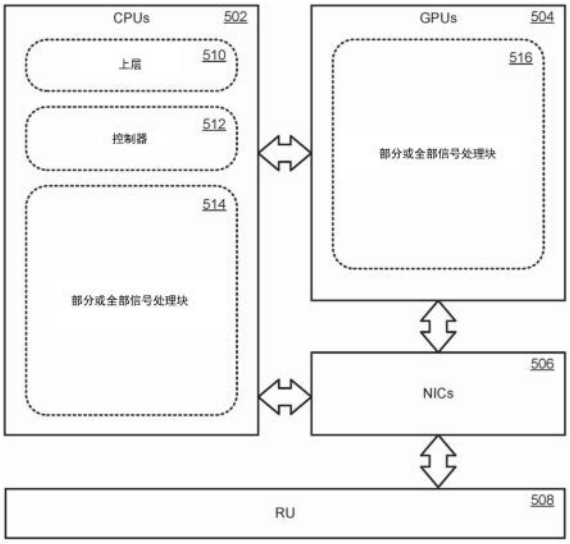
权利要求书2页 说明书102页 附图60页

(54)发明名称

5G NR信号信息的加速并行处理

(57)摘要

使用一个或更多个并行处理单元(PPU)在第五代(5G)新无线电(NR)网络中执行信号处理运算的装置、系统和技术。在至少一个实施例中,一个或更多个PPU在基带单元(BBU)中实现信号处理,该BBU单元在通信网络中执行5G NR物理层运算。



CN 114731163 A

1.一种处理器,包括:

一个或更多个电路,用于在一个或更多个并行处理单元上对第五代(5G)新无线电信号信息进行编码和解码。

2.根据权利要求1所述的处理器,其中所述一个或更多个电路用于通过执行一个或更多个物理层步骤来编码和解码所述5G新无线电信号信息,其中所述一个或更多个物理层步骤部分地由所述一个或更多个并行处理单元执行,以及部分地由中央处理单元执行。

3.根据权利要求1所述的处理器,其中对所述5G新无线电信号信息进行编码包括:启动一个或更多个并行处理单元内核以执行一个或更多个信号处理块。

4.根据权利要求3所述的处理器,其中所述一个或更多个信号处理块是物理上行链路共享信道(PUSCH)信号处理块。

5.根据权利要求3所述的处理器,其中所述一个或更多个信号处理块是物理下行链路共享信道(PDSCH)信号处理块。

6.根据权利要求1所述的处理器,其中所述一个或更多个并行处理单元直接从一个或更多个网络接口控制器发送和接收所述5G新无线电信号信息。

7.根据权利要求1所述的处理器,其中所述一个或更多个并行处理单元是图形处理单元。

8.一种系统,包括:

一个或更多个处理器,用于使用一个或更多个并行处理单元并行执行第五代(5G)信号处理的一个或更多个阶段。

9.根据权利要求8所述的系统,其中所述5G信号处理的一个或更多个阶段包括:使用部分地由所述一个或更多个并行处理单元执行的软件指令来执行物理层运算。

10.根据权利要求9所述的系统,其中由所述一个或更多个并行处理单元执行的所述软件指令由中央处理单元执行的附加指令控制。

11.根据权利要求9所述的系统,其中所述物理层运算是通过从CPU主机启动并行处理单元功能来执行的。

12.根据权利要求8所述的系统,所述5G信号处理的一个或更多个阶段属于5G物理层并且响应于来自与所述物理层不同的层的应用编程调用而被执行。

13.根据权利要求8所述的系统,其中所述5G信号处理的一个或更多个阶段是对由一个或更多个网络接口控制器发送或接收的一个或更多个信号执行的。

14.根据权利要求8所述的系统,其中所述并行处理单元直接从一个或更多个网络接口控制器接收一个或更多个信号用于5G信号处理。

15.根据权利要求8所述的系统,其中所述一个或更多个并行处理单元是图形处理单元。

16.一种机器可读介质,其上存储有一组指令,所述指令如果由一个或更多个处理器执行,则使所述一个或更多个处理器至少:

使用一个或更多个并行处理单元对第五代(5G)新无线电信号信息进行编码和解码。

17.根据权利要求16所述的机器可读介质,其中所述指令如果被执行,则进一步使所述一个或更多个处理器执行一个或更多个物理层步骤,以编码和解码所述5G新无线电信号信息,其中所述一个或更多个物理层步骤部分地由所述一个或更多个并行处理单元执行,以

及部分地由中央处理单元执行。

18.根据权利要求16所述的机器可读介质,其中对所述5G新无线电信号信息进行编码和解码包括:启动一个或更多个并行处理单元内核以在所述一个或更多个并行处理单元上执行一个或更多个信号处理块。

19.根据权利要求18所述的机器可读介质,其中所述一个或更多个信号处理块是物理上行链路共享信道(PUSCH)信号处理块。

20.根据权利要求18所述的机器可读介质,其中所述一个或更多个信号处理块是物理下行链路共享信道(PDSCH)信号处理块。

21.根据权利要求16所述的机器可读介质,其中所述一个或更多个并行处理单元直接从一个或更多个网络接口控制器发送和接收所述5G新无线电信号信息。

22.根据权利要求16所述的机器可读介质,其中通过启动以下项中的至少一个来执行对所述5G新无线电信号信息的编码:物理上行链路共享信道(PUSCH)波束成形内核、PUSCH前端内核、PUSCH解速率匹配内核、PUSCH低密度奇偶校验(LDPC)内核或循环冗余校验(CRC)内核。

23.根据权利要求16所述的机器可读介质,其中通过启动以下项中的至少一个来执行对所述5G新无线电信号信息的解码:物理下行链路共享信道(PDSCH)波束成形内核、PDSCH速率匹配内核、PDSCH低密度奇偶校验(LDPC)内核或循环冗余校验(CRC)内核。

24.根据权利要求16所述的机器可读介质,其中所述5G新无线电信号信息通过一个或更多个网络接口控制器发送和接收。

25.一种方法,包括:

使用一个或更多个并行处理单元并行执行第五代(5G)信号处理的一个或更多个阶段。

26.根据权利要求25所述的方法,其中所述5G信号处理的一个或更多个阶段进一步包括:使用由所述一个或更多个并行处理单元执行的软件函数来执行物理层运算。

27.根据权利要求26所述的方法,其中由所述一个或更多个并行处理单元执行的软件函数由中央处理单元控制。

28.根据权利要求26所述的方法,其中所述软件函数实现所述物理层的一个或更多个信号处理运算。

29.根据权利要求28所述的方法,其中所述一个或更多个信号处理运算是物理上行链路共享信道(PUSCH)信号处理运算。

30.根据权利要求28所述的方法,其中所述一个或更多个信号处理运算是物理下行链路共享信道(PDSCH)信号处理运算。

31.根据权利要求25所述的方法,其中所述一个或更多个并行处理单元直接从一个或更多个网络接口控制器接收一个或更多个信号用于5G信号处理。

32.根据权利要求25所述的方法,其中所述一个或更多个并行处理单元是图形处理单元。

5G NR信号信息的加速并行处理

[0001] 相关申请的交叉引用

[0002] 本申请要求于2019年12月16日提交的题为“5G NR信号信息的加速并行处理(ACCELERATED PARALLEL PROCESSING OF 5G NR SIGNAL INFORMATION)”的美国专利申请No.16/716,350的优先权,其全部内容通过引用整体并入本文并用于所有目的。

技术领域

[0003] 至少一个实施例涉及在第五代(5G)新无线电网络中处理信号。例如,至少一个实施例涉及根据本文描述的各种新技术的用于使用一个或更多个图形处理单元对5G信号进行编码和解码的处理器或计算系统。

背景技术

[0004] 随着越来越多的用户改为使用蜂窝设备作为他们的主要通信方式,移动数据流量正以指数速度增长。为了处理这种快速增长,已经开发了新的通信标准,这些通信标准引入了促进更大的无线网络访问和信息吞吐量的技术层。当前技术难以满足新无线标准的吞吐量和数据容量要求,因为网络基础设施的硬件能力必须能够处理具有挑战性的信息处理水平。实现这些新标准的技术难以满足极低延迟、极高吞吐量、低功耗和云就绪的要求。

附图说明

[0005] 图1示出了根据至少一个实施例的无线网络中的基带单元(BBU);

[0006] 图2示出了根据至少一个实施例的包含BBU的各种上行链路和下行链路信号处理组件的框图;

[0007] 图3示出了根据至少一个实施例的用于实现BBU的信号处理组件的分离架构;

[0008] 图4示出了根据至少一个实施例的包含诸如图形处理单元(GPU)之类的并行处理单元(PPU)的通用BBU架构;

[0009] 图5示出了根据至少一个实施例的用于在一个或更多个PPU(例如 BBU中的GPU)上部分或完全实现信号处理块的系统;

[0010] 图6示出了根据至少一个实施例的用于在一个或更多个PPU(诸如 GPU)上执行第五代(5G)新无线电(NR)基带运算的过程;

[0011] 图7示出了根据至少一个实施例的示例数据中心系统;

[0012] 图8A示出了根据至少一个实施例的自主车辆的示例;

[0013] 图8B示出了根据至少一个实施例的图8A的自主车辆的相机位置和视野的示例;

[0014] 图8C是根据至少一个实施例的示出图8A的自主车辆的示例系统架构的框图;

[0015] 图8D是根据至少一个实施例的示出用于一个或更多个基于云的服务器与图8A的自主车辆之间进行通信的系统的图;

[0016] 图9是根据至少一个实施例的示出计算机系统的框图;

[0017] 图10是根据至少一个实施例的示出计算机系统的框图;

[0018] 图11示出了根据至少一个实施例的计算机系统；

[0019] 图12示出了根据至少一个实施例的计算机系统；

[0020] 图13A示出了根据至少一个实施例的计算机系统；

[0021] 图13B示出了根据至少一个实施例的计算机系统；

[0022] 图13C示出了根据至少一个实施例的计算机系统；

[0023] 图13D示出了根据至少一个实施例的计算机系统；

[0024] 图13E和图13F示出了根据至少一个实施例的共享编程模型；

[0025] 图14示出了根据至少一个实施例的示例性集成电路和相关的图形处理器；

[0026] 图15A-图15B示出了根据至少一个实施例的示例性集成电路和相关联的图形处理器；

[0027] 图16A-图16B示出了根据至少一个实施例的附加的示例性图形处理器逻辑；

[0028] 图17示出了根据至少一个实施例的计算机系统；

[0029] 图18A示出了根据至少一个实施例的并行处理器；

[0030] 图18B示出了根据至少一个实施例的分区单元；

[0031] 图18C示出了根据至少一个实施例的处理集群；

[0032] 图18D示出了根据至少一个实施例的图形多处理器；

[0033] 图19示出了根据至少一个实施例的多图形处理单元(GPU)系统；

[0034] 图20示出了根据至少一个实施例的图形处理器；

[0035] 图21是根据至少一个实施例的示出用于处理器的处理器微架构的框图；

[0036] 图22示出了根据一个或更多个实施例的图形处理器的至少部分；

[0037] 图23示出了根据一个或更多个实施例的图形处理器的至少部分；

[0038] 图24示出了根据一个或更多个实施例的图形处理器的至少部分；

[0039] 图25是根据至少一个实施例的图形处理器的图形处理引擎的框图；

[0040] 图26是根据至少一个实施例的图形处理器核心的至少部分的框图；

[0041] 图27A-图27B示出了根据至少一个实施例的线程执行逻辑,其包括图形处理器核心的处理元件的阵列；

[0042] 图28示出了根据至少一个实施例的并行处理单元(“PPU”)；

[0043] 图29示出了根据至少一个实施例的通用处理集群(“GPC”)；

[0044] 图30示出了根据至少一个实施例的并行处理单元(“PPU”)的存储器分区单元；

[0045] 图31示出了根据至少一个实施例的流式多处理器；

[0046] 图32示出了根据至少一个实施例的用于在5G无线通信网络内通信数据的网络；

[0047] 图33示出了根据至少一个实施例的用于5G LTE无线网络的网络架构；

[0048] 图34是示出根据至少一个实施例的根据LTE和5G原理操作的移动电信网络/系统的一些基本功能的图；

[0049] 图35示出了根据至少一个实施例的可以是5G网络架构的部分的无线电接入网络；

[0050] 图36提供了根据至少一个实施例的其中使用多个不同类型的设备的 5G移动通信系统的示例图解；

[0051] 图37示出了根据至少一个实施例的示例高级系统；

[0052] 图38示出了根据至少一个实施例的网络系统的架构；

[0053] 图39示出了根据至少一个实施例的设备的示例组件；

[0054] 图40示出了根据至少一个实施例的基带电路的示例接口；

[0055] 图41示出了根据至少一个实施例的上行链路信道的示例；

[0056] 图42示出了根据至少一个实施例的网络系统的架构；

[0057] 图43示出了根据至少一个实施例的控制平面协议栈；

[0058] 图44示出了根据至少一个实施例的用户平面协议栈；

[0059] 图45示出了根据至少一个实施例的核心网络的组件；以及

[0060] 图46示出了根据至少一个实施例的支持网络功能虚拟化(NFV)的系统的组件。

具体实施方式

[0061] 图1示出了无线网络中的基带单元(BBU)112。在至少一个实施例中，多个无线设备102、104、106使用本文所述的第五代(5G)新无线电标准与连接到无线电单元(RU)108的天线进行通信。在至少一个实施例中，RU 108通过网络110与至少一个基带单元(BBU)112通信，所述网络诸如以太网、光纤或另一种合适的有线或无线网络介质。

[0062] 在至少一个实施例中，BBU 112包含一个或更多个计算设备，用于执行信号处理和其他运算以处理5G网络中的信息。在至少一个实施例中，BBU 112包含本文所述的一个或更多个并行处理单元(PPU)，用于执行包括与信号处理相关的那些的计算运算。在至少一个实施例中，BBU 112 负责解释通信系统中的基带频率，包括计算机网络、互联网、电话网络和无线电广播系统。在至少一个实施例中，BBU 112在多个位置之间切换网络数据业务中的通信。在至少一个实施例中，BBU 112对用纠错码编码的数据执行编码和解码。在至少一个实施例中，BBU 112利用一个或更多个 PPU来实现信号处理运算，例如作为5G通信方案的PHY或物理层的一部分来实现的那些。在至少一个实施例中，PPU包括图形处理单元(GPU)。在至少一个实施例中，BBU 112在多个位置之间切换网络数据业务中的通信。

[0063] 在至少一个实施例中，BBU 112连接到包含天线108的基站，天线108 也可以称为无线电单元(RU)、远程无线电单元(RRU)或远程无线电头 (RRH)。在至少一个实施例中，RU 108包含处理资源或计算资源。在至少一个实施例中，RU 108可以包含一个或更多个PPU。在至少一个实施例中，RU 108可以包含一个或更多个GPU，用于促进信号处理运算的计算。在至少一个实施例中，RU 108将包含用于促进与一个或更多个无线设备 102、104、106的无线通信的天线。

[0064] 在至少一个实施例中，RU 108天线可以是线状天线，例如短偶极天线、偶极天线、环形天线或单极天线。在至少一个实施例中，RU 108天线可以是对数周期天线，例如蝴蝶结天线、对数周期天线或对数周期偶极子阵列。在至少一个实施例中，RU 108天线可以是孔径天线，例如缝隙天线或喇叭天线。在至少一个实施例中，RU 108天线可以是微带天线，例如矩形微带贴片天线或四分之一波长贴片天线。在至少一个实施例中，RU 108天线可以是反射器天线，例如平板反射器天线、角反射器天线或抛物面反射器天线。在至少一个实施例中，RU 108天线可以是透镜天线。在至少一个实施例中，RU 108天线可以是行波天线，例如长线天线、八木天线、螺旋线天线或螺旋天线。在至少一个实施例中，RU 108天线可以是阵列天线，例如二元阵列天线、线性阵列天线或相控阵天线。在至少一个实施例中，RU 108 天线可以是促进长距离或短距离无线通信的任何类型的天线。

[0065] 在至少一个实施例中，无线设备102、104、106通过有噪声的无线信道将数据发送到包含一个或更多个天线的一个或更多个RU 108，如本文所述。在至少一个实施例中，无线设备102、104、106使用被确定为5G无线通信标准的一部分的技术对通信信号进行编码和解码。在至少一个实施例中，无线设备102、104、106对发送到一个或更多个RU 108的信息进行编码。在至少一个实施例中，由无线设备102、104、106发送到一个或更多个RU 108的编码信息由本文进一步描述的BBU 112进行处理。

[0066] 在至少一个实施例中，无线设备102、104、106包括蜂窝电话和其他移动电话设备。在至少一个实施例中，无线设备102、104、106是手持平板电脑或本文所述的其他便携式计算设备。在至少一个实施例中，无线设备102、104、106是膝上型计算机。在至少一个实施例中，无线设备102、 104、106是无线连接到基站或RU 108的桌上型计算设备。在至少一个实施例中，无线设备102、104、106包括通过5G新无线电无线通信方案发送和接收数据的任何计算设备。在至少一个实施例中，无线设备102、104、 106包括通过任何其他无线通信方案发送和接收数据的任何计算设备。

[0067] 图2示出了包含基带单元(BBU)的上行链路202和下行链路204信号处理组件的框图。在至少一个实施例中，基带单元的上行链路202和下行链路204信号处理组件包括第五代(5G)新无线电无线通信系统中的物理层(PHY)。在至少一个实施例中，PHY被组织成物理信道，例如上行链路202和下行链路204。在至少一个实施例中，每个信道执行数据的发送202和接收204的功能。在至少一个实施例中，每个信道执行控制信息的发送202和接收204、小区发现和初始接入的功能。在至少一个实施例中，上行链路202和下行链路204信号处理组件提供由特定于每个PHY信道的信号处理块组成的信号处理管线。在至少一个实施例中，对于BBU 实现发射器通信的下行链路204信道，信号处理块由第三代合作伙伴计划 (3GPP) NR标准规范确定。在至少一个实施例中，对于上行链路202信道，其中BBU正在实现接收器通信，信号处理块是特定于实现方式的并且不限于图2中指定的那些。

[0068] 在至少一个实施例中，PHY控制器206提供至BBU中的上层的在PHY 层信号处理运算之间的接口，例如包含在上行链路信道202中的那些和包含在下行链路信道204中的那些。在至少一个实施例中，PHY控制器206 充当实现信号处理运算的PHY或物理层与第三代合作伙伴计划3GPP协议栈的上层之间的接口。在至少一个实施例中，PHY控制器206提供应用编程接口(API)。在至少一个实施例中，PHY控制器206提供基于标准的 API，例如Small Cell Forum FAPI接口。在至少一个实施例中，PHY控制器206提供专有的API。

[0069] 在至少一个实施例中，PHY控制器206管理在任何并行处理单元(PPU) 上运行的软件。在至少一个实施例中，PHY控制器206管理用于执行一个或更多个特定运算的软件内核或软件指令段，实现信号处理运算，例如在 5G新无线电无线通信系统的PHY层中实现的那些，其中所述软件内核由一个或更多个图形处理单元(GPU)执行，如下所述。在至少一个实施例中，一个或更多个GPU可以实现信号处理管线中的所有功能。在至少一个实施例中，一个或更多个GPU可以用作信号处理管线的特定块的加速器。在至少一个实施例中，PHY控制器206提供软件来管理GPU软件调用。在至少一个实施例中，PHY控制器206向一个或更多个GPU执行的软件内核提供必要的参数。在至少一个实施例中，PHY控制器206管理来自正在由一个或更多个GPU执行的软件内核的输出。

[0070] 在至少一个实施例中，5G NR PHY信道202、204提供用于数据传输的信号处理组

件。在至少一个实施例中，数据传输组件包括物理上行链路共享信道(PUSCH)210。在至少一个实施例中，5G NR中的PUSCH 210 被指定用于承载多路复用的控制信息和用户应用数据，如本文进一步描述的。在至少一个实施例中，数据传输组件包括物理下行链路共享信道(PDSCH)224。在至少一个实施例中，PDSCH承载用户数据和高层信令，如本文进一步描述的。

[0071] 在至少一个实施例中，5G NR PHY信道202、204提供用于控制信息传输的信号处理组件。在至少一个实施例中，控制信息传输组件包括物理下行链路控制信道(PDCCH)226和物理上行链路控制信道(PUCCH)212。在至少一个实施例中，PDCCH 226和PUCCH 212承载关于与PDSCH 224 和PUSCH 210信道相关的传输格式和资源分配的信息，如本文进一步描述的。

[0072] 在至少一个实施例中，控制信息传输组件包括四个主要参考信号214、216、230。在至少一个实施例中，PHY控制信息传输组件中的四个主要参考信号是解调参考信号(DMRS)214、230、相位-跟踪参考信号(PTRS) 230、探测参考信号(SRS)216和信道状态信息参考信号(CSI-RS)230。在至少一个实施例中，DMRS 214、230用于估计用于解调的无线电信道，如本文进一步所述。在至少一个实施例中，PTRS 230用于实现振荡器相位噪声的补偿，如本文进一步描述的。在至少一个实施例中，SRS 216和 CSI-RS 230用于执行用于调度和链路自适应的信道状态信息(CSI)测量，如本文进一步描述的。

[0073] 在至少一个实施例中，5G NR PHY信道202、204为初始接入和小区发现提供信号处理组件。在至少一个实施例中，通过物理随机接入信道 (PRACH)218和物理广播信道(PBCH)228促进小区发现，如本文进一步描述的。在至少一个实施例中，可以广播同步信号块(SS块)232以选择服务小区，如本文进一步描述的。

[0074] 在至少一个实施例中，5G NR PHY信道202、204提供用于低物理(低 PHY)功能的信号处理组件，其对5G NR信号执行基本运算，如本文进一步描述的。在至少一个实施例中，低PHY功能包括快速傅立叶变换(FFT) 222和快速傅立叶逆变换(IFFT)236。在至少一个实施例中，FFT 222和 IFFT 236将基于频率的信号信息转换为基于时间的数据以用于处理，反之亦然，如本文进一步描述的。在至少一个实施例中，低PHY功能包括循环前缀(CP)插入236和移除222。在至少一个实施例中，CP插入236和移除222促进FFT 222和IFFT 236运算的执行以执行卷积，本文进一步描述。在至少一个实施例中，低PHY功能包括发射波束成形(Tx波束成形)234 和接收波束成形(Rx波束成形)220。在至少一个实施例中，波束成形220、234是在5G NR和其他无线网络中使用的信号滤波技术，如本文进一步所述。在至少一个实施例中，一个或更多个无线电单元中的一个或更多个天线208将发送和接收信号数据。

[0075] 图3示出了用于实现基带单元(BBU)304和无线电单元(RU)302 的信号处理组件的分离架构。在至少一个实施例中，所有物理信道(PHY) 功能都由RU 302或BBU 304执行。在至少一个实施例中，PHY功能在 RU 302和BBU 304之间被分段。在至少一个实施例中，PHY功能的分段由诸如第三代合作伙伴计划(3GPP)和/或开放无线电接入网络(ORAN) 之类的标准确定。在至少一个实施例中，PHY功能的分段可以由专有设计要求确定。

[0076] 作为示例，在至少一个实施例中，本文描述的低PHY功能在RU 302 上实现，而其余PHY功能在BBU 304上实现，BBU 304可以包含如本文所述的一个或更多个并行处理单元。在至少一个实施例中，低PHY功能包括快速傅立叶变换(FFT)和快速傅立叶逆变换(IFFT)306，它们执行卷积运算以将信号信息转换成可以由其他PHY信号处理运算进一步处理的数据。

在至少一个实施例中，低PHY功能包括循环前缀(CP)添加和移除306，用于促进卷积运算(例如FFT和IFFT 306)的执行。在至少一个实施例中，本文描述的波束成形运算308作为将低PHY功能分段到RU 302 的一部分来执行。

[0077] 在至少一个实施例中，当低PHY运算306、308被分段到RU 302时，剩余的PHY信号处理运算在BBU 304中实现，BBU 304包括一个或更多个并行处理单元(PPU)，PPU包括图形处理单元(GPU)。在至少一个实施例中，PHY控制器310管理在包括GPU的任何PPU上运行的软件。在至少一个实施例中，PHY控制器310管理用于执行一个或更多个特定运算的软件内核或软件指令段，实现信号处理运算以执行PHY信道处理312，其中所述软件内核由一个或更多个图形处理单元(GPU)执行，如下所述。在至少一个实施例中，PHY控制器310提供软件来管理GPU软件调用以执行PHY信道处理312。在至少一个实施例中，PHY控制器310向一个或更多个GPU执行的软件内核提供必要的参数。在至少一个实施例中，PHY 控制器310管理来自由一个或更多个GPU执行的软件内核的输出。在至少一个实施例中，BBU 304使用一个或更多个中央处理单元(CPU)和一个或更多个GPU来执行其他PHY信道处理运算312，如下所述。在至少一个实施例中，尽管使用RU 302对低PHY运算进行分段，但在BBU 304 上执行需要更多处理资源的一些低PHY运算，例如波束成形权重计算314。

[0078] 图4示出了包含一个或更多个图形处理单元(GPU)408的通用基带单元(BBU)402架构。在至少一个实施例中，BBU 402或实现第五代(5G) 新无线电(NR)物理层或信道(PHY)的其他系统包含用于执行信号处理运算以及接收信号信息的硬件或软件。在至少一个实施例中，通过有线网络连接从远程无线电单元(RRU)或其他无线电单元(RU)接收信号信息。

[0079] 在至少一个实施例中，实现5G NR PHY的BBU 402或其他系统包含一个或更多个中央处理单元(CPU)406。在至少一个实施例中，一个或更多个CPU 406实现或执行实现5G NR物理层的上层的软件指令。在至少一个实施例中，一个或更多个CPU 406实现或执行软件指令，这些软件指令实现5G NR PHY的一些或所有信号处理运算，这取决于这些运算在BBU 402和一个或更多个RU之间如何划分，如上所述。在至少一个实施例中，一个或更多个CPU 406实现或执行软件指令，这些软件指令实现5G NR PHY的一些或所有信号处理运算，其中执行的特定信号处理运算基于已指定哪些运算用于在一个或更多个图形处理单元(GPU)408上加速执行。

[0080] 在至少一个实施例中，实现5G NR PHY的BBU 402或其他系统包含一个或更多个GPU 408。在至少一个实施例中，一个或更多个GPU 408实现或执行实现全部或部分5G NR PHY的软件指令。在至少一个实施例中，一个或更多个GPU 408实现或执行实现5G NR PHY的一些或所有运算的软件指令，这取决于已指定哪些运算用于在一个或更多个CPU 406上实现或由其执行。在至少一个在实施例中，一个或更多个GPU 408实现全部的 5G NR PHY。在至少一个实施例中，一个或更多个GPU 408实现部分5G NR PHY，其余运算将由一个或更多个CPU 406执行。

[0081] 在至少一个实施例中，BBU 402或实现5G NR PHY的其他系统包含一个或更多个网络接口控制器(NIC)404。在至少一个实施例中，一个或更多个NIC 404将从一个或更多个RU或RRU接收的信息通过直接通信信道410(例如总线，或BBU 402或实现5G NR PHY的其他系统中的其他内部通信介质)直接传输到一个或更多个CPU 406和/或一个或更多个GPU 408。在至少一个实施例中，一个或更多个CPU 406和/或一个或更多个GPU 408与一个或更多个

NIC 404的通信是双向的并且用于数据的发送和接收。

[0082] 图5示出了用于在基带单元(BBU)中的一个或更多个并行处理单元 (PPU)(例如一个或更多个图形处理单元(GPU)504)上的第五代(5G) 新无线电(NR物理层或信道(PHY)中部分或全部实现信号处理块的系统。在至少一个实施例中,BBU包含一个或更多个中央处理单元(CPU) 502和一个或更多个PPU,例如GPU 504,其处理实现5G NR PHY的信号处理组件或运算的软件指令。在至少一个实施例中,信号处理组件或运算在一个或更多个软件块514、516中实现。在至少一个实施例中,软件块 514、516在由一个或更多个CPU 502和/或一个或更多个GPU 504执行或处理时执行5G NR信号的接收和发送中涉及的信号处理算法。在至少一个实施例中,软件块514、516实现低物理层(低PHY)运算,例如快速傅立叶变换(FFT)、快速傅立叶逆变换(IFFT)和波束成形,如本文所述。在至少一个实施例中,软件块514、516在由一个或更多个CPU 502和/或一个或更多个GPU 504执行或处理时,针对一个或更多个物理信道执行一些或所有5G NR PHY处理运算。

[0083] 在至少一个实施例中,软件块514、516在由一个或更多个CPU 502 和/或一个或更多个GPU 504执行或处理时,执行用于5G NR PHY的信号处理管线。在至少一个实施例中,信号处理管线由至少几个信号处理块514、 516组成。在至少一个实施例中,信号处理块特定于每个5G NR PHY信道,如本文所述。在至少一个实施例中,BBU实现发射器的下行链路信道包含由第三代合作伙伴计划(3GPP)指定的信号处理块。在至少一个实施例中, BBU实现接收器的上行链路信道、实现的信号处理块是灵活的。在至少一个实施例中,一个或更多个GPU 504将信号处理管线中的所有功能实现为软件信号处理块516。在至少一个实施例中,一个或更多个GPU 504用作 5G NR物理层中的信号处理块516的特定子集的加速器。

[0084] 在至少一个实施例中,5G NR PHY信道处理完全或部分实现为一个或更多个PPU(例如GPU 504)上的软件指令块516。在至少一个实施例中, PPU和/或GPU 504加速是通过实现从CPU 502主机调用的各个信号处理块或内核516来实现的。在至少一个实施例中,数据在一个或更多个网络接口控制器(NIC)506和与一个或更多个CPU 502相关联的存储器之间直接传输。在至少一个实施例中,数据在一个或更多个网络接口控制器 (NIC)506和与一个或更多个PPU(例如GPU 504)相关联的共享内存之间直接传输。

[0085] 在至少一个实施例中,一个或更多个CPU 502包含在被执行时实现 5G NR PHY的上层510的指令。在至少一个实施例中,上层510包括软件库和3GPP中指定的其他运算。在至少一个实施例中,一个或更多个CPU 502实现PHY控制器512。在至少一个实施例中,PHY控制器512提供5G NR PHY信号处理运算块514、516和BBU中的上层510之间的接口。在至少一个实施例中,PHY控制器512充当5G NR PHY信号处理运算514、516与3GPP协议栈的上层510之间的接口。在至少一个实施例中,PHY 控制器512提供应用编程接口(API)。在至少一个实施例中,PHY控制器512提供API用于与软件指令块或内核接口,该软件指令块或内核在由一个或更多个CPU 502和/或一个或更多个PPU(诸如GPU 504)执行时,实现5G NR PHY运算。在至少一个实施例中,PHY控制器512提供基于标准的API,例如Small Cell Forum FAPI接口。在至少一个实施例中,PHY 控制器512提供专有的API。

[0086] 在至少一个实施例中,PHY控制器512管理实现在诸如GPU 504之类的任何并行处理单元(PPU)上运行的5G NR PHY的软件指令块或内核。在至少一个实施例中,PHY控制器512管理用于执行一个或更多个特定运算的软件内核或软件指令段,实现信号处理运算,例

如在5G新无线电无线通信系统的PHY层中实现的那些,其中所述软件内核由一个或更多个PPU执行,例如GPU 504,如本文所述。在至少一个实施例中,一个或更多个PPU,例如GPU 504,可以实现信号处理管线中的所有功能。在至少一个实施例中,一个或更多个PPU,例如GPU 504,可以用作信号处理管线的特定块的加速器,例如5G NR PHY中的特定信号处理运算。在至少一个实施例中,PHY控制器512提供软件来管理GPU 504软件调用。在至少一个实施例中,PHY控制器512向一个或更多个GPU 504执行的软件内核提供必要的参数。在至少一个实施例中,PHY控制器512管理来自一个或更多个GPU 504执行的软件内核的输出。

[0087] 在至少一个实施例中,一个或更多个PPU 504执行软件指令块516,该软件指令块516实现了执行5G NR PHY的部分的一个或更多个运算。在至少一个实施例中,一个或更多个PPU是GPU 504,其为实现特定5G NR PHY运算的一些或所有函数的软件指令块或内核516提供加速。在至少一个实施例中,一个或更多个CPU 502负责管理软件内核和在一个或更多个 GPU 504上启动软件内核。在至少一个实施例中,一个或更多个CPU 502 构建5G NR PHY管线信号处理内核的执行图以最小化从CPU 502启动的 GPU 504软件内核的数量。

[0088] 在至少一个实施例中,物理上行链路共享信道(PUSCH)接收器在5G NR BBU中的一个或更多个PPU(例如GPU 504)上实现。在至少一个实施例中,5G NR PHY中的PUSCH被指定用于承载多路复用的控制信息和用户应用数据,如本文进一步描述的。在至少一个实施例中,实现为软件内核以供一个或更多个PPU(例如GPU 504)执行的特定PUSCH运算包括解调参考信号(DMRS)处理和信道估计、接收波束成形权重计算、均衡器权重计算、信号均衡、层和资源元素解映射、软解映射、解扰、解速率匹配、解交错和循环冗余校验(CRC)解码。在至少一个实施例中,特定的GPU内核实现PUSCH波束成形,其具体执行至少包括波束成形权重计算和波束成形权重应用的处理块。在至少一个实施例中,特定的GPU 内核实现PUSCH前端,其具体执行至少包括DMRS处理和信道估计、接收波束成形权重计算、均衡器权重计算、信号均衡、层解映射和软解映射的处理块。在至少一个实施例中,特定的GPU内核实现PUSCH解速率匹配,其具体执行至少包括资源元素解映射、解扰、解速率匹配和解交错的处理块。在至少一个实施例中,特定的GPU内核实现PUSCH低密度奇偶校验(LDPC)解码器,其具体执行至少包括LDPC解码的处理块。在至少一个实施例中,特定的GPU内核实现PUSCH CRC解码器,其具体执行至少包括CRC解码的处理块。

[0089] 在至少一个实施例中,物理下行链路共享信道(PDSCH)发射器在 5G NR BBU中的一个或更多个PPU(例如GPU 504)上实现。在至少一个实施例中,PDSCH承载用户数据和更高层信令,如本文进一步描述的。在至少一个实施例中,实现为软件内核以供一个或更多个PPU(例如GPU 504) 执行的特定PUSCH运算包括传输块分段、用于传输块和代码块的CRC编码、LDPC编码、速率匹配、交错、加扰、调制、资源元素和层映射、预编码、波束成形权重计算和波束成形权重应用。在至少一个实施例中,特定的GPU内核实现PDSCH CRC编码器,其具体执行至少包括CRC编码的处理块。在至少一个实施例中,特定GPU内核实现PDSCH LDPC编码器,其具体执行至少包括LDPC编码的处理块。在至少一个实施例中,特定的GPU内核实现PDSCH速率匹配,其具体执行至少包括速率匹配、交错、加扰和资源元素映射的处理块。在至少一个实施例中,特定的GPU内核实现PDSCH波束成形,其具体执行至少包括波束成形权重计算和波束成形权重应用的处理块。在至少一个实施例中,特定的GPU内核实现了 PDSCH前端,其具体执行至少包括DMRS处理和信道估计、接收波束成形权重计算、均衡器权重计算、

信号均衡、层解映射和软解映射的处理块。在至少一个实施例中，特定GPU内核实现PDSCH解速率匹配，其具体执行至少包括资源元素解映射、解扰、解速率匹配和解交错的处理块。在至少一个实施例中，特定的GPU内核实现PDSCH LDPC解码器，其具体执行至少包括LDPC解码的处理块。在至少一个实施例中，特定的GPU内核实现PDSCH CRC解码器，其具体执行至少包括CRC解码的处理块。

[0090] 在至少一个实施例中，物理上行链路控制信道(PUCCH)发射器和物理下行链路控制信道(PDCCH)接收器在5G NR BBU中的一个或更多个 PPU(例如GPU 504)上实现。在至少一个实施例中，PUCCH和PDCCH 运载关于与PUSCH和PDSCH信道相关的传输格式和资源分配的信息，如本文进一步描述的。在至少一个实施例中，被实现为软件内核以供诸如 GPU 504之类的一个或更多个PPU执行的特定PUCCH运算包括PUCCH 格式0接收器、PUCCH格式1接收器、PUCCH格式2接收器、PUCCH 格式3接收器、PUCCH格式4接收器、极性解码器和PUCCH解码器。在至少一个实施例中，特定的GPU内核实现PUCCH极性解码内核，其具体执行至少包括极性解码器的处理块。在至少一个实施例中，特定的GPU内核实现PUCCH格式0接收器内核，其具体执行至少包括PUCCH序列检测器和单工解码器的处理块。在至少一个实施例中，特定的GPU内核实现 PUCCH格式1接收器内核，其具体执行至少包括PUCCH格式1信道估计器、PUCCH格式1均衡和解调以及PUCCH格式1解码器的处理块。在至少一个实施例中，特定的GPU内核实现PUCCH格式2接收器内核，其具体执行至少包括PUCCH格式2序列生成器、PUCCH格式2解复用器、 PUCCH格式2均衡和解调以及PUCCH格式2解码器的处理块。在至少一个实施例中，特定的GPU内核实现PUCCH格式3接收器内核，其具体执行至少包括PUCCH格式3序列生成器、PUCCH格式3解复用器、PUCCH 格式3均衡和解调、以及PUCCH格式3解码器的处理块。在至少一个实施例中，特定的GPU内核实现PUCCH格式4接收器内核，其具体执行至少包括PUCCH格式4信道估计和PUCCH格式4均衡和解调的处理块。

[0091] 在至少一个实施例中，支持物理随机接入信道(PRACH)的组件在 5G NR BBU中的一个或更多个PPU(例如GPU 504)上实现。在至少一个实施例中，PRACH为促进小区发现提供支持，如本文进一步描述的。在至少一个实施例中，实现为软件内核以供诸如GPU 504之类的一个或更多个PPU执行的特定PRACH运算包括PRACH序列生成、PRACH相关器和阈值计算以及PRACH检测器。在至少一个实施例中，特定的GPU内核实现PRACH序列生成内核，其具体执行至少包括PRACH序列生成的处理块。在至少一个实施例中，特定的GPU内核实现PRACH相关器和阈值计算内核，其具体执行至少包括PRACH相关器和阈值计算的处理块。在至少一个实施例中，特定的GPU内核实现PRACH检测器内核，其具体执行至少包括PRACH检测器的处理块。

[0092] 在至少一个实施例中，其他5G NR PHY组件在5G NR BBU中的一个或更多个PPU(例如GPU 504)上实现。在至少一个实施例中，其他组件包括低PHY运算和控制信息传输组件，其包括LDPC编码和解码、快速傅立叶变换(FFT)、快速傅立叶逆变换(IFFT)、解调参考信号(DMRS)、相位跟踪参考信号(PTRS)、探测参考信号(SRS)和信道状态信息参考信号(CSI-RS)，以及执行低PHY和控制信息传输的其他运算，在此进一步描述。在至少一个实施例中，其他组件包括用于初始接入和小区发现的信号处理组件，包括物理广播信道(PBCH)和一个或更多个同步信号块(SS块)，在此进一步描述。在至少一个实施例中，其他组件进一步包括低PHY运算，例如循环前缀(CP)内插，用于促进卷积以及传输波束成形(Tx波束成形)和接收

波束成形(Rx波束成形),本文进一步描述。在至少一个实施例中,与Tx波束成形和Rx波束成形相关的其他组件包括实现为GPU 504内核的波束成形权重应用。

[0093] 在至少一个实施例中,一个或更多个远程无线电单元(RRU)或无线电单元(RU)508包含一个或更多个天线。在至少一个实施例中,一个或更多个RRU或RU 508将无线电信号转换为传输到一个或更多个网络接口控制器506的数据,如本文所述。在至少一个实施例中,一个或更多个NIC 506通过BBU中的通信信道(包括通信总线)向一个或更多个CPU 502 提供数据样本。在至少一个实施例中,一个或更多个NICS 506通过BBU 中的通信信道(包括通信总线)向一个或更多个PPU(包括GPU 506)的共享存储器提供数据样本。

[0094] 在至少一个实施例中,使用一个或更多个PPU(包括GPU 504)的 BBU被配置为处理不同的载波配置。在至少一个实施例中,BBU中的参数是可配置的,包括载波带宽、TDD或FDD运算、接收层的数量和接收器天线的数量508、发射层的数量和发射器天线的数量508、参数学、时隙格式、参考信号格式、调制和编码方案、传输块大小以及指定为由3GPP 动态、半静态和静态配置的任何其他参数。在至少一个实施例中,对于动态、半静态和静态配置,使用一个或更多个PPU(包括GPU 504)的BBU 通过在内核启动时将特定配置参数传递给如上所述的每个软件内核来配置。在至少一个实施例中,对于动态、半静态和静态配置,使用一个或更多个 PPU(包括GPU 504)的BBU通过选择如上所述的适当的软件内核来配置,以基于可用配置参数来启动。在至少一个实施例中,对于半静态和静态配置,通过为该配置选择硬件和/或虚拟化配置文件来配置使用一个或更多个 PPU(包括GPU 504)的BBU。

[0095] 在至少一个实施例中,小区扇区载波是在给定小区站点和扇区上传输的一个5G NR载波。在至少一个实施例中,使用一个或更多个PPU(包括 GPU 504)的BBU支持多个小区扇区载波。在至少一个实施例中,对应于给定小区站点的所有扇区载波由包含一个或更多个PPU(包括GPU 504) 的单个BBU处理。在至少一个实施例中,分布式地理位置中的小区扇区载波由包含一个或更多个PPU(包括GPU 504)的单个BBU处理。在至少一个实施例中,由包含一个或更多个PPU(包括GPU 504)的单个BBU 支持不同频带和具有不同配置的小区扇区载波,例如带宽、参数、接收和发射层的数量以及接收和发射天线的数量。在至少一个实施例中,多个小区扇区载波优化包括多个PHY信道管线,每个由多个PPU或GPU 516软件内核启动组成,其中每个处理一个小区扇区载波。在至少一个实施例中,多个小区扇区载波优化包括封装PHY信道管线的功能的C++编程语言类,其中多个C++类对象被实例化以处理多个小区扇区载波。在至少一个实施例中,多个小区扇区载波优化包括将C++类中的每个PHY管线对象与 CUDA流相关联,以使CPU 502能够独立于其他PHY管线将工作负载提交给包括GPU 504的PPU。在至少一个实施例中,多个小区扇区载波优化包括通过聚合输入由单个PHY信道管线处理多个小区扇区载波。在至少一个实施例中,多个小区扇区载波优化包括将接收到的信号堆叠在存储器中并且被视为具有更高带宽或等效地具有更高数量的PRB的单个管线,如本文所述。

[0096] 图6示出了用于在诸如图形处理单元(GPU)之类的一个或更多个并行处理单元(PPU)上执行第五代(5G)新无线电(NR)基带单元(BBU) 物理层(PHY)运算的过程。在至少一个实施例中,用于执行5G NR BBU PHY运算的过程通过确定信号是否被接收或将被发送604而从602开始。在至少一个实施例中,如果信号将被发送604,则用于传输的数据从上层 606接收,如本文所述。在至少一个实施例中,如果一个或更多个PPU(例如GPU)将用于加速5G

PHY运算608,则数据被复制到与一个或更多个 PPU(例如GPU 610)相关联的共享存储器。在至少一个实施例中,如果不使用加速608,则将数据复制到与一个或更多个中央处理单元(CPU) 相关联的共享存储器618。

[0097] 在至少一个实施例中,如果加速被用于数据传输,则CPU调用一个或更多个PPU(例如GPU)上的个体或一组函数或内核612,如上所述。在至少一个实施例中,如上所述,那些函数或内核在一个或更多个PPU(例如GPU)上执行5G PHY信号处理运算614。在至少一个实施例中,一旦加速的5G PHY处理完成,数据就直接或间接地从一个或更多个PPU(例如GPU)被复制到一个或更多个网络接口控制器(NIC)616。

[0098] 在至少一个实施例中,如果没有将加速用于数据传输,则在一个或更多个CPU 620上执行5G PHY信号处理运算。在至少一个实施例中,一旦一个或更多个CPU上的5G PHY处理完成,则信号数据从与一个或更多个 CPU相关联的存储器被复制到与一个或更多个NIC 622相关联的存储器,以用于传输。

[0099] 在至少一个实施例中,如果接收到信号604并且正在使用诸如GPU 之类的一个或更多个PPU的加速624,则将信号数据从与一个或更多个NIC 相关联的存储器复制到与一个或更多个PPU(例如GPU 626)相关联的存储器。在至少一个实施例中,一个或更多个CPU调用实现各个5G PHY信号处理运算628的函数或内核,如上所述,以执行5G PHY处理630。在至少一个实施例中,一旦在一个或更多个PPU(例如GPU)上的5G PHY 处理完成,则数据从与一个或更多个PPU(例如GPU)相关联的存储器直接或间接复制到与一个或更多个CPU相关联的存储器,以用于传播到上层 632,如上所述。

[0100] 在至少一个实施例中,如果接收到信号604并且没有使用诸如GPU 之类的一个或更多个PPU的加速624,则将信号数据从与一个或更多个NIC 相关联的存储器直接复制到与一个或更多个CPU 634相关联的存储器。在至少一个实施例中,一个或更多个CPU然后执行5G PHY信号处理636,并且数据被传播到上层638。在至少一个实施例中,一旦5G PHY层运算完成,用于执行5G NR BBU PHY运算的过程完成640。

[0101] 数据中心

[0102] 图7示出了可以使用至少一个实施例的示例数据中心700。在至少一个实施例中,数据中心700包括数据中心基础设施层710、框架层720、软件层730和应用程序层740。

[0103] 在至少一个实施例中,如图7所示,数据中心基础设施层710可以包括资源协调器712、分组的计算资源714和节点计算资源(“节点C.R.”) 716(1)-716(N),其中“N”表示任何整数、正整数。在至少一个实施例中,节点C.R.716(1)-716(N)可以包括但不限于任何数量的中央处理单元(“CPU”)或其他处理器(包括加速器、现场可编程门阵列(FPGA)、图形处理器等),存储器设备(例如动态只读存储器),存储设备(例如固态硬盘或磁盘驱动器),网络输入/输出(“NW I/O”)设备,网络交换机,虚拟机(“VM”),电源模块和冷却模块等。在至少一个实施例中,节点 C.R.716(1)-716(N)中的一个或更多个节点C.R.可以是具有一个或更多个上述计算资源的服务器。

[0104] 在至少一个实施例中,分组的计算资源714可以包括容纳在一个或更多个机架内的节点C.R.的单独分组(未示出),或者容纳在各个地理位置的数据中心内的许多机架(也未示出)。分组的计算资源714内的节点C.R. 的单独分组可以包括可以被配置或分配为支持一个或更多个工作负载的分组的计算、网络、存储器或存储资源。在至少一个实施例中,

可以将包括 CPU或处理器的几个节点C.R.分组在一个或更多个机架内，以提供计算资源来支持一个或更多个工作负载。在至少一个实施例中，一个或更多个机架还可以包括任何数量的电源模块、冷却模块和网络交换机，以任意组合。

[0105] 在至少一个实施例中，资源协调器712可以配置或以其他方式控制一个或更多个节点C.R.716(1)-716(N)和/或分组的计算资源714。在至少一个实施例中，资源协调器712可以包括用于数据中心700的软件设计基础结构(“SDI”)管理实体。在至少一个实施例中，资源协调器可以包括硬件、软件或其某种组合。

[0106] 在至少一个实施例中，如图7所示，框架层720包括作业调度器732、配置管理器734、资源管理器736和分布式文件系统738。在至少一个实施例中，框架层720可以包括支持软件层730的软件732和/或应用程序层740 的一个或更多个应用程序742的框架。在至少一个实施例中，软件732或应用程序742可以分别包括基于Web的服务软件或应用程序，例如由Amazon Web Services，Google Cloud和Microsoft Azure提供的服务或应用程序。在至少一个实施例中，框架层720可以是但不限于一种免费和开放源软件网络应用程序框架，例如可以利用分布式文件系统738来进行大范围数据处理(例如“大数据”)的Apache Spark™(以下称为“Spark”)。在至少一个实施例中，作业调度器732可以包括Spark驱动器，以促进对数据中心700的各个层所支持的工作负载进行调度。在至少一个实施例中，配置管理器734可以能够配置不同的层，例如软件层730和包括Spark和用于支持大规模数据处理的分布式文件系统738的框架层720。在至少一个实施例中，资源管理器736能够管理映射到或分配用于支持分布式文件系统738和作业调度器732的集群或分组计算资源。在至少一个实施例中，集群或分组计算资源可以包括数据中心基础设施层710上的分组计算资源 714。在至少一个实施例中，资源管理器736可以与资源协调器712协调以管理这些映射的或分配的计算资源。

[0107] 在至少一个实施例中，包括在软件层730中的软件732可以包括由节点C.R.716(1)-716(N)的至少一部分，分组的计算资源714和/或框架层720的分布式文件系统738使用的软件。一个或更多个类型的软件可以包括但不限于Internet网页搜索软件、电子邮件病毒扫描软件、数据库软件和流视频内容软件。

[0108] 在至少一个实施例中，应用程序层740中包括的一个或更多个应用程序742可以包括由节点C.R.716(1)-716(N)的至少一部分、分组计算资源714和/或框架层720的分布式文件系统738使用的一个或更多个类型的应用程序。一个或更多个类型的应用程序可以包括但不限于任何数量的基因组学应用程序、认知计算和机器学习应用程序，包括训练或推理软件，机器学习框架软件(例如PyTorch、TensorFlow、Caffe等)或其他与一个或更多个实施例结合使用的机器学习应用程序。

[0109] 在至少一个实施例中，配置管理器734、资源管理器736和资源协调器712中的任何一个可以基于以任何技术上可行的方式获取的任何数量和类型的数据来实现任何数量和类型的自我修改动作。在至少一个实施例中，自我修改动作可以减轻数据中心700的数据中心操作员做出可能不好的配置决定并且可以避免数据中心的未充分利用和/或执行差的部分。

[0110] 在至少一个实施例中，数据中心700可以包括工具、服务、软件或其他资源，以根据本文所述的一个或更多个实施例来训练一个或更多个机器学习模型或者使用一个或更多

个机器学习模型来预测或推理信息。例如,在至少一个实施例中,可以通过使用上文关于数据中心700描述的软件和计算资源,根据神经网络架构通过计算权重参数来训练机器学习模型。在至少一个实施例中,通过使用通过本文所述的一个或更多个训练技术计算出的权重参数,可以使用上面与关于数据中心700所描述的资源,使用对应于一个或更多个神经网络的经训练的机器学习模型来推理或预测信息。

[0111] 在至少一个实施例中,数据中心可以使用CPU、专用集成电路(ASIC)、GPU、FPGA或其他硬件来使用上述资源来执行训练和/或推理。此外,上述的一个或更多个软件和/或硬件资源可以配置成一种服务,以允许用户训练或执行信息推理,例如图像识别、语音识别或其他人工智能服务。

[0112] 自主车辆

[0113] 图8A示出了根据至少一个实施例的自动驾驶车辆800的示例。在至少一个实施例中,自动驾驶车辆800(在本文中可替代地称为“车辆800”) 可以是但不限于客运车辆,例如汽车、卡车、公共汽车和/或可容纳一个或更多个乘客的另一种类型的车辆。在至少一个实施例中,车辆800可以是用于拖运货物的半牵引车-拖车。在至少一个实施例中,车辆800可以是飞机、机器人车辆或其他类型的车辆。

[0114] 可以根据由美国运输部下属的国家公路交通安全管理局(“NHTSA”) 和汽车工程师学会(“SAE”)“与用于道路机动车辆的驾驶自动化系统有关的术语(Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles)”(例如,于2018年6月15日发布的标准号J3016-201806,于2016年9月30日发布的标准号J3016-201609,以及该版本的以前和将来的版本此标准)定义的自动化级别来描述自动驾驶汽车。在一个或更多个实施例中,车辆800可能能够根据自动驾驶级别的级别1-级别5中的一个或更多个来进行功能。例如,在至少一个实施例中,根据实施例,车辆800可能能够进行条件自动化(级别3)、高度自动化(级别4)和/或全自动(级别5)。

[0115] 在至少一个实施例中,车辆800可以包括但不限于组件,诸如底盘、车身、车轮(例如2、4、6、8、18等)、轮胎、车轴和车辆的其他组件。在至少一个实施例中,车辆800可以包括但不限于推进系统850,例如内燃机、混合动力装置、全电动发动机和/或另一种推进系统类型。在至少一个实施例中,推进系统850可以连接至车辆800的传动系,其可以包括但不限于变速器,以使得能够对车辆800进行推进。在至少一个实施例中,可以响应于从油门/加速器852接收信号以控制推进系统850。

[0116] 在至少一个实施例中,当推进系统850正在运行时(例如,当车辆行驶时),转向系统854(其可以包括但不限于方向盘)用于使车辆800转向(例如,沿着期望的路径或路线)。在至少一个实施例中,转向系统854 可以从转向致动器856接收信号。方向盘对于全自动化(级别5)功能可以是可选的。在至少一个实施例中,制动传感器系统846可以用于响应于从制动致动器848和/或制动传感器接收到的信号来操作车辆制动器。

[0117] 在至少一个实施例中,控制器836可以包括但不限于一个或更多个片上系统(“SoC”)(图8A中未示出)和/或图形处理单元(“GPU”)向车辆 800的一个或更多个组件和/或系统提供信号(例如,代表命令)。例如,在至少一个实施例中,控制器836可以发送信号以通过制动致动器848操作车辆制动,通过一个或更多个转向致动器856操作转向系统854,通过一个或更多个油门(throttle)/加速器852操作推进系统850。控制器836 可以包括一个

或更多个机载(例如,集成)计算设备(例如,超级计算机),其处理传感器信号并输出操作命令(例如,表示命令的信号)以实现自动驾驶和/或协助驾驶员驾驶车辆800。在至少一个实施例中,一个或更多个控制器836可以包括用于自动驾驶功能的第一控制器836,用于功能安全功能的第二控制器836,用于人工智能功能(例如计算机视觉)的第三控制器836,用于信息娱乐功能的第四控制器836,用于紧急情况下的冗余的第五控制器836和/或其他控制器。在至少一个实施例中,单个控制器836 可以处理上述功能中的两个或更多个,两个或更多控制器836可以处理单个功能和/或其任何组合。

[0118] 在至少一个实施例中,一个或更多个控制器836响应于从一个或更多个传感器(例如,传感器输入)接收到的传感器数据,提供用于控制车辆 800的一个或更多个组件和/或系统的信号。在至少一个实施例中,传感器数据可以从传感器接收,传感器类型例如但不限于一个或更多个全球导航卫星系统(“GNSS”)传感器858(例如,一个或更多个全球定位系统传感器)、一个或更多个RADAR传感器860、一个或更多个超声波传感器862、一个或更多个LIDAR传感器864、一个或更多个惯性测量单元(IMU)传感器866(例如,一个或更多个加速度计、一个或更多个陀螺仪、一个或更多个磁罗盘、一个或更多个磁力计等)、一个或更多个麦克风896、一个或更多个立体声相机868、一个或更多个广角相机870(例如鱼眼相机)、一个或更多个红外相机872、一个或更多个环绕相机874(例如,360度相机)、远程相机(图8A中未示出)、中程相机(图8A中未示出)、一个或更多个速度传感器844(例如,用于测量车辆800的速度)、一个或更多个振动传感器842、一个或更多个转向传感器840、一个或更多个制动传感器(例如,作为制动传感器系统846的一部分)和/或其他传感器类型接收。

[0119] 在至少一个实施例中,一个或更多个控制器836可以从车辆800的仪表板832接收输入(例如,由输入数据表示)并通过人机接口(“HMI”) 显示器834、声音信号器、扬声器和/或车辆800的其他组件提供输出(例如,由输出数据、显示数据等表示)。在至少一个实施例中,输出可包括信息,诸如车速、速度、时间、地图数据(例如,高清晰度地图(图8A 中未显示)、位置数据(例如,车辆800的位置,例如在地图上)、方向、其他车辆的位置(例如,占用光栅)、关于对象的信息以及由一个或更多个控制器836感知到的对象的状态等。例如,在至少一个实施例中,HMI 显示器834可以显示关于一个或更多个对象的存在的信息(例如,路牌、警告标志、交通信号灯变更等)和/或有关驾驶操作车辆已经、正在或将要制造的信息(例如,现在改变车道、在两英里内驶出34B出口等)。

[0120] 在至少一个实施例中,车辆800进一步包括网络接口824,其可以使用一个或更多个无线天线826和/或一个或更多个调制解调器通过一个或更多个网络进行通信。例如,在至少一个实施例中,网络接口824可能能够通过长期演进(“LTE”)、宽带码分多址(“WCDMA”)、通用移动电信系统(“UMTS”)、全球移动通信系统(“GSM”)、IMT-CDMA多载波(“CDMA2000”)网络等进行通信。在至少一个实施例中,一个或更多个无线天线826还可以使用一个或更多个局域网(例如Bluetooth、Bluetooth Low Energy(LE)、Z-Wave、ZigBee等)和/或一个或更多个低功耗广域网(以下简称“LPWAN”)(例如LoRaWAN、SigFox等协议),使环境中的对象(例如,车辆、移动设备)之间进行通信。

[0121] 图8B示出了根据至少一个实施例的图8A的自动驾驶车辆800的相机位置和视野的示例。在至少一个实施例中,相机和各自的视野是一个示例实施例,并且不旨在进行限制。例如,在至少一个实施例中,可以包括附加的和/或替代的相机和/或相机可以位于车辆800

上的不同位置。

[0122] 在至少一个实施例中，用于相机的相机类型可以包括但不限于可以适于与车辆800的组件和/或系统一起使用的数字相机。相机可以以汽车安全完整性等级(“ASIL”)B和/或其他ASIL进行操作。在至少一个实施例中，根据实施例，相机类型可以具有任何图像捕获速率，例如60帧每秒(fps)、120fps、240fps等。在至少一个实施例中，相机可以能够使用滚动快门、全局快门、另一种类型的快门或其组合。在至少一个实施例中，滤色器阵列可以包括红色透明透明(“RCCC”)滤色器阵列、红色透明透明蓝色 (“RCCB”)滤色器阵列、红色蓝色绿色透明(“RBGC”)滤色器阵列、Foveon X3滤色器阵列、拜耳(Bayer)传感器(“RGGB”)滤色器阵列、单色传感器滤色器阵列和/或其他类型的滤色器阵列。在至少一个实施例中，可以使用透明像素相机，例如具有RCCC、RCCB和/或RBGC滤色器阵列的相机，以努力提高光敏性。

[0123] 在至少一个实施例中，一个或更多个相机可以用于执行先进驾驶员辅助系统(“ADAS”)功能(例如，作为冗余或故障安全设计的一部分)。例如，在至少一个实施例中，可以安装多功能单声道相机以提供包括车道偏离警告、交通标志辅助和智能大灯控制的功能。在至少一个实施例中，一个或更多个相机(例如，所有相机)可以同时记录并提供图像数据(例如，视频)。

[0124] 在至少一个实施例中，可以将一个或更多个相机安装在安装组件中，例如定制设计的(三维(“3D”)打印的)组件，以便切出杂散光和来自在汽车内的反光(例如，仪表板的反射在挡风玻璃镜中反光)，其可能会干扰相机的图像数据捕获能力。关于后视镜安装组件，在至少一个实施例中，后视镜组件可以是3D打印定制的，使得相机安装板匹配后视镜的形状。在至少一个实施例中，一个或更多个相机可以被集成到后视镜中。在至少一个实施例中，对于侧视相机，一个或更多个相机也可以集成在汽车的每个角落的四个支柱内。

[0125] 在至少一个实施例中，具有包括车辆800前面的环境的部分的视野的相机(例如，前向相机)可以用于环视，以及在一个或更多个控制器836 和/或控制SoC的帮助下帮助识别向前的路径和障碍物，从而提供对于生成占用网格和/或确定优选的车辆路径至关重要的信息。在至少一个实施例中，前向相机可以用于执行许多与LIDAR相同的ADAS功能，包括但不限于紧急制动、行人检测和避免碰撞。在至少一个实施例中，前向相机也可以用于ADAS功能和系统，包括但不限于车道偏离警告(“LDW”)、自动巡航控制(“ACC”)和/或其他功能(例如交通标志识别)。

[0126] 在至少一个实施例中，各种相机可以用于前向配置，包括例如包括 CMOS(“互补金属氧化物半导体”)彩色成像器的单目相机平台。在至少一个实施例中，广角相机870可以用于感知从外围进入的对象(例如，行人、过马路或自行车)。尽管在图8B中仅示出了一个广角相机870，但是，在其他实施例中，车辆800上可以有任何数量(包括零)的广角相机。在至少一个实施例中，任何数量的远程相机898(例如，远程立体相机对) 可用于基于深度的对象检测，尤其是对于尚未训练神经网络的对象。在至少一个实施例中，远程相机898也可以用于对象检测和分类以及基本对象跟踪。

[0127] 在至少一个实施例中，任何数量的立体声相机868也可以包括在前向配置中。在至少一个实施例中，一个或更多个立体声相机868可以包括集成控制单元，该集成控制单元包括可缩放处理单元，该可缩放处理单元可以提供可编程逻辑(“FPGA”)和具有单个芯片上集成的控制器局域网 (“CAN”)或以太网接口的多核心微处理器。在至少一个实施例中，这样

的单元可以用于生成车辆800的环境的3D地图，包括对图像中所有点的距离估计。在至少一个实施例中，一个或更多个立体相机868可以包括但不限于紧凑型立体视觉传感器，其可以包括但不限于两个相机镜头(左右分别一个)和一个图像处理芯片，其可以测量从车辆800到目标对象的距离并使用所生成的信息(例如，元数据)来激活自主紧急制动和车道偏离警告功能。在至少一个实施例中，除了本文所述的那些之外，还可以使用其他类型的立体相机868。

[0128] 在至少一个实施例中，具有包括车辆800侧面的环境的一部分的视野的相机(例如，侧视相机)可以用于环绕查看，从而提供用于创建和更新占据网格的信息，以及产生侧面碰撞警告。例如，在至少一个实施例中，环绕相机874(例如，如图8B所示的四个环绕相机874)可以定位在车辆 800上。一个或更多个环绕相机874可以包括但不限于，任意数量和组合的广角相机870、一个或更多个鱼目镜头、一个或更多个360度相机和/或类似相机。例如，在至少一个实施例中，四个鱼目镜头相机可以位于车辆 800的前、后和侧面。在至少一个实施例中，车辆800可以使用三个环绕相机874(例如，左、右和后面)，并且可以利用一个或更多个其他相机 (例如，前向相机)作为第四个环视相机。

[0129] 在至少一个实施例中，具有包括车辆800后方的环境的一部分的视野的相机(例如，后视相机)可以用于停车辅助、环视、后方碰撞警告、以及创建和更新占用光栅。在至少一个实施例中，可以使用各种各样的相机，包括但不限于还适合作为一个或更多个前向相机的相机(例如，远程相机 898和/或一个或更多个中程相机876、一个或更多个立体相机868、一个或更多个红外相机872等)，如本文所述。

[0130] 图8C示出了根据至少一个实施例的图8A的自动驾驶车辆800的示例系统架构的框图。在至少一个实施例中，图8C中的车辆800的一个或更多个组件、一个或更多个特征和一个或更多个系统中的每一个都示出为经由总线802连接。在至少一个实施例中，总线802可包括但不限于CAN 数据接口(在本文中可替代地称为“CAN总线”)。在至少一个实施例中，CAN可以是车辆800内部的网络，用于帮助控制车辆800的各种特征和功能，例如制动器的致动、加速、制动、转向、雨刷等。在一个实施例中，总线802可以配置成具有数十个甚至数百个节点，每个节点具有其自己的唯一识别符(例如，CAN ID)。在至少一个实施例中，可以读取总线802 以找到方向盘角度、地面速度、发动机每分钟转动次数(“RPM”)、按钮位置和/或其他车辆状态指示器。在至少一个实施例中，总线802可以是符合ASIL B的CAN总线。

[0131] 在至少一个实施例中，除了CAN之外或来自CAN，可使用FlexRay 和/或以太网(Ethernet)。在至少一个实施例中，可以有任意数量的总线 802，其可以包括但不限于零或更多的CAN总线，零或更多的FlexRay总线，零或更多的以太网总线，和/或零或更多的使用其他协议的其他类型的总线。在至少一个实施例中，两个或更多个总线802可以用于执行不同的功能，和/或可以用于冗余。例如，第一总线802可以用于碰撞避免功能，并且第二总线802可以用于致动控制。在至少一个实施例中，每个总线802 可以与车辆800的任何组件通信，并且总线802的两个或更多个可以与相同的组件通信。在至少一个实施例中，任何数量的片上系统(“SoC”)804 中的每一个，一个或更多个控制器836中的每一个和/或车辆内的每个计算机都可以访问相同的输入数据(例如，来自车辆800的传感器的输入)，并且可以连接到公共总线，例如CAN总线。

[0132] 在至少一个实施例中，车辆800可以包括一个或更多个控制器836，诸如本文关于

图8A所描述的那些。控制器836可以用于多种功能。在至少一个实施例中，控制器836可以耦合到车辆800的各种其他组件和系统中的任何一个，并且可以用于控制车辆800、车辆800的人工智能、车辆 800的信息娱乐和/或其他功能。

[0133] 在至少一个实施例中，车辆800可以包括任何数量的SoC 804。SoC 804中的每一个可以包括但不限于中央处理单元(“一个或更多个CPU”) 806、图形处理单元(“一个或更多个GPU”)808、一个或更多个处理器810、一个或更多个高速缓存812、一个或更多个加速器814、一个或更多个数据存储816和/或其他未显示的组件和特征。在至少一个实施例中，一个或更多个SoC 804可以用于在各种平台和系统中控制车辆800。例如，在至少一个实施例中，一个或更多个SoC 804可以与高清晰度(“HD”)地图822 在系统(例如，车辆800的系统)中组合，该高清晰度地图822可以经由网络接口824从一个或更多个服务器(图8C中未示出)获得地图刷新和/ 或更新。

[0134] 在至少一个实施例中，一个或更多个CPU 806可以包括CPU集群或 CPU复合体(在本文中可替代地称为“CCPLEX”)。在至少一个实施例中，一个或更多个CPU 806可以包括多个核心和/或二级(“L2”)高速缓存。例如，在至少一个实施例中，一个或更多个CPU 806可以在相互耦合的多处理器配置中包括八个核心。在至少一个实施例中，一个或更多CPU 806可以包括四个双核心集群，其中每个集群具有专用的L2高速缓存(例如， 2MB L2高速缓存)。在至少一个实施例中，一个或更多CPU 806(例如， CCPLEX)可以配置成支持同时的集群操作，使得一个或更多CPU 806的集群的任何组合在任何给定的时间都可以是活跃的。

[0135] 在至少一个实施例中，一个或更多个CPU 806可以实现电源管理功能，这些功能包括但不限于以下特征中的一个或更多个：空闲时可以自动对各个硬件模块进行时钟门控以节省动态功率；当该核心由于执行等待中断(“WFI”)/事件等待(“WFE”)指令而未主动执行指令时，可以对每个核心时钟进行门控；每个核心都可以独立供电；当所有核心都被时钟门控或功率门控时，每个核心集群可以被独立地时钟门控；以及/或当所有核心都被功率门控时，每个核心集群可以被独立地功率门控。在至少一个实施例中，一个或更多CPU 806可以进一步实现用于管理功率状态的增强算法，其中指定了允许的功率状态和预期的唤醒时间，并且硬件/微码确定了针对核心、集群和CCPLEX输入的最佳功率状态。在至少一个实施例中，处理核心可以在软件中支持简化的功率状态输入序列，其中工作被分担给微码。

[0136] 在至少一个实施例中，一个或更多个GPU 808可以包括集成的GPU (在本文中或者称为“iGPU”)。在至少一个实施例中，一个或更多个GPU 808可以是可编程的，并且对于并行工作负载可以是有效的。在至少一个实施例中，一个或更多个GPU 808可以使用增强的张量指令集。在一个实施例中，一个或更多个GPU 808可以包括一个或更多个流式微处理器，其中每个流式微处理器可以包括一级(“L1”)高速缓存(例如，具有至少96KB 的存储容量的L1高速缓存)，以及两个或更多个流式微处理器可以共享 L2高速缓存(例如，具有512KB存储容量的L2高速缓存)。在至少一个实施例中，一个或更多个GPU 808可以包括至少八个流式微处理器。在至少一个实施例中，一个或更多个GPU 808可以使用计算应用程序编程接口(API)。在至少一个实施例中，一个或更多GPU 808可以使用一个或更多个并行计算平台和/或编程模型(例如，NVIDIA的CUDA)。

[0137] 在至少一个实施例中，一个或更多个GPU 808可以经功耗优化以在汽车和嵌入式用例中获得最佳性能。例如，在一个实施例中，可以在鳍式场效应晶体管(“FinFET”)上制造

一个或更多个GPU 808。在至少一个实施例中，每个流式微处理器可以包含多个划分为多个块的混合精度处理核心，。例如但不限于，可以将64个PF32核心和32个PF64核心划分为四个处理块。在至少一个实施例中，可以为每个处理块分配16个FP32核心、8个FP64核心、16个INT32核心、两个用于深度学习矩阵算术的混合精度NVIDIA张量核心、零级(“L0”)指令缓存、线程束调度器、分派单元和/或64KB寄存器文件。在至少一个实施例中，流式微处理器可以包括独立的并行整数和浮点数据路径来提供混合了计算和寻址运算的工作量的有效执行。在至少一个实施例中，流式微处理器可以包括独立的线程调度能力，以实现更细粒度的同步和并行线程之间的协作。在至少一个实施例中，流式微处理器可以包括组合的L1数据高速缓存和共享存储器单元，以便在简化编程的同时提高性能。

[0138] 在至少一个实施例中，一个或更多个GPU 808可以包括高带宽存储器(“HBM”)和/或16GB HBM2存储器子系统，以在一些示例中提供约 900GB/秒的峰值存储带宽。在至少一个实施例中，除了或替代于HBM存储器，可以使用同步图形随机存取存储器(“SGRAM”)，例如图形双倍数据速率类型的五同步随机存取存储器(“GDDR5”)。

[0139] 在至少一个实施例中，一个或更多个GPU 808可以包括统一存储器技术。在至少一个实施例中，地址转换服务(“ATS”)支持可以用于允许一个或更多个GPU 808直接访问一个或更多个CPU 806页表。在至少一个实施例中，当一个或更多个GPU 808存储器管理单元(“MMU”)经历未命中时，可以将地址转换请求发送到一个或更多个CPU 806。作为响应，在至少一个实施例中，一个或更多个CPU 806可以在其页面表中查找地址的虚拟-物理的映射并将转换传送回一个或更多个GPU 808。在至少一个实施例中，统一存储器技术可以允许单个统一虚拟地址空间用于一个或更多个CPU 806和一个或更多个GPU 808两者的存储器，从而简化了一个或更多个GPU 808的编程以及将应用程序移植到一个或更多个GPU 808。

[0140] 在至少一个实施例中，一个或更多个GPU 808可以包括任意数量的访问计数器，其可以跟踪一个或更多个GPU 808对其他处理器的存储器的访问频率。在至少一个实施例中，一个或更多个访问计数器可以帮助确保将存储器页移动到最频繁访问页面的处理器的物理存储器中，从而提高处理器之间共享的存储器范围的效率。

[0141] 在至少一个实施例中，一个或更多个SoC 804可以包括任何数量的高速缓存812，包括本文所述的那些。例如，在至少一个实施例中，一个或更多个高速缓存812可以包括可用于一个或更多个CPU 806和一个或更多个GPU 808(例如，连接到CPU 806和GPU 808)的三级(“L3”)高速缓存。在至少一个实施例中，一个或更多个高速缓存812可以包括回写式高速缓存，该回写式高速缓存可以例如通过使用高速缓存相干协议(例如，MEI、MESI、MSI等)来跟踪线的状态。在至少一个实施例中，尽管可以使用较小的高速缓存大小，根据实施例，L3高速缓存可以包括4MB或更多。

[0142] 在至少一个实施例中，一个或更多个SoC 804可以包括一个或更多个加速器814(例如，硬件加速器、软件加速器或其组合)。在至少一个实施例中，一个或更多个SoC 804可以包括硬件加速集群，其可以包括优化的硬件加速器和/或大的片上存储器。在至少一个实施例中，大的片上存储器(例如4MB的SRAM)可以使硬件加速集群能够加速神经网络和其他计算。在至少一个实施例中，硬件加速集群可以用于补充一个或更多个 GPU 808并且卸载一个或更多个GPU 808的一些任务(例如，释放一个或更多个GPU 808的更多周期以执行其他任务)。在至少一个实施例中，一个或更多个加速器814可以用于足够稳定以经得起加速

检验的目标工作负载(例如,感知、卷积神经网络(“CNN”)、递归神经网络(“RNN”)等)。在至少一个实施例中,CNN可以包括基于区域或区域卷积神经网络 (“RCNN”)和快速RCNN(例如,如用于对象检测)或其他类型的CNN。

[0143] 在至少一个实施例中,一个或更多个加速器814(例如,硬件加速集群)可以包括一个或更多个深度学习加速器(“DLA”)。一个或更多个DLA 可以包括但不限于一个或更多个Tensor处理单元(“TPU”),其可以配置成每秒提供额外的10万亿次操作用于深度学习应用程序和推理。在至少一个实施例中,TPU可以是配置成并被优化用于执行图像处理功能(例如,用于CNN、RCNN等)的加速器。可以针对神经网络类型和浮点运算以及推理的特定集合进一步优化一个或更多个DLA。在至少一个实施例中,一个或更多个DLA的设计可以提供比典型的通用GPU更高的每毫米性能,并且通常大大超过CPU的性能。在至少一个实施例中,一个或更多个TPU 可执行若干功能,包括支持例如INT8、INT16和FP16数据类型以用于特征和权重的单实例卷积功能以及后处理器功能的。在至少一个实施例中,一个或更多个DLA可以针对各种功能中的任何功能,在处理或未处理的数据上快速且有效地执行神经网络,尤其是CNN,包括例如但不限于:用于使用来自相机传感器的数据进行对象识别和检测的CNN;用于使用来自相机传感器的数据进行距离估算的CNN;用于使用来自麦克风896的数据进行紧急车辆检测以及识别和检测的CNN;用于使用来自相机传感器的数据进行人脸识别和车主识别的CNN;以及/或用于安全和/或安全相关事件的 CNN。

[0144] 在至少一个实施例中,DLA可以执行一个或更多个GPU 808的任何功能,并且通过使用推理加速器,例如,设计者可以将一个或更多个DLA 或一个或更多个GPU 808作为目标用于任何功能。例如,在至少一个实施例中,设计者可以将CNN的处理和浮点运算集中在一个或更多个DLA上,并将其他功能留给一个或更多个GPU 808和/或其他一个或更多个加速器 814。

[0145] 在至少一个实施例中,一个或更多个加速器814(例如,硬件加速集群)可以包括可编程视觉加速器(“PVA”),其在本文中可以可替代地称为计算机视觉加速器。在至少一个实施例中,一个或更多个PVA可以设计和配置为加速用于高级驾驶员辅助系统(“ADAS”)838、自动驾驶、增强现实(“AR”)应用程序和/或虚拟现实(“VR”)应用程序的计算机视觉算法。在至少一个实施例中,一个或更多个PVA可以在性能和灵活性之间取得平衡。例如,在至少一个实施例中,一个或更多个PVA中的每一个可以包括例如但不限于任何数量的精简指令集计算机(“RISC”)核心、直接存储器访问(“DMA”)和/或任意数量的向量处理器。

[0146] 在至少一个实施例中,RISC核心可以与图像传感器(例如,本文描述的任意相机的图像传感器)、图像信号处理器等交互。在至少一个实施例中,每个RISC核心可以包括任意数量的存储器。在至少一个实施例中,根据实施例,RISC核心可以使用多种协议中的任意一种。在至少一个实施例中,RISC核心可以执行实时操作系统(“RTOS”)。在至少一个实施例中,可以使用一个或更多个集成电路设备、专用集成电路(“ASIC”)和/ 或存储设备来实现RISC核心。例如,在至少一个实施例中,RISC核心可以包括指令高速缓存和/或紧密耦合的RAM。

[0147] 在至少一个实施例中,DMA可以使PVA的组件能够独立于一个或更多个CPU 806访问系统存储器。在至少一个实施例中,DMA可以支持用于向PVA提供优化的任何数量的特征,包括但不限于,支持多维寻址和 /或循环寻址。在至少一个实施例中,DMA可以支持多达六

个或更多个寻址的维度,其可以包括但不限于块宽度、块高度、块深度、水平块步进、垂直块步进和/或深度步进。

[0148] 在至少一个实施例中,向量处理器可以是可编程处理器,其可以设计为有效且灵活地执行用于计算机视觉算法并提供信号处理能力的编程。在至少一个实施例中,PVA可以包括PVA核心和两个向量处理子系统分区。在至少一个实施例中,PVA核心可以包括处理器子系统、DMA引擎(例如,两个DMA引擎)和/或其他外围设备。在至少一个实施例中,向量处理子系统可以用作PVA的主要处理引擎,并且可以包括向量处理单元("VPU")、指令高速缓存和/或向量存储器(例如"VMEM")。在至少一个实施例中,VPU核心可以包括数字信号处理器,例如,单指令多数据 ("SIMD")、超长指令字("VLIW")数字信号处理器。在至少一个实施例中,SIMD和VLIW的组合可以提高吞吐量和速度。

[0149] 在至少一个实施例中,每个向量处理器可以包括指令高速缓存并且可以耦合到专用存储器。结果,在至少一个实施例中,每个向量处理器可以配置为独立于其他向量处理器执行。在至少一个实施例中,特定PVA中包括的向量处理器可以配置为采用数据并行性。例如,在至少一个实施例中,单个PVA中包括的多个向量处理器可以执行相同计算机视觉算法,除了在图像的不同区域上之外。在至少一个实施例中,包括在特定PVA中的向量处理器可以在相同图像上同时执行不同的计算机视觉算法,或者甚至在序列图像或部分图像上执行不同的算法。在至少一个实施例中,除其他外,在硬件加速集群中可以包括任何数量的PVA,并且在每个PVA中可以包括任何数量的向量处理器。在至少一个实施例中,PVA可以包括附加的纠错码("ECC")存储器,以增强整体系统安全性。

[0150] 在至少一个实施例中,一个或更多个加速器814(例如,硬件加速集群)可以包括片上计算机视觉网络和静态随机存取存储器("SRAM"),用于为一个或更多个加速器814提供高带宽,低延迟SRAM。在至少一个实施例中,片上存储器可以包括至少4MB SRAM,其包括例如但不限于八个现场可配置的内存块,PVA和DLA均可以对其进行访问。在至少一个实施例中,每对存储块可以包括高级外围总线("APB")接口、配置电路、控制器和多路复用器。在至少一个实施例中,可以使用任何类型的存储器。在至少一个实施例中,PVA和DLA可以经由为PVA和DLA提供对存储器的高速访问的主干网来访问存储器。在至少一个实施例中,主干网可以包括片上计算机视觉网络,其将PVA和DLA互连到存储器(例如,使用 APB)。

[0151] 在至少一个实施例中,片上计算机视觉网络可以包括接口,该接口在传输任何控制信号/地址/数据之前确定PVA和DLA均提供就绪和有效信号。在至少一个实施例中,接口可以提供用于发送控制信号/地址/数据的单独的相位和单独的信道,以及用于连续数据传输的突发型通信。在至少一个实施例中,尽管可以使用其他标准和协议,但是接口可以符合国际标准化组织("ISO")26262或国际电工委员会("IEC")61508标准。

[0152] 在至少一个实施例中,一个或更多个SoC 804可以包括实时视线追踪硬件加速器。在至少一个实施例中,实时视线追踪硬件加速器可以用于快速且有效地确定对象的位置和范围(例如,在世界模型内),以生成实时可视化模拟,以用于RADAR信号解释,用于声音传播合成和/或分析,用于SONAR系统的模拟,用于一般的波传播模拟,与用于定位和/或其他功能的LIDAR数据进行比较,和/或用于其他用途。

[0153] 在至少一个实施例中,一个或更多个加速器814(例如,硬件加速集群)具有用于自动驾驶的广泛用途。在至少一个实施例中,PVA可以是可编程视觉加速器,其用于ADAS和自

动驾驶汽车中的关键处理阶段。在至少一个实施例中，在低功耗和低延迟下PVA的能力与需要可预测的处理的算法域良好匹配。换句话说，PVA在半密集或密集的常规计算中表现出色，即使在小型数据集上也是如此，这些数据集可能需要具有低延迟和低功耗的可预测的运行时间。在至少一个实施例中，自主车辆，诸如在车辆800 中PVA可能被设计为运行经典的计算机视觉算法，因为它们可以在对象检测和整数数学运算方面是有效的。

[0154] 例如，根据技术的至少一个实施例，PVA被用于执行计算机立体视觉。在至少一个实施例中，可以在一些示例中使用基于半全局匹配的算法，尽管这并不意味着限制性。在至少一个实施例中，用于3-5级自动驾驶的应用程序在运行中使用动态的估计/立体匹配(例如，从运动中恢复结构、行人识别、车道检测等)。在至少一个实施例中，PVA可以对来自两个单目相机的输入执行计算机立体视觉功能。

[0155] 在至少一个实施例中，PVA可以用于执行密集的光流。例如，在至少一个实施例中，PVA可以处理原始RADAR数据(例如，使用4D快速傅立叶变换)以提供处理后的RADAR数据。在至少一个实施例中，例如，通过处理原始飞行时间数据以提供处理后的飞行时间数据，将PVA用于飞行时间深度处理。

[0156] 在至少一个实施例中，DLA可用于运行任何类型的网络以增强控制和驾驶安全性，包括例如但不限于神经网络，其输出用于每个对象检测的置信度。在至少一个实施例中，可以将置信度表示或解释为概率，或者表示为提供每个检测相对于其他检测的相对“权重”。在至少一个实施例中，置信度使系统能够做出进一步的决定，即关于哪些检测应当被认为是真正的阳性检测而不是假阳性检测。例如，在至少一个实施例中，系统可以为置信度设置阈值，并且仅将超过阈值的检测视为真阳性检测。在使用自动紧急制动(“AEB”)系统的实施例中，假阳性检测将导致车辆自动执行紧急制动，这显然是不希望的。在至少一个实施例中，高度自信的检测可以被认为是AEB的触发。在至少一个实施例中，DLA可以运行用于回归置信度值的神经网络。在至少一个实施例中，神经网络可以将参数的至少一些子集作为其输入，例如包围盒尺寸，获得的地平面估计(例如，从另一子系统)，与从神经网络和/或其他传感器(例如，一个或更多个LIDAR 传感器864或一个或更多个RADAR传感器860)等获得的对象的车辆800 方向、距离、3D位置估计相关的一个或更多个IMU传感器866的输出。

[0157] 在至少一个实施例中，一个或更多个SoC 804(例如，硬件加速集群) 可以包括一个或更多个数据存储装置816(例如，存储器)。在至少一个实施例中，一个或更多个数据存储816可以是一个或更多个SoC 804的片上存储器，其可以存储要在一个或更多个GPU 808和/或DLA上执行的神经网络。在至少一个实施例中，一个或更多个数据存储816可以具有足够大的容量以存储神经网络的多个实例以用于冗余和安全。在至少一个实施例中，一个或更多个数据存储812可以包括L2或L3高速缓存。

[0158] 在至少一个实施例中，一个或更多个SoC 804可以包括任何数量的处理器810(例如，嵌入式处理器)。处理器810可以包括启动和电源管理处理器，该启动和电源管理处理器可以是专用处理器和子系统，以处理启动电源和管理功能以及相关的安全实施。在至少一个实施例中，启动和电源管理处理器可以是一个或更多个SoC 804启动序列的一部分，并且可以提供运行时电源管理服务。在至少一个实施例中，启动功率和管理处理器可以提供时钟和电压编程，辅助系统低功率状态转换，一个或更多个SoC 804热和温度传感器管理和/或一个或更多个SoC 804功率状态管理。在至少一个实施例中，每个温度传感器可以实现为

其输出频率与温度成比例的环形振荡器，并且一个或更多个SoC 804可以使用环形振荡器来检测一个或更多个CPU 806，一个或更多个GPU808和/或一个或更多个加速器814 的温度。在至少一个实施例中，如果确定温度超过阈值，则启动和电源管理处理器可以进入温度故障例程，并将一个或更多个SoC 804置于较低功耗状态和/或将车辆800置于司机的安全停车图案(例如，使车辆800安全停车)。

[0159] 在至少一个实施例中，一个或更多个处理器810可以进一步包括一组嵌入式处理器，其可以用作音频处理引擎。在至少一个实施例中，音频处理引擎可以是音频子系统，其能够通过多个接口以及广泛且灵活范围的音频I/O接口为硬件提供对多通道音频的完全硬件支持。在至少一个实施例中，音频处理引擎是专用处理器核心，其具有带专用RAM的数字信号处理器。

[0160] 在至少一个实施例中，一个或更多个处理器810可以进一步包括始终在线的处理器引擎。在至少一个实施例中，自动处理引擎可以提供必要的硬件特征以支持低功率传感器管理和唤醒用例。在至少一个实施例中，始终在线的处理器引擎上的处理器可以包括但不限于处理器核心、紧密耦合的RAM、支持外围设备(例如，定时器和中断控制器)、各种I/O控制器外围设备以及路由逻辑。

[0161] 在至少一个实施例中，一个或更多个处理器810可以进一步包括安全集群引擎，该安全集群引擎包括但不限于用于处理汽车应用程序的安全管理的专用处理器子系统。在至少一个实施例中，安全集群引擎可以包括但不限于两个或更多个处理器核心、紧密耦合的RAM、支持外围设备(例如，定时器、中断控制器等)和/或路由逻辑。在安全模式下，在至少一个实施例中，两个或更多个核心可以以锁步模式操作，并且可以用作具有用以检测其操作之间的任何差异的比较逻辑的单个核心。在至少一个实施例中，一个或更多个处理器810可以进一步包括实时相机引擎，该实时相机引擎可以包括但不限于用于处理实时相机管理的专用处理器子系统。在至少一个实施例中，一个或更多个处理器810可以进一步包括高动态范围信号处理器，该高动态范围信号处理器可以包括但不限于图像信号处理器，该图像信号处理器是作为相机处理管线的一部分的硬件引擎。

[0162] 在至少一个实施例中，一个或更多个处理器810可以包括视频图像合成器，该视频图像合成器可以是处理块(例如，在微处理器上实现)，该处理块实现视频回放应用程序产生最终的视频所需要的视频后处理功能，以产生用于播放器窗口的最终图像。在至少一个实施例中，视频图像合成器可以在一个或更多个广角相机870、一个或更多个环绕相机874和/或一个或更多个舱内监控相机传感器上执行透镜畸变校正。在至少一个实施例中，优选地，由在SoC 804的另一实例上运行的神经网络来监控舱室内监控相机传感器，该神经网络被配置为识别舱室事件并相应地做出响应。在至少一个实施例中，舱室内系统可以执行但不限于唇读以激活蜂窝服务和拨打电话、指示电子邮件、改变车辆的目的地、激活或改变车辆的信息娱乐系统和设置、或者提供语音激活的网上冲浪。在至少一个实施例中，当车辆以自主模式运行时，某些功能对于驾驶员是可用的，否则将其禁用。

[0163] 在至少一个实施例中，视频图像合成器可以包括用于同时空间和时间降噪的增强的时间降噪。例如，在至少一个实施例中，在运动发生在视频中的情况下，降噪适当地对空间信息加权，从而减小由相邻帧提供的信息的权重。在至少一个实施例中，在图像或图像的一部分不包括运动的情况下，由视频图像合成器执行的时间降噪可以使用来自先前图像的

信息来降低当前图像中的噪声。

[0164] 在至少一个实施例中，视频图像合成器还可以配置为对输入的立体透镜帧执行立体校正。在至少一个实施例中，当使用操作系统桌面时，视频图像合成器还可以用于用户接口合成，并且不需要一个或更多个GPU 808 来连续渲染新表面。在至少一个实施例中，当对一个或更多个GPU 808供电并使其活跃地进行3D渲染时，视频图像合成器可以被用于卸载一个或更多个GPU 808以改善性能和响应性。

[0165] 在至少一个实施例中，SoC 804中的一个或更多个SoC可以进一步包括用于从相机接收视频和输入的移动工业处理器接口(“MIPI”)相机串行接口、高速接口和/或可用于相机和相关像素输入功能的视频输入块。在至少一个实施例中，一个或更多个SoC 804可以进一步包括输入/输出控制器，该输入/输出控制器可以由软件控制并且可以被用于接收未提交给特定角色的I/O信号。

[0166] 在至少一个实施例中，SoC804中的一个或更多个SoC可以进一步包括广泛的外围接口，以使得能够与外围设备、音频编码器/解码器(“编解码器”)，电源管理和/或其他设备通信。一个或更多个SoC 804可用于处理来自(例如，通过千兆位多媒体串行链路和以太网连接)相机、传感器 (例如，一个或更多个LIDAR传感器864，一个或更多个RADAR传感器 860等，其可以通过以太网连接)的数据，来自总线802的数据(例如，车辆800的速度、方向盘位置等)，来自一个或更多个GNSS传感器858 的数据(例如，通过以太网总线或CAN总线连接)等。在至少一个实施例中，SoC 804中的一个或更多个可以进一步包括专用高性能海量存储控制器，其可以包括它们自己的DMA引擎，并且可以用于使一个或更多个 CPU 806摆脱常规数据管理任务。

[0167] 在至少一个实施例中，一个或更多个SoC 804可以是具有灵活架构的端到端平台，其跨越自动化级别3-5级，从而提供利用并有效使用计算机视觉和ADAS技术来实现多样性和冗余的综合的功能安全架构，其提供了可提供灵活、可靠的驾驶软件堆栈以及深度学习工具的平台。在至少一个实施例中，一个或更多个SoC 804可以比常规系统更快、更可靠，并且甚至在能量效率和空间效率上也更高。例如，在至少一个实施例中，一个或更多个加速器814当与一个或更多个CPU 806、一个或更多个GPU 808以及一个或更多个数据存储装置816结合时，可以提供用于3-5级自动驾驶车辆的快速、有效的平台。

[0168] 在至少一个实施例中，计算机视觉算法可以在CPU上执行，CPU可以使用高级编程语言(例如C编程语言)配置为在多种视觉数据上执行多种处理算法。然而，在至少一个实施例中，CPU通常不能满足许多计算机视觉应用程序的性能要求，例如与执行时间和功耗有关的性能要求。在至少一个实施例中，许多CPU不能实时执行复杂的对象检测算法，该算法被用于车载ADAS应用程序和实际3-5级自动驾驶车辆中。

[0169] 本文所述的实施例允许同时和/或序列地执行多个神经网络，并且允许将结果结合在一起以实现3-5级自动驾驶功能。例如，在至少一个实施例中，在DLA或离散GPU(例如，一个或更多个GPU 820)上执行的CNN 可包括文本和单词识别，从而允许超级计算机读取和理解交通标志，包括神经网络尚未被专门训练的标志。在至少一个实施例中，DLA还可包括神经网络，该神经网络能够识别、解释并提供符号的语义理解，并将该语义理解传递给在CPU Complex上运行的路径规划模块。

[0170] 在至少一个实施例中，对于3、4或5级的驱动，可以同时运行多个神经网络。例如，

在至少一个实施例中，由“警告标志包括：闪烁的灯指示结冰状况(Caution:flashing lights indicate icy conditions)”连通电灯一起组成的警告标志可以由多个神经网络独立地或共同地解释。在至少一个实施例中，可以通过第一部署的神经网络(例如，已经训练的神经网络)将该标志本身识别为交通标志，可以通过第二部署的神经网络来解释文本“闪烁的灯指示结冰状况(flashing lights indicate icy conditions)”，其通知车辆的路径规划软件(最好在CPU Complex上执行)：当检测到闪烁的灯光时，就会存在结冰状况。在至少一个实施例中，可以通过在多个帧上操作第三部署的神经网络来识别闪烁的灯，向车辆的路径规划软件通知存在(或不存在)闪烁的灯。在至少一个实施例中，所有三个神经网络可以同时运行，例如在DLA内和/或在一个或更多个GPU 808上。

[0171] 在至少一个实施例中，用于面部识别和车辆所有者识别的CNN可以使用来自相机传感器的数据来识别授权驾驶员和/或车辆800的所有者的存在。在至少一个实施例中，当所有者接近驾驶员门并打开灯时，常开传感器处理器引擎可用于解锁车辆，并且，在安全模式下，当所有者离开该车辆时，可用于禁用该车辆。以此方式，一个或更多个SoC 804提供防止盗窃和/或劫车的保障。

[0172] 在至少一个实施例中，用于紧急车辆检测和识别的CNN可以使用来自麦克风896的数据来检测和识别紧急车辆警报器。在至少一个实施例中，一个或更多个SoC 804使用CNN来对环境和城市声音进行分类，以及对视觉数据进行分类。在至少一个实施例中，训练在DLA上运行的CNN以识别紧急车辆的相对接近速度(例如，通过使用多普勒效应)。在至少一个实施例中，还可以训练CNN来识别针对车辆正在运行的区域的紧急车辆，如一个或更多个GNSS传感器858所识别。在至少一个实施例中，当在欧洲运行时，CNN将寻求检测欧洲警报器，而在美国时，CNN将寻求仅识别北美警报器。在至少一个实施例中，一旦检测到紧急车辆，就可以在一个或更多个超声波传感器862的辅助下使用控制程序来执行紧急车辆安全例程、减速车辆、将车辆驶至路边、停车、和/或使车辆闲置，直到紧急车辆通过。

[0173] 在至少一个实施例中，车辆800可以包括一个或更多个CPU 818(例如，一个或更多个离散CPU或一个或更多个dCPU)，其可以经由高速互连(例如PCIe)耦合到一个或更多个SoC 804。在至少一个实施例中，一个或更多个CPU 818可以包括X86处理器，例如一个或更多个CPU 818 可用于执行各种功能中的任何功能，例如包括在ADAS传感器和一个或更多个SoC 804之间潜在的仲裁不一致的结果，和/或一个或更多个监控控制器836的状态和健康和/或片上信息系统(“信息SoC”)830。

[0174] 在至少一个实施例中，车辆800可以包括一个或更多个GPU 820(例如，一个或更多个离散GPU或一个或更多个dGPU)，其可以经由高速互连(例如NVIDIA的NVLINK)耦合到一个或更多个SoC 804。在至少一个实施例中，一个或更多个GPU 820可以提供附加的人工智能功能，例如通过执行冗余和/或不同的神经网络，并且可以至少部分地基于来自车辆 800的传感器的输入(例如，传感器数据)来用于训练和/或更新神经网络。

[0175] 在至少一个实施例中，车辆800可以进一步包括网络接口824，其可以包括但不限于一个或更多个无线天线826(例如，用于不同通信协议的一个或更多个无线天线826，诸如蜂窝天线、蓝牙天线等)。在至少一个实施例中，网络接口824可以用于使能通过具有云的互联网(例如，采用服务器和/或其他网络设备)与其他车辆和/或计算设备(例如乘客的客户端设备)的无线连接。在至少一个实施例中，为了与其他车辆通信，可以在车辆800和其他车

辆之间建立直接链路和/或可以建立间接链路(例如,通过网络和互联网)。在至少一个实施例中,可以使用车辆到车辆的通信链路来提供直接链路。车辆到车辆的通信链路可以向车辆800提供关于车辆 800附近的车辆的信息(例如,车辆800前面、侧面和/或后面的车辆)。在至少一个实施例中,该前述功能可以是车辆800的协作自适应巡航控制功能的一部分。

[0176] 在至少一个实施例中,网络接口824可以包括SoC,其提供调制和解调功能并使一个或更多个控制器836能够通过无线网络进行通信。在至少一个实施例中,网络接口824可以包括射频前端,用于从基带到射频的上转换以及从射频到基带的下转换。在至少一个实施例中,可以以任何技术上可行的方式执行频率转换。例如,可以通过公知的过程和/或使用超外差过程来执行频率转换。在至少一个实施例中,射频前端功能可以由单独的芯片提供。在至少一个实施例中,网络接口可以包括用于通过LTE、WCDMA、UMTS、GSM、CDMA2000、蓝牙、蓝牙LE、Wi-Fi、Z-Wave、ZigBee、LoRaWAN和/或其他无线协议进行通信的无线功能。

[0177] 在至少一个实施例中,车辆800可以进一步包括一个或更多个数据存储828,其可以包括但不限于片外(例如,一个或更多个SoC 804)存储。在至少一个实施例中,一个或更多个数据存储828可以包括但不限于一个或更多个存储元件,包括RAM、SRAM、动态随机存取存储器(“DRAM”)、视频随机存取存储器(“VRAM”)、闪存、硬盘和/或其他组件和/或可以存储至少一位数据的设备。

[0178] 在至少一个实施例中,车辆800可以进一步包括一个或更多个GNSS 传感器858(例如,GPS和/或辅助GPS传感器),以辅助地图绘制、感知、占用光栅生成和/或路径规划功能。在至少一个实施例中,可以使用任何数量的GNSS传感器858,包括例如但不限于使用具有以太网的USB连接器连接到串行接口(例如RS-232)桥的GPS。

[0179] 在至少一个实施例中,车辆800可以进一步包括一个或更多个 RADAR传感器860。一个或更多个RADAR传感器860可以由车辆800 用于远程车辆检测,即使在黑暗和/或恶劣天气条件下。在至少一个实施例中,RADAR功能安全等级可以是ASIL B。一个或更多个RADAR传感器 860可以使用CAN总线和/或总线802(例如,以传输由一个或更多个 RADAR传感器860生成的数据)来进行控制和访问对象跟踪数据,在某些示例中可以访问以太网通道以访问原始数据。在至少一个实施例中,可以使用各种各样的RADAR传感器类型。例如但不限于,RADAR传感器 860中的一个或更多个传感器可适合于前、后和侧面RADAR使用。在至少一个实施例中,一个或更多个RADAR传感器860是脉冲多普勒RADAR 传感器。

[0180] 在至少一个实施例中,一个或更多个RADAR传感器860可以包括不同的配置,例如具有窄视野的远程、具有宽事业的近程、近程侧面覆盖等。在至少一个实施例中,远程RADAR可以用于自适应巡航控制功能。在至少一个实施例中,远程RADAR系统可以提供通过两次或更多次独立扫描 (例如在250m范围内)实现的宽广的视野。在至少一个实施例中,一个或更多个RADAR传感器860可以帮助在静态对象和运动对象之间区分,并且可以被ADAS系统838用于紧急制动辅助和向前碰撞警告。包括在远程RADAR系统中的一个或更多个传感器860可以包括但不限于具有多个 (例如六个或更多个)固定RADAR天线以及高速CAN和FlexRay接口的单基地多模式RADAR。在至少一个实施例中,具有六个天线、中央四个天线可以创建聚焦的波束图,该波束图设计为以较高的速度记录车辆800 的周围环境,而相邻车道的交通干扰最小。在至少一个实施例中,其他两个天线可以扩大视野,从而可以快速检测进入或离开车辆800的车道。

[0181] 在至少一个实施例中，作为示例，中程RADAR系统可包括例如高达 160m(前)或80m(后)的范围，以及高达42度(前)或150度(后) 的视野。在至少一个实施例中，短程RADAR系统可以包括但不限于设计成安装在后保险杠的两端的任意数量的RADAR传感器860。当安装在后保险杠的两端时，在至少一个实施例中，RADAR传感器系统可以产生两个光束，该两个光束不断地监测车辆后部和附近的盲点。在至少一个实施例中，短程RADAR系统可以在ADAS系统838中用于盲点检测和/或车道改变辅助。

[0182] 在至少一个实施例中，车辆800可以进一步包括一个或更多个超声传感器862。可以定位在车辆800的前、后和/或侧面的一个或更多个超声传感器862可以用于停车辅助和/或创建和更新占用光栅。在至少一个实施例中，可以使用各种各样的超声传感器862，并且可以将不同的超声传感器 862用于不同的检测范围(例如2.5m、4m)。在至少一个实施例中，超声传感器862可以在ASIL B的功能安全等级下操作。

[0183] 在至少一个实施例中，车辆800可以包括一个或更多个LIDAR传感器864。一个或更多个LIDAR传感器864可以用于对象和行人检测、紧急制动、避免碰撞和/或其他功能。在至少一个实施例中，一个或更多个 LIDAR传感器864可以是功能安全等级ASIL B。在至少一个实施例中，车辆800可以包括可以使用以太网的多个(例如，两个、四个、六个等) LIDAR传感器864(例如，将数据提供给千兆以太网交换机)。

[0184] 在至少一个实施例中，一个或更多个LIDAR传感器864可能能够提供针对360度视野的对象及其距离的列表。在至少一个实施例中，市售的一个或更多个LIDAR传感器864例如可以具有大约100m的广告范围，具有2cm-3cm的精度，并且支持100Mbps的以太网连接。在至少一个实施例中，可以使用一个或更多个非突出的LIDAR传感器864。在这样的实施例中，一个或更多个LIDAR传感器864可以作为嵌入到车辆800的前、后、侧面和/或拐角中的小型设备实施。在至少一个实施例中，一个或更多个 LIDAR传感器864，在这样的实施例中，即使对于低反射率的对象，也可以提供高达120度的水平视野和35度的垂直视野，并且具有200m的范围。在至少一个实施例中，可将前向一个或更多个LIDAR传感器864配置为用于45度至135度之间的水平视野。

[0185] 在至少一个实施例中，也可以使用LIDAR技术(诸如3D闪光LIDAR)。3D闪光LIDAR使用激光闪光作为传输源，以照亮车辆800周围大约200m。在至少一个实施例中，闪光LIDAR单元包括但不限于接收器，该接收器记录激光脉冲传播时间和每个像素上的反射光，该像素又对应于从车辆800 到对象的范围。在至少一个实施例中，闪光LIDAR可以允许利用每个激光闪光生成周围环境的高度准确且无失真的图像。在至少一个实施例中，可以部署四个闪光LIDAR传感器，在车辆800的每一侧部署一个传感器。在至少一个实施例中，3D闪光LIDAR系统包括但不限于除了风扇(例如非扫描LIDAR设备)以外没有移动部件的固态3D视线阵列LIDAR相机。在至少一个实施例中，闪光LIDAR设备可以每帧使用5纳秒的I类(人眼安全)激光脉冲，并且可以捕获反射激光，以3D测距点云的形式和共同登记的强度数据。

[0186] 在至少一个实施例中，车辆800还可包括一个或更多个IMU传感器 866。在至少一个实施例中，一个或更多个IMU传感器866可位于车辆800 的后轴中心。在至少一个实施例中，一个或更多个IMU传感器866可以包括，例如但不限于，一个或更多个加速度计、一个或更多个磁力计、一个或更多个陀螺仪、一个或更多个磁罗盘和/或其他传感器类型。在至少一个实施例中，例如在六轴应用程序中，一个或更多个IMU传感器866可以包括但不限于加

速度计和陀螺仪。在至少一个实施例中，例如在九轴应用程序中，一个或更多个IMU传感器866可以包括但不限于加速度计、陀螺仪和磁力计。

[0187] 在至少一个实施例中，一个或更多个IMU传感器866可以实现为结合了微机电系统(“MEMS”)惯性传感器，高灵敏度GPS接收器和先进的卡尔曼滤波算法的微型高性能GPS辅助惯性导航系统(“GPS/INS”)，以提供位置、速度和姿态的估算；在至少一个实施例中，一个或更多个IMU 传感器866可使车辆800估算航向而无需来自磁传感器通过直接观测和关联从GPS到一个或更多个IMU传感器866的速度变化来实现的输入。在至少一个实施例中，一个或更多个IMU传感器866和一个或更多个GNSS 传感器858可以组合在单个集成单元中。

[0188] 在至少一个实施例中，车辆800可以包括放置在车辆800内和/或周围的一个或更多个麦克风896。在至少一个实施例中，此外，一个或更多个麦克风896可以用于紧急车辆检测和识别。

[0189] 在至少一个实施例中，车辆800可以进一步包括任何数量的相机类型，包括一个或更多个立体相机868、一个或更多个广角相机870、一个或更多个红外相机872、一个或更多个环绕相机874、一个或更多个远程相机898、一个或更多个中程相机876和/或其他相机类型。在至少一个实施例中，相机可用于捕获车辆800的整个外围周围的图像数据。在至少一个实施例中，所使用的相机的类型取决于车辆800。在至少一个实施例中，相机类型的任何组合可以是用于在车辆800周围提供必要覆盖范围。在至少一个实施例中，部署的相机的数量可以根据实施例而不同。例如，在至少一个实施例中，车辆800可以包括六个相机、七个相机、十个相机、十二个相机或其他数量的相机。相机可以作为示例但不限于支持千兆位多媒体串行链路 (“GMSL”)和/或千兆位以太网。在至少一个实施例中，本文先前参照图 8A和图8B可以更详细地描述了每个相机。

[0190] 在至少一个实施例中，车辆800可以进一步包括一个或更多个振动传感器842。一个或更多个振动传感器842可以测量车辆800的部件(例如，轴)的振动。例如，在至少一个实施例中，振动的变化可以指示路面的变化。在至少一个实施例中，当使用两个或更多个振动传感器842时，振动之间的差异可以用于确定路面的摩擦或打滑(例如，当在动力驱动轴和自由旋转轴之间存在振动差异时)。

[0191] 在至少一个实施例中，车辆800可以包括ADAS系统838。ADAS系统838可以包括但不限于SoC。在至少一个实施例中，ADAS系统838可以包括但不限于任何数量的自主/自适应/自动巡航控制(“ACC”)系统、协作自适应巡航控制(“CACC”)系统、前撞警告(“FCW”)系统、自动紧急制动(“AEB”)系统、车道偏离警告(“LDW”)系统、车道保持辅助(“LKA”) 系统、盲区警告(“BSW”)系统、后方交叉交通警告(“RCTW”)系统、碰撞警告(“CW”)系统、车道对中(“LC”)系统和/或其他系统、特征和/ 或功能及其组合。

[0192] 在至少一个实施例中，ACC系统可以使用一个或更多个RADAR传感器860、一个或更多个LIDAR传感器864和/或任何数量的相机。在至少一个实施例中，ACC系统可以包括纵向ACC系统和/或横向ACC系统。在至少一个实施例中，纵向ACC系统监控并控制到紧邻车辆800的前面车辆的距离，并自动调节车辆800的速度以保持与前方车辆的安全距离。在至少一个实施例中，横向ACC系统执行距离保持，并在需要时建议车辆 800改变车道。在至少一个实施例中，横向ACC与其他ADAS应用程序有关，例如LC和CW。

[0193] 在至少一个实施例中，CACC系统使用来自其他车辆的信息，该信息可以经由网络

接口824和/或一个或更多个无线天线826从其他车辆接收经由无线链路或者间接经由网络连接(例如,经由互联网)接收。在至少一个实施例中,直接链路可以由车辆到车辆("V2V")的通信链路提供,而间接链路可以由基础设施到车辆("I2V")的通信链路提供。通常,V2V通信概念提供关于紧接在前的车辆(例如,紧接在车辆800之前并与之在同一车道上的车辆)的信息,而I2V通信概念提供关于更前方交通的信息。在至少一个实施例中,CACC系统可以包括I2V和V2V信息源之一或两者。在至少一个实施例中,在给定车辆800之前的车辆的信息的情况下,CACC 系统可以更可靠,并且具有改善交通流的平滑度并减少道路拥堵的潜力。

[0194] 在至少一个实施例中,FCW系统被设计成警告驾驶员危险,以便该驾驶员可以采取纠正措施。在至少一个实施例中,FCW系统使用前向相机和/或一个或更多个RADAR传感器860,其耦合至专用处理器、DSP、FPGA 和/或ASIC,其电耦合至驾驶员反馈,例如显示器、扬声器和/或振动组件。在至少一个实施例中,FCW系统可以提供警告,例如以声音、视觉警告,振动和/或快速制动脉冲的形式。

[0195] 在至少一个实施例中,AEB系统检测到与另一车辆或其他对象的即将发生的向前碰撞,并且如果驾驶员在指定的时间或距离参数内未采取纠正措施,则可以自动施加制动。在至少一个实施例中,AEB系统可以使用耦合到专用处理器、DSP、FPGA和/或ASIC的一个或更多个前向相机和/ 或一个或更多个RADAR传感器860。在至少一个实施例中,当AEB系统检测到危险时,AEB系统通常首先警告驾驶员采取纠正措施以避免碰撞,并且,如果该驾驶员没有采取纠正措施,则该AEB系统可以自动施加制动器以试图防止或至少减轻预测碰撞的影响。在至少一个实施例中,AEB系统可以包括诸如动态制动器支持和/或即将发生碰撞的制动的技术。

[0196] 在至少一个实施例中,当车辆800越过车道标记时,LDW系统提供视觉、听觉和/或触觉警告,例如方向盘或座椅振动,以警告驾驶员。在至少一个实施例中,当驾驶员诸如通过激活转向信号灯指示有意的车道偏离时,LDW系统不活跃。在至少一个实施例中,LDW系统可以使用耦合到专用处理器、DSP、FPGA和/或ASIC的面向正面的相机,其被电耦合至诸如显示器、扬声器和/或振动组件之类的驾驶员反馈。在至少一个实施例中,LKA系统是LDW系统的一种变型。如果车辆800开始离开车道,则 LKA系统提供转向输入或制动以校正车辆800。

[0197] 在至少一个实施例中,BSW系统检测并警告汽车盲区中的车辆驾驶员。在至少一个实施例中,BSW系统可以提供视觉、听觉和/或触觉警报,以指示合并或改变车道是不安全的。在至少一个实施例中,当驾驶员使用转向灯时,BSW系统可以提供附加警告。在至少一个实施例中,BSW系统可以使用耦合到专用处理器、DSP、FPGA和/或ASIC的一个或更多个朝后侧的相机和/或一个或更多个RADAR传感器860,其电耦合到驾驶员反馈,例如显示器、扬声器和/或振动组件。

[0198] 在至少一个实施例中,当在车辆800倒车时在后相机范围之外检测到对象时,RCTW系统可以提供视觉、听觉和/或触觉通知。在至少一个实施例中,RCTW系统包括AEB系统,以确保应用程序车辆制动器以避免碰撞。在至少一个实施例中,RCTW系统可以使用一个或更多个面向后方的 RADAR传感器860,其耦合到专用处理器、DSP、FPGA和/或ASIC,其被电耦合至诸如显示器、扬声器和/或振动组件之类的驾驶员反馈。

[0199] 在至少一个实施例中,常规的ADAS系统可能易于产生误报结果,这可能使驾驶员

烦恼和分散注意力,但通常不是灾难性的,因为常规的 ADAS系统会警告驾驶员并允许该驾驶员决定安全状况是否真正存在并采取相应动作。在至少一个实施例中,在结果冲突的情况下,车辆800本身决定是否听从主计算机或副计算机(例如,第一控制器836或第二控制器836)的结果。例如,在至少一个实施例中,ADAS系统838可以是用于将感知信息提供给备份计算机合理性模块的备用和/或辅助计算机。在至少一个实施例中,备用计算机合理性监控器可以在硬件组件上运行冗余的各种软件,以检测感知和动态驾驶任务中的故障。在至少一个实施例中,可以将来自ADAS系统838的输出提供给监控MCU。在至少一个实施例中,如果来自主计算机和辅助计算机冲突的输出,则监督MCU决定如何协调冲突以确保安全操作。

[0200] 在至少一个实施例中,主计算机可以配置为向监督MCU提供置信度分数,以指示该主计算机对所选结果的置信度。在至少一个实施例中,如果该置信度得分超过阈值,则该监督MCU可以遵循该主计算机的指示,而不管该辅助计算机是否提供冲突或不一致的结果。在至少一个实施例中,在置信度得分不满足阈值的情况下,并且在主计算机和辅助计算机指示不同的结果(例如,冲突)的情况下,监督MCU可以在计算机之间仲裁以确定适当的结果。

[0201] 在至少一个实施例中,监督MCU可以配置为运行神经网络,该神经网络被训练和配置为至少部分地基于来自主计算机和辅助计算机的输出来确定该辅助计算机提供错误警报的条件。在至少一个实施例中,监督MCU 中的神经网络可以学习何时可以信任辅助计算机的输出,以及何时不能信任。例如,在至少一个实施例中,当该辅助计算机是基于RADAR的FCW 系统时,该监督MCU中的神经网络可以学习FCW系统何时识别实际上不是危险的金属对象,例如会触发警报的排水格栅或井盖。在至少一个实施例中,当辅助计算机是基于相机的LDW系统时,当存在骑自行车的人或行人并且实际上车道偏离是最安全的操作时,监督MCU中的神经网络可以学会覆盖LDW。在至少一个实施例中,监督MCU可以包括适合于运行具有相关联的存储器的神经网络的DLA或GPU中的至少一个。在至少一个实施例中,监督MCU可以包括和/或被包括为一个或更多个SoC 804的组件。

[0202] 在至少一个实施例中,ADAS系统838可以包括使用传统的计算机视觉规则执行ADAS功能的辅助计算机。在至少一个实施例中,该辅助计算机可以使用经典计算机视觉规则(如果-则),并且监督MCU中的神经网络的存在可以提高可靠性、安全性和性能。例如,在至少一个实施例中,多样化的实现方式和有意的非同一性使得整个系统更加容错,尤其是对于由软件(或软件-硬件接口)功能引起的故障。例如,在至少一个实施例中,如果在主计算机上运行的软件中存在软件漏洞或错误,并且在辅助计算机上运行的不相同的软件代码提供了相同的总体结果,则监督MCU可以更有把握地认为总体结果是正确,并且该主计算机上的软件或硬件中的漏洞不会导致重大错误。

[0203] 在至少一个实施例中,可以将ADAS系统838的输出输入到主计算机的感知模块和/或主计算机的动态驾驶任务模块中。例如,在至少一个实施例中,如果ADAS系统838由于正前方的对象而指示向前碰撞警告,则感知块可以在识别对象时使用该信息。在至少一个实施例中,如本文所述,辅助计算机可以具有其自己的神经网络,该神经网络经过训练从而降低了误报的风险。

[0204] 在至少一个实施例中,车辆800可以进一步包括信息娱乐SoC 830(例如,车载信息娱乐系统(IVI))。尽管被示出和描述为SoC,但是在至少一个实施例中,信息娱乐系统830可以不是SoC,并且可以包括但不限于两个或更多个分立组件。在至少一个实施例中,信息娱

乐SoC 830可以包括但不限于硬件和软件的组合，其可以用于提供音频(例如，音乐、个人数字助理、导航指令、新闻、广播等)、视频(例如，电视、电影、流媒体等)、电话(例如，免提通话)、网络连接(例如，LTE、WiFi等)和 /或信息服务(例如，导航系统、后停车辅助、无线电数据系统、与车辆相关的信息，例如燃油水平、总覆盖距离、制动燃油水平、油位、车门打开/关闭、空气滤清器信息等)到车辆800。例如，信息娱乐SoC 830可以包括收音机、磁盘播放器、导航系统、视频播放器、USB和蓝牙连接、汽车、车载娱乐系统、WiFi、方向盘音频控制、免提语音控制、抬头显示器 (“HUD”)、HMI显示器834、远程信息处理设备、控制面板(例如，用于控制各种组件、特征和/或系统和/或与之交互)和/或其他组件。在至少一个实施例中，信息娱乐SoC 830可以进一步用于向车辆的用户提供信息 (例如，视觉和/或听觉的)，诸如来自ADAS系统838的信息、自动驾驶信息(诸如计划的车辆操纵)、轨迹、周围环境信息(例如，交叉路口信息、车辆信息、道路信息等)和/或其他信息。

[0205] 在至少一个实施例中，信息娱乐SoC 830可以包括任何数量和类型的 GPU功能。在至少一个实施例中，信息娱乐SoC 830可以通过总线802(例如，CAN总线、以太网等)与车辆800的其他设备、系统和/或组件通信。在至少一个实施例中，信息娱乐SoC 830可以是耦合到监控MCU，使得信息娱乐系统的GPU可以在主控制器836(例如，车辆800的主计算机和/ 或备用计算机)发生故障的情况下执行一些自动驾驶功能。在至少一个实施例中，信息娱乐SoC 830可以使车辆800进入司机到安全停止模式，如本文所述。

[0206] 在至少一个实施例中，车辆800可以进一步包括仪表板832(例如，数字仪表板、电子仪表板、数字仪表操纵板等)。仪表板832可以包括但不限于控制器和/或超级计算机(例如，离散控制器或超级计算机)。在至少一个实施例中，仪表板832可以包括但不限于一组仪表的任何数量和组合，例如车速表、燃料水平、油压、转速表、里程表、转弯指示器、换档位置指示器、一个或更多个安全带警告灯、一个或更多个驻车制动警告灯、一个或更多个发动机故障灯、辅助约束系统(例如安全气囊)信息、照明控件、安全系统控件、导航信息等。在某些示例中，信息可能是在信息娱乐SoC 830和仪表板832之间显示和/或共享。在至少一个实施例中，仪表板832可以被包括作为信息娱乐SoC 830的一部分，反之亦然。

[0207] 图8D是根据至少一个实施例的在基于云的服务器与图8A的自动驾驶车辆800之间进行通信的系统876的图。在至少一个实施例中，系统876 可以包括但不限于一个或更多个服务器878、一个或更多个网络890以及任何数量和类型的车辆，包括车辆800。一个或更多个服务器878可以包括但不限于，多个GPU 884(A)-884(H)(在本文中统称为GPU 884)、PCIe交换机882(A)-882(D)(在本文中统称为PCIe交换机882)，和 /或CPU 880(A)-880(B)(在本文中统称为CPU 880)、GPU 884、CPU 880和PCIe交换机882可以与高速连接线互连，例如但不限于，由NVIDIA 开发的NVLink接口888和/或PCIe连接886。GPU 884通过NVLink和/或NVSwitchSoC连接，GPU 884和PCIe交换机882通过PCIe互连连接。在至少一个实施例中，尽管示出了八个GPU 884、两个CPU 880和四个PCIe 交换机882，但这并不旨在进行限制。在至少一个实施例中，一个或更多个服务器878中的每一个可以包括但不限于任意数量的GPU 884、CPU 880 和/或PCIe交换机882的任何组合。例如，在至少一个实施例中，一个或更多个服务器878可各自包括八个、十六个、三十二个和/或更多个GPU 884。

[0208] 在至少一个实施例中，一个或更多个服务器878可以通过一个或更多个网络890并从车辆接收表示图像的图像数据，该图像示出了意外的或改变的道路状况，例如最近开始

的道路工程。在至少一个实施例中，一个或更多个服务器878可以通过一个或更多个网络890并且向车辆传输神经网络892，更新的神经网络892，和/或地图信息894，包括但不限于关于交通和道路状况的信息。在至少一个实施例中，对地图信息894的更新可以包括但不限于对HD地图822的更新，例如关于建筑工地、坑洼、便道、洪水和/或其他障碍物的信息。在至少一个实施例中，神经网络892，更新的神经网络892和/或地图信息894可能是由从环境中的任何数量的车辆接收的数据中表示的新训练和/或经验产生的，和/或至少基于在数据中心执行的训练(例如，使用一个或更多个服务器878和/或其他服务器)。

[0209] 在至少一个实施例中，一个或更多个服务器878可以用于至少部分地基于训练数据来训练机器学习模型(例如，神经网络)。训练数据可以由车辆产生，和/或可以在模拟中产生(例如，使用游戏引擎)。在至少一个实施例中，标记任何数量的训练数据(例如，在相关的神经网络受益于监督学习的情况下)和/或经历其他预处理。在至少一个实施例中，没有对任何数量的训练数据进行标记和/或预处理(例如，在相关联的神经网络不需要监督学习的情况下)。在至少一个实施例中，一旦机器学习模型被训练，机器学习模型就可以被车辆使用(例如，通过一个或更多个网络890传输到车辆，和/或机器学习模型可以被一个或更多个服务器878使用以远程监控车辆。

[0210] 在至少一个实施例中，一个或更多个服务器878可以从车辆接收数据并且将数据应用程序于最新的实时神经网络以用于实时智能推理。在至少一个实施例中，一个或更多个服务器878可以包括由一个或更多个GPU 884供电的深度学习超级计算机和/或专用AI计算机，例如由NVIDIA开发的DGX和DGX Station机器。然而，在至少一个实施例中，一个或更多个服务器878可以包括使用CPU供电的数据中心的深度学习基础设施。

[0211] 在至少一个实施例中，一个或更多个服务器878的深度学习基础结构可能能够进行快速、实时的推理，并且可以使用该能力来评估和验证车辆 800中处理器、软件和/或相关硬件的健康。例如，在至少一个实施例中，深度学习基础设施可以从车辆800接收周期性更新，例如车辆800在该图像序列中所定位的图像序列和/或对象(例如，通过计算机视觉和/或其他机器学习对象分类技术)。在至少一个实施例中，深度学习基础设施可以运行其自己的神经网络以识别对象并将它们与车辆800所识别的对象进行比较，并且，如果结果不匹配和深度学习基础设施断定车辆800中的AI正在发生故障，则一个或更多个服务器878可以将信号发送到车辆800，以指示车辆800的故障安全计算机采取控制、通知乘客并完成安全停车操作。

[0212] 在至少一个实施例中，一个或更多个服务器878可以包括一个或更多个GPU 884和一个或更多个可编程推理加速器(例如NVIDIA的TensorRT 3)。在至少一个实施例中，GPU驱动的服务器和推理加速的组合可以使实时响应成为可能。在至少一个实施例中，例如在性能不太关键的情况下，可以将由CPU、FPGA和其他处理器驱动的服务器用于推理。在至少一个实施例中，硬件结构715用于执行一个或更多个实施例。本文结合图7A 和/或图7B提供关于硬件结构715的细节。

[0213] 计算机系统

[0214] 图9是示出根据至少一个实施例示例性计算机系统的框图，该示例性计算机系统可以是具有互连的设备和组件的系统，片上系统(SOC)或它们的某种形成有处理器的组合，该处理器可以包括执行单元以执行指令。在至少一个实施例中，根据本公开，例如本文所述

的实施例，计算机系统 900可以包括但不限于组件，例如处理器902，其执行单元包括逻辑以执行用于过程数据的算法。在至少一个实施例中，计算机系统900可以包括处理器，例如可从加利福尼亚圣塔克拉拉的英特尔公司(Intel Corporation of Santa Clara, California)获得的PENTIUM®处理器家族、Xeon™、Itanium®、 XScale™和/或StrongARM™，Intel®Core™或Intel ®Nervana™微处理器，尽管也可以使用其他系统(包括具有其他微处理器的PC、工程工作站、机顶盒等)。在至少一个实施例中，计算机系统900可以执行可从华盛顿州雷蒙德市的微软公司(Microsoft Corporation of Redmond,Wash.)获得的 WINDOWS操作系统版本，尽管其他操作系统(例如UNIX和Linux)、嵌入式软件和/或图形用户界面也可以使用。

[0215] 实施例可以用在其他设备中，例如手持设备和嵌入式应用程序。手持设备的一些示例包括蜂窝电话、互联网协议(Internet Protocol)设备、数码相机、个人数字助理(“PDA”)和手持PC。在至少一个实施例中，嵌入式应用程序可以包括微控制器、数字信号处理器(“DSP”)、片上系统、网络计算机(“NetPC”)、机顶盒、网络集线器、广域网(“WAN”)交换机，或根据至少一个实施例可以执行一个或更多个指令的任何其他系统。

[0216] 在至少一个实施例中，计算机系统900可包括但不限于处理器902，该处理器902可包括但不限于一个或更多个执行单元908，以根据本文描述的技术执行机器学习模型训练和/或推理。在至少一个实施例中，系统9 是单处理器台式机或服务器系统，但是在另一个实施例中，系统9可以是多处理器系统。在至少一个实施例中，处理器902可以包括但不限于复杂指令集计算机(“CISC”)微处理器、精简指令集计算(“RISC”)微处理器、超长指令字(“VLIW”)微处理器、实现指令集组合的处理器，或任何其他处理器设备，例如数字信号处理器。在至少一个实施例中，处理器902可以耦合到处理器总线910，该处理器总线910可以在处理器902与计算机系统900中的其他组件之间传输数据信号。

[0217] 在至少一个实施例中，处理器902可以包括但不限于1级(“L1”)内部高速缓存存储器(“cache”)904。在至少一个实施例中，处理器902可以具有单个内部高速缓存或多级内部缓存。在至少一个实施例中，高速缓存存储器可以驻留在处理器902的外部。根据特定的实现和需求，其他实施例也可以包括内部和外部高速缓存的组合。在至少一个实施例中，寄存器文件906可以在各种寄存器中存储不同类型的数据，包括但不限于整数寄存器、浮点寄存器、状态寄存器和指令指针寄存器。

[0218] 在至少一个实施例中，包括但不限于执行整数和浮点运算的逻辑的执行单元908，其也位于处理器902中。处理器902还可以包括微码(“ucode”) 只读存储器(“ROM”)，用于存储某些宏指令的微代码。在至少一个实施例中，执行单元908可以包括用于处理封装指令集909的逻辑。在至少一个实施例中，通过将封装指令集909包括在通用处理器的指令集中，以及要执行指令的相关电路，可以使用通用处理器902中的封装数据来执行许多多媒体应用程序使用的操作。在一个或更多个实施例中，可以通过使用处理器的数据总线的全宽度来在封装的数据上执行操作来加速和更有效地执行许多多媒体应用程序，这可能不需要在该处理器的数据总线上传输较小的数据单元来一次执行一个数据元素的一个或更多个操作。

[0219] 在至少一个实施例中，执行单元908也可以用在微控制器、嵌入式处理器、图形设备、DSP和其他类型的逻辑电路中。在至少一个实施例中，计算机系统900可以包括但不限于存储器920。在至少一个实施例中，存储器920可以实现为动态随机存取存储器(“DRAM”)设

备、静态随机存取存储器(“SRAM”)设备、闪存设备或其他存储设备。存储器920可以存储由处理器902可以执行的由数据信号表示的指令919和/或数据921。

[0220] 在至少一个实施例中,系统逻辑芯片可以耦合到处理器总线910和存储器920。在至少一个实施例中,系统逻辑芯片可以包括但不限于存储器控制器集线器(“MCH”)916,并且处理器902可以经由处理器总线910 与MCH 916通信。在至少一个实施例中,MCH 916可以提供到存储器920 的高带宽存储器路径918以用于指令和数据存储以及用于图形命令、数据和纹理的存储。在至少一个实施例中,MCH 916可以在处理器902、存储器920和计算机系统900中的其他组件之间启动数据信号,并且在处理器总线910、存储器920和系统I/O 922之间桥接数据信号。在至少一个实施例中,系统逻辑芯片可以提供用于耦合到图形控制器的图形端口。在至少一个实施例中,MCH 916可以通过高带宽存储器路径918耦合到存储器920,并且图形/视频卡912可以通过加速图形端口(Accelerated Graphics Port) (“AGP”)互连914耦合到MCH 916。

[0221] 在至少一个实施例中,计算机系统900可以使用系统I/O 922其为专有集线器接口总线来将MCH 916耦合到I/O控制器集线器(“ICH”)930。在至少一个实施例中,ICH 930可以通过本地I/O总线提供与某些I/O设备的直接连接。在至少一个实施例中,本地I/O总线可以包括但不限于用于将外围设备连接到存储器920、芯片组和处理器902的高速I/O总线。示例可以包括但不限于音频控制器929、固件集线器(“Flash BIOS”)928、无线收发器926、数据存储924、包含用户输入和键盘接口的传统I/O控制器923、串行扩展端口927(例如通用串行总线(USB))和网络控制器 934。数据存储924可以包括硬盘驱动器、软盘驱动器、CD-ROM设备、闪存设备或其他大容量存储设备。

[0222] 在至少一个实施例中,图9示出了包括互连的硬件设备或“芯片”的系统,而在其他实施例中,图9可以示出片上系统(SoC)。在至少一个实施例中,图9中示出的设备可以与专有互连、标准化互连(例如,PCIe) 或其某种组合互连。在至少一个实施例中,系统900的一个或更多个组件使用计算快速链路(CXL)互连来互连。

[0223] 在至少一个实施例中,图5示出了包括互连硬件设备或“芯片”的系统,而在其他实施例中,图5可示出示例性片上系统(“SoC”)。

[0224] 图10是示出根据至少一个实施例的用于利用处理器1010的电子设备 1000的框图。在至少一个实施例中,电子设备1000可以是例如但不限于,笔记本电脑、塔式服务器、机架服务器、刀片服务器、膝上型计算机、台式机、平板电脑、移动设备、电话、嵌入式计算机或任何其他合适的电子设备。

[0225] 在至少一个实施例中,系统1000可以包括但不限于通信地耦合到任何合适数量或种类的组件、外围设备、模块或设备的处理器1010。在至少一个实施例中,处理器1010使用总线或接口耦合,诸如I℃总线、系统管理总线(“SMBus”)、低引脚计数(LPC)总线、串行外围接口(“SPI”)、高清音频(“HDA”)总线、串行高级技术附件(“SATA”)总线、通用串行总线(“USB”)(1、2、3版等)或通用异步接收器/发送器(“UART”)总线。在至少一个实施例中,图10示出了系统,该系统包括互连的硬件设备或“芯片”,而在其他实施例中,图10可以示出示例性片上系统(SoC)。在至少一个实施例中,图10中所示的设备可以与专有互连线、标准化互连 (例如,PCIe)或其某种组合互连。在至少一个实施例中,图10的一个或更多个组件使用计算快速链路(CXL)互连线来互连。

[0226] 在至少一个实施例中，图10可以包括显示器1024、触摸屏1025、触摸板1030、近场通信单元(“NFC”)1045、传感器集线器1040、热传感器 1046、快速芯片组(“EC”)1035、可信平台模块(“TPM”)1038、BIOS/ 固件/闪存(“BIOS，FW Flash”)1022、DSP1060、驱动器“SSD或HDD” 1020(例如固态磁盘(“SSD”)或硬盘驱动器(“HDD”))、无线局域网单元(“WLAN”)1050、蓝牙单元1052、无线广域网单元(“WWAN”)1056、全球定位系统(GPS)1055、相机(“USB 3.0相机”)1054(例如USB 3.0 相机)或以例如LPDDR3标准实现的低功耗双倍数据速率(“LPDDR”)存储器单元(“LPDDR3”)1015。这些组件可以各自以任何合适的方式实现。

[0227] 在至少一个实施例中，其他组件可以通过如上所述的组件通信地耦合到处理器1010。在至少一个实施例中，加速度计1041、环境光传感器(“ALS”) 1042、罗盘1043和陀螺仪1044可以可通信地耦合到传感器集线器1040。在至少一个实施例中，热传感器1039、风扇1037、键盘1046和触摸板1030 可以通信地耦合到EC1035。在至少一个实施例中，扬声器1063、耳机1064 和麦克风(“mic”)1065可以通信地耦合到音频单元(“音频编解码器和D 类放大器”)1064，其又可以通信地耦合到DSP1060。在至少一个实施例中，音频单元1064可以包括例如但不限于音频编码器/解码器(“编解码器”) 和D类放大器。在至少一个实施例中，SIM卡(“SIM”)1057可以通信地耦合到WWAN单元1056。在至少一个实施例中，组件(诸如WLAN单元1050和蓝牙单元1052以及WWAN单元1056)可以被实现为下一代形式因素(NGFF)。

[0228] 图11示出了根据至少一个实施例的计算机系统1100。在至少一个实施例中，计算机系统1100配置为实现贯穿本公开描述的各种过程和方法。

[0229] 在至少一个实施例中，计算机系统1100包括但不限于至少一个中央处理单元(“CPU”)1102，该中央处理单元(“CPU”)1102连接到使用任何合适协议实现的通信总线1110，诸如PCI(“外围设备互联”)、外围组件互连Express(“PCI-Express”)、AGP(“加速图形端口”)、超传输或任何其他总线或点对点通信协议。在至少一个实施例中，计算机系统1100包括但不限于主存储器1104和控制逻辑(例如，实现为硬件、软件或其组合)，并且数据可以采取随机存取存储器(“RAM”)的形式存储在主存储器1104中。在至少一个实施例中，网络接口子系统(“网络接口”)1122 提供到其他计算设备和网络的接口，用于从计算机系统1100接收数据并将数据传输到其他系统。

[0230] 在至少一个实施例中，计算机系统1100在至少一个实施例中包括但不限于输入设备1108、并行处理系统1112和显示设备1106，它们可以使用常规的阴极视线管(“CRT”)、液晶显示器(“LCD”)、发光二极管(“LED”)、等离子显示器或其他合适的显示技术实现。在至少一个实施例中，从输入设备1108(诸如键盘、鼠标、触摸板、麦克风等)接收用户输入。在至少一个实施例中，前述模块中的每一个可以位于单个半导体平台上以形成处理系统。

[0231] 图12示出了根据至少一个实施例的计算机系统1200。在至少一个实施例中，计算机系统1200包括但不限于计算机1210和USB棒1220。在至少一个实施例中，计算机1210可以包括但不限于任何数量和类型的处理器(未示出)和存储器(未示出)。在至少一个实施例中，计算机1210 包括但不限于服务器、云实例、膝上型计算机和台式计算机。

[0232] 在至少一个实施例中，USB棒1220包括但不限于处理单元1230、USB接口1240和USB接口逻辑1250。在至少一个实施例中，处理单元 1230可以是任何指令执行系统、装置或能够执行指令的设备。在至少一个实施例中，处理单元1230可以包括但不限于任何数量和类型的处理核心 (未示出)。在至少一个实施例中，处理核心1230包括专用集成电路

(“ASIC”),该专用集成电路被优化为执行与机器学习相关联的任何数量和类型的操作。例如,在至少一个实施例中,处理核心1230是张量处理单元(“TPC”),其被优化以执行机器学习推理操作。在至少一个实施例中,处理核心1230是视觉处理单元(“VPU”),其被优化以执行机器视觉和机器学习推理操作。

[0233] 在至少一个实施例中,USB接口1240可以是任何类型的USB连接器或USB插座。例如,在至少一个实施例中,USB接口1240是用于数据和电源的USB 3.0Type-C插座。在至少一个实施例中,USB接口1240是 USB 3.0Type-A连接器。在至少一个实施例中,USB接口逻辑1250可以包括使处理单元1230能够经由USB连接器1240与设备(例如计算机1210) 相连接的任何数量和类型的逻辑。

[0234] 图13A示出了示例性架构,其中多个GPU 1310-1313通过高速链路 1340-1343(例如,总线/点对点互连等)通信地耦合到多个多核心处理器 1305-1306。在一个实施例中,高速链路1340-1343支持4GB/s、30GB/s、80GB/s或更高的通信吞吐量。可以使用各种互连协议,包括但不限于PCIe 4.0或5.0以及NVLink 2.0。

[0235] 此外,在一个实施例中,两个或更多个GPU 1310-1313通过高速链路1329-1330互连,该高速链路可以使用与用于高速链路1340-1343的协议/链路相同或不同的协议/链路来实现。类似地,两个或更多个多核心处理器1305-1306可以通过高速链路1328连接,该高速链路可以是以20GB/s、30GB/s、120GB/s或更高的速度运行的对称多处理器(SMP)总线。可替代地,可以使用相同的协议/链路(例如,通过公共互连结构)来完成图13A 中所示的各种系统组件之间的所有通信。

[0236] 在一个实施例中,每个多核心处理器1305-1306分别经由存储器互连 1326-1327通信地耦合到处理器存储器1301-1302,并且每个GPU 1310-1313分别通过GPU存储器互连1350-1353通信地耦合到GPU存储器 1320-1323。存储器互连1326-1327和1350-1353可以利用相同或不同的存储器访问技术。作为示例而非限制,处理器存储器1301-1302和GPU存储器1320-1323可以是易失性存储器,诸如动态随机存取存储器(DRAM) (包括堆叠的DRAM)、图形DDR SDRAM(GDDR)(例如GDDR5、GDDR6),或高带宽存储器(HBM),和/或可以是非易失性存储器,例如3D XPoint或Nano-Ram。在一个实施例中,处理器存储器1301-1302的某些部分可以是易失性存储器,而另一部分可以是非易失性存储器(例如,使用两级存储器(2LM)层次结构)。

[0237] 如本文所述,尽管各种处理器1305-1306和GPU 1310-1313可以分别物理地耦合到特定存储器1301-1302、1320-1323,可以实现统一存储器架构,其中相同虚拟系统地址空间(也称为“有效地址”空间)分布在各个物理存储器之间。例如,处理器存储器1301-1302可以各自包含64GB的系统存储器地址空间,并且GPU存储器1320-1323可以各自包含32GB的系统存储器地址空间(在该示例中导致总计256GB的可寻址存储器大小)。

[0238] 图13B示出了根据一个示例性实施例的用于多核心处理器1307和图形加速模块1346之间互连的附加细节。图形加速模块1346可以包括集成在线路卡上的一个或更多个GPU芯片,该线路卡经由高速链路1340耦合到处理器1307。可选地,图形加速模块1346可以集成在作为处理器1307 的相同封装或芯片上。

[0239] 在至少一个实施例中,示出的处理器1307包括多个核心 1360A-1360D,每个核心都具有转换后备缓冲区1361A-1361D和一个或更多个高速缓存1362A-1362D。在至少一个实

施例中，核心1360A-1360D可以包括未示出的各种其他组件，用于执行指令和处理数据。在至少一个实施例中，高速缓存1362A-1362D可以包括级别1(L1)和级别2(L2)高速缓存。此外，一个或更多个共享高速缓存1356可以被包括在高速缓存 1362A-1362D中，并且由各组核心1360A-1360D共享。例如，处理器1307 的一个实施例包括24个核心，每个核心具有其自己的L1高速缓存，十二个共享的L2高速缓存，和十二个共享的L3高速缓存。在该实施例中，两个相邻核心共享一个或更多个L2和L3高速缓存。处理器1307和图形加速模块1346与系统存储器1314连接，该系统存储器1314可以包括图13A 中的处理器存储器1301-1302。

[0240] 通过一致性总线1364经由核心间通信为存储在各个高速缓存 1362A-1362D、1356和系统存储器1314中的数据和指令维护一致性。例如，每个高速缓存可以具有与其相关联的高速缓存一致性逻辑/电路，以响应于检测到对特定高速缓存行的读取或写入通过一致性总线1364进行通信。在一个实现中，通过一致性总线1364实现高速缓存监听协议，以监听(snoop) 高速缓存访问。

[0241] 在一个实施例中，代理电路1325将图形加速模块1346通信地耦合到一致性总线1364，从而允许图形加速模块1346作为核心1360A-1360D的对等方参与高速缓存一致性协议。特别地，在至少一个实施例中，接口1335 通过高速链路1340(例如，PCIe总线、NVLink等)提供到代理电路1325 的连接，并且接口1337将图形加速模块1346连接到链路1340。

[0242] 在一个实现中，加速器集成电路1336代表图形加速模块的多个图形处理引擎1331，1332，N提供高速缓存管理、存储器访问、上下文管理和中断管理服务。图形处理引擎1331，1332，N可各自包括单独的图形处理单元(GPU)。可选地，图形处理引擎1331，1332，N选择性地可以包括 GPU内的不同类型的图形处理引擎，诸如图形执行单元、媒体处理引擎(例如，视频编码器/解码器)、采样器和blit引擎。在至少一个实施例中，图形加速模块1346可以是具有多个图形处理引擎1331-1332，N的GPU，或者图形处理引擎1331-1332，N可以是集成在通用封装、线路卡或芯片上的各个GPU。

[0243] 在一个实施例中，加速器集成电路1336包括存储器管理单元(MMU) 1339，用于执行各种存储器管理功能，例如虚拟到物理存储器转换(也称为有效到真实存储器转换)，还包括用于访问系统存储器1314的存储器访问协议。MMU 1339还可包括转换后备缓冲区(“TLB”)(未示出)，用于高速缓存虚拟/有效到物理/真实地址转换。在一个实现中，高速缓存1338 可以存储命令和数据，用于图形处理引擎1331-1332，N有效地访问。在至少一个实施例中，将存储在高速缓存1338和图形存储器1333-1334，M中的数据与核心高速缓存1362A-1362D、1356和系统存储器1314保持一致。如前所述，可以经由代表高速缓存1338和图形存储器1333-1334，M的代理电路1325来完成该任务(例如，将与处理器高速缓存1362A-1362D、1356 上的高速缓存行的修改/访问有关的更新发送到高速缓存1338，并从高速缓存1338接收更新)。

[0244] 一组寄存器1345存储由图形处理引擎1331，1332，N执行的线程的上下文数据，并且上下文管理电路1348管理线程上下文。例如，上下文管理电路1348可以执行保存和恢复操作，以在上下文切换期间保存和恢复各个线程的上下文(例如，其中保存第一线程并且存储第二线程，以便可以由图形处理引擎执行第二线程)。例如，上下文管理电路1348在上下文切换时，可以将当前寄存器值存储到存储器中的(例如，由上下文指针识别的)指定区域。然后，当返回上下文时可以恢复寄存器值。在一个实施例中，中断管理电路1347接收并处理

从系统设备接收的中断。

[0245] 在一个实现中,MMU 1339将来自图形处理引擎1331的虚拟/有效地址转换为系统存储器1314中的真实/物理地址。加速器集成电路1336的一个实施例支持多个(例如,4、8、16)图形加速器模块1346和/或其他加速器设备。图形加速器模块1346可以专用于在处理器1307上执行的单个应用程序,或者可以在多个应用程序之间共享。在一个实施例中,呈现了虚拟化的图形执行环境,其中图形处理引擎1331-1332,N的资源与多个应用程序或虚拟机(VM)共享。在至少一个实施例中,可以基于处理要求和与VM和/或应用程序相关联的优先级,将资源细分为“切片”,其被分配给不同的VM和/或应用程序。

[0246] 在至少一个实施例中,加速器集成电路1336作为图形加速模块1346 的系统的桥来执行,并提供地址转换和系统存储器高速缓存服务。另外,加速器集成电路1336可以为主机处理器提供虚拟化设施,以管理图形处理引擎1331-1332的虚拟化、中断和存储器管理。

[0247] 由于图形处理引擎1331-1332,N的硬件资源被明确地映射到主机处理器1307看到的真实地址空间,因此任何主机处理器都可以使用有效地址值直接寻址这些资源。在至少一个实施例中,加速器集成电路1336的一个功能是物理分离图形处理引擎1331-1332,N,使得它们在系统看来为独立的单元。

[0248] 在至少一个实施例中,一个或更多个图形存储器1333-1334,M分别耦合到每个图形处理引擎1331-1332,N。图形存储器1333-1334,M存储指令和数据,所述指令和数据由每个图形处理引擎1331-1332,N处理。图形存储器1333-1334,M可以是易失性存储器,例如DRAM(包括堆叠的 DRAM)、GDDR存储器(例如,GDDR5,GDDR6)或HBM,和/或可以是非易失性存储器,例如3D XPoint或Nano-Ram。

[0249] 在一个实施例中,为了减少链路1340上的数据流量,可以使用偏置技术以确保存储在图形存储器1333-1334,M中的数据是图形处理引擎 1331-1332,N最常使用的,并且最好核心1360A-1360D不使用(至少不经常使用)的数据。类似地,在至少一个实施例中,偏置机制试图将核心 (并且优选地不是图形处理引擎1331-1332,N)需要的数据保持在核心的高速缓存1362A-1362D、1356和系统存储器1314中。

[0250] 图13C示出了另一个示例性实施例,其中加速器集成电路1336被集成在处理器1307内。在该实施例中,图形处理引擎1331-1332,N经由接口1337和接口1335(同样可以利用任何形式的总线或接口协议)通过高速链路1340直接与加速器集成电路1336通信。加速器集成电路1336可以执行与关于图13B描述的操作相同的操作。但是由于它紧密靠近一致性总线1364和高速缓存1362A-1362D、1356,可能具有更高的吞吐量。一个实施例支持不同的编程模型,包括专用进程编程模型(无图形加速模块虚拟化)和共享编程模型(具有虚拟化),所述编程模型可以包括由加速器集成电路1336控制的编程模型和由图形加速模块1346控制的编程模型。

[0251] 在至少一个实施例中,图形处理引擎1331-1332,N专用于单个操作系统下的单个应用程序或进程。在至少一个实施例中,单个应用程序可以将其他应用程序请求汇聚(funnel)到图形处理引擎1331-1332,N,从而在VM/分区内提供虚拟化。

[0252] 在至少一个实施例中,图形处理引擎1331-1332,N可以被多个VM/ 应用程序分区共享。在至少一个实施例中,共享模型可以使用系统管理程序来虚拟化图形处理引擎1331-1332,N,以允许每个操作系统进行访问。对于没有管理程序的单分区系统,操作系统拥有图

形处理引擎1331-1332,N。在至少一个实施例中,操作系统可以虚拟化图形处理引擎1331-1332, N,以提供对每个进程或应用程序的访问。

[0253] 在至少一个实施例中,图形加速模块1346或个体图形处理引擎 1331-1332,N使用进程句柄来选择进程元素。在一个实施例中,进程元素被存储在系统存储器1314中,并且可使用本文所述的有效地址到真实地址转换技术来寻址。在至少一个实施例中,进程句柄可以是特定于实现方式的值,其在向图形处理引擎1331-1332,N注册其上下文时提供给主机进程 (即,调用系统软件以将进程元素添加到进程元素链接列表)。在至少一个实施例中,进程句柄的较低16位可以是进程元素在进程元素链接列表中的偏移量。

[0254] 图13D示出了示例性加速器集成切片1390。如这里所用的,“切片”包括加速器集成电路1336的处理资源的指定部分。应用程序是系统存储器 1314中的有效地址空间1382,其存储进程元素1383。在一个实施例中,响应于来自在处理器1307上执行的应用程序1380的GPU调用1381,存储进程元素1383。进程元素1383包含相应的应用程序1380的进程状态。包含在进程元素1383中的工作描述符(WD)1384可以是由应用程序请求的单个作业,或者可以包含指向作业队列的指针。在至少一个实施例中, WD 1384是指向应用程序的地址空间1382中的作业请求队列的指针。

[0255] 图形加速模块1346和/或各个图形处理引擎1331-1332,N可以由系统中所有进程或进程子集共享。在至少一个实施例中,可以包括用于设置进程状态并将WD 1384发送到图形加速模块1346以在虚拟化环境中开始作业的基础设施。

[0256] 在至少一个实施例中,专用进程编程模型是特定于实现方式的。在该模型中,单个进程拥有图形加速模块1346或个体图形处理引擎1331。由于图形加速模块1346由单个进程拥有时,管理程序初始化用于所拥有的分区的加速器集成电路,当指派了图形加速模块1346时,操作系统初始化用于所拥有的进程的加速器集成电路1336。

[0257] 在操作中,加速器集成切片1390中的WD获取单元1391获取下一个WD 1384,其包括要由图形加速模块1346的一个或更多个图形处理引擎完成的工作的指示。来自WD 1384的数据可以存储在寄存器1345中,并由MMU 1339、中断管理电路1347和/或上下文管理电路1348使用,如图所示。例如,MMU 1339的一个实施例包括用于访问OS虚拟地址空间1385内的段/页表1386的段/页漫游电路。中断管理电路1347可以处理从图形加速模块1346接收的中断事件1392。当执行图形操作时,由图形处理引擎1331-1332,N生成的有效地址1393被MMU 1339转换为真实地址。

[0258] 在一个实施例中,为每个图形处理引擎1331-1332,N和/或图形加速模块1346复制寄存器1345的相同集合,并且所述寄存器1345可以由管理程序或操作系统初始化。这些复制的寄存器中的每一个可以被包括在加速器集成切片1390中。可以由管理程序初始化的示例性寄存器在表1中示出。

[0259]

表1–管理程序初始化的寄存器	
1	切片控制寄存器
2	实地址（RA）调度的进程区域指针
3	权限屏蔽覆写寄存器
4	中断向量表条目偏移量
5	中断向量表条目限制
6	状态寄存器
7	逻辑分区 ID
8	实地址（RA）管理程序加速器利用率记录指针
9	存储描述寄存器

[0260] 表2中示出了可由操作系统初始化的示例性寄存器。

[0261]

表2–操作系统初始化的寄存器	
1	进程和线程识别
2	有效地址（EA）上下文保存/恢复指针
3	虚拟地址（VA）加速器利用率记录指针
4	虚拟地址（VA）存储段表指针
5	权限屏蔽
6	工作描述符

[0262] 在一个实施例中，每个WD 1384特定于特定的图形加速模块1346和 /或图形处理引擎1331-1332,N。它包含图形处理引擎1331-1332,N完成工作所需的所有信息，或者它可以是指向存储器位置的指针，在该存储器位置应用程序已经设置了要完成的工作的命令队列。

[0263] 图13E示出了共享模型的一个示例性实施例的附加细节。该实施例包括管理程序实地址空间1398，其中存储了进程元素列表1399。可经由管理程序1396来访问管理程序实地址空间1398，所述管理程序1396虚拟化用于操作系统1395的图形加速模块引擎。

[0264] 在至少一个实施例中，共享编程模型允许来自系统中全部分区或分区子集的全部进程或进程子集使用图形加速模块1346。存在两种编程模型，其中图形加速模块1346由多个进程和分区共享：时间切片共享和图形定向共享。

[0265] 在该模型中，系统管理程序1396拥有图形加速模块1346，并使其功能可用于所有操作系统1395。对于图形加速模块1346通过系统管理程序 1396支持虚拟化，图形加速模块1346可以遵守以下：(1)应用程序的作业请求必须是自主的(即，不需要在作业之间保持状态)，或者图形加速模块1346必须提供上下文保存和恢复机制，(2)图形加速模块1346保证

应用程序的作业请求在指定的时间量内完成，包括任何转换错误，或者图形加速模块1346提供了抢占作业处理的能力，(3)在有向共享编程模型中进行操作时，必须确保图形加速模块1346进程之间的公平性。

[0266] 在至少一个实施例中，需要应用程序1380使用图形加速模块1346 类型、工作描述符(WD)、权限屏蔽寄存器(AMR)值和上下文保存/ 恢复区域指针(CSRP)进行操作系统1395系统调用。在至少一个实施例中，图形加速模块1346类型描述了用于系统调用的目标加速函数。在至少一个实施例中，图形加速模块1346类型可以是系统特定的值。在至少一个实施例中，WD是专门为图形加速模块1346格式化的，并且可以采用图形加速模块1346命令、指向用户定义的结构的有效地址指针、指向命令队列的有效地址指针的形式，或描述要由图形加速模块1346完成的工作的任何其他数据结构。在一个实施例中，AMR值是用于当前进程的AMR状态。在至少一个实施例中，传递给操作系统的值与设置AMR的应用程序类似。如果加速器集成电路1336和图形加速模块1346的实现不支持用户权限屏蔽覆写寄存器(UAMOR)，则在管理程序调用中传递AMR之前，操作系统可以将当前UAMOR值应用于AMR值。管理程序1396可以在将AMR 放入进程元素1383中之前选择性地应用当前权限屏蔽覆写寄存器(AMOR)值。在至少一个实施例中，CSRP是寄存器1345中的一个，所述寄存器包含应用程序的地址空间1382中的区域的有效地址，供图形加速模块1346 保存和恢复上下文状态。如果不需要在作业之间保存状态或者当作业被抢占时，则该指针是可选的。在至少一个实施例中，上下文保存/恢复区域可以是固定的系统存储器。

[0267] 在接收到系统调用时，操作系统1395可以验证应用程序1380已经注册并且被授予使用图形加速模块1346的权限。然后，操作系统1395使用表3中所示的信息来调用管理程序1396。

[0268]

表3–操作系统到管理程序的调用参数	
1	工作描述符（WD）
2	权限屏蔽寄存器（AMR）值（可能被屏蔽）
3	有效地址（EA）上下文保存/恢复区域指针（CSRP）
4	进程 ID（PID）和可选线程 ID（TID）
5	虚拟地址（VA）加速器利用率记录指针（AURP）
6	存储段表指针的虚拟地址（SSTP）
7	逻辑中断服务号（LISN）

[0269] 在接收到管理程序调用时，管理程序1396验证操作系统1395已注册并被授予使用图形加速模块1346的权限。然后，管理程序1396将进程元素1383放入相应的图形加速模块1346类型的进程元素链接列表中。进程元素可以包括表4中所示的信息。

[0270]

表4–进程元素信息	
1	工作描述符（WD）
2	权限屏蔽寄存器（AMR）值（可能被屏蔽）
3	有效地址（EA）上下文保存/恢复区域指针（CSRP）
4	进程ID（PID）和可选线程ID（TID）
5	虚拟地址（VA）加速器利用率记录指针（AURP）
6	存储段表指针的虚拟地址（SSTP）
7	逻辑中断服务号（LISN）
8	从管理程序调用参数派生的中断向量表
9	状态寄存器（SR）值
10	逻辑分区ID（LPID）
11	真实地址（RA）管理程序加速器利用率记录指针

[0271]

12	存储描述符寄存器（SDR）

[0272] 在至少一个实施例中，管理程序初始化多个加速器集成切片1390寄存器1345。

[0273] 如图13F所示，在至少一个实施例中，使用统一存储器，所述统一存储器可经由用于访问物理处理器存储器1301-1302和GPU存储器 1320-1323的公共虚拟存储器地址空间来寻址。在该实现方式中，在GPU 1310-1313上执行的操作利用相同的虚拟/有效存储器地址空间来访问处理器存储器1301-1302，反之亦然，从而简化了可编程性。在至少一个实施例中，虚拟/有效地址空间的第一部分被分配给处理器存储器1301，第二部分被分配给第二处理器存储器1302，第三部分被分配给GPU存储器1320，以此类推。在至少一个实施例中，整个虚拟/有效存储器空间(有时称为有效地址空间)由此分布在处理器存储器1301-1302和GPU存储器1320-1323 的每一个中，从而允许任何处理器或GPU采用映射到任何物理存储器的虚拟地址访问该存储器。

[0274] 在一个实施例中，一个或更多个MMU 1339A-1339E内的偏置/一致性管理电路1394A-1394E确保一个或更多个主机处理器(例如，1305)与 GPU 1310-1313的高速缓存之间的高速缓存一致性，并实现指示应在其中存储某些类型的数据的物理存储器的偏置技术。虽然在图13F中示出了偏置/一致性管理电路1394A-1394E的多个实例，但可以在一个或更多个主机处理器1305的MMU内和/或在加速器集成电路1336内实现偏置/一致性电路。

[0275] 一个实施例允许将GPU附接存储器1320-1323映射为系统存储器的一部分，并使用共享虚拟存储器(SVM)技术进行访问，但不会遭受与完整系统高速缓存一致性相关的性能缺陷。在至少一个实施例中，将GPU附接存储器1320-1323作为系统存储器来访问而无需繁重的高速缓存一致性开销的能力为GPU卸载提供了有利的操作环境。该布置允许主机处理器 1305的软件设置操作数并访问计算结果，而没有传统的I/O DMA数据拷贝的开销。这样的传统拷贝包括驱动程序调用、中断和存储器映射I/O (MMIO)访问，相对于简单的存储器

访问而言,这些访问效率均较低。在至少一个实施例中,在没有高速缓存一致性开销的情况下访问GPU附接存储器1320-1323的能力对于卸载的计算的执行时间可能是关键的。例如,在具有大量流式写入存储器流量的情况下,高速缓存一致性开销可以显著降低GPU 1310-1313所看到的有效写入带宽。在至少一个实施例中,操作数设置的效率、结果访问的效率和GPU计算的效率可能会在确定GPU卸载的有效性方面发挥作用。

[0276] 在至少一个实施例中,GPU偏置和主机处理器偏置的选择由偏置跟踪器数据结构驱动。例如,可以使用偏置表,所述偏置表可以是页面粒度结构(例如,以存储器页面的粒度来控制),该页面粒度结构包括每个GPU 附加的存储器页面1或2位。在至少一个实施例中,在GPU 1310-1313中具有或不具有偏置高速缓存(例如,用于高速缓存偏置表的频繁/最近使用的条目)的情况下,可以在一个或更多个GPU附接存储器1320-1323的被盗存储器范围中实现偏置表。替代地,可以在GPU内维护整个偏置表。

[0277] 在至少一个实施例中,在实际访问GPU存储器之前,访问与对GPU 附加存储器1320-1323的每次访问相关联的偏置表条目,从而引起以下操作。首先,来自GPU 1310-1313的在GPU偏置中找到其页面的本地请求被直接转发到对应的GPU存储器1320-1323。来自GPU的在主机偏置中找到其页面的本地请求被转发至处理器1305(例如,通过如上所述的高速链路)。在一个实施例中,来自处理器1305的在主机处理器偏置中找到所请求页面的请求完成了与正常存储器读取类似的请求。替代地,可以将指向GPU偏置页面的请求转发到GPU1310-1313。在至少一个实施例中,如果GPU当前不使用页面,则GPU可随后将页面迁移到主机处理器偏置。在至少一个实施例中,页面的偏置状态可以通过基于软件的机制、基于硬件辅助的软件的机制、或者在有限的情况下通过纯粹基于硬件的机制来改变。

[0278] 一种用于改变偏置状态的机制采用API调用(例如OpenCL),所述 API调用随后调用GPU的设备驱动程序,所述设备驱动程序随后发送消息 (或使命令描述符入队)到GPU,引导GPU改变偏置状态,并在某些迁移中在主机中执行高速缓存刷新操作。在至少一个实施例中,高速缓存刷新操作用于从主机处理器1305偏置到GPU偏置的迁移,但是不用于相反的迁移。

[0279] 在一个实施例中,高速缓存一致性是通过暂时渲染主机处理器1305 无法高速缓存的GPU偏置页面来维护的。为了访问这些页面,处理器1305 可以请求来自GPU 1310的访问,GPU 1310可以或可以不立即授予访问权限。因此,为了减少处理器1305和GPU 1310之间的通信,确保GPU偏置页面是GPU所需的页面而不是主机处理器1305所需的页面是有益的,反之亦然。

[0280] 一个或更多个硬件结构715用于执行一个或更多个实施例。在本文中可以结合图7A和/或图7B提供关于一个或更多个硬件结构715的细节。

[0281] 图14示出了根据本文所述的各个实施例的示例性集成电路和相关联的图形处理器,其可以使用一个或更多个IP核心来制造。除了图示之外,在至少一个实施例中可以包括其他逻辑和电路,包括附加的图形处理器/ 核心、外围接口控制器或通用处理器核心。

[0282] 图14是示出根据至少一个实施例的可使用一个或更多个IP核心制造的芯片集成电路1400上的示例性系统的框图。在至少一个实施例中,集成电路1400包括一个或更多个应用程序处理器1405(例如,CPU)、至少一个图形处理器1410,并且可以另外包括图像处理器1415和/或视频处理器1420,其中任意一个可能是模块化IP核心。在至少一个实施例中,

集成电路1400包括外围或总线逻辑，其包括USB控制器1425、UART控制器1430、SPI/SDIO控制器1435和I.sup.2S/I.sup.2C控制器1440。在至少一个实施例中，集成电路1400可以包括显示设备1445耦合到高清多媒体接口(HDMI)控制器1450和移动工业处理器接口(MIPI)显示接口1455 中的一个或更多个。在至少一个实施例中，存储可以由闪存子系统1460 提供，包括闪存和闪存控制器。在至少一个实施例中，可以经由存储器控制器1465提供存储器接口以用于访问SDRAM或SRAM存储器设备。在至少一个实施例中，一些集成电路还包括嵌入式安全引擎1470。

[0283] 图15A-15B示出了根据本文所述的各个实施例的示例性集成电路和相关联的图形处理器，其可以使用一个或更多个IP核心来制造。除了图示之外，在至少一个实施例中可以包括其他逻辑和电路，包括附加的图形处理器/核心、外围接口控制器或通用处理器核心。

[0284] 图15A-15B是示出根据本文描述的实施例的在SoC内使用的示例性图形处理器的框图。图15A示出了根据至少一个实施例的可以使用一个或更多个IP核心制造的片上系统集成电路的示例性图形处理器1510。图15B 示出了根据至少一个实施例的片上系统集成电路的另外示例性图形处理器 1540，其可以使用一个或更多个IP核心来制造。在至少一个实施例中，图 15A的图形处理器1510是低功耗图形处理器核心。在至少一个实施例中，图15B的图形处理器1540是更高性能的图形处理器核心。在至少一个实施例中，每个图形处理器1510、1540可以是图14的图形处理器1410的变体。

[0285] 在至少一个实施例中，图形处理器1510包括顶点处理器1505和一个或更多个片段处理器1515A-1515N(例如1515A、1515B、1515C、1515D 至1515N-1和1515N)。在至少一个实施例中，图形处理器1510可以经由单独的逻辑来执行不同的着色器程序，使得顶点处理器1505被优化以执行针对顶点着色器程序的操作，而一个或更多个片段处理器1515A-1515N执行片段(例如，像素)着色操作用于片段或像素或着色器程序。在至少一个实施例中，顶点处理器1505执行3D图形管线的顶点处理阶段并生成图元和顶点数据。在至少一个实施例中，一个或更多个片段处理器 1515A-1515N使用由顶点处理器1505生成的图元和顶点数据生成在显示设备上显示的帧缓冲区。在至少一个实施例中，一个或更多个片段处理器1515A-1515N被优化以执行如在OpenGL API中所提供的片段着色器程序，其可以用于执行与在Direct 3D API中所提供的像素着色器程序类似的操作。

[0286] 在至少一个实施例中，图形处理器1510附加地包括一个或更多个存储器管理单元(MMU)1520A-1520B、一个或更多个高速缓存1525A-1525B 和一个或更多个电路互连1530A-1530B。在至少一个实施例中，一个或更多个MMU 1520A-1520B提供用于图形处理器1510的虚拟到物理地址的映射，包括用于顶点处理器1505和/或片段处理器1515A-1515N，其可以引用存储在存储器中的顶点或图像/纹理数据，除了存储在一个或更多个高速缓存1525A-1525B中的顶点或图像/纹理数据之外。在至少一个实施例中，一个或更多个MMU 1520A-1520B可以与系统内的其他MMU同步，包括与图14的一个或更多个应用程序处理器1405、图像处理器1415和/或视频处理器1420相关联的一个或更多个MMU，使得每个处理器1405-1420可以参与共享或统一的虚拟存储器系统。在至少一个实施例中，一个或更多个电路互连1530A-1530B使图形处理器1510能够经由SoC的内部总线或经由直接连接与SoC内的其他IP核心相连接。

[0287] 在至少一个实施例中，图形处理器1540包括图15A的图形处理器 1510的一个或更

多个MMU 1520A-1520B、高速缓存1525A-1525B和电路互连1530A-1530B。在至少一个实施例中，图形处理器1540包括一个或更多个着色器核心1555A-1555N(例如，1555A、1555B、1555C、1555D、1555E、1555F到1555N-1和1555N)，其提供了统一的着色器核心架构，其中单个核心或类型或核心可以执行所有类型的可编程着色器代码，包括用于实现顶点着色器、片段着色器和/或计算着色器的着色器程序代码。在至少一个实施例中，多个着色器核心可以变化。在至少一个实施例中，图形处理器1540包括核心间任务管理器1545，其充当线程分派器以将执行线程分派给一个或更多个着色器核心1555A-1555N和分块单元1558，以加速基于图块渲染的分块操作，其中在图像空间中细分了场景的渲染操作，例如，以利用场景内的局部空间一致性或优化内部缓存的使用。

[0288] 图16A-16B示出了根据本文描述的实施例的附加示例性图形处理器逻辑。在至少一个实施例中，图16A示出了可以包括在图14的图形处理器1410内的图形核心1600，并且在至少一个实施例中，其可以是如图15B 所示的统一着色器核心1555A-1555N。图16B示出了在至少一个实施例中的适用于在多芯片模块上部署的高度并行的通用图形处理单元1630。

[0289] 在至少一个实施例中，图形核心1600包括共享指令高速缓存1602、纹理单元1618和高速缓存/共享存储器1620，它们对于图形核心1600内的执行资源是通用的。在至少一个实施例中，图形核心1600可包括多个切片 1601A-1601N或每个核心的分区，并且图形处理器可包括图形核心1600 的多个实例。切片1601A-1601N可包括支持逻辑，所述逻辑包括本地指令高速缓存1604A-1604N、线程调度器1606A-1606N、线程分派器 1608A-1608N和一组寄存器1610A-1610N。在至少一个实施例中，切片 1601A-1601N可以包括一组附加功能单元(AFU 1612A-1612N)、浮点单元(FPU 1614A-1614N)、整数算术逻辑单元(ALU 1616A-1616N)、地址计算单元(ACU 1613A-1613N)、双精度浮点单元(DPFPU 1615A-1615N) 和矩阵处理单元(MPU 1617A-1617N)。

[0290] 在至少一个实施例中，FPU 1614A-1614N可以执行单精度(32位) 和半精度(16位)浮点运算，而DPFPU 1615A-1615N则执行双精度(64 位)浮点运算点操作。在至少一个实施例中，ALU 1616A-1616N可以以8 位、16位和32位精度执行可变精度整数运算，并且可以配置为混合精度运算。在至少一个实施例中，MPU 1617A-1617N还可被配置用于混合精度矩阵运算，包括半精度浮点运算和8位整数运算。在至少一个实施例中， MPU 1617-1617N可以执行各种矩阵运算以加速机器学习应用程序框架，包括使得能够支持加速的通用矩阵到矩阵乘法(GEMM)。在至少一个实施例中，AFU 1612A-1612N可以执行浮点数或整数单元不支持的附加逻辑运算，包括三角运算(例如，正弦，余弦等)。

[0291] 图16B示出了在至少一个实施例中的通用处理单元(GPGPU)1630，其可以被配置为使得高度并行的计算操作能够由一组图形处理单元来执行。在至少一个实施例中，GPGPU 1630可以直接链接到GPGPU 1630的其他实例，以创建多GPU集群以提高用于深度神经网络的训练速度。在至少一个实施例中，GPGPU 1630包括主机接口1632，以实现与主机处理器的连接。在至少一个实施例中，主机接口1632是PCI Express接口。在至少一个实施例中，主机接口1632可以是厂商专用的通信接口或通信结构。在至少一个实施例中，GPGPU 1630接收主机处理器的命令，并使用全局调度器1634，以将与那些命令相关联的执行线程分配给一组计算集群 1636A-1636H。在至少一个实施例中，计算群集1636A-1636H共享高速缓存存储器1638。在至少一个实施例中，高速缓存存储器1638可以用作计算群集1636A-1636H内的高

速缓存存储器的更高级别的高速缓存。

[0292] 在至少一个实施例中，GPGPU 1630包括存储器1644A-1644B，所述存储器1644A-1644B经由一组存储器控制器1642A-1642B与计算集群 1636A-1636H耦合。在至少一个实施例中，存储器1644A-1644B可以包括各种类型的存储器设备，包括动态随机存取存储器(DRAM)或图形随机存取存储器，例如同步图形随机存取存储器(SGRAM)，其包括图形双倍数据速率(GDDR)存储器。

[0293] 在至少一个实施例中，计算集群1636A-1636H每个都包括一组图形核心，例如图16A的图形核心1600，所述图形核心可以包括多种类型的整数和浮点逻辑单元，所述逻辑单元可以在计算机各种精度范围上执行计算操作，包括适用于机器学习计算的精度。例如，在至少一个实施例中，每个计算集群1636A-1636H中的浮点单元的至少一个子集可以被配置为执行 16位或32位浮点运算，而浮点单元的不同子集可以配置为执行64位浮点运算。

[0294] 在至少一个实施例中，GPGPU 1630的多个实例可以被配置为用作计算集群。在至少一个实施例中，计算集群1636A-1636H用于同步和数据交换的通信在实施例之间变化。在至少一个实施例中，GPGPU 1630的多个实例通过主机接口1632进行通信。在至少一个实施例中，GPGPU 1630包括I/O集线器1639，所述集线器将GPGPU 1630与GPU链路1640耦合，使得能够直接连接到GPGPU 1630的其他实例。在至少一个实施例中，GPU 链路1640耦合到专用GPU到GPU桥，所述桥使得GPGP 1630的多个实例之间能够通信和同步。在至少一个实施例中，GPU链路1640与高速互连耦合，以向其他GPGPU或并行处理器发送和接收数据。在至少一个实施例中，GPGPU 1630的多个实例位于单独的数据处理系统中，并通过可通过主机接口1632访问的网络设备进行通信。在至少一个实施例中，GPU 链路1640可被配置为使得能够连接到主机除主机接口1632之外或作为其替代的处理器。

[0295] 在至少一个实施例中，GPGPU 1630可以被配置为训练神经网络。在至少一个实施例中，可以在推理平台内使用GPGPU 1630。在至少一个实施例中，在其中使用GPGPU 1630进行推理的情况下，相对于使用GPGPU 训练神经网络时，GPGPU可以包括更少的计算集群1636A-1636H。在至少一个实施例中，与存储器1644A-1644B相关联的存储器技术可以在推理和训练配置之间有所不同，其中更高带宽的存储器技术专用于训练配置。在至少一个实施例中，GPGPU 1630的推理配置可以支持推理特定指令。例如，在至少一个实施例中，推理配置可以提供对一个或更多个8位整数点积指令的支持，该指令可以在部署的神经网络的推理操作期间使用。

[0296] 图17示出了根据至少一个实施例的计算机系统1700的框图。在至少一个实施例中，计算机系统1700包括具有一个或更多个处理器1702的处理子系统1701和系统存储器1704，所述系统存储器1704经由可包括存储器集线器1705的互连路径通信。在至少一个实施例中，存储器集线器1705 可以是芯片组部件内的单独部件，或者可以集成在一个或更多个处理器 1702内。在至少一个实施例中，存储器集线器1705通过通信链路1706与 I/O子系统1711耦合。在一个实施例中，I/O子系统1711包括I/O集线器 1707，所述I/O集线器可以使计算机系统1700能够接收来自一个或更多个输入设备1708的输入。在至少一个实施例中，I/O集线器1707可以使显示控制器向一个或更多个显示设备1710A提供输出，所述显示控制器可以包括在一个或更多个处理器1702中。在至少一个实施例中，与I/O集线器 1707耦合的一个或更多个显示设备1710A可以包括本地，内部或嵌入式显示设备。

[0297] 在至少一个实施例中，处理子系统1701包括经由总线或其他通信链路1713耦合到存储器集线器1705的一个或更多个并行处理器1712中。在至少一个实施例中，通信链路1713可以是任何一种许多基于标准的通信链路技术或协议，例如但不限于PCI Express，或者可以是特定于供应商的通信接口或通信结构。在至少一个实施例中，一个或更多个并行处理器1712 形成计算集中的并行或向量处理系统，所述系统可以包括大量处理核心和/或处理集群，例如多集成核心(MIC)处理器。在至少一个实施例中，一个或更多个并行处理器1712形成图形处理子系统，所述图形处理子系统可以将像素输出到经由I/O集线器1707耦合的一个或更多个显示设备1710A 之一。在至少一个实施例中，一个或更多个并行处理器1712还可以包括显示控制器和显示接口(未示出)，以使得能够直接连接到一个或更多个显示设备1710B。

[0298] 在至少一个实施例中，系统存储单元1714可以连接到I/O集线器1707，以提供用于计算机系统1700的存储机制。在至少一个实施例中，I/O交换机1716可以用于提供一个接口机制，以实现I/O集线器1707与其他组件之间的连接，例如可以集成到平台中的网络适配器1718和/或无线网络适配器1717，以及可以通过一个或更多个附加设备1720添加的各种其他设备。在至少一个实施例中，网络适配器1718可以是以太网适配器或另一有线网络适配器。在至少一个实施例中，无线网络适配器1719可以包括Wi-Fi、蓝牙、近场通信(NFC)中的一个或更多个，或包括一个或更多个无线电设备的其他网络设备。

[0299] 在至少一个实施例中，计算机系统1700可以包括未明确示出的其他组件，所述其他组件包括USB或其他端口连接、光学存储驱动器、视频捕获设备等，所述其他组件也可以连接到I/O集线器1707。在至少一个实施例中，可以使用任何合适的协议(例如基于PCI(外围组件互连)的协议 (例如PCI-Express)或其他总线或点对点通信接口和/或协议)来实现互连图17中各个组件的通信路径，例如NV-Link高速互连或互连协议。

[0300] 在至少一个实施例中，一个或更多个并行处理器1712包括为图形和视频处理而优化的电路，所述电路包括例如视频输出电路，并构成图形处理单元(GPU)。在至少一个实施例中，一个或更多个并行处理器1712 包括为通用处理而优化的电路。在至少一个实施例中，计算机系统1700 的组件可以与单个集成电路上的一个或更多个其他系统元件集成。例如，在至少一个实施例中，一个或更多个并行处理器1712、存储器集线器1705、处理器1702和I/O集线器1707，可以被集成到片上系统(SoC)集成电路中。在至少一个实施例中，计算机系统1700的组件可以被集成到单个封装中，以形成系统级封装(SIP)配置。在至少一个实施例中，计算机系统 1700的组件的至少一部分可以被集成到多芯片模块(MCM)中，所述多芯片模块可以与其他多芯片模块互连到模块化计算机系统中。

[0301] 处理器

[0302] 图18A示出了根据至少一个实施例的并行处理器1800。在至少一个实施例中，并行处理器1800的各种组件可以使用一个或更多个集成电路设备来实现，例如可编程处理器、专用集成电路(ASIC)或现场可编程门阵列(FPGA)。在至少一个实施例中，所示的并行处理器1800是根据示例性实施例的图17所示的一个或更多个并行处理器1712的变体。

[0303] 在至少一个实施例中，并行处理器1800包括并行处理单元1802。在至少一个实施例中，并行处理单元1802包括I/O单元1804，其使得能够与其他设备进行通信，包括并行处理单元1802的其他实例。在至少一个实施例中，I/O单元1804可以直接连接到其他设备。在

至少一个实施例中，I/O单元1804通过使用集线器或交换机接口(例如,存储器集线器1705) 与其他设备连接。在至少一个实施例中,存储器集线器1705与I/O单元 1804之间的连接形成通信链路1713。在至少一个实施例中,I/O单元1804 与主机接口1806和存储器交叉开关1816连接,其中主机接口1806接收用于执行处理操作的命令,而存储器交叉开关1816接收用于执行存储器操作的命令。

[0304] 在至少一个实施例中,当主机接口1806经由I/O单元1804接收命令缓冲区时,主机接口1806可以引导工作操作以执行那些命令到前端1808。在至少一个实施例中,前端1808与调度器1810耦合,调度器1810配置成将命令或其他工作项分配给处理集群阵列1812。在至少一个实施例中,调度器1810确保在将任务分配给处理集群阵列1812之前,处理集群阵列 1812被正确地配置并且处于有效状态。在至少一个实施例中,调度器1810 通过在微控制器上执行的固件逻辑来实现。在至少一个实施例中,微控制器实现的调度器1810可配置成以粗粒度和细粒度执行复杂的调度和工作分配操作,从而实现对在处理阵列1812上执行的线程的快速抢占和上下文切换。在至少一个实施例中,主机软件可以证明用于通过多个图形处理门铃之一在处理阵列1812上进行调度的工作负载。在至少一个实施例中,工作负载然后可以由包括调度器1810的微控制器内的调度器1810逻辑在处理阵列1812上自动分配。

[0305] 在至少一个实施例中,处理集群阵列1812可以包括多达“N”个处理集群(例如,集群1814A、集群1814B到集群1814N)。在至少一个实施例中,处理集群阵列1812的每个集群1814A-1814N可以执行大量并发线程。在至少一个实施例中,调度器1810可以使用各种调度和/或工作分配算法将工作分配给处理集群阵列1812的集群1814A-1814N,其可以根据每种程序或计算类型产生的工作负载而变化。在至少一个实施例中,调度可以由调度器1810动态地处理,或者可以在配置为由处理集群阵列1812执行的程序逻辑的编译期间部分地由编译器逻辑来辅助。在至少一个实施例中,可将处理集群阵列1812的不同的集群1814A-1814N分配用于处理不同类型的程序或用于执行不同类型的计算。

[0306] 在至少一个实施例中,处理集群阵列1812可以配置成执行各种类型的并行处理操作。在至少一个实施例中,处理集群阵列1812配置成执行通用并行计算操作。例如,在至少一个实施例中,处理集群阵列1812可以包括执行处理任务的逻辑,该处理任务包括对视频和/或音频数据的过滤,执行建模操作,包括物理操作以及执行数据转换。

[0307] 在至少一个实施例中,处理集群阵列1812配置成执行并行图形处理操作。在至少一个实施例中,处理集群阵列1812可以包括附加逻辑以支持这种图形处理操作的执行,包括但不限于执行纹理操作的纹理采样逻辑,以及镶嵌逻辑和其他顶点处理逻辑。在至少一个实施例中,处理集群阵列 1812可以配置成执行与图形处理有关的着色器程序,例如但不限于顶点着色器、曲面细分着色器、几何着色器和像素着色器。在至少一个实施例中,并行处理单元1802可以经由I/O单元1804从系统存储器传送数据以进行处理。在至少一个实施例中,在处理期间,可以在处理期间将传送的数据存储到片上存储器(例如,并行处理器存储器1822),然后将其写回到系统存储器。

[0308] 在至少一个实施例中,当并行处理单元1802用于执行图形处理时,调度器1810可以配置成将处理工作负载划分为近似相等大小的任务,以更好地将图形处理操作分配给处理集群阵列1812的多个集群1814A-1814N。在至少一个实施例中,处理集群阵列1812的部分

可以配置成执行不同类型的处理。例如，在至少一个实施例中，第一部分可以配置成执行顶点着色和拓扑生成，第二部分可以配置成执行镶嵌和几何着色，并且第三部分可以配置成执行像素着色或其他屏幕空间操作，以生成用于显示的渲染图像。在至少一个实施例中，可以将由集群1814A-1814N中的一个或更多个产生的中间数据存储在缓冲区中，以允许在集群1814A-1814N之间传输中间数据以进行进一步处理。

[0309] 在至少一个实施例中，处理集群阵列1812可以经由调度器1810接收要执行的处理任务，该调度器1810从前端1808接收定义处理任务的命令。在至少一个实施例中，处理任务可以包括要被处理的数据的索引，例如，表面(补丁)数据、原始数据、顶点数据和/或像素数据，以及状态参数和定义如何处理数据的命令(例如，要执行什么程序)。在至少一个实施例中，调度器1810可以配置成获取与任务相对应的索引，或者可以从前端 1808接收索引。在至少一个实施例中，前端1808可以配置成确保在启动由传入命令缓冲区(例如，批缓冲区(batch-buffer)、推送缓冲区等)指定的工作负载之前，处理集群阵列1812配置成有效状态。

[0310] 在至少一个实施例中，并行处理单元1802的一个或更多个实例中的每一个可以与并行处理器存储器1822耦合。在至少一个实施例中，可以经由存储器交叉开关1816访问并行处理器存储器1822，所述存储器交叉开关1816可以接收来自处理集群阵列1812以及I/O单元1804的存储器请求。在至少一个实施例中，存储器交叉开关1816可以经由存储器接口1818访问并行处理器存储器1822。在至少一个实施例中，存储器接口1818可以包括多个分区单元(例如，分区单元1820A、分区单元1820B到分区单元1820N)，其可各自耦合至并行处理器存储器1822的一部分(例如，存储器单元)。在至少一个实施例中，多个分区单元1820A-1820N为配置为等于存储器单元的数量，使得第一分区单元1820A具有对应的第一存储器单元1824A，第二分区单元1820B具有对应的存储器单元1824B，第N分区单元1820N具有对应的第N存储器单元1824N。在至少一个实施例中，分区单元1820A-1820N的数量可以不等于存储器设备的数量。

[0311] 在至少一个实施例中，存储器单元1824A-1824N可以包括各种类型的存储器设备，包括动态随机存取存储器(DRAM)或图形随机存取存储器，例如同步图形随机存取存储器(SGRAM)，包括图形双倍数据速率(GDDR)存储器。在至少一个实施例中，存储器单元1824A-1824N还可包括3D堆叠存储器，包括但不限于高带宽存储器(HBM)。在至少一个实施例中，可以跨存储器单元1824A-1824N来存储诸如帧缓冲区或纹理映射的渲染目标，从而允许分区单元1820A-1820N并行地写入每个渲染目标的部分，以有效地使用并行处理器存储器1822的可用带宽。在至少一个实施例中，可以排除并行处理器存储器1822的本地实例，以有利于利用系统存储器与本地高速缓存存储器结合的统一存储器设计。

[0312] 在至少一个实施例中，处理集群阵列1812的集群1814A-1814N中的任何一个都可以处理将被写入并行处理器存储器1822内的任何存储器单元1824A-1824N中的数据。在至少一个实施例中，存储器交叉开关1816 可以配置为将每个集群1814A-1814N的输出传输到任何分区单元 1820A-1820N或另一个集群1814A-1814N，集群1814A-1814N可以对输出执行其他处理操作。在至少一个实施例中，每个集群1814A-1814N可以通过存储器交叉开关1816与存储器接口1818通信，以从各种外部存储设备读取或写入各种外部存储设备。在至少一个实施例中，存储器交叉开关 1816具有到存储器接口1818的连接以与I/O单元1804通信，

以及到并行处理器存储器1822的本地实例的连接，从而使不同处理集群1814A-1814N 内的处理单元与系统存储器或不是并行处理单元1802本地的其他存储器进行通信。在至少一个实施例中，存储器交叉开关1816可以使用虚拟通道来分离集群1814A-1814N和分区单元1820A-1820N之间的业务流。

[0313] 在至少一个实施例中，可以在单个插入卡上提供并行处理单元1802 的多个实例，或者可以将多个插入卡互连。在至少一个实施例中，并行处理单元1802的不同实例可以配置成相互操作，即使不同实例具有不同数量的处理核心，不同数量的本地并行处理器存储器和/或其他配置差异。例如，在至少一个实施例中，并行处理单元1802的一些实例可以包括相对于其他实例而言更高精度的浮点单元。在至少一个实施例中，结合并行处理单元1802或并行处理器1800的一个或更多个实例的系统可以以各种配置和形式因素来实现，包括但不限于台式机、膝上型计算机或手持式个人计算机、服务器、工作站、游戏机和/或嵌入式系统。

[0314] 图18B是根据至少一个实施例的分区单元1820的框图。在至少一个实施例中，分区单元1820是图18A的分区单元1820A-1820N之一的实例。在至少一个实施例中，分区单元1820包括L2高速缓存1821、帧缓冲区接口1825和ROP 1826(光栅操作单元)。L2高速缓存1821是读/写高速缓存，其配置成执行从存储器交叉开关1816和ROP 1826接收的加载和存储操作。在至少一个实施例中，L2高速缓存1821将读取未命中和紧急回写请求输出到帧缓冲区接口1825以进行处理。在至少一个实施例中，还可以经由帧缓冲区接口1825将更新发送到帧缓冲区以进行处理。在至少一个实施例中，帧缓冲区接口1825与并行处理器存储器中的存储器单元(诸如图 18A的存储器单元1824A-1824N(例如，在并行处理器存储器1822内)) 之一相互作用。

[0315] 在至少一个实施例中，ROP 1826是一种处理单元，其执行光栅操作，诸如模版、z测试、混合等。在至少一个实施例中，ROP 1826然后输出存储在图形存储器中的处理后的图形数据。在至少一个实施例中，ROP 1826 包括压缩逻辑以压缩被写入存储器的深度或颜色数据并解压缩从存储器读取的深度或颜色数据。在至少一个实施例中，压缩逻辑可以是利用多种压缩算法中的一个或更多个的无损压缩逻辑。ROP 1826执行的压缩的类型可以基于要压缩的数据的统计特性而变化。例如，在至少一个实施例中，基于每图块基础上的深度和颜色数据执行增量颜色压缩。

[0316] 在至少一个实施例中，ROP 1826包括在每个处理集群内(例如，图 18A的集群1814A-1814N)，而不是在分区单元1820内。在至少一个实施例中，通过存储器交叉开关1816而不是像素片段数据传输对像素数据的读取和写入请求。在至少一个实施例中，经处理的图形数据可以在显示设备上(诸如图17的一个或更多个显示设备1710之一)显示，由处理器1702 路由以供进一步处理，或者由图18A的并行处理器1800内的处理实体之一路由以供进一步处理。

[0317] 图18C是根据至少一个实施例的并行处理单元内的处理集群1814的框图。在至少一个实施例中，处理集群是图18A的处理集群1814A-1814N 之一的实例。在至少一个实施例中，处理集群1814可以配置成并行执行许多线程，其中术语“线程”是指在特定的一组输入数据上执行的特定程序的实例。在至少一个实施例中，单指令多数据(SIMD)指令发布技术用于支持大量线程的并行执行而无需提供多个独立的指令单元。在至少一个实施例中，使

用单指令多线程(SIMT)技术来支持并行执行大量一般同步的线程,这使用了公共指令单元,该公共指令单元配置成向每个处理集群内的一组处理引擎发出指令。

[0318] 在至少一个实施例中,可以通过将处理任务分配给SIMT并行处理器的管线管理器1832来控制处理集群1814的操作。在至少一个实施例中,管线管理器1832从图18A的调度器1810接收指令,通过图形多处理器 1834和/或纹理单元1836管理这些指令的执行。在至少一个实施例中,图形多处理器1834是SIMT并行处理器的示例性实例。然而,在至少一个实施例中,处理集群1814内可以包括不同架构的各种类型的SIMT并行处理器。在至少一个实施例中,在处理集群1814内可以包括图形多处理器1834 的一个或更多个实例。在至少一个实施例中,图形多处理器1834可以处理数据,并且数据交叉开关1840可以用于将处理后的数据分发到多个可能的目的(包括其他着色器单元)地之一。在至少一个实施例中,管线管理器 1832可以通过指定要经由数据交叉开关1840分配的处理后的数据的目的地来促进处理后的数据的分配。

[0319] 在至少一个实施例中,处理集群1814内的每个图形多处理器1834 可以包括相同的一组功能执行逻辑(例如,算术逻辑单元、加载存储单元等)。在至少一个实施例中,可以以管线方式配置功能执行逻辑,其中可以在先前的指令完成之前发出新的指令。在至少一个实施例中,功能执行逻辑支持多种操作,包括整数和浮点算术、比较操作、布尔运算、移位和各种代数函数的计算。在至少一个实施例中,可以利用相同的功能单元硬件来执行不同的操作,并且可以存在功能单元的任何组合。

[0320] 在至少一个实施例中,传送到处理集群1814的指令构成线程。在至少一个实施例中,跨一组并行处理引擎执行的一组线程是线程组。在至少一个实施例中,线程组在不同的输入数据上执行程序。在至少一个实施例中,线程组内的每个线程可被分配给图形多处理器1834内的不同处理引擎。在至少一个实施例中,线程组可包括比图形多处理器1834内的多个处理引擎更少的线程。在至少一个实施例中,当线程组包括的线程数少于处理引擎的数量时,一个或更多个处理引擎在正在处理该线程组的循环期间可能是空闲的。在至少一个实施例中,线程组还可以包括比图形多处理器1834 内的多个处理引擎更多的线程。在至少一个实施例中,当线程组包括比图形多处理器1834内的处理引擎的数量更多的线程时,可以在连续的时钟周期内执行处理。在至少一个实施例中,可以在图形多处理器1834上同时执行多个线程组。

[0321] 在至少一个实施例中,图形多处理器1834包括内部高速缓存存储器,以执行加载和存储操作。在至少一个实施例中,图形多处理器1834可以放弃内部高速缓存并使用处理集群1814内的高速缓存存储器(例如,L1高速缓存1848)。在至少一个实施例中,每个图形多处理器1834还可以访问分区单元(例如,图18A的分区单元1820A-1820N)内的L2高速缓存,这些分区单元在所有处理集群1814之间共享并且可以用于在线程之间传输数据。在至少一个实施例中,图形多处理器1834还可以访问片外全局存储器,其可以包括本地并行处理器存储器和/或系统存储器中的一个或更多个。在至少一个实施例中,并行处理单元1802外部的任何存储器都可以用作全局存储器。在至少一个实施例中,处理集群1814包括图形多处理器 1834的多个实例,它们可以共享可以存储在L1高速缓存1848中的公共指令和数据。

[0322] 在至少一个实施例中,每个处理集群1814可以包括配置成将虚拟地址映射为物理地址的存储器管理单元(“MMU”)1845。在至少一个实施例中,MMU 1845的一个或更多个实例

可以驻留在图18A的存储器接口1818 内。在至少一个实施例中,MMU 1845包括一组页表条目(PTE),其用于将虚拟地址映射到图块(更多地谈论平铺)的物理地址以及可选地映射到高速缓存行索引。在至少一个实施例中,MMU 1845可以包括地址转换后备缓冲区(TLB)或可以驻留在图形多处理器1834或L1高速缓存或处理集群1814内的高速缓存。在至少一个实施例中,处理物理地址以分配表面数据访问局部性,以便在分区单元之间进行有效的请求交织。在至少一个实施例中,高速缓存行索引可以用于确定对高速缓存线的请求是命中还是未命中。

[0323] 在至少一个实施例中,可以配置处理集群1814,使得每个图形多处理器1834耦合到纹理单元1836,以执行纹理映射操作,所述操作确定纹理样本位置、读取纹理数据以及过滤纹理数据。在至少一个实施例中,根据需要从内部纹理L1高速缓存(未示出)或从图形多处理器1834内的L1 高速缓存中读取纹理数据,并从L2高速缓存、本地并行处理器存储器或系统存储器中获取纹理数据。在至少一个实施例中,每个图形多处理器 1834将处理后的任务输出到数据交叉开关1840,以将处理后的任务提供给另一处理集群1814以进行进一步处理或将处理后的任务存储在L2高速缓存、本地并行处理器存储器、或经由存储器交叉开关1816的系统存储器中。在至少一个实施例中,PreROP 1842(光栅前操作单元)配置成从图形多处理器1834接收数据,将数据引导至ROP单元,该ROP单元可以与本文所述的分区单元(例如,图18A的分区单元1820A-1820N)一起定位。在至少一个实施例中,PreROP 1842单元可以执行用于颜色混合的优化、组织像素颜色数据以及执行地址转换。

[0324] 在至少一个实施例中,处理集群1814可以被配置为使得每个图形多处理器1834耦合到纹理单元1836,用于执行图5中实现的基带单元(BBU) 操作。

[0325] 图18D示出了根据至少一个实施例的图形多处理器1834。在至少一个实施例中,图形多处理器1834与处理集群1814的管线管理器1832耦合。在至少一个实施例中,图形多处理器1834具有执行管线,该执行管线包括但不限于指令高速缓存1852、指令单元1854、地址映射单元1856、寄存器文件1858、一个或更多个通用图形处理单元(GPGPU)核心1862和一个或更多个加载/存储单元1866。GPGPU核心1862和加载/存储单元1866 与高速缓存存储器1872和共享存储器1870通过存储器和高速缓存互连 1868耦合。

[0326] 在至少一个实施例中,指令高速缓存1852从管线管理器1832接收要执行的指令流。在至少一个实施例中,将指令高速缓存在指令高速缓存 1852中并将其分派以供指令单元1854执行。在一个实施例中,指令单元 1854可以分派指令作为线程组(例如,线程束),将线程组的每个线程分配给GPGPU核心1862内的不同执行单元。在至少一个实施例中,指令可以通过在统一地址空间内指定地址来访问任何本地、共享或全局地址空间。在至少一个实施例中,地址映射单元1856可以用于将统一地址空间中的地址转换成可以由加载/存储单元1866访问的不同的存储器地址。

[0327] 在至少一个实施例中,寄存器文件1858为图形多处理器1834的功能单元提供了一组寄存器。在至少一个实施例中,寄存器文件1858为连接到图形多处理器1834的功能单元(例如,GPGPU核心1862、加载/存储单元1866)的数据路径的操作数提供了临时存储。在至少一个实施例中,在每个功能单元之间划分寄存器文件1858,使得为每个功能单元分配寄存器文件1858的专用部分。在至少一个实施例中,寄存器文件1858在图形多处理器1834正在执行的不同线程束之间划分。

[0328] 在至少一个实施例中,GPGPU核心1862可以各自包括用于执行图形多处理器1834的指令的浮点单元(FPU)和/或整数算术逻辑单元(ALU)。GPGPU核心1862在架构上可以相似或架构可能有所不同。在至少一个实施例中,GPGPU核心1862的第一部分包括单精度FPU和整数ALU,而 GPGPU核心的第二部分包括双精度FPU。在至少一个实施例中,FPU可以实现用于浮点算法的IEEE 754-2008标准或启用可变精度浮点算法。在至少一个实施例中,图形多处理器1834可以另外包括一个或更多个固定功能或特殊功能单元,以执行特定功能,诸如复制矩形或像素混合操作。在至少一个实施例中,GPGPU核心中的一个或更多个也可以包括固定或特殊功能逻辑。

[0329] 在至少一个实施例中,GPGPU核心1862包括能够对多组数据执行单个指令的SIMD逻辑。在一个实施例中,GPGPU核心1862可以物理地执行SIMD4、SIMD8和SIMD16指令,并且在逻辑上执行SIMD1、SIMD2 和SIMD32指令。在至少一个实施例中,用于GPGPU核心的SIMD指令可以在编译时由着色器编译器生成,或者在执行针对单程序多数据(SPMD) 或SIMT架构编写和编译的程序时自动生成。在至少一个实施例中,可以通过单个SIMD指令来执行为SIMT执行模型配置的程序的多个线程。例如,在至少一个实施例中,可以通过单个SIMD8逻辑单元并行执行执行相同或相似操作的八个SIMT线程。

[0330] 在至少一个实施例中,存储器和高速缓存互连1868是将图形多处理器1834的每个功能单元连接到寄存器文件1858和共享存储器1870的互连网络。在至少一个实施例中,存储器和高速缓存互连1868是交叉开关互连,其允许加载/存储单元1866在共享存储器1870和寄存器文件1858之间实现加载和存储操作。在至少一个实施例中,寄存器文件1858可以以与 GPGPU核心1862相同的频率操作,从而在GPGPU核心1862和寄存器文件1858之间进行数据传输的延迟非常低。在至少一个实施例中,共享存储器1870可以用于启用在图形多处理器1834内的功能单元上执行的线程之间的通信。在至少一个实施例中,高速缓存存储器1872可以用作例如数据高速缓存,以高速缓存在功能单元和纹理单元1836之间通信的纹理数据。在至少一个实施例中,共享存储器1870也可以用作程序管理的高速缓存。在至少一个实施例中,除了存储在高速缓存存储器1872中的自动高速缓存的数据之外,在GPGPU核心1862上执行的线程还可以以编程方式将数据存储在共享存储器中。

[0331] 在至少一个实施例中,如本文所述的并行处理器或GPGPU通信地耦合到主机/处理器核心,以加速图形操作、机器学习操作、图案分析操作以及各种通用GPU(GPGPU)功能。在至少一个实施例中,GPU可以通过总线或其他互连(例如,诸如PCIe或NVLink的高速互连)通信地耦合到主机处理器/核心。在至少一个实施例中,GPU可以与核心集成在相同封装或芯片上,并通过内部处理器总线/互连(即,封装或芯片的内部)通信地耦合到核心。在至少一个实施例中,不管GPU连接的方式如何,处理器核心可以以工作描述符中包含的命令/指令序列的形式向该GPU分配工作。在至少一个实施例中,该GPU然后使用专用电路/逻辑来有效地处理这些命令/指令。

[0332] 图19示出了根据至少一个实施例的多GPU计算系统1900。在至少一个实施例中,多GPU计算系统1900可以包括经由主机接口交换机1904 耦合到多个通用图形处理单元(GPGPU)1906A-D的处理器1902。在至少一个实施例中,主机接口交换机1904是将处理器1902耦合到PCI Express 总线的PCI Express交换机设备,处理器1902可以通过PCI Express总线与 GPGPU 1906A-D通信。GPGPU 1906A-D可以经由一组高速P2P GPU到 GPU链

路1916互连。在至少一个实施例中，GPU到GPU链路1916经由专用GPU链路连接到GPGPU 1906A-D中的每一个。在至少一个实施例中， P2P GPU链路1916使得能够在每个GPGPU 1906A-D之间进行直接通信，而无需通过处理器1902所连接的主机接口总线1904进行通信。在至少一个实施例中，在GPU到GPU业务定向到P2P GPU链路1916的情况下，主机接口总线1904保持可用于系统存储器访问或例如经由一个或更多个网络设备与多GPU计算系统1900的其他实例进行通信。虽然在至少一个实施例中，GPGPU 1906A-D经由主机接口交换机1904连接到处理器1902，但是在至少一个实施例中，处理器1902包括对P2P GPU链路1916的直接支持，并且可以直接连接到GPGPU 1906A-D。

[0333] 图20是根据至少一个实施例的图形处理器2000的框图。在至少一个实施例中，图形处理器2000包括环形互连2002、管线前端2004、媒体引擎2037和图形核心2080A-2080N。在至少一个实施例中，环形互连2002 将图形处理器2000耦合到其他处理单元，所述处理单元包括其他图形处理器或一个或更多个通用处理器核心。在至少一个实施例中，图形处理器 2000是集成在多核心处理系统内的许多处理器之一。

[0334] 在至少一个实施例中，图形处理器2000经由环形互连2002接收多批命令。在至少一个实施例中，输入的命令由管线前端2004中的命令流转化器(command streamer)2003解释。在至少一个实施例中，图形处理器2000 包括可扩展执行逻辑，用于经由图形核心2080A-2080N执行3D几何处理和媒体处理。在至少一个实施例中，对于3D几何处理命令，命令流转化器2003将命令提供给几何管线2036。在至少一个实施例中，对于至少一些媒体处理命令，命令流转化器2003将命令提供给视频前端2034，该视频前端与媒体引擎2037耦合。在至少一个实施例中，媒体引擎2037包括用于视频和图像后处理的视频质量引擎(VQE)2030，以及用于提供硬件加速的媒体数据编码和解码的多格式编码/解码(MFX)2033引擎。在至少一个实施例中，几何管线2036和媒体引擎2037各自生成用于由至少一个图形核心2080A提供的线程执行资源的执行线程。

[0335] 在至少一个实施例中，图形处理器2000包括具有(featuring)模块化核心2080A-2080N(有时被称为核心切片)的可扩展线程执行资源，每个图形核心具有多个子核心2050A-2050N，2060A-2060N(有时称为核心子切片)。在至少一个实施例中，图形处理器2000可以具有任意数量的图形核心2080A至2080N。在至少一个实施例中，图形处理器2000包括具有至少第一子核心2050A和第二子核心2060A的图形核心2080A。在至少一个实施例中，图形处理器2000是具有单个子核心(例如2050A)的低功率处理器。在至少一个实施例中，图形处理器2000包括多个图形核心 2080A-2080N，每个图形核心包括一组第一子核心2050A-2050N和一组第二子核心2060A-2060N。在至少一个实施例中，第一子核心2050A-2050N 中的每个子核心至少包括第一组执行单元2052A-2052N和媒体/纹理采样器2054A-2054N。在至少一个实施例中，第二子核心2060A-2060N中的每个子核心至少包括第二组执行单元2062A-2062N和采样器2064A-2064N。在至少一个实施例中，每个子核心2050A-2050N，2060A-2060N共享一组共享资源2070A-2070N。在至少一个实施例中，共享资源包括共享高速缓存存储器和像素操作逻辑。

[0336] 图21是根据至少一个实施例的说明用于处理器2100的微架构的框图，该处理器2100可以包括用于执行指令的逻辑电路。在至少一个实施例中，处理器2100可以执行指令，包括x86指令、ARM指令、用于专用集成电路(ASIC)的专用指令等。在至少一个实施例中，处

理器2100可以包括用于存储封装数据的寄存器，例如作为加利福尼亚州圣克拉拉市英特尔公司采用MMX技术启用的微处理器中的64位宽MMX™寄存器。在至少一个实施例中，整数和浮点数形式可用的MMX寄存器可以与封装的数据元素一起运行，所述封装的数据元素伴随单指令多数据(“SIMD”)和流式 SIMD扩展(“SSE”)指令。在至少一个实施例中，与SSE2、SSE3、SSE4、AVX或更高版本(一般称为“SSEx”)技术有关的128位宽XMM寄存器可以保存此类封装数据操作数。在至少一个实施例中，处理器2110可以执行指令以加速机器学习或深度学习算法、训练或推理。

[0337] 在至少一个实施例中，处理器2100包括有序前端(“前端”)2101，以提取要执行的指令并准备稍后在处理器管线中使用的指令。在至少一个实施例中，前端2101可以包括几个单元。在至少一个实施例中，指令预取器2126从存储器中获取指令并将指令提供给指令解码器2128，指令解码器2128又对指令进行解码或解释。例如，在至少一个实施例中，指令解码器2128将接收到的指令解码为机器可执行的所谓的“微指令”或“微操作” (也称为“微操作”或“微指令”)的一个或更多个操作。在至少一个实施例中，指令解码器2128将指令解析为操作码以及相应的数据和控制字段，其可以由微架构用来使用以根据至少一个实施例来执行操作。在至少一个实施例中，跟踪高速缓存2130可以将解码的微指令组装成微指令队列2134 中的程序排序的序列或追踪以供执行。在至少一个实施例中，当追踪高速缓存2130遇到复杂指令时，微码ROM 2132提供完成操作所需的微指令。

[0338] 在至少一个实施例中，可以将一些指令转换成单个微操作，而另一些指令则需要几个微操作来完成全部操作。在至少一个实施例中，如果需要多于四个的微指令来完成一条指令，则指令解码器2128可以访问微码 ROM 2132以执行该指令。在至少一个实施例中，可以将指令解码为少量的微指令以在指令解码器2128处进行处理。在至少一个实施例中，如果需要多个微指令完成该操作，则可以将指令存储在微码ROM 2132中。在至少一个实施例中，追踪高速缓存器2130参考入口点可编程逻辑阵列(“PLA”) 以确定正确的微指令指针，用于根据至少一个实施例从微码ROM 2132读取微码序列以完成一个或更多个指令。在至少一个实施例中，在微码ROM 2132完成对指令的微操作排序之后，机器的前端2101可以恢复从追踪高速缓存2130获取微操作。

[0339] 在至少一个实施例中，乱序执行引擎(“乱序引擎”)2103可以准备用于执行的指令。在至少一个实施例中，乱序执行逻辑具有多个缓冲区，以使指令流平滑并重新排序，以在指令沿管线下降并被调度执行时优化性能。乱序执行引擎2103包括但不限于分配器/寄存器重命名器2140、存储器微指令队列2142、整数/浮点微指令队列2144、存储器调度器2146、快速调度器2102、慢速/通用浮点调度器(“慢速/通用FP调度器”)2104和简单浮点调度器(“简单FP调度器”)2106。在至少一个实施例中，快速调度器2102、慢速/通用浮点调度器2104和简单浮点调度器2106也统称为“微指令调度器2102、2104、2106”。分配器/寄存器重命名器2140分配每个微指令所需要的机器缓冲区和资源以便执行。在至少一个实施例中，分配器/ 寄存器重命名器2140将逻辑寄存器重命名为寄存器文件中的条目。在至少一个实施例中，分配器/寄存器重命名器2140还为两个微指令队列之一中的每个微指令分配条目，存储器微指令队列2142用于存储器操作和整数/ 浮点微指令队列2144用于非存储器操作，在存储器调度器2146和微指令调度器2102、2104、2106的前面。在至少一个实施例中，微指令调度器 2102、2104、2106基于它们的从属输入寄存器操作数源的就绪性和需要完成

的执行资源微指令的可用性来确定何时准备好执行微指令。至少一个实施例的快速调度器2102可以在主时钟周期的每个一半上调度，而慢速/通用浮点调度器2104和简单浮点调度器2106可以在每个主处理器时钟周期调度一次。在至少一个实施例中，微指令调度器2102、2104、2106对调度端口进行仲裁，以调度用于执行的微指令。

[0340] 在至少一个实施例中，执行块2111包括但不限于整数寄存器文件/ 支路网络2108、浮点寄存器文件/支路网络(“FP寄存器文件/支路网络”) 2110、地址生成单元(“AGU”)2112和2114、快速算术逻辑单元(“快速 ALU”)2116和2118、慢速算术逻辑单元(“慢速ALU”)2120、浮点ALU (“FP”)2122和浮点移动单元(“FP移动”)2124。在至少一个实施例中，整数寄存器文件/支路网络2108和浮点寄存器文件/旁路网络2110在本文中也称为“寄存器文件2108、2110”。在至少一个实施例中，AGU 2112和2114、快速ALU 2116和2118、慢速ALU 2120、浮点ALU 2122和浮点移动单元 2124在本文中也称为“执行单元2112、2114、2116、2118、2120、2122和2124”。在至少一个实施例中，执行块2111可以包括但不限于任意数量(包括零)和类型的寄存器文件、支路网络、地址生成单元和执行单元(以任何组合)。

[0341] 在至少一个实施例中，寄存器文件2108、2110可以布置在微指令调度器2102、2104、2106与执行单元2112、2114、2116、2118、2120、2122 和2124之间。在至少一个实施例中，整数寄存器文件/支路网络2108执行整数运算。在至少一个实施例中，浮点寄存器文件/支路网络2110执行浮点操作。在至少一个实施例中，寄存器文件2108、2110中的每一个可以包括但不限于支路网络，该支路网络可以绕过或转发尚未写入寄存器文件中的刚刚完成的结果到新的从属对象。在至少一个实施例中，寄存器文件 2108、2110可以彼此通信数据。在至少一个实施例中，整数寄存器文件/ 支路网络2108可以包括但不限于两个单独的寄存器文件、一个寄存器文件用于低阶32位数据，第二寄存器文件用于高阶32位数据。在至少一个实施例中，浮点寄存器文件/支路网络2110可以包括但不限于128位宽的条目，因为浮点指令通常具有宽度为64至128位的操作数。

[0342] 在至少一个实施例中，执行单元2112、2114、2116、2118、2120、 2122、2124可以执行指令。在至少一个实施例中，寄存器文件2108、2110 存储微指令需要执行的整数和浮点数据操作数值。在至少一个实施例中，处理器2100可以包括但不限于任何数量的执行单元2112、2114、2116、 2118、2120、2122、2124及其组合。在至少一个实施例中，浮点ALU 2122和浮点移动单元2124，可以执行浮点、MMX、SIMD、AVX和SSE或其他操作，包括专门的机器学习指令。在至少一个实施例中，浮点ALU 2122 可以包括但不限于64位乘64位浮点除法器，以执行除法、平方根和余数微操作。在至少一个实施例中，可以用浮点硬件来处理涉及浮点值的指令。在至少一个实施例中，可以将ALU操作传递给快速ALU 2116、2118。在至少一个实施例中，快速ALU 2116、2118可以以半个时钟周期的有效延迟执行快速操作。在至少一个实施例中，大多数复杂的整数运算进入慢速 ALU 2120，因为慢速ALU 2120可以包括但不限于用于长延迟类型操作的整数执行硬件，例如乘法器、移位、标志逻辑和分支处理。在至少一个实施例中，存储器加载/存储操作可以由AGUS 2112、2114执行。在至少一个实施例中，快速ALU 2116、快速ALU 2118和慢速ALU 2120可以对64 位数据操作数执行整数运算。在至少一个实施例中，可以实现快速ALU 2116、快速ALU 2118和慢速ALU 2120以支持包括十六、三十二、128、 256等的各种数据位大小。在至少一个实施例中，浮点ALU 2122和浮点移动单元2124可以实现为支持具有各种宽度的位的一定范围的操作数。在至少一个实施例中，浮点

ALU 2122和浮点移动单元2124可以结合SIMD和多媒体指令对128位宽封装数据操作数进行操作。

[0343] 在至少一个实施例中，微指令调度器2102、2104、2106在父加载完成执行之前调度从属操作。在至少一个实施例中，由于可以在处理器2100 中推测性地调度和执行微指令，处理器2100还可以包括用于处理存储器未命中的逻辑。在至少一个实施例中，如果数据高速缓存中的数据加载未命中，则可能存在在管线中正在运行的从属操作，其使调度器暂时没有正确的数据。在至少一个实施例中，一种重放机制追踪踪并重新执行使用不正确数据的指令。在至少一个实施例中，可能需要重放从属操作并且可以允许完成独立操作。在至少一个实施例中，处理器的至少一个实施例的调度器和重放机制也可以设计为捕获用于文本串比较操作的指令序列。

[0344] 在至少一个实施例中，术语“寄存器”可以指代可以用作识别操作数的指令的一部分的机载处理器存储位置。在至少一个实施例中，寄存器可以是那些可以从处理器外部使用的寄存器(从程序员的角度来看)。在至少一个实施例中，寄存器可能不限于特定类型的电路。相反，在至少一个实施例中，寄存器可以存储数据、提供数据并执行本文描述的功能。在至少一个实施例中，本文描述的寄存器可以通过处理器内的电路使用多种不同技术来实现，例如专用物理寄存器、使用寄存器重命名动态分配的物理寄存器、专用和动态分配的物理寄存器的组合等。在至少一个实施例中，整数寄存器存储32位整数数据。至少一个实施例的寄存器文件还包含八个用于封装数据的多媒体SIMD寄存器。

[0345] 图22示出了根据至少一个实施例的处理系统的框图。在至少一个实施例中，系统2200包括一个或更多个处理器2202和一个或更多个图形处理器2208，并且可以是单处理器台式机系统、多处理器工作站系统或具有大量处理器2202或处理器核心2207的服务器系统。在至少一个实施例中，系统2200是结合在片上系统(SoC)集成电路内的处理平台，以在移动、手持或嵌入式设备使用。

[0346] 在至少一个实施例中，系统2200可以包括或结合在基于服务器的游戏平台中，包括游戏和媒体控制台的游戏控制台、移动游戏控制台、手持游戏控制台或在线游戏控制台。在至少一个实施例中，系统2200是移动电话、智能电话、平板计算设备或移动互联网设备。在至少一个实施例中，处理系统2200还可包括与可穿戴设备耦合或集成在可穿戴设备中，例如智能手表可穿戴设备、智能眼镜设备、增强现实设备或虚拟现实设备。在至少一个实施例中，处理系统2200是电视或机顶盒设备，其具有一个或更多个处理器2202以及由一个或更多个图形处理器2208生成的图形界面。

[0347] 在至少一个实施例中，一个或更多个处理器2202每个包括一个或更多个处理器核心2207，以处理指令，该指令在被执行时执行针对系统和用户软件的操作。在至少一个实施例中，一个或更多个处理器核心2207中的每一个被配置为处理特定指令集2209。在至少一个实施例中，指令集2209 可以促进复杂指令集计算(CISC)、精简指令集计算(RISC)，或通过超长指令字(VLIW)进行计算。在至少一个实施例中，处理器核心2207可以各自处理不同的指令集2209，该指令集可以包括有助于仿真其他指令集的指令。在至少一个实施例中，处理器核心2207还可以包括其他处理设备，例如数字信号处理器(DSP)。

[0348] 在至少一个实施例中，处理器2202包括高速缓存存储器2204。在至少一个实施例中，处理器2202可以具有单个内部高速缓存或多个级别的内部高速缓存。在至少一个实施

例中，高速缓存存储器在处理器2202的各个组件之间共享。在至少一个实施例中，处理器2202还使用外部高速缓存(例如，三级(L3)高速缓存或最后一级高速缓存(LLC))(未示出)，可以使用已知的高速缓存一致性技术在处理器核心2207之间共享该外部高速缓存。在至少一个实施例中，处理器2202中另外包括寄存器文件2206，处理器可以包括用于存储不同类型的数据的不同类型的寄存器(例如，整数寄存器、浮点寄存器、状态寄存器和指令指针寄存器)。在至少一个实施例中，寄存器文件2206可以包括通用寄存器或其他寄存器。

[0349] 在至少一个实施例中，一个或更多个处理器2202与一个或更多个接口总线2210耦合，以在处理器2202与系统2200中的其他组件之间传输通信信号，例如地址、数据或控制信号。在至少一个实施例中，接口总线2210 在一个实施例中可以是处理器总线，例如直接媒体接口(DMI)总线的版本。在至少一个实施例中，接口2210不限于DMI总线，并且可以包括一个或更多个外围组件互连总线(例如，PCI，PCI Express)、存储器总线或其他类型的接口总线。在至少一个实施例中，处理器2202包括集成存储器控制器2216和平台控制器集线器2230。在至少一个实施例中，存储器控制器2216促进存储器设备与处理系统2200的其他组件之间的通信，而平台控制器集线器(PCH)2230通过本地I/O总线提供到输入/输出(I/O)设备的连接。

[0350] 在至少一个实施例中，存储器设备2220可以是动态随机存取存储器 (DRAM)设备、静态随机存取存储器(SRAM)设备、闪存设备、相变存储设备或具有适当的性能以用作处理器存储器。在至少一个实施例中，存储设备2220可以用作处理系统2200的系统存储器，以存储数据2222 和指令2221，以在一个或更多个处理器2202执行应用程序或过程时使用。在至少一个实施例中，存储器控制器2216还与可选的外部图形处理器2212 耦合，其可以与处理器2202中的一个或更多个图形处理器2208通信以执行图形和媒体操作。在至少一个实施例中，显示设备2211可以连接至处理器2202。在至少一个实施例中，显示设备2211可以包括内部显示设备中的一个或更多个，例如在移动电子设备或膝上型设备或通过显示器接口(例如显示端口(DisplayPort)等)连接的外部显示设备中。在至少一个实施例中，显示设备2211可以包括头戴式显示器(HMD)，诸如用于虚拟现实(VR)应用或增强现实(AR)应用中的立体显示设备。

[0351] 在至少一个实施例中，平台控制器集线器2230使外围设备能够通过高速I/O总线连接到存储设备2220和处理器2202。在至少一个实施例中， I/O外围设备包括但不限于音频控制器2246、网络控制器2234、固件接口 2228、无线收发器2226、触摸传感器2225、数据存储设备2224(例如，硬盘驱动器、闪存等)。在至少一个实施例中，数据存储设备2224可以经由存储接口(例如，SATA)或经由外围总线来连接，诸如外围组件互连总线(例如，PCI、PCIe)。在至少一个实施例中，触摸传感器2225可以包括触摸屏传感器、压力传感器或指纹传感器。在至少一个实施例中，无线收发器2226可以是Wi-Fi收发器、蓝牙收发器或移动网络收发器，诸如 3G、4G或长期演进(LTE)收发器。在至少一个实施例中，固件接口2228 使能与系统固件的通信，并且可以是例如统一可扩展固件接口(UEFI)。在至少一个实施例中，网络控制器2234可以启用到有线网络的网络连接。在至少一个实施例中，高性能网络控制器(未示出)与接口总线2210耦合。在至少一个实施例中，音频控制器2246是多通道高清晰度音频控制器。在至少一个实施例中，处理系统2200包括可选的传统(legacy)I/O控制器2240，用于将传统(例如，个人系统2(PS/2))设备耦合到系统2200。在至少一个实施例中，平

台控制器集线器2230还可以连接到一个或更多个通用串行总线(USB)控制器2242,该控制器连接输入设备,诸如键盘和鼠标2243组合、相机2244或其他USB输入设备。

[0352] 在至少一个实施例中,存储器控制器2216和平台控制器集线器2230 的实例可以集成到离散的外部图形处理器中,例如外部图形处理器2212。在至少一个实施例中,平台控制器集线器2230和/或存储器控制器2216可以在一个或更多个处理器2202的外部。例如,在至少一个实施例中,系统 2200可以包括外部存储器控制器2216和平台控制器集线器2230,其可以配置成在与处理器2202通信的系统芯片组中的存储器控制器集线器和外围控制器集线器。

[0353] 图23是根据至少一个实施例的具有一个或更多个处理器核心 2302A-2302N、集成存储器控制器2314和集成图形处理器2308的处理器 2300的框图。在至少一个实施例中,处理器2300可以包含附加核心,多达并包括以虚线框表示的附加核心2302N。在至少一个实施例中,每个处理器核心2302A-2302N包括一个或更多个内部高速缓存单元2304A-2304N。在至少一个实施例中,每个处理器核心还可以访问一个或更多个共享高速缓存单元2306。

[0354] 在至少一个实施例中,内部高速缓存单元2304A-2304N和共享高速缓存单元2306表示处理器2300内的高速缓存存储器层次结构。在至少一个实施例中,高速缓存存储器单元2304A-2304N可以包括每个处理器核心内的至少一级指令和数据高速缓存以及共享中级高速缓存中的一级或更多级缓存,例如2级(L2)、3级(L3)、4级(L4)或其他级别的高速缓存,其中将外部存储器之前的最高级别的高速缓存归类为LLC。在至少一个实施例中,高速缓存一致性逻辑维持各种高速缓存单元2306和 2304A-2304N之间的一致性。

[0355] 在至少一个实施例中,处理器2300还可包括一组一个或更多个总线控制器单元2316和系统代理核心2310。在至少一个实施例中,一个或更多个总线控制器单元2316管理一组外围总线,例如一个或更多个PCI或 PCIe总线。在至少一个实施例中,系统代理核心2310为各种处理器组件提供管理功能。在至少一个实施例中,系统代理核心2310包括一个或更多个集成存储器控制器2314,以管理对各种外部存储器设备(未示出)的访问。

[0356] 在至少一个实施例中,一个或更多个处理器核心2302A-2302N包括对多线程同时进行的支持。在至少一个实施例中,系统代理核心2310包括用于在多线程处理期间协调和操作核心2302A-2302N的组件。在至少一个实施例中,系统代理核心2310可以另外包括电源控制单元(PCU),该电源控制单元包括用于调节处理器核心2302A-2302N和图形处理器2308的一个或更多个电源状态的逻辑和组件。

[0357] 在至少一个实施例中,处理器2300还包括用于执行图处理操作的图形处理器2308。在至少一个实施例中,图形处理器2308与共享高速缓存单元2306和包括一个或更多个集成存储器控制器2314的系统代理核心 2310耦合。在至少一个实施例中,系统代理核心2310还包括用于驱动图形处理器输出到一个或更多个耦合的显示器的显示器控制器2311。在至少一个实施例中,显示器控制器2311也可以是经由至少一个互连与图形处理器2308耦合的独立模块,或者可以集成在图形处理器2308内。

[0358] 在至少一个实施例中,基于环的互连单元2312用于耦合处理器2300 的内部组件。在至少一个实施例中,可以使用替代性互连单元,例如点对点互连、交换互连或其他技术。在至少一个实施例中,图形处理器2308 经由I/O链路2313与环形互连2312耦合。

[0359] 在至少一个实施例中,I/O链路2313代表多种I/O互连中的至少一种,包括促进各

种处理器组件与高性能嵌入式存储器模块2318(例如eDRAM 模块)之间的通信的封装I/O互连。在至少一个实施例中,处理器核心 2302A-2302N和图形处理器2308中的每一个使用嵌入式存储器模块2318 作为共享的最后一级高速缓存。

[0360] 在至少一个实施例中,处理器核心2302A-2302N是执行公共指令集架构的同质核心。在至少一个实施例中,处理器核心2302A-2302N在指令集架构(ISA)方面是异构的,其中一个或更多个处理器核心2302A-2302N 执行公共指令集,而一个或更多个其他处理器核心2302A-2302N执行公共指令集的子集或不同指令集。在至少一个实施例中,就微架构而言,处理器核心2302A-2302N是异构的,其中具有相对较高功耗的一个或更多个核心与具有较低功耗的一个或更多个功率核心耦合。在至少一个实施例中,处理器2300可以在一个或更多个芯片上实现或被实现为SoC集成电路。

[0361] 图24是图形处理器2400的框图,该图形处理器可以是分立的图形处理单元,或者可以是与多个处理核心集成的图形处理器。在至少一个实施例中,图形处理器2400经由存储器映射的I/O接口与图形处理器2400上的寄存器以及放置在存储器中的命令进行通信。在至少一个实施例中,图形处理器2400包括用于访问存储器的存储器接口2414。在至少一个实施例中,存储器接口2414是到本地存储器、一个或更多个内部高速缓存、一个或更多个共享的外部高速缓存和/或到系统存储器的接口。

[0362] 在至少一个实施例中,图形处理器2400还包括用于将显示输出数据驱动到显示设备2420的显示控制器2402。在至少一个实施例中,显示控制器2402包括用于显示设备2420的一个或更多个覆盖平面的硬件以及多层视频或用户接口元素的组合。在至少一个实施例中,显示设备2420可以是内部或外部显示设备。在至少一个实施例中,显示设备2420是头戴式显示设备,例如虚拟现实(VR)显示设备或增强现实(AR)显示设备。在至少一个实施例中,图形处理器2400包括视频编解码器引擎2406,以将媒体编码、解码或转码为一个或更多个媒体编码格式,从一个或更多个媒体编码格式编码、解码或转码,或在一个或更多个媒体编码格式之间进行编码、解码或转码,所述媒体编码格式包括但不限于运动图像专家组(MPEG)格式(例如MPEG-2),高级视频编码(AVC)格式(例如 H.264/MPEG-4AVC,以及美国电影电视工程师协会(SMPTE)421M/VC-1) 和联合图像专家组(JPEG)格式(例如JPEG)和Motion JPEG(MJPEG) 格式。

[0363] 在至少一个实施例中,图形处理器2400包括块图像传送(BLIT)引擎2404,以执行二维(2D)光栅化器操作,包括例如位边界块传送。但是,在至少一个实施例中,使用图形处理引擎(GPE)2410的一个或更多个组件来执行2D图形操作。在至少一个实施例中,GPE 2410是用于执行图形操作(包括三维(3D)图形操作和媒体操作)的计算引擎。

[0364] 在至少一个实施例中,GPE 2410包括用于执行3D操作的3D管线 2412,例如使用对3D图元形状(例如,矩形、三角形等)进行操作的处理功能来渲染三维图像和场景。3D管线2412包括执行各种任务和/或产生到3D/媒体子系统2415的执行线程的可编程和固定功能元素。虽然3D管线2412可用于执行媒体操作,但是在至少一个实施例中,GPE 2410还包括媒体管线2416,其用于执行媒体操作,诸如视频后处理和图像增强。

[0365] 在至少一个实施例中,媒体管线2416包括固定功能或可编程逻辑单元,用于执行一个或更多个专门的媒体操作,例如视频解码加速,视频去隔行和视频编码加速,代替或代表视频编解码器引擎2406。在至少一个实施例中,媒体管线2416还包括线程产生单元,用于

产生线程以在3D/媒体子系统2415上执行。在至少一个实施例中，产生的线程在3D/媒体子系统 2415中包含的一个或更多个图形执行单元上执行媒体操作的计算。

[0366] 在至少一个实施例中，3D/媒体子系统2415包括用于执行3D管线 2412和媒体管线2416产生的线程的逻辑。在至少一个实施例中，3D管线 2412和媒体管线2416将线程执行请求发送到3D/媒体子系统2415，其包括用于仲裁各种请求并将其分派给可用线程执行资源的线程分派逻辑。在至少一个实施例中，执行资源包括用于处理3D和媒体线程的图形执行单元的阵列。在至少一个实施例中，3D/媒体子系统2415包括用于线程指令和数据的一个或更多个内部高速缓存。在至少一个实施例中，子系统2415 还包括共享存储器，其包括寄存器和可寻址存储器，以在线程之间共享数据并存储输出数据。

[0367] 图25是根据至少一个实施例的图形处理器的图形处理引擎2510的框图。在至少一个实施例中，图形处理引擎(GPE)2510是图24中所示的 GPE 2410的版本。在至少一个实施例中，媒体管线2516是可选的，并且可以不显式地包括在GPE 2510中。在至少一个实施例中，单独的媒体和/ 或图像处理器耦合到GPE 2510。

[0368] 在至少一个实施例中，GPE 2510耦合到或包括命令流转化器2503，其向3D管线2512和/或媒体管线2516提供命令流。在至少一个实施例中，命令流转化器2503耦合到存储器，所述存储器可以是系统存储器，也可以是内部高速缓存存储器和共享高速缓存存储器中的一个或更多个。在至少一个实施例中，命令流转化器2503从存储器接收命令，并且将命令发送到 3D管线2512和/或媒体管线2516。在至少一个实施例中，命令是从环形缓冲区中获取的指令、基元或微操作，该环形缓冲区存储用于3D管线2512 和媒体管线2516的命令。在至少一个实施例中，环形缓冲区还可以包括存储各批多个命令的批命令缓冲区。在至少一个实施例中，用于3D管线2512 的命令还可以包括对存储在存储器中的数据的引用，例如但不限于用于3D 管线2512的顶点和几何数据和/或用于媒体管线2516的图像数据和存储器对象。在至少一个实施例中，3D管线2512和媒体管线2516通过执行操作或通过将一个或更多个执行线程分派到图形核心阵列2514，来处理命令和数据。在至少一个实施例中，图形核心阵列2514包括一个或更多个图形核心块(例如，一个或更多个图形核心2515A、一个或更多个图形核心2515B)，每个块包括一个或更多个图形核心。在至少一个实施例中，每个图形核心包括一组图形执行资源，所述图形执行资源包括通用和图形特定的执行逻辑，用于执行图形和计算操作，以及固定功能纹理处理和/或机器学习和人工智能加速逻辑。

[0369] 在至少一个实施例中，3D管线2512包括固定功能和可编程逻辑，用于通过处理指令并将执行线程分派到图形核心阵列2514，来处理一个或更多个着色器程序，例如顶点着色器、几何着色器、像素着色器、片段着色器、计算着色器或其他着色器程序。在至少一个实施例中，图形核心阵列 2514提供统一的执行资源块，所述执行资源块用于处理着色器程序。在至少一个实施例中，在图形核心阵列2514的图形核心2515A-2515B内的多用途执行逻辑(例如，执行单元)包括对各种3D API着色器语言的支持，并且可以执行与多个着色器关联的多个同时执行线程。

[0370] 在至少一个实施例中，图形核心阵列2514还包括执行逻辑，用于执行媒体功能，诸如视频和/或图像处理。在至少一个实施例中，除了图形处理操作之外，执行单元还包括可编程以执行并行通用计算操作的通用逻辑。

[0371] 在至少一个实施例中，输出数据可以将数据输出到统一返回缓冲区 (URB)2518中

的存储器，所述输出数据由在图形核心阵列2514上执行的线程生成。在至少一个实施例中，URB 2518可以存储多个线程的数据。在至少一个实施例中，URB 2518可以用于在图形核心阵列2514上执行的不同线程之间发送数据。在至少一个实施例中，URB 2518还可用于图形核心阵列2514上的线程与共享功能逻辑2520内的固定功能逻辑之间的同步。

[0372] 在至少一个实施例中，图形核心阵列2514是可缩放的，使得图形核心阵列2514包括可变数量的图形核心，每个图形核心具有基于GPE 2510 的目标功率和性能水平的可变数量的执行单元。在至少一个实施例中，执行资源是动态可伸缩的，使得执行资源可以根据需要被启用或禁用。

[0373] 在至少一个实施例中，图形核心阵列2514耦合到共享功能逻辑2520，该共享功能逻辑包括在图形核心阵列2514中的图形核心之间共享的多个资源。在至少一个实施例中，由共享功能逻辑2520执行的共享功能体现在向图形核心阵列2514提供专门的补充功能的硬件逻辑单元中。在至少一个实施例中，共享功能逻辑2520包括但不限于采样器2521、数学2522和线程间通信(ITC)逻辑2523。在至少一个实施例中，一个或更多个高速缓存2525被包含在或耦合到共享功能逻辑2520中。

[0374] 在至少一个实施例中，如果对专用功能的需求不足以包含在图形核心阵列2514中，则使用共享功能。在至少一个实施例中，专用功能的单个实例在共享功能逻辑2520中使用，并且在图形核心阵列2514内的其他执行资源之间共享。在至少一个实施例中，特定共享功能可以包括在图形核心阵列2514内的共享功能逻辑2520内，所述特定共享功能在图形核心阵列 2514广泛使用的共享功能逻辑2516内。在至少一个实施例中，图形核心阵列2514内的共享功能逻辑2516可包括共享功能逻辑2520内的一些或全部逻辑。在至少一个实施例中，共享功能逻辑2520内的所有逻辑元件可在图形核心阵列2514的共享功能逻辑2516内复制。在至少一个实施例中，排除共享功能逻辑2520，以支持图形核心阵列2514内的共享功能逻辑2516。

[0375] 图26是根据本文所述的至少一个实施例的图形处理器核心2600的硬件逻辑的框图。在至少一个实施例中，图形处理器核心2600被包括在图形核心阵列内。在至少一个实施例中，图形处理器核心2600(有时称为核心切片)可以是模块化图形处理器内的一个或更多个图形核心。在至少一个实施例中，图形处理器核心2600是一个图形核心切片的示例，并且本文所述的图形处理器可以基于目标功率和性能包络线包括多个图形核心切片。在至少一个实施例中，每个图形核心2600可以包括与多个子核心 2601A-2601F耦合的固定功能块2630，也称为子切片，其包括通用和固定功能逻辑的模块块。

[0376] 在至少一个实施例中，固定功能块2630包括几何和固定功能管线 2636，例如，在较低性能和/或较低功率的图形处理器实施方式中，该几何和固定功能管线2636可以由图形处理器2600中的所有子核心共享。在至少一个实施例中，几何和固定功能管线2636包括3D固定功能管线、视频前端单元，线程产生器和线程分派器以及管理统一返回缓冲区的统一返回缓冲区管理器。

[0377] 在固定的至少一个实施例中，固定功能块2630还包括图形SoC接口 2637、图形微控制器2638和媒体管线2639。图形SoC接口2637提供了图形核心2600以及片上集成电路系统中的其他处理器核心之间的接口。在至少一个实施例中，图形微控制器2638是可编程子处理器，其可配置为管理图形处理器2600的各种功能，包括线程分派、调度和抢占。在至少

一个实施例中，媒体管线2639包括有助于对包括图像和视频数据的多媒体数据进行解码、编码、预处理和/或后处理的逻辑。在至少一个实施例中，媒体管线2639经由对子核心2601-2601F内的计算或采样逻辑的请求来实现媒体操作。

[0378] 在至少一个实施例中，SoC接口2637使图形核心2600能够与通用应用程序处理器核心(例如，CPU)和/或SoC内的其他组件通信，包括存储器层次结构元素，诸如共享的最后一级高速缓存、系统RAM和/或嵌入式片上或封装DRAM。在至少一个实施例中，SoC接口2637还可以使得能够与SoC内的固定功能设备(例如，相机成像管线)进行通信，并且使得能够使用和/或实现可以在图形核心2600和SoC内部的CPU之间共享的全局存储器原子。在至少一个实施例中，SoC接口2637还可以实现用于图形核心2600的电源管理控制，并且启用图形核心2600的时钟域与SoC内的其他时钟域之间的接口。在至少一个实施例中，SoC接口2637使得能够从命令流转化器和全局线程分派器接收命令缓冲区，其配置为向图形处理器内的一个或更多个图形核心中的每一个提供命令和指令。在至少一个实施例中，当要执行媒体操作时，可以将命令和指令分派给媒体管线2639，或者当要执行图形处理操作时，可以将其分配给几何形状和固定功能管线(例如，几何形状和固定功能管线2636，和/或几何形状和固定功能管线2614)。

[0379] 在至少一个实施例中，图形微控制器2638可以配置为对图形核心 2600执行各种调度和管理任务。在至少一个实施例中，图形微控制器2638 可以在子核心2601A-2601F中的执行单元(EU)阵列2602A-2602F、2604A-2604F内的各种图形并行引擎上执行图形和/或计算工作负载调度。在至少一个实施例中，在包括图形核心2600的SoC的CPU核心上执行的主机软件可以提交多个图形处理器门铃之一的工作负载，其调用适当的图形引擎上的调度操作。在至少一个实施例中，调度操作包括确定接下来要运行哪个工作负载、将工作负载提交给命令流转化器、抢先在引擎上运行的现有工作负载、监控工作负载的进度以及在工作负载完成时通知主机软件。在至少一个实施例中，图形微控制器2638还可以促进图形核心2600 的低功率或空闲状态，从而为图形核心2600提供在图形核心2600内独立于操作系统和/或系统上的图形驱动程序软件的跨低功率状态转换的保存和恢复寄存器的能力。

[0380] 在至少一个实施例中，图形核心2600可以具有比所示的子核心 2601A-2601F多或少达N个模块化子核心。对于每组N个子核心，在至少一个实施例中，图形核心2600还可以包括共享功能逻辑2610、共享和/或高速缓存存储器2612、几何/固定功能管线2614以及附加的固定功能逻辑 2616以加速各种图形和计算处理操作。在至少一个实施例中，共享功能逻辑2610可以包括可由图形核心2600内的每个N个子核心共享的逻辑单元 (例如，采样器、数学和/或线程间通信逻辑)。在至少一个实施例中，共享和/或高速缓存存储器2612可以是图形核心2600内的N个子核心 2601A-2601F的最后一级高速缓存，并且还可以用作可由多个子核心访问的共享存储器。在至少一个实施例中，可以包括几何/固定功能管线2614 来代替固定功能块2630内的几何/固定功能管线2636，并且可以包括相同或相似的逻辑单元。

[0381] 在至少一个实施例中，图形核心2600包括附加的固定功能逻辑2616，其可以包括供图形核心2600使用的各种固定功能加速逻辑。在至少一个实施例中，附加的固定功能逻辑2616包括用于仅位置着色中使用的附加的几何管线。在仅位置着色中，存在至少两个几何管线，而在几何和固定功能管线2614、2636内的完整几何管线和剔除管线中，其是可以包括在附加的固定功能逻辑2616中的附加几何管线。在至少一个实施例中，剔除管线是完整

几何管线的修整版。在至少一个实施例中，完整管线和剔除管线可以执行应用程序的不同实例，每个实例具有单独的环境。在至少一个实施例中，仅位置着色可以隐藏被丢弃的三角形的长剔除运行，从而在某些情况下可以更早地完成着色。例如，在至少一个实施例中，附加固定功能逻辑 2616中的剔除管线逻辑可以与主应用程序并行执行位置着色器，并且通常比完整管线更快地生成关键结果，因为剔除管线获取并遮蔽顶点的位置属性，无需执行光栅化和将像素渲染到帧缓冲区。在至少一个实施例中，剔除管线可以使用生成的临界结果来计算所有三角形的可见性信息，而与这些三角形是否被剔除无关。在至少一个实施例中，完整管线(在这种情况下可以称为重播管线)可以消耗可见性信息来跳过剔除的三角形以仅遮盖最终传递到光栅化阶段的可见三角形。

[0382] 在至少一个实施例中，附加的固定功能逻辑2616还可包括机器学习加速逻辑，例如固定功能矩阵乘法逻辑，用于实现包括用于机器学习训练或推理的优化。

[0383] 在至少一个实施例中，在每个图形子核心2601A-2601F内包括一组执行资源，其可用于响应于图形管线、媒体管线或着色器程序的请求来执行图形、媒体和计算操作。在至少一个实施例中，图形子核心2601A-2601F 包括多个EU阵列2602A-2602F、2604A-2604F，线程分派和线程间通信 (TD/IC)逻辑2603A-2603F，3D(例如，纹理)采样器2605A-2605F，媒体采样器2606A-2606F，着色器处理器2607A-2607F和共享本地存储器 (SLM)2608A-2608F。在至少一个实施例中，EU阵列2602A-2602F、 2604A-2604F每个都包含多个执行单元，这些执行单元是通用图形处理单元，能够为图形、媒体或计算操作提供服务，执行浮点和整数/定点逻辑运算，包括图形、媒体或计算着色器程序。在至少一个实施例中，TD/IC逻辑2603A-2603F为子核心内的执行单元执行本地线程分派和线程控制操作，并促进在子核心的执行单元上执行的线程之间的通信。在至少一个实施例中，3D采样器2605A-2605F可以将与纹理或其他3D图形相关的数据读取到存储器中。在至少一个实施例中，3D采样器可以基于与给定纹理相关联的配置的采样状态和纹理格式来不同地读取纹理数据。在至少一个实施例中，媒体采样器2606A-2606F可以基于与媒体数据相关联的类型和格式来执行类似的读取操作。在至少一个实施例中，每个图形子核心2601A-2601F 可以可替代地包括统一的3D和媒体采样器。在至少一个实施例中，在每个子核心2601A-2601F内的执行单元上执行的线程可以利用每个子核心内的共享本地存储器2608A-2608F，以使在线程组内执行的线程能够使用片上存储器的公共池来执行。

[0384] 图27A-27B示出了根据至少一个实施例的包括图形处理器核心的处理元件的阵列的线程执行逻辑2700。图27A示出了至少一个实施例，其中使用了线程执行逻辑2700。图27B示出了根据至少一个实施例的执行单元的示例性内部细节。

[0385] 如图27A中所示，在至少一个实施例中，线程执行逻辑2700包括着色器处理器2702、线程分派器2704、指令高速缓存2706、包括多个执行单元2708A-2708N的可缩放执行单元阵列、采样器2710、数据高速缓存 2712和数据端口2714。在至少一个实施例中，可缩放执行单元阵列可以例如基于工作负载的计算要求，通过启用或禁用一个或更多个执行单元(例如，执行单元2708A,2708B,2708C,2708D，至2708N-1和2708N中的任意一个)来动态缩放。在至少一个实施例中，可缩放执行单元通过链路到每个执行单元的互连结构互连。在至少一个实施例中，线程执行逻辑2700 包括通过指令高速缓存2706、数据端口2714、采样器2710和执行单元 2708A-2708N中的一个或更多个到存储器(诸如系统存储器或高速缓存存

储器)的一个或更多个连接。在至少一个实施例中,每个执行单元(例如 2708A)是独立的可编程通用计算单元,其能够执行多个同时的硬件线程,同时针对每个线程并行处理多个数据元素。在至少一个实施例中,执行单元2708A-2708N的阵列可缩放以包括任意数量的单独执行单元。

[0386] 在至少一个实施例中,执行单元2708A-2708N主要用于执行着色器程序。在至少一个实施例中,着色器处理器2702可以处理各种着色器程序并经由线程分派器2704来分派与着色器程序相关联的执行线程。在至少一个实施例中,线程分派器2704包括用于仲裁来自图形和媒体管线的线程初始化庆祝以及在执行单元2708A-2708N中的一个或更多个执行单元上实例化请求的线程的逻辑。例如,在至少一个实施例中,几何管线可以将顶点、镶嵌或几何着色器分派到线程执行逻辑以进行处理。在至少一个实施例中,线程分派器2704还可以处理来自执行着色器程序的运行时线程产生请求。

[0387] 在至少一个实施例中,执行单元2708A-2708N支持一种指令集,该指令集包括对许多标准3D图形着色器指令的本机支持,从而使图形库(例如Direct 3D和OpenGL)中的着色器程序只需最少的转换即可执行。在至少一个实施例中,执行单元支持顶点和几何处理(例如,顶点程序、几何程序、顶点着色器)、像素处理(例如,像素着色器、片段着色器)和通用处理(例如,计算和媒体着色器)。在至少一个实施例中,每个执行单元2708A-2708N包括一个或更多个算术逻辑单元(ALU),能够执行多发出单指令多数据(SIMD),并且多线程操作实现了高效的执行环境尽管有更高的延迟存储器访问。在至少一个实施例中,每个执行单元内的每个硬件线程具有专用的高带宽寄存器文件和相关的独立线程状态。在至少一个实施例中,执行是每个时钟到管线的多次发出,管线能够进行整数、单精度和双精度浮点运算、SIMD分支功能、逻辑运算、先验运算和其他其他运算。在至少一个实施例中,在等待来自存储器或共享功能之一的数据时,执行单元2708A-2708N内的依赖性逻辑使等待线程休眠直到返回了所请求的数据。在至少一个实施例中,当等待线程正在休眠时,硬件资源可以专用于处理其他线程。例如,在至少一个实施例中,在与顶点着色器操作相关联的延迟期间,执行单元可以对像素着色器、片段着色器或另一类型的着色器程序(包括不同的顶点着色器)执行操作。

[0388] 在至少一个实施例中,执行单元2708A-2708N中的每一个执行单元在数据元素的阵列上进行操作。在至少一个实施例中,多个数据元素是“执行大小”或指令的通道数。在至少一个实施例中,执行通道是用于指令内的数据元素访问、屏蔽和流控制的执行的逻辑单元。在至少一个实施例中,多个通道可以独立于用于特定图形处理器的多个物理算术逻辑单元(ALU) 或浮点单元(FPU)。在至少一个实施例中,执行单元2708A-2708N支持整数和浮点数据类型。

[0389] 在至少一个实施例中,执行单元指令集包括SIMD指令。在至少一个实施例中,各种数据元素可以作为封装数据类型存储在寄存器中,并且执行单元将基于那些元素的数据大小来处理各种元素。例如,在至少一个实施例中,当对256位宽的向量进行操作时,将向量的256位存储在寄存器中,并且执行单元对向量进行操作,作为四个单独的64位封装数据元素(四字(QW)大小数据元素)、八个单独的32位封装数据元素(双字(DW) 大小数据元素)、十六个单独的16位封装数据元素(单词(W)大小数据元素)或三十二个单独的8位数据元素(字节(B)大小的数据元素)。然而,在至少一个实施例中,不同的向量宽度和寄存器大小是可能

的。

[0390] 在至少一个实施例中，一个或更多个执行单元可以被组合成具有通常融合EU的线程控制逻辑(2707A-2707N)的融合执行单元2709A-2709N。在至少一个实施例中，可以将多个EU融合成一个EU组。在至少一个实施例中，融合EU组中的EU的数量可以配置为执行单独的SIMD硬件线程。融合的EU组中的EU的数量可能根据各个实施例而变化。在至少一个实施例中，每个EU可以执行各种SIMD宽度，包括但不限于SIMD8、SIMD16 和SIMD32。在至少一个实施例中，每个融合图形执行单元2709A-2709N 包括至少两个执行单元。例如，在至少一个实施例中，融合执行单元2709A 包括第一EU 2708A、第二EU 2708B以及第一EU 2708A和第二EU 2708B 共有的线程控制逻辑2707A。在至少一个实施例中，线程控制逻辑2707A 控制在融合图形执行单元2709A上执行的线程，从而允许融合执行单元 2709A-2709N内的每个EU使用公共指令指针寄存器来执行。

[0391] 在至少一个实施例中，一个或更多个内部指令高速缓存(例如2706) 被包括在线程执行逻辑2700中以高速缓存用于执行单元的线程指令。在至少一个实施例中，包括一个或更多个数据高速缓存(例如2712)以在线程执行期间高速缓存线程数据。在至少一个实施例中，包括采样器2710以提供用于3D操作的纹理采样和用于媒体操作的媒体采样。在至少一个实施例中，采样器2710包括专门的纹理或媒体采样功能，以在将采样数据提供给执行单元之前在采样过程中处理纹理或媒体数据。

[0392] 在执行期间，在至少一个实施例中，图形和媒体管线通过线程产生和分派逻辑将线程发起请求发送到线程执行逻辑2700。在至少一个实施例中，一旦一组几何对象已经被处理并光栅化成像素数据，则在着色器处理器 2702内的像素处理器逻辑(例如，像素着色器逻辑、片段着色器逻辑等) 被调用以进一步计算输出信息并且导致将结果写入输出表面(例如，颜色缓冲区、深度缓冲区、模板缓冲区等)。在至少一个实施例中，像素着色器或片段着色器计算要在光栅化对象上插值的各种顶点属性的值。在至少一个实施例中，着色器处理器2702内的像素处理器逻辑然后执行应用程序接口(API)提供的像素或片段着色器程序。在至少一个实施例中，为了执行着色器程序，着色器处理器2702经由线程分派器2704将线程分派到执行单元(例如2708A)。在至少一个实施例中，着色器处理器2702使用采样器2710中的纹理采样逻辑来访问存储在存储器中的纹理贴图中的纹理数据。在至少一个实施例中，对纹理数据和输入几何数据的算术运算为每个几何片段计算像素颜色数据，或者丢弃一个或更多个像素以进行进一步处理。

[0393] 在至少一个实施例中，数据端口2714提供了一种用于线程执行逻辑 2700的存储器访问机制，以将处理后的数据输出到存储器以在图形处理器输出管线上进行进一步处理。在至少一个实施例中，数据端口2714包括或耦合到一个或更多个高速缓存存储器(例如，数据高速缓存2712)以高速缓存数据以便经由数据端口进行存储器访问。

[0394] 如图27B所示，在至少一个实施例中，图形执行单元2708可以包括指令获取单元2737、通用寄存器文件阵列(GRF)2724、架构寄存器文件阵列(ARF)2726、线程仲裁器2722、发送单元2730、分支单元2732、一组SIMD浮点单元(FPU)2734，以及一组专用整数SIMD ALU 2735。在至少一个实施例中，GRF 2724和ARF 2726包括一组与可以在图形执行单元2708中活跃的每个同时硬件线程相关联的通用寄存器文件和架构寄存器文件。在至少一个实施例中，在ARF 2726中维护每个线程架构状态，而在线程执行期间使用的数据存储在GRF 2724

中。在至少一个实施例中，每个线程的执行状态，包括每个线程的指令指针，可以被保存在ARF 2726 中的线程专用寄存器中。

[0395] 在至少一个实施例中，图形执行单元2708具有一种架构，该架构是同时多线程(SMT)和细粒度交错多线程(IMT)的组合。在至少一个实施例中，架构具有模块化配置，该模块化配置可以在设计时基于同时线程的目标数量和每个执行单元的寄存器数量来进行微调，其中执行单元资源在用于执行多个同时线程的逻辑上分配。

[0396] 在至少一个实施例中，图形执行单元2708可以共同发布多个指令，每个指令可以是不同的指令。在至少一个实施例中，图形执行单元线程 2708的线程仲裁器2722可以将指令分派到发送单元2730、分支单元2742 或SIMD FPU 2734之一以供执行。在至少一个实施例中，每个执行线程可以访问GRF 2724中的128个通用寄存器，其中每个寄存器可以存储32个字节，可以作为32位数据元素的SIMD 8元素向量进行访问。在至少一个实施例中，每个执行单元线程可以访问GRF 2724中的4KB，尽管实施例不限于此，并且在其他实施例中可以提供更多或更少的寄存器资源。在至少一个实施例中，尽管每个执行单元的线程数量也可以根据实施例而变化，但是最多可以同时执行七个线程。在其中七个线程可以访问4KB的至少一个实施例中，GRF 2724可以存储总共28KB。在至少一个实施例中，灵活的寻址模式可以允许将寄存器一起寻址以有效地建立更宽的寄存器或表示跨步的矩形块数据结构。

[0397] 在至少一个实施例中，经由由消息传递发送单元2730执行的“发送”指令来调度存储器操作、采样器操作和其他更长延迟的系统通信。在至少一个实施例中，将分支指令分派到专门的分支单元2732促进SIMD发散和最终收敛。

[0398] 在至少一个实施例中，图形执行单元2708包括一个或更多个SIMD 浮点单元(FPU)2734，以执行浮点操作。在至少一个实施例中，一个或更多个FPU 2734还支持整数计算。在至少一个实施例中，一个或更多个 FPU 2734可以SIMD执行多达M个32位浮点(或整数)运算，或者SIMD 执行多达2M个16位整数或16位浮点运算。在至少一个实施例中，至少一个FPU提供扩展的数学能力以支持高吞吐量的先验数学函数和双精度64位浮点。在至少一个实施例中，还存在一组8位整数SIMD ALU 2735，并且可以被专门优化以执行与机器学习计算相关的操作。

[0399] 在至少一个实施例中，可以在图形子核心分组(例如，子切片)中实例化图形执行单元2708的多个实例的阵列。在至少一个实施例中，执行单元2708可以跨多个执行通道执行指令。在至少一个实施例中，在图形执行单元2708上执行的每个线程在不同的通道上执行。

[0400] 在至少一个实施例中，可以在图形子核心分组(例如，子切片)504 中实例化图形执行单元2708的多个实例的阵列，以执行信号处理操作。

[0401] 图28示出了根据至少一个实施例的并行处理单元(“PPU”)2800。在至少一个实施例中，PPU 2800配置有机器可读代码，该机器可读代码如果由PPU 2800执行，则使得PPU 2800执行贯穿本公开描述的一些或全部过程和技术。在至少一个实施例中，PPU 2800是在一个或更多个集成电路设备上实现的多线程处理器，并且利用多线程作为被设计为处理在多个线程上并行执行的计算机可读指令(也称为机器可读指令或简单的指令)的延迟隐藏技术。在至少一个实施例中，线程是指执行线程，并且是被配置为由PPU 2800执行的一组指令的实例。在至少一个实施例中，PPU 2800 是图形处理单元(“GPU”)，图形处理单元配置为

实现用于处理三维(“3D”) 图形数据的图形渲染管线,以便生成用于在显示设备(诸如液晶显示器 (“LCD”)设备)上显示的二维(“2D”)图像数据。在至少一个实施例中, PPU 2800用于执行计算,诸如线性代数运算和机器学习运算。图28仅出于说明性目的示出了示例并行处理器,并且应被解释为在本公开的范围内设想的处理器架构的非限制性示例,并且可以采用任何适当的处理器来对其进行补充和/或替代。

[0402] 在至少一个实施例中,一个或更多个PPU 2800配置成加速高性能计算(“HPC”)、数据中心和机器学习应用程序。在至少一个实施例中,PPU 2800配置成加速深度学习系统和应用程序,包括以下非限制性示例:自动驾驶汽车平台、深度学习、高精度语音、图像、文本识别系统、智能视频分析、分子模拟、药物发现、疾病诊断、天气预报、大数据分析、天文学、分子动力学模拟、财务建模、机器人技术、工厂自动化、实时语言转换、在线搜索优化以及个性化用户推荐等。

[0403] 在至少一个实施例中,PPU 2800包括但不限于输入/输出(“I/O”)单元2806、前端单元2810、调度器单元2812、工作分配单元2814、集线器2816、交叉开关(“Xbar”)2820、一个或更多个通用处理集群(“GPC”) 2818和一个或更多个分区单元(“存储器分区单元”)2822。在至少一个实施例中,PPU 2800通过一个或更多个高速GPU互连(“GPU互连”)2808 连接到主机处理器或其他PPU 2800。在至少一个实施例中,PPU 2800通过互连2802连接到主机处理器或其他外围设备。在一个实施例中,PPU 2800连接到包括一个或更多个存储器设备(“存储器”)2804的本地存储器。在至少一个实施例中,存储器设备2804包括但不限于一个或更多个动态随机存取存储器(“DRAM”)设备。在至少一个实施例中,一个或更多个DRAM 设备配置和/或可配置为高带宽存储器(“HBM”)子系统,并且在每个设备内堆叠有多个DRAM管芯。

[0404] 在至少一个实施例中,高速GPU互连2808可以指代系统使用其来进行缩放的基于线的多通道通信链路,并包括与一个或更多个中央处理单元结合的一个或更多个PPU 2800(“CPU”),支持PPU 2800和CPU之间的缓存相干以及CPU主控。在至少一个实施例中,高速GPU互连2808通过集线器2816将数据和/或命令传输到PPU 2800的其他单元,例如一个或更多个复制引擎、视频编码器、视频解码器、电源管理单元和/或在图28中可能未明确示出的其他组件。

[0405] 在至少一个实施例中,I/O单元2806配置为通过系统总线2802从主机处理器(图28中未示出)发送和接收通信(例如,命令、数据)。在至少一个实施例中,I/O单元2806直接通过系统总线2802或通过一个或更多个中间设备(例如存储器桥)与主机处理器通信。在至少一个实施例中, I/O单元2806可以经由系统总线2802与一个或更多个其他处理器(例如一个或更多个PPU 2800)通信。在至少一个实施例中,I/O单元2806实现外围组件互连Express(“PCIe”)接口,用于通过PCIe总线进行通信。在至少一个实施例中,I/O单元2806实现用于与外部设备通信的接口。

[0406] 在至少一个实施例中,I/O单元2806对经由系统总线2802接收的分组进行解码。在至少一个实施例中,至少一些分组表示被配置为使PPU 2800执行各种操作的命令。在至少一个实施例中,I/O单元2806如命令所指定的那样将解码的命令发送到PPU 2800的各种其他单元。在至少一个实施例中,命令被发送到前端单元2810和/或被发送到集线器2816或PPU 2800的其他单元,例如一个或更多个复制引擎、视频编码器、视频解码器、电源管理单

元等(图28中未明确示出)。在至少一个实施例中,I/O单元 2806配置为在PPU 2800的各种逻辑单元之间路由通信。

[0407] 在至少一个实施例中,由主机处理器执行的程序在缓冲区中对命令流进行编码,该缓冲区将工作负载提供给PPU 2800以进行处理。在至少一个实施例中,工作负载包括指令和要由那些指令处理的数据。在至少一个实施例中,缓冲区是可由主机处理器和PPU 2800两者访问(例如,读/写) 的存储器中的区域—主机接口单元可以配置为访问经由I/O单元2806通过系统总线2802传输的存储器请求连接到系统总线2802的系统存储器中的缓冲区。在至少一个实施例中,主机处理器将命令流写入缓冲区,然后将指示命令流开始的指针发送给PPU 2800,使得前端单元2810接收指向一个或更多个命令流指针并管理一个或更多个命令流,从命令流中读取命令并将命令转发到PPU 2800的各个单元。

[0408] 在至少一个实施例中,前端单元2810耦合到调度器单元2812,该调度器单元2812配置各种GPC 2818以处理由一个或更多个命令流定义的任务。在至少一个实施例中,调度器单元2812配置为跟踪与调度器单元2812 管理的各种任务有关的状态信息,其中状态信息可以指示任务被分配给哪个GPC 2818,任务是活跃的还是非活跃的,与任务相关联的优先级等等。在至少一个实施例中,调度器单元2812管理在一个或更多个GPC 2818上执行的多个任务。

[0409] 在至少一个实施例中,调度器单元2812耦合到工作分配单元2814,该工作分配单元2814配置为分派任务以在GPC 2818上执行。在至少一个实施例中,工作分配单元2814跟踪从调度器单元2812接收到的多个调度任务并且工作分配单元2814管理每个GPC 2818的待处理任务池和活动任务池。在至少一个实施例中,待处理任务池包括多个时隙(例如32个时隙),这些时隙包含分配给要由特定的GPC 2818处理的任务;活动任务池可包括用于由GPC 2818主动处理的任务的多个时隙(例如4个时隙),以使随着GPC 2818中的一个完成任务的执行,该任务将从GPC 2818的活动任务池中逐出,并且从待处理任务池中选择另一个任务,并安排其在GPC 2818上执行。在至少一个实施例中,如果活动任务在GPC 2818上处于空闲状态,例如在等待数据依赖性解决时,则活动任务从GPC 2818中驱逐并返回到待处理任务池,同时选择了待处理任务池中的另一个任务并调度在GPC 2818上执行。

[0410] 在至少一个实施例中,工作分配单元2814经由XBar2820与一个或更多个GPC 2818通信。在至少一个实施例中,XBar2820是互连网络,其将PPU 2800的许多单元耦合到PPU 2800的其他单元,并且可以配置为将工作分配单元2814耦合到特定的GPC2818。在至少一个实施例中,一个或更多个PPU 2800的其他单元也可以通过集线器2816连接到XBar2820。

[0411] 在至少一个实施例中,任务由调度器单元2812管理,并由工作分配单元2814分配给GPC 2818之一。在至少一个实施例中,GPC 2818配置为处理任务并产生结果。在至少一个实施例中,结果可以由GPC 2818中的其他任务消耗,通过XBar2820路由到不同的GPC 2818或存储在存储器 2804中。在至少一个实施例中,结果可以通过分区单元2822写到存储器2804中,其实现了用于向存储器2804写入数据或从存储器2804读取数据的存储器接口。在至少一个实施例中,结果可以经由高速GPU互连2808 传输到另一PPU 2804或CPU。在至少一个实施例中,PPU 2800包括但不限于U个分区单元2822,其等于耦合到PPU 2800的分离且不同的存储器设备2804的数量。在至少一个实施例中,分区单元2822将结合图30更详细地描述。

[0412] 在至少一个实施例中，主机处理器执行驱动器核心，该驱动程序核心实现应用程序编程接口(API)，该应用程序编程接口使在主机处理器上执行的一个或更多个应用程序能够调度操作以在PPU 2800上执行。在一个实施例中，多个计算应用程序由PPU 2800同时执行，并且PPU 2800为多个计算应用程序提供隔离、服务质量(“QoS”)和独立的地址空间。在至少一个实施例中，应用程序生成指令(例如，以API调用的形式)，该指令使驱动器核心生成一个或更多个任务以供PPU 2800执行，并且驱动器核心将任务输出至由PPU 2800处理的一个或更多个流。在至少一个实施例中，每个任务包括一个或更多个相关线程组，其可以被称为线程束 (warp)。在至少一个实施例中，线程束包括可以并行执行的多个相关线程(例如28个线程)。在至少一个实施例中，协作线程可以指代多个线程，包括用于执行任务并且通过共享存储器交换数据的指令，结合图30根据至少一个实施例更详细地描述了线程和协作线程。

[0413] 图29示出了根据至少一个实施例的通用处理集群(“GPC”)2900。在至少一个实施例中，GPC 2900是图28的GPC 2818。在至少一个实施例中，每个GPC 2900包括但不限于用于处理任务的多个硬件单元，并且每个GPC 2900包括但不限于管线管理器2902、预光栅操作单元(“PROP”) 2904、光栅引擎2908、工作分配交叉开关(“WDX”)2916、存储器管理单元(“MMU”)2918、一个或更多个数据处理集群(“DPC”)2906，以及部件的任何合适组合。

[0414] 在至少一个实施例中，GPC 2900的操作由管线管理器2902控制。在至少一个实施例中，管线管理器2902管理一个或更多个DPC 2906的配置，以处理分配给GPC 2900的任务。在至少一个实施例中，管线管理器2902 配置一个或更多个DPC 2906中的至少一个以实现图形渲染管线的至少一部分。在至少一个实施例中，DPC 2906配置为在可编程流式多处理器 (“SM”)2914上执行顶点着色器程序。在至少一个实施例中，管线管理器2902配置为将从工作分配单元接收的数据分组路由到GPC 2900内的适当逻辑单元，以及在至少一个实施例中，可以将一些数据分组路由到PROP 2904和/或光栅引擎2908中的固定功能硬件单元，而可以将其他数据分组路由到DPC 2906以由原始引擎2912或SM2914进行处理。在至少一个实施例中，管线管理器2902配置DPC 2906中的至少一个以实现神经网络模型和/或计算管线。

[0415] 在至少一个实施例中，PROP单元2904配置为在至少一个实施例中将由光栅引擎2908和DPC 2906生成的数据路由到分区单元2822中的光栅操作(“ROP”)单元，上面结合图28更详细地描述。在至少一个实施例中，PROP单元2904配置为执行用于颜色混合的优化、组织像素数据、执行地址转换等等。在至少一个实施例中，光栅引擎2908包括但不限于配置为执行各种光栅操作的多个固定功能硬件单元，并且在至少一个实施例中，光栅引擎2908包括但不限于设置引擎、粗光栅引擎、剔除引擎、裁剪引擎、精细光栅引擎、图块合计引擎及其任意合适的组合。在至少一个实施例中，设置引擎接收变换后的顶点并生成与由顶点定义的几何图元相关联的平面方程；平面方程式被传送到粗光栅引擎以生成基本图元的覆盖信息(例如，图块的x、y覆盖范围掩码)；粗光栅引擎的输出将传输到剔除引擎，在剔除引擎中与z测试失败的图元相关联的片段将被剔除，并传输到剪切引擎，在剪切引擎中剪切位于视锥范围之外的片段。在至少一个实施例中，将经过裁剪和剔除的片段传递给精细光栅引擎，以基于设置引擎生成的平面方程式生成像素片段的属性。在至少一个实施例中，光栅引擎2908的输出包括将由任何适当的实体(例如，由在DPC 2906内实现的片段着色器)处理的片段。

[0416] 在至少一个实施例中，包括在GPC 2900中的每个DPC 2906包括但不限于M管线控制器(“MPC”)2910；图元引擎2912；一个或更多个SM 2914；及其任何合适的组合。在至少一个实施例中，MPC 2910控制DPC 2906的操作，将从管线管理器2902接收的分组路由到DPC 2906中的适当单元。在至少一个实施例中，将与顶点相关联的分组路由到图元引擎2912，图元引擎2912配置为从存储器中获取与顶点关联的顶点属性；相反，可以将与着色器程序相关联的数据分组发送到SM 2914。

[0417] 在至少一个实施例中，SM 2914包括但不限于可编程流式处理器，其配置为处理由多个线程表示的任务。在至少一个实施例中，SM 2914是多线程的并且配置为同时执行来自特定线程组的多个线程(例如32个线程)，并且实现单指令、多数据(“SIMD”)架构，其中将一组线程(例如，线程束)中的每个线程配置为基于相同的指令集来处理不同的数据集。在至少一个实施例中，线程组中的所有线程执行相同指令。在至少一个实施例中， SM 2914实施单指令、多线程(“SIMT”)架构，其中一组线程中的每个线程配置为基于相同指令集来处理不同的数据集，但是其中线程组中的各个线程允许在执行期间发散。在至少一个实施例中，为每个线程束维护程序计数器、调用栈和执行状态，从而当线程束中的线程发散时，实现线程束和线程束内的串行执行之间的并发性。在另一个实施例中，为每个单独的线程维护程序计数器、调用栈和执行状态，从而使得在线程束内和线程束之间的所有线程之间具有相等的并发性。在至少一个实施例中，为每个单独的线程维持执行状态，并且可以收敛并并行地执行执行相同指令的线程以提高效率。本文更详细地描述SM 2914的至少一个实施例。

[0418] 在至少一个实施例中，MMU 2918在GPC 2900和存储器分区单元(例如，图28的分区单元2822)之间提供接口，并且MMU 2918提供虚拟地址到物理地址的转换、存储器保护以及存储器请求的仲裁。在至少一个实施例中，MMU 2918提供一个或更多个转换后备缓冲区(“TLB”)，用于执行虚拟地址到存储器中的物理地址的转换。

[0419] 图30示出了根据至少一个实施例的并行处理单元(“PPU”)的存储器分区单元3000。在至少一个实施例中，存储器分区单元3000包括但不限于光栅操作(“ROP”)单元3002；二级(“L2”)高速缓存3004；存储器接口3006；及其任何合适的组合。存储器接口3006耦合到存储器。存储器接口3006可以实现30、64、128、1024位数据总线，或者类似的实现方式用于高速数据传输。在至少一个实施例中，PPU包括U个存储器接口3006，每对分区单元3000一个存储器接口3006，其中每对分区单元3000连接到对应的存储器设备。例如，在至少一个实施例中，PPU可以连接至多达Y 个存储器设备，例如高带宽存储器堆栈或图形双数据速率版本5同步动态随机存取存储器(“GDDR5 SDRAM”)。

[0420] 在至少一个实施例中，存储器接口3006实现高带宽存储器第二代 (“HBM2”)存储器接口，并且Y等于U的一半。在至少一个实施例中， HBM2存储器堆栈位于相同物理封装上作为PPU，与传统的GDDR5 SDRAM系统相比，可提供大量功率并节省面积。在至少一个实施例中，每个HBM2堆栈包括但不限于四个存储器管芯，且Y＝4，每个HBM2堆栈包括每个管芯两个128位通道，用于总共8个通道和1024位的数据总线宽度。在至少一个实施例中，存储器支持单错误校正双错误检测 (“SECDED”)错误校正码(“ECC”)以保护数据。ECC可以为对数据损坏敏感的计算应用程序提供更高的可靠性。

[0421] 在至少一个实施例中，PPU实现了多级存储器层次结构。在至少一个实施例中，存储器分区单元3000支持统一存储器以为中央处理单元 (“CPU”)和PPU存储器提供单个统一

虚拟地址空间，从而实现虚拟存储器系统之间的数据共享。在至少一个实施例中，追踪PPU对位于其他处理器上的存储器的访问频率，以确保将存储器页面移动到更频繁地访问页面的PPU的物理存储器。在至少一个实施例中，高速GPU互连2808支持地址转换服务，其允许PPU直接访问CPU的页表，并通过PPU提供对CPU 存储器的完全访问。

[0422] 在至少一个实施例中，复制引擎在多个PPU之间或PPU与CPU之间传输数据。在至少一个实施例中，复制引擎可以为未被映射到页表中的地址生成页面错误，并且存储器分区单元3000然后为页面错误提供服务，将地址映射到页表中，之后复制引擎执行传输。在至少一个实施例中，为多个处理器之间的多个复制引擎操作固定(即不可分页)存储器，从而实质上减少了可用存储器。在至少一个实施例中，在硬件页面故障的情况下，可以将地址传递给复制引擎，而无需考虑是否驻留存储器页，并且复制过程是透明的。

[0423] 根据至少一个实施例，来自图28的存储器2804或其他系统存储器的数据由存储器分区单元3000获取，并将其存储在L2高速缓存3004中， L2高速缓存3004位于芯片上并且在各种GPC之间共享。在至少一个实施例中，每个存储器分区单元3000包括但不限于与对应的存储器设备相关联的L2高速缓存的至少一部分。在至少一个实施例中，在GPC内的各个单元中实现较低级别的高速缓存。在至少一个实施例中，图29的每个SM 2914可以实现一级(“L1”)高速缓存，其中L1高速缓存是专用于特定SM 2914的私有存储器，并且从L2高速缓存3004中获取数据并将其存储在每个L1高速缓存中，用于在SM 2914的功能单元中进行处理。在至少一个实施例中，L2高速缓存3004耦合到存储器接口3006和XBar 2820。

[0424] 在至少一个实施例中，ROP单元3002执行与像素颜色有关的图形光栅操作，诸如颜色压缩、像素混合等。在至少一个实施例中，ROP单元3002 结合光栅引擎2908实施深度测试，从光栅引擎2908的剔除引擎接收与像素片段相关联的样本位置的深度。在至少一个实施例中，针对在与片段关联的样本位置的深度缓冲区中的相应深度测试深度。在至少一个实施例中，如果该片段通过了针对该样本位置的该深度测试，则ROP单元3002更新深度缓冲区，并将该深度测试的结果发送给光栅引擎2908。将意识到，分区单元3000的数量可以不同于GPC的数量，因此，可以在至少一个实施例中将每个ROP单元3002耦合到每个GPC。在至少一个实施例中，ROP 单元3002追踪从不同GPC接收到的分组，并且确定ROP单元3002生成的结果是否要通过XBar 2820路由到。

[0425] 图31示出了根据至少一个实施例的流式多处理器(“SM”)3100。在至少一个实施例中，SM 3100是图29的SM。在至少一个实施例中，SM 3100 包括但不限于指令高速缓存3102；一个或更多个调度器单元3104；寄存器文件3108；一个或更多个处理核心(“核心”)3110；一个或更多个特殊功能单元(“SFU”)3112；一个或更多个加载/存储单元(“LSU”)3114；互连网络3116；共享存储器/一级(“L1”)高速缓存3118；以及/或其任何合适的组合。在至少一个实施例中，工作分配单元调度任务以在并行处理单元(“PPU”)的通用处理集群(“GPC”)上执行，并且每个任务被分配给 GPC内部的特定数据处理集群(“DPC”)，并且如果任务与着色器程序相关联，则将该任务分配给SM 3100之一。在至少一个实施例中，调度器单元3104从工作分配单元接收任务并管理分配给SM 3100的一个或更多个线程块的指令调度。在至少一个实施例中，调度器单元3104调度线程块以作为并行线程的线程束来执行，其中每个线程块被分配至少一个线程束。在至少一个实施例中，每个线程束执行线程。在至少一个实施例中，调度器单元3104管理多个不同的线程块，将线程束分配给不同的线程块，然后在每个时

钟周期内将来自多个不同的协作组的指令分派给各种功能单元 (例如,处理核心3110、SFU 3112和LSU 3114)。

[0426] 在至少一个实施例中,协作组可以指用于组织通信线程组的编程模型,其允许开发者表达线程正在通信的粒度,从而能够表达更丰富、更有效的并行分解。在至少一个实施例中,协作启动API支持线程块之间的同步以执行并行算法。在至少一个实施例中,常规编程模型的应用程序提供了用于同步协作线程的单一、简单的构造:跨线程块的所有线程的屏障(例如, syncthreads()函数)。但是,在至少一个实施例中,程序员可以在小于线程块粒度的情形下来定义线程组,并在所定义的组内进行同步,以实现更高的性能、设计灵活性以及以集合组范围功能接口的形式实现软件重用。在至少一个实施例中,协作组使程序员能够以子块(即,小到单个线程) 和多块粒度明确定义线程组,并执行集合操作,例如对协作组中的线程进行同步。该编程模型支持跨软件边界的干净组合,从而库和实用程序功能可以在其本地环境中安全地同步,而不必进行关于收敛的假设。在至少一个实施例中,协作组图元使协作并行的新图案成为可能,包括但不限于生产者-消费者并行,机会主义并行以及整个线程块网格上的全局同步。

[0427] 在至少一个实施例中,调度单元3106配置为将指令发送到功能单元中的一个或更多个,并且调度器单元3104并包括但不限于两个调度单元 3106,该两个调度单元3106使得来自相同线程束的两个不同指令能够在每个时钟周期被调度。在至少一个实施例中,每个调度器单元3104包括单个调度单元3106或附加调度单元3106。

[0428] 在至少一个实施例中,每个SM 3100在至少一个实施例中包括但不限于寄存器文件3108,该寄存器文件3108为SM 3100的功能单元提供了一组寄存器。在至少一个实施例中,寄存器文件3108在每个功能单元之间划分,从而为每个功能单元分配寄存器文件3108的专用部分。在至少一个实施例中,寄存器文件3108在由SM 3100执行的不同线程束之间划分,并且寄存器文件3108为连接到功能单元的数据路径的操作数提供临时存储。在至少一个实施例中,每个SM 3100包括但不限于多个L个处理核心3110,其中L是正整数。在至少一个实施例中,SM 3100包括但不限于大量(例如128个或更多)不同的处理核心3110。在至少一个实施例中,每个处理核心3110包括但不限于全管线、单精度、双精度和/或混合精度处理单元,其包括但不限于浮点算术逻辑单元和整数算术逻辑单元。在至少一个实施例中,浮点算术逻辑单元实现用于浮点算术的IEEE 754-2008标准。在至少一个实施例中,处理核心3110包括但不限于64个单精度(32 位)浮点核心、64个整数核心、32个双精度(64位)浮点核心和8个张量核心。

[0429] 根据至少一个实施例,张量核心配置为执行矩阵运算。在至少一个实施例中,一个或更多个张量核心包括在处理核心3110中。在至少一个实施例中,张量核心配置为执行深度学习矩阵算术,例如用于神经网络训练和推理的卷积运算。在至少一个实施例中,每个张量核心在4×4矩阵上操作并且执行矩阵乘法和累加运算D＝A×B+C,其中A、B、C和D是4×4矩阵。

[0430] 在至少一个实施例中,矩阵乘法输入A和B是16位浮点矩阵,并且累加矩阵C和D是16位浮点或32位浮点矩阵。在至少一个实施例中,张量核心对16位浮点输入数据进行32位浮点累加运算。在至少一个实施例中,16位浮点乘法使用64个运算,并得到全精度乘积,然后使用32位浮点加法与其他中间乘积累加起来,以进行4x4x4矩阵乘法。在至少一个实施例

中,张量核心用于执行由这些较小元件构成的更大的二维或更高维度的矩阵运算。在至少一个实施例中,API(诸如CUDA 9C++API)公开专门的矩阵加载、矩阵乘法和累加以及矩阵存储操作,以有效地使用来自 CUDA-C++程序的张量核心。在至少一个实施例中,在CUDA级别,线程束级别接口假定跨越所有32个线程束线程的16×16大小的矩阵。

[0431] 在至少一个实施例中,每个SM 3100包括但不限于执行特殊功能(例如,属性评估、倒数平方根等)的M个SFU 3112。在至少一个实施例中, SFU 3112包括但不限于配置为遍历分层树数据结构的树遍历单元。在至少一个实施例中,SFU 3112包括但不限于配置为执行纹理映射过滤操作的纹理单元。在至少一个实施例中,纹理单元配置为从存储器中加载纹理映射 (例如,纹理像素的2D阵列)和采样纹理映射,以产生采样的纹理值以供由SM 3100执行的着色器程序使用。在至少一个实施例中,将纹理映射存储在共享存储器/L1高速缓存3118中。在至少一个实施例中,根据至少一个实施例,纹理单元使用mip映射(mip-maps)(例如,细节级别不同的纹理映射)来实现纹理操作(诸如过滤操作)。在至少一个实施例中,每个SM 3100包括但不限于两个纹理单元。

[0432] 在至少一个实施例中,每个SM 3100包括但不限于实现共享存储器/ L1高速缓存3118与寄存器文件3108之间的加载和存储操作的N个LSU 3114。在至少一个实施例中,每个SM 3100包括但不限于互连网络3116 将每个功能单元连接到寄存器文件3108,并且LSU 3114连接到寄存器文件3108和共享存储器/L1高速缓存3118。在至少一个实施例中,互连网络3116是交叉开关,其可以配置为将任何功能单元连接到寄存器文件3108 中的任何寄存器,并且将LSU 3114连接到寄存器文件3108和共享存储器 /L1高速缓存3118中的存储器位置。

[0433] 在至少一个实施例中,共享存储器/L1高速缓存3118是片上存储器的阵列,其在至少一个实施例中允许SM 3100与图元引擎之间以及SM 3100中的线程之间的数据存储和通信。在至少一个实施例中,共享存储器 /L1高速缓存3118包括但不限于128KB的存储容量,并且位于从SM 3100 到分区单元的路径中。在至少一个实施例中,共享存储器/L1高速缓存3118 在至少一个实施例中用于高速缓存读取和写入。在至少一个实施例中,共享存储器/L1高速缓存3118、L2高速缓存和存储器中的一个或更多个是后备存储。

[0434] 在至少一个实施例中,将数据高速缓存和共享存储器功能组合到单个存储器块中,为两种类型的存储器访问提供了改进的性能。在至少一个实施例中,容量由不使用共享存储器的程序使用或将其用作高速缓存,例如如果共享存储器配置为使用一半容量,并且纹理和加载/存储操作可以使用剩余容量。根据至少一个实施例,在共享存储器/L1高速缓存3118内的集成使共享存储器/L1高速缓存3118能够用作用于流传输数据的高吞吐量管线,同时提供对频繁重用的数据的高带宽和低延迟访问。在至少一个实施例中,当配置用于通用并行计算时,与图形处理相比,可以使用更简单的配置。在至少一个实施例中,绕过固定功能图形处理单元,从而创建了更加简单的编程模型。在至少一个实施例中,在通用并行计算配置中,工作分配单元直接将线程的块分配和分布给DPC。在至少一个实施例中,块中的线程执行相同程序,在计算中使用唯一的线程ID以确保每个线程生成唯一的结果,使用SM 3100执行程序并执行计算,使用共享存储器/L1高速缓存3118在线程之间进行通信,以及使用LSU 3114通过共享存储器/L1 高速缓存3118和存储器分区单元来读写全局存储器。在至少一个实施例中,当被配置用于通用并行计算时,SM 3100向调度器单元3104写入可以

用来在DPC上启动新工作的命令。

[0435] 在至少一个实施例中，PPU被包括在台式计算机、膝上型计算机、平板电脑、服务器、超级计算机、智能电话(例如，无线、手持设备)、个人数字助理(“PDA”)、数码相机、车辆、头戴式显示器、手持式电子设备等中或与之耦合。在至少一个实施例中，PPU被实现在单个半导体衬底上。在至少一个实施例中，PPU与一个或更多个其他设备(例如附加的 PPU、存储器、精简指令集计算机(“RISC”)CPU，一个或更多个存储器管理单元(“MMU”)、数模转换器(“DAC”)等)一起被包括在片上系统 (“SoC”)中。

[0436] 在至少一个实施例中，PPU可以被包括在包括一个或更多个存储设备的图形卡上。该图形卡可以配置为与台式计算机主板上的PCIe插槽相连接。在至少一个实施例中，该PPU可以是包括在主板的芯片组中的集成图形处理单元(“iGPU”)。

[0437] 在至少一个实施例中，单个半导体平台可以指唯一的单一基于半导体的集成电路或芯片。在至少一个实施例中，可以使用具有增加的连接性的多芯片模块，其模拟芯片上的操作，并且相对于利用传统的中央处理单元 (“CPU”)和总线实现方式进行了实质性的改进。在至少一个实施例中，根据用户的需求，各种模块也可以分开放置或以半导体平台的各种组合放置。

[0438] 在至少一个实施例中，机器可读的可执行代码或计算机控制逻辑算法形式的计算机程序被存储在主存储器1104和/或辅助存储中。根据至少一个实施例，如果由一个或更多个处理器执行，则计算机程序使系统1100 能够执行各种功能。存储器1104、存储和/或任何其他存储是计算机可读介质的可能示例。在至少一个实施例中，辅助存储可以指代任何合适的存储设备或系统，例如硬盘驱动器和/或可移除存储驱动器，其代表软盘驱动器、磁带驱动器、光盘驱动器、数字多功能盘(“DVD”)驱动器、记录设备、通用串行总线(“USB”)闪存等。在至少一个实施例中，各个先前附图的架构和/或功能是在CPU 1102；并行处理系统1112；能够具有两个CPU 1102 的至少部分能力的集成电路；并行处理系统1112；芯片组(例如，设计成作为执行相关功能的单元工作并出售的一组集成电路等)；以及集成电路的任何适当组合的环境中实现的。

[0439] 在至少一个实施例中，各个先前附图的架构和/或功能在通用计算机系统、电路板系统、专用于娱乐目的的游戏控制台系统、专用系统等的环境中实现。在至少一个实施例中，计算机系统1100可以采取台式计算机、膝上型计算机、平板电脑、服务器、超级计算机、智能电话(例如，无线、手持设备)、个人数字助理(“PDA”)、数码相机、车辆、头戴式显示器、手持式电子设备、移动电话设备、电视、工作站、游戏机、嵌入式系统和/ 或任何其他类型的逻辑的形式。

[0440] 在至少一个实施例中，并行处理系统1112包括但不限于多个并行处理单元(“PPU”)1114和相关联的存储器1116。在至少一个实施例中，PPU 1114经由互连1118和交换机1120或多路复用器连接到主机处理器或其他外围设备。在至少一个实施例中，并行处理系统1112在可并行化的PPU 1114上分配计算任务，例如，作为跨多个图形处理单元(“GPU”)线程块的计算任务分布的一部分。在至少一个实施例中，在PPU 1114中的一些或全部之间共享和访问存储器(例如，用于读取和/或写入访问)，尽管这种共享存储器可能引发相对于使用本地存储器和驻留在PPU1114上的寄存器的性能损失。在至少一个实施例中，通过使用命令(诸如__syncthreads ())来同步PPU 1114的操作，其中块中的所有线程(例如，跨多

个PPU 1114执行)在进行之前到达某个代码执行点。

[0441] 网络

[0442] 图32示出了根据至少一个实施例的用于在5G无线通信网络内传送数据的网络3200。在至少一个实施例中,网络3200包括具有覆盖区域3204 的基站3206、多个移动设备3208和回程网络3202。在至少一个实施例中,如图所示,基站3206建立上行链路和/或下行链路与移动设备3208连接,用于将数据从移动设备3208传送到基站3206,反之亦然。在至少一个实施例中,通过上行链路/下行链路连接承载的数据可以包括在移动设备3208 之间传送的数据,以及通过回程网络3202的方式传送到/来自远程端(未示出)的数据。在至少一个实施例中,术语"基站"是指被配置为提供对网络的无线接入的任何组件(或组件的集合),例如增强型基站(eNB)、宏小区、毫微微小区、Wi-Fi接入点(AP)或其他启用无线功能的设备。在至少一个实施例中,基站可以根据一种或更多种无线通信协议提供无线接入,例如长期演进(LTE)、高级LTE(LTE-A)、高速分组接入(HSPA)、Wi-Fi 802.11a/b/g/n/ac等。在至少一个实施例中,术语"移动设备"是指能够与基站建立无线连接的任何组件(或组件的集合),例如用户设备(UE)、移动站(STA)和其他启用无线功能的设备。在一些实施例中,网络3200可以包括各种其他无线设备,例如中继、低功率节点等。

[0443] 图33示出了根据至少一个实施例的用于5G无线网络的网络架构 3300。在至少一个实施例中,如图所示,网络架构3300包括无线电接入网 (RAN)3304、演进分组核心(EPC)3302,其可以被称为核心网络,以及 UE3308的归属网络3316尝试接入RAN 3304。在至少一个实施例中,RAN 3304和EPC 3302形成服务无线网络。在至少一个实施例中,RAN 3304 包括基站3306,并且EPC 3302包括移动性管理实体(MME)3312、服务网关(SGW)3310和分组数据网络(PDN)网关(PGW)3314。在至少一个实施例中,归属网络3316包括应用服务器3318和归属用户服务器(HSS)3320。在至少一个实施例中,HSS 3320可以是归属网络3316、EPC 3302和/或其变体的一部分。

[0444] 在至少一个实施例中,MME 3312是网络中的终端点,用于NAS信令的加密/完整性保护并处理安全密钥管理。在至少一个实施例中,应当理解,术语"MME"用在4G LTE网络中,并且5G LTE网络可以包括执行类似功能的安全锚节点(SEAN)或安全访问功能(SEAF)。在至少一个实施例中,术语"MME"、"SEAN"和"SEAF"可以互换使用。在至少一个实施例中,MME 3312还提供用于LTE和2G/3G接入网络之间的移动性的控制平面功能,以及到漫游UE的归属网络的接口。在至少一个实施例中,SGW 3310路由和转发用户数据分组,同时在切换期间还充当用户平面的移动锚点。在至少一个实施例中,PGW 3314通过作为UE业务的出口和入口点来提供从UE到外部分组数据网络的连接性。在至少一个实施例中,HSS 3320是包含用户相关和订阅相关信息的中央数据库。在至少一个实施例中,应用服务器3318是中央数据库,其包含关于可以利用网络架构3300并经由网络架构3300进行通信的各种应用的用户相关信息。

[0445] 图34是示出根据至少一个实施例的根据LTE和5G原理操作的移动电信网络/系统的一些基本功能的图。在至少一个实施例中,移动电信系统包括基础设施设备,该基础设施设备包括连接到核心网络3402的基站3414,核心网络3402根据熟悉通信技术的人将理解的常规布置进行操作。在至少一个实施例中,基础设施设备3414也可以被称为例如基站、网元、增强型节点B(eNodeB)或协调实体,并且为覆盖区域内的一个或更多个通信设备提供无

线接入接口或由虚线3404表示的小区，其可以称为无线接入网。在至少一个实施例中，一个或更多个移动通信设备3406可以使用无线接入接口经由表示数据的信号的发送和接收来传送数据。在至少一个实施例中，核心网络3402还可以为网络实体所服务的通信设备提供包括认证、移动性管理、计费等功能。

[0446] 在至少一个实施例中，图34的移动通信设备还可以被称为通信终端、用户设备(UE)、终端设备等，并且被配置为经由网络实体与由相同或不同覆盖区域服务的一个或更多个其他通信设备进行通信。在至少一个实施例中，这些通信可以通过使用无线接入接口在双向通信链路上发送和接收表示数据的信号来执行。

[0447] 在至少一个实施例中，如图34所示，更详细地示出了eNodeB 3414a 之一包括用于经由无线接入接口向一个或更多个通信设备或UE 3406发送信号的发射器3412，以及用于从覆盖区域3404内的一个或更多个UE接收信号的接收器3410。在至少一个实施例中，控制器3408控制发送器3412 和接收器3410通过无线接入接口发送和接收信号。在至少一个实施例中，控制器3408可以执行控制无线接入接口的通信资源元素的分配的功能，并且在一些示例中可以包括用于调度经由无线接入接口用于上行链路和下行链路的传输的调度器。

[0448] 在至少一个实施例中，示例UE 3406a被更详细地示出为包括用于在无线接入接口的上行链路上向eNodeB 3414发送信号的发射器3420和用于经由无线接入接口在下行链路上接收由eNodeB 3414发送的信号的接收器 3418。在至少一个实施例中，发射器3420和接收器3418由控制器3416 控制。

[0449] 在至少一个实施例中，发射器3420和接收器3418由控制器3416控制，并在诸如图2的基带单元中使用。

[0450] 图35示出了根据至少一个实施例的可以是5G网络架构的一部分的无线电接入网络3500。在至少一个实施例中，无线电接入网络3500覆盖地理区域，该地理区域划分为多个蜂窝区域(小区)，用户设备(UE)基于从一个接入点或者基站在地理区域上广播的标识来唯一地标识这些蜂窝区域(cell)。在至少一个实施例中，宏小区3540、3528和3516以及小小区3530可以包括一个或更多个扇区。在至少一个实施例中，扇区是小区的子区域并且一个小区内的所有扇区都由同一个基站服务。在至少一个实施例中，属于该扇区的单个逻辑标识可以标识扇区内的无线电链路。在至少一个实施例中，小区内的多个扇区可以由天线组形成，每个天线负责与小区的一部分中的UE进行通信。

[0451] 在至少一个实施例中，每个小区由基站(BS)服务。在至少一个实施例中，基站是无线电接入网络中的网络元件，负责在一个或更多个小区中去往或来自UE的无线电传输和接收。在至少一个实施例中，基站还可以称为基站收发台(BTS)、无线电基站、无线电收发机、收发功能、基本服务集 (BSS)、扩展服务集(ESS))、接入点(AP)、节点B(NB)、eNodeB(eNB)、gNodeB(gNB)或一些其他合适的术语。在至少一个实施例中，基站可以包括用于与网络的回程部分进行通信的回程接口。在至少一个实施例中，基站具有集成天线或通过馈电电缆连接到天线或远程无线电头端(RRH)。

[0452] 在至少一个实施例中，回程可以提供基站和核心网络之间的链路，并且在一些示例中，回程可以提供各个基站之间的互连。在至少一个实施例中，核心网络是无线通信系统的一部分，其通常独立于无线电接入网络中使用的无线电接入技术。在至少一个实施例中，可以采用各种类型的回程接口，例如使用任何合适的传输网络的直接物理连接、虚拟网络

等。在至少一个实施例中，一些基站可以被配置为集成接入和回程(IAB)节点，其中无线频谱既可以用于接入链路(即与UE的无线链路)，也可以用于回程链路，这有时是称为无线自回程。在至少一个实施例中，通过无线自回程，用于基站和UE之间通信的无线频谱可以被用于回程通信，从而能够快速和容易地部署高密度小小区网络，而不是需要每个新基站站部署配备自己的硬线回程连接。

[0453] 在至少一个实施例中，高功率基站3536和3520显示在小区3540和 3528中，并且高功率基站3510显示为控制小区3516中的远程无线电头端 (RRH)3512。在至少一个实施例中，小区3540、3528和3516可以被称为大尺寸小区或宏小区。在至少一个实施例中，低功率基站3534示于小型小区3530(例如，微小区、微微小区、毫微微小区、家庭基站、家庭节点B、家庭eNodeB等)中，其可以与一个或更多个宏小区重叠，并且可以被称为小小区或小尺寸小区。在至少一个实施例中，可以根据系统设计以及组件约束来确定小区尺寸。在至少一个实施例中，可以部署中继节点以扩展给定小区的尺寸或覆盖区域。在至少一个实施例中，无线电接入网络3500 可以包括任意数量的无线基站和小区。在至少一个实施例中，基站3536、3520、3510、3534为任意数量的移动装置提供到核心网络的无线接入点。

[0454] 在至少一个实施例中，四轴飞行器或无人机3542可以被配置为用作基站。在至少一个实施例中，小区不一定是静止的，小区的地理区域可以根据移动基站(诸如四轴飞行器3542)的位置而移动。

[0455] 在至少一个实施例中，无线电接入网络3500支持多个移动装置的无线通信。在至少一个实施例中，移动装置通常被称为用户设备(UE)，但也可以被称为移动站(MS)、用户站、移动单元、用户单元、无线单元、远程单元、移动设备、无线设备、无线通信设备、远程设备、移动用户站、接入终端(AT)、移动终端、无线终端、远程终端、手机、终端、用户代理、移动客户端、客户端或其他一些合适的术语。在至少一个实施例中，UE 可以是向用户提供对网络服务的访问的装置。

[0456] 在至少一个实施例中，“移动”装置不必具有移动的能力，并且可以是静止的。在至少一个实施例中，移动装置或移动设备泛指各种不同的设备和技术。在至少一个实施例中，移动装置可以是手机、蜂窝(cell)电话、智能电话、会话发起协议(SIP)电话、膝上型计算机、个人计算机(PC)、笔记本、上网本、智能本、平板电脑、个人数字助理(PDA)、广泛的嵌入式系统，例如对应于“物联网”(IoT)、汽车或其他交通工具、远程传感器或执行器、机器人或机器人设备、卫星无线电、全球定位系统(GPS)设备、对象跟踪设备、无人机、多旋翼飞行器、四旋翼飞行器、遥控设备、消费者和/或可穿戴设备，例如眼镜、可穿戴相机、虚拟现实设备、智能手表、健康或健身追踪器、数字音频播放器(例如MP3播放器)、相机、游戏机、数字家庭或智能家居设备(例如家庭)音频、视频和/或多媒体设备、电器、自动售货机、智能照明、家庭安全系统、智能手机等，安全设备，太阳能电池板或太阳能电池板，控制电力(例如，智能电网)，照明，水等的市政基础设施设备，工业自动化和企业设备，物流控制器，农业设备，军事防御设备、车辆、飞机、船舶和武器等。

[0457] 在至少一个实施例中，移动装置可以提供连接的医学或远程医疗支持，即远距离的医疗保健。在至少一个实施例中，远程医疗设备可以包括远程医疗监控设备和远程医疗管理设备，其通信可以被给予优先处理或优先于其他类型信息的访问，例如，在关键服务数据传输的优先访问方面，和/ 或用于传输关键服务数据的相关QoS。

[0458] 在至少一个实施例中,无线电接入网络3500的小区可以包括可以与每个小区的一个或更多个扇区通信的UE。在至少一个实施例中,UE 3514 和3508可以通过RRH 3512与基站3510通信;UE 3522和3526可以与基站3520通信;UE 3532可以与低功率基站3534通信;UE 3538和3518可以与基站3536通信;UE 3544可以与移动基站3542通信。在至少一个实施例中,每个基站3510、3520、3534、3536和3542可以被配置为为各个小区中的所有UE提供到所有的核心网络(未示出)的接入点以及从基站 (例如,基站3536)到一个或更多个UE(例如,UE 3538和3518)的传输可以称为下行链路(DL)传输,而从UE(例如,UE 3538)到基站可以被称为上行链路(UL)传输。在至少一个实施例中,下行链路可以指点对多点传输,其可以被称为广播信道复用。在至少一个实施例中,上行链路可以指点对点传输。

[0459] 在至少一个实施例中,可以被称为移动网络节点的四轴飞行器3542 可以被配置为通过与基站3536通信来充当小区3540内的UE。在至少一个实施例中,多个UE(例如,UE 3522和3526)可以使用对等(P2P)或旁链路信号3524彼此通信,其可以绕过基站(诸如基站3520)。

[0460] 在至少一个实施例中,UE在移动时与其位置无关地进行通信的能力被称为移动性。在至少一个实施例中,移动性管理实体(MME)建立、维护和释放UE和无线电接入网络之间的各种物理信道。在至少一个实施例中,无线电接入网络3500可以利用基于DL的移动性或基于UL的移动性来实现移动性和切换(即,将UE的连接从一个无线电信道转移到另一无线电信道)。在至少一个实施例中,UE在配置用于基于DL的移动性的网络中可以监视来自其服务小区的信号的各种参数以及相邻小区的各种参数,并且根据这些参数的质量,UE可以保持与一个或更多个相邻小区的通信。在至少一个实施例中,如果在给定的时间量内来自相邻小区的信号质量超过来自服务小区的信号质量,或者如果UE从一个小区移动到另一小区,则 UE可以进行从服务小区到相邻(或目标)小区的切换或切换。在至少一个实施例中,UE 3518(图示为车辆,但可以使用任何合适形式的UE)可以从对应于小区(例如服务小区3540)的地理区域移动到对应于相邻小区 (例如相邻小区3516)的地理区域。在至少一个实施例中,在给定的时间内,当来自相邻小区3516的信号强度或质量超过其服务小区3540的信号强度或质量时,UE 3518可以向其服务基站3536发送报告消息以指示其状况。在至少一个实施例中,UE 3518可以接收切换命令,并且可以经历切换到小区3516。

[0461] 在至少一个实施例中,来自每个UE的UL参考信号可以被配置用于基于UL的移动性的网络用来为每个UE选择服务小区。在至少一个实施例中,基站3536、3520和3510/3512可以广播统一同步信号(例如统一主同步信号(PSS)、统一次同步信号(SSS)和统一物理广播信道(PBCH))。在至少一个实施例中,UE 3538、3518、3522、3526、3514和3508可以接收统一的同步信号,从同步信号中导出载频和时隙定时,并且响应于导出的定时,发送上行链路导频或参考信号。在至少一个实施例中,无线电接入网络3500内的两个或更多个小区(例如,基站3536和3510/3512)可以同时接收由UE(例如,UE 3518)发送的上行链路导频信号。在至少一个实施例中,小区可以测量导频信号的强度,并且无线电接入网络(例如,基站3536 和3510/3512中的一个或更多个和/或核心网络内的中央节点)可以确定UE 3518的服务小区。在至少一个实施例中,当UE 3518通过无线电接入网络 3500移动时,网络可以继续监视由UE 3518发送的上行链路导频信号。在至少一个实施例中,当相邻小区测量的导频信号的信号强度或质量超过服务小区测量的信号强度或质量时,网络3500可以将UE 3518从服务小

区切换给相邻小区，通知或不通知UE 3518。

[0462] 在至少一个实施例中，由基站3536、3520和3510/3512发送的同步信号可以是统一的，但是可以不识别特定小区，而是可以识别以相同频率和/或在相同的时间操作的多个小区的区域。在至少一个实施例中，5G网络或其他下一代通信网络中的区域启用基于上行链路的移动性框架并提高 UE和网络的效率，因为需要在UE和网络之间交换的移动性消息的数量可能被减少。

[0463] 在至少一个实施例中，无线电接入网络3500中的空中接口可以利用未授权频谱、授权频谱或共享频谱。在至少一个实施例中，未授权频谱提供频谱的一部分的共享使用而无需政府授予的许可，然而，虽然通常仍然需要遵守一些技术规则才能访问未授权频谱，通常，任何运营商或设备可能获得访问权限。在至少一个实施例中，授权频谱提供对频谱的一部分的排他使用，通常依靠移动网络运营商从政府监管机构购买许可。在至少一个实施例中，共享频谱可能介于授权频谱和未授权频谱之间，其中可能需要技术规则或限制来访问频谱，但频谱仍可能由多个运营商和/或多个RAT 共享。例如，在至少一个实施例中，授权频谱的一部分的许可的持有者可以提供许可共享接入(LSA)以与其他方共享该频谱，例如，以合适许可确定条件来获得接入。

[0464] 图36提供了根据至少一个实施例的其中使用多种不同类型的设备的 5G移动通信系统的示例图示。在至少一个实施例中，如图36所示，第一基站3618可以被提供给信号传输超过几公里的大型小区或宏小区。然而，在至少一个实施例中，系统还可以支持经由非常小的小区的传输，例如由第二基础设施设备3616传输，第二基础设施设备3616在数百米的距离上发送和接收信号，从而形成所谓的“微微”小区。在至少一个实施例中，第三类基础设施设备3612可以在数十米的距离上发送和接收信号，因此可以用于形成所谓的“毫微微”小区。

[0465] 在至少一个实施例中，也在图36中示出，不同类型的通信设备可以用于经由不同类型的基础设施设备3612、3616、3618来发送和接收信号，并且可以根据使用不同通信参数的不同类型的基础设施设备来适配数据通信。在至少一个实施例中，传统上，移动通信设备可以被配置为经由网络的可用通信资源向和从移动通信网络传送数据。在至少一个实施例中，无线接入系统被配置为向诸如智能电话3606的设备提供最高数据速率。在至少一个实施例中，可以提供“物联网”，其中低功率机器类型通信设备以非常低的功率、低带宽和可能具有低复杂性发送和接收数据。在至少一个实施例中，这种机器类型通信设备3614的示例可以经由微微小区3616进行通信。在至少一个实施例中，非常高的数据速率和低移动性可以是与例如电视3604通信的特征，它可以通过微微(Pico)小区进行通信。在至少一个实施例中，虚拟现实耳机3608可能需要非常高的数据速率和低延迟。在至少一个实施例中，可以部署中继设备3610以扩展给定小区或网络的大小或覆盖区域。

[0466] 图37示出了示例高级系统3700，其中可以使用至少一个实施例。在至少一个实施例中，高级系统3700包括应用程序3702、系统软件+库3704、框架软件3706和数据中心基础设施+资源协调器3708。在至少一个实施例中，高级系统3700可以被实现为云服务、物理服务、虚拟服务、网络服务和/或其变体。

[0467] 在至少一个实施例中，如图37所示，数据中心基础设施+资源协调器 3708可以包括5G无线电资源协调器3710、GPU分组处理和I/O 3712、以及节点计算资源(“节点C.R.”)

3716(1)-3716(N),其中“N”代表任何整数,正整数。在至少一个实施例中,节点CR3716(1)-3716(N)可以包括但不限于任何数量的中央处理单元(“CPU”)或其他处理器(包括加速器、现场可编程门阵列(FPGA)、图形处理器(“GPU”)等)、内存设备(例如,动态只读内存)、存储设备(例如,固态或磁盘驱动器)、网络输入/输出 (“NW I/O”)设备、网络交换机、虚拟机(“VM”)、电源模块和冷却模块等。在至少一个实施例中,节点C.R.3716(1)-3716(N)中的一个或更多个节点CR可以是服务器具有上述一种或更多种计算资源。

[0468] 在至少一个实施例中,5G无线电资源协调器3710可以配置或以其他方式控制一个或更多个节点C.R.3716(1)-3716(N)和/或5G网络架构可以包括的其他各种组件和资源。在至少一个实施例中,5G无线电资源协调器 3710可以包括用于高级系统3700的软件设计基础设施(“SDI”)管理实体。在至少一个实施例中,5G无线电资源协调器3710可以包括硬件、软件或其某种组合。在至少一个实施例中,5G无线电资源协调器3710可用于配置或以其他方式控制各种媒体接入控制子层、无线电接入网络、物理层或子层、和/或其变体,其可为5G网络架构的一部分。在至少一个实施例中, 5G无线电资源协调器3710可以配置或分配分组的计算、网络、存储器或存储资源以支持可以作为5G网络架构的一部分执行的一个或更多个工作负载。

[0469] 在至少一个实施例中,GPU分组处理和I/O 3712可以配置或以其他方式处理各种输入和输出,以及诸如数据分组的分组,其可以作为5G网络架构的一部分被发送/接收,可以由高层系统3700实现。在至少一个实施例中,分组可以是格式化为由网络提供的数据,并且通常可以分为控制信息和有效载荷(即,用户数据)。在至少一个实施例中,数据分组的类型可以包括互联网协议版本4(IPv4)数据分组、互联网协议版本6(IPv6) 数据分组和以太网II帧数据分组。在至少一个实施例中,数据分组的控制数据可以分为数据完整性字段和语义字段。在至少一个实施例中,可以在其上接收数据分组的网络连接包括局域网、广域网、虚拟专用网、因特网、内联网、外联网、公共交换电话网、红外网络、无线网络、卫星网络及其任意组合。

[0470] 在至少一个实施例中,框架软件3706包括AI模型架构+训练+用例 3722。在至少一个实施例中,AI模型架构+训练+用例3722可以包括工具、服务、软件或其他资源以根据一个或更多个实施例使用一个或更多个机器学习模型训练一个或更多个机器学习模型或预测或推理信息。例如,在至少一个实施例中,机器学习模型可以通过使用上文关于高级系统3700描述的软件和计算资源根据神经网络架构计算权重参数来训练。在至少一个实施例中,通过使用通过一种或更多种训练技术计算的权重参数,对应于一个或更多个神经网络的训练的机器学习模型可以用于使用上文关于高级系统3700描述的资源来推理或预测信息。在至少一个实施例中,框架软件 3706可以包括支持系统软件+库3704和应用程序3702的框架。

[0471] 在至少一个实施例中,系统软件+库3704或应用程序3702可分别包括基于网络的服务软件或应用程序,例如由亚马逊网络服务、谷歌云和微软Azure提供的那些。在至少一个实施例中,框架软件3706可以包括但不限于一种类型的免费和开源软件网络应用程序框架,例如Apache Spark™ (以下简称“Spark”)。在至少一个实施例中,系统软件+库3704可以包括由节点C.R.3716(1)-3716(N)的至少部分使用的软件。在至少一个实施例中,一种或更多种类型的软件可以包括但不限于互联网网页搜索软件、电子邮件病毒扫描软件、数据库

软件和流媒体视频内容软件。

[0472] 在至少一个实施例中,PHY3718是一组系统软件和库,被配置为提供与无线技术的物理层的接口,该物理层可以是诸如5G新无线电(NR)物理层的物理层。在至少一个实施例中,NR物理层利用灵活且可扩展的设计并且可以包括各种组件和技术,例如调制方案、波形结构、帧结构、参考信号、多天线传输和信道编码。

[0473] 在至少一个实施例中,NR物理层支持正交相移键控(QPSK)、16正交幅度调制(QAM)、64QAM和256QAM调制格式。在至少一个实施例中, NR物理层中还可以包括针对不同用户实体(UE)类别的不同调制方案。在至少一个实施例中,NR物理层可以在上行链路(UL)和下行链路(DL) 中利用的具有可扩展的数字(子载波间隔,循环前缀)循环前缀正交频分复用(CP-OFDM)高达至少52.6吉赫兹。在至少一个实施例中,NR物理层可以支持UL中的离散傅立叶变换扩展正交频分复用(DFT-SOFDM),用于覆盖受限的场景,具有单流传输(即,没有空间复用)。

[0474] 在至少一个实施例中,NR帧支持时分双工(TDD)和频分双工(FDD) 传输以及在授权和未授权频谱中的操作,这实现了非常低的延迟、快速的混合自动重传请求(HARQ)确认、动态TDD、与LTE共存和可变长度传输 (例如,超可靠低延迟通信(URLLC)的短持续时间和增强型移动宽带 (eMBB)的长持续时间)。在至少一个实施例中,NR帧结构遵循三个关键设计原则以增强前向兼容性并减少不同特征之间的交互。

[0475] 在至少一个实施例中,第一原则是传输是自包含的,这可以指一种方案,其中时隙和波束中的数据可独立解码而不依赖于其他时隙和波束。在至少一个实施例中,这意味着数据解调所需的参考信号被包括在给定的时隙和给定的波束中。在至少一个实施例中,第二原则是传输在时间和频率上被很好地限制,这导致可以引入与传统传输并行的新型传输的方案。在至少一个实施例中,第三原则是避免跨时隙和跨不同传输方向的静态和/ 或严格定时关系。在至少一个实施例中,第三原则的使用可能需要利用异步混合自动重传请求(HARQ)而不是预定义的重传时间。

[0476] 在至少一个实施例中,NR帧结构还允许快速HARQ确认,其中当从 DL接收切换到UL传输时,在DL数据的接收期间执行解码并且由UE在保护时段期间准备HARQ确认。在至少一个实施例中,为了获得低延迟,在时隙(或时隙组)的开始处,时隙(或时隙聚合的情况下的一组时隙) 被前置加载有控制信号和参考信号。

[0477] 在至少一个实施例中,NR具有超精简设计,其最小化始终在线的传输以提高网络能效并确保前向兼容性。在至少一个实施例中,仅在必要时才传输NR中的参考信号。在至少一个实施例中,四个主要参考信号是解调参考信号(DMRS)、相位跟踪参考信号(PTRS)、探测参考信号(SRS)和信道状态信息参考信号(CSI-RS)。

[0478] 在至少一个实施例中,DMRS用于估计用于解调的无线电信道。在至少一个实施例中,DMRS是特定于UE的,可以波束成形,限制在调度资源中,并且仅在必要时在DL和UL中传输。在至少一个实施例中,为了支持多层多输入多输出(MIMO)传输,可以调度多个正交DMRS端口,每层一个。在至少一个实施例中,基本DMRS模式是前置的,因为DMRS 设计考虑了早期解码要求以支持低延迟应用。在至少一个实施例中,对于低速场景,DMRS在时域中使用低密度。然而,在至少一个实施例中,对于高速场景,增加DMRS的时间密度以跟踪无线电信道中的快速变化。

[0479] 在至少一个实施例中，在NR中引入PTRS以实现振荡器相位噪声的补偿。在至少一个实施例中，典型地，相位噪声作为振荡器载波频率的函数而增加。在至少一个实施例中，因此可以在高载波频率(例如毫米波) 下利用PTRS来减轻相位噪声。在至少一个实施例中，PTRS是特定于UE 的，被限制在调度的资源中并且可以被波束成形。在至少一个实施例中，PTRS可根据振荡器的质量、载波频率、OFDM子载波间隔以及用于传输的调制和编码方案来配置。

[0480] 在至少一个实施例中，SRS在UL中传输以执行主要用于调度和链路自适应的信道状态信息(CSI)测量。在至少一个实施例中，对于NR，SRS 还用于大规模MIMO和UL波束管理的基于互易性的预编码器设计。在至少一个实施例中，SRS具有模块化和灵活的设计以支持不同的过程和UE 能力。在至少一个实施例中，信道状态信息参考信号(CSI-RS)的方法是类似的。

[0481] 在至少一个实施例中，NR根据频谱的哪一部分用于其操作而采用不同的天线解决方案和技术。在至少一个实施例中，对于较低频率，假设有低到中等数量的有源天线(多达约32个发射机链)并且FDD操作是常见的。在至少一个实施例中，CSI的获取需要DL中的CSI-RS的传输和UL 中的CSI报告。在至少一个实施例中，在该频率区域中可用的有限带宽需要通过多用户MIMO(MU-MIMO)和更高阶空间复用实现的高频谱效率，与 LTE相比，这是通过更高分辨率的CSI报告实现的。

[0482] 在至少一个实施例中，对于更高的频率，可以在给定孔径中采用更多数量的天线，这增加了波束成形和多用户(MU)-MIMO的能力。在至少一个实施例中，本文，频谱分配是TDD类型并且假设基于互易的操作。在至少一个实施例中，通过UL信道探测获取显式信道估计形式的高分辨率CSI。在至少一个实施例中，这种高分辨率CSI使得能够在基站(BS)处采用复杂的预编码算法。在至少一个实施例中，对于更高的频率(在毫米波范围内)，当前通常需要模拟波束成形实现，这将传输限制为每个时间单位和无线电链的单个波束方向。在至少一个实施例中，由于载波波长短，各向同性天线元件在该频率区域中非常小，因此需要大量天线元件来保持覆盖。在至少一个实施例中，需要在发射器和接收器端应用波束成形以对抗增加的路径损耗，即使对于控制信道传输也是如此。

[0483] 在至少一个实施例中，为了支持这些不同的用例，NR具有高度灵活但统一的CSI框架，其中与LTE相比，在NR中减少了CSI测量、CSI报告和实际DL传输之间的耦合。在至少一个实施例中，NR还支持更高级的方案，例如多点传输和协调。在至少一个实施例中，控制和数据传输遵循自包含原则，其中解码传输所需的所有信息(例如伴随的DMRS)包含在传输本身内。在至少一个实施例中，因此，网络可以随着UE在网络中移动而无缝地改变传输点或波束。

[0484] 在至少一个实施例中，MAC 3720是一组系统软件和库，被配置为提供具有媒体访问控制(MAC)层的接口，其可以是5G网络架构的一部分。在至少一个实施例中，MAC层控制负责与有线、光或无线传输介质交互的硬件。在至少一个实施例中，MAC为传输介质提供流量控制和复用。

[0485] 在至少一个实施例中，MAC子层提供物理层的抽象，使得物理链路控制的复杂性对于逻辑链路控制(LLC)和网络堆栈的上层是不可见的。在至少一个实施例中，任何LLC子层(和更高层)可以与任何MAC一起使用。在至少一个实施例中，任何MAC都可以与任何物理层

一起使用,而与传输介质无关。在至少一个实施例中,MAC子层在向网络上的另一设备发送数据时,将高层帧封装成适合传输介质的帧,添加帧校验序列以识别传输错误,然后在适当的信道访问方法允许时将数据转发到物理层。在至少一个实施例中,如果检测到拥塞信号,MAC还负责补偿冲突,其中MAC可以发起重传。

[0486] 在至少一个实施例中,应用程序3702可以包括由节点 C.R.3716(1)-3716(N)和/或框架软件3706的至少部分使用的一种或更多种类型的应用程序。在至少一个实施例中,一种或更多种类型的应用程序可以包括但不限于任何数量的基因组学应用程序、认知计算和机器学习应用程序,包括训练或推理软件、机器学习框架软件(例如、PyTorch、TensorFlow、Caffe等)或与一个或更多个实施例结合使用的其他机器学习应用程序。

[0487] 在至少一个实施例中,RAN API 3714可以是一组子例程定义、通信协议和/或软件工具,其提供与无线电接入网络(RAN)的组件通信的方法,该组件可以是5G网络架构的一部分。在至少一个实施例中,无线电接入网络是网络通信系统的一部分并且可以实现无线电接入技术。在至少一个实施例中,无线电接入网络功能通常由位于核心网络和用户设备两者中的硅芯片提供。关于无线电接入网络的更多信息可以在图35的描述中找到。

[0488] 在至少一个实施例中,高级系统3700可以使用CPU、专用集成电路 (ASIC)、GPU、FPGA或其他硬件来使用上述资源来执行训练、推理和/或其他各种过程。此外,在至少一个实施例中,上述一个或更多个软件和/ 或硬件资源可以被配置为允许用户训练或执行信息推理的服务,例如图像识别、语音识别或其他人工智能服务,以及其他服务,例如允许用户配置和实施5G网络架构各个方面的服务。

[0489] 在至少一个实施例中,高级系统3700可以使用CPU、专用集成电路 (ASIC)、GPU、FPGA或其他硬件来执行基带单元操作,例如5G NR PHY 信号处理。

[0490] 图38示出了根据至少一个实施例的网络系统3800的架构。在至少一个实施例中,系统3800被示为包括用户设备(UE)3802和UE 3804。在至少一个实施例中,UE 3802和3804被示为智能手机(例如,可连接到一个或更多个蜂窝网络的手持触摸屏移动计算设备),但也可以包括任何移动或非移动计算设备,例如个人数据助理(PDA)、寻呼机、膝上型计算机、台式计算机、无线手机或任何包括无线通信接口的计算设备。

[0491] 在至少一个实施例中,UE 3802和3804中的任一个可以包括物联网 (IoT)UE,其可以包括为利用短暂UE连接的低功率IoT应用而设计的网络接入层。在至少一个实施例中,IoT UE可以利用诸如机器对机器(M2M)或机器类型通信(MTC)之类的技术来通过公共陆地移动网络(PLMN)、基于临近服务(ProSe)或设备到设备(D2D)通信、传感器网络或IoT网络与MTC服务器或设备交换数据。在至少一个实施例中,M2M或MTC数据交换可以是机器发起的数据交换。在至少一个实施例中,IoT网络描述了互连IoT UE,其可以包括具有短期连接的唯一可识别的嵌入式计算设备(在互联网基础设施内)。在至少一个实施例中,IoT UE可以执行后台应用程序(例如,保持活动消息、状态更新等)以促进IoT网络的连接。

[0492] 在至少一个实施例中,UE 3802和3804可以被配置为与无线电接入网络(RAN)3816连接,例如,通信耦合。在至少一个实施例中,RAN 3816 可以是,例如,演进的通用移动电信系统(UMTS)陆地无线电接入网络 (E-UTRAN)、下一代RAN(NGRAN)或某种其他类型的RAN。在至少一个实施例中,UE 3802和3804分别利用连接3812和3814,每个连接包括物理通信接口或层。在至少一个实施例中,连接3812和3814被示为空中接口以实现通信耦合,并且可以与

蜂窝通信协议一致，例如全球移动通信系统(GSM)协议、码分多址(CDMA)网络协议、一键通(PTT)协议、蜂窝上PTT(POC)协议、通用移动电信系统(UMTS)协议、3GPP长期演进(LTE)协议、第五代(5G)协议、新无线电(NR)协议及其变体。

[0493] 在至少一个实施例中，UE 3802和3804可以进一步经由ProSe接口 3806直接交换通信数据。在至少一个实施例中，ProSe接口3806可以替代地被称为包括一个或更多个逻辑信道的侧链接口，包括但不限于物理侧链控制信道(PSCCH)、物理侧链共享信道(PSSCH)、物理侧链发现信道 (PSDCH)和物理侧链广播信道(PSBCH)。

[0494] 在至少一个实施例中，UE 3804被示为经配置以经由连接3808接入接入点(AP)3810。在至少一个实施例中，连接3808可包括本地无线连接，例如与任何IEEE 802.11协议，其中AP 3810将包括无线保真(WiFi®)路由器。在至少一个实施例中，AP 3810被示为连接到因特网而不连接到无线系统的核心网络。

[0495] 在至少一个实施例中，RAN 3816可以包括启用连接3812和3814的一个或更多个接入节点。在至少一个实施例中，这些接入节点(AN)可以被称为基站(BS)、NodeB、演进NodeB(eNB)、下一代NodeB(gNB)、RAN节点等等，并且可以包括提供地理区域(例如，小区)内的覆盖的地面站(例如，地面接入点)或卫星站。在至少一个实施例中，RAN 3816 可以包括一个或更多个用于提供宏小区的RAN节点，例如宏RAN节点 3818，以及一个或更多个用于提供毫微微小区或微微小区的RAN节点(例如，具有较小覆盖区域、较小用户容量、或与宏小区相比更高的带宽)，例如低功率(LP)RAN节点3820。

[0496] 在至少一个实施例中，RAN节点3818和3820中的任一个可以终止空中接口协议并且可以是UE 3802和3804的第一接触点。在至少一个实施例中，RAN节点3818和3820中的任一个可以实现RAN 3816的各种逻辑功能，包括但不限于无线网络控制器(RNC)功能，例如无线承载管理、上行链路和下行链路动态无线资源管理和数据分组调度以及移动性管理。

[0497] 在至少一个实施例中，UE 3802和3804可以被配置为根据各种通信技术使用正交频分复用(“OFDM”)通信信号在多载波通信信道上彼此通信或者与RAN节点3818和3820中的任一个通信，例如但不限于正交频分多址(OFDMA)通信技术(例如，用于下行链路通信)或单载波频分多址(SC-FDMA)通信技术(例如，用于上行链路和ProSe或侧链通信)，和/或其变体。在至少一个实施例中，OFDM信号可以包括多个正交子载波。

[0498] 在至少一个实施例中，下行链路资源网格可以用于从RAN节点3818 和3820中的任一个到UE 3802和3804的下行链路传输，而上行链路传输可以利用类似的技术。在至少一个实施例中，网格可以是时频网格，称为资源网格或时频资源网格，是每个时隙下行链路中的物理资源。在至少一个实施例中，这样的时频平面表示是OFDM系统的常见做法，这使得无线电资源分配直观。在至少一个实施例中，资源网格的每一列和每一行分别对应一个OFDM符号和一个OFDM子载波。在至少一个实施例中，时域中资源网格的持续时间对应于无线电帧中的一个时隙。在至少一个实施例中，资源网格中的最小时频单元被表示为资源元素。在至少一个实施例中，每个资源网格包括多个资源块，其描述了某些物理信道到资源元素的映射。在至少一个实施例中，每个资源块包括资源元素的集合。在至少一个实施例中，在频域中，这可以表示当前可以分配的最小资源量。在至少一个实施例中，存在使用这样的资源块传送的若干不同的物理下行链路信道。

[0499] 在至少一个实施例中，物理下行链路共享信道(PDSCH)可以携带用户数据和到UE

3802和3804的高层信令。在至少一个实施例中，物理下行链路控制信道(PDCCH)可以携带关于与PDSCH信道等相关的传输格式和资源分配的信息。在至少一个实施例中，它还可以将与上行链路共享信道相关的传输格式、资源分配和HARQ(混合自动重传请求)信息通知给UE 3802和3804。在至少一个实施例中，典型地，可以基于从UE 3802和3804 中的任一个反馈的信道质量信息在RAN节点3818和3820中的任一个处执行下行链路调度(向小区内的UE 3802分配控制和共享信道资源块)在至少一个实施例中，可以在用于(例如，分配给)UE 3802和3804中的每一个的PDCCH上发送下行链路资源分配信息。

[0500] 在至少一个实施例中，PDCCH可以使用控制信道元素(CCE)来传达控制信息。在至少一个实施例中，在被映射到资源元素之前，PDCCH复值符号可以首先被组织成四元组，然后可以使用子块交织器对其进行置换以进行速率匹配。在至少一个实施例中，可以使用这些CCE中的一个或更多个来传送每个PDCCH，其中每个CCE可以对应于被称为资源元素组(REG)的九组四个物理资源元素。在至少一个实施例中，四个正交相移键控(QPSK)符号可以映射到每个REG。在至少一个实施例中，取决于下行链路控制信息(DCI)的大小和信道条件，可以使用一个或更多个CCE来发送PDCCH。在至少一个实施例中，LTE中可以定义具有不同数量的 CCE(例如，聚合级别，L＝1、2、4或8)的四种或更多种不同的PDCCH格式。

[0501] 在至少一个实施例中，使用PDSCH资源的增强型物理下行链路控制信道(EPDCCH)可以用于控制信息传输。在至少一个实施例中，可以使用一个或更多个增强控制信道元素(ECCE)来传送EPDCCH。在至少一个实施例中，每个ECCE可以对应于被称为增强资源元素组(EREG)的九组四个物理资源元素。在至少一个实施例中，ECCE在一些情况下可以具有其他数量的EREG。

[0502] 在至少一个实施例中，RAN 3816被示为经由S1接口3822通信地耦合到核心网络(CN)3838。在至少一个实施例中，CN 3838可以是演进分组核心(EPC)网络，NextGen分组核心(NPC)网络或某些其他类型的CN。在至少一个实施例中，S1接口3822被分成两部分：S1-U接口3826，其在 RAN节点3818和3820与服务网关(S-GW)3830之间承载流量数据，以及 S1-移动性管理实体(MME)接口3824，其是RAN节点3818和3820与MME 3828之间的信令接口。

[0503] 在至少一个实施例中，CN 3838包括MME 3828、S-GW 3830、分组数据网络(PDN)网关(P-GW)3834和归属用户服务器(HSS)3832。在至少一个实施例中，MME 3828在功能上可以类似于传统服务通用分组无线电服务(GPRS)支持节点(SGSN)的控制平面。在至少一个实施例中，MME 3828 可以管理接入中的移动性方面，例如网关选择和跟踪区域列表管理。在至少一个实施例中，HSS 3832可以包括用于网络用户的数据库，包括订阅相关信息以支持网络实体对通信会话的处理。在至少一个实施例中，CN 3838 可以包括一个或更多个HSS 3832，这取决于移动用户的数量、设备的容量、网络的组织等。在至少一个实施例中，HSS 3832可以提供支持路由/漫游、身份验证、授权、命名/寻址解析、位置依赖等。

[0504] 在至少一个实施例中，S-GW 3830可以终止通向RAN 3816的S1接口3822，并且在RAN 3816和CN 3838之间路由数据分组。在至少一个实施例中，S-GW 3830可以是本地移动锚RAN间节点切换点，也可以为3GPP 间移动性提供锚点。在至少一个实施例中，其他职责可以包括合法拦截、收费和一些策略执行。

[0505] 在至少一个实施例中，P-GW 3834可以终止朝向PDN的SGi接口。在至少一个实施例中，P-GW 3834可以在EPC网络3838和外部网络(诸如包括应用服务器3840(或者称为应用功

能(AF))之间经由互联网协议(IP)接口3842路由数据分组。在至少一个实施例中,应用服务器3840 可以是提供使用具有核心网络的IP承载资源的应用的元件(例如,UMTS 分组服务(PS)域、LTEPS数据服务等)。在至少一个实施例中,P-GW 3834 被示为经由IP通信接口3842通信地耦合到应用服务器3840。在至少一个实施例中,应用服务器3840还可以被配置为经由CN 3838为UE 3802和 3804提供支持一种或更多种通信服务(例如,网络电话(VoIP)会话、PTT 会话、群组通信会话、社交网络服务等)。

[0506] 在至少一个实施例中,P-GW 3834还可以是用于策略实施和计费数据收集的节点。在至少一个实施例中,策略和计费执行功能(PCRF)3836是 CN 3838的策略和计费控制元件。在至少一个实施例中,在非漫游场景中,家庭公共土地移动网络(HPLMN)中可能存在单个PCRF与UE的互联网协议连接接入网络(IP-CAN)会话相关联。在至少一个实施例中,在具有本地流量中断的漫游场景中,可能存在与UE的IP-CAN会话相关联的两个 PCRF:HPLMN内的归属PCRF(H-PCRF)和访问公共陆地移动网络 (VPLMN)内的拜访PCRF(V-PCRF)。在至少一个实施例中,PCRF 3836 可以通过P-GW 3834通信耦合到应用服务器3840。在至少一个实施例中,应用服务器3840可以用信号通知PCRF 3836以指示新的服务流并选择适当的服务质量(QoS))和充电参数。在至少一个实施例中,PCRF 3836可以将该规则提供到策略和计费执行功能(PCEF)(未示出)中,并具有适当的业务流模板(TFT)和标识符的QoS类别(QCI),其开始QoS和计费由应用程序服务器3840指定。

[0507] 图39示出了根据至少一个实施例的设备3900的示例组件。在至少一个实施例中,设备3900可以包括至少如图所示耦合在一起的应用电路3904、基带电路3908、射频(RF)电路3910、前端模块(FEM)电路3902、一个或更多个天线3912和电源管理电路(PMC)3906。在至少一个实施例中,所示设备3900的组件可以包括在UE或RAN节点中。在至少一个实施例中,设备3900可以包括更少的元件(例如,RAN节点可以不利用应用电路3904,而是包括处理器/控制器来处理从EPC接收的IP数据)。在至少一个实施例中,设备3900可以包括附加元件,例如存储器/存储、显示器、相机、传感器或输入/输出(I/O)接口。在至少一个实施例中,以下描述的组件可以被包括在一个以上的设备中(例如,对于云-RAN(C-RAN)实现,所述电路可以被单独地包括在一个以上的设备中)。

[0508] 在至少一个实施例中,应用电路3904可以包括一个或更多个应用处理器。在至少一个实施例中,应用电路3904可以包括电路,诸如但不限于一个或更多个单核或多核处理器。在至少一个实施例中,处理器可以包括通用处理器和专用处理器(例如,图形处理器、应用处理器等)的任意组合。在至少一个实施例中,处理器可以与存储器/存储器耦合或者可以包括存储器/存储器并且可以被配置为执行存储在存储器/存储器中的指令以使得各种应用或操作系统能够在设备3900上运行。在至少一个实施例中,处理器应用电路3904可以处理从EPC接收的IP数据分组。

[0509] 在至少一个实施例中,基带电路3908可以包括电路,诸如但不限于一个或更多个单核或多核处理器。在至少一个实施例中,基带电路3908 可以包括一个或更多个基带处理器或控制逻辑以处理从RF电路3910的接收信号路径接收的基带信号并且为RF电路3910的发射信号路径生成基带信号。在至少一个实施例中,基带处理电路3908可以与应用电路3904接口以用于生成和处理基带信号以及用于控制RF电路3910的操作。在至少一个实施例中,基带电路3908可以包括第三代(3G)基带处理器3908A、第四代(4G)基带处理器3908B、第

五代(5G)基带处理器3908C或用于其他现有代、正在开发或将被开发的代(例如,第二代(2G),第六代(6G)等)的其他基带处理器3908D。在至少一个实施例中,基带电路3908(例如,基带处理器3908A-D中的一个或更多个)可以处理各种无线电控制功能,这些功能使得能够通过RF电路3910与一个或更多个无线电网络进行通信。在至少一个实施例中,基带处理器3908A-D的一些或全部功能可以包括在存储在存储器3908G中的模块中并且经由中央处理单元(CPU)3908E执行。在至少一个实施例中,无线电控制功能可以包括但不限于信号调制/解调、编码/解码、射频移位等。在至少一个实施例中,基带电路3908的调制/解调电路可以包括快速傅立叶变换(FFT)、预编码或星座映射/解映射功能。在至少一个实施例中,基带电路3908的编码/解码电路可以包括卷积、咬尾卷积、turbo、Viterbi或低密度奇偶校验(LDPC)编码器/解码器功能。

[0510] 在至少一个实施例中,基带电路3908可以包括一个或更多个音频数字信号处理器(DSP)3908F。在至少一个实施例中,音频DSP 3908F可以包括用于压缩/解压缩和回声消除的元件,并且在其他实施例中可以包括其他合适的处理元件。在至少一个实施例中,基带电路的组件可以适当地组合在单个芯片、单个芯片组中,或者在一些实施例中设置在同一电路板上。在至少一个实施例中,基带电路3908和应用电路3904的一些或所有组成组件可以一起实现,例如在片上系统(SOC)上。

[0511] 在至少一个实施例中,基带电路3908可以提供与一种或更多种无线电技术兼容的通信。在至少一个实施例中,基带电路3908可以支持与演进通用陆地无线电接入网(EUTRAN)或其他无线城域网(WMAN)、无线局域网(WLAN)、无线个域网(WPAN)的通信。在至少一个实施例中,基带电路 3908被配置为支持一种以上无线协议的无线电通信并且可以被称为多模基带电路。

[0512] 在至少一个实施例中,RF电路3910可以通过非固体介质使用调制电磁辐射来实现与无线网络的通信。在至少一个实施例中,RF电路3910可以包括交换机、滤波器、放大器等以促进与无线网络的通信。在至少一个实施例中,RF电路3910可以包括接收信号路径,该接收信号路径可以包括对从FEM电路3902接收的RF信号进行下转换并将基带信号提供给基带电路3908的电路。在至少一个实施例中,RF电路3910还可以包括发射信号路径,其可以包括用于对由基带电路3908提供的基带信号进行上转换并将RF输出信号提供给FEM电路3902以用于发射的电路。

[0513] 在至少一个实施例中,RF电路3910的接收信号路径可以包括混频器电路3910a、放大器电路3910b和滤波器电路3910c。在至少一个实施例中, RF电路3910的发射信号路径可以包括滤波器电路3910c和混频器电路 3910a。在至少一个实施例中,RF电路3910还可包括用于合成频率以供接收信号路径和发射信号路径的混频器电路3910a使用的合成器电路3910d。在至少一个实施例中,接收信号路径的混频器电路3910a可以被配置为基于合成器电路3910d提供的合成频率对从FEM电路3902接收的RF信号进行下转换。在至少一个实施例中,放大器电路3910b可以被配置为放大下转换信号并且滤波器电路3910c可以是低通滤波器(LPF)或带通滤波器 (BPF),被配置为从下转换信号中去除不需要的信号产生输出基带信号。在至少一个实施例中,可以将输出基带信号提供给基带电路3908以供进一步处理。在至少一个实施例中,输出基带信号可以是零频基带信号,尽管这不是必需的。在至少一个实施例中,接收信号路径的混频器电路3910a可以包括无源混频器。

[0514] 在至少一个实施例中，发射信号路径的混频器电路3910a可以被配置为基于由合成器电路3910d提供的合成频率对输入基带信号进行上转换以生成用于FEM电路3902的RF输出信号。在一个实施例中，基带信号可以由基带电路3908提供并且可以由滤波器电路3910c滤波。

[0515] 在至少一个实施例中，接收信号路径的混频器电路3910a和发射信号路径的混频器电路3910a可以包括两个或更多个混频器并且可以分别被布置用于正交下转换和上转换。在至少一个实施例中，接收信号路径的混频器电路3910a和发射信号路径的混频器电路3910a可以包括两个或更多个混频器并且可以被布置用于镜像抑制(例如，哈特利图像抑制)。在至少一个实施例中，接收信号路径的混频器电路3910a和混频器电路3910a可以分别布置用于直接下转换和直接上转换。在至少一个实施例中，接收信号路径的混频器电路3910a和发射信号路径的混频器电路3910a可以被配置用于超外差操作。

[0516] 在至少一个实施例中，输出基带信号和输入基带信号可以是模拟基带信号。在至少一个实施例中，输出基带信号和输入基带信号可以是数字基带信号。在至少一个实施例中，RF电路3910可以包括模数转换器(ADC) 和数模转换器(DAC)电路，并且基带电路3908可以包括与RF电路3910 通信的数字基带接口。

[0517] 在至少一个实施例中，可以提供单独的无线电IC电路来处理每个频谱的信号。在至少一个实施例中，合成器电路3910d可以是分数N合成器或分数N/N+1合成器。在至少一个实施例中，合成器电路3910d可以是 delta-sigma合成器、倍频器或包括具有分频器的锁相环的合成器。

[0518] 在至少一个实施例中，合成器电路3910d可以被配置为基于频率输入和分频器控制输入合成供RF电路3910的混频器电路3910a使用的输出频率。在至少一个实施例中，合成器电路3910d可以是分数N/N+1合成器。

[0519] 在至少一个实施例中，频率输入可以由压控振荡器(VCO)提供。在至少一个实施例中，分频器控制输入可以由基带电路3908或应用处理器3904 根据期望的输出频率提供。在至少一个实施例中，分频器控制输入(例如， N)可以基于应用处理器3904所指示的信道从查找表中确定。

[0520] 在至少一个实施例中，RF电路3910的合成器电路3910d可以包括分频器、延迟锁定环(DLL)、多路复用器和相位累加器。在至少一个实施例中，分频器可以是双模分频器(DMD)并且相位累加器可以是数字相位累加器(DPA)。在至少一个实施例中，DMD可以被配置为将输入信号除以N或 N+1(例如，基于进位)以提供小数除法比。在至少一个实施例中，DLL可以包括一组级联的可调延迟元件、相位检测器、电荷泵和D型触发器。在至少一个实施例中，延迟元件可以被配置为将VCO周期分成Nd个相等的相位分组，其中Nd是延迟线中的延迟元件的数量。在至少一个实施例中，以此方式，DLL提供负反馈以帮助确保通过延迟线的总延迟为一个VCO 周期。

[0521] 在至少一个实施例中，合成器电路3910d可以被配置为生成载波频率作为输出频率，而在其他实施例中，输出频率可以是载波频率的倍数(例如，载波频率的两倍，四倍载波频率)并与正交发生器和分频器电路结合使用，以在载波频率上生成多个信号，这些信号彼此之间具有多个不同的相位。在至少一个实施例中，输出频率可以是LO频率(fLO)。在至少一个实施例中，RF电路3910可以包括IQ/极性转换器。

[0522] 在至少一个实施例中,FEM电路3902可以包括接收信号路径,该接收信号路径可以包括被配置为对从一个或更多个天线3912接收的RF信号进行操作、放大接收信号并且将接收信号的放大版本提供给RF电路3910 的电路作进一步处理。在至少一个实施例中,FEM电路3902还可以包括发射信号路径,该发射信号路径可以包括被配置为放大由RF电路3910提供的用于发射的信号以便由一个或更多个天线3912中的一个或更多个发射的电路。在至少一个实施例中,通过发射或接收信号路径的放大可以单独在RF电路3910中、单独在FEM 3902中或在RF电路3910和FEM 3902 两者中完成。

[0523] 在至少一个实施例中,FEM电路3902可以包括TX/RX交换机以在发射模式和接收模式操作之间切换。在至少一个实施例中,FEM电路可以包括接收信号路径和发射信号路径。在至少一个实施例中,FEM电路的接收信号路径可以包括LNA以放大接收的RF信号并且提供放大的接收的 RF信号作为输出(例如,到RF电路3910)。在至少一个实施例中,FEM电路3902的发射信号路径可以包括功率放大器(PA)以放大输入RF信号(例如,由RF电路3910提供),以及一个或更多个滤波器以生成用于后续发射的RF信号(例如,通过一个或更多个天线3912中的一个或更多个)。

[0524] 在至少一个实施例中,PMC 3906可以管理提供给基带电路3908的功率。在至少一个实施例中,PMC 3906可以控制电源选择、电压缩放、电池充电或DC-DC转换。在至少一个实施例中,当设备3900能够由电池供电时,例如,当设备被包括在UE中时,可以经常包括PMC 3906。在至少一个实施例中,PMC 3906可以提高功率转换效率同时提供期望的实现尺寸和散热特性。

[0525] 在至少一个实施例中,PMC 3906可以附加地或替代地与其他组件(例如但不限于应用电路3904、RF电路3910或FEM 3902)耦合,并为其执行类似的功率管理操作。

[0526] 在至少一个实施例中,PMC 3906可以控制设备3900的各种省电机制或以其他方式为其一部分。在至少一个实施例中,如果设备3900处于 RRC连接状态,则它仍然连接到RAN节点预计很快就会接收流量,然后它可能会在一段时间不活动后进入称为不连续接收模式(DRX)的状态。在至少一个实施例中,在该状态期间,设备3900可以在短暂的时间间隔断电,从而节省电力。

[0527] 在至少一个实施例中,如果在延长的时间段内没有数据业务活动,则设备3900可以转换到RRC空闲状态,在那里它与网络断开连接并且不执行操作(诸如信道质量反馈、切换等)。在至少一个实施例中,设备3900 进入非常低的功率状态并且它执行寻呼,其中它再次周期性地唤醒以收听网络,然后再次断电。在至少一个实施例中,设备3900在该状态下可能不接收数据,为了接收数据,它必须转换回RRC连接状态。

[0528] 在至少一个实施例中,附加的省电模式可以允许设备在比寻呼间隔长的时间段内(范围从几秒到几小时)对网络不可用。在至少一个实施例中,在此期间,设备完全无法接入网络并且可能完全断电。在至少一个实施例中,在此期间发送的任何数据都会导致大延迟,并且假设延迟是可以接受的。

[0529] 在至少一个实施例中,应用电路3904的处理器和基带电路3908的处理器可用于执行协议栈的一个或更多个实例的元素。在至少一个实施例中,基带电路3908的处理器可以单独或组合地用于执行第3层、第2层或第1 层功能,而应用电路3908的处理器可以利用从这些接收到的数据(例如,分组数据)层并进一步执行第4层功能(例如,传输通信协议(TCP)

和用户数据报协议(UDP)层)。在至少一个实施例中,第3层可以包括无线电资源控制(RRC)层。在至少一个实施例中,第2层可以包括媒体访问控制(MAC) 层、无线电链路控制(RLC)层和分组数据会聚协议(PDCP)层。在至少一个实施例中,第1层可以包括UE/RAN节点的物理(PHY)层。

[0530] 图40示出了根据至少一个实施例的基带电路的示例接口。在至少一个实施例中,如上所述,图39的基带电路3908可以包括处理器 3908A-3908E和由所述处理器使用的存储器3908G。在至少一个实施例中,处理器3908A-3908E中的每一个可以分别包括存储器接口4002A-4002E,以向/从存储器3908G发送/接收数据。

[0531] 在至少一个实施例中,基带电路3908还可以包括一个或更多个接口以通信地耦合到其他电路/设备,例如存储器接口4004(例如,向/从基带电路3908外部存储器发送/接收数据的接口)、应用电路接口4006(例如,向/从图39的应用电路3904发送/接收数据的接口)、RF电路接口4008(例如,向/从图39的RF电路3910发送/接收数据的接口)、无线硬件连接接口4010(例如,向/从近场通信(NFC)组件、蓝牙®组件(例如Bluetooth® Low Energy)、Wi-Fi®组件和其他通信组件),以及电源管理接口4012(例如,向/从PMC 3906发送/接收电源或控制信号的接口。

[0532] 图41示出了根据至少一个实施例的上行链路信道的示例。在至少一个实施例中,图41示出了在5G NR中的物理上行链路共享信道(PUSCH) 内发送和接收数据,其可以是移动设备网络的物理层的一部分。

[0533] 在至少一个实施例中,5G NR中的物理上行链路共享信道(PUSCH) 被指定为承载复用控制信息和用户应用数据。在至少一个实施例中,5G NR 与其前身相比提供了更多的灵活性和可靠性,在一些示例中可以称为4G LTE,包括更具弹性的导频布置和对循环前缀(CP)-OFDM和离散傅立叶变换扩展(DFT-s)-OFDM波形的支持。在至少一个实施例中,标准引入的滤波OFDM(f-OFDM)技术被用来添加额外的滤波以减少带外发射并提高更高调制阶的性能。在至少一个实施例中,前向纠错(FEC)中的修改被强加于用准循环低密度奇偶校验(QC-LDPC)码替换4G LTE中使用的Turbo码,这被证明可以实现更好的传输速率并提供机会以获得更高效的硬件实现。

[0534] 在至少一个实施例中,5G NR下行链路和上行链路数据的传输被组织成持续时间为10毫秒的帧,每个帧被分成10个子帧,每个子帧为1毫秒。在至少一个实施例中,子帧由可变数量的时隙组成,这取决于在5G NR 中参数化的所选子载波间隔。在至少一个实施例中,时隙是由14个OFDMA 符号构建的,每个符号都带有循环前缀。在至少一个实施例中,位于通带内并被指定用于传输的子载波被称为资源元素(RE)。在至少一个实施例中,同一符号中的一组12个相邻RE形成物理资源块(PRB)。

[0535] 在至少一个实施例中,5G NR标准定义了与PUSCH信道内的传输相关联的两种类型的参考信号。在至少一个实施例中,解调参考信号(DMRS) 是具有高频率密度的用户特定参考信号。在至少一个实施例中,DMRS仅在专用正交频分多址(OFDMA)符号内传输并且被指定用于频率选择性信道估计。在至少一个实施例中,根据配置,一个时隙内的DMRS符号数量可以在1和4之间变化,其中更密集的DMRS符号时间间隔被指定用于快速时变信道,以在信道的相干时间内获得更准确的估计。在至少一个实施例中,在频域中,DMRS PRB被映射在整个传输分配内。在至少一个实施例中,为同一个天线端口(AP)分配的DMRS资源元素(RE)之间

的间距可以在2和3之间选择。在至少一个实施例中,在2-2多输入多输出(MIMO) 的情况下,标准允许在AP之间正交分配RE。在至少一个实施例中,接收器可以在MIMO均衡之前基于DMRS RE执行部分单输入多输出(SIMO) 信道估计,忽略空间相关。

[0536] 在至少一个实施例中,第二类型的参考信号是相位跟踪参考信号 (PTRS)。在至少一个实施例中,PTRS子载波以在时域中具有高密度的梳状结构排列。在至少一个实施例中,它主要用于毫米波频段以跟踪和校正相位噪声,这是性能损失的一个重要来源。在至少一个实施例中,PTRS 的使用是可选的,因为当相位噪声的影响可以忽略时,它可能会降低传输的总频谱效率。

[0537] 在至少一个实施例中,对于数据的传输,可以从MAC层生成传输块并且将其提供给物理层。在至少一个实施例中,传输块可以是意图被传输的数据。在至少一个实施例中,物理层中的传输以分组的资源数据开始,其可以被称为传输块。在至少一个实施例中,传输块由循环冗余校验 (CRC)4102接收。在至少一个实施例中,循环冗余校验被附加到每个传输块以用于错误检测。在至少一个实施例中,循环冗余校验用于传输块中的错误检测。在至少一个实施例中,使用整个传输块来计算CRC奇偶校验位,然后将这些奇偶校验位附加到传输块的末尾。在至少一个实施例中,最小和最大代码块大小被指定以便块大小与进一步的处理兼容。在至少一个实施例中,当输入块大于最大代码块大小时,输入块被分段。

[0538] 在至少一个实施例中,传输块由低密度奇偶校验(LDPC)编码4104接收和编码。在至少一个实施例中,NR采用低密度奇偶校验(LDPC)码用于数据信道和控制信道的极化码。在至少一个实施例中,LDPC码由它们的奇偶校验矩阵定义,每一列代表一个编码位,每一行代表一个奇偶校验方程。在至少一个实施例中,通过以迭代方式在变量和奇偶校验之间交换消息来解码LDPC码。在至少一个实施例中,为NR提出的LDPC码使用准循环结构,其中奇偶校验矩阵由较小的基矩阵定义。在至少一个实施例中,基矩阵的每个条目表示ZxZ零矩阵或移位的ZxZ单位矩阵。

[0539] 在至少一个实施例中,编码传输块由速率匹配4106接收。在至少一个实施例中,编码块用于创建具有期望码率的输出位流。在至少一个实施例中,速率匹配4106用于创建要以期望的码率传输的输出位流。在至少一个实施例中,从缓冲器中选择和修剪位以创建具有期望码率的输出位流。在至少一个实施例中,结合了混合自动重复请求(HARQ)纠错方案。

[0540] 在至少一个实施例中,在加扰4108中,输出位被加扰,这可以有助于隐私。在至少一个实施例中,码字与正交序列和UE特定加扰序列逐位相乘。在至少一个实施例中,加扰4108的输出可以被输入到调制/映射/预编码和其他过程4110中。在至少一个实施例中,执行各种调制、映射和预编码过程。

[0541] 在至少一个实施例中,从加扰4108输出的位用调制方案进行调制,从而产生调制符号块。在至少一个实施例中,加扰码字使用调制方案QPSK、16QAM、64QAM之一进行调制,从而产生调制符号块。在至少一个实施例中,可以利用信道交织器过程来实现调制符号到发射波形的第一时间映射,同时确保HARQ信息存在于两个时隙上。在至少一个实施例中,调制符号基于发射天线被映射到各个层。在至少一个实施例中,符号可以被预编码,其中它们被分成组,并且可以执行快速傅立叶逆变换。在至少一个实施例中,可以执行传输数据和控制复用,使得HARQ确认(ACK)信息存在于两个时隙中并且被映射到解调参考信号周围的资源。在至少一个实施例中,执行各种预编码过程。

[0542] 在至少一个实施例中，符号被映射到资源元素映射4112中分配的物理资源元素。在至少一个实施例中，分配大小可以被限制为质因数为2、3 和5的值。在至少一个实施例中，符号以从副载波开始的递增顺序进行映射。在至少一个实施例中，子载波映射的调制符号数据是通过OFDMA调制4114中的IFFT操作进行正交频分多址(OFDMA)调制的。在至少一个实施例中，每个符号的时域表示使用发射FIR滤波器进行连接和滤波，以衰减由于相位不连续性和使用不同数字学而导致的相邻频带的不需要的带外发射。在至少一个实施例中，OFDMA调制4114的输出可以被传送以被另一系统接收和处理。

[0543] 在至少一个实施例中，传输可以由OFDMA解调4116接收。在至少一个实施例中，传输可以通过蜂窝网络从用户移动设备发起，但是可能存在其他情况。在至少一个实施例中，可以通过IFFT处理来解调传输。在至少一个实施例中，一旦完成通过IFFT处理的OFDMA解调，就可以执行残余采样时间偏移(STO)和载波频率偏移(CFO)的估计和校正。在至少一个实施例中，CFO和STO校正都必须在频域中执行，因为接收信号可以是来自在频率上复用的多个UE的传输的叠加，每个UE遭受特定的残余同步错误。在至少一个实施例中，残余CFO被估计为属于不同OFDM符号的导频子载波之间的相位旋转并且通过频域中的循环卷积操作来校正。

[0544] 在至少一个实施例中，OFDMA解调4116的输出可以由资源元素解映射4118接收。在至少一个实施例中，资源元素解映射4118可以从分配的物理资源元素确定符号和解映射符号。在至少一个实施例中，在信道估计4120中执行信道估计和均衡以补偿多径传播的影响。在至少一个实施例中，可以利用信道估计4120来最小化源自各种传输层和天线的噪声的影响。在至少一个实施例中，信道估计4120可以从资源元素解映射4118的输出生成均衡符号。在至少一个实施例中，解调/解映射4122可以从信道估计 4120接收均衡符号。在至少一个实施例中，均衡符号是通过层解映射操作进行解映射和置换。在至少一个实施例中，最大后验概率(MAP)解调方法可用于产生表示关于接收位为0或1的置信度的值，以对数似然比(LLR) 的形式表达。

[0545] 在至少一个实施例中，软解调位使用各种操作来处理，包括在LDPC 解码之前使用循环缓冲器解扰、解交织和与LLR软组合的速率不匹配。在至少一个实施例中，解扰4124可以涉及反转加扰4108的一个或更多个过程的过程。在至少一个实施例中，速率不匹配4126可以涉及反转速率匹配 4106的一个或更多个过程的过程。在至少一个实施例中，解扰器4124可以接收来自解调/解映射4122的输出，并对接收到的位进行解扰。在至少一个实施例中，速率不匹配4126可以接收解扰位，并且在LDPC解码4128 之前利用循环缓冲器利用LLR软组合。

[0546] 在至少一个实施例中，在实际应用中对LDPC码的解码是基于迭代置信传播算法来完成的。在至少一个实施例中，LDPC码可以以二部图的形式表示，其中大小为M×N的奇偶校验矩阵H是定义图节点之间的连接的双邻接矩阵。在至少一个实施例中，矩阵H的M行对应于奇偶校验节点，而N列对应于变量节点，即接收的码字位。在至少一个实施例中，置信传播算法的原理基于迭代消息交换，其中更新变量和校验节点之间的后验概率，直到获得有效码字。在至少一个实施例中，LDPC解码4128可以输出包括数据的传输块。

[0547] 在至少一个实施例中，CRC校验4130可以基于附加到接收的传输块的奇偶校验位来确定错误并执行一个或更多个动作。在至少一个实施例中， CRC校验4130可以分析和处

理附加到接收的传输块的奇偶校验位，或者与CRC相关联的任何信息。在至少一个实施例中，CRC校验4130可以将处理后的传输块发送到MAC层以供进一步处理。

[0548] 应当注意，在各种实施例中，可以是传输块或其其他变体的发送和接收数据可以包括图41中未描绘的各种过程。在至少一个实施例中，图41 中描绘的过程并非旨在是详尽无遗的，并且进一步处理(诸如附加调制、映射、复用、预编码、星座映射/解映射、MIMO检测、检测、解码及其变体)可以在作为网络的一部分发送和接收数据中利用。

[0549] 图42示出了根据一些实施例的网络的系统4200的架构。在至少一个实施例中，系统4200被示为包括UE 4202、5G接入节点或RAN节点(示为(R)AN节点4208)、用户平面功能(示为UPF 4204)、数据网络(DN 4206)，例如可以是运营商服务、互联网接入或第3方服务，以及5G核心网(5GC) (显示为CN 4210)。

[0550] 在至少一个实施例中，CN 4210包括认证服务器功能(AUSF 4214)；核心访问和移动性管理功能(AMF 4212)；会话管理功能(SMF 4218)；网络暴露功能(NEF 4216)；策略控制功能(PCF 4222)；网络功能(NF)存储库功能(NRF 4220)；统一数据管理(UDM 4224)；以及应用程序功能(AF 4226)。在至少一个实施例中，CN 4210还可以包括未示出的其他元素，例如结构化数据存储网络功能(SDSF)、非结构化数据存储网络功能(UDSF) 及其变体。

[0551] 在至少一个实施例中，UPF 4204可以充当RAT内和RAT间移动性的锚点、与DN 4206互连的外部PDU会话点以及支持多宿主PDU会话的分支点。在至少一个实施例中，UPF 4204还可以执行分组路由和转发、分组检查、执行策略规则的用户平面部分、合法拦截分组(UP收集)；流量使用报告，为用户平面执行QoS处理(例如数据分组过滤、门控、UL/DL 速率执行)、执行上行链路流量验证(例如，SDF到QoS流映射)、上行链路和下行链路中的传输级数据分组标记以及下行链路数据分组缓冲和下行数据通知触发。在至少一个实施例中，UPF 4204可以包括上行链路分类器以支持将业务流路由到数据网络。在至少一个实施例中，DN 4206可以代表各种网络运营商服务、因特网接入或第三方服务。

[0552] 在至少一个实施例中，AUSF 4214可以存储用于UE 4202的认证的数据并且处理认证相关的功能。在至少一个实施例中，AUSF 4214可以促进用于各种访问类型的通用认证框架。

[0553] 在至少一个实施例中，AMF 4212可以负责注册管理(例如，用于注册UE 4202等)、连接管理、可达性管理、移动性管理和AMF相关事件的合法拦截，以及接入认证和授权。在至少一个实施例中，AMF 4212可以为SMF 4218的SM消息提供传输，并且充当用于路由SM消息的透明代理。在至少一个实施例中，AMF 4212还可以为UE 4202和SMS功能 (SMSF)(图42未示出)之间的短消息服务(SMS)消息提供传输。在至少一个实施例中，AMF 4212可以充当安全锚定功能(SEA)，其可以包括与AUSF 4214和UE 4202的交互以及作为UE 4202认证过程的结果而建立的中间密钥的接收。在使用基于USIM的认证的至少一个实施例中，AMF 4212可以从AUSF 4214检索安全材料。在至少一个实施例中，AMF 4212还可以包括安全上下文管理(SCM)功能，其从SEA接收密钥，它用于导出访问网络特定的密钥。此外，在至少一个实施例中，AMF 4212可以是RANCP 接口的终结点(N2参考点)、NAS(NI)信令的终结点，并执行NAS加密和完整性保护。

[0554] 在至少一个实施例中，AMF 4212还可以支持通过N3互通功能(IWF) 接口与UE 4202的NAS信令。在至少一个实施例中，N3IWF可用于提供对不可信实体的访问。在至少一个

实施例中,N3IWF可以分别是用于控制平面和用户平面的N2和N3接口的终止点,并且因此可以处理来自SMF 和AMF的用于PDU会话和QoS的N2信令,用于IPSec和N3隧道的封装/解封装分组,在上行链路中标记N3用户平面数据分组,并考虑与通过 N2接收的此类标记相关的QoS要求,强制执行与N3数据分组标记相对应的QoS。在至少一个实施例中,N3IWF还可以在UE 4202和AMF 4212之间中继上行链路和下行链路控制平面NAS(NI)信令,并且在UE 4202和 UPF 4204之间中继上行链路和下行链路用户平面分组。在至少一个实施例中,N3IWF还提供了与UE 4202建立IPsec隧道的机制。

[0555] 在至少一个实施例中,SMF 4218可以负责会话管理(例如,会话建立、修改和释放,包括UPF和AN节点之间的隧道维护);UEIP地址分配和管理(包括可选的授权);UP功能的选择与控制;在UPF配置流量转向以将流量路由到正确的目的地;终止对策略控制功能的接口;控制部分策略执行和QoS;合法拦截(用于SM事件和LI系统接口);终止NAS 消息的SM部分;下行数据通知;AN特定SM信息的发起者,通过N2上的AMF发送到AN;确定会话的SSC模式。在至少一个实施例中,SMF 4218 可以包括以下漫游功能:处理本地实施以应用QoSSLAB(VPLMN);计费数据采集和计费接口(VPLMN);合法拦截(在VPLMN中用于SM事件和与LI系统的接口);支持与外部DN交互,以传输外部DN的PDU会话授权/认证信令。

[0556] 在至少一个实施例中,NEF 4216可以提供用于安全地暴露由3GPP 网络功能为第三方提供的服务和能力、内部暴露/再暴露、应用功能(例如, AF 4226)、边缘计算或雾计算系统等的手段。在至少一个实施例中,NEF 4216可以认证、授权和/或节流AF。在至少一个实施例中,NEF 4216还可以翻译与AF 4226交换的信息和与内部网络功能交换的信息。在至少一个实施例中,NEF 4216可以在AF-服务-标识符和内部5GC信息之间进行转换。在至少一个实施例中,NEF 4216还可以基于其他网络功能的暴露能力从其他网络功能(NF)接收信息。在至少一个实施例中,该信息可以作为结构化数据存储在NEF 4216中,或者使用标准化接口存储在数据存储NF中。在至少一个实施例中,存储的信息然后可以被NEF 4216重新暴露给其他 NF和AF,和/或用于其他目的,例如分析。

[0557] 在至少一个实施例中,NRF 4220可以支持服务发现功能,从NF实例接收NF发现请求,并且将发现的NF实例的信息提供给NF实例。在至少一个实施例中,NRF 4220还维护可用NF实例及其支持的服务的信息。

[0558] 在至少一个实施例中,PCF 4222可以向控制平面功能提供策略规则以实施它们,并且还可以支持统一的策略框架来管理网络行为。在至少一个实施例中,PCF 4222还可以实现前端(FE)以访问与UDM 4224的UDR 中的策略决定相关的订阅信息。

[0559] 在至少一个实施例中,UDM 4224可以处理订阅相关信息以支持网络实体对通信会话的处理,并且可以存储UE 4202的订阅数据。在至少一个实施例中,UDM 4224可以包括两个部分,应用程序FE和用户数据存储库 (UDR)。在至少一个实施例中,UDM可以包括UDMFE,负责凭证的处理、位置管理、订阅管理等。在至少一个实施例中,几个不同的前端可以在不同的交易中为同一用户服务。在至少一个实施例中,UDM-FE访问存储在 UDR中的订阅信息并进行认证凭证处理;用户身份处理;访问授权;注册 /移动管理;以及订阅管理。在至少一个实施例中,UDR可以与PCF 4222 交互。在至少一个实施例中,UDM 4224还可以支持SMS管理,其中 SMS-FE实现与之前讨论的类似的应用逻辑。

[0560] 在至少一个实施例中,AF 4226可以提供对流量路由、对网络能力暴露(NCE)的访

问的应用影响，并且与用于策略控制的策略框架交互。在至少一个实施例中，NCE可以是一种允许5GC和AF 4226通过NEF 4216相互提供信息的机制，其可以用于边缘计算实现。在至少一个实施例中，网络运营商和第三方服务可以托管在UE 4202接入点附近，以通过减少端到端延迟和传输网络上的负载来实现高效的服务交付。在至少一个实施例中，对于边缘计算实现，5GC可以选择靠近UE 4202的UPF 4204并通过N6 接口执行从UPF 4204到DN 4206的流量导向。在至少一个实施例中，这可以基于UE订阅数据、UE位置和AF 4226提供的信息。在至少一个实施例中，AF 4226可以影响UPF(重)选择和业务路由。在至少一个实施例中，基于运营商部署，当AF 4226被认为是可信实体时，网络运营商可以允许AF 4226直接与相关NF交互。

[0561] 在至少一个实施例中，CN 4210可以包括SMSF，其可以负责SMS 订阅检查和验证，以及向/从UE 4202向/从其他实体中继SM消息，例如 SMS-GMSC/IWMSC/SMS-路由器。在至少一个实施例中，SMS还可以与 AMF 4212和UDM 4224交互以用于通知过程UE 4202可用于SMS传输(例如，设置UE不可达标志，并在UE 4202可用于SMS时通知UDM 4224)。

[0562] 在至少一个实施例中，系统4200可以包括以下基于服务的接口：Namf：AMF展示的基于服务的接口；Nsmf：SMF展示的基于服务的接口；Nnef：NEF展示的基于服务的接口；Npcf：PCF展示的基于服务的接口；Nudm：UDM展示的基于服务的接口；Naf：AF展示的基于服务的接口；Nnrf：NRF展示的基于服务的接口；以及Nausf：AUSF展示的基于服务的接口。

[0563] 在至少一个实施例中，系统4200可以包括以下参考点：N1：UE和 AMF之间的参考点；N2：(R)AN和AMF之间的参考点；N3：(R)AN和UPF之间的参考点；N 4：SMF和UPF之间的参考点；以及N6：UPF和数据网络之间的参考点。在至少一个实施例中，在NF中的NF服务之间可以有更多的参考点和/或基于服务的接口，然而，为了清楚起见省略了这些接口和参考点。在至少一个实施例中，NS参考点可以在PCF和AF之间；N7参考点可能在PCF和SMF之间；N11参考点在AMF和SMF之间；等。在至少一个实施例中，CN 4210可以包括Nx接口，其是MME和AMF 4212 之间的CN间接口，以实现CN 4210和CN7242之间的互通。

[0564] 在至少一个实施例中，系统4200可以包括多个RAN节点(例如(R)AN 节点4208)，其中Xn接口被定义在连接到5GC 410的两个或更多个(R)AN 节点4208(例如gNB)之间，在连接到CN 4210和eNB(例如，宏RAN节点)的(R)AN节点4208(例如，gNB)之间，和/或在连接到CN 4210的两个eNB之间。

[0565] 在至少一个实施例中，Xn接口可以包括Xn用户平面(Xn-U)接口和 Xn控制平面(Xn-C)接口。在至少一个实施例中，Xn-U可以提供用户平面 PDU的无保证递送并且支持/提供数据转发和流量控制功能。在至少一个实施例中，Xn-C可以提供管理和错误处理功能、管理Xn-C接口的功能；UE 4202在连接模式(例如，CM-CONNECTED)中的移动性支持，包括管理一个或更多个(R)AN节点4208之间的连接模式的UE移动性的功能。在至少一个实施例中，移动性支持可以包括来自旧(源)服务(R)AN节点4208到新(目标)服务(R)AN节点4208的上下文转移；以及控制旧 (源)服务(R)AN节点4208到新(目标)服务(R)AN节点4208之间的用户平面隧道。

[0566] 在至少一个实施例中，Xn-U的协议栈可以包括建立在互联网协议(IP) 传输层上的传输网络层，以及在UDP和/或IP层之上的GTP-U层，用于承载用户平面PDU。在至少一个实施例中，Xn-C协议栈可以包括应用层信令协议(称为Xn应用协议(Xn-AP))和建立在SCTP层

上的传输网络层。在至少一个实施例中，SCTP层可以在IP层之上。在至少一个实施例中，SCTP层提供应用层消息的有保证的传递。在至少一个实施例中，在传输 IP层中使用点对点传输来传递信令PDU。在至少一个实施例中，Xn-U协议栈和/或Xn-C协议栈可以与在本文示出和描述的用户平面和/或控制平面协议栈相同或相似。

[0567] 图43是根据一些实施例的控制平面协议栈的图示。在至少一个实施例中，控制平面4300被示为UE 3802(或备选地，UE 3804)、RAN 3816和 MME 3828之间的通信协议栈。

[0568] 在至少一个实施例中，PHY层4302可以通过一个或更多个空中接口发送或接收MAC层4304使用的信息。在至少一个实施例中，PHY层4302 可以进一步执行链路自适应或自适应调制和编码(AMC)、功率控制、小区搜索(例如，用于初始同步和切换目的)以及由更高层使用的其他测量，例如RRC层4310。在至少一个实施例中，PHY层4302还可以进一步对传输信道执行错误检测、传输信道的前向纠错(FEC)编码/解码、物理信道的调制/解调、交织、速率匹配、映射到物理信道，以及多输入多输出(MIMO) 天线处理。

[0569] 在至少一个实施例中，MAC层4304可以执行逻辑信道和传输信道之间的映射，将来自一个或更多个逻辑信道的MAC服务数据单元(SDU)复用到传输块(TB)上以经由传输信道传输到PHY，将MAC SDU从通过传输信道从PHY传送的传输块(TB)解复用到一个或更多个逻辑信道，将MAC SDU复用到TB，调度信息报告，通过混合自动重复请求(HARD)进行纠错，以及逻辑信道优先排序。

[0570] 在至少一个实施例中，RLC层4306可以在多种操作模式下操作，包括：透明模式(TM)、未确认模式(UM)和确认模式(AM)。在至少一个实施例中，RLC层4306可以执行上层协议数据单元(PDU)的传输、通过用于 AM数据传输的自动重复请求(ARQ)的纠错以及用于UM和AM数据的 RLC SDU的级联、分段和重组转让。在至少一个实施例中，RLC层4306 还可以执行用于AM数据传输的RLC数据PDU的重新分段、为UM和 AM数据传输重新排序RLC数据PDU、检测用于UM和AM数据传输的重复数据、丢弃用于UM和AM数据传输的RLC SDU，检测AM数据传输的协议错误，并执行RLC重建。

[0571] 在至少一个实施例中，PDCP层4308可以执行IP数据的报头压缩和解压缩、维护PDCP序列号(SN)、在重建下层时执行上层PDU的顺序传送、消除重复重新建立映射在RLCAM上的无线电承载的低层SDU，加密和解密控制平面数据，执行控制平面数据的完整性保护和完整性验证，控制基于定时器的数据丢弃，并执行安全操作(例如、加密、解密、完整性保护、完整性验证等)。

[0572] 在至少一个实施例中，RRC层4310的主要服务和功能可以包括系统信息的广播(例如，包括在与非接入层(NAS)相关的主信息块(MIB)或系统信息块(SIB)中)、与接入层(AS)相关的系统信息的广播、UE和 E-UTRAN之间RRC连接的寻呼、建立、维护和释放(例如，RRC连接寻呼、RRC连接建立、RRC连接修改，和RRC连接释放)，点对点无线承载的建立、配置、维护和释放，安全功能包括密钥管理、无线接入技术(RAT) 间移动性和用于UE测量报告的测量配置。在至少一个实施例中，所述MIB 和SIB可以包括一个或更多个信息元素(IE)，其每个可以包括单独的数据字段或数据结构。

[0573] 在至少一个实施例中，UE 3802和RAN 3816可以利用Uu接口(例如，LTE-Uu接口)经由包括PHY层4302、MAC层4304、RLC层4306、 PDCP层4308和RRC层4310的协议栈交换控制平面数据。

[0574] 在至少一个实施例中，非接入层(NAS)协议(NAS协议4312)形成UE 3802和MME 3828之间的控制平面的最高层。在至少一个实施例中，NAS 协议4312支持UE 3802的移动性和会话管理过程，以在UE 3802和P-GW 3834之间建立和维护IP连接。

[0575] 在至少一个实施例中，Si应用协议(S1-AP)层(Si-AP层4322)可以支持 Si接口的功能并且包括基本过程(EP)。在至少一个实施例中，EP是RAN 3816和CN 3828之间的交互单元。在至少一个实施例中，S1-AP层服务可以包括两组：UE相关服务和非UE相关服务。在至少一个实施例中，这些服务执行的功能包括但不限于：E-UTRAN无线接入承载(E-RAB)管理、UE能力指示、移动性、NAS信令传输、RAN信息管理(RIM)和配置传输。

[0576] 在至少一个实施例中，流控制传输协议(SCTP)层(或者称为流控制传输协议/互联网协议(SCTP/IP)层)(SCTP层4320)可以确保信令消息在RAN 3816和MME 3828部分地基于IP协议可靠地传送，由IP层4318支持。在至少一个实施例中，L2层4316和L1层4314可以指通信链路(例如，有线或无线)被RAN节点和MME用来交换信息。

[0577] 在至少一个实施例中，RAN 3816和MME 3828可以利用S1-MME 接口经由包括L1层4314、L2层4316、IP层4318、SCTP层4320和Si-AP 层4322的协议栈来交换控制平面数据。

[0578] 图44是根据至少一个实施例的用户平面协议栈的图示。在至少一个实施例中，用户平面4400被示为UE 3802、RAN 3816、S-GW 3830和P-GW 3834之间的通信协议栈。在至少一个实施例中，用户平面4400可以使用相同的协议层作为控制平面4300。例如，在至少一个实施例中，UE 3802 和RAN 3816可以利用Uu接口(例如，LTE-Uu接口)经由包括PHY层 4302、MAC层4304、RLC层4306、PDCP层4308的协议栈来交换用户平面数据。

[0579] 在至少一个实施例中，用于用户平面(GTP-U)层(GTP-U层4404)的通用分组无线服务(GPRS)隧道协议可用于在GPRS核心网络内以及在无线接入网络和核心网之间承载用户数据。在至少一个实施例中，传输的用户数据可以是例如IPv4、IPv6或PPP格式中的任一种的分组。在至少一个实施例中，UDP和IP安全(UDP/IP)层(UDP/IP层4402)可以提供用于数据完整性的校验和、用于在源和目的地寻址不同功能的端口号以及对选定数据流的加密和认证。在至少一个实施例中，RAN 3816和S-GW 3830可以利用 S1-U接口经由包括L1层4314、L2层4316、UDP/IP层4402和GTP-U层 4404的协议栈来交换用户平面数据。在至少一个实施例中，S-GW 3830和 P-GW 3834可以利用S5/S8a接口经由包括L1层4314、L2层4316、UDP/IP 层4402和GTP-U层4404的协议栈来交换用户平面数据。在至少一个实施例中，如上文关于图43所讨论的，NAS协议支持UE 3802的移动性和会话管理过程，以在UE 3802和P-GW 3834之间建立和维持IP连接。

[0580] 图45示出了根据至少一个实施例的核心网络的组件4500。在至少一个实施例中，CN 3838的组件可以在一个物理节点或单独的物理节点中实现，包括从机器可读或计算机可读介质(例如，非暂时性机器可读存储介质)读取和执行指令的组件。在至少一个实施例中，网络功能虚拟化(NFV) 用于经由存储在一个或更多个计算机可读存储介质中的可执行指令来虚拟化上述网络节点功能中的任何一个或所有(下面进一步详细描述)。在至少一个实施例中，CN 3838的逻辑实例可以被称为网络切片4502(例如，网络切片4502被示为包括HSS 3832、MME 3828和S-GW 3830)。在至少一个实施例中，CN 3838的一部分的逻辑实例可以被称为网络子切片4504(例如，网络子切片4504被示为包括P-GW 3834和PCRF 3836)。

[0581] 在至少一个实施例中，NFV架构和基础设施可用于将一个或更多个网络功能虚拟

化,或者由专有硬件执行到物理资源上,该物理资源包括行业标准服务器硬件、存储硬件或交换机的组合。在至少一个实施例中,NFV 系统可用于执行一个或更多个EPC组件/功能的虚拟或可重新配置的实现。

[0582] 图46是说明根据至少一个实施例的支持网络功能虚拟化(NFV)的系统4600的组件的框图。在至少一个实施例中,系统4600被示为包括虚拟化基础设施管理器(示为VIM 4602)、网络功能虚拟化基础设施(示为NFVI 4604)、VNF管理器(示为VNFM 4606)、虚拟化网络功能(示为VNF 4608)、元素管理器(示为EM 4610)、NFVO协调器(示为NFVO 4612)和网络管理器(示为NM 4614)。

[0583] 在至少一个实施例中,VIM 4602管理NFVI4604的资源。在至少一个实施例中,NFVI4604可以包括用于执行系统4600的物理或虚拟资源和应用程序(包括管理程序)。在至少一个实施例中,VIM 4602可以使用 NFVI4604管理虚拟资源的生命周期(例如,与一个或更多个物理资源相关联的虚拟机(VM)的创建、维护和拆除),跟踪VM实例,跟踪VM实例的性能、故障和安全性和关联的物理资源,并将VM实例和关联的物理资源暴露给其他管理系统。

[0584] 在至少一个实施例中,VNFM 4606可以管理VNF 4608。在至少一个实施例中,VNF 4608可以用于执行EPC组件/功能。在至少一个实施例中, VNFM 4606可以管理VNF 4608的生命周期并跟踪VNF 4608的虚拟方面的性能、故障和安全性。在至少一个实施例中,EM 4610可以跟踪VNF 4608 的功能方面的性能、故障和安全性。在至少一个实施例中,来自VNFM 4606 和EM 4610的跟踪数据可以包括,例如,由VIM 4602或NFVI4604使用的性能测量(PM)数据。在至少一个实施例中,VNFM 4606和EM 4610都可以扩展上/下系统4600的VNF数量。

[0585] 在至少一个实施例中,NFVO 4612可以协调、授权、释放和使用NFVI 4604的资源以提供请求的服务(例如,执行EPC功能、组件或切片)。在至少一个实施例中,NM 4614可以提供负责管理网络的最终用户功能包,其可以包括具有VNF、非虚拟化网络功能或两者的网络元素(VNF的管理可以通过EM 4610)。

[0586] 其他变化在本公开的精神内。因此,虽然所公开的技术易于进行各种修改和替代构造,但其某些图示实施例在附图中示出并且已经在上面详细描述。然而,应当理解,无意将公开内容限制于所公开的特定形式或形式,而是相反,意图是涵盖落入公开内容的精神和范围内的所有修改、替代构造和等同物,如所附权利要求所定义。

[0587] 鉴于以下条款,可以描述本公开的至少一个实施例:

[0588] 1.一种处理器,包括:

[0589] 一个或更多个电路用于在一个或更多个并行处理单元上对第五代5G 新无线电信号信息进行编码和解码。

[0590] 2.根据条款1所述的处理器,其中所述一个或更多个电路通过执行一个或更多个物理层步骤来编码和解码所述5G新无线电信号信息,其中所述一个或更多个物理层步骤部分地由所述一个或更多个并行处理单元执行,以及部分地由中央处理单元执行。

[0591] 3.根据条款1或2所述的处理器,其中对所述5G新无线电信号信息进行编码包括启动一个或更多个并行处理单元内核以执行一个或更多个信号处理块。

[0592] 4.根据条款3所述的处理器,其中所述一个或更多个信号处理块是物理上行链路共享信道PUSCH信号处理块。

[0593] 5.根据条款3或4所述的处理器，其中所述一个或更多个信号处理块是物理下行链路共享信道PDSCH信号处理块。

[0594] 6.根据条款1-5中任一项所述的处理器，其中所述一个或更多个并行处理单元直接从一个或更多个网络接口控制器发送和接收所述5G新无线电信号信息。

[0595] 7.根据条款1-6中任一项所述的处理器，其中所述一个或更多个并行处理单元是图形处理单元。

[0596] 8.一种系统，包括：

[0597] 一个或更多个处理器，使用一个或更多个并行处理单元并行执行第五代5G信号处理的一个或更多个阶段。

[0598] 9.根据条款8所述的系统，其中所述5G信号处理的一个或更多个阶段包括使用部分地由所述一个或更多个并行处理单元执行的软件指令来执行物理层运算。

[0599] 10.根据条款9所述的系统，其中由所述一个或更多个并行处理单元执行的所述软件指令由中央处理单元执行的附加指令控制。

[0600] 11.根据条款9或10所述的系统，其中所述物理层运算是通过从CPU 主机启动并行处理单元功能来执行的。

[0601] 12.根据条款8-11中任一项所述的系统，所述5G信号处理的一个或更多个阶段属于5G物理层并且响应于来自与所述物理层不同的层的应用编程调用而被执行。

[0602] 13.根据条款8-12中任一项所述的系统，其中所述5G信号处理的一个或更多个阶段是对一个或更多个网络接口控制器发送或接收的一个或更多个信号执行的。

[0603] 14.根据条款8-13中任一项所述的系统，其中所述并行处理单元直接从一个或更多个网络接口控制器接收用于5G信号处理的一个或更多个信号。

[0604] 15.根据条款8-14中任一项所述的系统，其中所述一个或更多个并行处理单元是图形处理单元。

[0605] 16.一种机器可读介质，其上存储有一组指令，所述指令如果由一个或更多个处理器执行，则使一个或更多个处理器至少：

[0606] 使用一个或更多个并行处理单元对第五代5G新无线电信号信息进行编码和解码。

[0607] 17.根据条款16所述的机器可读介质，其中所述指令如果被执行，则进一步使所述一个或更多个处理器执行一个或更多个物理层步骤，以编码和解码所述5G新无线电信号信息，其中所述一个或更多个物理层步骤部分地由所述一个或更多个并行处理单元执行，以及部分地由中央处理单元执行。

[0608] 18.根据条款16或17所述的机器可读介质，其中对所述5G新无线电信号信息进行编码和解码包括：启动一个或更多个并行处理单元内核以在所述一个或更多个并行处理单元上执行一个或更多个信号处理块。

[0609] 19.根据条款18所述的机器可读介质，其中所述一个或更多个信号处理块是物理上行链路共享信道PUSCH信号处理块。

[0610] 20.根据条款18或19所述的机器可读介质，其中所述一个或更多个信号处理块是物理下行链路共享信道PDSCH信号处理块。

[0611] 21.根据条款16-20中任一项所述的机器可读介质，其中所述一个或更多个并行处理单元直接从一个或更多个网络接口控制器发送和接收所述 5G新无线电信号信息。

[0612] 22.根据条款16-21中任一项所述的机器可读介质,其中通过启动以下项中的至少一个来对所述5G新无线电信号信息进行编码:物理上行链路共享信道(PUSCH)波束成形内核、PUSCH前端内核、PUSCH解速率匹配内核、PUSCH低密度奇偶校验(LDPC)内核或循环冗余校验(CRC) 内核。

[0613] 23.根据条款16-22中任一项所述的机器可读介质,其中通过启动以下项中的至少一个来执行对所述5G新无线电信号信息的解码:物理下行链路共享信道(PDSCH)波束成形内核、PDSCH速率匹配内核、PDSCH 低密度奇偶校验(LDPC)内核或循环冗余校验(CRC)内核。

[0614] 24.根据条款16-23中任一项所述的机器可读介质,其中所述5G新无线电信号信息通过一个或更多个网络接口控制器发送和接收。

[0615] 25.一种方法,包括:

[0616] 使用一个或更多个并行处理单元并行执行第五代(5G)信号处理的一个或更多个阶段。

[0617] 26.根据条款25所述的方法,其中所述5G信号处理的一个或更多个阶段进一步包括:使用由所述一个或更多个并行处理单元执行的软件函数来执行物理层运算。

[0618] 27.根据条款26所述的方法,其中由所述一个或更多个并行处理单元执行的软件函数由中央处理单元控制。

[0619] 28.根据条款26或27所述的方法,其中所述软件函数实现所述物理层的一个或更多个信号处理运算。

[0620] 29.根据条款28所述的方法,其中所述一个或更多个信号处理运算是物理上行链路共享信道PUSCH信号处理运算。

[0621] 30.根据条款28或29所述的方法,其中所述一个或更多个信号处理运算是物理下行链路共享信道PDSCH信号处理运算。

[0622] 31.根据条款25-30中任一项所述的方法,其中所述一个或更多个并行处理单元直接从一个或更多个网络接口控制器接收用于5G信号处理的一个或更多个信号。

[0623] 32.根据条款25-31中任一项所述的方法,其中所述一个或更多个并行处理单元是图形处理单元。

[0624] 除非本文另有说明,否则在描述公开的实施例的上下文中(尤其是在以下权利要求的上下文中)使用术语“一个(a)”和“一个(an)”和“该 (the)”以及类似的指代应被解释为涵盖单数和复数。显然与上下文矛盾,而不是作为术语的定义。除非另有说明,否则术语“包括(comprising)”、“具有(having)”、“包括(including)”和“包含(containing)”应被解释为开放式术语(意为“包括但不限于”)。术语“连接”,当未经修改并指物理连接时,应被解释为部分或全部包含在内部、连接到或连接在一起,即使有一些干预。除非本文另有说明,否则本文对数值范围的引用仅旨在作为单独提及落入范围内的每个单独值的速记方法,并且将每个单独值并入说明书中,就好像它在本文中单独引用一样。除非另有说明或与上下文相矛盾,否则术语“集合”(例如,“一组项目”)或“子集”的使用将被解释为包括一个或更多个成员的非空集合。此外,除非另有说明或与上下文相矛盾,否则对应集合的术语“子集”不一定表示对应集合的真子集,而子集与对应集合可以相等。

[0625] 连接语言,例如“A、B和C中的至少一个”或“A、B和C中的至少一个”形式的短语,除非另有特别说明或与上下文明显矛盾,否则为结合上下文理解为一般用于表示项目、术语

等可以是A或B或C，或A和B 和C的集合的任何非空子集。例如，在具有以下特征的集合的说明性示例中三个成员，连词“A、B和C中的至少一个”和“A、B和C中的至少一个”是指以下任一集合：{A}、{B}、{C}、{A,B}、{A,C}、{B,C}、{A,B,C}。因此，这样的连词一般不旨在暗示某些实施例需要A中的至少一个、B中的至少一个和C中的至少一个每个都存在。此外，除非另有说明或与上下文相矛盾，否则术语“多个”表示复数的状态(例如，“多个项目”表示多个项目)。多个项目的数量至少为两个，但是当明确地或通过上下文如此指示时可以更多。此外，除非另有说明或上下文清楚，否则短语“基于”是指“至少部分基于”而不是“仅基于”。

[0626] 除非本文另有说明或与上下文明显矛盾，否则本文描述的过程的操作可以以任何合适的顺序执行。在至少一个实施例中，诸如本文描述的那些过程(或其变体和/或组合)的过程在配置有可执行指令的一个或更多个计算机系统的控制下执行并且被实现为代码(例如，可执行指令、一个或更多个)多个计算机程序或一个或更多个应用程序)通过硬件或其组合在一个或更多个处理器上共同执行。在至少一个实施例中，代码例如以包括可由一个或更多个处理器执行的多个指令的计算机程序的形式存储在计算机可读存储介质上。在至少一个实施例中，计算机可读存储介质是不包括瞬时信号(例如，传播的瞬时电或电磁传输)但包括非瞬时数据存储电路(例如，缓冲器、缓存和队列)在瞬态信号的收发器内。在至少一个实施例中，代码(例如，可执行代码或源代码)被存储在其上存储了可执行指令(或用于存储可执行指令的其他存储器)的一组一个或更多个非暂时性计算机可读存储介质上，当由计算机系统的一个或更多个处理器执行(即，作为被执行的结果)导致计算机系统执行本文描述的操作。一组非暂时性计算机可读存储介质，在至少一个实施例中，包括多个非暂时性计算机可读存储介质和多个非暂时性计算机可读存储介质中的一个或更多个单独的非暂时性存储介质。所有代码，而多个非暂时性计算机可读存储介质共同存储所有代码。在至少一个实施例中，可执行指令被执行以使得不同的指令由不同的处理器执行——例如，非暂时性计算机可读存储介质存储指令并且主中央处理单元(“CPU”)执行一些指令，同时图形处理单元(“GPU”) 执行其他指令。在至少一个实施例中，计算机系统的不同组件具有单独的处理器并且不同的处理器执行不同的指令子集。

[0627] 因此，在至少一个实施例中，计算机系统被配置为实现一种或更多种服务，这些服务单独或共同执行本文描述的过程的操作，并且这样的计算机系统被配置有能够执行操作的适用硬件和/或软件。此外，实现本公开的至少一个实施例的计算机系统是单个设备，并且在另一个实施例中，是分布式计算机系统，包括多个操作不同的设备，使得分布式计算机系统执行本文描述的操作并且使得单个设备不执行所有操作。

[0628] 本文提供的任何和所有示例或示例性语言(例如，“诸如”)的使用仅旨在更好地阐明公开的实施例并且不对公开的范围构成限制，除非另有声明。说明书中的任何语言都不应被解释为表明任何未要求保护的元素对于公开的实践是必不可少的。

[0629] 在此引用的所有参考文献，包括出版物、专利申请和专利，都以相同的程度并入本文作为参考，就好像每篇参考文献被单独地和具体地指示为以引用方式并入并在此整体阐述一样。

[0630] 在描述和权利要求中，可以使用术语“耦合”和“连接”及其派生词。应当理解，这些术语可能并非旨在作为彼此的同义词。相反，在特定示例中，“连接”或“耦合”可用于指示两个或更多个元件彼此直接或间接物理或电接触。“耦合”也可能意味着两个或多个元件彼此

不直接接触，但仍然相互合作或相互作用。

[0631] 除非另外特别说明，否则可以理解，贯穿说明书的术语诸如“处理”、“计算(computing)”、“计算(calculating)”、“确定”等，指的是计算机或计算系统的动作和/或过程或类似的电子计算设备，用于操作和/ 或转换表示为物理量的数据，例如计算系统的寄存器和/或存储器中的电子量，转换为类似地表示为计算系统的存储器、寄存器或其他此类信息存储、传输或显示设备中的物理量的其他数据。

[0632] 以类似的方式，术语“处理器”可以指处理来自寄存器和/或存储器的电子数据并将该电子数据转换成可以存储在寄存器和/或存储器中的其他电子数据的任何设备或设备的一部分。记忆。作为非限制性示例，“处理器”可以是CPU或GPU。“计算平台”可以包括一个或更多个处理器。如本文所使用的，“软件”进程可以包括例如随时间执行工作的软件和/或硬件实体，例如任务、线程和智能代理。此外，每个进程可以指多个进程，用于依次或并行、连续或间歇地执行指令。术语“系统”和“方法”在本文中可互换使用，只要系统可以体现一种或更多种方法并且方法可以被认为是系统。

[0633] 在本文件中，可以参考获得、获取、接收模拟或数字数据或将模拟或数字数据输入到子系统、计算机系统或计算机实现的机器中。获取、获得、接收或输入模拟和数字数据的过程可以通过多种方式来完成，例如通过接收数据作为函数调用的参数或对应用程序接口的调用。在一些实施方式中，可以通过串行或并行接口传输数据来完成获取、获得、接收或输入模拟或数字数据的过程。在另一种实现方式中，获取、获得、接收或输入模拟或数字数据的过程可以通过计算机网络从提供实体向获取实体传输数据来完成。还可以参考提供、输出、传输、发送或呈现模拟或数字数据。在各种示例中，提供、输出、传输、发送或呈现模拟或数字数据的过程可以通过将数据作为函数调用的输入或输出参数、应用程序编程接口的参数或进程间通信机制的参数传输来完成。

[0634] 虽然上面的讨论阐述了所描述技术的示例实现，但是其他架构可以用于实现所描述的功能，并且旨在落入本公开的范围内。此外，虽然以上为了讨论的目的定义了具体的职责分配，但是根据情况，各种功能和职责可能以不同的方式分配和划分。

[0635] 此外，虽然主题已经以结构特征和/或方法行为特定的语言进行了描述，但是应当理解，在所附权利要求中要求保护的主题不一定限于所描述的特定特征或行为。相反，具体特征和动作被公开为实施权利要求的示例性形式。

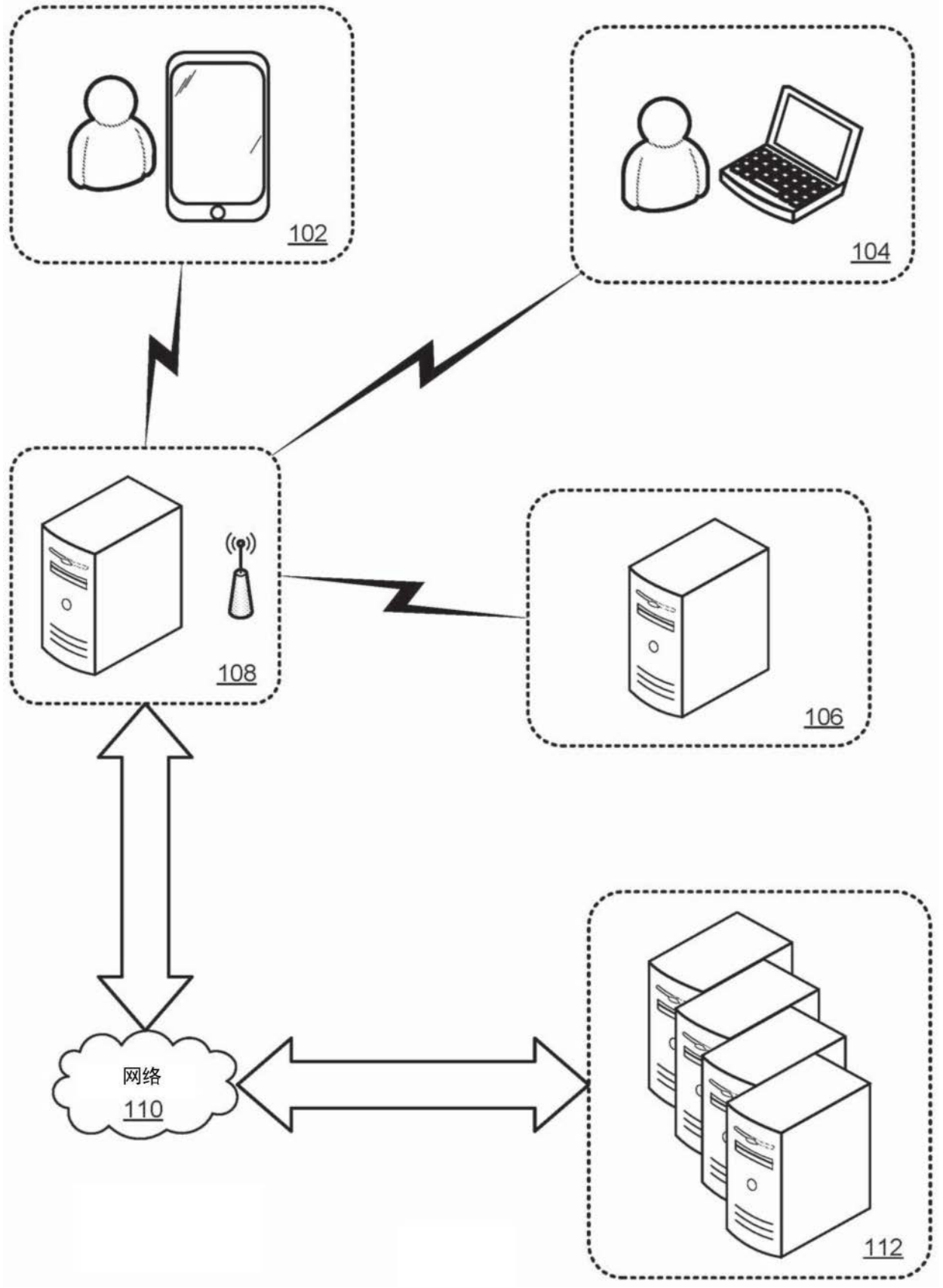


图1

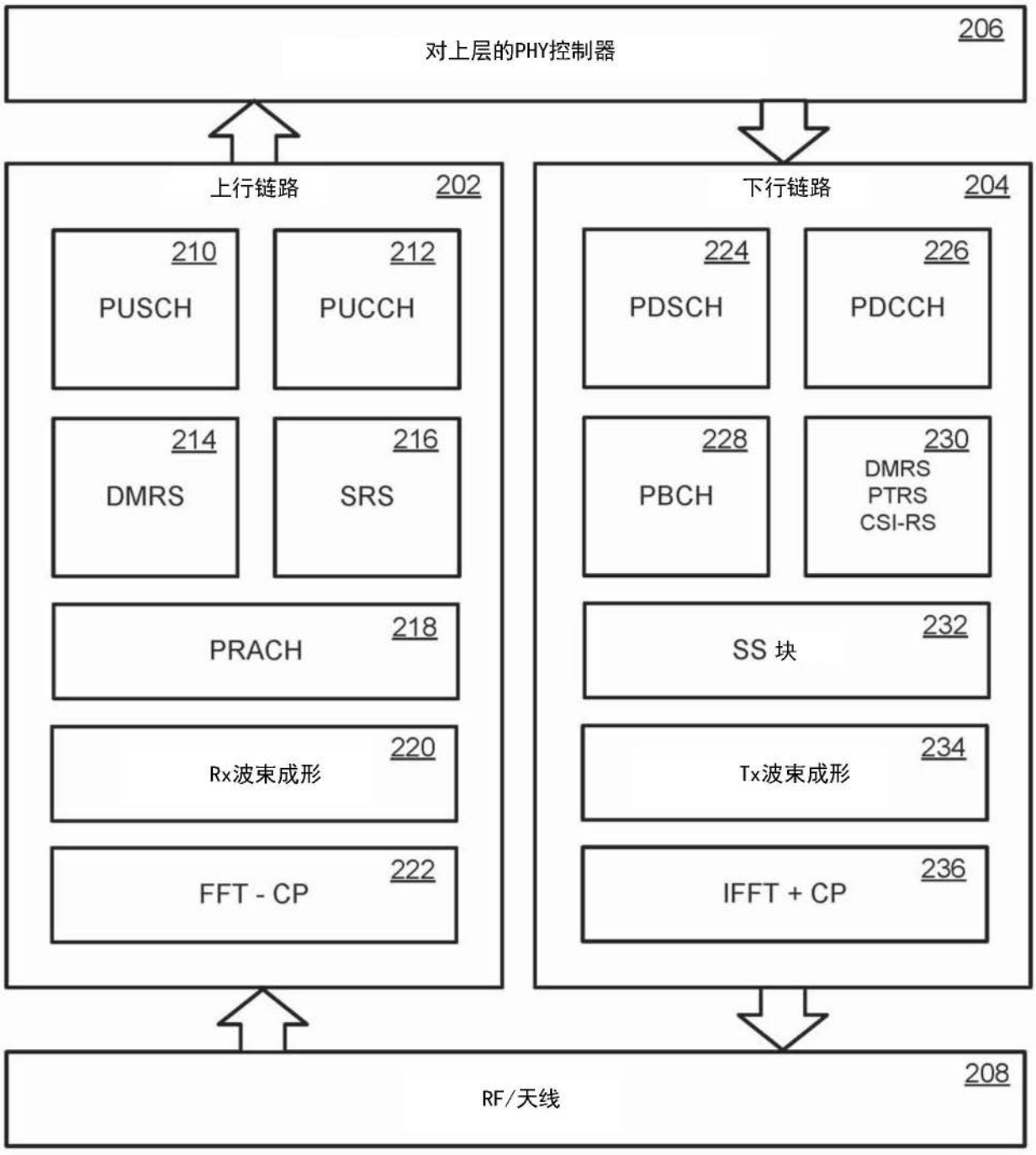

206
对上层的PHY控制器
上行链路
202
下行链路
204
210
PUSCH
212
PUCCH
224
PDSCH
226
PDCCH
214
DMRS
216
SRS
228
PBCH
230
DMRS
PTRS
CSI-RS
218
PRACH
232
SS 块
220
Rx波束成形
234
Tx波束成形
222
FFT - CP
236
IFFT + CP
208
RF/天线

图2

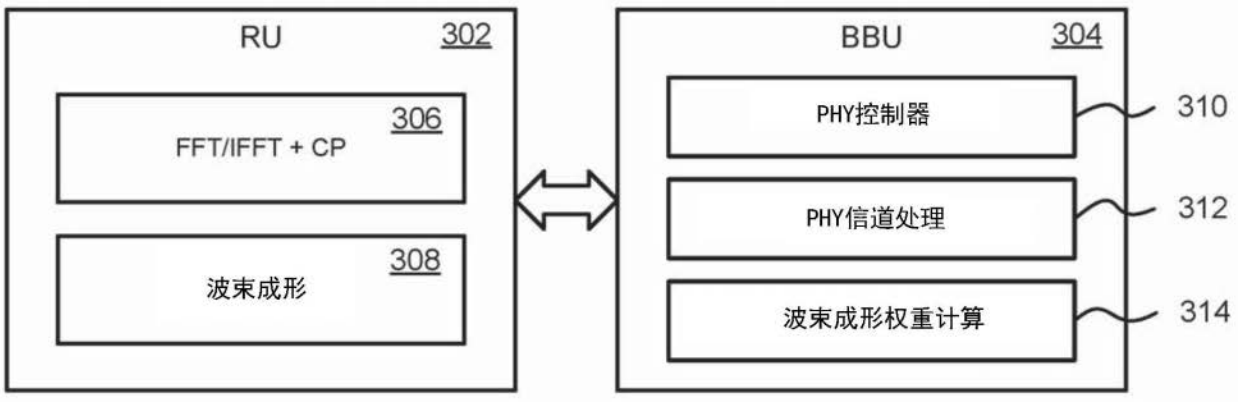

RU
302
306
FFT/IFFT + CP
308
波束成形
BBU
304
PHY控制器
310
PHY信道处理
312
波束成形权重计算
314

图3

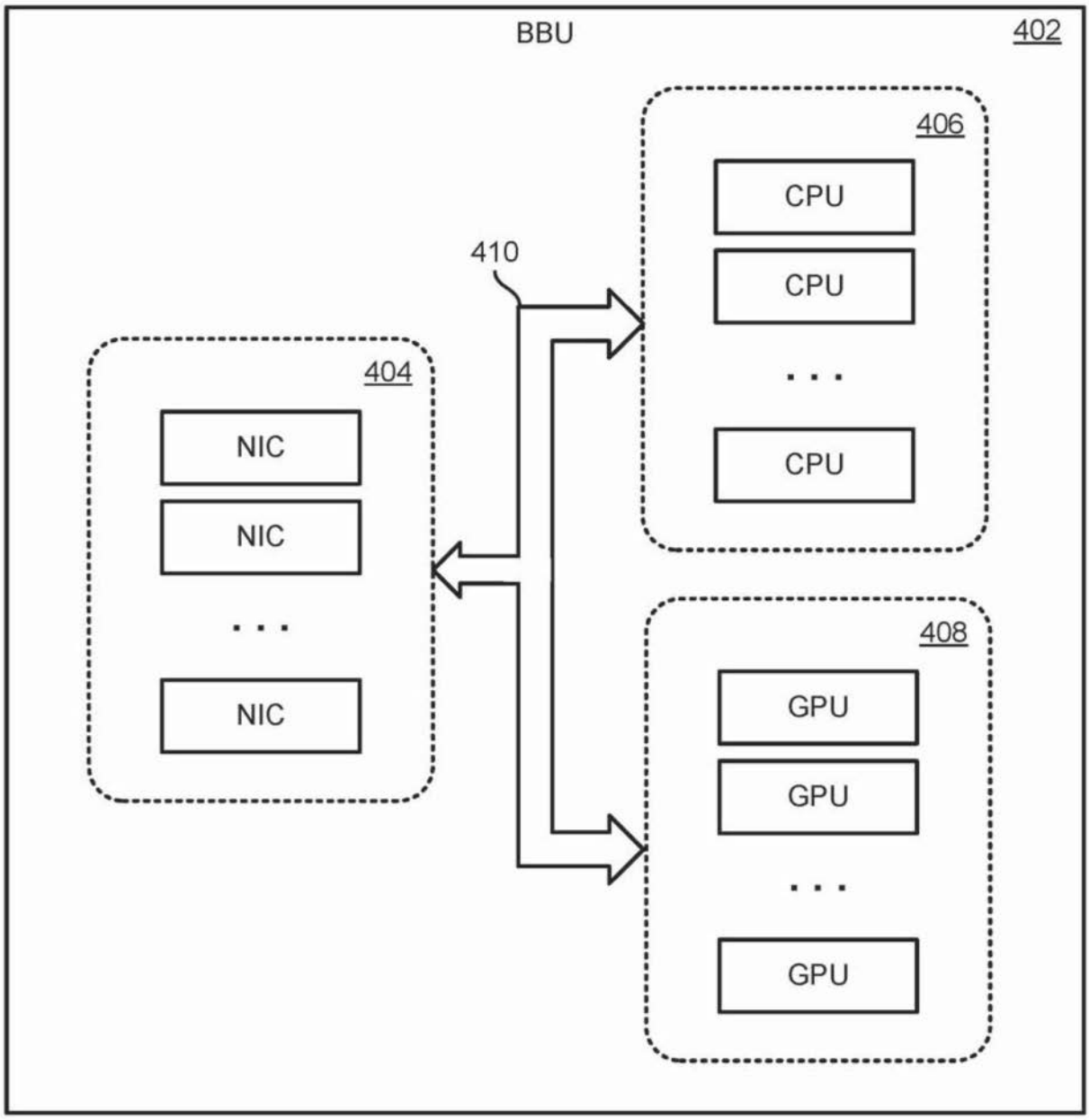


图4

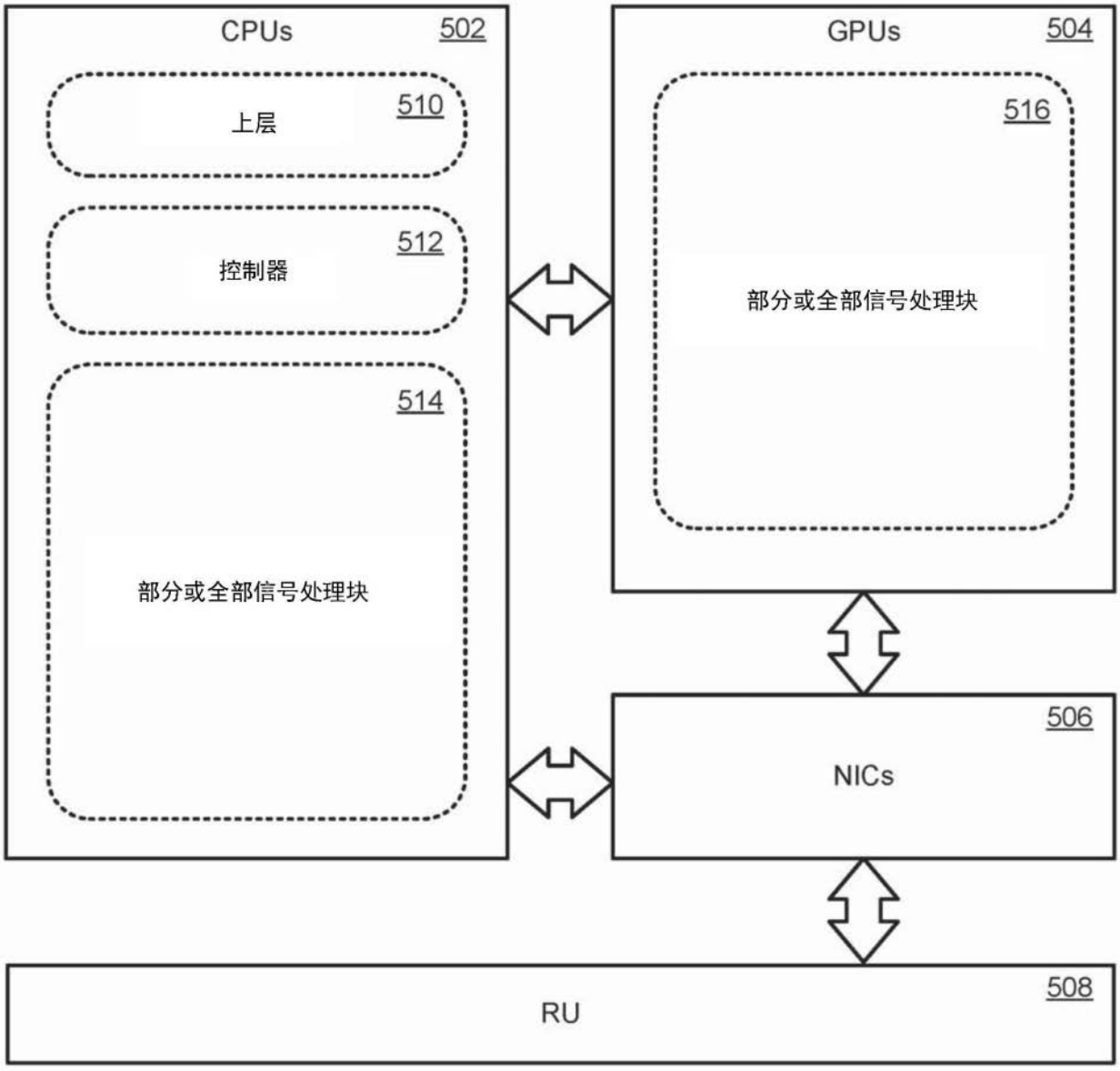

CPUs
502
上层
510
控制器
512
514
部分或全部信号处理块
GPUs
504
516
部分或全部信号处理块
506
NICs
508
RU

图5

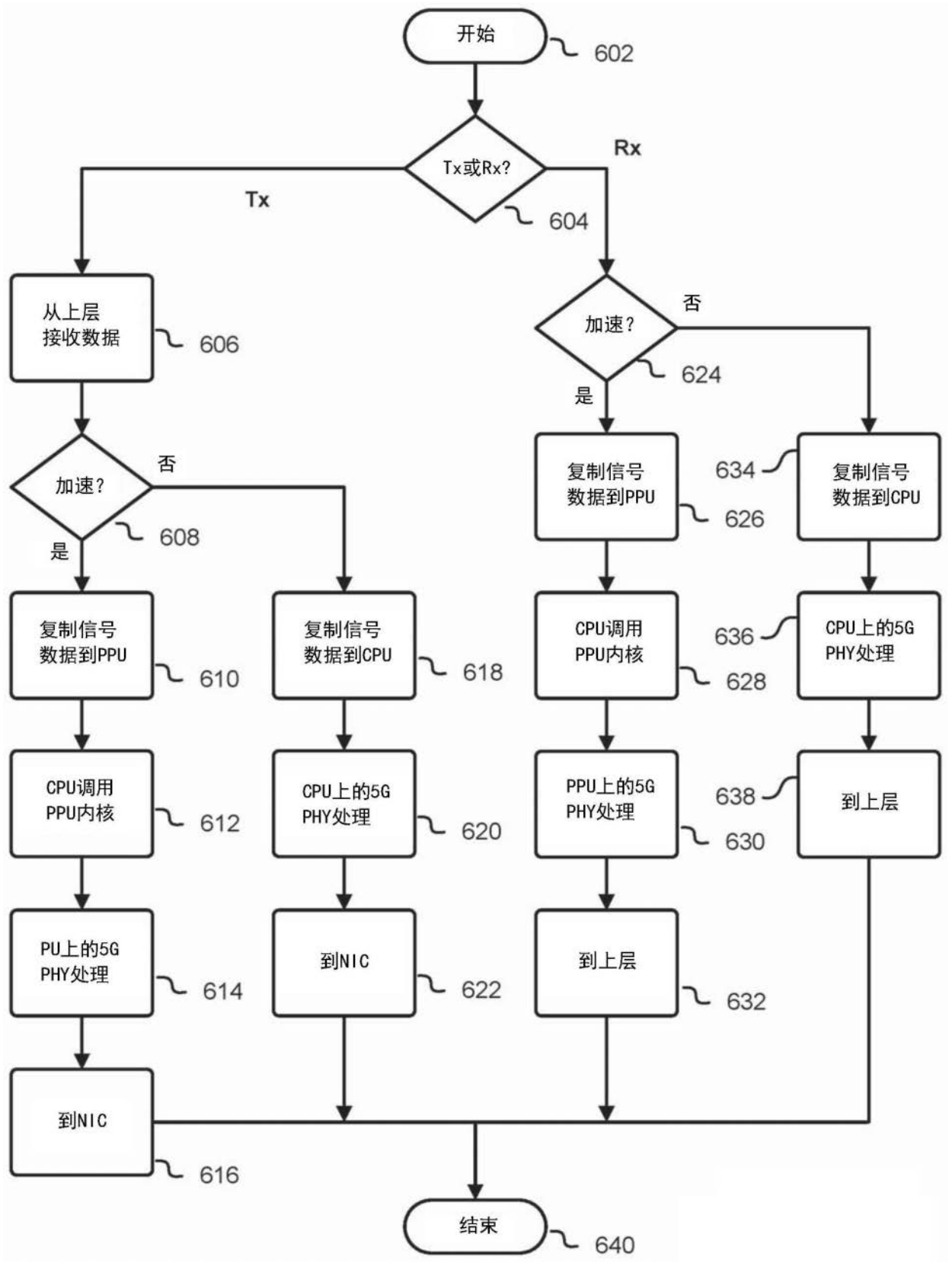

开始
602
Tx或Rx?
604
Tx
Rx
从上层
接收数据
606
加速?
608
否
是
复制信号
数据到PPU
610
CPU调用
PPU内核
612
PU上的5G
PHY处理
614
复制信号
数据到CPU
618
CPU上的5G
PHY处理
620
到NIC
622
到NIC
616
加速?
624
否
是
复制信号
数据到PPU
626
CPU调用
PPU内核
628
PPU上的5G
PHY处理
630
到上层
632
复制信号
数据到CPU
634
CPU上的5G
PHY处理
636
到上层
638
结束
640

图6

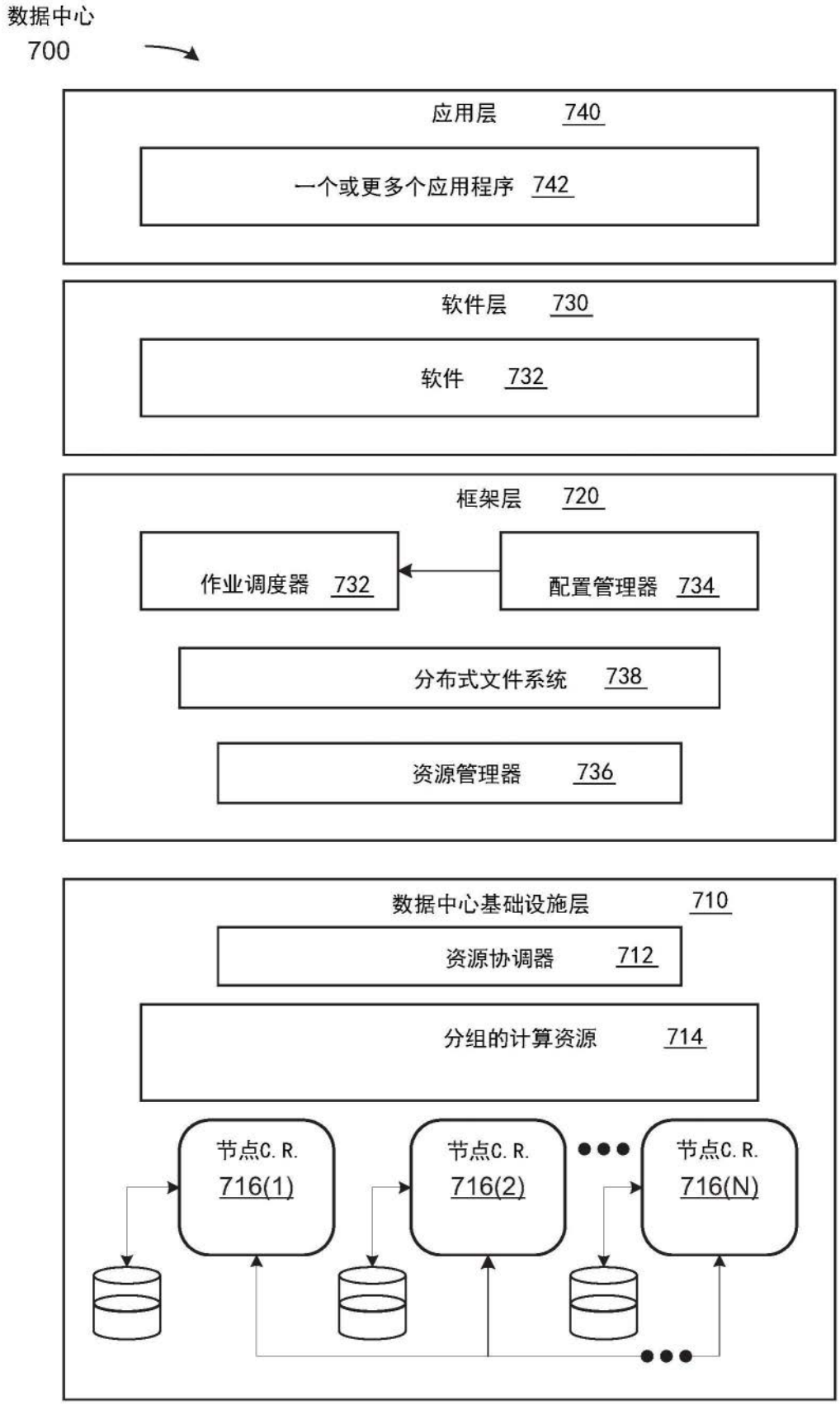

数据中心
700
应用层 740
一个或更多个应用程序 742
软件层 730
软件 732
框架层 720
作业调度器 732
配置管理器 734
分布式文件系统 738
资源管理器 736
数据中心基础设施层 710
资源协调器 712
分组的计算资源 714
节点C. R.
716(1)
节点C. R.
716(2)
节点C. R.
716(N)

图7

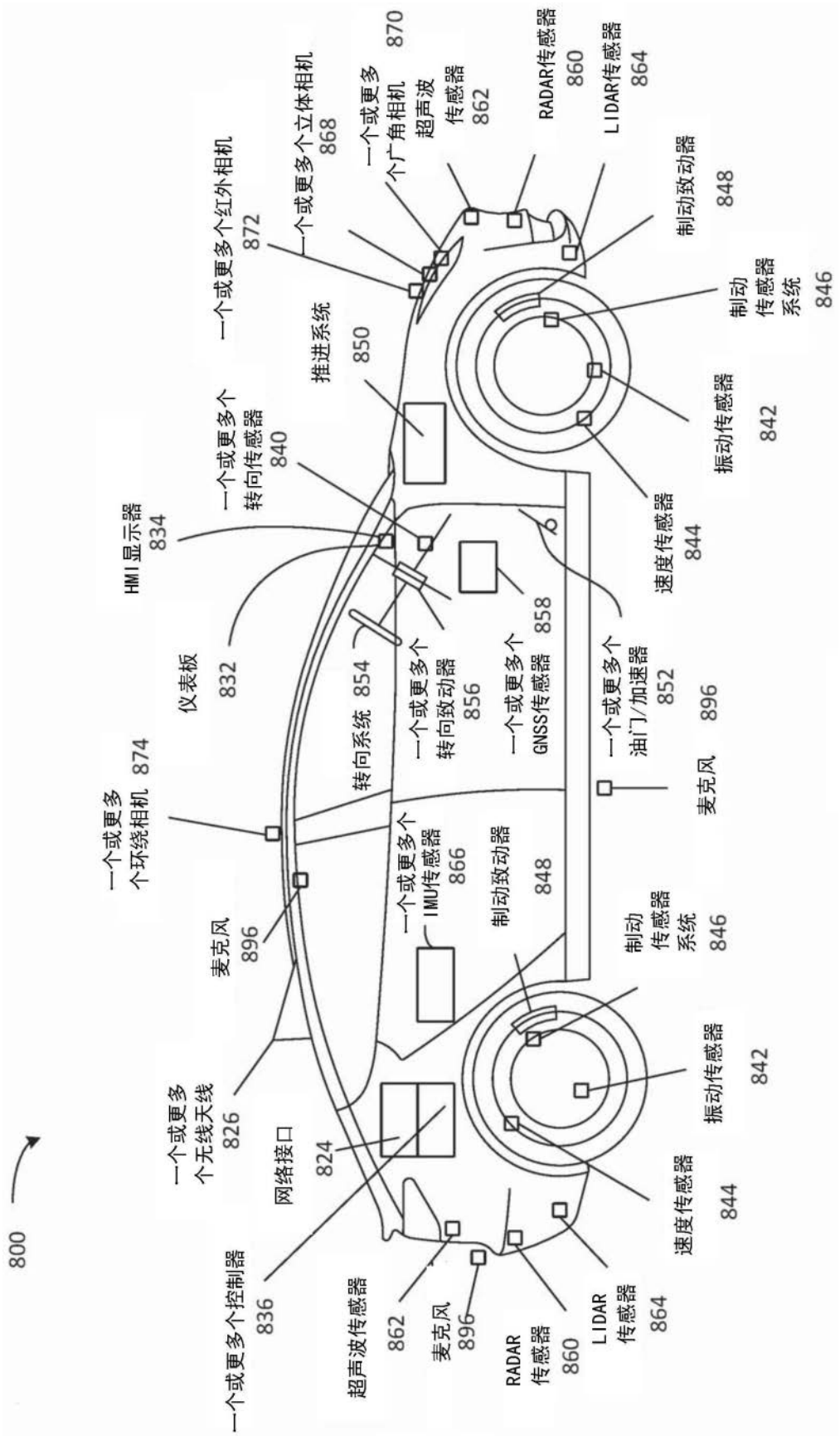

800
一个或更多个控制器 836
超声波传感器 862
麦克风 896
RADAR 传感器 860
LIDAR 传感器 864
速度传感器 844
振动传感器 842
制动传感器系统 846
制动致动器 848
网络接口 824
一个或更多个无线天线 826
麦克风 896
一个或更多个环绕相机 874
一个或更多个IMU传感器 866
制动致动器 848
转向系统 854
一个或更多个转向致动器 856
一个或更多个GNSS传感器 858
一个或更多个油门/加速器 852
麦克风 896
仪表板 832
HMI 显示器 834
一个或更多个转向传感器 840
推进系统 850
一个或更多个红外相机 872
一个或更多个立体相机 868
一个或更多个广角相机 超声波传感器 862 870
RADAR传感器 860
LIDAR传感器 864
制动致动器 848
制动传感器系统 846
振动传感器 842
速度传感器 844

图8A

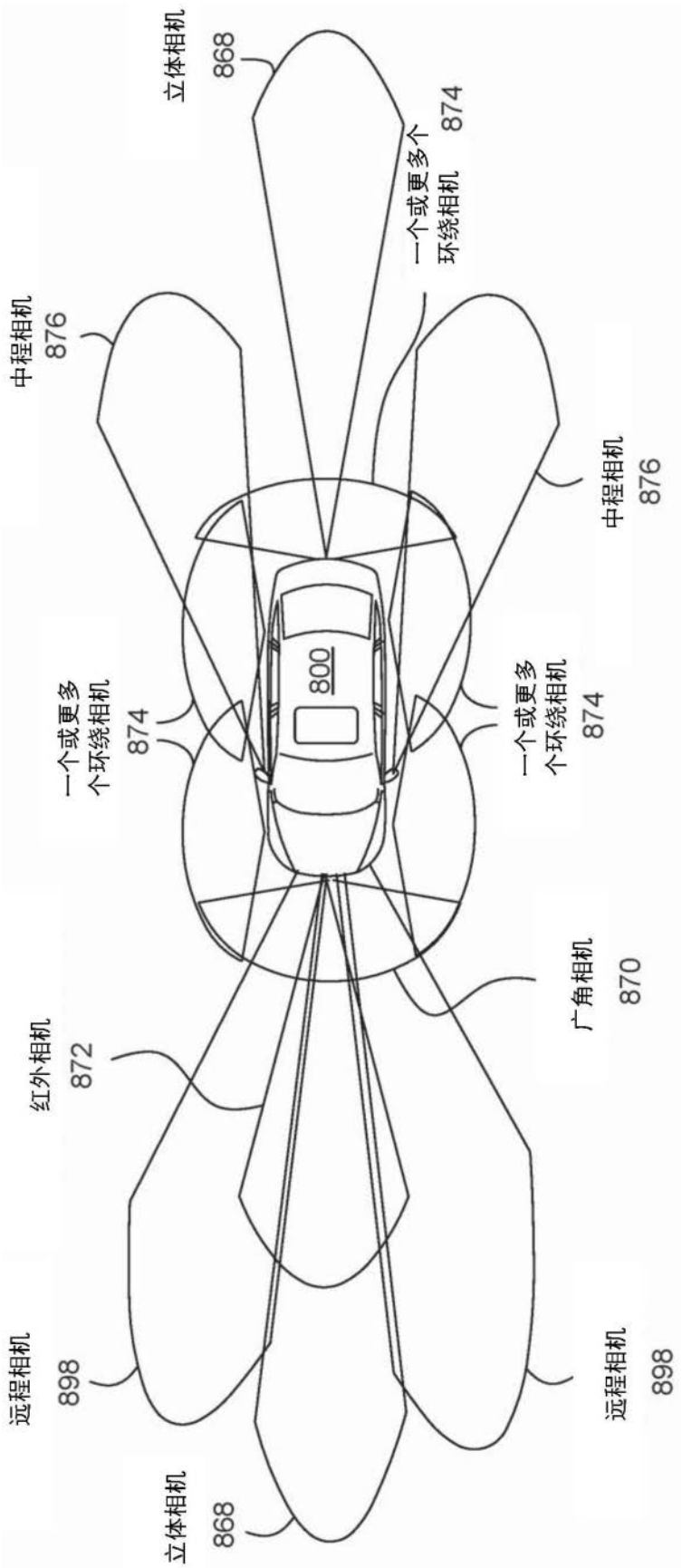

立体相机
868
一个或更多个
环绕相机
874
中程相机
876
800
一个或更多
个环绕相机
874
一个或更多
个环绕相机
874
中程相机
876
红外相机
872
广角相机
870
远程相机
898
远程相机
898
立体相机
868

图8B

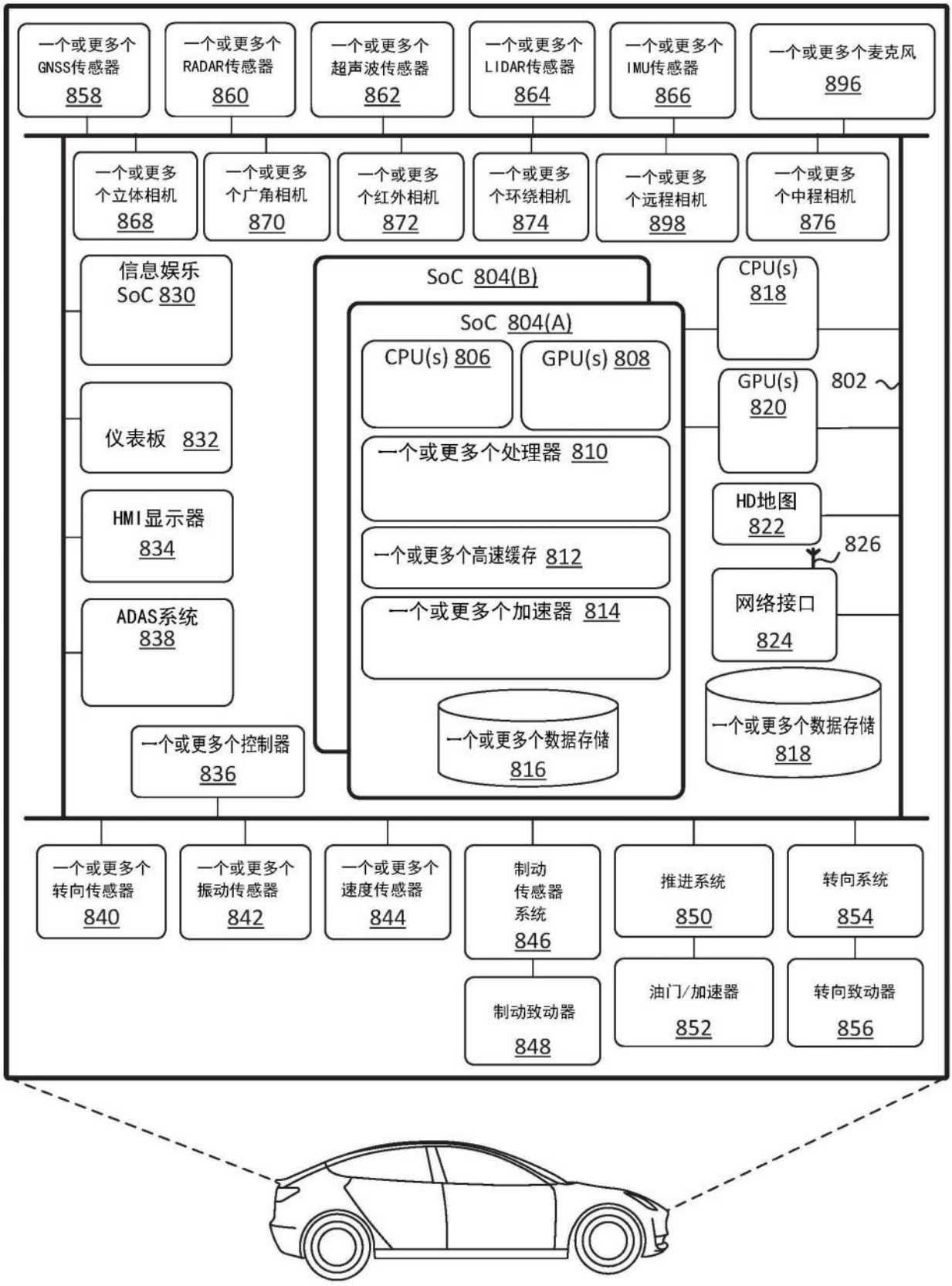

一个或更多个 GNSS传感器 858
一个或更多个 RADAR传感器 860
一个或更多个 超声波传感器 862
一个或更多个 LIDAR传感器 864
一个或更多个 IMU传感器 866
一个或更多个麦克风 896
一个或更多个立体相机 868
一个或更多个广角相机 870
一个或更多个红外相机 872
一个或更多个环绕相机 874
一个或更多个远程相机 898
一个或更多个中程相机 876
信息娱乐 SoC 830
仪表板 832
HMI显示器 834
ADAS系统 838
一个或更多个控制器 836
SoC 804(B)
SoC 804(A)
CPU(s) 806
GPU(s) 808
一个或更多个处理器 810
一个或更多个高速缓存 812
一个或更多个加速器 814
一个或更多个数据存储 816
CPU(s) 818
GPU(s) 820
802
HD地图 822
826
网络接口 824
一个或更多个数据存储 818
一个或更多个 转向传感器 840
一个或更多个 振动传感器 842
一个或更多个 速度传感器 844
制动 传感器 系统 846
制动致动器 848
推进系统 850
油门/加速器 852
转向系统 854
转向致动器 856

图8C

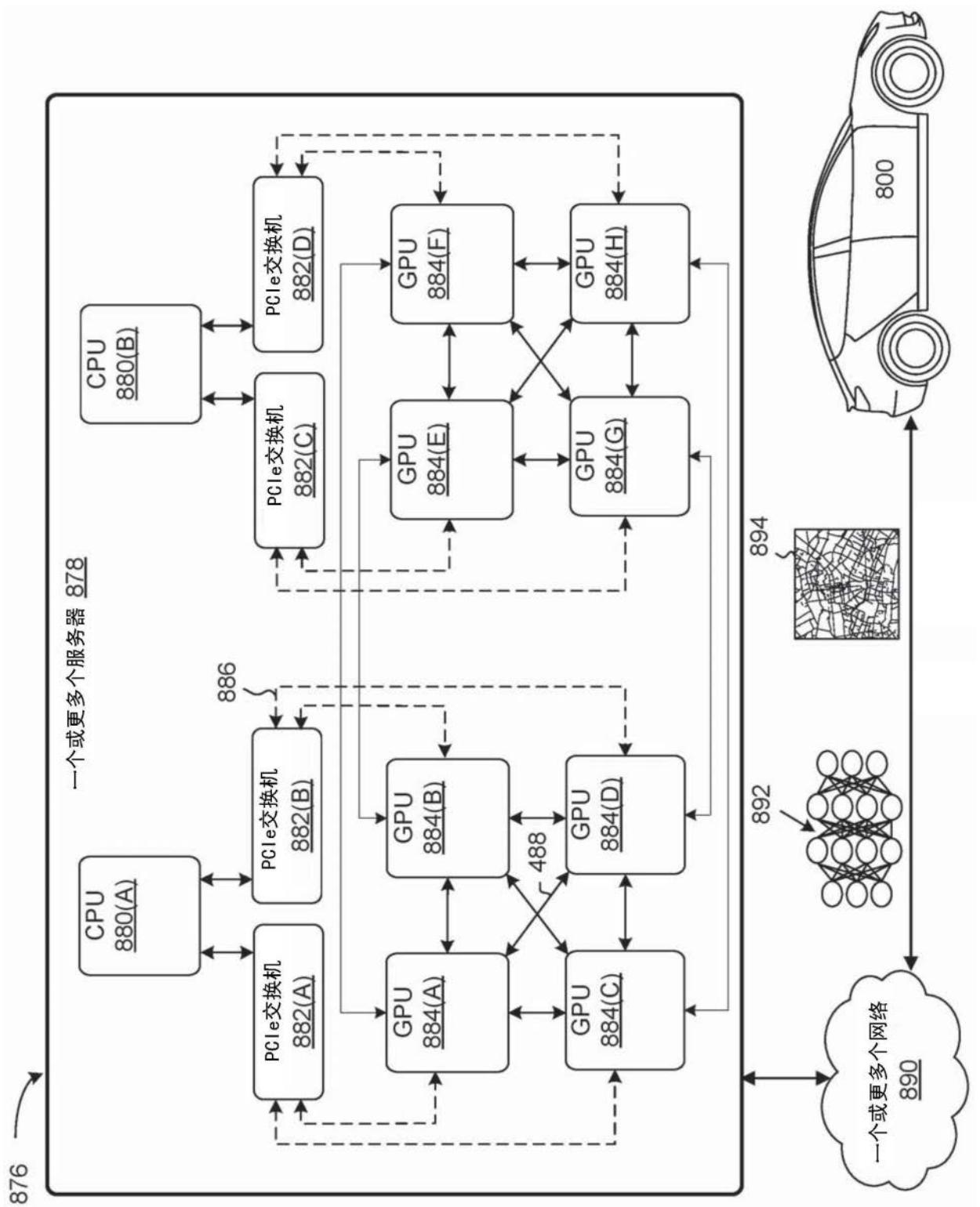

876
一个或更多个服务器 878
CPU 880(A)
CPU 880(B)
PCIe交换机 882(A)
PCIe交换机 882(B)
PCIe交换机 882(C)
PCIe交换机 882(D)
886
GPU 884(A)
GPU 884(B)
GPU 884(C)
GPU 884(D)
GPU 884(E)
GPU 884(F)
GPU 884(G)
GPU 884(H)
488
800
894
892
一个或更多个网络 890

图8D

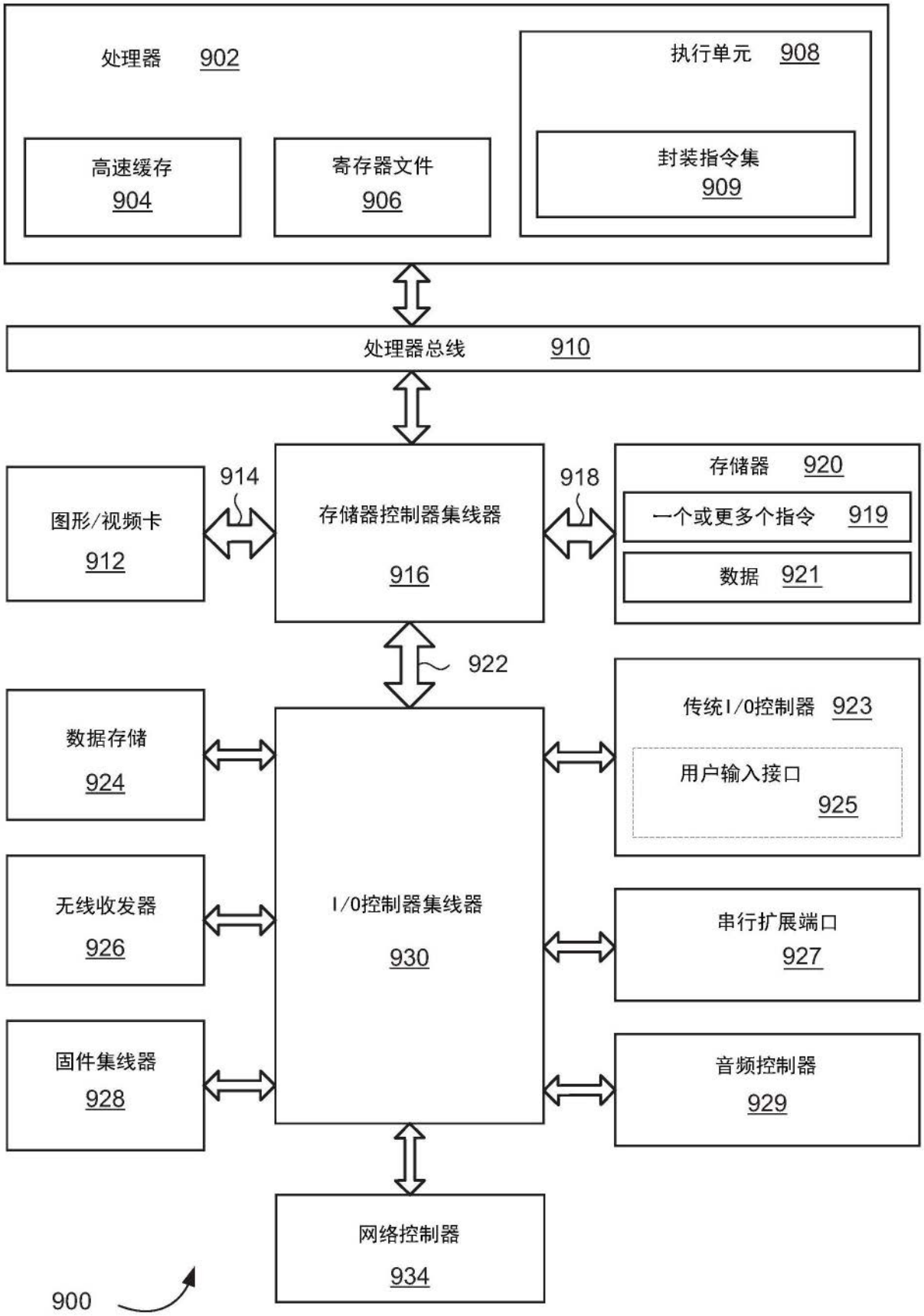

处理器 902
高速缓存 904
寄存器文件 906
执行单元 908
封装指令集 909
处理器总线 910
914
918
图形/视频卡 912
存储器控制器集线器 916
存储器 920
一个或更多个指令 919
数据 921
922
数据存储 924
传统I/O控制器 923
用户输入接口 925
无线收发器 926
I/O控制器集线器 930
串行扩展端口 927
固件集线器 928
音频控制器 929
网络控制器 934
900

图9

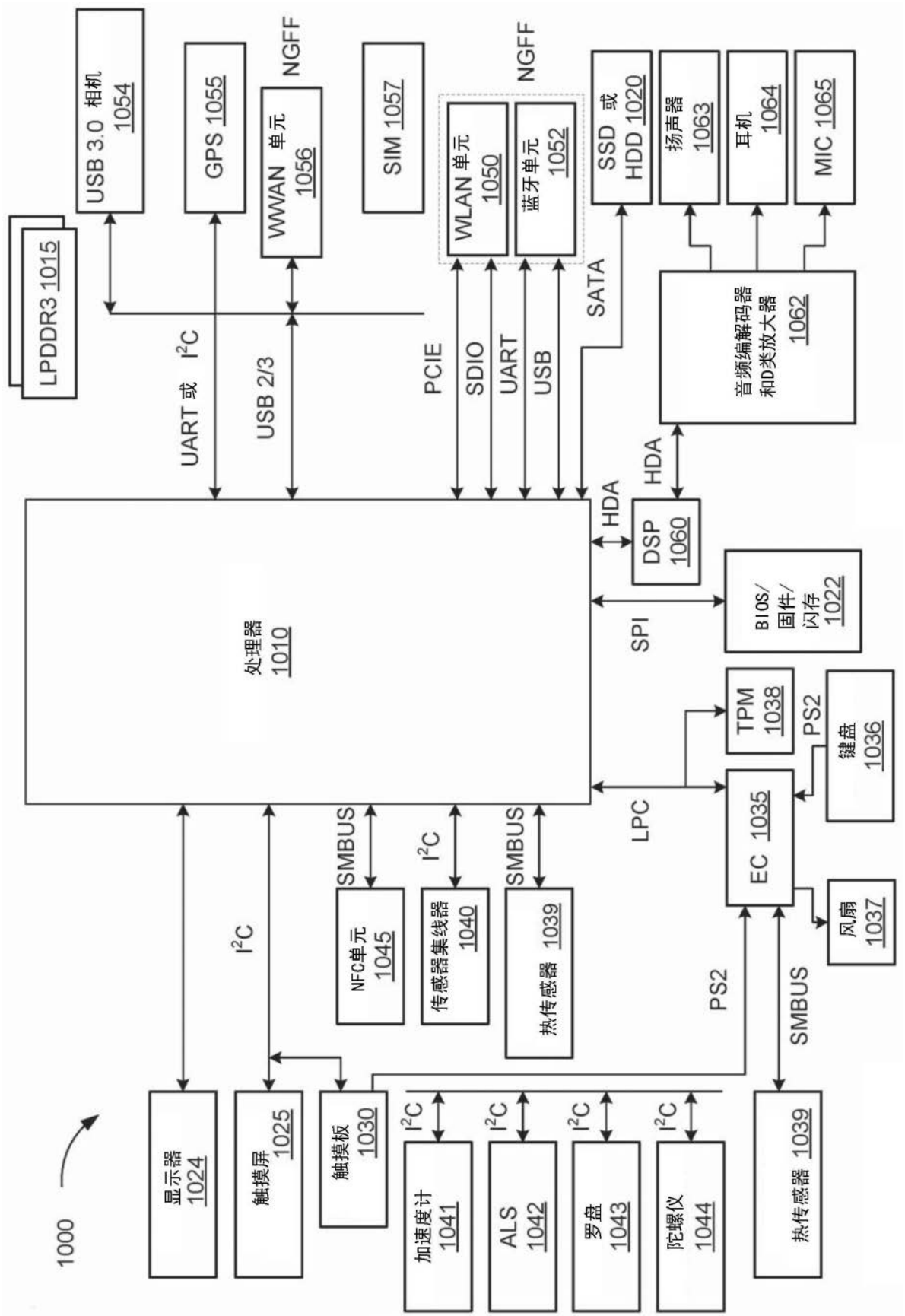

1000
处理器 1010
LPDDR3 1015
USB 3.0 相机 1054
GPS 1055
WWAN 单元 1056
NGFF
SIM 1057
WLAN 单元 1050
蓝牙单元 1052
NGFF
SSD 或 HDD 1020
扬声器 1063
耳机 1064
MIC 1065
音频编解码器和D类放大器 1062
UART 或 I2C
USB 2/3
PCIE
SDIO
UART
USB
SATA
HDA
HDA
DSP 1060
BIOS/固件/闪存 1022
SPI
TPM 1038
PS2
键盘 1036
LPC
EC 1035
风扇 1037
SMBUS
I2C
SMBUS
NFC单元 1045
传感器集线器 1040
热传感器 1039
I2C
PS2
SMBUS
显示器 1024
触摸屏 1025
触摸板 1030
I2C
I2C
I2C
I2C
加速度计 1041
ALS 1042
罗盘 1043
陀螺仪 1044
热传感器 1039

图10

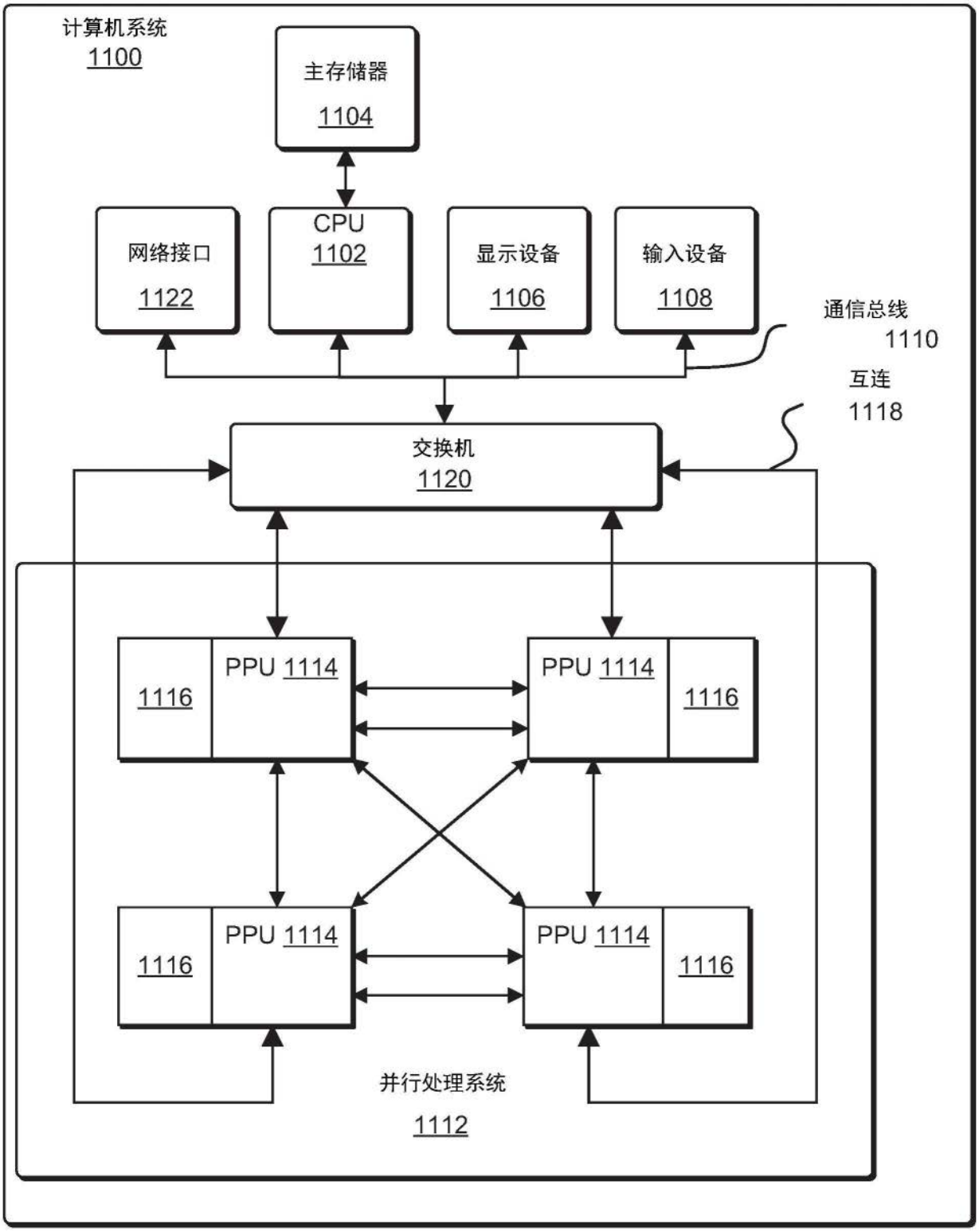

计算机系统
1100
主存储器
1104
CPU
1102
网络接口
1122
显示设备
1106
输入设备
1108
通信总线
1110
互连
1118
交换机
1120
1116
PPU 1114
PPU 1114
1116
1116
PPU 1114
PPU 1114
1116
并行处理系统
1112

图11

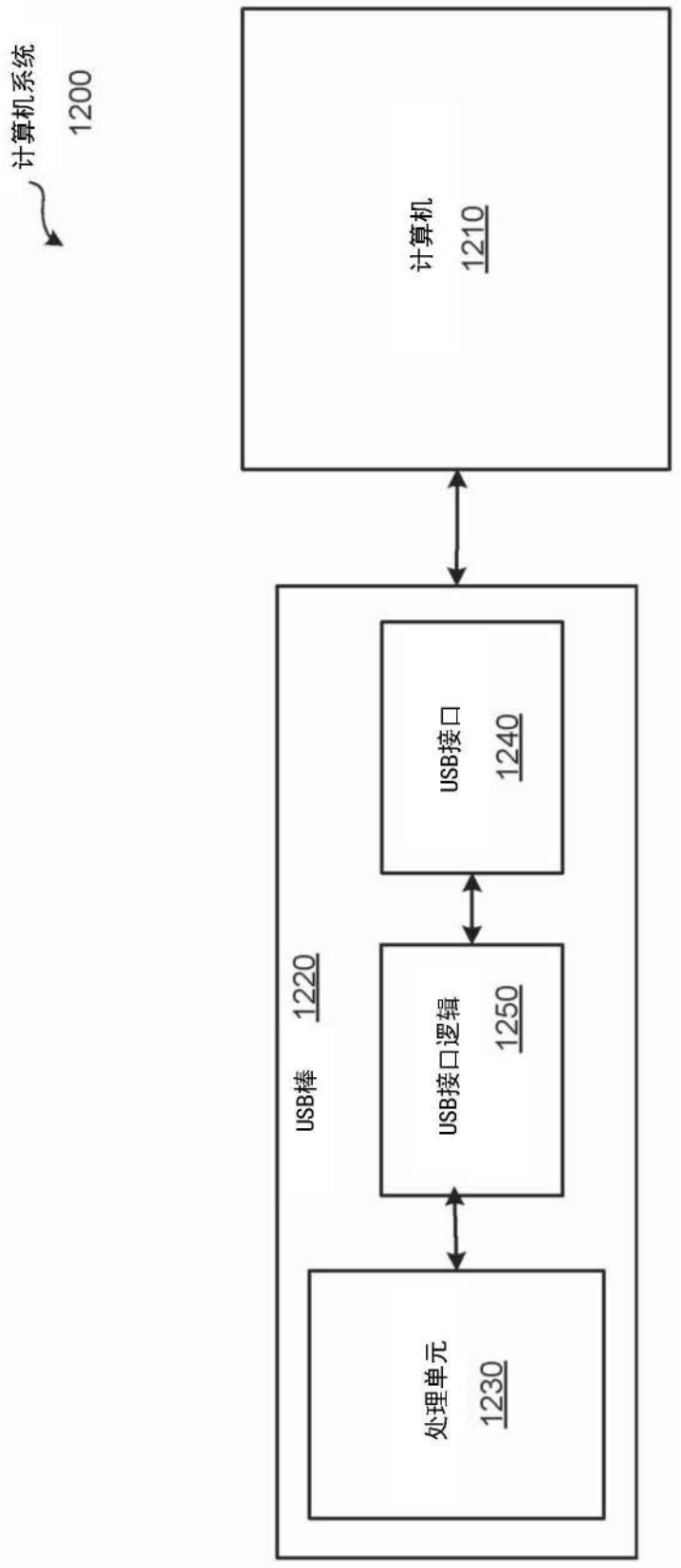

计算机系统
1200
计算机
1210
USB棒
1220
USB接口
1240
USB接口逻辑
1250
处理单元
1230

图12

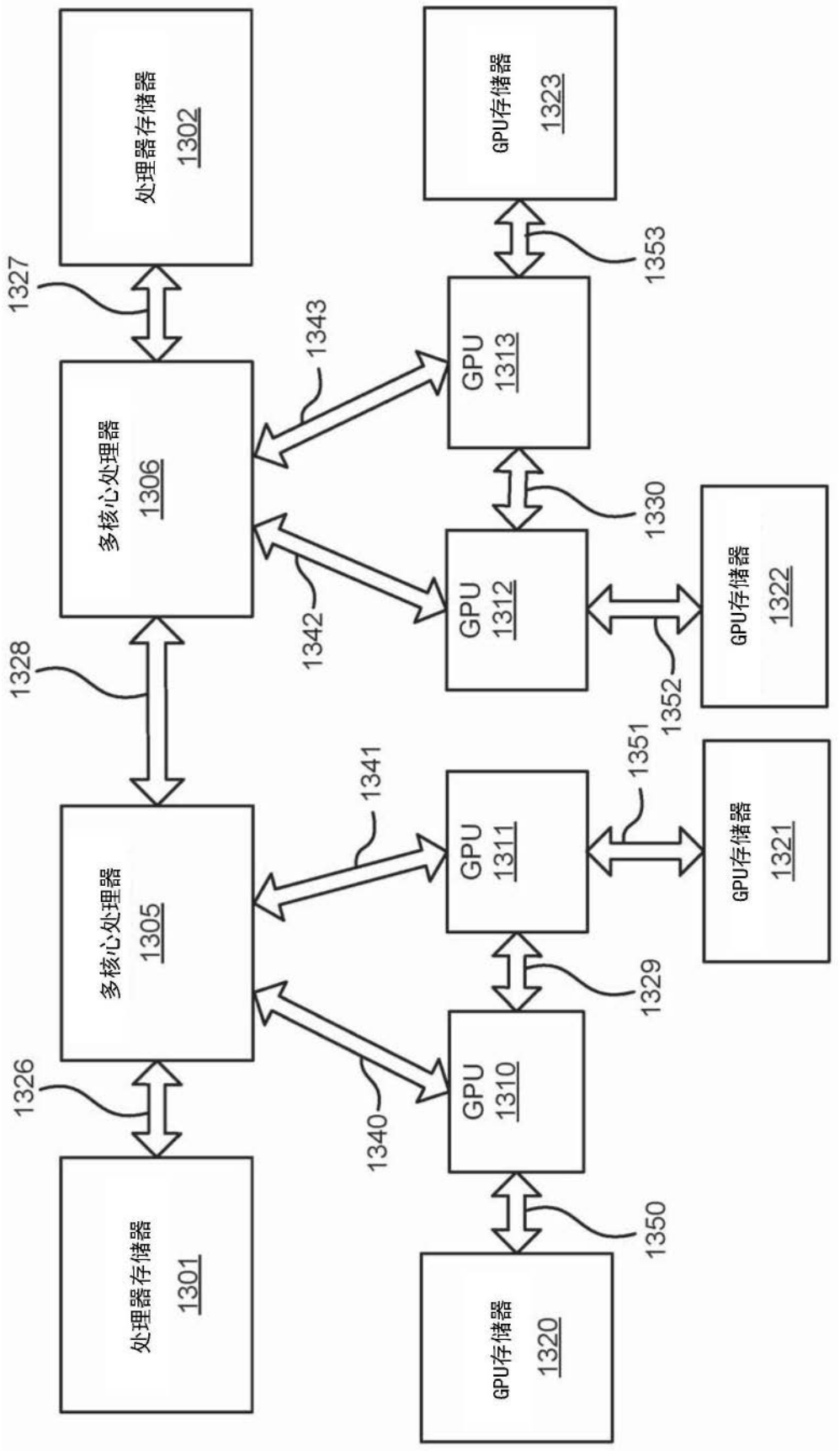

处理器存储器
1302
GPU存储器
1323
1327
1353
1343
多核心处理器
1306
GPU
1313
1330
1328
1342
GPU
1312
GPU存储器
1322
1352
1341
1351
多核心处理器
1305
GPU
1311
GPU存储器
1321
1329
1326
1340
GPU
1310
处理器存储器
1301
1350
GPU存储器
1320

图13A

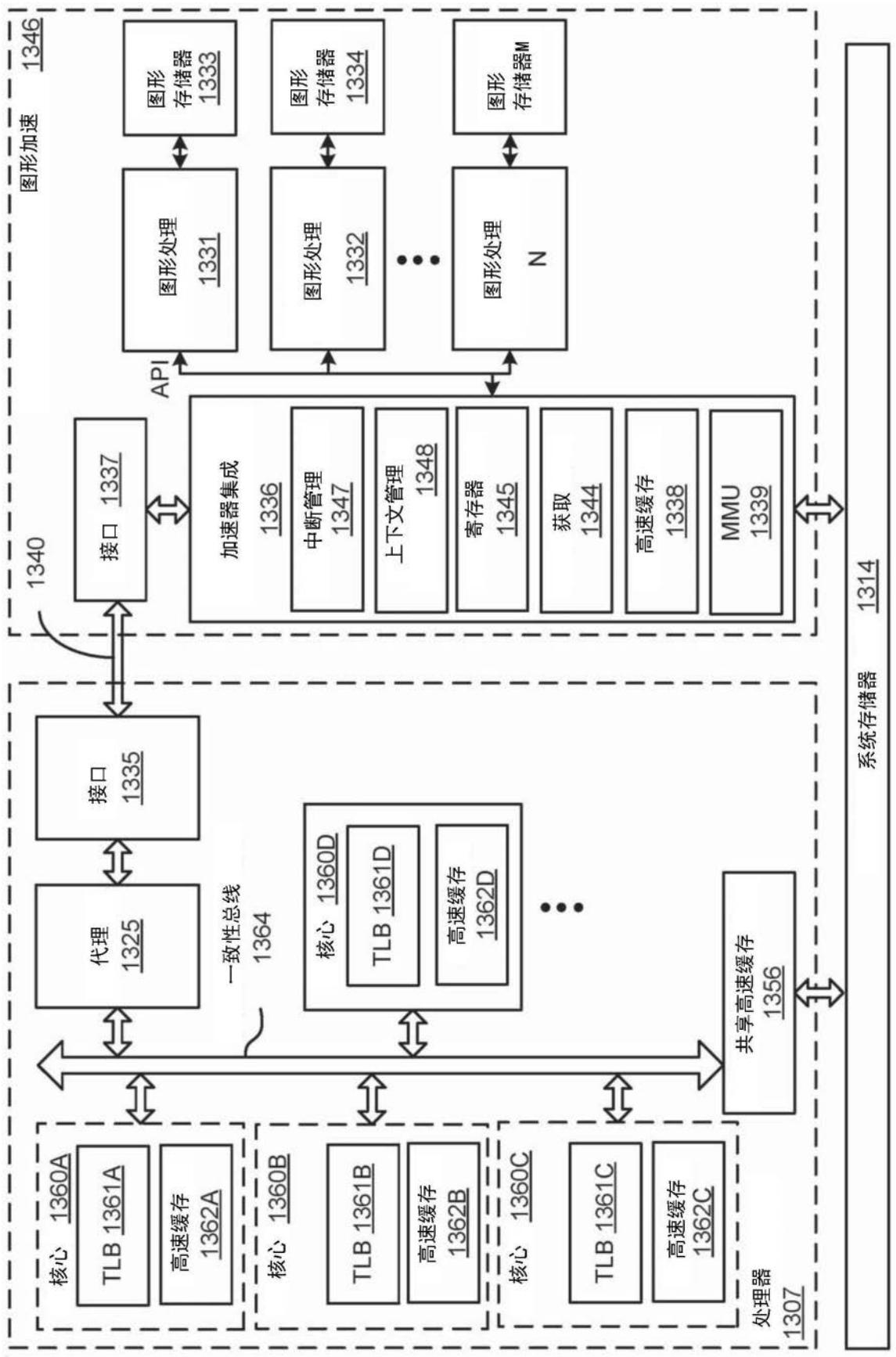

图形加速 1346
图形处理 1331
图形处理 1332
图形处理 N
图形存储器 1333
图形存储器 1334
图形存储器M
API
1340
接口 1337
加速器集成 1336
中断管理 1347
上下文管理 1348
寄存器 1345
获取 1344
高速缓存 1338
MMU 1339
接口 1335
代理 1325
一致性总线 1364
核心 1360D
TLB 1361D
高速缓存 1362D
共享高速缓存 1356
系统存储器 1314
核心 1360A
TLB 1361A
高速缓存 1362A
核心 1360B
TLB 1361B
高速缓存 1362B
核心 1360C
TLB 1361C
高速缓存 1362C
处理器 1307

图13B

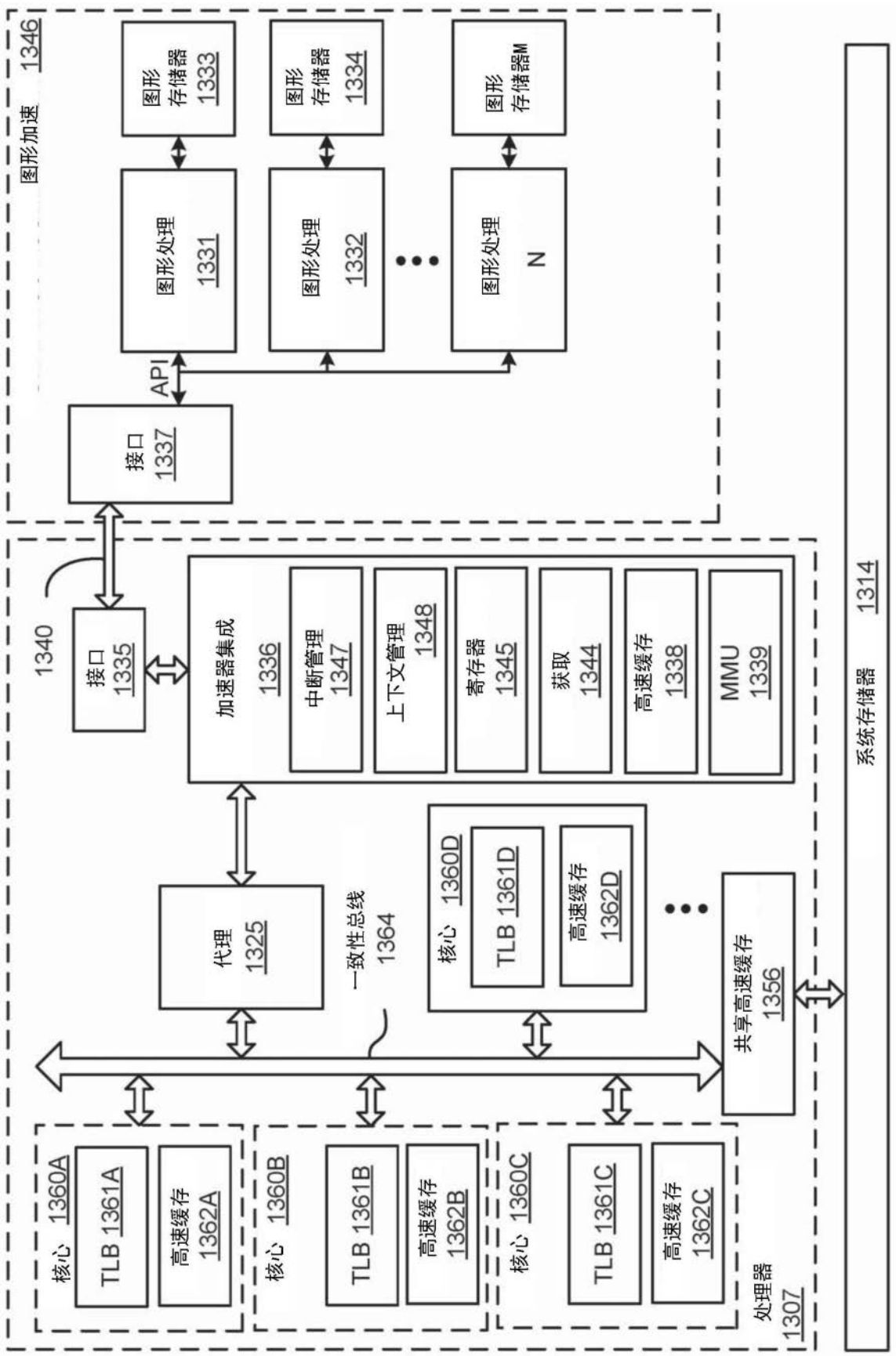

图形加速 1346
图形存储器 1333
图形存储器 1334
图形存储器M
图形处理 1331
图形处理 1332
图形处理 N
API
接口 1337
1340
接口 1335
加速器集成 1336
中断管理 1347
上下文管理 1348
寄存器 1345
获取 1344
高速缓存 1338
MMU 1339
系统存储器 1314
代理 1325
一致性总线 1364
核心 1360D
TLB 1361D
高速缓存 1362D
共享高速缓存 1356
核心 1360A
TLB 1361A
高速缓存 1362A
核心 1360B
TLB 1361B
高速缓存 1362B
核心 1360C
TLB 1361C
高速缓存 1362C
处理器 1307

图13C

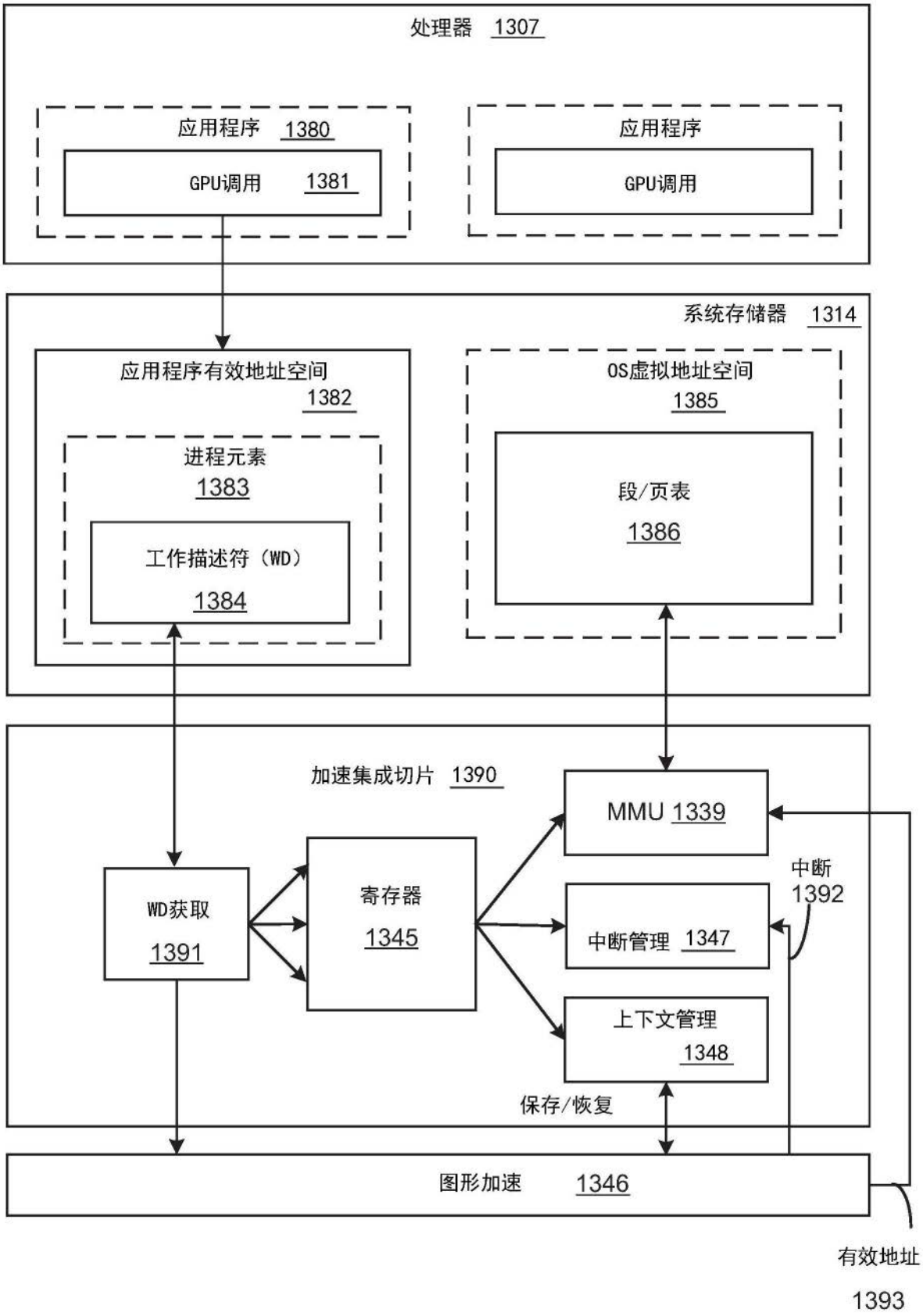

处理器 1307
应用程序 1380
GPU调用 1381
应用程序
GPU调用
系统存储器 1314
应用程序有效地址空间
1382
进程元素
1383
工作描述符（WD）
1384
OS虚拟地址空间
1385
段/页表
1386
加速集成切片 1390
MMU 1339
WD获取
1391
寄存器
1345
中断
1392
中断管理 1347
上下文管理
1348
保存/恢复
图形加速 1346
有效地址
1393

图13D

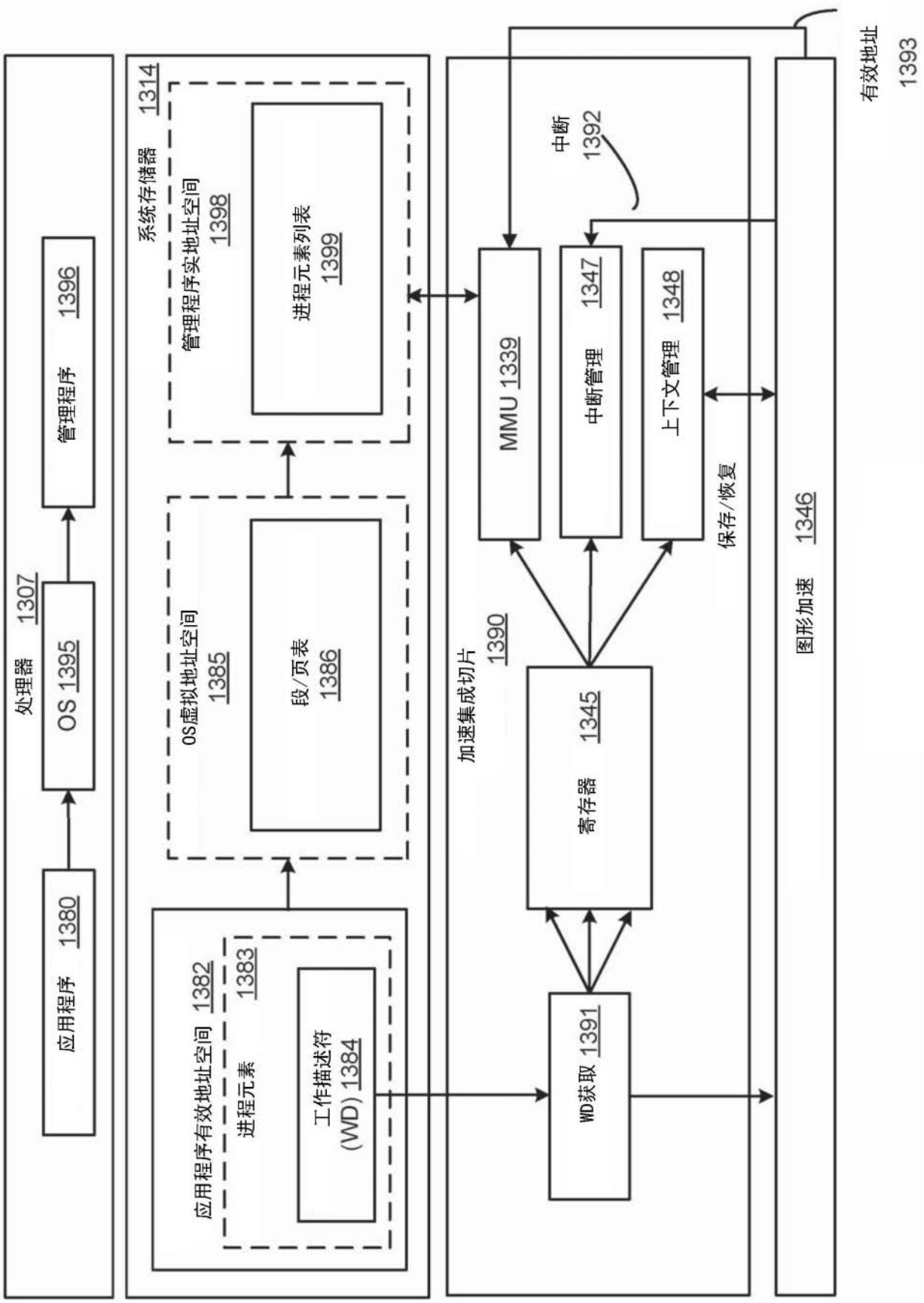

处理器 1307
应用程序 1380
OS 1395
管理程序 1396
系统存储器 1314
管理程序实地址空间 1398
进程元素列表 1399
OS虚拟地址空间 1385
段/页表 1386
应用程序有效地址空间 1382
进程元素 1383
工作描述符 (WD) 1384
加速集成切片 1390
MMU 1339
中断管理 1347
上下文管理 1348
寄存器 1345
WD获取 1391
保存/恢复
图形加速 1346
中断 1392
有效地址 1393

图13E

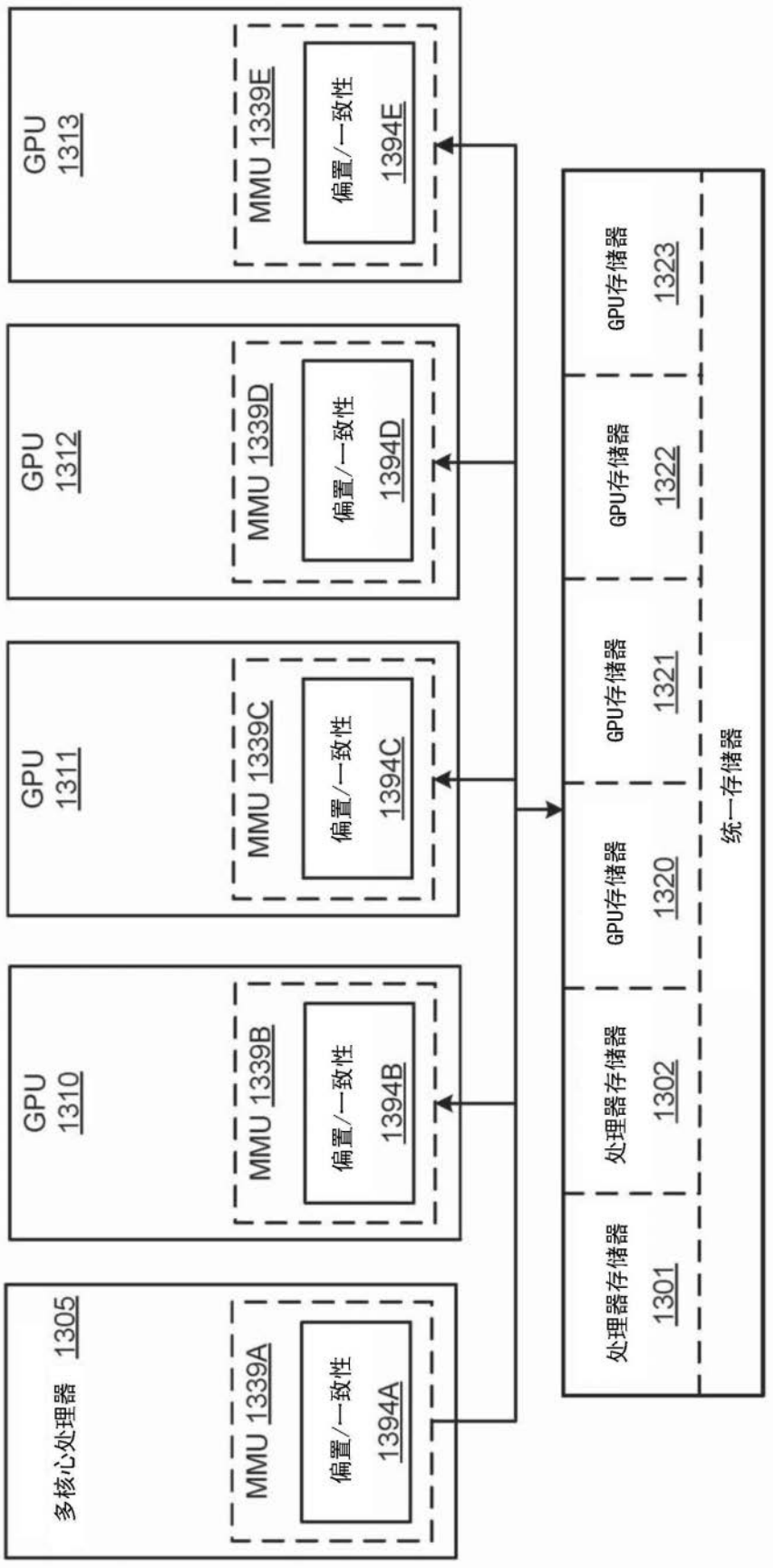

多核心处理器 1305
MMU 1339A
偏置/一致性 1394A
GPU 1310
MMU 1339B
偏置/一致性 1394B
GPU 1311
MMU 1339C
偏置/一致性 1394C
GPU 1312
MMU 1339D
偏置/一致性 1394D
GPU 1313
MMU 1339E
偏置/一致性 1394E
处理器存储器 1301
处理器存储器 1302
GPU存储器 1320
GPU存储器 1321
GPU存储器 1322
GPU存储器 1323
统一存储器

图13F

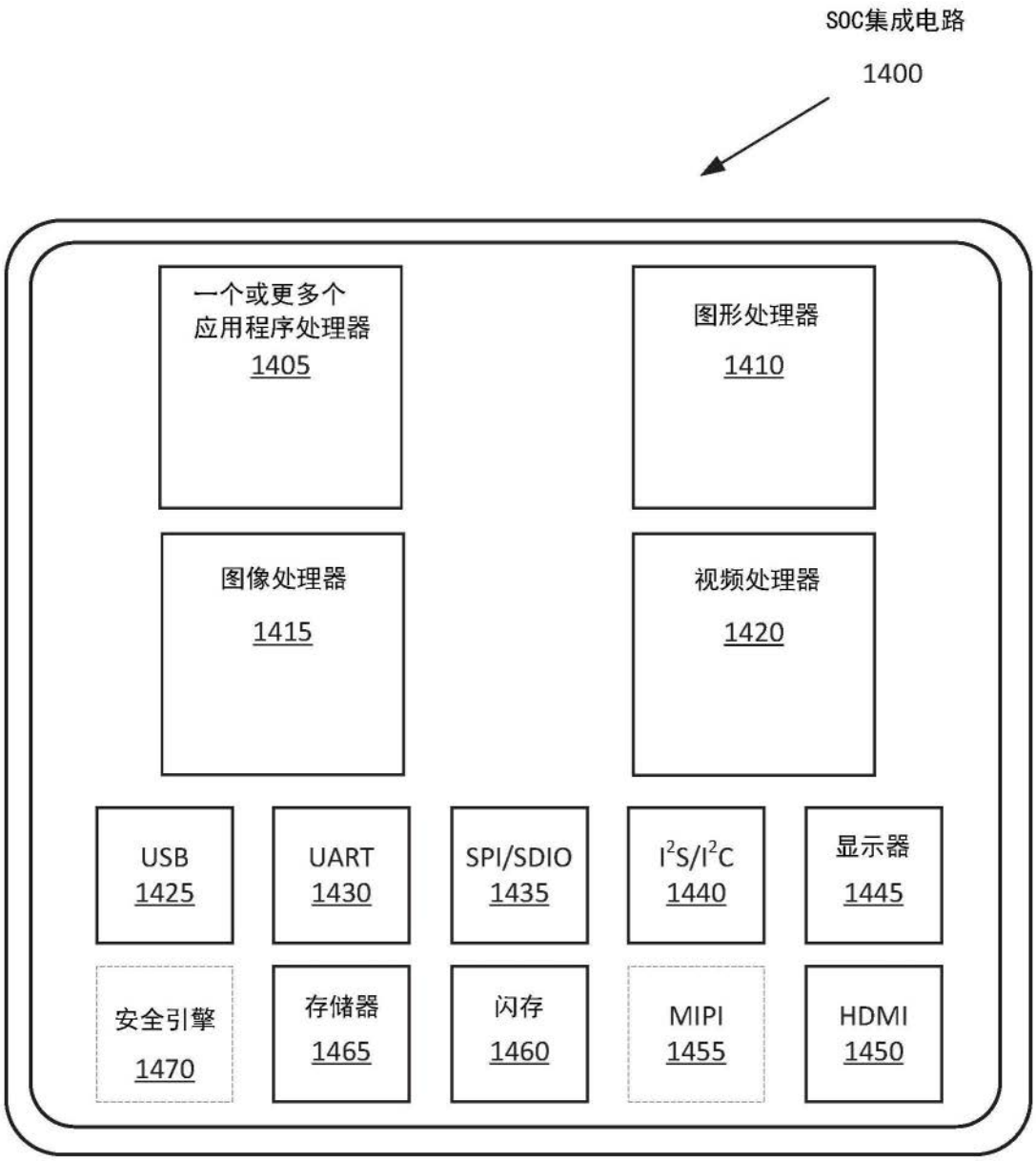

SOC集成电路
1400
一个或更多个
应用程序处理器
1405
图形处理器
1410
图像处理器
1415
视频处理器
1420
USB
1425
UART
1430
SPI/SDIO
1435
I2S/I2C
1440
显示器
1445
安全引擎
1470
存储器
1465
闪存
1460
MIPI
1455
HDMI
1450

图14


图形处理器
1510
顶点处理器
1505
片段处理器
1515A
片段处理器
1515C
片段处理器
1515N-1
片段处理器
1515B
片段处理器
1515D
片段处理器
1515N
MMU
1520A
MMU
1520B
高速缓存
1525A
高速缓存
1525B
互连
1530A
互连
1530B

图15A

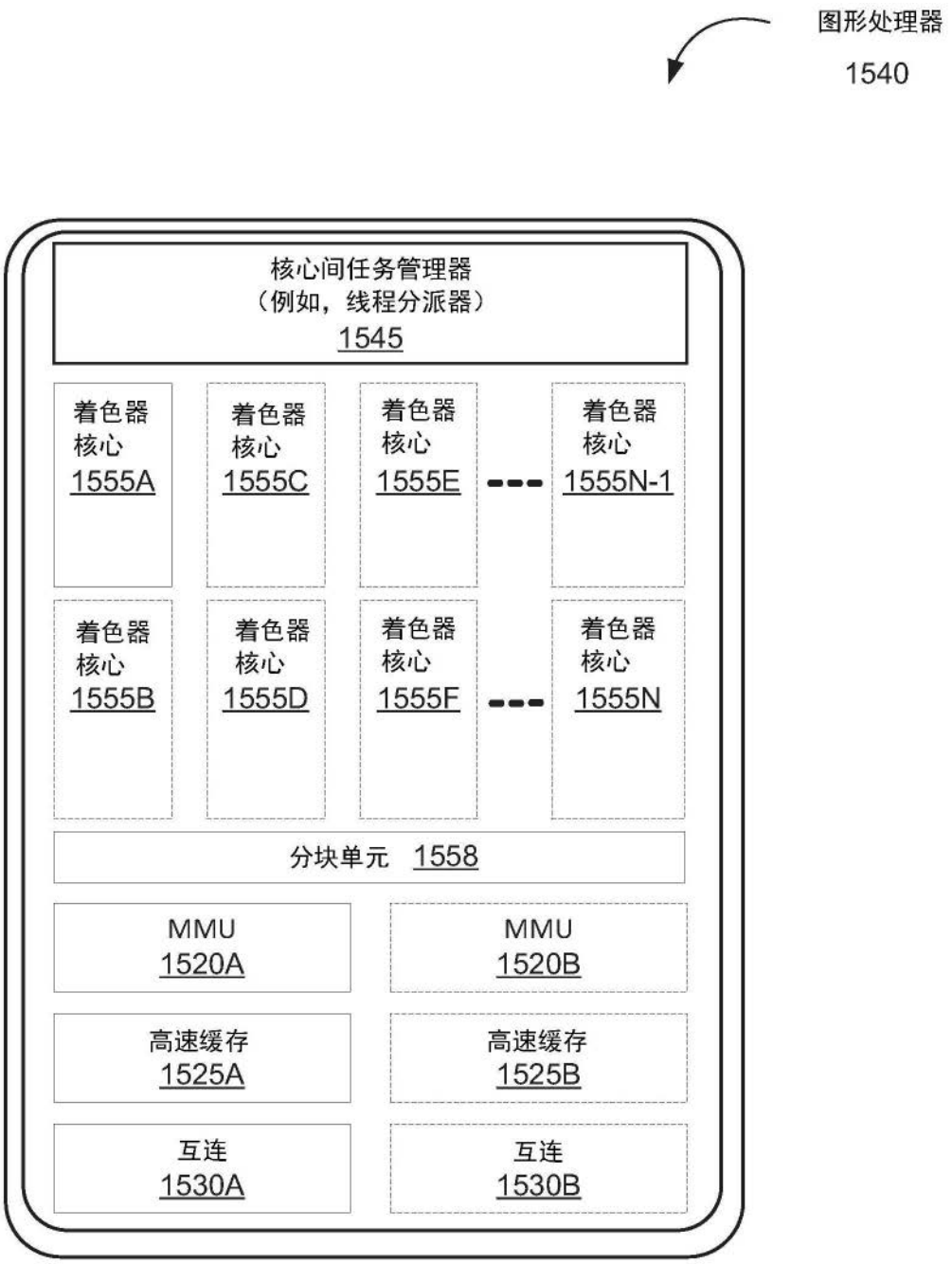

图形处理器
1540
核心间任务管理器
（例如，线程分派器）
1545
着色器核心 1555A
着色器核心 1555C
着色器核心 1555E
着色器核心 1555N-1
着色器核心 1555B
着色器核心 1555D
着色器核心 1555F
着色器核心 1555N
分块单元 1558
MMU 1520A
MMU 1520B
高速缓存 1525A
高速缓存 1525B
互连 1530A
互连 1530B

图15B

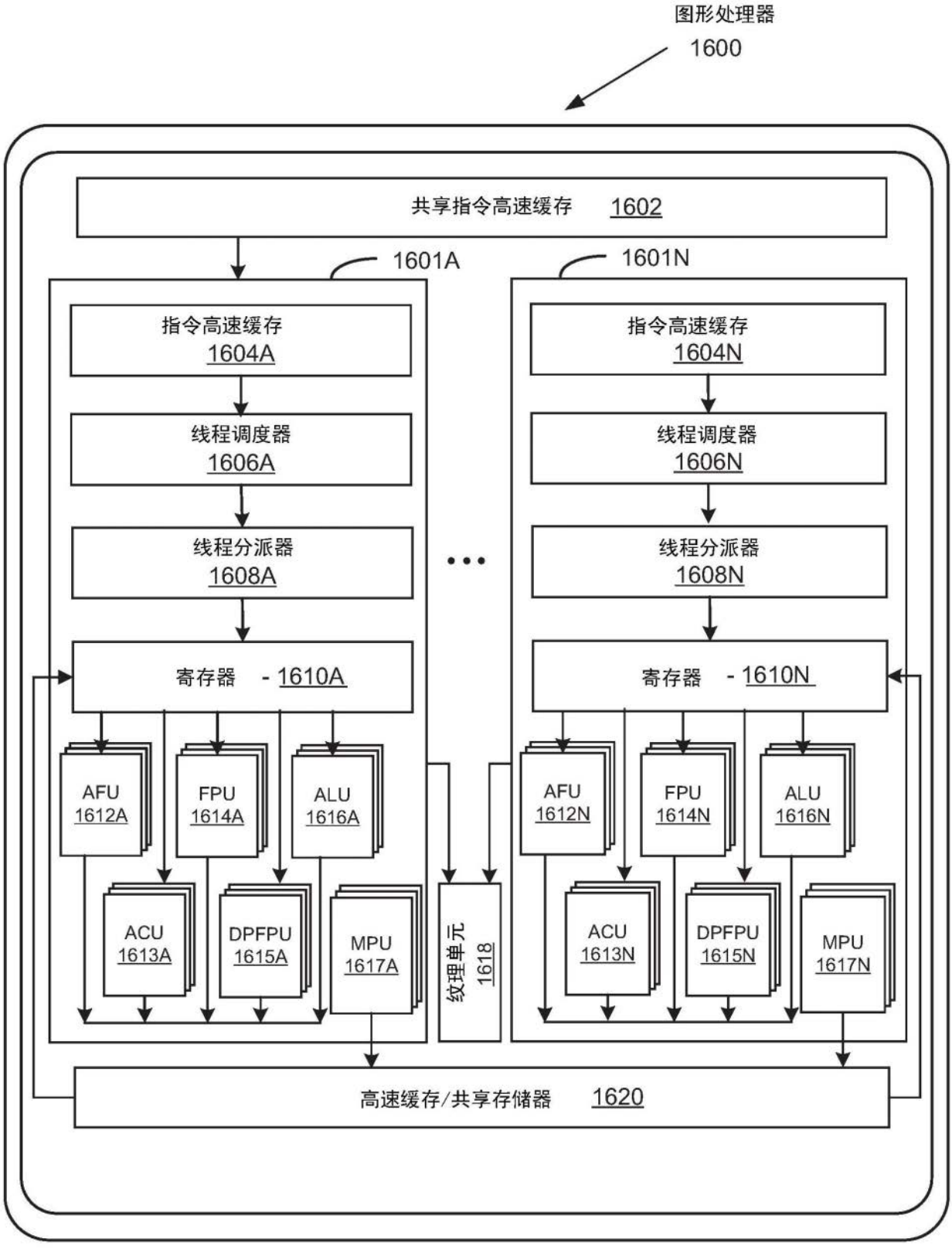

图形处理器
1600
共享指令高速缓存 1602
1601A
1601N
指令高速缓存 1604A
线程调度器 1606A
线程分派器 1608A
寄存器 - 1610A
AFU 1612A
FPU 1614A
ALU 1616A
ACU 1613A
DPFPU 1615A
MPU 1617A
纹理单元 1618
指令高速缓存 1604N
线程调度器 1606N
线程分派器 1608N
寄存器 - 1610N
AFU 1612N
FPU 1614N
ALU 1616N
ACU 1613N
DPFPU 1615N
MPU 1617N
高速缓存/共享存储器 1620

图16A

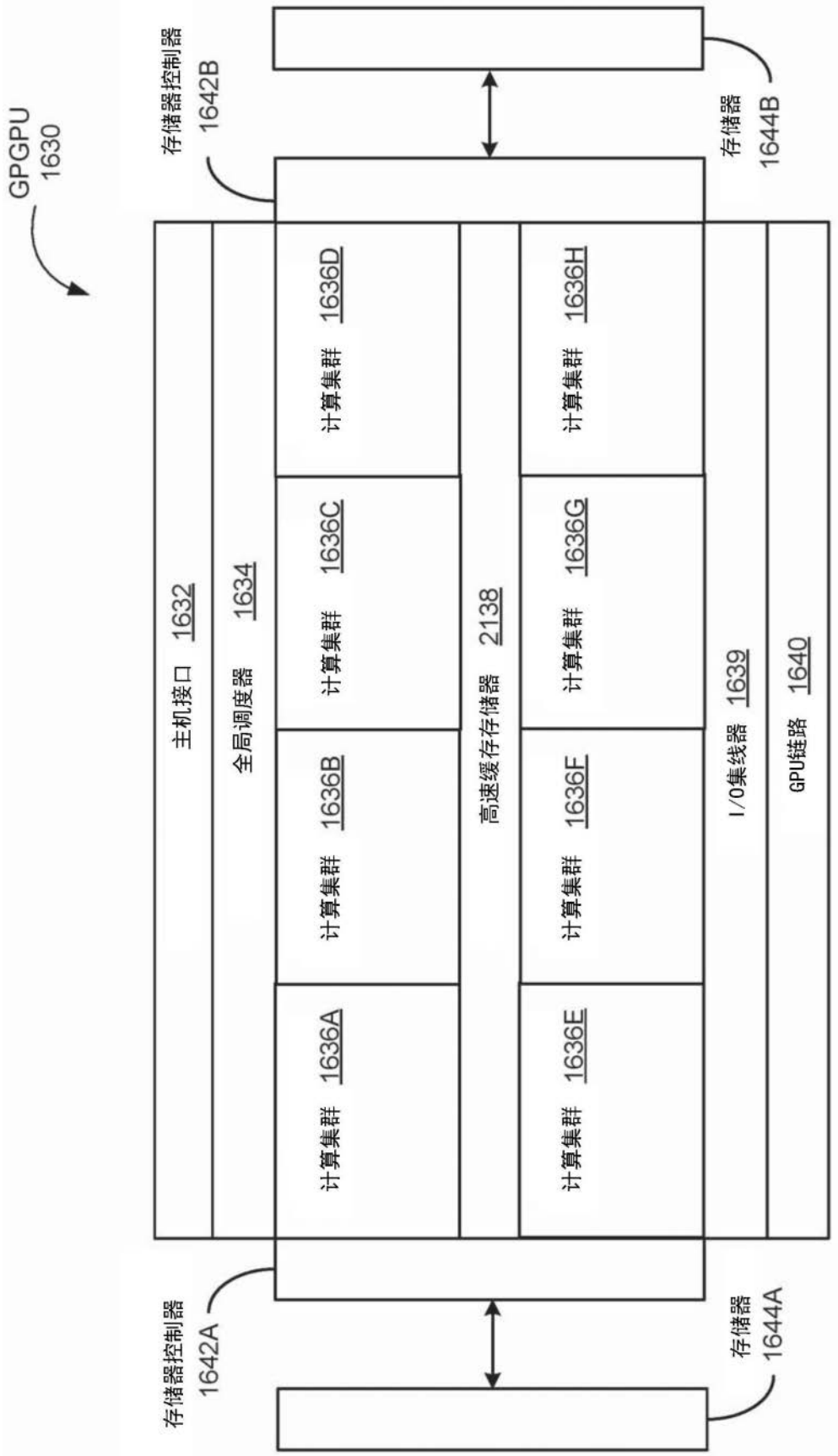

GPGPU
1630
存储器控制器
1642B
存储器
1644B
主机接口 1632
全局调度器 1634
计算集群 1636A
计算集群 1636B
计算集群 1636C
计算集群 1636D
高速缓存存储器 2138
计算集群 1636E
计算集群 1636F
计算集群 1636G
计算集群 1636H
I/O集线器 1639
GPU链路 1640
存储器控制器
1642A
存储器
1644A

图16B

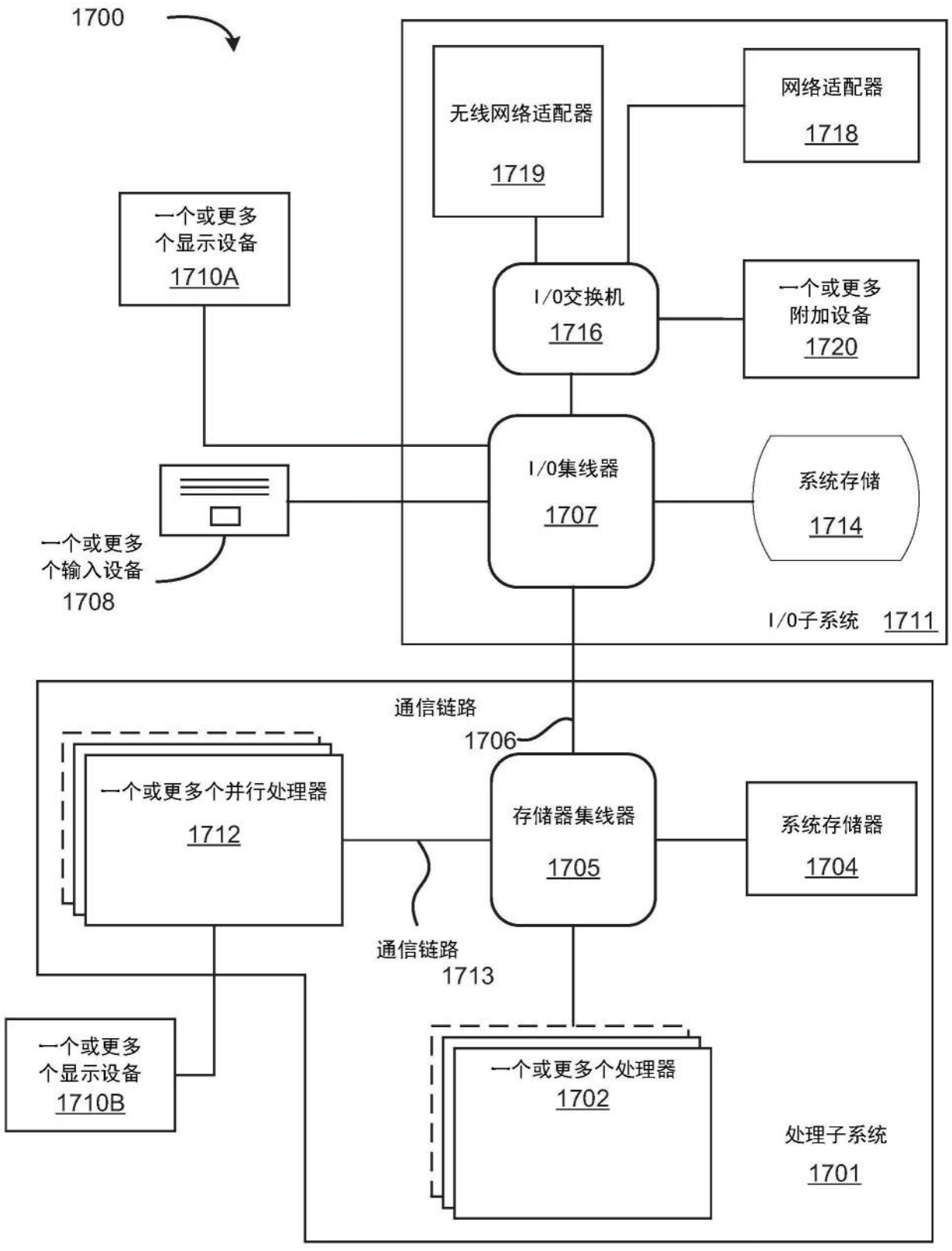

1700
无线网络适配器
1719
网络适配器
1718
一个或更多
个显示设备
1710A
I/O交换机
1716
一个或更多
附加设备
1720
I/O集线器
1707
系统存储
1714
一个或更多
个输入设备
1708
I/O子系统 1711
通信链路
1706
一个或更多个并行处理器
1712
存储器集线器
1705
系统存储器
1704
通信链路
1713
一个或更多
个显示设备
1710B
一个或更多个处理器
1702
处理子系统
1701

图17

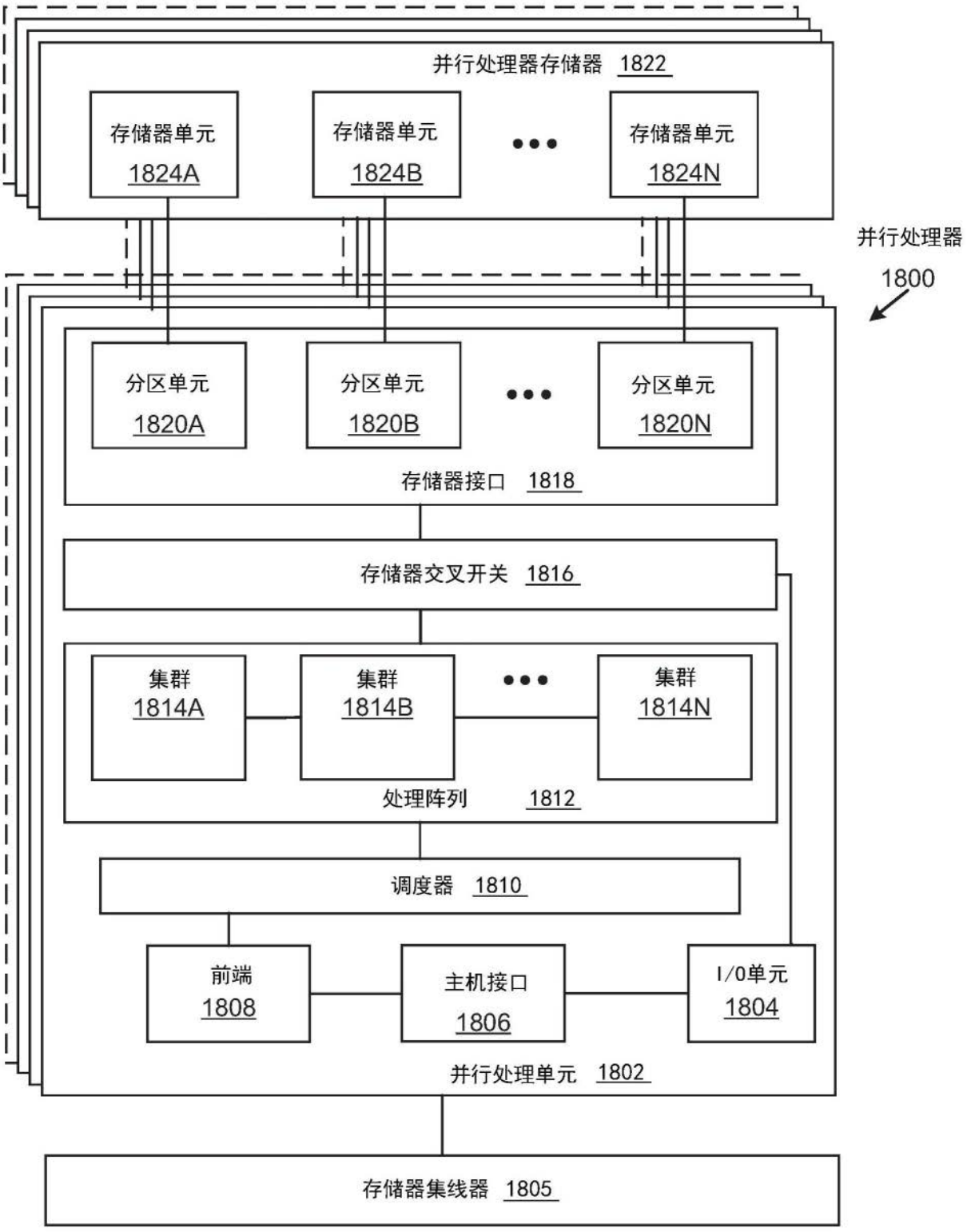

并行处理器存储器 1822
存储器单元 1824A
存储器单元 1824B
存储器单元 1824N
并行处理器 1800
分区单元 1820A
分区单元 1820B
分区单元 1820N
存储器接口 1818
存储器交叉开关 1816
集群 1814A
集群 1814B
集群 1814N
处理阵列 1812
调度器 1810
前端 1808
主机接口 1806
I/O单元 1804
并行处理单元 1802
存储器集线器 1805

图18A

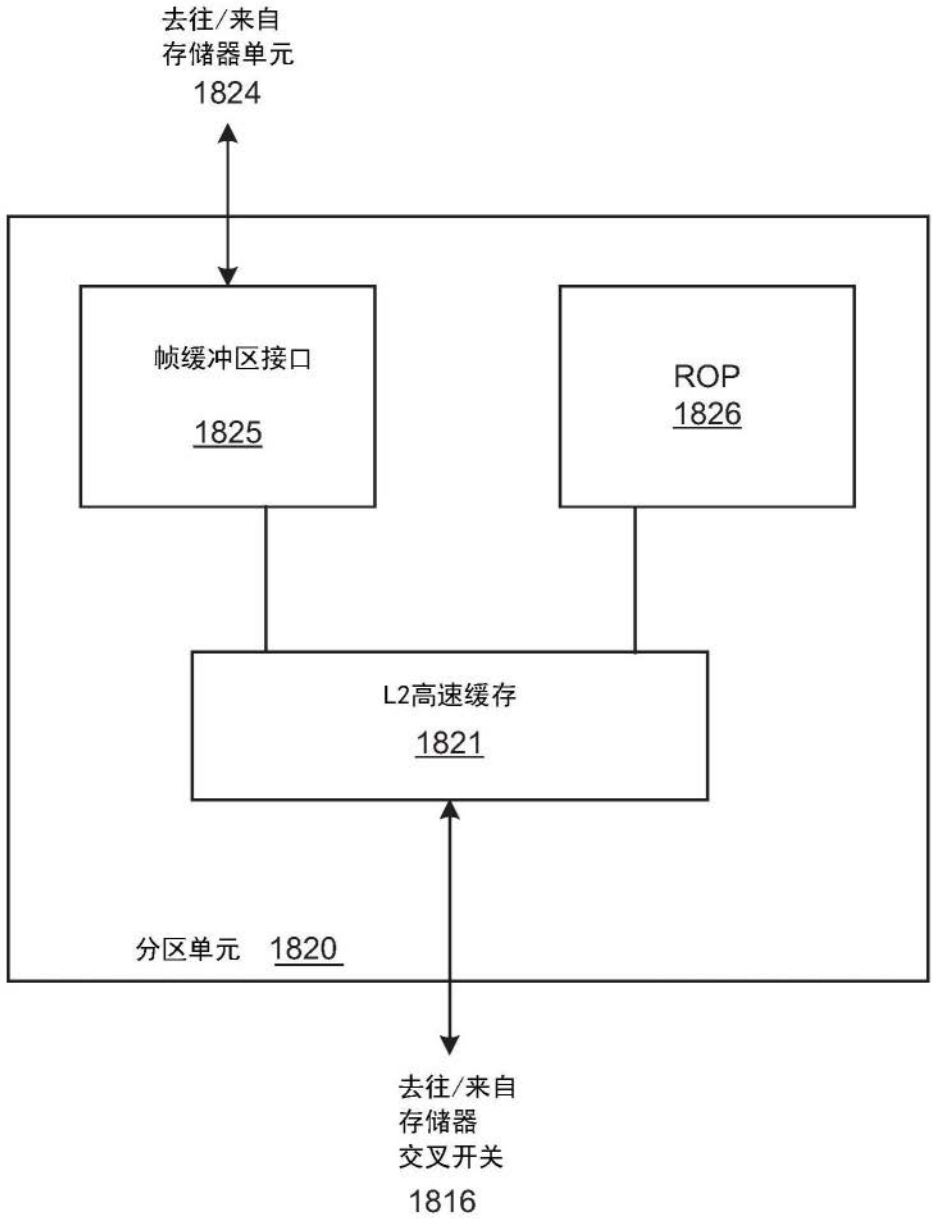

去往/来自
存储器单元
1824
帧缓冲区接口
1825
ROP
1826
L2高速缓存
1821
分区单元 1820
去往/来自
存储器
交叉开关
1816

图18B

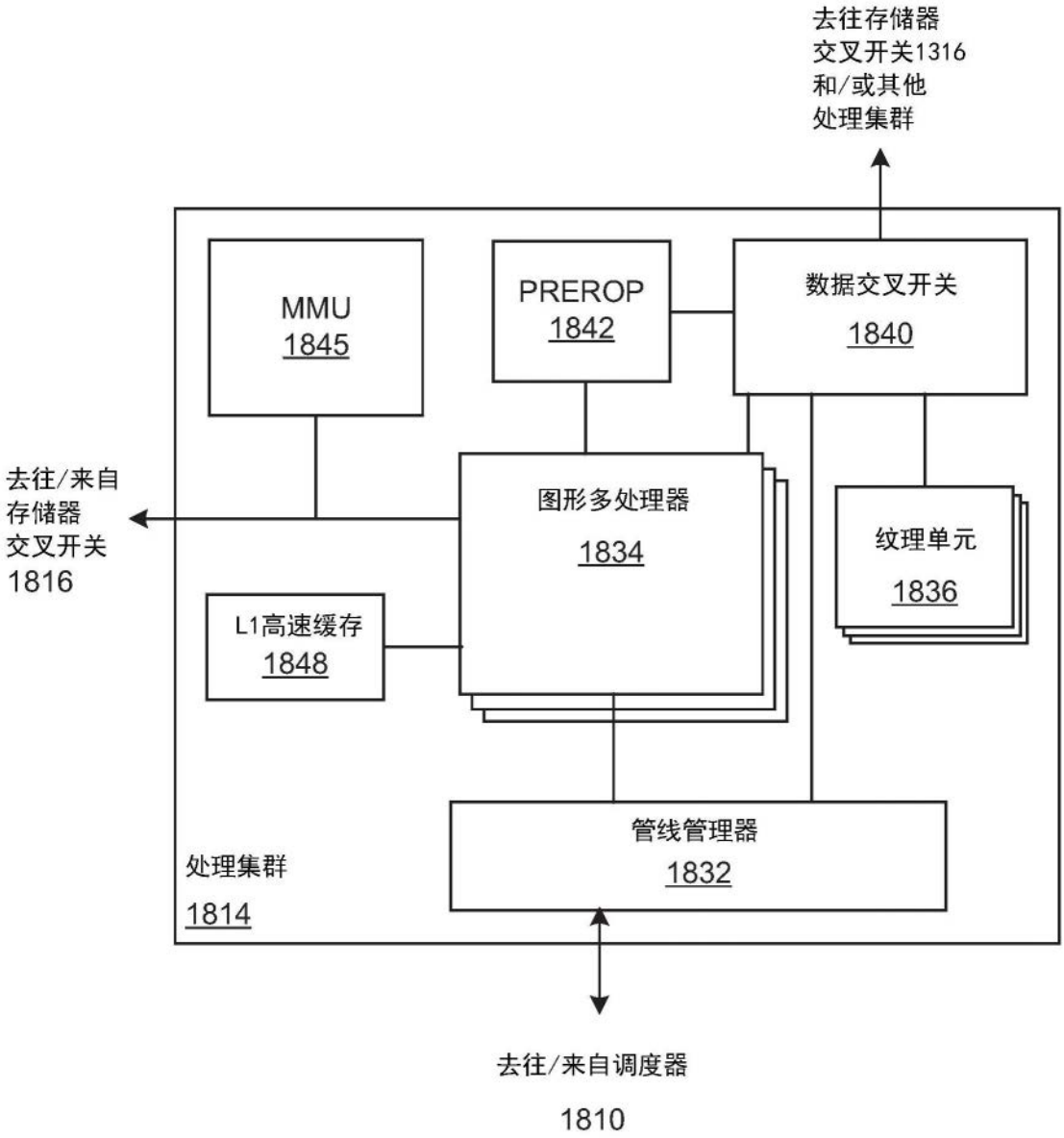

去往存储器
交叉开关1316
和/或其他
处理集群
MMU
1845
PREROP
1842
数据交叉开关
1840
去往/来自
存储器
交叉开关
1816
图形多处理器
1834
纹理单元
1836
L1高速缓存
1848
管线管理器
1832
处理集群
1814
去往/来自调度器
1810

图18C

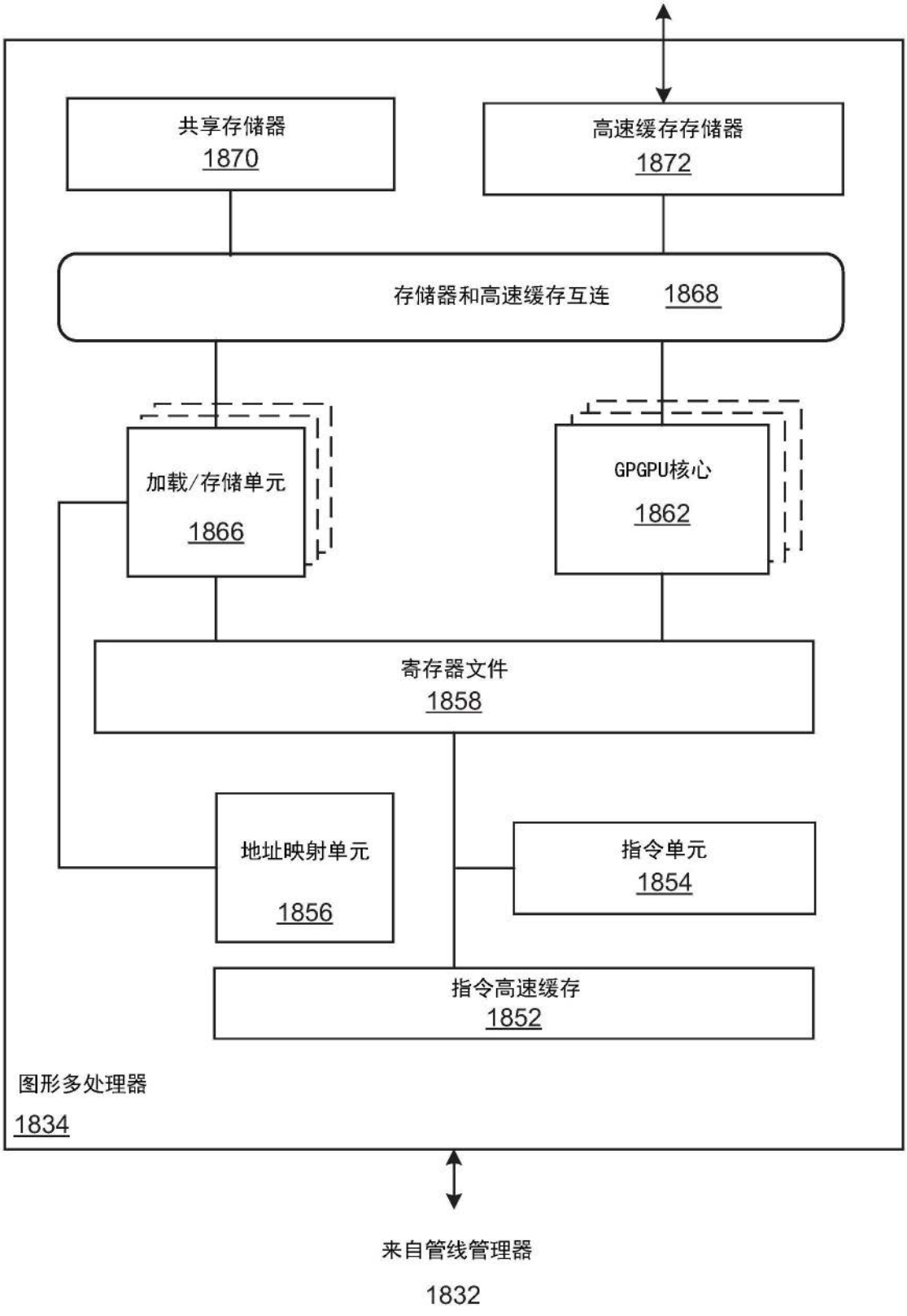

共享存储器
1870
高速缓存存储器
1872
存储器和高速缓存互连 1868
加载/存储单元
1866
GPGPU核心
1862
寄存器文件
1858
地址映射单元
1856
指令单元
1854
指令高速缓存
1852
图形多处理器
1834
来自管线管理器
1832

图18D

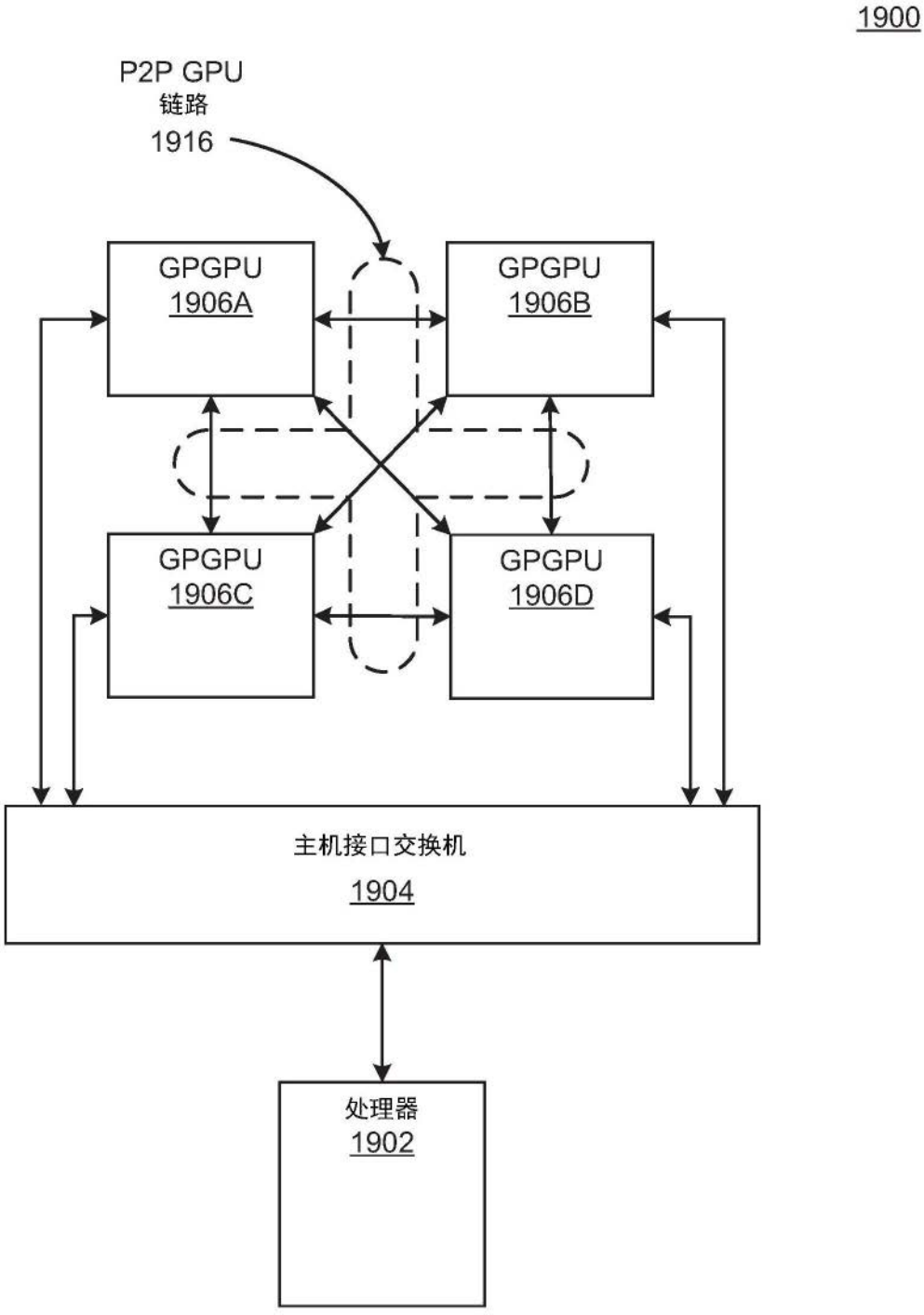

1900
P2P GPU
链路
1916
GPGPU
1906A
GPGPU
1906B
GPGPU
1906C
GPGPU
1906D
主机接口交换机
1904
处理器
1902

图19

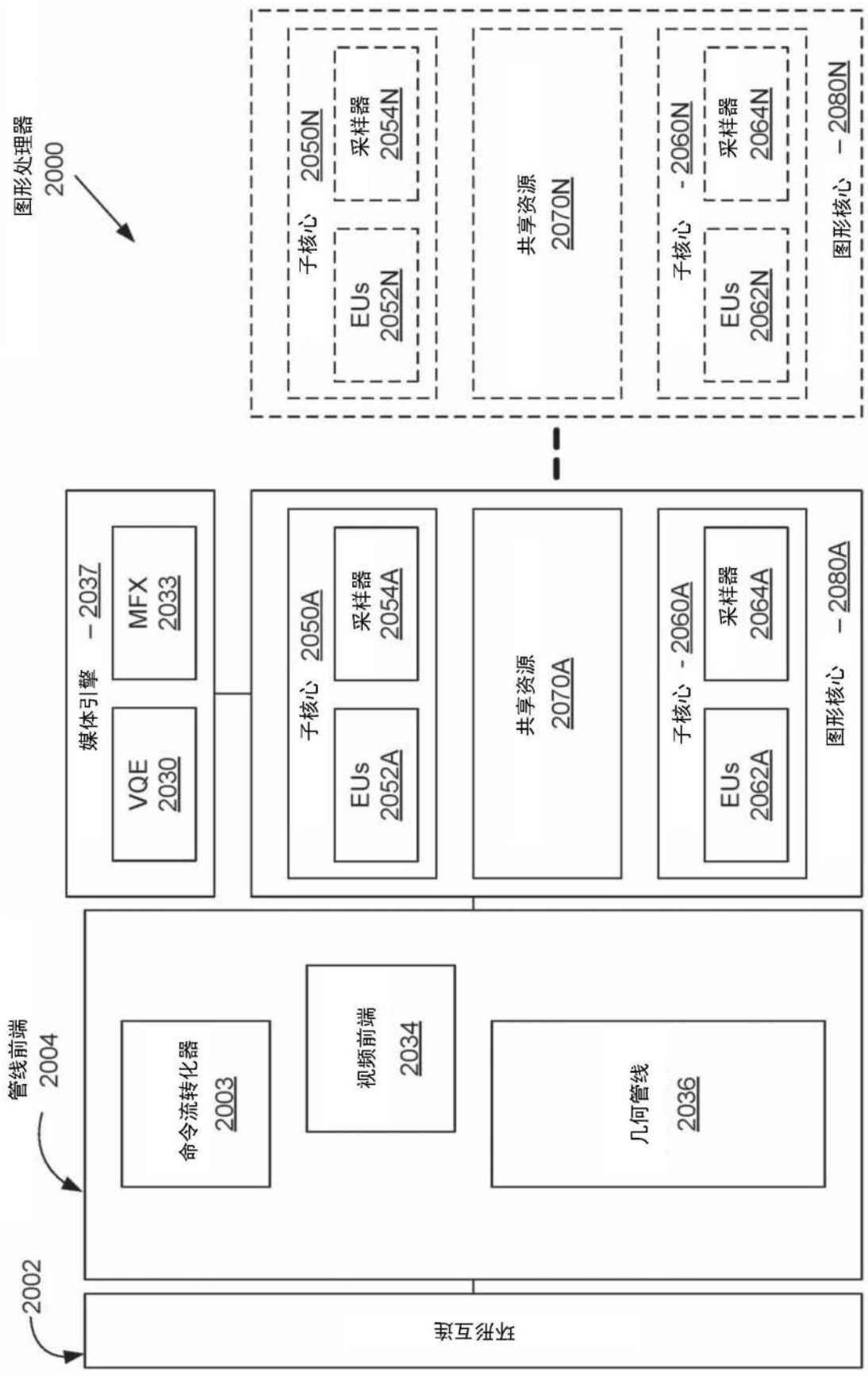

图形处理器
2000
环形互连
2002
管线前端
2004
命令流转化器
2003
视频前端
2034
几何管线
2036
媒体引擎
2037
VQE
2030
MFX
2033
子核心
2050A
EUs
2052A
采样器
2054A
共享资源
2070A
子核心
2060A
EUs
2062A
采样器
2064A
图形核心
2080A
子核心
2050N
EUs
2052N
采样器
2054N
共享资源
2070N
子核心
2060N
EUs
2062N
采样器
2064N
图形核心
2080N

图20

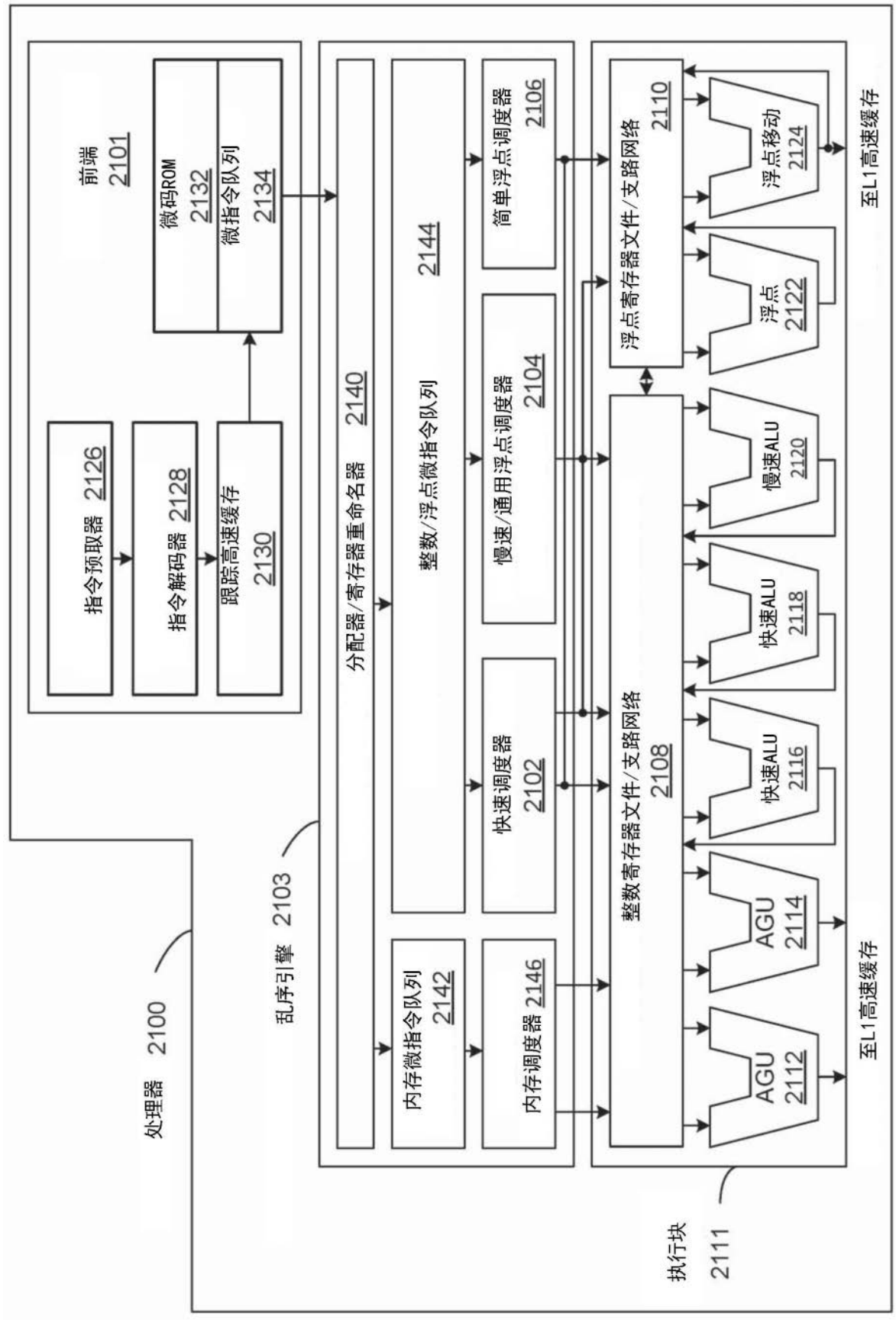

处理器 2100
乱序引擎 2103
前端 2101
指令预取器 2126
指令解码器 2128
跟踪高速缓存 2130
微码ROM 2132
微指令队列 2134
分配器/寄存器重命名器 2140
整数/浮点微指令队列 2144
简单浮点调度器 2106
慢速/通用浮点调度器 2104
快速调度器 2102
内存微指令队列 2142
内存调度器 2146
浮点寄存器文件/支路网络 2110
整数寄存器文件/支路网络 2108
浮点移动 2124
浮点 2122
慢速ALU 2120
快速ALU 2118
快速ALU 2116
AGU 2114
AGU 2112
至L1高速缓存
至L1高速缓存
执行块 2111

图21

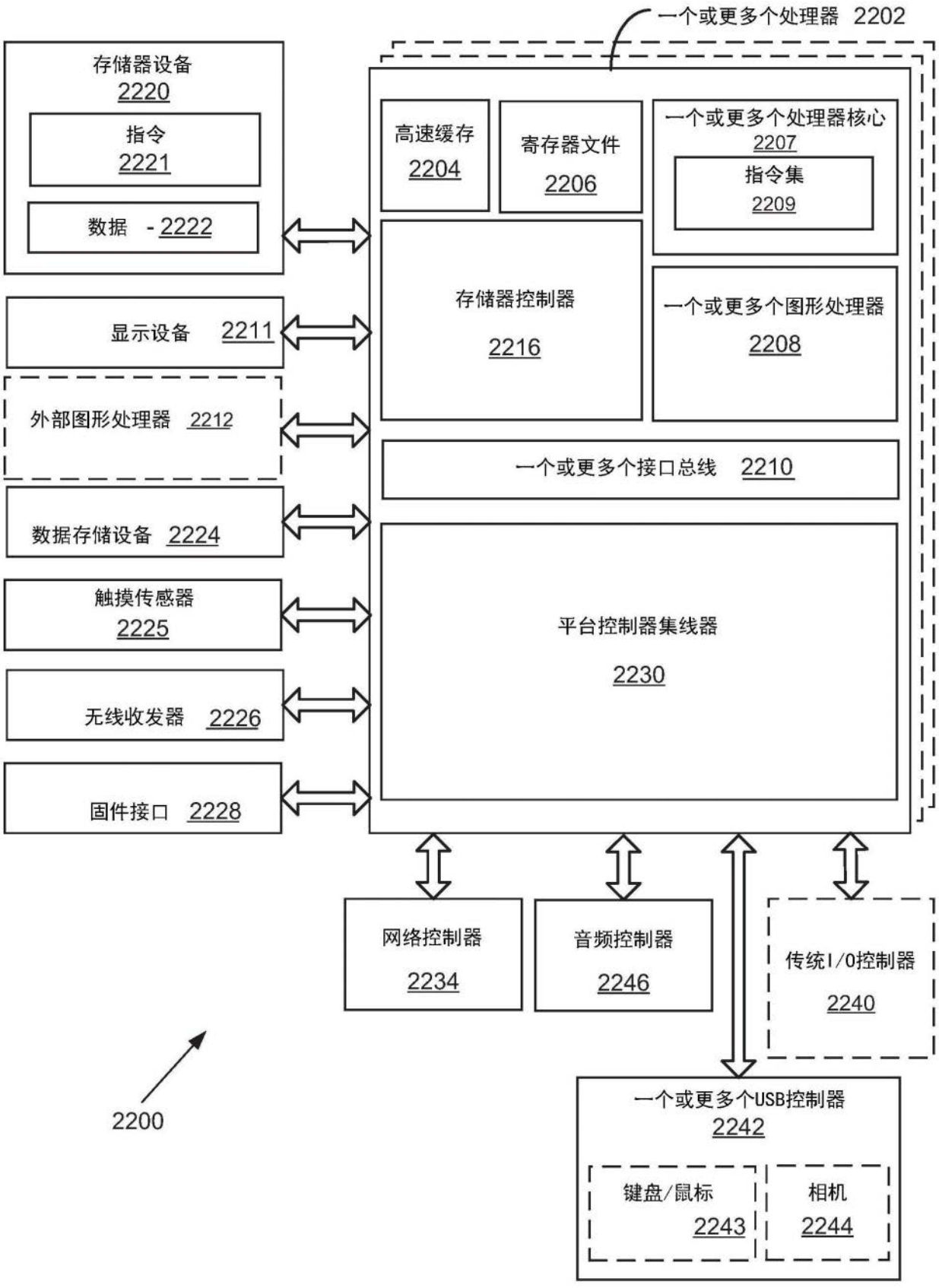

一个或更多个处理器 2202
存储器设备
2220
指令
2221
数据 - 2222
显示设备 2211
外部图形处理器 2212
数据存储设备 2224
触摸传感器
2225
无线收发器 2226
固件接口 2228
高速缓存
2204
寄存器文件
2206
一个或更多个处理器核心
2207
指令集
2209
存储器控制器
2216
一个或更多个图形处理器
2208
一个或更多个接口总线 2210
平台控制器集线器
2230
网络控制器
2234
音频控制器
2246
传统I/O控制器
2240
一个或更多个USB控制器
2242
键盘/鼠标
2243
相机
2244
2200

图22

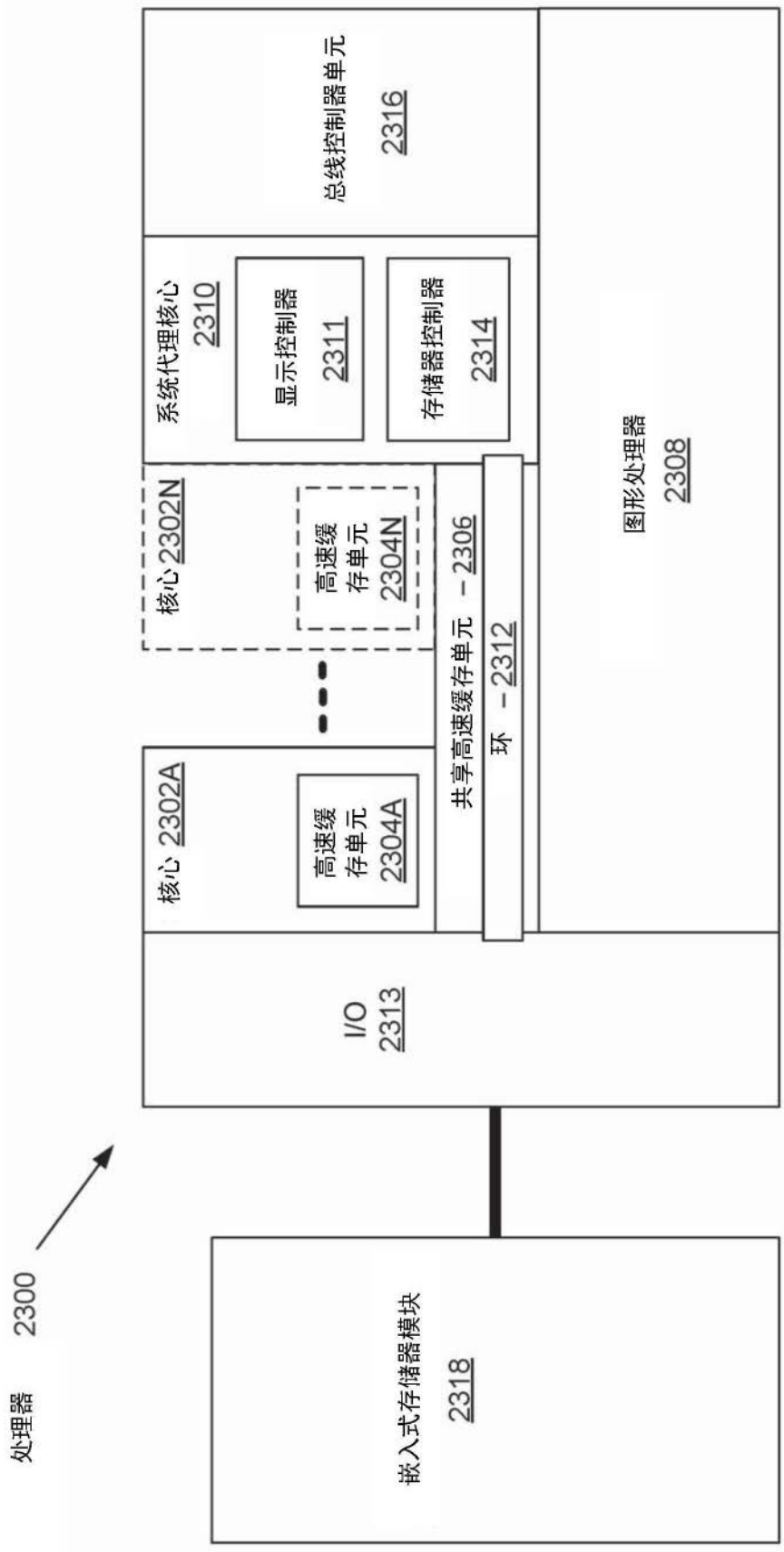

处理器 2300
嵌入式存储器模块 2318
I/O 2313
核心 2302A
高速缓存单元 2304A
核心 2302N
高速缓存单元 2304N
共享高速缓存单元 2306
环 2312
图形处理器 2308
系统代理核心 2310
显示控制器 2311
存储器控制器 2314
总线控制器单元 2316

图23

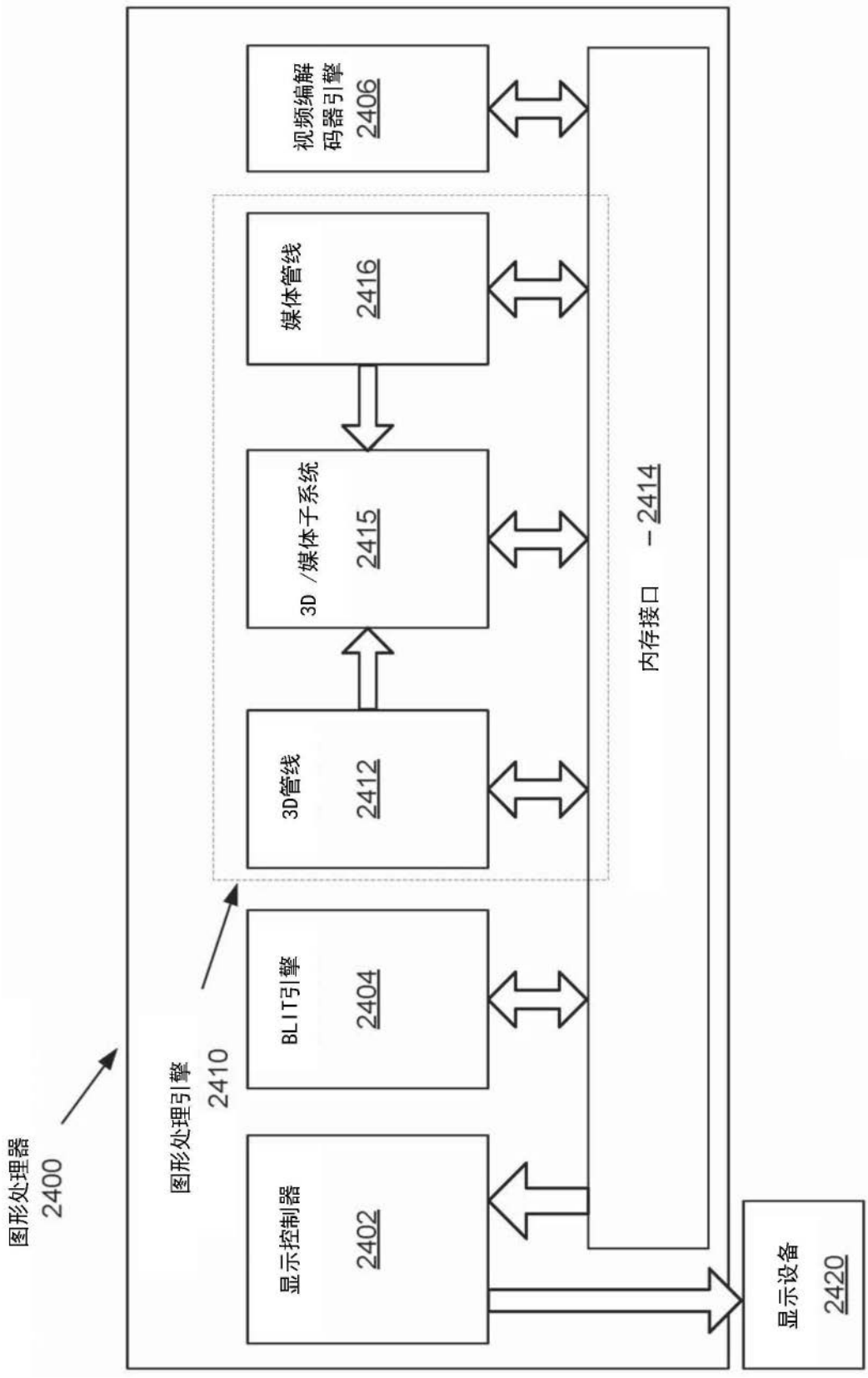

图形处理器
2400
图形处理引擎
2410
显示控制器
2402
BLIT引擎
2404
3D管线
2412
3D /媒体子系统
2415
媒体管线
2416
视频编解码器引擎
2406
内存接口
2414
显示设备
2420

图24

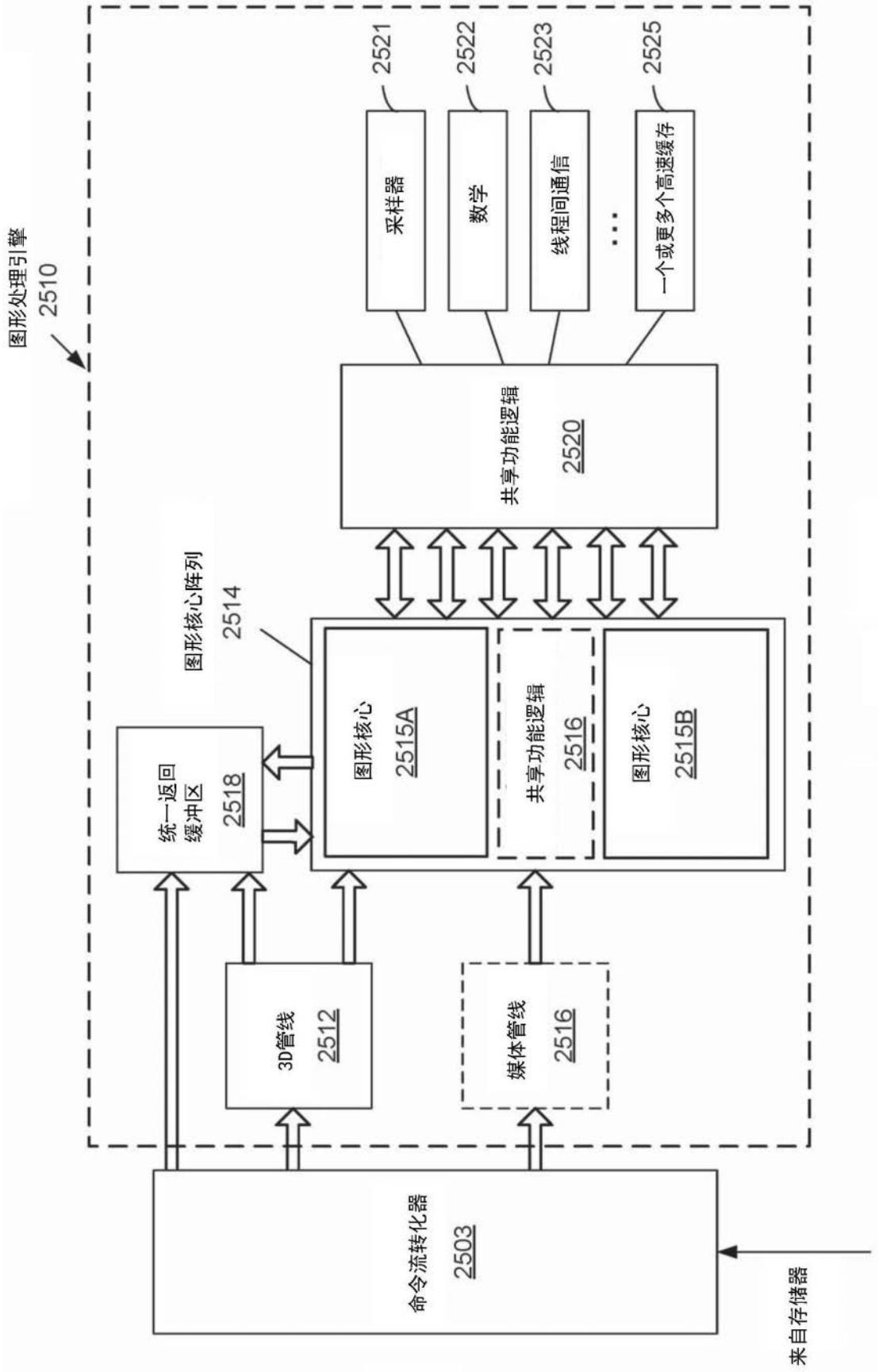

图形处理引擎
2510
采样器
2521
数学
2522
线程间通信
2523
...
一个或更多个高速缓存
2525
共享功能逻辑
2520
图形核心阵列
2514
图形核心
2515A
共享功能逻辑
2516
图形核心
2515B
统一返回
缓冲区
2518
3D管线
2512
媒体管线
2516
命令流转化器
2503
来自存储器

图25

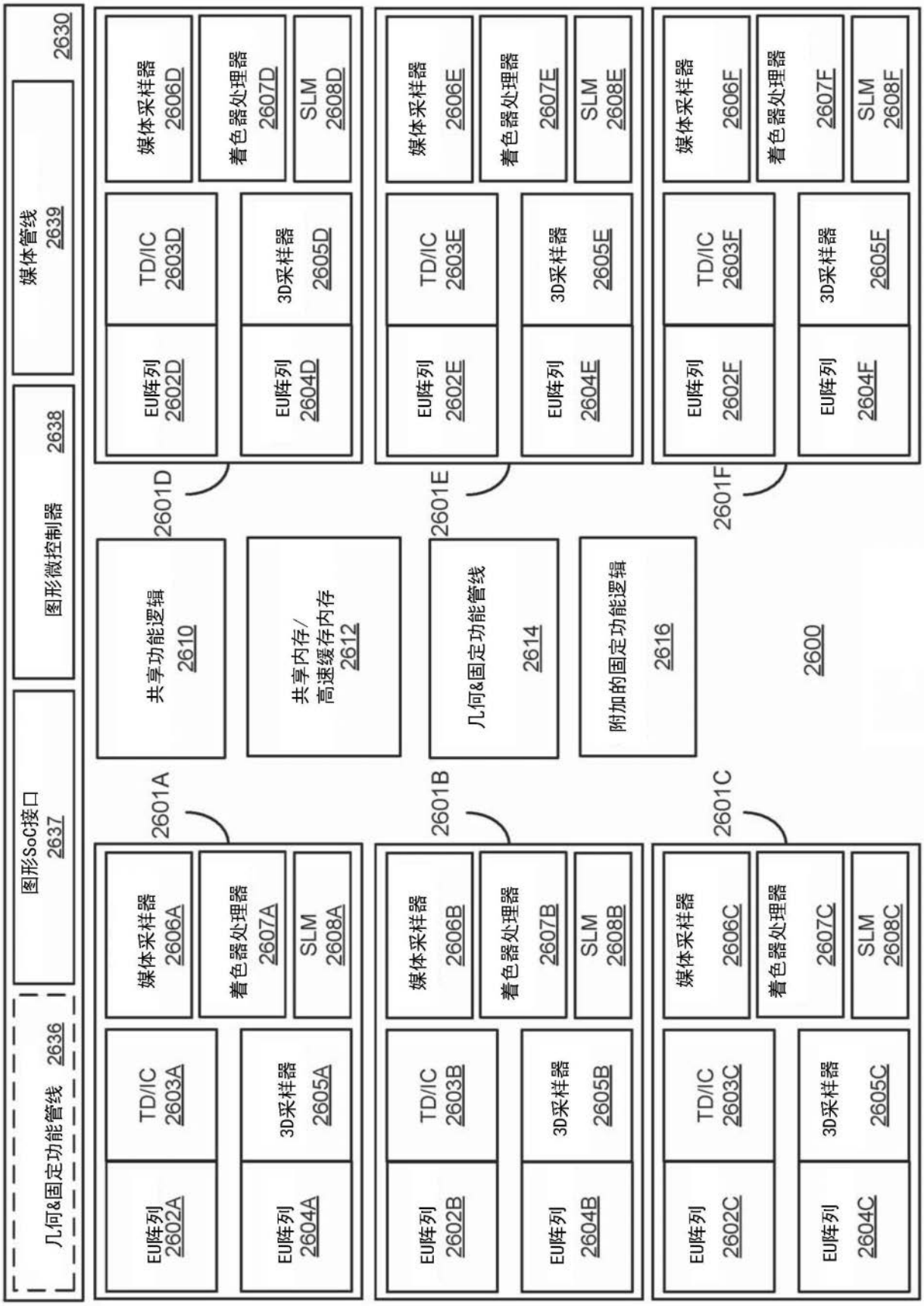

2600
2601A
2601B
2601C
2601D
2601E
2601F
几何&固定功能管线 2636
图形SoC接口 2637
图形微控制器 2638
媒体管线 2639
2630
共享功能逻辑 2610
共享内存/高速缓存内存 2612
几何&固定功能管线 2614
附加的固定功能逻辑 2616
EU阵列 2602A
TD/IC 2603A
媒体采样器 2606A
EU阵列 2604A
3D采样器 2605A
着色器处理器 2607A
SLM 2608A
EU阵列 2602B
TD/IC 2603B
媒体采样器 2606B
EU阵列 2604B
3D采样器 2605B
着色器处理器 2607B
SLM 2608B
EU阵列 2602C
TD/IC 2603C
媒体采样器 2606C
EU阵列 2604C
3D采样器 2605C
着色器处理器 2607C
SLM 2608C
EU阵列 2602D
TD/IC 2603D
媒体采样器 2606D
EU阵列 2604D
3D采样器 2605D
着色器处理器 2607D
SLM 2608D
EU阵列 2602E
TD/IC 2603E
媒体采样器 2606E
EU阵列 2604E
3D采样器 2605E
着色器处理器 2607E
SLM 2608E
EU阵列 2602F
TD/IC 2603F
媒体采样器 2606F
EU阵列 2604F
3D采样器 2605F
着色器处理器 2607F
SLM 2608F

图26

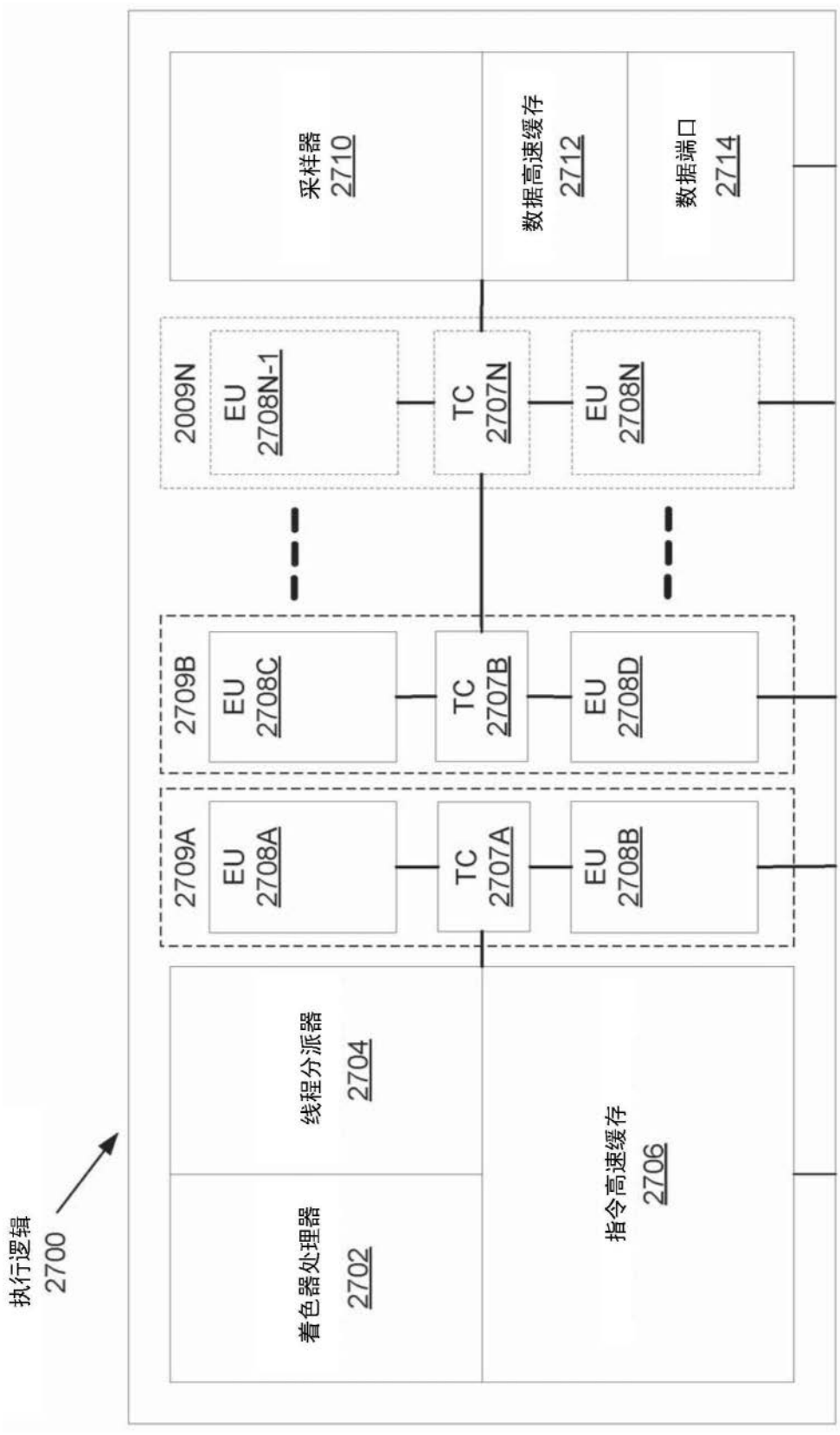

执行逻辑
2700
着色器处理器
2702
线程分派器
2704
指令高速缓存
2706
2709A
EU
2708A
TC
2707A
EU
2708B
2709B
EU
2708C
TC
2707B
EU
2708D
2009N
EU
2708N-1
TC
2707N
EU
2708N
采样器
2710
数据高速缓存
2712
数据端口
2714

图27A

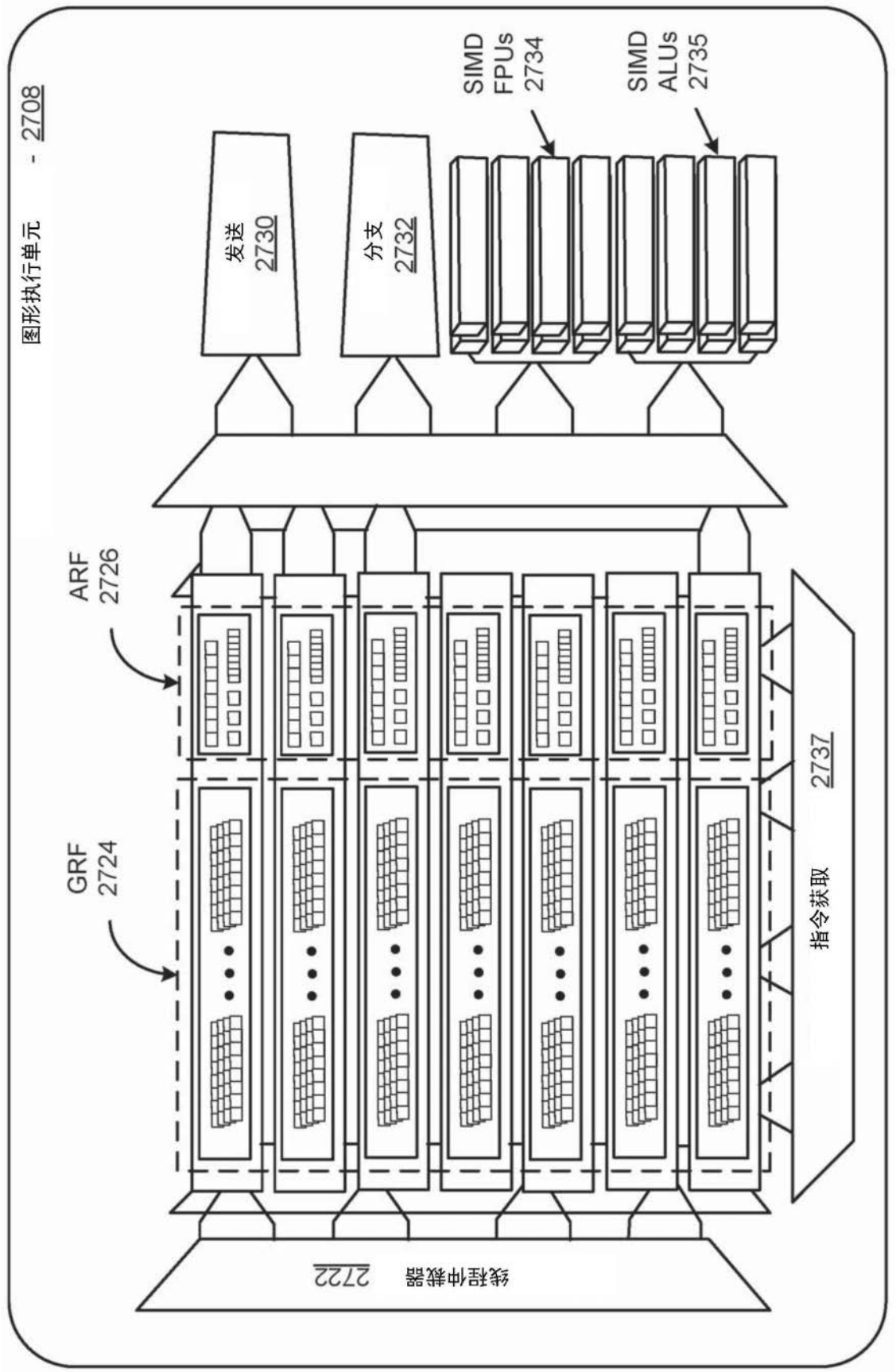

图形执行单元 - 2708
SIMD FPUs 2734
SIMD ALUs 2735
发送 2730
分支 2732
ARF 2726
GRF 2724
指令获取 2737
线程仲裁器 2722

图27B

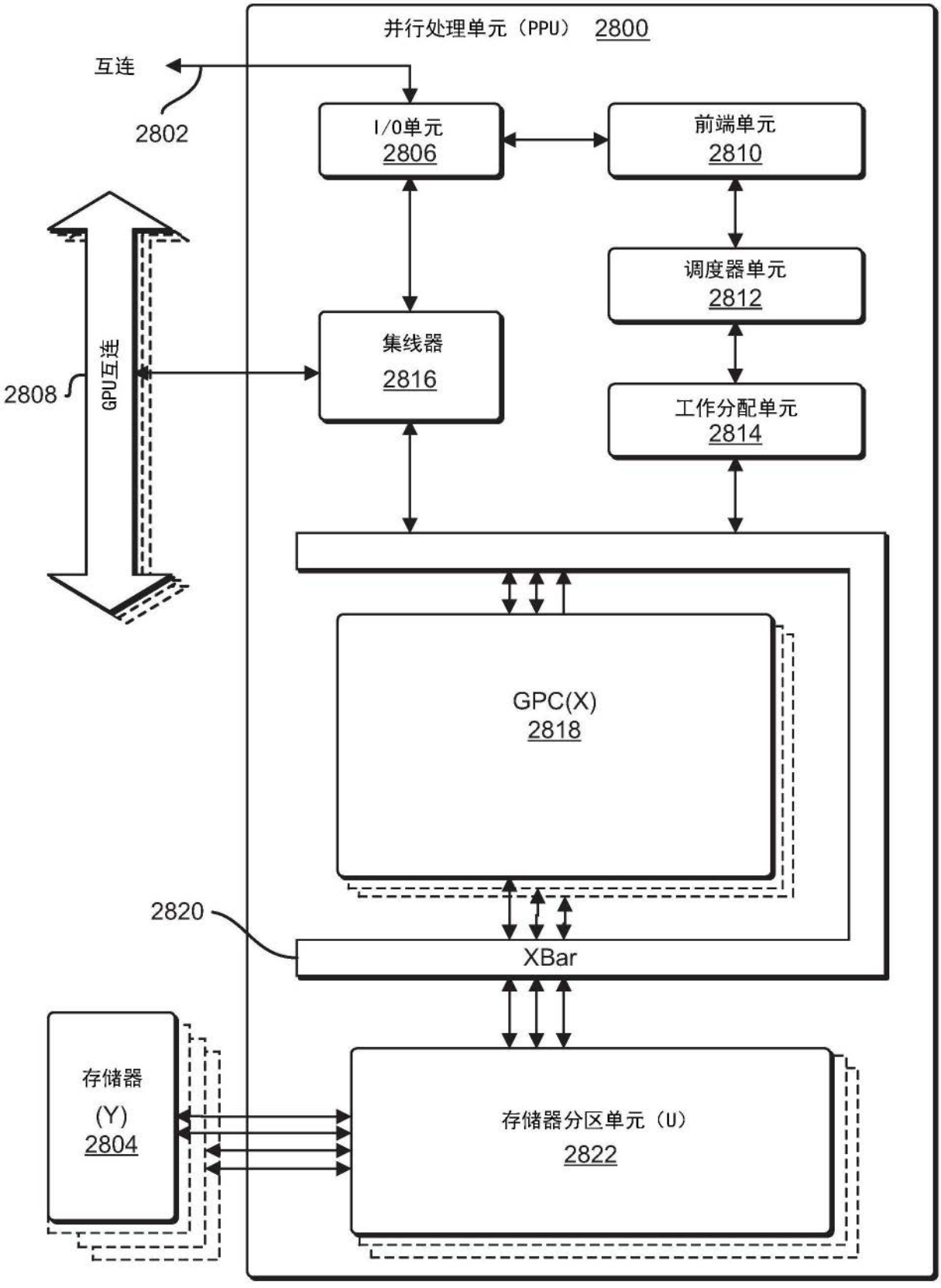

并行处理单元（PPU） 2800
互连
2802
I/O单元
2806
前端单元
2810
调度器单元
2812
集线器
2816
工作分配单元
2814
2808
GPU互连
GPC(X)
2818
2820
XBar
存储器
(Y)
2804
存储器分区单元（U）
2822

图28

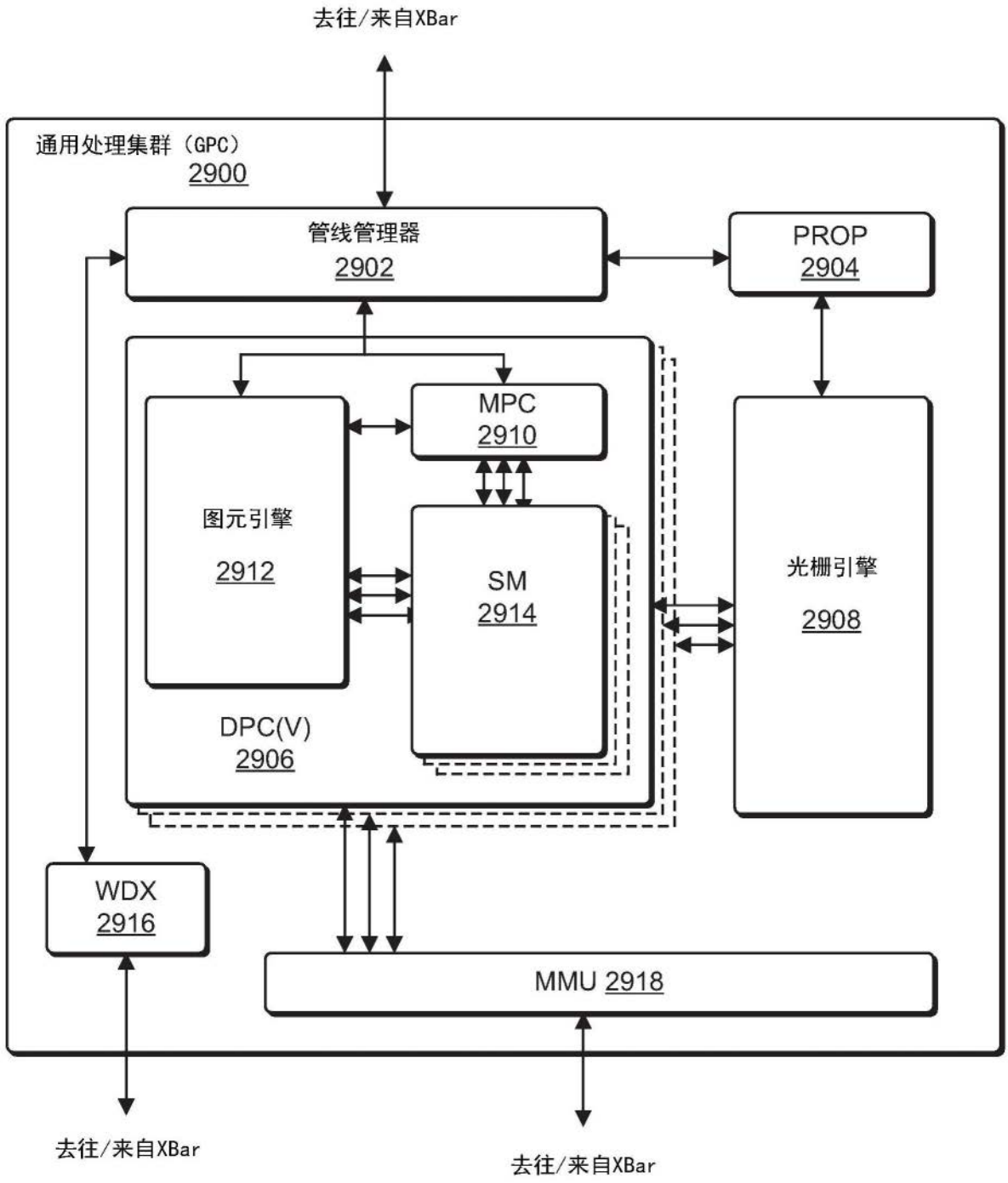

去往/来自XBar
通用处理集群（GPC）
2900
管线管理器
2902
PROP
2904
MPC
2910
图元引擎
2912
SM
2914
光栅引擎
2908
DPC(V)
2906
WDX
2916
MMU 2918
去往/来自XBar
去往/来自XBar

图29

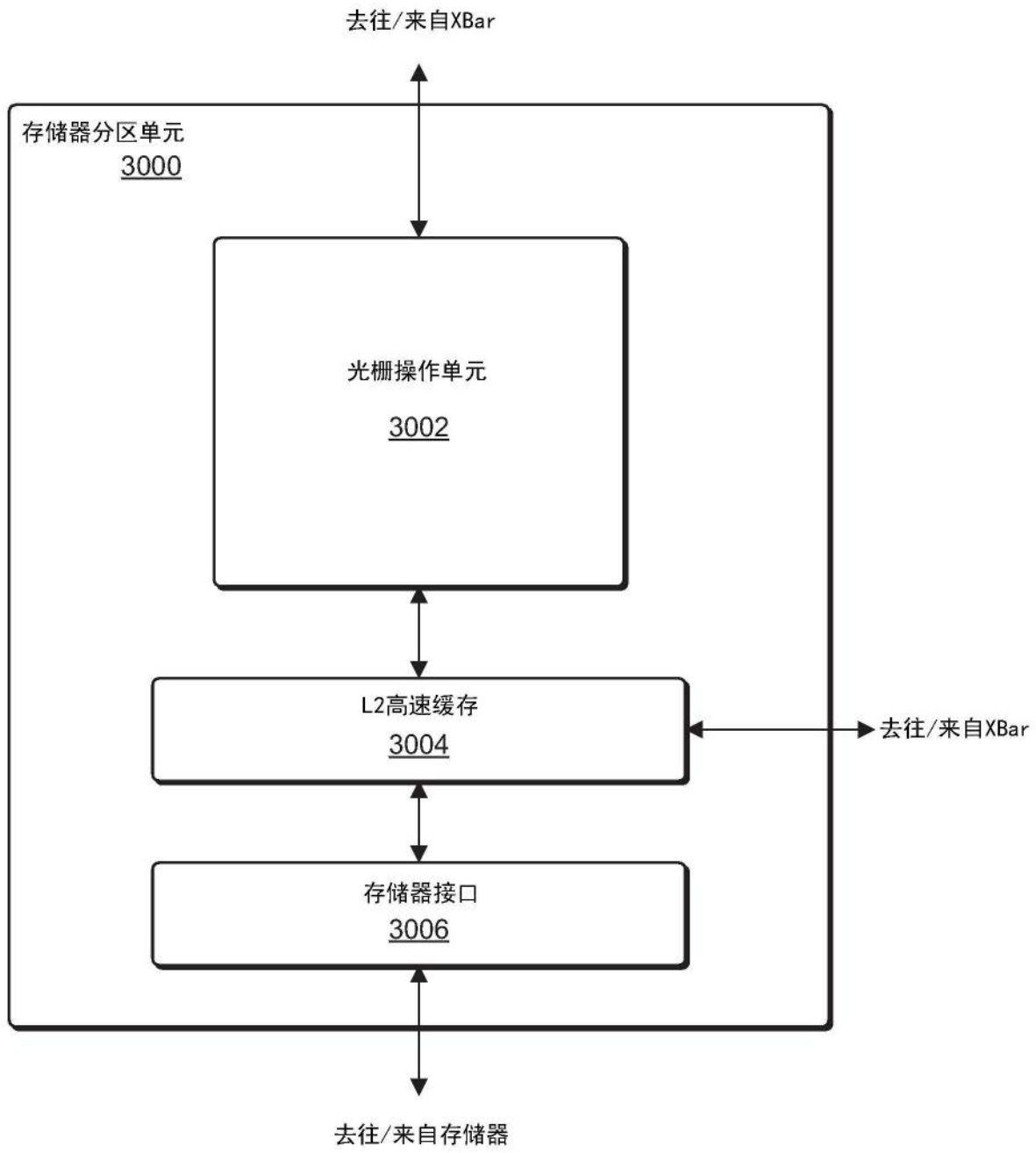

去往/来自XBar
存储器分区单元
3000
光栅操作单元
3002
L2高速缓存
3004
去往/来自XBar
存储器接口
3006
去往/来自存储器

图30

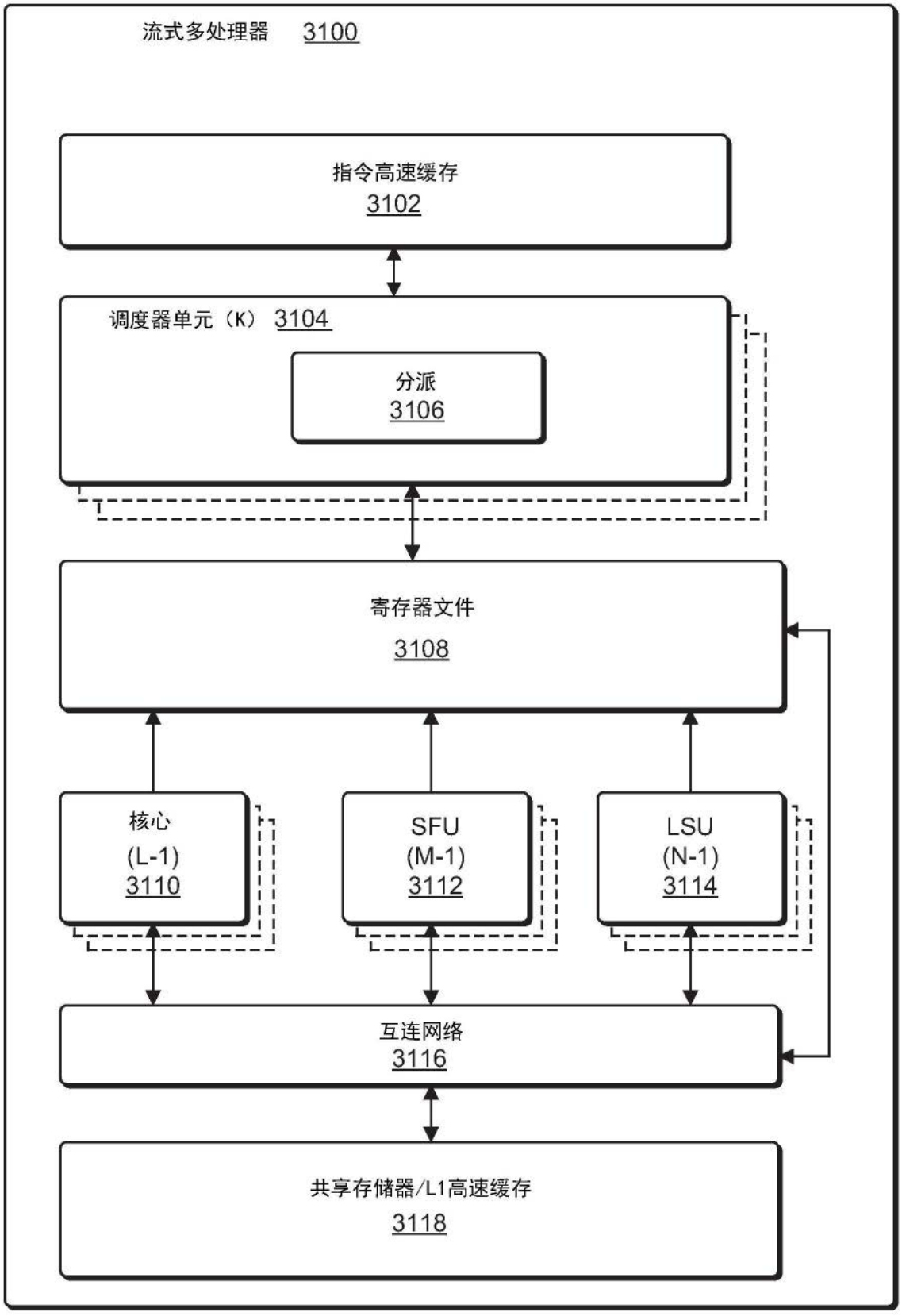

流式多处理器 3100
指令高速缓存
3102
调度器单元（K） 3104
分派
3106
寄存器文件
3108
核心
(L-1)
3110
SFU
(M-1)
3112
LSU
(N-1)
3114
互连网络
3116
共享存储器/L1高速缓存
3118

图31

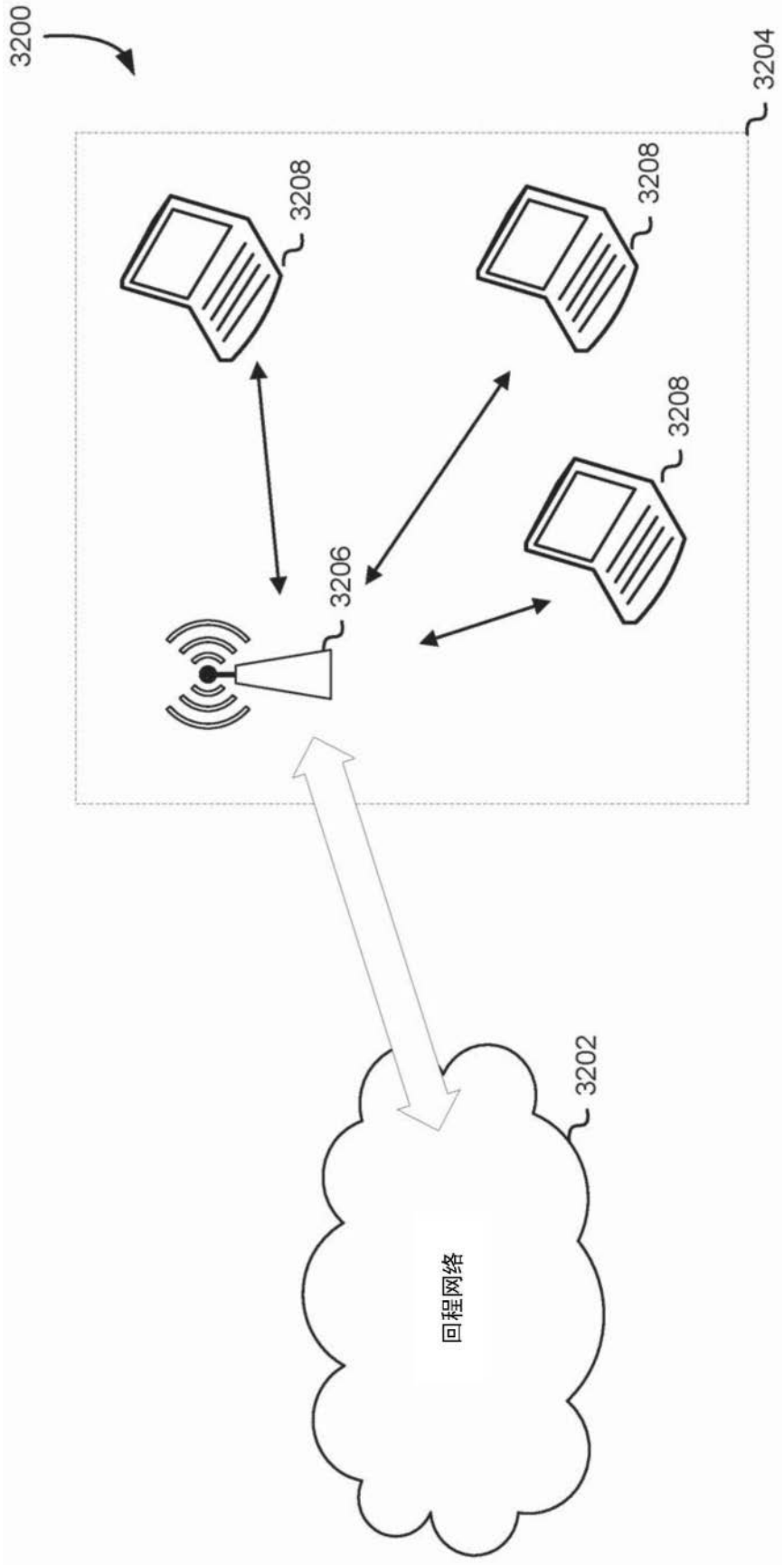


图32

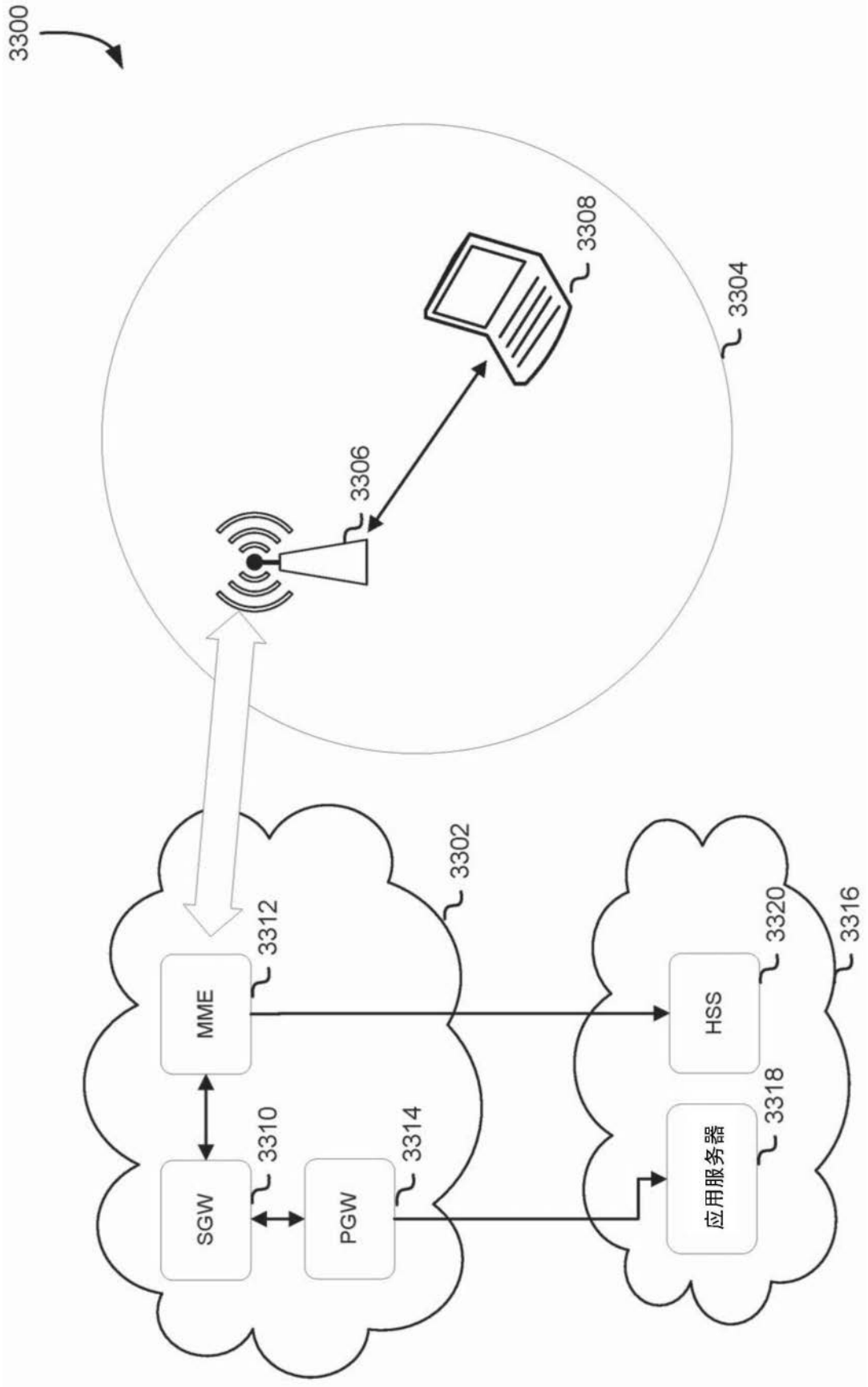

3300
3308
3304
3306
3302
3312
MME
3320
HSS
3316
3310
SGW
3314
PGW
3318
应用服务器

图33

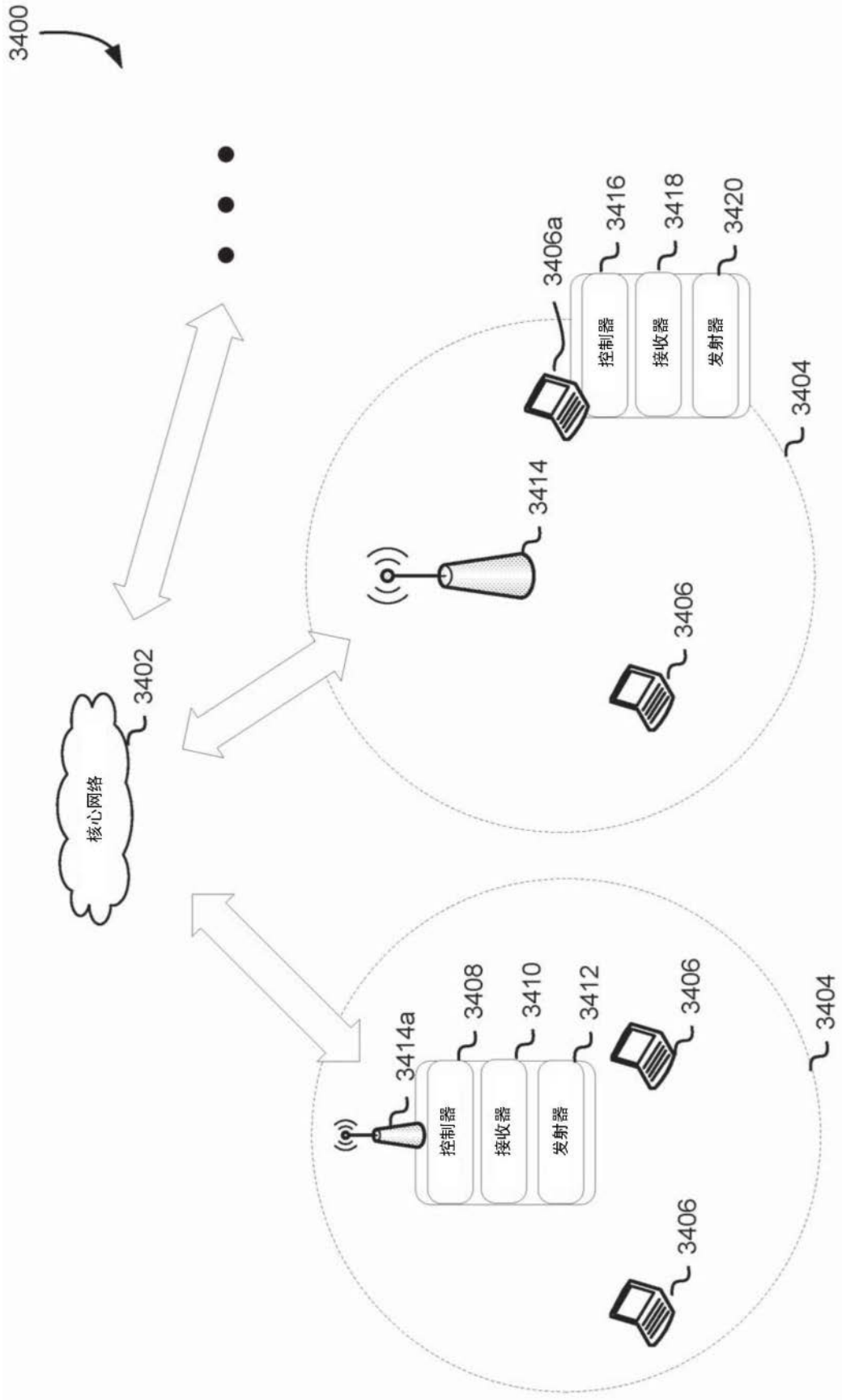

3400
3402
核心网络
3404
3406
3406a
3408
3410
3412
3414
3414a
3416
3418
3420
控制器
接收器
发射器

图34

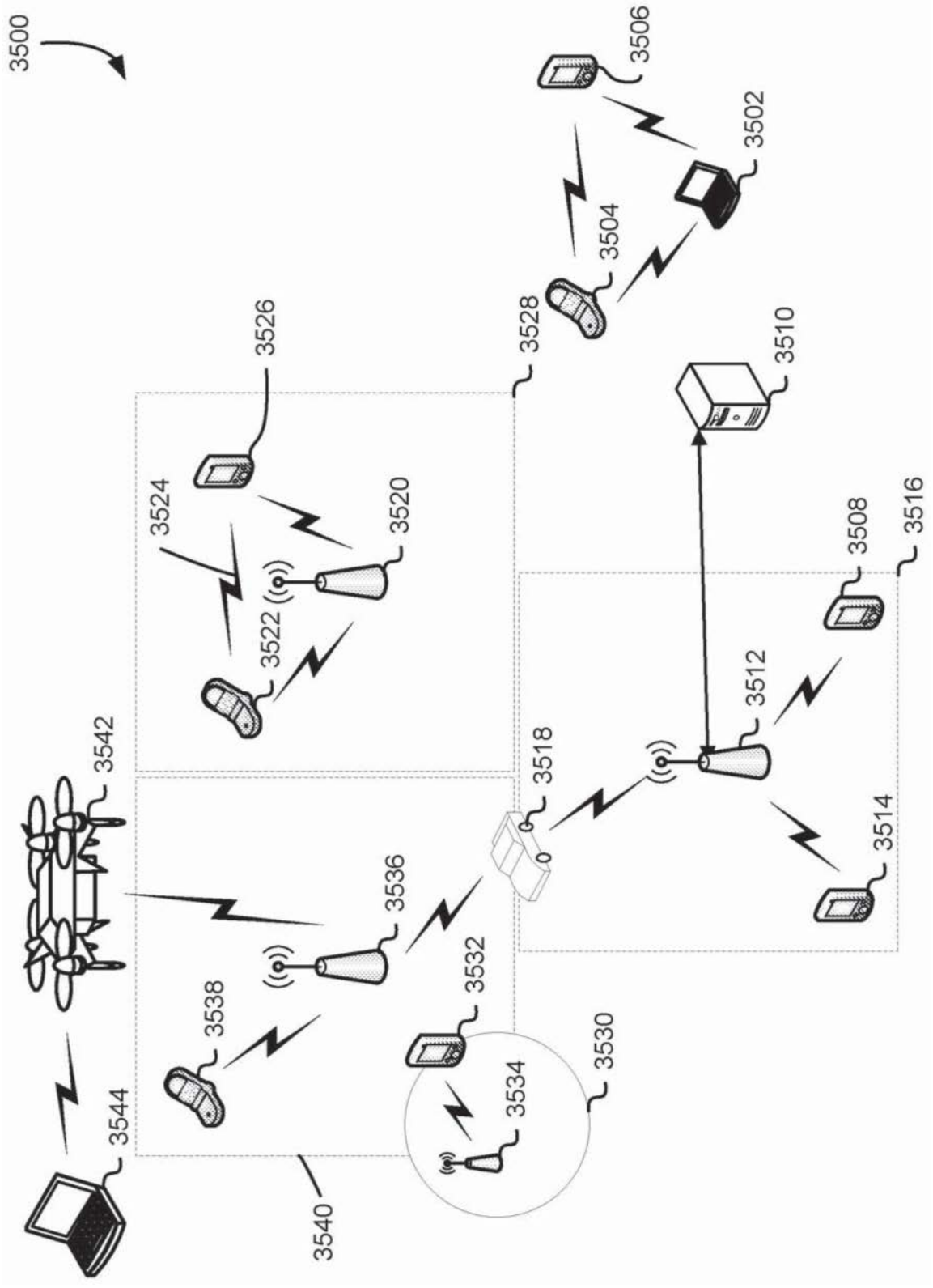

3500
3506
3502
3504
3528
3526
3510
3524
3520
3522
3508
3516
3512
3542
3518
3514
3536
3532
3538
3530
3534
3544
3540

图35

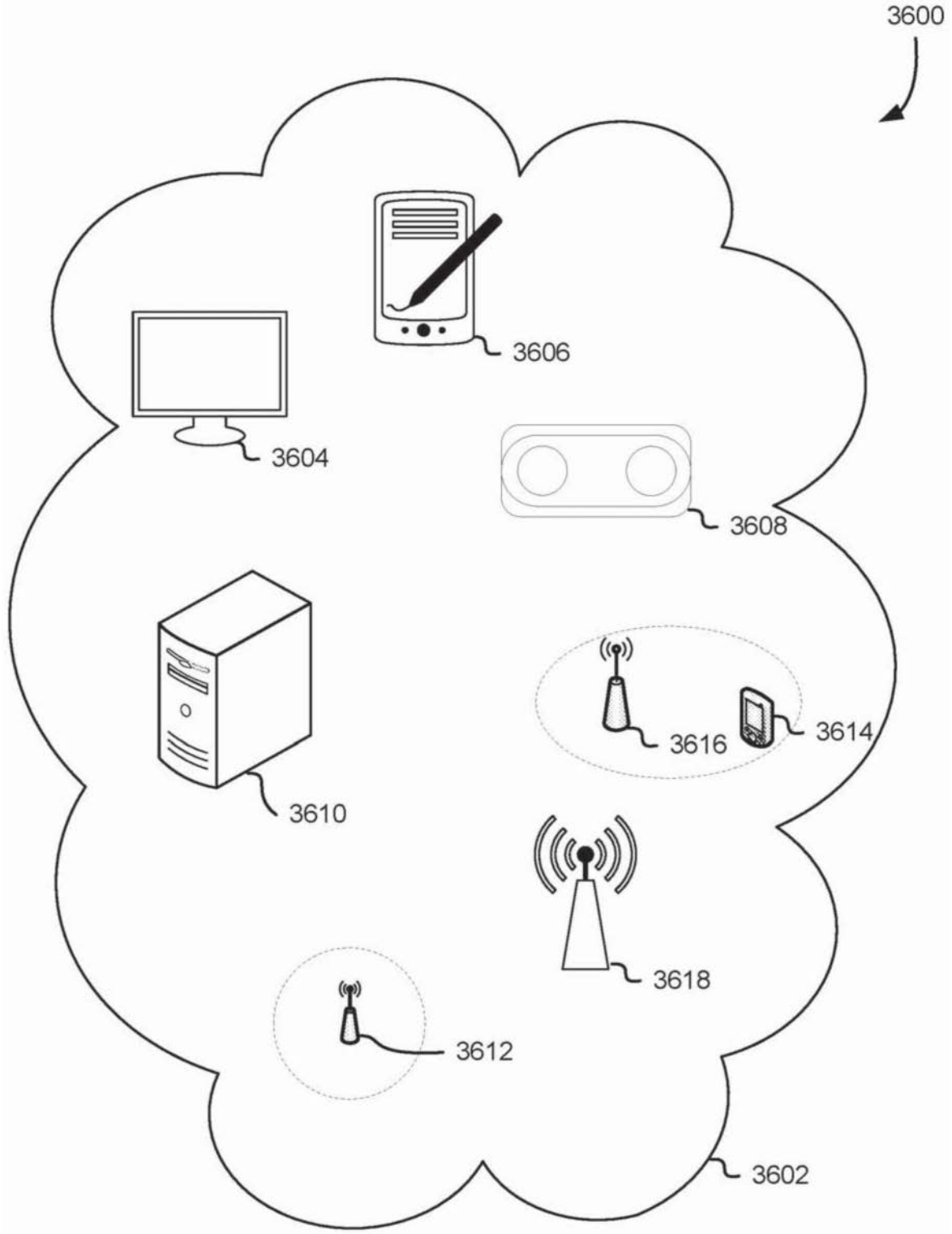


图36

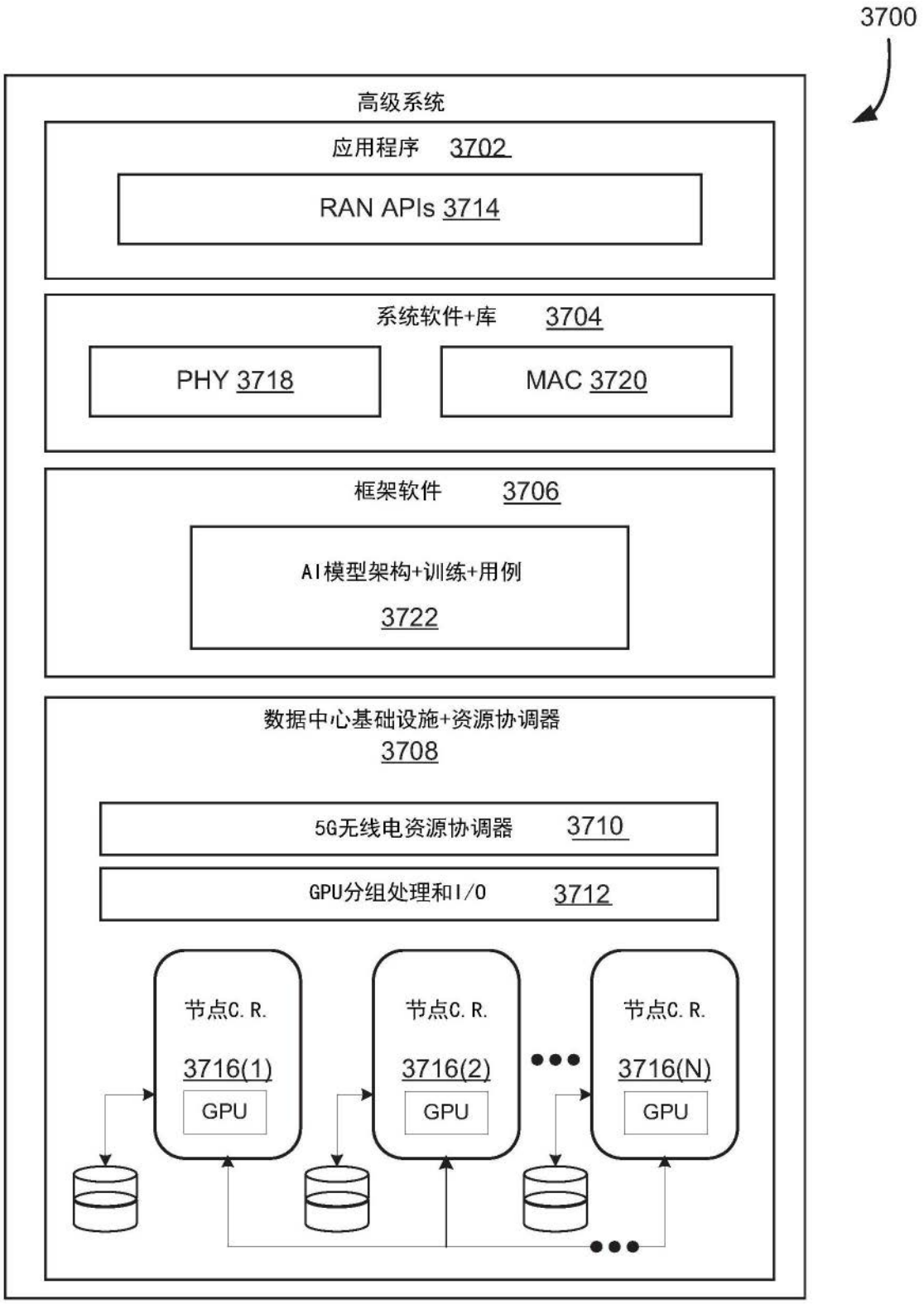

3700
高级系统
应用程序 3702
RAN APIs 3714
系统软件+库 3704
PHY 3718
MAC 3720
框架软件 3706
AI模型架构+训练+用例
3722
数据中心基础设施+资源协调器
3708
5G无线电资源协调器 3710
GPU分组处理和I/O 3712
节点C. R.
3716(1)
GPU
节点C. R.
3716(2)
GPU
节点C. R.
3716(N)
GPU

图37

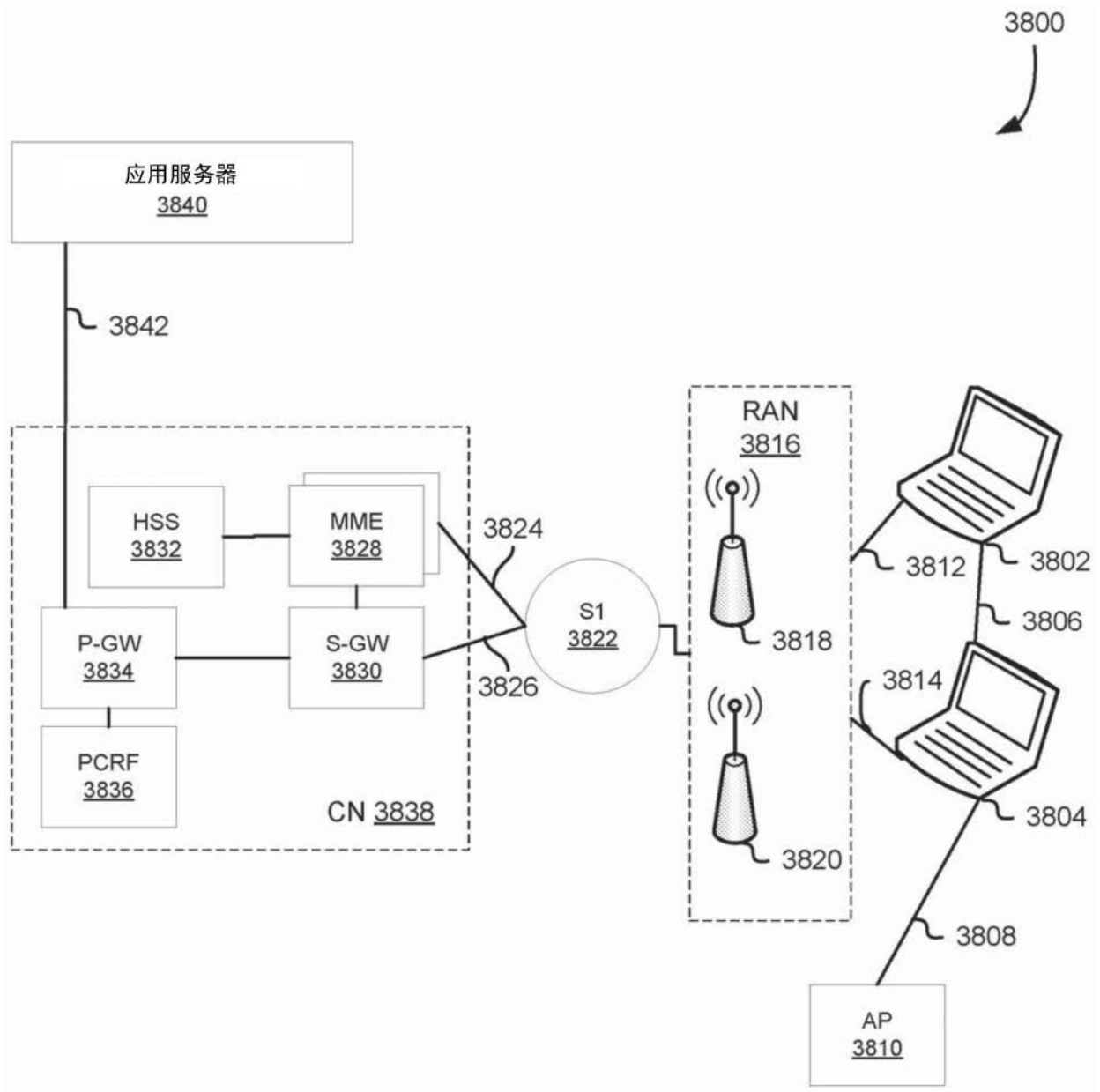

3800
应用服务器
3840
3842
HSS
3832
MME
3828
3824
P-GW
3834
S-GW
3830
3826
PCRF
3836
CN 3838
S1
3822
RAN
3816
3818
3820
3812
3802
3806
3814
3804
3808
AP
3810

图38

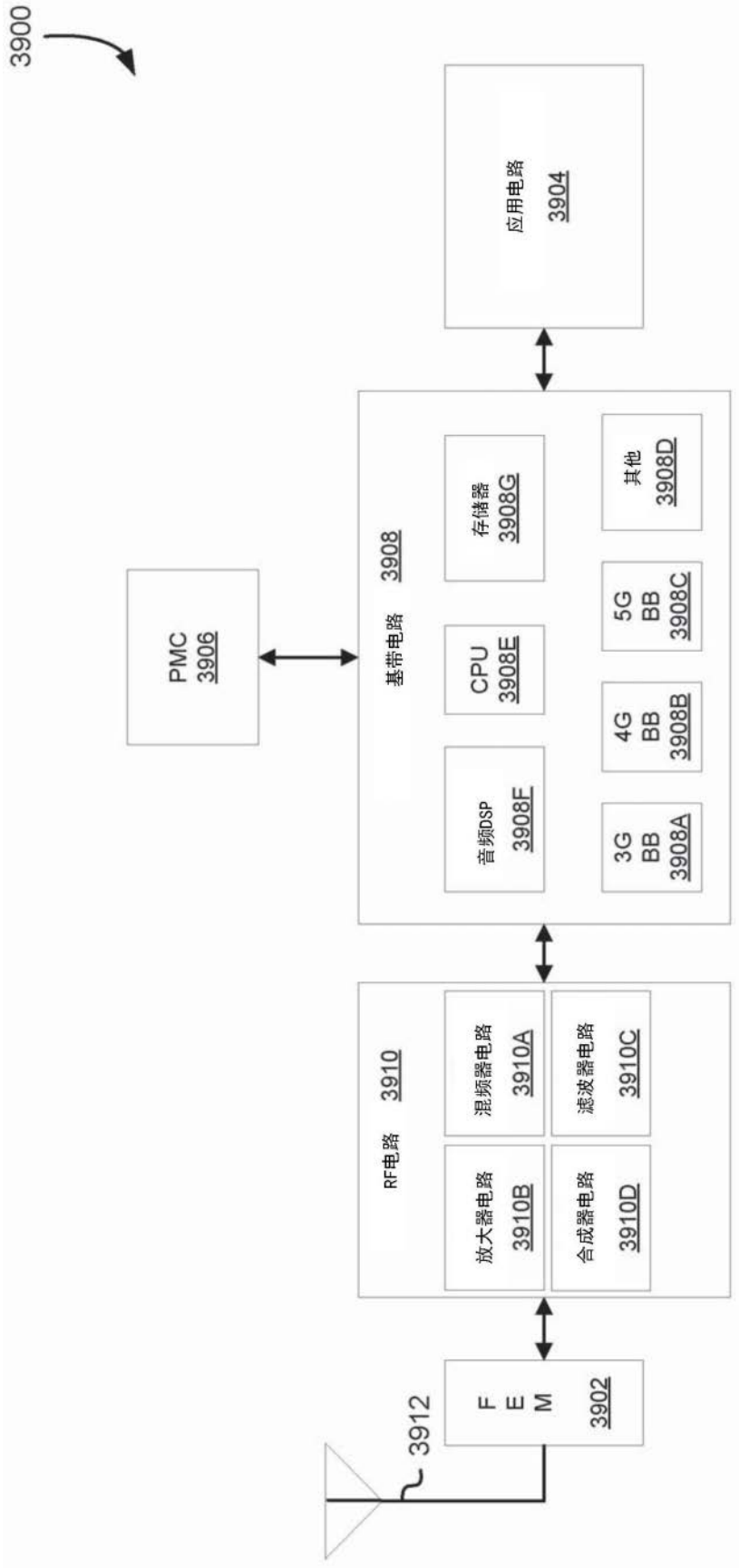

3900
应用电路
3904
PMC
3906
基带电路
3908
存储器
3908G
其他
3908D
CPU
3908E
5G
BB
3908C
4G
BB
3908B
音频DSP
3908F
3G
BB
3908A
RF电路
3910
混频器电路
3910A
滤波器电路
3910C
放大器电路
3910B
合成器电路
3910D
F
E
M
3902
3912

图39

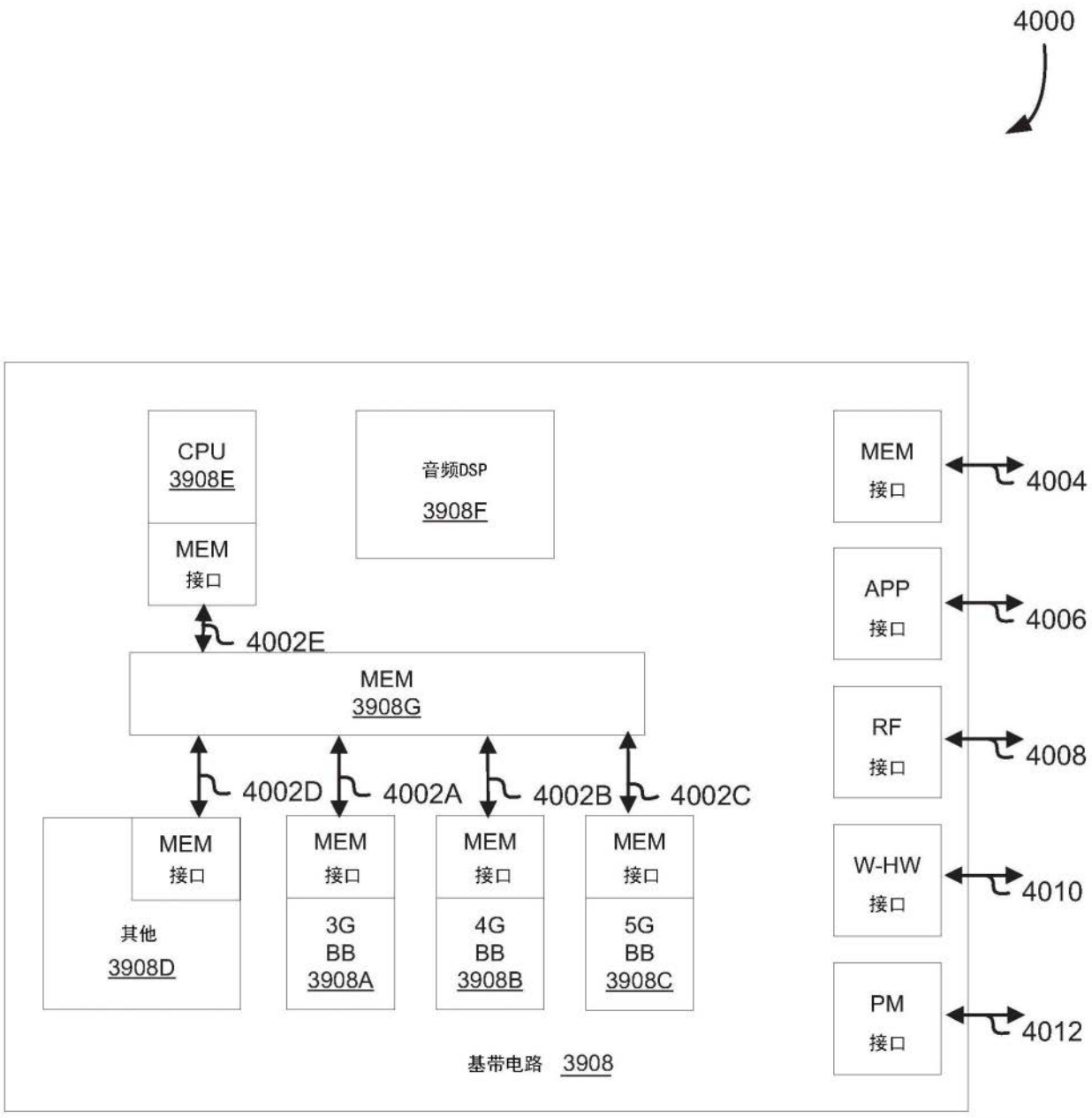

4000
CPU
3908E
MEM
接口
音频DSP
3908F
MEM
接口
4004
APP
接口
4006
4002E
MEM
3908G
RF
接口
4008
4002D
4002A
4002B
4002C
MEM
接口
MEM
接口
MEM
接口
MEM
接口
W-HW
接口
4010
其他
3908D
3G
BB
3908A
4G
BB
3908B
5G
BB
3908C
PM
接口
4012
基带电路 3908

图40

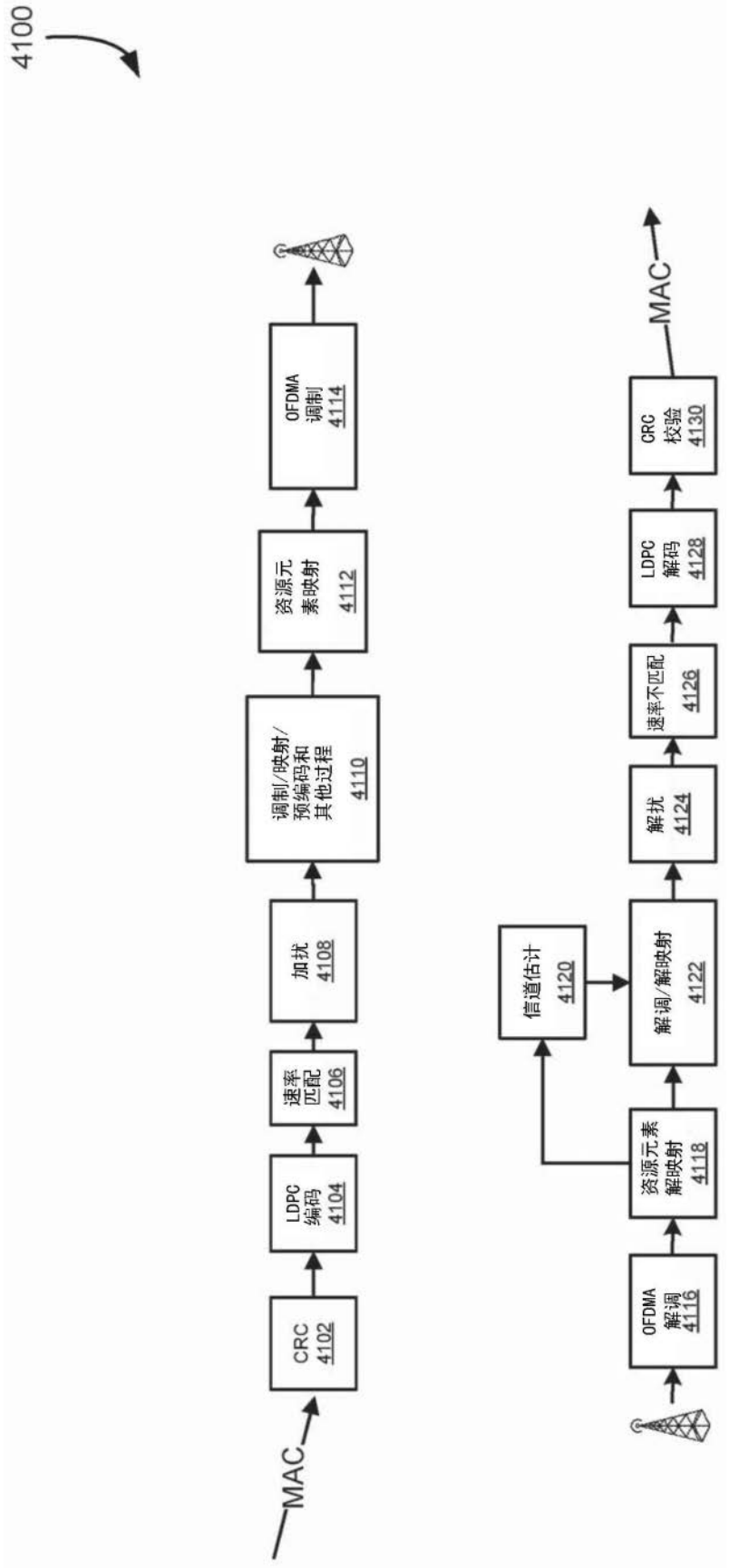

4100
MAC
CRC 4102
LDPC 编码 4104
速率匹配 4106
加扰 4108
调制/映射/预编码和其他过程 4110
资源元素映射 4112
OFDMA 调制 4114
OFDMA 解调 4116
资源元素解映射 4118
信道估计 4120
解调/解映射 4122
解扰 4124
速率不匹配 4126
LDPC 解码 4128
CRC 校验 4130
MAC

图41

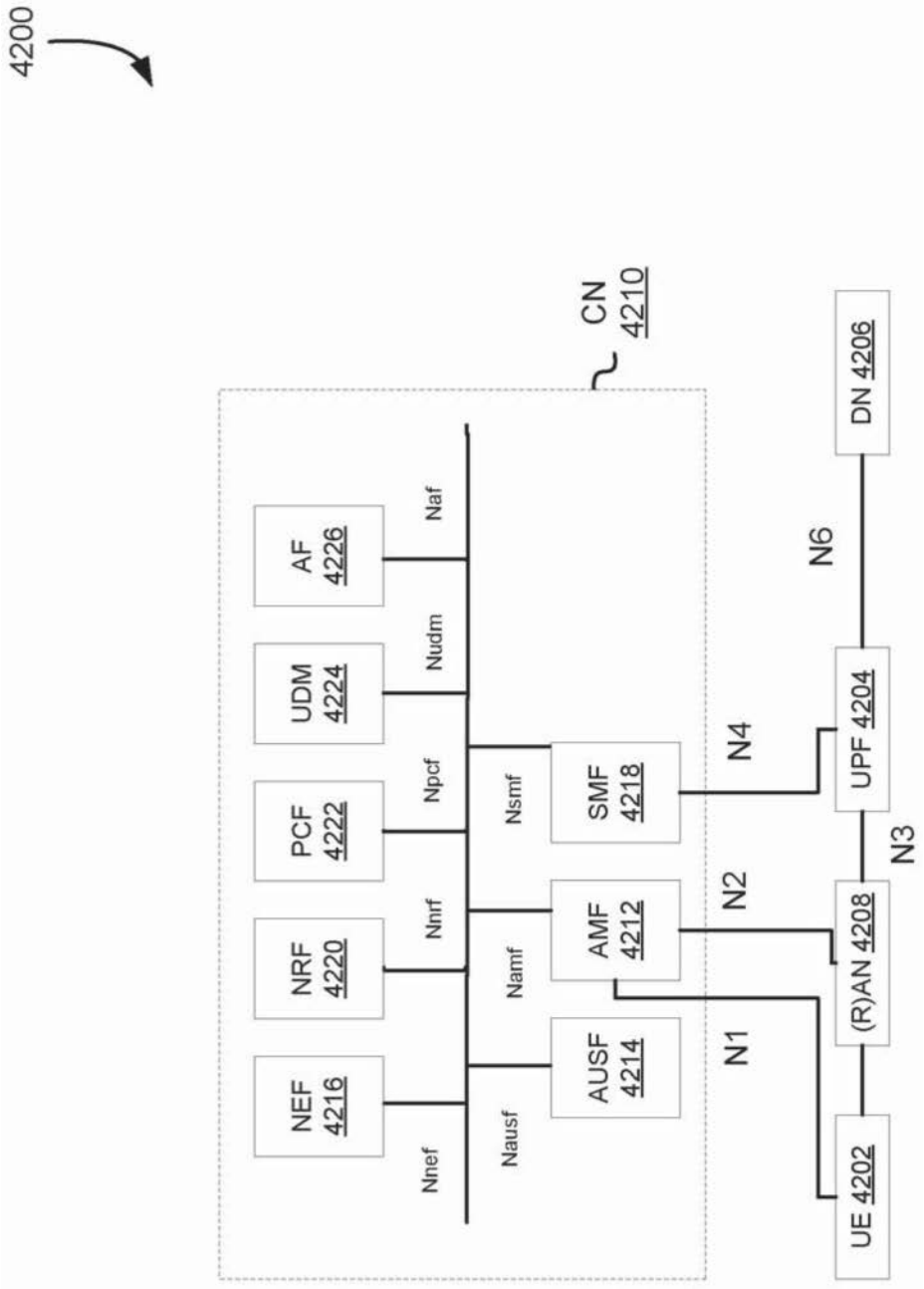

4200
CN 4210
AF 4226
UDM 4224
PCF 4222
NRF 4220
NEF 4216
Naf
Nudm
Npcf
Nnrf
Nnef
Nsmf
Namf
Nausf
SMF 4218
AMF 4212
AUSF 4214
N4
N2
N1
N6
N3
DN 4206
UPF 4204
(R)AN 4208
UE 4202

图42

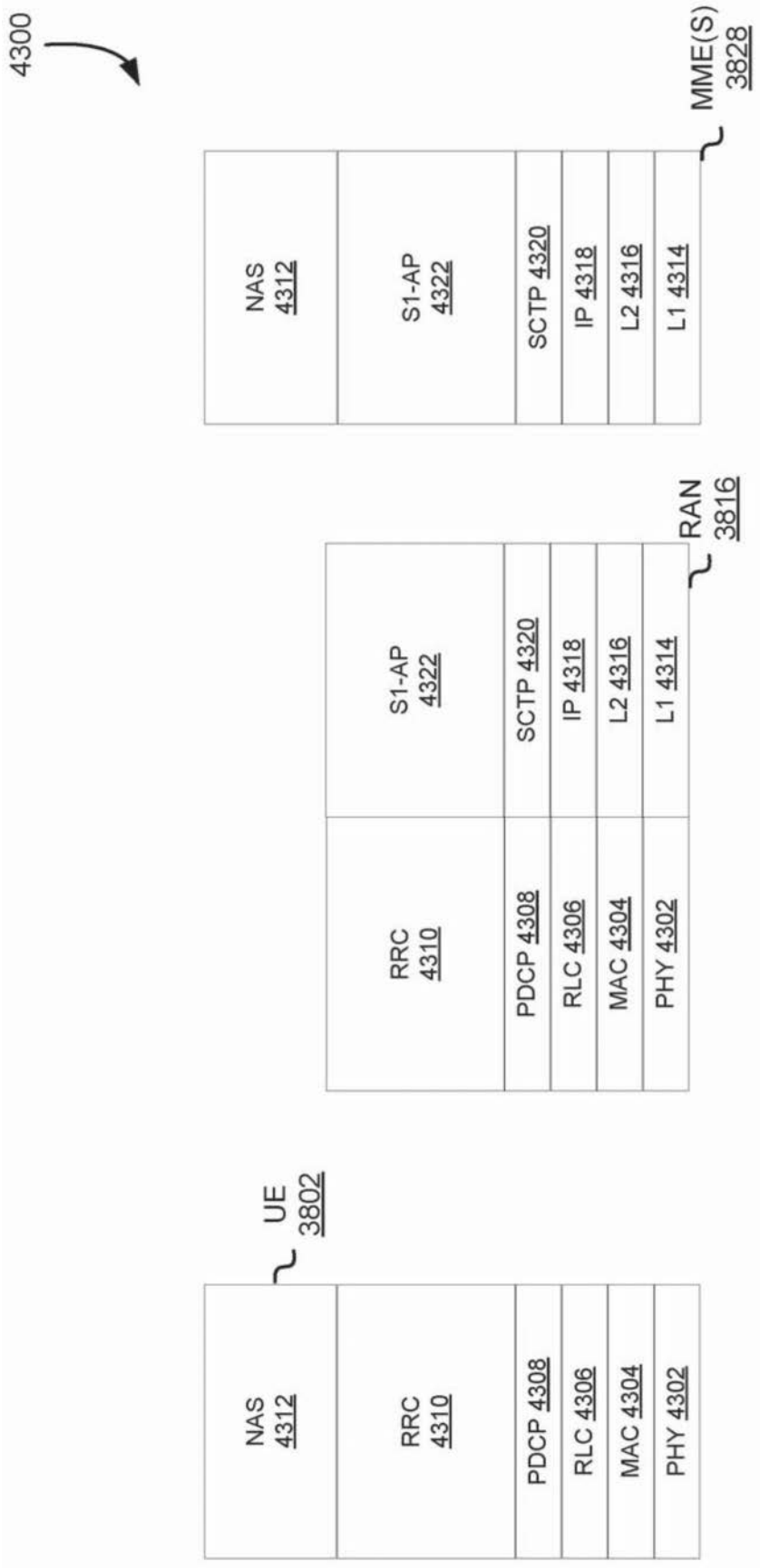

4300
MME(S) 3828
NAS 4312
S1-AP 4322
SCTP 4320
IP 4318
L2 4316
L1 4314
RAN 3816
S1-AP 4322
SCTP 4320
IP 4318
L2 4316
L1 4314
RRC 4310
PDCP 4308
RLC 4306
MAC 4304
PHY 4302
UE 3802
NAS 4312
RRC 4310
PDCP 4308
RLC 4306
MAC 4304
PHY 4302

图43

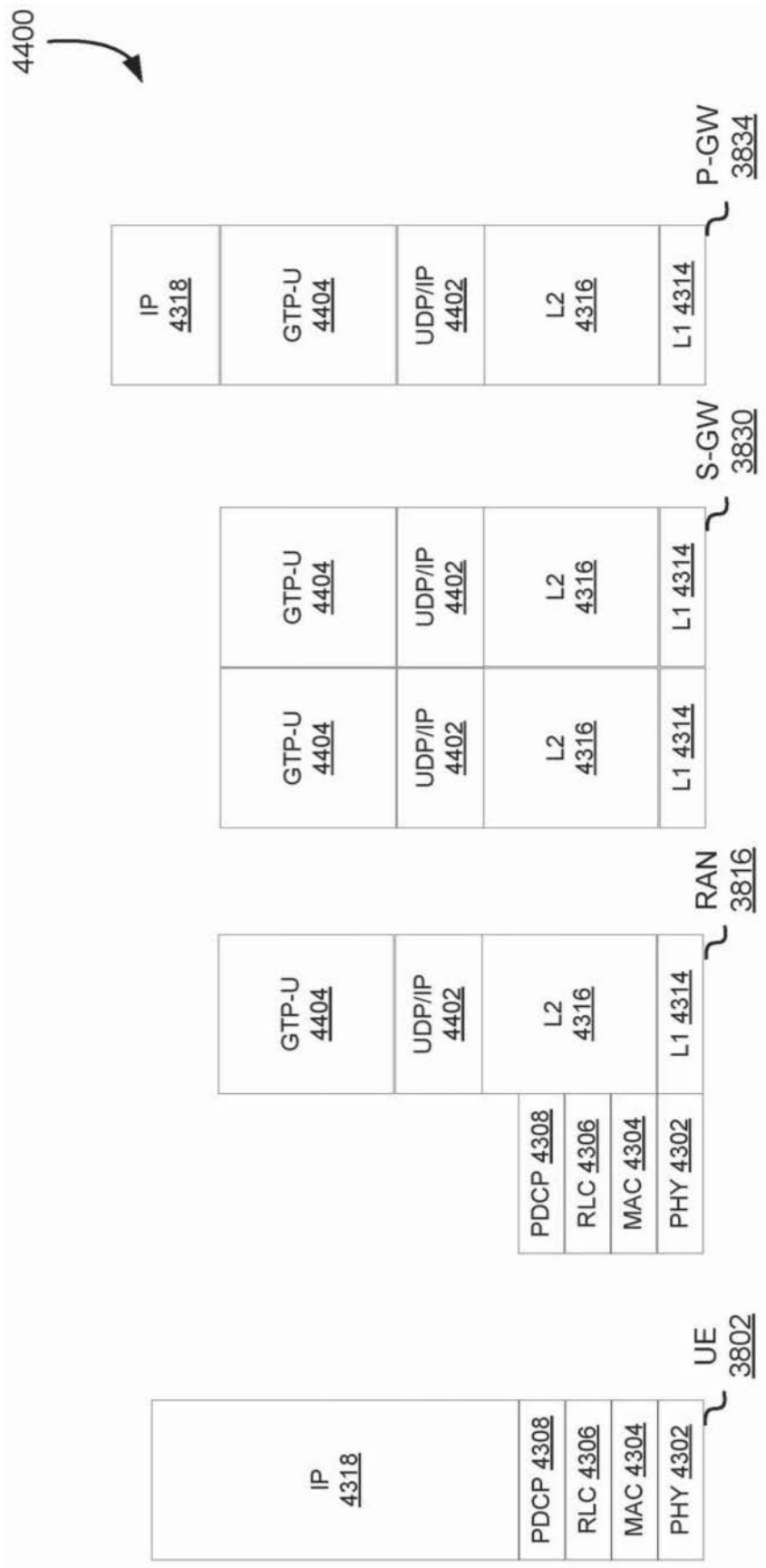

4400
P-GW 3834
IP 4318
GTP-U 4404
UDP/IP 4402
L2 4316
L1 4314
S-GW 3830
GTP-U 4404
UDP/IP 4402
L2 4316
L1 4314
RAN 3816
GTP-U 4404
UDP/IP 4402
L2 4316
L1 4314
PDCP 4308
RLC 4306
MAC 4304
PHY 4302
UE 3802
IP 4318
PDCP 4308
RLC 4306
MAC 4304
PHY 4302

图44

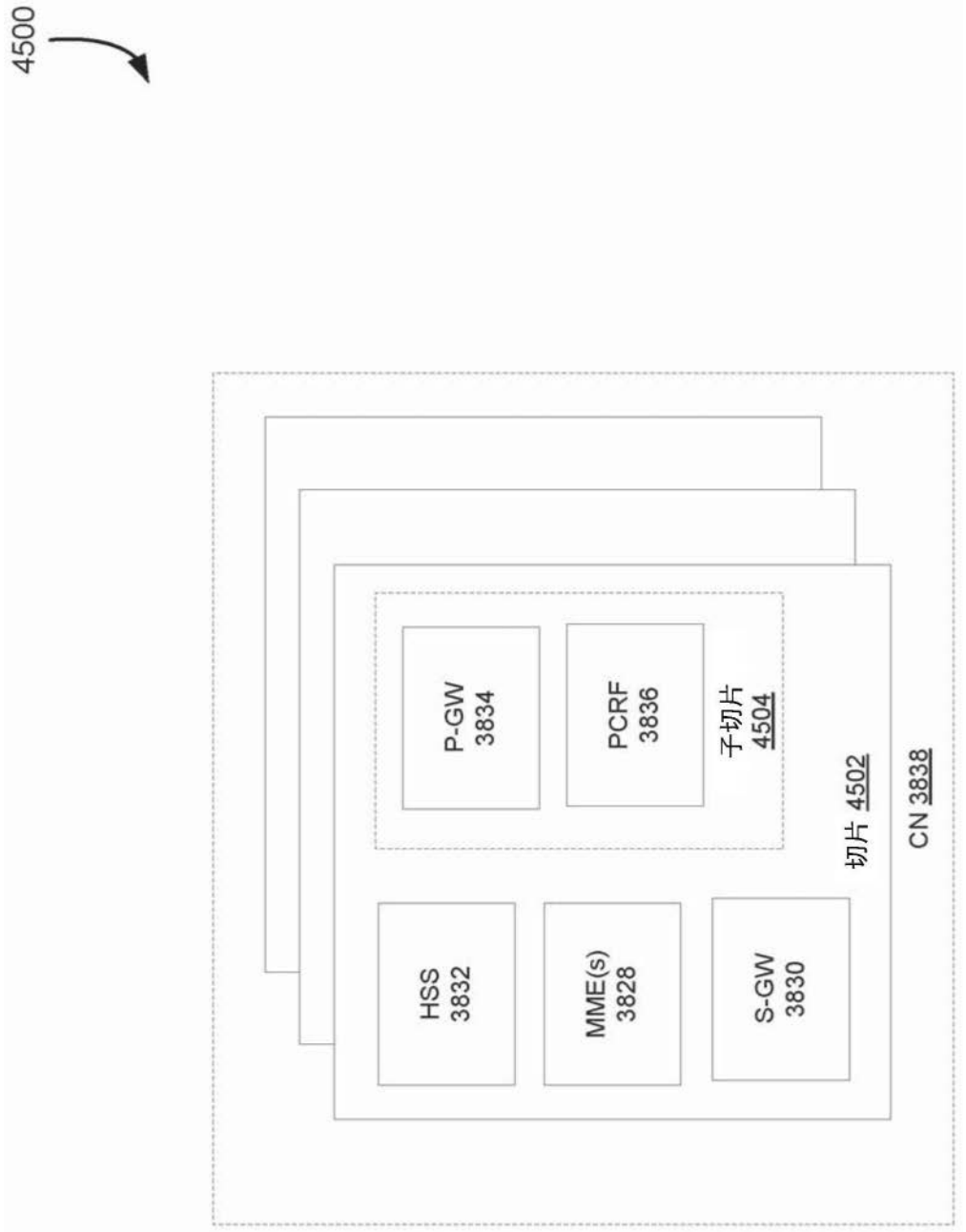

4500
CN 3838
切片 4502
子切片 4504
P-GW 3834
PCRF 3836
HSS 3832
MME(s) 3828
S-GW 3830

图45

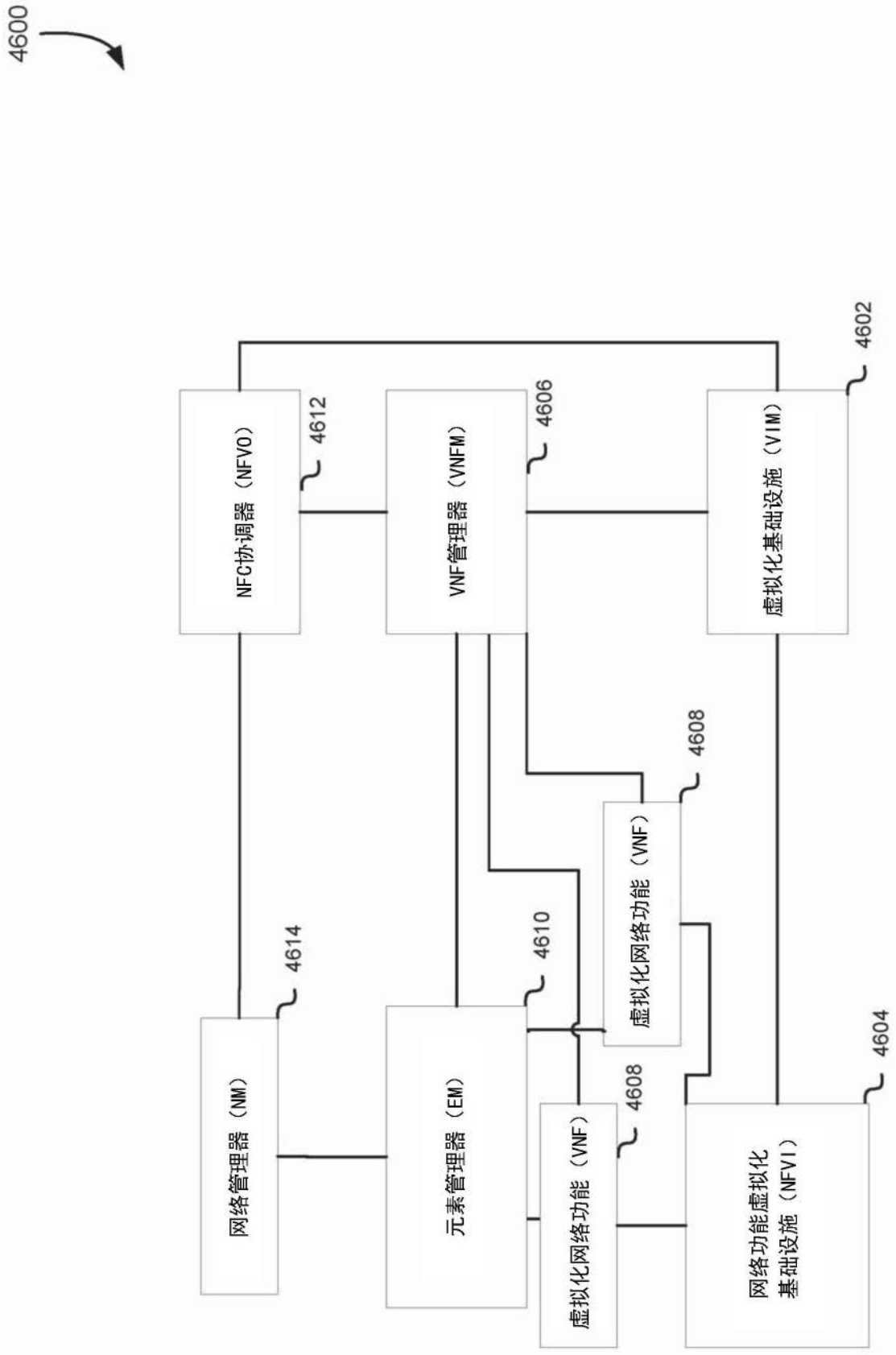

4600
NFC协调器（NFVO）
4612
VNF管理器（VNFM）
4606
虚拟化基础设施（VIM）
4602
4608
虚拟化网络功能（VNF）
网络管理器（NM）
4614
元素管理器（EM）
4610
虚拟化网络功能（VNF）
4608
网络功能虚拟化
基础设施（NFVI）
4604

图46