(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 113723586 A
(43)申请公布日 2021.11.30

(21)申请号 202110453866.8

(22)申请日 2021.04.26

(30)优先权数据
16/860,175 2020.04.28 US

(71)申请人 辉达公司
地址 美国加利福尼亚州

(72)发明人 S·帕德什 P·P·科塔里
V·V·盖克瓦德

(74)专利代理机构 北京市磐华律师事务所 11336
代理人 高伟

(51)Int.Cl.
G06N 3/04(2006.01)
G06N 3/08(2006.01)
G10L 15/08(2006.01)
G10L 15/183(2013.01)
G10L 15/26(2006.01)

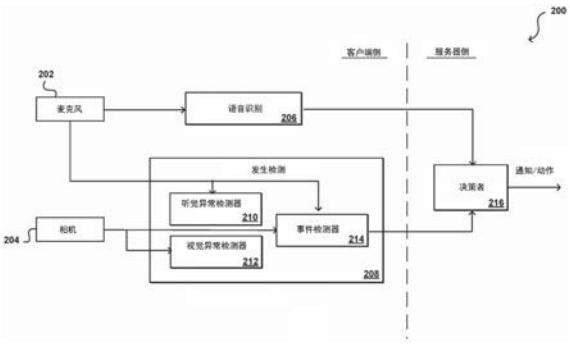
权利要求书3页 说明书99页 附图53页

(54)发明名称

使用一个或更多个神经网络确定的通知

(57)摘要

本发明公开了使用一个或更多个神经网络确定的通知，具体提出了用于确定针对数据异常将采取的动作的装置、系统和技术。在至少一个实施例中，可以分析针对用户的环境捕获的音频和视频数据，以检测一个或更多个数据异常，并且根据异常是否适用于该用户来确定是否通知该用户。



CN 113723586 A

1.一种处理器,包括:

一个或更多个电路,用于使用一个或更多个神经网络来检测一个或更多个数据异常,并使得一个或更多个用户被通知所述一个或更多个异常,其中所述一个或更多个用户被通知取决于所述一个或更多个数据异常是否适用于所述一个或更多个用户。

2.根据权利要求1所述的处理器,其中所述一个或更多个神经网络包括用于提供所述一个或更多个数据异常的实例和置信数据的音频异常检测器和视频异常检测器,所述音频异常检测器将针对所述一个或更多个用户的环境捕获的音频数据作为输入,并且所述视频异常检测器将针对所述一个或更多个用户的所述环境捕获的视频数据作为输入。

3.根据权利要求2所述的处理器,其中所述一个或更多个神经网络包括用于确定所述一个或更多个数据异常中的每一个数据异常的分类的事件检测器。

4.根据权利要求3所述的处理器,其中所述一个或更多个神经网络包括用于确定是否使得所述一个或更多个用户被通知的决策者网络,所述决策者网络使用所述实例和置信数据以及用于所述一个或更多个数据异常中的每一个数据异常的所述分类,来确定所述一个或更多个数据异常是否适用于所述一个或更多个用户。

5.根据权利要求4所述的处理器,其中所述一个或更多个电路进一步用于使用语音识别模块来向所述决策者网络提供与所述一个或更多个数据异常相关的检测到的语音的文本。

6.根据权利要求4所述的处理器,其中所述决策者网络进一步确定是否采取动作以减小所述一个或更多个用户的体验的沉浸性。

7.一种系统,包括:

一个或更多个处理器,用于使用一个或更多个神经网络来检测视频数据和音频数据中的一个或更多个数据异常,并使得一个或更多个用户被通知所述一个或更多个异常,其中所述一个或更多个用户被通知取决于所述一个或更多个数据异常是否适用于所述一个或更多个用户。

8.根据权利要求7所述的系统,其中所述一个或更多个神经网络包括用于提供所述一个或更多个数据异常的实例和置信数据的音频异常检测器和视频异常检测器,所述音频异常检测器将针对所述一个或更多个用户的环境捕获的音频数据作为输入,并且所述视频异常检测器将针对所述一个或更多个用户的所述环境捕获的视频数据作为输入。

9.根据权利要求8所述的系统,其中所述一个或更多个神经网络包括用于确定所述一个或更多个数据异常中的每一个数据异常的分类的事件检测器。

10.根据权利要求9所述的系统,其中所述一个或更多个神经网络包括用于确定是否使得所述一个或更多个用户被通知的决策者网络,所述决策者网络使用所述实例和置信数据以及用于所述一个或更多个数据异常中的每一个数据异常的所述分类,来确定所述一个或更多个数据异常是否适用于所述一个或更多个用户。

11.根据权利要求10所述的系统,其中所述一个或更多个处理器进一步使用语音识别模块来向所述决策者网络提供与所述一个或更多个数据异常相关的检测到的语音的文本。

12.根据权利要求10所述的系统,其中所述决策者网络进一步确定是否采取动作以减小所述一个或更多个用户的体验的沉浸性。

13.一种方法,包括:

使用一个或更多个神经网络来检测视频数据和音频数据中的一个或更多个数据异常，并使得一个或更多个用户被通知所述一个或更多个异常，其中所述一个或更多个用户被通知取决于所述一个或更多个数据异常是否适用于所述一个或更多个用户。

14.根据权利要求13所述的方法，其中所述一个或更多个神经网络包括用于提供所述一个或更多个数据异常的实例和置信数据的音频异常检测器和视频异常检测器，所述音频异常检测器将针对所述一个或更多个用户的环境捕获的音频数据作为输入，并且所述视频异常检测器将针对所述一个或更多个用户的所述环境捕获的视频数据作为输入。

15.根据权利要求14所述的方法，其中所述一个或更多个神经网络包括用于确定所述一个或更多个数据异常中的每一个数据异常的分类的事件检测器。

16.根据权利要求15所述的方法，其中所述一个或更多个神经网络包括用于确定是否使得所述一个或更多个用户被通知的决策者网络，所述决策者网络使用所述实例和置信数据以及用于所述一个或更多个数据异常中的每一个数据异常的所述分类，来确定所述一个或更多个数据异常是否适用于所述一个或更多个用户。

17.根据权利要求16所述的方法，进一步包括：

使用语音识别模块向所述决策者网络提供与所述一个或更多个数据异常有关的检测到的语音的文本。

18.根据权利要求16所述的方法，其中所述决策者网络进一步确定是否采取动作以减小所述一个或更多个用户的体验的沉浸性。

19.一种机器可读介质，具有存储在其上的指令集，所述指令集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：

使用一个或更多个神经网络来检测视频数据和音频数据中的一个或更多个数据异常，并使得一个或更多个用户被通知所述一个或更多个异常，其中所述一个或更多个用户被通知取决于所述一个或更多个数据异常是否适用于所述一个或更多个用户。

20.根据权利要求19所述的机器可读介质，其中所述一个或更多个神经网络包括用于提供所述一个或更多个数据异常的实例和置信数据的音频异常检测器和视频异常检测器，所述音频异常检测器将针对所述一个或更多个用户的环境捕获的音频数据作为输入，并且所述视频异常检测器将针对所述一个或更多个用户的所述环境捕获的视频数据作为输入。

21.根据权利要求20所述的机器可读介质，其中所述一个或更多个神经网络包括用于确定所述一个或更多个数据异常中的每一个数据异常的分类的事件检测器。

22.根据权利要求21所述的机器可读介质，其中所述一个或更多个神经网络包括用于确定是否使得所述一个或更多个用户被通知的决策者网络，所述决策者网络使用所述实例和置信数据以及用于所述一个或更多个数据异常中的每一个数据异常的所述分类，来确定所述一个或更多个数据异常是否适用于所述一个或更多个用户。

23.根据权利要求22所述的机器可读介质，其中所述一个或更多个处理器进一步使用语音识别模块来向所述决策者网络提供与所述一个或更多个数据异常相关的检测到的语音的文本。

24.根据权利要求22所述的机器可读介质，其中所述决策者网络进一步确定是否采取动作以减小所述一个或更多个用户的体验的沉浸性。

25.一种用户通知系统，包括：

相机,用于捕获视频数据;

麦克风,用于捕获音频数据;

一个或更多个处理器,用于使用一个或更多个神经网络来检测所述视频数据和音频数据中的一个或更多个数据异常,并使得一个或更多个用户被通知所述一个或更多个异常,其中所述一个或更多个用户被通知取决于所述一个或更多个数据异常是否适用于所述一个或更多个用户;以及

存储器,用于存储所述一个或更多个神经网络的网络参数。

26.根据权利要求25所述的用户通知系统,其中所述一个或更多个神经网络包括用于提供所述一个或更多个数据异常的实例和置信数据的音频异常检测器和视频异常检测器。

27.根据权利要求26所述的用户通知系统,其中所述一个或更多个神经网络包括用于确定所述一个或更多个数据异常中的每一个数据异常的分类的事件检测器。

28.根据权利要求27所述的用户通知系统,其中所述一个或更多个神经网络包括用于确定是否使得所述一个或更多个用户被通知的决策者网络,所述决策者网络使用所述实例和置信数据以及用于所述一个或更多个数据异常中的每一个数据异常的所述分类,来确定所述一个或更多个数据异常是否适用于所述一个或更多个用户。

29.根据权利要求28所述的用户通知系统,其中所述一个或更多个处理器进一步使用语音识别模块来向所述决策者网络提供与所述一个或更多个数据异常相关的检测到的语音的文本。

30.根据权利要求28所述的用户通知系统,其中所述决策者网络进一步确定是否采取动作以减小所述一个或更多个用户的体验的沉浸性。

使用一个或更多个神经网络确定的通知

技术领域

[0001] 至少一个实施例涉及用于执行和促进人工智能的处理资源。例如,至 少一个实施例涉及用于根据本文描述的各种新颖技术来训练神经网络的处 理器或计算系统。

背景技术

[0002] 用户越来越多地参与沉浸式活动,如观看虚拟现实(VR)内容或乘 坐自主车辆。对于这些沉浸式活动中的许多活动,用户可能不关注外部事 件,或者可能至少部分由于沉浸而不能看到或听到这些事件。如果不知道 这些外部事件中的一些,可能至少对于一些用户是不方便的,而不知道其 他外部事件的情况下,可能为用户提供安全或其他风险。

附图说明

[0003] 将参照附图描述根据本公开的各个实施例,在附图中:

[0004] 图1示出了根据至少一个实施例的用户环境;

[0005] 图2示出了根据至少一个实施例的异常确定系统的部件;

[0006] 图3A和图3B示出了根据至少一个实施例的可以为用户提供的通知;

[0007] 图4示出了根据至少一个实施例的用于检测异常并确定要采取的动作 的第一过程;

[0008] 图5示出了根据至少一个实施例的用于检测数据异常的第二过程;

[0009] 图6A示出了根据至少一个实施例的推理和/或训练逻辑;

[0010] 图6B示出了根据至少一个实施例的推理和/或训练逻辑;

[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8示出了根据至少一个实施例的计算机系统;

[0013] 图9示出了根据至少一个实施例的计算机系统;

[0014] 图10示出了根据至少一个实施例的计算机系统;

[0015] 图11示出了根据至少一个实施例的计算机系统;

[0016] 图12A示出了根据至少一个实施例的计算机系统;

[0017] 图12B示出了根据至少一个实施例的计算机系统;

[0018] 图12C示出了根据至少一个实施例的计算机系统;

[0019] 图12D示出了根据至少一个实施例的计算机系统;

[0020] 图12E和图12F示出了根据至少一个实施例的共享编程模型;

[0021] 图13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;

[0022] 图14A-14B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;

[0023] 图15A-15B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;

[0024] 图16示出了根据至少一个实施例的计算机系统;

[0025] 图17A示出了根据至少一个实施例的并行处理器;

[0026] 图17B示出了根据至少一个实施例的分区单元；

[0027] 图17C示出了根据至少一个实施例的处理集群；

[0028] 图17D示出了根据至少一个实施例的图形多处理器；

[0029] 图18示出了根据至少一个实施例的多图形处理单元(GPU)系统；

[0030] 图19示出了根据至少一个实施例的图形处理器；

[0031] 图20示出了根据至少一个实施例的处理器的微架构；

[0032] 图21示出了根据至少一个实施例的深度学习应用处理器；

[0033] 图22示出了根据至少一个实施例的示例神经元形态处理器；

[0034] 图23和图24示出了根据至少一个实施例的图形处理器的至少部分；

[0035] 图25示出了根据至少一个实施例的图形处理器核心的至少部分；

[0036] 图26A-26B示出了根据至少一个实施例的图形处理器核心的至少部 分；

[0037] 图27示出了根据至少一个实施例的并行处理单元(“PPU”)；

[0038] 图28示出了根据至少一个实施例的通用处理集群(“GPC”)；

[0039] 图29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；

[0040] 图30示出了根据至少一个实施例的流式多处理器；

[0041] 图31是根据至少一个实施例的用于高级计算管线的示例数据流程图；

[0042] 图32是根据至少一个实施例的用于在高级计算管线中训练、适配、示例化和部署机器学习模型的示例系统的系统图；

[0043] 图33包括根据至少一个实施例的用于处理成像数据的高级计算管线 的示例图示；

[0044] 图34A包括根据至少一个实施例的支持超声设备的虚拟仪器的示例 数据流程图；

[0045] 图34B包括根据至少一个实施例的支持CT扫描仪的虚拟仪器的示例 数据流程图；

[0046] 图35A示出了根据至少一个实施例的用于训练机器学习模型的过程 的数据流程图；

[0047] 图35B是根据至少一个实施例的用于利用预训练的注释模型来增强 注释工具的客户端-服务器架构的示例图示；

[0048] 图36A示出了根据至少一个实施例的自主车辆的示例；

[0049] 图36B根据至少一个实施例示出了图36A的自主车辆的相机位置和 视野的示例；

[0050] 图36C示出了根据至少一个实施例的用于图36A的自主车辆的示例 系统架构；以及

[0051] 图36D示出了根据至少一个实施例的用于在基于云的一个或更多个 服务器与图36A的自主车辆之间通信的系统。

具体实施方式

[0052] 在至少一个实施例中，用户可以参与或加入部分或全部沉浸式活动，如图1中的环境100的视图所示。在至少一个实施例中，用户可能正在玩 游戏或访问虚拟现实(VR)内容，其中用户正在关注该内容的音频和视频 方面，而不一定来自外部源。在至少一个实施例中，用户102可以正在访 问客户端设备104上呈现的内容，但是也可以佩戴着可以进一步帮助创建 沉浸式环境的附加设备或组件。在至少一个实施例中，这可以包括头戴式 耳机

106,其可以包括噪声消除头戴式耳机,因此用户102主要仅听到与 所呈现的沉浸式内容相关联的音频。在至少一个实施例中,用户102还可 以佩戴VR头盔或VR护目镜,其屏蔽该用户的视图免受该沉浸式体验外 部的内容的影响,因此用户102主要仅知晓作为沉浸式体验的一部分经由 该客户端设备呈现的音频、视频和其他内容。在至少一个实施例中,这样 的沉浸式体验可以尝试滤除尽可能多的外部内容,以便改进这种沉浸的真 实性。在至少一个实施例中,沉浸式环境112可以被认为是包括音频、视 觉、触觉和与游戏、媒体或体验有关的其他反馈或交互,该游戏、媒体或 体验是至少部分沉浸式的,因为用户应当主要仅意识到通过该沉浸式环境 提供的内容。

[0053] 然而,在至少一个实施例中,可能存在发生在该沉浸式环境之外的发 生(occurrence)或事件,对于该沉浸式环境,应当至少通知用户102。在 至少一个实施例中,这可以包括与用户102一起进入房间的另一个人114。

[0054] 在至少一个实施例中,这可包括呼叫用户102的姓名116的某人。在至少 一个实施例中,这可包括电话振铃120或与消息或通信相关联的其他通知。 在至少一个实施例中,这可包括来自诸如烟雾检测器、时钟、定时器或火 灾警报器之类的设备的警报118。在至少一个实施例中,这还可以包括所 关注的大声噪声122,诸如爆炸或螺丝刀。在至少一个实施例中,可存在 用户想要至少被通知或以其他方式意识到的不同其他类型的发生或事件。

[0055] 在至少一个实施例中,这可以针对其中用户可能不关注或意识到周围环境 中的事件的任何环境是相关的。

[0056] 在至少一个实施例中,可以捕获在该沉浸式环境112外部并且可能在 其内部提供发生或事件的信息的数据。在至少一个实施例中,诸如用户102 的头戴式耳机106上的网络摄像头110或摄像头之类的摄像头可以捕获该 环境的附近部分或视图的图像或视频数据。在至少一个实施例中,头戴式 耳机106上的麦克风108、网络摄像头110、客户端设备或其他此类源也可 捕获此环境的音频。在至少一个实施例中,也可提供其他类型的信息,诸 如来自生物或环境传感器、光传感器、运动传感器、智能设备等的信息。 在至少一个实施例中,这样的信息可被分析以尝试检测用户102可能感兴 趣或以其他方式适用于用户102的发生或事件。在至少一个实施例中,事 件或发生最初可被检测为所捕获的音频或视频数据中的任何异常。在至少 一个实施例中,可以针对任何这种异常做出关于是通知用户还是采取动作 的决定,其中通知或动作的类型可以至少部分地取决于检测到的异常的类 型、以及可以为该用户个性化的通知或动作的偏好。

[0057] 在至少一个实施例中,可以使用用于确定异常和相关动作的系统200, 如图2所示。在至少一个实施例中,这样的系统200可包括用于提供关于 附近发生或事件的信息的一个或更多个传感器或数据捕获设备。在至少一 个实施例中,这可包括至少一个麦克风202和至少一个相机204,其被定 位和配置成分别捕获一个或更多个感兴趣的用户所位于的环境的音频和视 频数据。在至少一个实施例中,由这些设备捕获的音频和视频(或图像)数据可由发生检测模块208分析以尝试检测异常。在至少一个实施例中, 一组感兴趣的事件可以被预定义或指定为异常。在至少一个实施例中,这 些事件在训练时间(用于初始或继续学习)被定义,并且可以包括事件或 发生,诸如由在当前视野内部或外部的人对用户姓名的会话提及。在至少 一个实施例中,这些感兴趣的事件还可以包括诸如跌倒、警报、电话呼叫、门铃或尖叫等异常声音,以及可视视野内的突然的或意外的改变。在至少 一个实

施例中,用户还可以具有将感兴趣的事件定义为针对至少某些类型 的事件的异常的能力。

[0058] 在至少一个实施例中,系统200可以提供对来自这些异常类别的选择 性操作和过滤功能两者,从而实现进一步的用户个性化程度。在至少一个 实施例中,系统200可包括一个决策者模型206和多个特征检测模型,诸 如所示的四个特征模型,以便以这种方式检测和处理事件。在至少一个实 施例中,特征检测模型可包括用于语音识别、听觉新颖检测、视觉异常检 测和事件检测和分类的模型,其中事件检测和分类可尝试对与检测到的场 景、对象、动作或识别出的音频相关的发生或事件进行分类。

[0059] 在至少一个实施例中,语音识别模块206可接收由麦克风202捕获的 音频数据。在至少一个实施例中,语音识别模块206可分析捕获的音频数 据并尝试将语音数据转换成文本数据。在至少一个实施例中,语音识别模 块206可尝试获得对该语音的上下文理解,并且可利用命名实体识别(NER) 来尝试检测对该用户(或相关人)的名字的说出(utterance)。在至少一个 实施例中,这样的过程也可以对在该语音数据中命名的实体进行分类。在 至少一个实施例中,诸如SOTA WaveNet模型之类的模型可以用于语音识 别和语音到文本转换。在至少一个实施例中,可训练来自变换器的双向编 码器表示(BERT)模型以检测可与所确定的用户相关或可应用于所确定的 用户的上下文信息。在至少一个实施例中,包括用户的名字或可以直接寻 址到用户的转换可以持有更高的权重。在至少一个实施例中,该文本数据 与任何确定的上下文或分类信息一起,可作为输入被提供给决策者模型 216。

[0060] 在至少一个实施例中,发生检测模块208或用于执行经分类的音频- 视觉异常检测的模块可以利用一个或更多个异常检测模型,诸如听觉异常 检测器210和视觉异常检测器212。在至少一个实施例中,还可以使用事 件检测器或分类器,其可以尝试对由这些异常检测器识别出的异常进行分 类或提供上下文。在至少一个实施例中,可以使用自动编码器(AE)来实 现异常检测。在至少一个实施例中,可以在音频、视频信号或两者中检测 异常。在至少一个实施例中,还可以利用统计异常检测。

[0061] 在至少一个实施例中,听觉新颖检测模块210利用一个或更多个对抗 式自动编码器来区分正常事件和异常事件,其中异常事件可被视为新颖 (novelty)。在至少一个实施例中,处理输入音频信号以计算特征,诸如 Log-Mel(来自Mel-Frequency Cepstrum,MFC)频谱特征。在至少一个实 施例中,对各个“正常”听觉信号训练的自动编码器可以用于从此输入音 频中的音频片段检测“异常”事件。在至少一个实施例中,对抗式策略用 于训练该自动编码器,其可涉及训练单独的鉴别器网络,以将该自动编码 器的输出与训练分布区分开,以及使用为该鉴别器的输出计算的二进制交 叉熵损失来训练该自动编码器。在至少一个实施例中,如果该自动编码器 的该输入和输出之间的欧几里德距离超过确定或指定的阈值,则检测到异 常。在至少一个实施例中,可以使用深度自回归网络来实现原始音频中的 这种异常检测。

[0062] 在至少一个实施例中,听觉异常检测器可以利用环境声音分类,所述 环境声音分类可以旨在识别在周围环境中检测到的声音的类型。在至少一 个实施例中,这种识别可以利用1-D、2-D或3-D卷积神经网络(CNN)。在至少一个实施例中,该CNN可学习将所捕获的声音转换成2D表示,或 者可直接从音频信号学习特征。在至少一个实施例中,长-短期记忆网络 (LSTM)架构也可以用于分类。

[0063] 在至少一个实施例中,视觉异常检测模块212可接受图像或视频数据 作为输入。

在至少一个实施例中,视觉异常检测基于对抗式视觉不规则性 检测(AVID),AVID可利用生成式对抗网络(GAN)来检测图像或视频 中的异常。在至少一个实施例中,这种GAN可以包括至少两个网络:生 成器网络和鉴别器网络。在至少一个实施例中,该生成器网络是卷积神经 网络(CNN),其学习从输入视频信号中去除逐像素不规则性,诸如通过 像素级修复(in-painting)。在至少一个实施例中,此鉴别器网络是学习检 测补丁级不规则性的全卷积神经网络(FCN)。在至少一个实施例中,基于 此生成器和此鉴别器的输出的交集来检测异常。

[0064] 在至少一个实施例中,事件检测器214或分类器可以包括充当单个分 类器模块的一组模型,其被配置为对输入信号中的该场景、对象、动作和 音频元数据进行标记。在至少一个实施例中,事件检测器214可与听觉异 常检测器和视觉异常检测器一起使用,以实现检测到的异常的音频-视觉异 常检测和分类。在至少一个实施例中,事件检测器可分析环境数据,诸如 表示环境的视频和音频数据,以尝试确定关于可适用于特定用户的该环境的信息。在至少一个实施例中,这包括分析环境的至少一部分的视频,以 尝试识别诸如人、对象和在该环境内发生或出现的动作之类的方面。在至 少一个实施例中,当对该环境中的一个或更多个检测到的异常进行分类或 以其他方式描述或涉及该环境中的一个或更多个检测到的异常时,可以生 成与这些方面相关联的关键字并将其提供给决策者模型216。在至少一个 实施例中,这可包括检测出现在用户后面或接近用户的人、环境中的烟雾 或火、动物的外观、或诸如自主车辆可见的外部危险。在至少一个实施例 中,该视频或图像数据可被分析以识别动作,诸如人向用户挥动或向用户 投掷某物。在至少一个实施例中,声音也可用于识别特定动作,诸如接近 该用户的人说出与该用户相关的语音的情况。在至少一个实施例中,这些 生成的关键字可以用于帮助决策者模型确定数据异常的类型,并且确定要 采取的适当动作。在至少一个实施例中,诸如通知之类的动作可包括描述 中的这些关键字中的一个或更多个关键字或提供上下文。在至少一个实施 例中,可以使用基于计算机视觉和基于机器学习的技术来处理此环境的此 捕获的音频和视频内容。在至少一个实施例中,一个或更多个经训练的人 工神经网络(ANN)可用于从所捕获的音频和视频中识别特征(诸如人、对象和动作),以生成对应于该环境的关键字集合。在至少一个实施例中, 经训练的ANN是从ResNet(残差神经网络)导出的网络或其他这样的网 络或其导出物。

[0065] 在至少一个实施例中,决策者模型216可以包括从这些个体特征模型 收集输入并且可以执行评估的前馈神经网络。在至少一个实施例中,这些 评估可考虑诸如这些输入是包括异常事件还是重大事件、该事件的潜在严 重性以及相关用户对该事件感兴趣或以其他方式适用于相关用户的可能性 的条件。在至少一个实施例中,该决策者模型216的输出是对应于所确定 的事件的这些以上状态的向量。在至少一个实施例中,该模型可以被馈送来自一个或更多个特征模型的异常输入,并且该模型可以使用有监督的机 器学习来训练,以便不同地处理来自每个模型的输入。在至少一个实施例 中,语音识别输出可特别地使用针对此用户的名字的命名实体识别(NER) 或包含此用户的上下文,指示输入语音被引向用户的概率。在至少一个实 施例中,听觉新颖和视觉异常检测模型仅指示该输入是否是异常的,并且 当检测到这样的异常时,可以用这些相同的输入来激活实体检测器214以 对它们进行分类。在至少一个实施例中,如可以基于具有这个事件检测器 输入结合这两个模型的有监督机器学习,检测到的异常的严重性可以由此 决策者模型216来估计。在至少一个实施例中,该估计的严重性可指示该 事件是具有低至中等严重性的通知级别事件,或具有高

严重性的警报级别 事件。在至少一个实施例中，如果通知事件被特定用户忽略，则此模型可 使用此作为反馈来过滤未来通知。在至少一个实施例中，可以存在分配给 不同类别的事件的预先确定的严重性，这可以提供更受控制的体验。在至 少一个实施例中，用户可指定对于某些类别的事件应做什么。在至少一个 实施例中，火警可以具有“紧急”严重性，而正在叫的某人的姓名可以具 有较低或正常的“会话”严重性级别。在至少一个实施例中，门铃或电话 可以具有“注意”严重性。在至少一个实施例中，用户可以指定要通知或 具有采取动作的严重性级别，并且可以能够针对如所提及的不同严重性分 类指定不同的动作或通知。

[0066] 在至少一个实施例中，可以采取动作，如图3A和图3B中所示的那 些。在至少一个实施例中，通知302可以被提供给用户，如图3A的界面 300所示。在至少一个实施例中，检测到未被确定为严重的数据异常，因 此确定简单的通知动作是适当的。在至少一个实施例中，这可包括通知相 关用户某人被确定为呼叫该用户的名字，因此该用户可确定是否采取适当 的动作，诸如暂停该游戏或降低音量。在至少一个实施例中，被确定为严 重(或以其他方式满足相关阈值或标准)的动作可导致内容被暂停或体验 的沉浸性被打破。在至少一个实施例中，这可包括暂停游戏或播放，以及 提供关于此暂停的信息，如图3B的界面350中所示。在至少一个实施例 中，用户可以具有恢复该沉浸式体验的选项。在至少一个实施例中，取决 于数据异常的类型，用户可能不能恢复此体验，直到可验证此数据异常的 安全性确定或中止为止。

[0067] 在至少一个实施例中，可以训练异常确定模型或决策者模型以在正常 数据活动与异常数据活动之间进行区分。在至少一个实施例中，这可包括 在房间中的两个其他人之间的交谈中出现的用户姓名与叫出该用户姓名的 人之间进行区分的能力。在至少一个实施例中，这可以考虑诸如体(volume) 和周围文本上下文之类的信息。在至少一个实施例中，这还可涉及分析视 频数据以尝试确定此语音是否指向此用户。在至少一个实施例中，这还可 包括确定特定人是否通常在该环境中，诸如办公室或家中的其他人，这些 人正进行期望的例行程序，使得该人的存在或动作不表示异常。在至少一 个实施例中，这可以基于用户反馈或动作随时间学习。在至少一个实施例 中，出于确定特定动作、事件或发生是否表示异常的目的，可以保持动作 的日志。在至少一个实施例中，用户还可能够指定哪些事件是异常，以及 要对每个事件采取的动作。

[0068] 在至少一个实施例中，自动编码器可以被训练成识别“正常”声音，所述“正常”声音可以是系统特定的或环境特定的。在至少一个实施例中，可针对多个环境或场景(诸如家庭环境、咖啡厅环境和游戏厅环境)训练 该模型。在至少一个实施例中，这样的模型可用于确定对于该特定类型的 环境不正常的声音。在至少一个实施例中，这可以包括音量水平的异常以 及声音的类型。在至少一个实施例中，可以使用具有对抗式训练的单独的 鉴别器来区分正常与“异常”或异常音频。在至少一个实施例中，可以使 用交叉熵损失来确定有效性。在至少一个实施例中，可以使用欧几里德距 离来确定事件是否应当被认为是异常，如本文其他地方所讨论的。在至少 一个实施例中，深度回归网络也可以用于异常检测。在至少一个实施例中，可训练模型以学习某些用户敏感度，以用于确定是否采取特定动作的目的。

[0069] 在至少一个实施例中，可以使用用于确定异常和适当动作的过程400，如图4所

示。在至少一个实施例中，可以使得402用户能够参与沉浸式体 验。在至少一个实施例中，这可包括用户佩戴头戴式耳机以观看VR内容、在环绕声音环境中观看电影、或乘坐自主车辆，其中该用户不必注意周围 环境并且可能不能看到或听到该车辆外部的发生或事件。在至少一个实施 例中，可以捕获404表示该用户的环境的音频和视频数据。在至少一个实 施例中，这可以包括用户参与沉浸式活动的房间，或者可以包括自主车辆 正在其中操作的外部环境。在至少一个实施例中，也可以捕获其他数据， 其他数据可以涉及光、运动、压力或温度等。在至少一个实施例中，可使 用一组异常检测器或能够检测多种类型的异常的一个检测器来分析406此 捕获的音频和视频数据，以从此捕获的数据识别一个或更多个异常。在至 少一个实施例中，还可使用语音检测器来分析该音频数据以生成表示说出 的语音的文本。在至少一个实施例中，可以使用事件分类器来分析408检 测到的异常的音频和视频数据，以便至少提供所检测到的异常的分类。在 至少一个实施例中，事件分类器可以提供与这些检测到的异常相关的关键 词集。在至少一个实施例中，一个或更多个异常的指示和置信度值可以连 同检测到的语音的文本一起提供410给系统、设备或服务(诸如决策模块)。 在至少一个实施例中，该决策模块可确定412是否应针对一个或更多个检 测到的异常采取一个或更多个动作。在至少一个实施例中，该确定可基于 诸如异常的类型和检测该异常的置信度之类信息。在至少一个实施例中， 如果确定414不采取动作，则此过程可用新捕获的环境数据来继续。在至 少一个实施例中，如果确定要采取动作，则可以确定416要采取的这些一 个或更多个动作。在至少一个实施例中，这可包括使用决策器模块来接受 异常的类型和置信度值作为输入，确定该异常的严重性，以及确定要采取 的一个或更多个适当动作。在至少一个实施例中，该过程然后可以使得418 采取这些一个或更多个动作，诸如通过指示生成通知或暂停沉浸式体验， 以及本文讨论和建议的其他可能动作。在至少一个实施方式中，耳机或头 戴式耳机中的噪声消除可暂停一段时间，以便用户听到某些环境噪声或声 音。

[0070] 在至少一个实施例中，可以使用用于确定数据异常并提供通知的过程 500，如图5所示。在至少一个实施例中，可以获得502与环境相关联的数 据，如可以包括音频和视频数据。在至少一个实施例中，该数据可用于确 定504一个或更多个数据异常。在至少一个实施例中，如果这些一个或更 多个数据异常适用于这些一个或更多个用户，则可以确定506通知一个或 更多个用户。

[0071] 在至少一个实施例中，在适当的情况下，可以在要安装在客户端设备 或服务器设备上的软件中提供这种功能。在至少一个实施例中，可提供所 有相关模型，其中用于检测模型的所确定的初始权重被提供有客户端软件。 在至少一个实施例中，可以使这些客户端本地地或通过网络向服务器提供 异常数据，该服务器可以做出适当的确定。在至少一个实施例中，至少轻 量级决策者模型可以被包括在客户端软件中，以便使得能够在客户端上做出至少一些确定。在至少一个实施例中，这些检测模型的输出向量可被发 送到服务器侧决策者模型，该服务器侧决策者模型可确定检测到的异常的 适当动作过程。

[0072] 在至少一个实施例中，此类过程可以实现经分类和个性化的音频视觉 异常检测，这可以进一步允许选择性地去除自主或沉浸式环境(如VR) 中的沉浸。在至少一个实施例中，沉浸的自动和选择性去除使得能够为VR 和自主车辆使用者实现更大程度的安全性和方便性，这可以帮助增加该技 术的采用。在至少一个实施例中，这可以为自主车辆的乘客

提供被动地监 测其环境并且决定是否对异常事件做出反应的能力,诸如帮助车辆已经出故障的某人或为搭便车者停车。在至少一个实施例中,对当前环境的视听 特征以及围绕先前暂停事件和用户定义的偏好的条件的监测使得能够选择 性地移除自主或沉浸式环境中的沉浸。

[0073] 推理和训练逻辑

[0074] 图6A示出了用于执行与一个或更多个实施例相关联的推理和/或训练 操作的推理和/或训练逻辑615。下面结合图6A和/或图6B提供关于推理 和/或训练逻辑615的细节。

[0075] 在至少一个实施例中,推理和/或训练逻辑615可以包括但不限于,代 码和/或数据存储器601,用于存储前向和/或输出权重和/或输入/输出数据, 和/或在一个或更多个实施例的各方面中用于配置被训练和/或用于推理的 神经网络的神经元或层的其他参数。在至少一个实施例中,训练逻辑615 可以包括或被耦合到代码和/或数据存储器601以存储图形代码或用于控制 时序和/或顺序的其他软件,其中将加载权重和/或其他参数信息来配置逻辑, 包括整数和/或浮点单元(统称为算术逻辑单元(ALU))。在至少一个实施 例中,代码(如图形代码)基于这个代码所对应的神经网络的架构将权重 或其他参数信息加载到处理器ALU中。在至少一个实施例中,代码和/或 数据存储器601存储神经网络的每个层的权重参数和/或输入/输出数据,该 神经网络在使用一个或更多个实施例的各方面的训练和/或推理期间在输 入/输出数据和/或权重参数的前向传播期间与一个或更多个实施例结合训 练或使用。在至少一个实施例中,代码和/或数据存储器601的任何部分可 与其他片上或片外数据存储(包括处理器的L1、L2或L3高速缓存或系统 存储器)包括在一起。

[0076] 在至少一个实施例中,代码和/或数据存储器601的任何部分可以在一 个或更多个处理器或其他硬件逻辑设备或电路的内部或外部。在至少一个 实施例中,代码和/或数据存储器601可以是高速缓存存储器、动态随机可 寻址存储器(“DRAM”)、静态随机可寻址存储器(“SRAM”)、非易失性 存储器(例如,闪存)或其他存储器。在至少一个实施例中,代码和/或数 据存储器601是处理器内部还是外部的选择,例如,或包括DRAM、SRAM、闪存或一些其他存储类型可取决于片上相对于片外的可用存储,正在执行 的训练和/或推理功能的等待时间要求、在神经网络的推理和/或训练中使用 的数据的批量大小,或这些因素的一些组合。

[0077] 在至少一个实施例中,推理和/或训练逻辑615可以包括但不限于:代 码和/或数据存储器605,所述代码和/或数据存储器605用于存储与在一个 或更多个实施例的方面中被训练和/或用于推理的神经网络的神经元或层 相对应的后向和/或输出权重和/或输入/输出数据。在至少一个实施例中, 代码和/或数据存储器605存储神经网络的每个层的权重参数和/或输入/输 出数据,该神经网络在使用一个或更多个实施例的各方面的训练和/或推理 期间的输入/输出数据和/或权重参数的向后传播期间与一个或更多个实施 例结合训练或使用。在至少一个实施例中,训练逻辑615可包括或被耦合 到代码和/或数据存储器605以存储图形代码或其他软件以控制时间和/或 顺序,其中将加载权重和/或其他参数信息来配置逻辑,包括整数和/或浮点 单元(统称为算术逻辑单元(ALU))。在至少一个实施例中,代码(如图 形代码)基于这个代码所对应的神经网络的架构将权重或其他参数信息加载到处理器ALU中。在至少一个实施例中,代码和/或数据存储605的任 何部分可与其他片上或片外数据存储(包括处理器的L1、L2或L3高速缓 存或系统存储器)包括在一起。在至少

一个实施例中,代码和/或数据存储 器605的任何部分可以是在一个或更多个处理器或其他硬件逻辑设备或电 路上的内部或外部的。在至少一个实施例中,代码和/或数据存储器605可 以是高速缓存存储器、DRAM、SRAM、非易失性存储器(例如,闪存) 或其他存储器。在至少一个实施例中,代码和/或数据存储器605是处理器 内部还是外部的选择,例如,或包括DRAM、SRAM、闪存或一些其他存 储类型可取决于片上相对于片外的可用存储器、正在执行的训练和/或推理 功能的等待时间要求、在神经网络的推理和/或训练中使用的数据的批量大 小或这些因素的一些组合。

[0078] 在至少一个实施例中,代码和/或数据存储器601和代码和/或数据存 储器605可以是单独的存储结构。在至少一个实施例中,代码和/或数据存 储器601和代码和/或数据存储器605可以是相同的存储结构。在至少一个 实施例中,代码和/或数据存储器601和代码和/或数据存储器605可以是部 分相同的存储结构和部分分开的存储结构。在至少一个实施例中,代码和/ 或数据存储器601和代码和/或数据存储器605的任何部分可与其他片上或 片外数据存储器(包括处理器的L1、L2或L3高速缓存或系统存储器)包 括在一起。

[0079] 在至少一个实施例中,推理和/或训练逻辑615可以包括但不限于,一 个或更多个算术逻辑单元(“ALU”)610,包括整数和/或浮点单元,至少 部分地基于或指示来执行逻辑和/或数学运算,训练和/或推理代码(例如, 图形代码),其结果可以产生存储在激活存储器620中的激活(例如,来自 神经网络内的层或神经元的输出值),所述激活是存储在代码和/或数据存 储器601和/或代码和/或数据存储器605中的输入/输出和/或权重参数数据的函数。在至少一个实施例中,存储在激活存储器620中的激活是根据由 一个或更多个ALU610响应于执行指令或其他代码而执行的线性代数和/ 或基于矩阵的数学运算来生成的,其中存储在代码和/或数据存储器605和 /或代码和/或数据存储器601中的权重值与其他值一起被用作操作数,诸如 偏置值、梯度信息、动量值或其他参数或超参数,这些参数或超参数中的 任一者或全部可被存储在代码和/或数据存储器605或代码和/或数据存储 器601或片上或片外的另一存储器中。

[0080] 在至少一个实施例中,一个或更多个ALU610被包括在一个或更多个 处理器或其他硬件逻辑设备或电路内,而在另一个实施例中,一个或更多 个ALU610可以在处理器或使用它们的其他硬件逻辑设备或电路(例如, 协处理器)的外部。在至少一个实施例中,ALU610可以包括在处理器的 执行单元内或者以其他方式包括在可由处理器的执行单元访问的ALU组 内,所述处理器的执行单元在同一处理器内或分布在不同类型的不同处理 器(例如,中央处理单元、图形处理单元、固定功能单元等)之间。在至 少一个实施例中,代码和/或数据存储器601、代码和/或数据存储器605, 并且激活存储器620可以在同一处理器或其他硬件逻辑设备或电路上,而 在另一个实施例中,它们可以在不同的处理器或其他硬件逻辑器件或电路 中,或相同和不同处理器或其他硬件逻辑设备或电路的某种组合。在至少 一个实施例中,激活存储器620的任何部分可以与其他片上或片外数据存 储器包括在一起,所述其他片上或片外数据存储器包括处理器的L1、L2 或L3高速缓存或系统存储器。此外,推理和/或训练代码可与处理器或其 他硬件逻辑或电路可访问的其他代码一起存储,并使用处理器的获取、解 码、调度、执行、引退和/或其他逻辑电路来获取和/或处理。

[0081] 在至少一个实施例中,激活存储器620可以是高速缓存存储器、DRAM、 SRAM、非易失性存储器(例如,闪存)或其他存储器。在至少一个实施 例中,激活存储器620可以完全地

或部分地在一个或更多个处理器或其他 逻辑电路之内或之外。在至少一个实施例中，选择激活存储器620是处理 器内部还是外部，例如，或包括DRAM、SRAM、闪存或一些其他存储类型可取决于片上相对于片外的可用存储，正在执行的训练和/或推理功能的 等待时间要求、在神经网络的推理和/或训练中使用的数据的批量大小、或 这些因素的一些组合。在至少一个实施例中，图6A中所示出的推理和/或 训练逻辑615可以与专用集成电路(“ASIC”)结合使用，如来自谷歌的 Tensorflow®处理单元、来自Graphcore™的推理处理单元(IPU)、或来自 英特尔公司的Nervana®(例如，“湖牙”)处理器。在至少一个实施例中， 图6A中所示出的推理和/或训练逻辑615可以与中央处理单元(“CPU”) 硬件、图形处理单元(“GPU”)硬件或其他硬件(如现场可编程门阵列 (“FPGA”))结合使用。

[0082] 图6B示出了根据至少一个实施例的推理和/或训练逻辑615。在至少 一个实施例中，推理和/或训练逻辑615可包括但不限于其中计算资源是专 用的或以其他方式结合与神经网络内的一个或更多个神经元层相对应的权 重值或其他信息来专门使用的硬件逻辑。在至少一个实施例中，图6B中 所示出的推理和/或训练逻辑615可以与专用集成电路(ASIC)结合使用， 如来自谷歌的Tensorflow®处理单元、来自Graphcore™的推理处理单元 (IPU)、或来自英特尔公司的Nervana®(例如，“Lake Crest”)处理器。 在至少一个实施例中，图6B中所示出的推理和/或训练逻辑615可以与中 央处理单元(CPU)硬件、图形处理单元(GPU)硬件或其他硬件(如现 场可编程门阵列(FPGA))结合使用。在至少一个实施例中，推理和/或训 练逻辑615包括但不限于代码和/或数据存储器601和代码和/或数据存储器 605，它们可以用于存储代码(例如，图形代码)、权重值和/或其他信息， 包括偏置值、梯度信息、动量值和/或其他参数或超参数信息。在图6B中 所示出的至少一个实施例中，代码和/或数据存储器601和代码和/或数据存 储器605中的每一个分别与专用计算资源(如计算硬件602和计算硬件606) 相关联。在至少一个实施例中，计算硬件602和计算硬件606中的每一个 包括一个或更多个ALU，该一个或更多个ALU仅分别对存储在代码和/或 数据存储器601和代码和/或数据存储器605中的信息执行数学函数(诸如 线性代数函数)，其结果被存储在激活存储器620中。

[0083] 在至少一个实施例中，代码和/或数据存储器601和605以及对应的计 算硬件602和606中的每一个，对应地，对应于神经网络的不同层，从而 使得来自代码和/或数据存储器601和计算硬件602的一个“存储/计算对 601/602”的所得激活被提供作为到代码和/或数据存储器605和计算硬件 606的“存储/计算对605/606”的输入，以便反映神经网络的概念组织。在 至少一个实施例中，存储/计算对601/602和605/606中的每一个可对应于 多于一个神经网络层。在至少一个实施例中，继存储计算对601/602和 605/606之后或与存储计算对601/602和605/606并行的附加存储器/计算对 (未示出)可被包括在推理和/或训练逻辑615中。

[0084] 数据中心

[0085] 图7示出了其中可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和应用层740。

[0086] 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 括资源编排器712、分组计算资源714和节点计算资源(“节点C.R.”)716 (1)-716(N)，其中“N”表示任何完

整的正整数。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)、存储器(例如，固态或磁盘驱动器)、网络输入/输出(“NW I/O”)装置、网络交换机、虚 拟机(“VM”)、功率模块和冷却模块等等。在至少一个实施例中，来自节 点C.R.716(1)-716(N)中的一个或更多个节点C.R.可以是具有上述计 算资源中的一个或更多个的服务器。

[0087] 在至少一个实施例中，已分组计算资源714可包括容纳在一个或更多 个机架(未示出)内的节点C.R.的单独分组或容纳在不同地理位置(也未 示出)处的数据中心内的许多机架。分组计算资源714内的节点C.R.的单 独分组可包括可被配置或分配来支持一个或更多个工作负荷的分组计算、网络、存储器或存储资源。在至少一个实施例中，包括CPU或处理器的若 干节点C.R.可以分组在一个或更多个机架内，以提供计算资源来支持一个 或更多个工作负荷。在至少一个实施例中，一个或更多个机架还可包括任 意组合的任意数量的功率模块、冷却模块和网络交换机。

[0088] 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，资源编排器712可以包括用于数据中心700的软件设计 基础设施(“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。

[0089] 在至少一个实施例中，如图7所示，框架层720包括工作调度器722、配置管理器724、资源管理器726和分布式文件系统728。在至少一个实施 例中，框架层720可包括支持软件层730的软件732和/或应用层740的一 个或更多个应用742的框架。在至少一个实施例中，软件732或应用742 可分别包括基于web的服务软件或应用，诸如由Amazon Web Services、Google Cloud和Microsoft Azure提供的那些。在至少一个实施例中，框架 层720可以是但不限于可以利用分布式文件系统728进行大规模数据处理 (例如，“大数据”)的自由和开源软件Web应用框架的类型，如Apache Spark™(下文称为“Spark”)。在至少一个实施例中，工作调度器722可 以包括Spark驱动器，用于促进由数据中心700的各层支持的工作负荷的 调度。在至少一个实施例中，配置管理器724可以能够配置不同层，诸如 软件层730和框架层720，其包括用于支持大规模数据处理的火花和分布 式文件系统728。在至少一个实施例中，资源管理器726可以能够管理被 映射到分布式文件系统728和作业调度器722或被分配用于支持分布式文 件系统728和作业调度器722的集群或分组的计算资源。在至少一个实施 例中，群集的或分组的计算资源可包括数据中心基础设施层710处的分组的计算资源714。在至少一个实施例中，资源管理器726可与资源协调器 712协调以管理这些映射或分配的计算资源。

[0090] 在至少一个实施例中，包括在软件层730中的软件732可包括由框架 层720的节点C.R.716(1)-716(N)、经分组的计算资源714和/或分布式 文件系统728的至少部分使用的软件。一种或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、数据库软件和流式 视频内容软件。

[0091] 在至少一个实施例中，包括在应用层740中的一个或更多个应用742 可包括由框架层720的节点C.R.716(1)-716(N)、经分组的计算资源714 和/或分布式文件系统728的至少部分使用的一个或更多个类型的应用。一 种或多种类型的应用可以包括但不限于任何

数量的基因组学应用、认知计 算和机器学习应用，包括训练或推理软件、机器学习框架软件(例如，PyTorch、TensorFlow、Caffe等)或结合一个或更多个实施例使用的其他 机器学习应用。

[0092] 在至少一个实施例中，配置管理器724、资源管理器726和资源编排 器712中的任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修改动作。在至少一个实施例中，自 修改动作可使数据中心700的数据中心操作者免于做出可能较差的配置决 策和可能避免数据中心的未充分利用和/或较差执行部分。

[0093] 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 他资源以训练一个或更多个机器学习模型或使用根据本文描述的一个或更 多个实施例的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，可以通过使用上文相对于数据中心700所描述的软件和 计算资源根据神经网络架构来计算权重参数来训练机器学习模型。在至少 一个实施例中，对应于一个或更多个神经网络的经训练的机器学习模型可用于通过使用经由本文描述的一个或更多个训练技术计算的权重参数来使 用上文相对于数据中心700描述的资源来推理或预测信息。

[0094] 在至少一个实施例中，数据中心可使用CPU、专用集成电路(ASIC)、GPU、FPGA或其他硬件来使用上述资源执行训练和/或推理。此外，上文 所描述的一个或更多个软件和/或硬件资源可经配置为允许用户训练或执 行对信息的推理的服务，例如图像辨识、语音识别或其他人工智能服务。

[0095] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中，推理和/或训练逻辑615可以在系统图 7中用于至少部分地基于使用在此描述的神经网络训练操作、神经网络功 能和/或架构或神经网络使用情况计算的权重参数来推理或预测操作。

[0096] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中，此逻辑可与这些图的组件一起使 用，以确定数据异常和适当的响应动作。

[0097] 计算机系统

[0098] 图8是示出根据至少一个实施例的示例性计算机系统的框图，该计算 机系统可以是具有互连的设备和部件的系统、片上系统(SOC)或用可以 包括执行指令的执行单元的处理器形成的其某种组合800。在至少一个实 施例中，计算机系统800可包括但不限于诸如处理器802的组件，其采用 包括根据本公开(诸如在本文中描述的实施例中)执行用于处理数据的算 法的逻辑的执行单元。在至少一个实施例中，计算机系统800可以包括处 理器，如可以从加利福尼亚州圣克拉拉市的英特尔公司获得的PENTIUM®处理器族、Xeon™、Itanium®、XScale™和/或StrongARM™、Intel®Core™或Intel®Nervana™微处理器，但是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800可执行可从华盛顿州雷蒙德市的微软公司获得的WINDOWS的操作系统的版本，但也可使用其他操作系统(例如UNIX和Linux)、嵌入式 软件和/或图形用户界面。

[0099] 实施例可以用在其他设备中，诸如手持设备和嵌入式应用。手持式设 备的一些示

例包括蜂窝电话、互联网协议设备、数码相机、个人数字助理 (“PDA”)以及手持式PC。在至少一个实施例中,嵌入式应用可以包括微 控制器、数字信号处理器(“DSP”)、片上系统、网络计算机(“NetPC”)、机顶盒、网络中枢、广域网(“WAN”)交换机、或可以根据至少一个实施例执行一个或更多个指令的任何其他系统。

[0100] 在至少一个实施例中,计算机系统800可以包括但不限于处理器802,所述处理器可以包括但不限于根据本文描述的技术执行机器学习模型训练 和/或推理的一个或更多个执行单元808。在至少一个实施例中,计算机系 统800是单处理器桌面或服务器系统,但是在另一个实施例中,计算机系 统800可以是多处理器系统。在至少一个实施例中,处理器802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、实现指令集的组合的处理 器或任何其他处理器设备,例如像数字信号处理器。在至少一个实施例中, 处理器802可以耦合到处理器总线810,处理器总线810可以在处理器802 和计算机系统800中的其他组件之间传输数据信号。

[0101] 在至少一个实施例中,处理器802可以包括但不限于1级(“L1”)内 部高速缓存存储器(“高速缓存”)804。在至少一个实施例中,处理器802 可具有单个内部高速缓存或多级内部高速缓存。在至少一个实施例中,高 速缓冲存储器可驻留在处理器802外部。其他实施例还可包括内部和外部 高速缓存两者的组合,这取决于特定实现和需要。在至少一个实施例中, 寄存器文件806可以将不同类型的数据存储在各种寄存器中,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。

[0102] 在至少一个实施例中,执行单元808(包括但不限于用于执行整数和 浮点操作的逻辑)也驻留在处理器802中。在至少一个实施例中,处理器 802还可以包括存储用于某些宏指令的微代码的微代码(“ucode”)只读存 储器(“ROM”)。在至少一个实施例中,执行单元808可以包括用于处理 打包指令集809的逻辑。在至少一个实施例中,通过在通用处理器802的 指令集中包括打包指令集809以及用于执行指令的相关联的电路,许多多 媒体应用所使用的操作可以使用通用处理器802中的打包数据来执行。在 一个或更多个实施例中,可通过使用处理器的数据总线的全宽度对紧缩数 据执行操作来更加有效地加速和执行许多多媒体应用,这可消除跨处理器 的数据总线传送较小的数据单元以一次一个数据元素执行一个或更多个操 作的需要。

[0103] 在至少一个实施例中,执行单元808还可以用于微控制器、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800可包括但不限于存储器820。在至少一个实施例中,存储 器820可以被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器设备。在至少一个实 施例中,存储器820可存储可由处理器802执行的数据信号表示的指令819 和/或数据821。

[0104] 在至少一个实施例中,系统逻辑芯片可耦接至处理器总线810和存储 器820。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”)816,并且处理器802可以经由处理器总线810与 MCH816通信。在至少一个实施例中,MCH816可以向存储器820提供高 带宽存储器路径818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,MCH816可以在处理器802、存储器820 和计算机系统800中的其他部件之间引导数据信号,并且在处理器总线810、 存储器820和系统I/O822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,MCH816可以通过高带宽存储器路径818耦合到存储器820,并且图 形/视频卡812可以通过加速图形端口(“AGP”)互连814耦合到MCH816。

[0105] 在至少一个实施例中,计算机系统800可以使用作为专用集线器接口 总线的系统I/O 822将MCH 816耦合至I/O控制器集线器(“ICH”)830。在至少一个实施例中,ICH 830可以经由本地I/O总线提供到一些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820、芯片组和处理器802的高速I/O总线。示 例可以包括但不限于音频控制器829、固件中枢(“闪存BIOS”)828、无 线收发器826、数据存储器824、包含用户输入和键盘接口825的传统I/O 控制器823、串行扩展端口827(如通用串行总线(“USB”)、和网络控制 器834)。数据存储824可包括硬盘驱动器、软盘驱动器、CD-ROM设备、闪存设备或其他大容量存储器。

[0106] 在至少一个实施例中,图8示出了一种系统,该系统包括互连的硬件 设备或“芯片”,而在其他实施例中,图8可以示出示范性片上系统(“SoC”)。在至少一个实施例中,图8中所示的设备可以与专用互连、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,计算机系统800的 一个或更多个组件使用快速计算链路(CXL)互连来互连。

[0107] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中,推理和/或训练逻辑615可以在系统图 8中用于至少部分地基于使用在此描述的神经网络训练操作、神经网络功 能和/或架构或神经网络使用情况计算的权重参数来推理或预测操作。

[0108] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中,此逻辑可与这些图的组件一起使 用,以确定数据异常和适当的响应动作。

[0109] 图9是示出根据至少一个实施例的用于利用处理器910的电子设备 900的框图。在至少一个实施例中,电子设备900可以是例如但不限于笔 记本、塔式服务器、机架式服务器、刀片式服务器、膝上型计算机、台式 计算机、平板计算机、移动设备、电话、嵌入式计算机、或任何其他合适 的电子设备。

[0110] 在至少一个实施例中,系统900可以包括但不限于通信地耦合到任何 合适数量或种类的部件、外围设备、模块或设备上的处理器910。在至少 一个实施例中,处理器910使用总线或接口(诸如1℃总线)耦合,系 统管理总线(“SMBus”)、低引脚计数(LPC)总线、串行外围接口(“SPI”), 高清晰度音频(“HDA”)总线、串行高级技术附件(“SATA”)总线、通 用串行总线(“USB”)(版本1、2、3),或通用异步接收器/发送器(“UART”) 总线。在至少一个实施例中,图9示出了一种系统,该系统包括互连的硬 件设备或“芯片”,而在其他实施例中,图9可以示出示范性片上系统 (“SoC”)。在至少一个实施例中,图9中所示出的设备可以与专用互连、 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,图 9的一个或更多个部件使用计算快速链路(CXL)互连来互连。

[0111] 在至少一个实施例中,图9可以包括显示器924、触摸屏925、触摸 板930、近场通信单元(“NFC”)945、传感器集线器940、热传感器946、 快速芯片组(“EC”)935、可信平台模块(“TPM”)938、BIOS/固件/闪存 (“BIOS,FW闪存”)922、DSP 960、驱动器920(如固态磁盘(“SSD”) 或硬盘驱动器(“HDD”))、无线局域网单元(“WLAN”)950、蓝牙单元 952、无线广域

网单元(“WWAN”)956、全球定位系统(GPS)955、相机 (“USB 3.0相机”)954(如USB 3.0相机)和/或在其中实施的低功率双倍 数据速率(“LPDDR”)存储器单元(“LPDDR3”)915,例如,LPDDR3 标准。这些组件可以各自以任何合适的方式实现。

[0112] 在至少一个实施例中,其他组件可通过上文所论述的组件通信地耦合 到处理器910。在至少一个实施例中,加速度计941、环境光传感器(“ALS”) 942、罗盘943和陀螺仪944可以通信地耦合至传感器集线器940。在至少 一个实施例中,热传感器939、风扇937、键盘936和触摸板930可通信地 耦合到EC935。在至少一个实施例中,扬声器963、耳机964和麦克风(“mic”) 965可以通信地耦合至音频单元(“音频编解码器和d类放大器”)962,该 音频单元进而可以通信地耦合至DSP960。在至少一个实施例中,音频单 元964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”)957可以通信地耦合到 WWAN单元956。在至少一个实施例中,可以在下一代形状因子(“NGFF”) 中实现诸如WLAN单元950和蓝牙单元952以及WWAN单元956之类的 组件。

[0113] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中,推理和/或训练逻辑615可以在系统图 9中用于至少部分地基于使用在此描述的神经网络训练操作、神经网络功 能和/或架构或神经网络使用情况计算的权重参数来推理或预测操作。

[0114] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中,此逻辑可与这些图的组件一起使 用,以确定数据异常和适当的响应动作。

[0115] 图10示出了根据至少一个实施例的计算机系统1000。在至少一个实 施例中,计算机系统1000被配置为实现贯穿本公开所描述的不同过程和方 法。

[0116] 在至少一个实施例中,计算机系统1000包括但不限于至少一个中央 处理单元(“CPU”)1002,该中央处理单元连接至使用任何合适的协议实 现的通信总线1010,如PCI(“外围部件互连”)、快速外围部件互连 (“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1000包括但不限于主存 储器1004和控制逻辑(例如,被实现为硬件、软件或其组合),并且数据 存储在主存储器1004中,该主存储器可以采取随机存取存储器(“RAM”) 的形式。在至少一个实施例中,网络接口子系统(“网络接口”)1022提供 到其他计算装置和网络的接口,用于从计算机系统1000接收数据和向其他 系统传输数据。

[0117] 在至少一个实施例中,计算机系统1000在至少一个实施例中包括但 不限于输入设备1008、并行处理系统1012、以及可以使用常规阴极射线管 (“CRT”)、液晶显示器(“LCD”)、发光二极管(“LED”)、等离子体显示 器、或其他适合的显示技术来实施的显示设备1006。在至少一个实施例中, 从诸如键盘、鼠标、触摸板、麦克风等输入设备1008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。

[0118] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中,推理和/或训练逻辑615可以在图10 的系统中用于至少部分地基于使用在此描述

的神经网络训练操作、神经网 络功能和/或架构或神经网络使用情况计算的权重参数来推理或预测操作。

[0119] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中,此逻辑可与这些图的组件一起使 用,以确定数据异常和适当的响应动作。

[0120] 图11示出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包括但不限于计算机1110和USB杆1120。在至少 一个实施例中,计算机1110可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1110包括但不 限于服务器、云示例、膝上型计算机和台式计算机。

[0121] 在至少一个实施例中,USB杆1120包括但不限于处理单元1130、USB 接口1140和USB接口逻辑1150。在至少一个实施例中,处理单元1130 可以是能够执行指令的任何指令执行系统、装置或设备。在至少一个实施 例中,处理单元1130可以包括但不限于任何数量和类型的处理单元(未示 出)。在至少一个实施例中,处理单元1130包括被优化用于执行与机器学 习相关联的任何量和类型的操作的专用集成电路(“ASIC”)。例如,在至 少一个实施例中,处理单元1130是被优化为执行机器学习推理操作的张量 处理单元(“TPC”)。在至少一个实施例中,处理单元1130是视觉处理单 元(“VPU”),其被优化用于执行机器视觉和机器学习推理操作。

[0122] 在至少一个实施例中,USB接口1140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1140是用于数据和 电力的USB3.0类型C插座。在至少一个实施例中,USB接口1140是USB3.0 类型A连接器。在至少一个实施例中,USB接口逻辑1150可以包括使得 处理单元1130能够经由USB连接器1140与设备(例如,计算机1110)接 口连接的任何量和类型的逻辑。

[0123] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中,推理和/或训练逻辑615可以在图11 的系统中用于至少部分地基于使用在此描述的神经网络训练操作、神经网 络功能和/或架构或神经网络使用情况计算的权重参数来推理或预测操作。

[0124] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中,此逻辑可与这些图的组件一起使 用,以确定数据异常和适当的响应动作。

[0125] 图12A示出了示范性架构,其中多个GPU1210-1213通过高速链路 1240-1243(例如,总线、点对点互连等)通信地耦合至多个多核处理器 1205-1206。在一个实施例中,高速链路1240-1243支持4GB/s、30GB/s、 80GB/s或更高的通信吞吐量。可以使用不同互连协议,包括但不限于 PCIe4.0或5.0和NVLink2.0。

[0126] 此外,并且在一个实施例中,GPU1210-1213中的两个或更多个GPU 在高速链路1229-1230上互连,所述高速链路1229-1230可以使用与用于 高速链路1240-1243的那些协议/链路相同或不同的协议/链路来实现。类似 地,多核处理器1205-1206中的两者或两者以上可经由高速链路1228连接, 所述高速链路1228可为以20GB/s、30GB/s、120GB/s或更高的速度操作 的对称多处理器(SMP)总线。可替代地,图12A中所示的不同系统部件 之间的

所有通信可以使用相同的协议/链路(例如,通过公共互连结构)来 实现。

[0127] 在一个实施例中,每个多核处理器1205-1206分别经由存储器互连 1226-1227通信地耦合到处理器存储器1201-1202,并且每个GPU1210-1213 分别通过GPU存储器互连1250-1253通信地耦合到GPU存储器1220-1223。存储器互连1226-1227和1250-1253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1201-1202和GPU存储器1220-1223 可以是易失性存储器,诸如动态随机存取存储器(DRAM)(包括堆叠式 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,诸如3D XPoint或 Nano-Ram。在一个实施例中,处理器存储器1201-1202的一些部分可以是 易失性存储器,并且另一部分可以是非易失性存储器(例如,使用两级存 储器(2LM)层级)。

[0128] 如下所述,尽管各个处理器1205-1206和GPU1210-1213可以分别物 理地耦合至特定存储器1201-1202、1220-1223,但是可以实现统一存储器 架构,其中相同的虚拟系统地址空间(也被称为"有效地址"空间)分布 在各个物理存储器中。例如,处理器存储器1201-1202可各自包括64GB 的系统存储器地址空间,且GPU存储器1220-1223可各自包括32GB的系 统存储器地址空间(在此示例中导致总共256GB可寻址存储器)。

[0129] 图12B示出了根据一个示范性实施例的多核处理器1207与图形加速 模块1246之间的互连的附加细节。图形加速模块1246可包含集成在线卡 上的一个或更多个GPU芯片,所述线卡经由高速链路1240耦合到处理器 1207。或者,图形加速模块1246可与处理器1207集成在同一封装或芯片 上。

[0130] 在至少一个实施例中,所示出的处理器1207包括多个核1260A-1260D, 每个核具有转换后备缓冲器1261A-1261D和一个或更多个高速缓存 1262A-1262D。在至少一个实施例中,核1260A-1260D可以包括用于执行 指令和处理数据的不同其他组件(未示出)。高速缓存1262A-1262D可包 括1级(L1)和2级(L2)高速缓存。另外,一个或更多个共享高速缓存1256可被包括在高速缓存1262A-1262D中并且由核集合1260A-1260D共 享。例如,处理器1207的一个实施例包括24个核,每个核具有其自己的 L1高速缓存、十二个共享L2高速缓存和十二个共享L3高速缓存。在该 实施例中,一个或更多个L2和L3高速缓存由两个相邻核共享。处理器1207 和图形加速模块1246与系统存储器1214连接,系统存储器1214可以包括图12A的处理器存储器1201-1202。

[0131] 经由一致性总线1264上的核间通信,为存储在不同高速缓存 1262A-1262D、1256和系统存储器1214中的数据和指令维持一致性。例如, 每个高速缓存可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1264进行通信。在 一个实现中,在一致性总线1264上实现高速缓存监听协议以监听高速缓存 访问。

[0132] 在一个实施例中,代理电路1225将图形加速模块1246通信地耦合至 一致性总线1264,从而允许图形加速模块1246作为核1260A-1260D的对 等体参与高速缓存一致性协议。具体地,接口1235通过高速链路1240(例 如,PCIe总线、NV链路等)提供到代理电路1225的连接,并且接口1237 将图形加速模块1246连接到链路1240。

[0133] 在一个实现方式中,加速器集成电路1236代表图形加速模块1246的 多个图形处理引擎1231、1232、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1231、1232、N可以各自包括单独 的图形处理单元(GPU)。可替代地,图形处理引擎1231、1232、N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)、采样器和位块传送引擎。在至少一个实 施例中,图形加速模块1246可以是具有多个图形处理引擎1231-1232、N 的GPU或图形处理引擎1231-1232、N可以是集成在公共封装、线卡或芯 片上的单独的GPU。

[0134] 在一个实施例中,加速器集成电路1236包括存储器管理单元(MMU) 1239,用于执行不同存储器管理功能,诸如虚拟到物理存储器转换(也称 为有效到真实存储器转换)和用于访问系统存储器1214的存储器访问协议。MMU1239还可包含用于高速缓存虚拟/有效到物理/真实地址转换的转换 后援缓冲器(TLB)(未示出)。在一个实现中,高速缓存1238存储供图形 处理引擎1231-1232、N高效访问的命令和数据。在一个实施例中,存储在 高速缓存1238和图形存储器1233-1234、M中的数据与核高速缓存 1262A-1262D、1256和系统存储器1214保持一致。如上所述,这可经由代 理电路1225代表高速缓存1238和存储器1233-1234、M(例如,将与处理 器高速缓存1262A-1262D、1256上的高速缓存行的修改/存取有关的更新发 送到高速缓存1238,以及从高速缓存1238接收更新)来完成。

[0135] 寄存器组1245存储由图形处理引擎1231-1232、N执行的线程的上下 文数据,并且上下文管理电路1248管理线程上下文。例如,上下文管理电 路1248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不同线 程的上下文(例如,在保存第一线程并且存储第二线程以使得图形处理引 擎可以执行第二线程的情况下)。例如,在上下文切换时,上下文管理电路 1248可将当前寄存器值存储到存储器中的指定区域(例如,由上下文指针 识别)。其然后可以在返回到上下文时恢复寄存器值。在一个实施例中,中 断管理电路1247接收并处理从系统设备接收的中断。

[0136] 在一个实施例中,来自图形处理引擎1231的虚拟/有效地址由 MMU1239转化为系统存储器1214中的真实/物理地址。加速器集成电路 1236的一个实施例支持多个(例如,4、8、16)图形加速器模块1246和/ 或其他加速器设备。图形加速器模块1246可以专用于在处理器1207上执 行的单个应用,或者可以在多个应用之间共享。在一个实施例中,呈现了 虚拟化图形执行环境,其中图形处理引擎1231-1232、N的资源与多个应用 或虚拟机(VM)共享。在至少一个实施例中,资源可以被细分为“片”, 这些片基于与VM和/或应用相关联的处理要求和优先级被分配给不同的 VM和/或应用。

[0137] 在至少一个实施例中,加速器集成电路1236作为用于图形加速模块 1246的系统的桥执行,并且提供地址转换和系统存储器高速缓存服务。此 外,加速器集成电路1236可以为主机处理器提供虚拟化设施以管理图形处 理引擎1231-1232、N的虚拟化中断和存储器管理。

[0138] 因为图形处理引擎1231-1232、N的硬件资源被显式地映射到由主处 理器1207看到的真实地址空间,所以任何主处理器可以直接使用有效地址 值来寻址这些资源。在一个实施例中,加速器集成电路1236的一个功能是 图形处理引擎1231-1232、N的物理分离,使得它们对于系统而言表现为独 立的单元。

[0139] 在至少一个实施例中,一个或更多个图形存储器1233-1234、M分别 耦合到图形处理引擎1231-1232、N中的每一个。图形存储器1233-1234、 M存储由图形处理引擎1231-1232、N中的每一个处理的指令和数据。图 形存储器1233-1234、M可以是易失性存储器,诸

如DRAM(包括堆叠式 DRAM)、GDDR存储器(例如,GDDR5、GDDR6)或HBM,和/或可以 是非易失性存储器,诸如3D XPoint或Nano-Ram。

[0140] 在一个实施例中,为了减少链路1240上的数据流量,使用偏置技术 来确保存储在图形存储器1233-1234、M中的数据是将由图形处理引擎 1231-1232、N最频繁使用并且优选地不被核1260A-1260D(至少不频繁) 使用的数据。类似地,偏置机制尝试保持核的高速缓存1262A-1262D、1256 内的核(并且优选地不是图形处理引擎1231-1232、N)和系统存储器1214 所需的数据。

[0141] 图12C示出了加速器集成电路1236被集成在处理器1207内的另一示 范性实施例。在至少这个实施例中,图形处理引擎1231-1232、N经由接口 1237和接口1235(其同样可以利用任何形式的总线或接口协议)在高速链 路1240上直接与加速器集成电路1236通信。加速器集成电路1236可执行 与关于图12B所描述的那些操作相同的操作,但是考虑到其非常接近一致 性总线1264和高速缓存1262A-1262D、1256,可能以更高的吞吐量执行操 作。至少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),其可以包括由加速器 集成电路1236控制的编程模型和由图形加速模块1246控制的编程模型。

[0142] 在至少一个实施例中,图形处理引擎1231-1232、N专用于单个操作 系统下的单个应用或进程。在至少一个实施例中,单个应用可以将其他应 用请求汇集到图形处理引擎1231-1232、N,从而提供VM/分区内的虚拟化。

[0143] 在至少一个实施例中,图形处理引擎1231-1232、N可以由多个VM/ 应用分区共享。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1231-1232、N,以允许每个操作系统访问。对于没有 管理程序的单分区系统,图形处理引擎1231-1232、N由操作系统拥有。在 至少一个实施例中,操作系统可以虚拟化图形处理引擎1231-1232、N以提 供对每个进程或应用的访问。

[0144] 在至少一个实施例中,图形加速模块1246或单独的图形处理引擎 1231-1232、N使用进程句柄来选择处理元件。在至少一个实施例中,处理 元件被存储在系统存储器1214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,进程句柄可以是在将其上下 文注册到图形处理引擎1231-1232、N(即,调用系统软件以将进程元件添 加到进程元件链表)时提供给主机进程的实现方式特定的值。在至少一个 实施例中,进程句柄的较低16位可以是处理元件链表内的处理元件的偏移。

[0145] 图12D示出了示例性加速器集成片1290。如本文中所使用的,“切片” 包括加速器集成电路1236的处理资源的指定部分。系统存储器1214内的 应用有效地址空间1282存储处理元件1283。在一个实施例中,响应于来 自在处理器1207上执行的应用1280的GPU调用1281而存储处理元件 1283。处理元件1283包含对应应用1280的处理状态。包含在处理元件1283 中的工作描述符(WD)1284可以是由应用请求的单个作业,或者可以包 含指向作业队列的指针。在至少一个实施例中,WD1284是指向应用的地 址空间1282中的作业请求队列的指针。

[0146] 图形加速模块1246和/或单独的图形处理引擎1231-1232、N可以被 系统中的进程的全部或子集共享。在至少一个实施例中,可以包括用于设 置进程状态并将WD1284发送到图形加速模块1246以开始虚拟化环境中 的作业的基础结构。

[0147] 在至少一个实施例中，专用进程编程模型是实现特定的。在该模型中，单个进程拥有图形加速模块1246或单独的图形处理引擎1231。因为图形 加速模块1246由单个进程拥有，所以当图形加速模块1246被分配时，管 理程序初始化用于拥有分区的加速器积分电路1236，并且操作系统初始化 用于拥有进程的加速器积分电路1236。

[0148] 在操作中，加速器集成片1290中的WD取回单元1291取回下一 WD1284，下一WD1284包括要由图形加速模块1246的一个或更多个图形 处理引擎完成的工作的指示。来自WD1284的数据可存储在寄存器1245 中且由MMU1239、中断管理电路1247和/或上下文管理电路1248使用， 如所说明。例如，MMU1239的一个实施例包括用于访问OS虚拟地址空间1285内的区段/页表1286的区段/页行走电路。中断管理电路1247可以处 理从图形加速模块1246接收的中断事件1292。当执行图形操作时，由图 形处理引擎1231-1232、N生成的有效地址1293，由MMU1239转换成真 实地址。

[0149] 在一个实施例中，相同的寄存器组1245针对每个图形处理引擎 1231-1232、N和/或图形加速模块1246被复制并且可以由管理程序或操作 系统初始化。这些复制寄存器的每一个可以包括在加速器集成片1290中。可由管理程序初始化的示范性寄存器在表1中示出。

[0150] 表1-管理程序初始化寄存器

[0151]

1	切片控制寄存器
2	真实地址（RA）调度进程区域指针
3	权限掩码覆盖寄存器
4	中断向量表条目偏移
5	中断向量表条目限制

[0152]

6	状态寄存器
7	逻辑分区标识
8	真实地址（RA）管理程序加速器利用记录指针
9	存储描述寄存器

[0153] 可由操作系统初始化的示范性寄存器在表2中示出。

[0154] 表2-操作系统初始化寄存器

[0155]

1	过程和线程识别
2	有效地址(EA)上下文保存/恢复指针
3	虚拟地址(VA)加速器利用记录指针
4	虚拟地址(VA)存储段表指针
5	权限掩码
6	工作描述符

[0156] 在一个实施例中，每个WD1284专用于特定图形加速模块1246和/ 或图形处理引擎

1231-1232、N。它包含图形处理引擎1231-1232、N进行 工作所需的所有信息，或者它可以是指向应用已经建立要完成的工作的命 令队列的存储器位置的指针。

[0157] 图12E示出了共享模型的一个示范性实施例的额外细节。该实施例包 括管理程序真实地址空间1298，处理元件列表1299存储在管理程序真实 地址空间1298中。管理程序真实地址空间1298可经由管理程序1296来访 问，管理程序1296虚拟化操作系统1295的图形加速模块引擎。

[0158] 在至少一个实施例中，共享的编程模型允许来自系统中的分区的全部 或子集的进程的全部或子集使用图形加速模块1246。存在图形加速模块 1246由多个进程和分区共享的两个编程模型：时间分片共享和图形导向共 享。

[0159] 在该模型中，系统管理程序1296拥有图形加速模块1246，并使其功 能对所有操作系统1295可用。为了使图形加速模块1246支持系统管理程 序1296的虚拟化，图形加速模块1246可以遵守以下各项：1)应用的作业 请求必须是自主的(也就是说，在作业之间不需要维持状态，或图形加速 模块1246必须提供上下文保存和恢复机制。应用的作业请求由图形加速模 块1246保证在指定的时间量(包括任何转换故障)内完成，或图形加速模 块1246提供抢先于作业的处理的能力。当在定向共享编程模型中操作时， 图形加速模块1246必须保证进程之间的公平性。

[0160] 在至少一个实施例中，应用1280被要求利用图形加速模块1246类型、工作描述符(WD)、权限掩码寄存器(AMR)值和上下文保存/恢复区域 指针(CSRP)来进行操作系统1295系统调用。在至少一个实施例中，图 形加速模块1246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1246类型可以是系统特定的值。在至少一个实施例 中，WD被专门格式化以用于图形加速模块1246，并且可以是图形加速模 块1246命令的形式，指向用户定义结构的有效地址指针、指向命令队列的 有效地址指针，或者用于描述要由图形加速模块1246完成的工作的任何其 他数据结构。在一个实施例中，AMR值是用于当前过程的AMR状态。在 至少一个实施例中，传递给操作系统的值类似于设置AMR的应用。如果 加速器集成电路1236和图形加速模块1246实现不支持用户授权掩码覆盖 寄存器(UAMOR)，则操作系统可在传递管理程序调用中的AMR之前将 当前UAMOR值应用于AMR值。管理程序1296可在将AMR放置到过程 元件1283中之前任选地应用当前权限掩码覆盖寄存器(AMOR)值。在至 少一个实施例中，CSRP是寄存器1245中的一个，其包含用于图形加速模 块1246保存和恢复上下文状态的应用的有效地址空间1282中的区域的有 效地址。如果在作业之间不需要保存状态或者当作业被抢占时，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。

[0161] 在接收到系统调用时，操作系统1295可验证应用1280已注册和已被 给予使用图形加速模块1246的权限。操作系统1295随后用表3中所示的 信息调用管理程序1296。

[0162] 表3-OS到管理程序的调用参数

[0163]

1	工作描述符（WD）

[0164]

2	权限掩码寄存器（AMR）值（潜在地被掩码）
3	有效地址（EA）上下文保存/恢复区域指针（CSRP）
4	进程 ID（PID）和可选的线程 ID（TID）
5	虚拟地址（VA）加速器使用记录指针（AURP）
6	存储段表指针的虚拟地址（SSTP）
7	逻辑中断服务号码（LISN）

[0165] 在接收到管理程序调用时，管理程序1296验证操作系统1295已经注 册并且已经给予使用图形加速模块1246的权限。管理程序1296然后将处 理元件1283放入相应图形加速模块1246类型的处理元件链表中。处理元 素可以包括表4中所示的信息。

[0166] 表4-处理元件信息

[0167]

1	工作描述符(WD)
2	权限掩码寄存器(AMR)值(潜在地被掩码)。
3	有效地址(EA)上下文保存/恢复区域指针(CSRP)
4	进程ID(PID)和可选的线程ID(TID)
5	虚拟地址(VA)加速器使用记录指针(AURP)
6	存储段表指针的虚拟地址(SSTP)
7	7逻辑中断服务号码(LISN)
8	中断向量表，从管理程序调用参数导出
9	状态寄存器(SR)值
10	逻辑分区ID(LPID)
11	真实地址(RA)管理程序加速器使用记录指针
12	存储描述符寄存器(SDR)

[0168] 在至少一个实施例中，管理程序初始化多个加速器集成片1290寄存 器1245。

[0169] 如图12F中所示，在至少一个实施例中，使用统一存储器，该统一存 储器可经由用于访问物理处理器存储器1201-1202和GPU存储器 1220-1223的公共虚拟存储器地址空间来寻址。在这种实现方式中，在GPU 1210-1213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1201-1202并且反之亦然，由此简化可编程性。在一个实施例中， 虚拟/有效地址空间的第一部分被分配给处理器存储器1201，第二部分被分 配给第二处理器存储器1202，第三部分被分配给GPU存储器1220，等等。 在至少一个实施例中，整个虚拟/有效存储器空间(有时被称为有效地址空 间)由此跨处理器存储器1201-1202和GPU存储器1220-1223中的每个分 布，从而允许任何处理器或GPU访问具有映射到该存储器的虚拟地址的任 何物理存储器。

[0170] 在一个实施例中，MMU 1239A-1239E中的一个或更多个MMU内的 偏置/一致性管理电路1294A-1294E确保一个或更多个主机处理器(例如， 1205)的高速缓存和GPU1210-1213之间的高速缓存一致性，并且实现指 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12F 中示出了偏置/一致性管理电路1294A-1294E的多个示例，但是偏置/

一致 性电路可以在一个或更多个主机处理器1205的MMU内和/或在加速器集 成电路1236内实现。

[0171] 一个实施例允许附接GPU的存储器1220-1223被映射为系统存储器 的一部分，并且使用共享虚拟存储器(SVM)技术来访问，但不遭受与全 系统高速缓存一致性相关联的性能缺陷。在至少一个实施例中，在没有繁 重的高速缓存一致性开销的情况下将GPU附接的存储器1220-1223作为系 统存储器访问的能力为GPU卸载提供了有益的操作环境。这种布置允许主 机处理器1205软件设置操作数并访问计算结果，而没有传统I/ODMA数 据副本的开销。此类传统副本涉及相对于简单存储器存取全部低效的驱动 器调用、中断和存储器映射I/O(MMIO)存取。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附接的存储器1220-1223的能 力对于已卸载计算的执行时间可以是关键的。在具有大量流式传输写入存 储器流量的情况下，例如，高速缓存相干性开销可显著减小GPU1210-1213 所见的有效写入带宽。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计算的效率可在确定GPU卸载的有效性中起作用。

[0172] 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追踪 器数据结构驱动。例如，可以使用偏置表，其可以是每附接GPU的存储器 页包含1或2位的页粒度结构(即，在存储器页的粒度下控制)。在至少一 个实施例中，可以在GPU附接的存储器1220-1223的便宜的存储器范围中 实现偏置表，在GPU1210-1213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)。或者，可在GPU内维持整个 偏置表。

[0173] 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 接的存储器1220-1223的每次访问相关联的偏置表项，引起以下操作。首 先，来自GPU1210-1213的在GPU偏置中找到其页的本地请求被直接转发 到对应的GPU存储器1220-1223。来自GPU的在主机偏置中找到其页的 本地请求被转发到处理器1205(例如，通过如上文所论述的高速链路)。 在一个实施例中，来自处理器1205的在主处理器中找到所请求的页的请求 偏置完成像正常存储器读取的请求。或者，可将指向GPU偏置页的请求转 发到GPU1210-1213。在至少一个实施例中，如果GPU当前不使用页，那 么GPU可接着将页转变到主机处理器偏置。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对于 有限的情况集合纯粹的基于硬件的机制来改变。

[0174] 一种用于改变偏置状态的机制采用API调用(例如，OpenCL)，该 API调用进而调用GPU的设备驱动器，该GPU的设备驱动器进而向GPU 发送消息(或使命令描述符入队)以引导其改变偏置状态并且对于一些转 换在主机中执行高速缓存刷新操作。在至少一个实施例中，高速缓存清空 操作用于从主机处理器1205偏置到GPU偏置的转变，但不用于相反的转变。

[0175] 在一个实施例中，通过暂时使GPU偏置的页不可由主机处理器1205 高速缓存来维持高速缓存相干性。为了访问这些页，处理器1205可以从 GPU1210请求访问，GPU1210可以或可以不立刻授权访问。由此，为了减 少处理器1205与GPU1210之间的通信，确保GPU偏置的页是GPU而非 主机处理器1205所需的页且反之亦然是有益的。

[0176] 推理和/或训练逻辑615用于执行一个或更多个实施例。下面结合图 6A和/或图6B提供关于推理和/或训练逻辑615的细节。

[0177] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中,此逻辑可与这些图的组件一起使 用,以确定数据异常和适当的响应动作。

[0178] 图13示出了根据本文描述的不同实施例的可以使用一个或更多个IP 核制造的示范性集成电路和相关联的图形处理器。除了所示出的内容之外, 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、外围接口控制器或通用处理器核。

[0179] 图13是根据至少一个实施例示出了可以使用一个或更多个IP核制造 的示例性片上系统集成电路1300的框图。在至少一个实施例中,集成电路 1300包括一个或更多个应用处理器1305(例如,CPU)、至少一个图形处 理器1310,并且可以另外包括图像处理器1315和/或视频处理器1320,它 们中的任何一个可以是模块化IP核。在至少一个实施例中,集成电路1300 包括外围或总线逻辑,该外围或总线逻辑包括USB控制器1325、UART 控制器1330、SPI/SDIO控制器1335和I2S/I2C控制器1340。在至少一个 实施例中,集成电路1300可包括耦合到高清晰度多媒体接口(HDMI)控 制器1350和移动行业处理器接口(MIPI)显示接口1355中的一个或更多 个的显示设备1345。在至少一个实施例中,存储可以由包括闪存和闪存控 制器的闪存子系统1360提供。在至少一个实施例中,存储器接口可以经由 存储器控制器1365提供,用于访问SDRAM或SRAM存储器设备。在至 少一个实施例中,一些集成电路另外包括嵌入式安全引擎1370。

[0180] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中,可以在集成电路1300中使用推理和/ 或训练逻辑615,用于至少部分地基于使用在此描述的神经网络训练操作、神经网络功能和/或架构或神经网络使用情况计算的权重参数来推理或预 测操作。

[0181] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中,此逻辑可与这些图的组件一起使 用,以确定数据异常和适当的响应动作。

[0182] 图14A-14B示出了根据本文描述的各种实施例的可使用一个或更多 个IP核制造的示例性集成电路和相关联的图形处理器。除了所示出的内容 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、外围接口控制器或通用处理器核。

[0183] 图14A-14B是示出了根据本文描述的实施例的在SoC内使用的示范 性图形处理器的框图。图14A示出了根据至少一个实施例的可使用一个或 更多个IP核制造的片上系统集成电路的示例性图形处理器1410。图14B 示出了根据至少一个实施例的可以使用一个或更多个IP核制造的片上系 统集成电路的附加示范性图形处理器1440。在至少一个实施例中,图14A 的图形处理器1410是低功率图形处理器核。在至少一个实施例中,图14B 的图形处理器1440是更高性能的图形处理器核。在至少一个实施例中,图 形处理器1410、1440中的每个图形处理器可以是图13的图形处理器1310 的变体。

[0184] 在至少一个实施例中,图形处理器1410包括顶点处理器1405和一个 或更多个片段处理器1415A-1415N(例如,1415A、1415B、1415C、1415D 至1415N-1和1415N)。在至少一个实施例中,图形处理器1410可经由单 独逻辑执行不同着色器程序,使得顶点处理器1405经优化以执行用于顶点 着色器程序的操作,而一个或更多个片段处理器1415A到1415N执行

用于 片段或像素着色器程序的片段(例如,像素)着色操作。在至少一个实施 例中,顶点处理器1405执行3D图形管线的顶点处理级且产生图元和顶点 数据。在至少一个实施例中,一个或更多个片段处理器1415A-1415N使用 由顶点处理器1405生成的图元和顶点数据来产生显示在显示设备上的帧 缓冲器。在至少一个实施例中,片段处理器1415A-1415N经优化以执行如 OpenGL API中提供的片段着色器程序,其可用以执行与如直接3D API中 提供的像素着色器程序类似的操作

[0185] 在至少一个实施例中,图形处理器1410另外包括一个或更多个存储 器管理单元(MMU)1420A-1420B、一个或更多个高速缓存1425A-1425B 和一个或更多个电路互连1430A-1430B。在至少一个实施例中,一个或更 多个MMU 1420A-1420B为图形处理器1410提供虚拟到物理地址映射,包 含顶点处理器1405和/或片段处理器1415A到1415N,其可参考存储在存 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 1425A-1425B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 1420A-1420B可与系统内的其他MMU同步,包含与一个 或更多个应用程序处理器1305、图像处理器1315相关联的一个或更多个 MMU,和/或图13的视频处理器1320,从而使得每个处理器1305-1320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1430A-1430B使得图形处理器1410能够经由SoC的内部总线 或经由直接连接与SoC内的其他IP核对接。

[0186] 在至少一个实施例中,图形处理器1440包括图14A的图形处理器 1410的一个或更多个MMU 1420A-1420B、一个或更多个高速缓存 1425A-1425B和一个或更多个电路互连1430A-1430B。在至少一个实施例 中,图形处理器1440包含一个或更多个着色器核1455A到1455N(例如, 1455A、1455B、1455C、1455D、1455E、1455F到1455N-1和1455N), 其提供统一着色器核架构,其中单个核或类型或核可执行所有类型的可编 程着色器代码,包含用以实施顶点着色器、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,着色器核的数量可以变化。在至 少一个实施例中,图形处理器1440包括核间任务管理器1445,其用作线 程分派器以将执行线程分派到一个或更多个着色器核1455A-1455N和平铺单元1458以加速用于基于平铺块的再现的平铺操作,其中在图像空间中细 分场景的渲染操作,例如,以利用场景内的局部空间相干性或优化内部高 速缓存的使用。

[0187] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中,可以在集成电路14A和/或14B中使用 推理和/或训练逻辑615,用于至少部分地基于使用在此描述的神经网络训 练操作、神经网络功能和/或架构或神经网络使用情况计算的权重参数来推 理或预测操作。推理和/或训练逻辑615用于执行与一个或更多个实施例相 关联的推理和/或训练操作。在至少一个实施例中,此逻辑可与这些图的组 件一起使用,以确定数据异常和适当的响应动作。

[0188] 图15A-15B示出了根据本文所描述的实施例的附加的示范性图形处 理器逻辑。图15A示出了在至少一个实施例中可以包括在图13的图形处 理器1310内并且在至少一个实施例中可以是如图14B中的统一着色器核 1455A-1455N的图形核1500。图15B示出了在至少一个实施例中适合于部 署在多芯片模块上的高度并行的通用图形处理单元1530。

[0189] 在至少一个实施例中,图形核1500包括对于图形核1500内的执行 资源共用的共

享指令高速缓存1502、纹理单元1518、和高速缓存/共享存 储器1520。在至少一个实施例中，图形核1500可以包括多个切片 1501A-1501N或用于每个核的分区，并且图形处理器可以包括图形核1500 的多个示例。切片1501A-1501N可包括支持逻辑，该支持逻辑包括本地指 令高速缓存1504A-1504N、线程调度器1506A-1506N、线程分派器 1508A-1508N和寄存器组1510A-1510N。在至少一个实施例中，片 1501A-1501N可包括一组附加功能单元(AFU1512A-1512N)、浮点单元 (FPU1514A-1514N)、整数算术逻辑单元(ALU1516-1516N)、地址计算 单元(ACU1513A-1513N)、双精度浮点单元(DPFPU1515A-1515N)和矩 阵处理单元(MPU1517A-1517N)。

[0190] 在至少一个实施例中，FPU 1514A-1514N可执行单精度(32位)和 半精度(16位)浮点运算，而DPFPU 1515A-1515N执行双精度(64位) 浮点运算。在至少一个实施例中，ALU 1516A-1516N能够以8位、16位 和32位精度执行可变精度整数操作，并且能够被配置用于混合精度操作。 在至少一个实施例中，MPU 1517A-1517N还可被配置用于混合精度矩阵运算，包括半精度浮点和8位整数运算。在至少一个实施例中，MPU 1517A-1517N可以执行各种矩阵操作以加速机器学习应用框架，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 1512A-1512N可执行浮点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)。

[0191] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中，推理和/或训练逻辑615可以在图形核 1500中用于至少部分地基于使用在此描述的神经网络训练操作、神经网络 功能和/或架构或神经网络使用情况计算的权重参数来推理或预测操作。

[0192] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中，此逻辑可与这些图的组件一起使 用，以确定数据异常和适当的响应动作。

[0193] 图15B示出了在至少一个实施例中的通用处理单元(GPGPU)1530， 所述GPGPU可以被配置成用于使得能够由图形处理单元阵列执行高度并 行的计算操作。在至少一个实施例中，GPGPU1530可直接链接到 GPGPU1530的其他示例以创建多GPU群集以改进深度神经网络的训练速 度。在至少一个实施例中，GPGPU1530包含主机接口1532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1532是PCI高速接口。在 至少一个实施例中，主机接口1532可以是供应商特定的通信接口或通信结 构。在至少一个实施例中，GPGPU1530从主机处理器接收命令且使用全局 调度器1534来将与那些命令相关联的执行线程分配到计算群集 1536A-1536H的集合。在至少一个实施例中，计算集群1536A-1536H共享 高速缓存存储器1538。在至少一个实施例中，高速缓存存储器1538可以 充当计算集群1536A-1536H内的高速缓存存储器的高级高速缓存。

[0194] 在至少一个实施例中，GPGPU1530包含经由存储器控制器1542A到 1542B的集合与计算群集1536A到1536H耦合的存储器1544A到1544B。 在至少一个实施例中，存储器1544A-1544B可以包括不同类型的存储器设 备，包括动态随机存取存储器(DRAM)或图形随机存取存储器，诸如同 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。

[0195] 在至少一个实施例中，计算集群1536A-1536H各自包括图形核的集 合，如图15A的

图形核1500,所述图形核可以包括多种类型的整数和浮点 逻辑单元,所述整数和浮点逻辑单元可以在包括适合于机器学习计算的精 度范围内执行计算操作。例如,在至少一个实施例中,计算集群1536A-1536H中的每一个中的至少一个浮点单元子集可被配置成执行16 位或32位浮点运算,而不同的浮点单元子集可被配置成执行64位浮点运 算。

[0196] 在至少一个实施例中,GPGPU1530的多个示例可经配置以作为计算 群集操作。在至少一个实施例中,由计算集群1536A-1536H用于同步和数 据交换的通信跨实施例而变化。在至少一个实施例中,GPGPU1530的多个 示例通过主机接口1532通信。在至少一个实施例中,GPGPU1530包含I/O 集线器1539,其将GPGPU1530与GPU链路1540耦合,GPU链路1540使得能够直接连接到GPGPU1530的其他示例。在至少一个实施例中,GPU 链路1540耦合到使得能够在GPGPU1530的多个示例之间进行通信和同步 的专用GPU到GPU桥。在至少一个实施例中,GPU链路1540与高速互 连耦合以向其他GPGPU或并行处理器传输和接收数据。在至少一个实施 例中,GPGPU1530的多个示例位于单独的数据处理系统中并且经由可经由 主机接口1532访问的网络装置进行通信。在至少一个实施例GPU中,除 了主机接口1532之外或作为主机接口1532的替代,链路1540可经配置以 实现到主机处理器的连接。

[0197] 在至少一个实施例中,GPGPU1530可经配置以训练神经网络。在至 少一个实施例中,可以在推理平台内使用GPGPU1530。在其中GPGPU1530 用于推理的至少一个实施例中,相对于在GPGPU用于训练神经网络时, GPGPU可包含较少的计算群集1536A-1536H。在至少一个实施例中,与存 储器1544A-1544B相关联的存储器技术可以在推理配置和训练配置之间不 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, GPGPU1530的推理配置可支持推理特定指令。例如,在至少一个实施例中, 推理配置可提供对一个或更多个8位整数点积指令的支持,其可在针对所 部署的神经网络的推理操作期间使用。

[0198] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中,推理和/或训练逻辑615可以在 GPGPU1530中用于至少部分地基于使用在此描述的神经网络训练操作、神 经网络功能和/或架构或神经网络使用情况计算的权重参数来推理或预测 操作。

[0199] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中,此逻辑可与这些图的组件一起使 用,以确定数据异常和适当的响应动作。

[0200] 图16是示出根据至少一个实施例的计算系统1600的框图。在至少 一个实施例中,计算系统1600包括处理子系统1601,该处理子系统具有 经由互连路径进行通信的一个或更多个处理器1602和系统存储器1604, 该互连路径可以包括存储器集线器1605。在至少一个实施例中,存储器集 线器1605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1602内。在至少一个实施例中,存储器集线器1605经由通信链路 1606与I/O子系统1611耦合。在至少一个实施例中,I/O子系统1611包 括能够使计算系统1600能够从一个或更多个输入设备1608接收输入的I/O 中枢1607。在至少一个实施例中,I/O集线器1607可以使可以包括在一个 或更多个处理器1602中的显示控制器能够向一个或更多个显示设备 1610A提供输出。在至少一个实施例中,与I/O中枢1607耦合的一个或更 多个显示设备1610A可以包括本地、内部或嵌入式显示设备。

[0201] 在至少一个实施例中，处理子系统1601包括经由总线或其他通信链 路1613耦接到存储器集线器1605的一个或更多个并行处理器1612。在至 少一个实施例中，通信链路1613可以是任何数量的基于标准的通信链路技 术或协议之一，诸如但不限于PCI高速，或者可以是供应商特定的通信接 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1612形成 计算聚焦的并行或向量处理系统，该系统可以包括大量处理单元和/或处理集群，诸如许多集成核(MIC)处理器。在至少一个实施例中，一个或更 多个并行处理器1612形成图形处理子系统，该图形处理子系统可以将像素 输出到经由I/O中枢1607耦合的一个或更多个显示设备1610A中的一个。 在至少一个实施例中，一个或更多个并行处理器1612还可以包括显示控制 器和显示接口(未示出)以实现到一个或更多个显示设备1610B的直接连 接。

[0202] 在至少一个实施例中，系统存储单元1614可连接到I/O集线器1607 以提供用于计算系统1600的存储机制。在至少一个实施例中，I/O交换机 1616可以用于提供接口机制以启用I/O集线器1607与其他部件之间的连 接，诸如可以集成到一个或更多个平台中的网络适配器1618和/或无线网 络适配器1619，以及可经由一个或更多个附加设备1620添加的各种其他 设备。在至少一个实施例中，网络适配器1618可以是以太网适配器或另一 个有线网络适配器。在至少一个实施例中，无线网络适配器1619可以包括 Wi-Fi、蓝牙、近场通信(NFC)或包括一个或更多个无线电的其他网络设 备中的一个或更多个。

[0203] 在至少一个实施例中，计算系统1600可以包括未明确示出的其他部 件，包括USB或其他端口连接、光学存储驱动器、视频采集设备等，还可 以连接至I/O集线器1607。在至少一个实施例中，可以使用任何合适的协 议(如基于PCI(外围部件互连)的协议(例如，PCI快速))或其他总线 或点对点通信接口和/或一个或更多个协议(如NV链路高速互连)或互连 协议来实现将图16中的各个部件互连的通信路径。

[0204] 在至少一个实施例中，一个或更多个并行处理器1612合并对于图形 和视频处理优化的电路，包括例如视频输出电路，并且构成图形处理单元 (GPU)。在至少一个实施例中，一个或更多个并行处理器1612合并为通 用处理优化的电路。在至少一个实施例中，计算系统1600的组件可以与一 个或更多个其他系统元件集成在单个集成电路上。例如，在至少一个实施 例中，一个或更多个并行处理器1612、存储器集线器1605、一个或更多个 处理器1602和I/O集线器1607可以集成到片上系统(SoC)集成电路中。 在至少一个实施例中，计算系统1600的组件可以集成到单个包中以形成系 统级包(SIP)配置。在至少一个实施例中，计算系统1600的组件的至少 一部分可以集成到多芯片模块(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系统中。

[0205] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中，推理和/或训练逻辑615可以在系统1600 中用于至少部分地基于使用在此描述的神经网络训练操作、神经网络功能 和/或架构或神经网络使用情况计算的权重参数来推理或预测操作。

[0206] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中，此逻辑可与这些图的组件一起使 用，以确定数据异常和适当的响应动作。

[0207] 处理器

[0208] 图17A示出了根据至少一个实施例的并行处理器1700。在至少一个 实施例中，并行处理器1700的不同部件可使用一个或更多个集成电路设备 (诸如可编程处理器、专用集成电路(ASIC)或现场可编程门阵列(FPGA)) 来实现。在至少一个实施例中，所示出的并行处理器1700是根据示范性实 施例的图16中所示的一个或更多个并行处理器1612的变体。

[0209] 在至少一个实施例中，并行处理器1700包括并行处理单元1702。在 至少一个实施例中，并行处理单元1702包括允许与其他设备(包括并行处 理单元1702的其他示例)通信的I/O单元1704。在至少一个实施例中，I/O 单元1704可以直接连接到其他设备。在至少一个实施例中，I/O单元1704 经由使用集线器或交换机接口(诸如存储器集线器1605)与其他设备连接。 在至少一个实施例中，存储器集线器1605与I/O单元1704之间的连接形 成通信链路1613。在至少一个实施例中，I/O单元1704与主机接口1706 和纵横存储器1716连接，其中主机接口1706接收指向执行处理操作的命 令，并且纵横存储器1716接收指向执行存储器操作的命令。

[0210] 在至少一个实施例中，当主机接口1706经由I/O单元1704接收命令 缓冲器时，主机接口1706可以将执行那些命令的工作操作引导到前端1708。 在至少一个实施例中，前端1708与调度器1710耦合，该调度器被配置用 于向处理集群阵列1712分发命令或其他工作项。在至少一个实施例中，调 度器1710确保在任务被分发到处理集群阵列1712之前处理集群阵列1712 被适当地配置并且处于有效状态。在至少一个实施例中，通过在微控制器 上执行的固件逻辑来实现调度器1710。在至少一个实施例中，微控制器实 现的调度器1710可配置成以粗和细粒度执行复杂的调度和工作分配操作， 使得能够快速抢占和在处理阵列1712上执行的线程的上下文切换。在至少 一个实施例中，主机软件可以经由多个图形处理门铃之一证明用于在处理 阵列1712上调度的工作负荷。在至少一个实施例中，然后可以由包括调度 器1710的微控制器内的调度器1710逻辑跨处理阵列1712自动地分配工作 负荷。

[0211] 在至少一个实施例中，处理集群阵列1712可以包括多达“N”个处 理集群(例如，集群1714A、集群1714B至集群1714N)。在至少一个实 施例中，处理集群阵列1712的每个集群1714A-1714N可以执行大量并发 线程。在至少一个实施例中，调度器1710可以使用不同调度和/或工作分 配算法来将工作分配给处理集群阵列1712的集群1714A-1714N，这些调度和/或工作分配算法可以取决于针对每种类型的程序或计算而产生的工作 负荷而变化。在至少一个实施例中，调度可以由调度器1710动态地处理， 或者可以在配置用于由处理集群阵列1712执行的程序逻辑的编译期间部 分地由编译器逻辑辅助。在至少一个实施例中，处理集群阵列1712的不同 集群1714A-1714N可以被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。

[0212] 在至少一个实施例中，处理集群阵列1712可以被配置成用于执行不 同类型的并行处理操作。在至少一个实施例中，处理集群阵列1712被配置 为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1712 可以包括用于执行处理任务的逻辑，所述处理任务包括视频和/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。

[0213] 在至少一个实施例中，处理集群阵列1712被配置为执行并行图形处 理操作。在至少一个实施例中，处理集群阵列1712可以包括用于支持这样 的图形处理操作的执行的附

加逻辑,包括但不限于用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1712可以被配置成用于执行图形处理相关的着色器程序,如但 不限于顶点着色器、曲面细分着色器、几何形状着色器和像素着色器。在至少一个实施例中,并行处理单元1702可以经由I/O单元1704从系统存 储器传送数据以用于处理。在至少一个实施例中,在处理期间,可以在处 理期间将所传输的数据存储到片上存储器(例如,并行处理器存储器1722), 然后写回到系统存储器。

[0214] 在至少一个实施例中,当并行处理单元1702用于执行图形处理时, 调度器1710可以被配置为将处理工作负荷划分为大致相等大小的任务,以 更好地实现将图形处理操作分配到处理集群阵列1712的多个集群 1714A-1714N。在至少一个实施例中,处理集群阵列1712的部分可以被配 置成用于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 经配置以执行顶点着色和拓扑产生,第二部分可经配置以执行曲面细分和 几何形状着色,且第三部分可经配置以执行像素着色或其他屏幕空间操作 以产生经再现图像以供显示。在至少一个实施例中,由集群1714A-1714N 中的一个或更多个集群产生的中间数据可以存储在缓冲器中,以允许在集 群1714A-1714N之间传输中间数据以供进一步处理。

[0215] 在至少一个实施例中,处理集群阵列1712可以经由调度器1710接 收要执行的处理任务,调度器1710从前端1708接收定义处理任务的命令。 在至少一个实施例中,处理任务可以包括有待处理的数据的索引,例如, 表面(补丁)数据、图元数据、顶点数据、和/或像素数据、以及定义如何 处理数据的状态参数和命令(例如,执行什么程序)。在至少一个实施例中, 调度器1710可以被配置为提取对应于任务的索引或可以从前端1708接收 索引。在至少一个实施例中,前端1708可以被配置为确保处理集群阵列 1712被配置为在由进入的命令缓冲器(例如,批量缓冲器、推送缓冲器等) 指定的工作负荷被启动之前处于有效状态。

[0216] 在至少一个实施例中,并行处理单元1702的一个或更多个示例中的 每一个可与并行处理器存储器1722耦合。在至少一个实施例中,可以经由 存储器交叉1716访问并行处理器存储器1722,存储器交叉1716可以从处 理集群阵列1712以及I/O单元1704接收存储器请求。在至少一个实施例 中,存储器交叉1716可以经由存储器接口1718来访问并行处理器存储器 1722。在至少一个实施例中,存储器接口1718可包括多个分区单元(例如, 分区单元1720A、分区单元1720B至分区单元1720N),每个分区单元可 耦接至并行处理器存储器1722的一部分(例如,存储器单元)。在至少一 个实施例中,分区单元1720A-1720N的数量被配置为等于存储器单元的数 量,使得第一分区单元1720A具有对应的第一存储器单元1724A,第二分 区单元1720B具有对应的存储器单元1724B,第N分区单元1720N具有对 应的第N存储单元1724N。在至少一个实施例中,分区单元1720A-1720N 的数量可不等于存储装置的数量。

[0217] 在至少一个实施例中,存储器单元1724A-1724N可以包括不同类型 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,诸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1724A-1724N还可 以包括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1724A-1724N存储渲染目标(诸如帧缓冲器 或纹理图),从而允许分区单元1720A-1720N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1722的可用带宽。在至少一个实施例 中,并行处理器存储器1722的本地示例可被排除,以有利于结合本地高速 缓存存储器利用系统存储器的统一存储器设计。

[0218] 在至少一个实施例中,处理群集阵列1712的群集1714A-1714N中的 任一者可处理将被写入到并行处理器存储器1722内的存储器单元 1724A-1724N中的任一者的数据。在至少一个实施例中,内存交叉1716 可以被配置成用于将每个集群1714A-1714N的输出转移至任何分区单元 1720A-1720N或转移至另一个集群1714A-1714N,该集群可以对输出执行 额外的处理操作。在至少一个实施例中,每个集群1714A-1714N可以通过 存储器交叉1716与存储器接口1718通信以从不同外部存储器设备读取或 向其写入。在至少一个实施例中,存储器交叉1716具有到存储器接口1718 的连接,以与I/O单元1704通信,以及到并行处理器存储器1722的本地 示例的连接,使得不同处理集群1714A-1714N内的处理单元能够与系统存 储器或对并行处理单元1702而言不是本地的其他存储器通信。在至少一个 实施例中,内存纵横1716可以使用虚拟信道来分离集群1714A-1714N与 分区单元1720A-1720N之间的业务流。

[0219] 在至少一个实施例中,并行处理单元1702的多个示例可设置在单个 附加卡上,或多个附加卡可互连。在至少一个实施例中,并行处理单元1702 的不同示例可被配置成相互操作,即使不同示例具有不同数量的处理单元、不同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1702的一些示例可包括相对于其他示例更高精度的 浮点单元。在至少一个实施例中,合并并行处理单元1702或并行处理器1700的一个或更多个示例的系统可以用各种配置和形状因子来实现,包括 但不限于台式、膝上型或手持式个人计算机、服务器、工作站、游戏控制 台和/或嵌入式系统。

[0220] 图17B是根据至少一个实施例的分区单元1720的框图。在至少一个 实施例中,分区单元1720是图17A的分区单元1720A-1720N中的一个的 示例。在至少一个实施例中,分区单元1720包含L2高速缓冲存储器1721、帧缓冲器接口1725和光栅操作单元(“ROP”)1726。L2高速缓冲存储器 1721为经配置以执行从存储器交叉1716和ROP1726接收的加载和存储操 作的读取/写入高速缓冲存储器。在至少一个实施例中,读取未命中和紧急 回写请求由L2高速缓存1721输出到帧缓冲器接口1725用于处理。在至 少一个实施例中,更新还可经由帧缓冲器接口1725发送到帧缓冲器以供处 理。在至少一个实施例中,帧缓冲器接口1725与并行处理器存储器中的存 储器单元之一(诸如图17的存储器单元1724A-1724N(例如,在并行处理 器存储器1722内))对接。

[0221] 在至少一个实施例中,ROP1726为执行光栅操作(例如,模板、z 测试、混合等等)的处理单元。在至少一个实施例中,ROP1726接着输出 存储于图形存储器中的经处理图形数据。在至少一个实施例中,ROP1726 包含用以压缩写入到存储器的深度或色彩数据以及解压缩从存储器读取的 深度或色彩数据的压缩逻辑。在至少一个实施例中,压缩逻辑可以是利用 多个压缩算法中的一个或更多个的无损压缩逻辑。由ROP1726执行的压缩 逻辑可基于待压缩数据的统计特性而变化。例如,在至少一个实施例中, 基于每图块对深度和颜色数据执行增量颜色压缩。

[0222] 在至少一个实施例中,ROP1726包括在每个处理集群(例如,图17A 的集群1714A-1714N)内,而不是分区单元1720内。在至少一个实施例中, 通过存储器交叉1716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

可以被显示在显示设备(如 图16的一个或更多个显示设备1610之一)上,被路由以供处理器1602 进一步处理,或被路由以供图17A的并行处理器1700内的处理实体之一 进一步处理。

[0223] 图17C是根据至少一个实施例的并行处理单元内的处理集群1714的 框图。在至少一个实施例中,处理集群是图17A的处理集群1714A-1714N 之一的示例。在至少一个实施例中,一个或更多个处理集群1714中的一个 或更多个可以被配置成用于并行执行许多线程,其中"线程"指对特定数 据集执行的特定程序的示例。在至少一个实施例中,使用单指令多数据 (SIMD)指令发布技术来支持大量线程的并行执行,而不提供多个独立的 指令单元。在至少一个实施例中,单指令多线程(SIMT)技术用于使用公 共指令单元来支持大量一般同步的线程的并行执行,该公共指令单元被配 置成用于向处理集群中的每个处理集群内的处理引擎集合发布指令。

[0224] 在至少一个实施例中,处理集群1714的操作可以经由将处理任务分 配给SIMT并行处理器的管线管理器1732来控制。在至少一个实施例中, 管线管理器1732从图17A的调度器1710接收指令并且经由图形多处理器 1734和/或纹理单元1736管理那些指令的执行。在至少一个实施例中,图 形多处理器1734是SIMT并行处理器的示范性示例。然而,在至少一个实 施例中,在处理集群1714内可以包括不同类型的不同架构的SIMT并行处 理器。在至少一个实施例中,图形多处理器1734的一个或更多个示例可以 包括在处理集群1714内。在至少一个实施例中,图形多处理器1734可以 处理数据,并且数据交叉1740可以用于将经处理的数据分配到多个可能的 目的地(包括其他着色器单元)中的一个。在至少一个实施例中,管线管 理器1732可通过指定要被分发的已处理数据的目的地与数据交叉1740面 对面来促进已处理数据的分发。

[0225] 在至少一个实施例中,处理集群1714内的每个图形多处理器1734 可以包括一组相同的功能执行逻辑(例如,算术逻辑单元、加载-存储单元 等)。在至少一个实施例中,功能执行逻辑可以以管线方式配置,其中可以 在先前的指令完成之前发布新指令。在至少一个实施例中,功能执行逻辑 支持不同运算,包括整数和浮点算术、比较运算、布尔运算、位移位和不 同代数函数的计算。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。

[0226] 在至少一个实施例中,传输到处理集群1714的指令构成一个线程。 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,线程组对不同的输入数据执行程序。在至少一个实施 例中,可以将线程组内的每个线程分配给图形多处理器1734内的不同处理 引擎。在至少一个实施例中,线程组可以包括比图形处理器1734内的多个 处理引擎更少的线程。在至少一个实施例中,当线程组包括比多个处理引擎更少的线程时,一个或更多个处理引擎可以在该线程组被处理的周期期 间是空闲的。在至少一个实施例中,线程组还可以包括比图形处理器1734 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 形处理器1734内的处理引擎更多的线程时,可以在连续的时钟周期上执行 处理。在至少一个实施例中,可以在图形多处理器1734上同时执行多个线 程组。

[0227] 在至少一个实施例中,图形多处理器1734包括用于执行加载和存储 操作的内部高速缓存存储器。在至少一个实施例中,图形多处理器1734 可以放弃内部高速缓存并使用

处理集群1714内的高速缓存存储器(例如，L1高速缓存1748)。在至少一个实施例中，每个图形多处理器1734还可 访问在所有处理集群1714之间共享的并且可用于在线程之间传输数据的 分区单元(例如，图17A的分区单元1720A-1720N)内的L2缓存。在至 少一个实施例中，图形多处理器1734还可以访问片外全局存储器，片外全 局存储器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1702外部的任何存储器可以用 作全局存储器。在至少一个实施例中，处理集群1714包括图形多处理器 1734的多个示例可以共享公共指令和数据，它们可以存储在L1高速缓存 1748中。

[0228] 在至少一个实施例中，每个处理集群1714可以包括被配置成用于将 虚拟地址映射到物理地址的存储器管理单元(“MMU”)1745。在至少一个 实施例中，MMU1745的一个或更多个示例可驻留在图17A的存储器接口 1718内。在至少一个实施例中，MMU1745包含用于将虚拟地址映射到瓦 片的物理地址的页表条目(PTE)集合和任选地高速缓存行索引。在至少一个实施例中，MMU1745可以包括可以驻留在图形多处理器1734或L1 高速缓存或处理集群1714内的地址转换后备缓冲器(TLB)或高速缓存。 在至少一个实施例中，处理物理地址以分布表面数据访问局部性，从而允 许分区单元之间的有效请求交错。在至少一个实施例中，高速缓存行索引 可用于确定对高速缓存行的请求是命中还是未中。

[0229] 在至少一个实施例中，处理集群1714可以被配置为使得每个图形多 处理器1734耦合到纹理单元1736用于执行纹理映射操作，例如，确定纹 理样本位置、读取纹理数据和过滤纹理数据。在至少一个实施例中，从内 部纹理L1高速缓存(未示出)或从图形多处理器1734内的L1高速缓存 读取纹理数据，并且根据需要从L2高速缓存、本地并行处理器存储器或 系统存储器获取纹理数据。在至少一个实施例中，每个图形多处理器1734 向数据交叉1740输出经处理的任务，以向另一处理集群1714提供经处理 的任务一个或更多个以供进一步处理，或者经由存储器交叉1716将经处理 的任务一个或更多个存储在L2高速缓存、本地并行处理器存储器或系统 存储器中。在至少一个实施例中，预ROP1742(预光栅操作单元)经配置 以从图形多处理器1734接收数据，将数据引导到ROP单元，ROP单元可 位于如本文描述的分区单元(例如，图17A的分区单元1720A-1720N)中。 在至少一个实施例中，预ROP1742单元可执行用于颜色混合的优化、组织 像素颜色数据和执行地址转换。

[0230] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中，推理和/或训练逻辑615可以在图形处 理集群1714中用于至少部分地基于使用在此描述的神经网络训练操作、神 经网络功能和/或架构或神经网络使用情况计算的权重参数来推理或预测 操作。

[0231] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中，此逻辑可与这些图的组件一起使 用，以确定数据异常和适当的响应动作。

[0232] 图17D示出了根据至少一个实施例的图形多处理器1734。在至少一 个实施例中，图形多处理器1734与处理集群1714的管线管理器1732耦合。 在至少一个实施例中，图形多处理器1734具有执行管线，包括但不限于指 令高速缓存1752、指令单元1754、地址映射单元1756、寄存器文件1758、 一个或更多个通用图形处理单元(GPGPU)核1762以及一个或更

多个加 载/存储单元1766。GPGPU核1762和加载/存储单元1766经由存储器和高 速缓存互连1768与高速缓存存储器1772和共享存储器1770耦合。

[0233] 在至少一个实施例中,指令高速缓存1752从管线管理器1732接收 要执行的指令流。在至少一个实施例中,指令被高速缓存在指令高速缓存 1752中并且被分派以供指令单元1754执行。在至少一个实施例中,指令 单元1754可以将指令分派为线程组(例如,经线),其中每个线程组被指 派到一个或更多个GPGPU核1762内的不同执行单元。在至少一个实施例 中,指令可通过指定统一地址空间内的地址来访问本地、共享或全局地址 空间中的任一个。在至少一个实施例中,地址映射单元1756可用于将统一 地址空间中的地址转换成可由加载/存储单元1766访问的不同存储器地址 在至少一个实施例中,寄存器文件1758为图形多处理器1734的功 能单元提供一组寄存器。在至少一个实施例中,寄存器文件1758为连接到 图形处理器1734的功能单元(例如,GPGPU核1762、加载/存储单元1766) 的数据路径的操作数提供临时存储。在至少一个实施例中,寄存器文件 1758在每个功能单元之间划分,使得每个功能单元被分配寄存器文件1758 的专用部分。在至少一个实施例中,寄存器文件1758在由图形多处理器 1734执行的不同经线之间划分。

[0234] 在至少一个实施例中,GPGPU核1762可以各自包括用于执行图形 多处理器1734的指令的浮点单元(FPU)和/或整数算术逻辑单元(ALU)。GPGPU核1762可在架构上类似或可在架构上不同。在至少一个实施例中, GPGPU核的第一部分1762包括单精度FPU和整数ALU,而GPGPU核的 第二部分包括双精度FPU。在至少一个实施例中,FPU可实现用于浮点运 算的IEEE754-2008标准或实现可变精度浮点运算。在至少一个实施例中, 图形多处理器1734还可以包括一个或更多个固定功能或特殊功能单元,以 执行诸如复制矩形或像素混合操作的特定功能。在至少一个实施例中, GPGPU核中的一个或更多个还可以包括固定或特殊功能逻辑。

[0235] 在至少一个实施例中,GPGPU核1762包括能够对多个数据集执行 单个指令的SIMD逻辑。在至少一个实施例中,GPGPU核1762可以物理 地执行SIMD4、SIMD8和SIMD16指令,并且逻辑地执行SIMD1、SIMD2 和SIMD32指令。在至少一个实施例中,用于GPGPU核的SIMD指令可 在编译时间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,被配 置用于SIMT执行模型的程序的多个线程可以经由单个SIMD指令来执行。 例如,在至少一个实施例中,可以经由单个SIMD8逻辑单元并行地执行其 执行相同或相似操作的八个SIMT线程。

[0236] 在至少一个实施例中,存储器和高速缓存互连1768是将图形多处理 器1734的每个功能单元连接到寄存器文件1758和共享存储器1770的互连 网络。在至少一个实施例中,存储器和高速缓存互连1768是允许加载/存 储单元1766实现共享存储器1770与寄存器文件1758之间的加载和存储操 作的纵横互连。在至少一个实施例中,寄存器文件1758可以以与GPGPU 核1762相同的频率操作,因此在GPGPU核1762和寄存器文件1758之间 的数据传送是非常低的等待时间。在至少一个实施例中,可以使用共享存 储器1770来实现在图形多处理器1734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,高速缓冲存储器1772可用作例如数据高速缓 冲存储器,以对在功能单元与纹理单元1736之间传送的纹理数据进行高速 缓存。在至少一个实施例中,共享存储器1770还可以用作程序管理的高速缓存。在至少一个实施例中,在GPGPU核心1762上执行的线程除了存储 在高速缓冲存储器

1772内的自动高速缓存的数据之外还可编程地将数据 存储在共享存储器内。

[0237] 在至少一个实施例中,如本文中所描述的并行处理器或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、模式分析操作 和不同通用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,高速互连,如PCIe或NVLink)通信地耦合至主 机处理器/核。在至少一个实施例中,GPU可以集成在与核相同的封装或芯 片上,并且通过内部处理器总线/互连(即,封装或芯片内部)通信地耦合 到核。在至少一个实施例中,不管GPU连接的方式如何,处理器核心都可 以工作描述符中包含的命令/指令序列的形式向GPU分配工作。在至少一 个实施例中,GPU然后使用专用电路/逻辑来有效地处理这些命令/指令。

[0238] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑615的细节。在至少一个实施例中,推理和/或训练逻辑615可在图形多处 理器1734中用于至少部分地基于使用本文描述的神经网络训练操作、神经 网络功能和/或架构或神经网络使用情况计算的权重参数来推理或预测操 作。

[0239] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中,此逻辑可与这些图的组件一起使 用,以确定数据异常和适当的响应动作。

[0240] 图18示出了根据至少一个实施例的多GPU计算系统1800。在至少 一个实施例中,多GPU计算系统1800可包含处理器1802,其经由主机接 口交换机1804耦合到多个通用图形处理单元(GPGPU)1806A-D。在至少 一个实施例中,主机接口交换机1804是将处理器1802耦合到PCI快速总 线的PCI快速交换机设备,处理器1802可通过PCI快速总线与 GPGPU1806A-D通信。GPGPU1806A-D可经由高速点对点GPU到GPU 链路1816的集合互连。在至少一个实施例中,GPU到GPU链路1816经 由专用GPU链路连接到GPGPU1806A-D中的每一者。在至少一个实施例 中,P2PGPU链路1816实现GPGPU1806A-D中的每一个之间的直接通信, 而不需要处理器1802连接到的主机接口总线1804上的通信。在至少一个 实施例中,通过指向P2PGPU链路1816的GPU到GPU流量,主机接口 总线1804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装置 与多GPU计算系统1800的其他示例通信。虽然在至少一个实施例中,GPGPU1806A-D经由主机接口交换机1804连接到处理器1802,但是在至 少一个实施例中,处理器1802包括对P2PGPU链路1816的直接支持,并 且可以直接连接到GPGPU1806A-D。

[0241] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中,推理和/或训练逻辑615可以在多GPU 计算系统1800中用于至少部分地基于使用在此描述的神经网络训练操作、神经网络功能和/或架构或神经网络使用情况计算的权重参数来推理或预 测操作。

[0242] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中,此逻辑可与这些图的组件一起使 用,以确定数据异常和适当的响应动作。

[0243] 图19是根据至少一个实施例的图形处理器1900的框图。在至少一 个实施例中,图形处理器1900包括环形互连1902、管线前端1904、媒体 引擎1937和图形核1980A-1980N。在

至少一个实施例中,环形互连1902 将图形处理器1900耦合到其他处理单元,包括其他图形处理器或一个或更 多个通用处理器核。在至少一个实施例中,图形处理器1900是集成在多核 处理系统内的许多处理器之一。

[0244] 在至少一个实施例中,图形处理器1900经由环形互连1902接收批 量命令。在至少一个实施例中,传入命令由管线前端1904中的命令流转化 器1903解释。在至少一个实施例中,图形处理器1900包括用于经由一个 或更多个图形核1980A-1980N执行3D几何处理和媒体处理的可扩展执行 逻辑。在至少一个实施例中,对于3D几何处理命令,命令流转化器1903 向几何管线1936供应命令。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1903向与媒体引擎1937耦合的视频前端1934提供命 令。在至少一个实施例中,媒体引擎1937包括用于视频和图像后处理的视 频质量引擎(VQE)1930和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)1933引擎。在至少一个实施例中,几何管线1936 和媒体引擎1937各自为由至少一个图形核1980A提供的线程执行资源生 成执行线程。

[0245] 在至少一个实施例中,图形处理器1900包括以模块化核 1980A-1980N(有时称为核切片)为特征的可缩放线程执行资源,每个模 块化核具有多个子核1950A-1950N、1960A-1960N(有时称为核子切片)。 在至少一个实施例中,图形处理器1900可以具有任何数量的图形核1980A 至1980N。在至少一个实施例中,图形处理器1900包括具有至少第一子核1950A和第二子核1960A的图形核1980A。在至少一个实施例中,图形处 理器1900是具有单个子核的低功率处理器(例如,1950A)。在至少一个 实施例中,图形处理器1900包括多个图形核1980A-1980N,每个图形核包 括第一子核1950A-1950N的集合和第二子核1960A-1960N的集合。在至少 一个实施例中,第一子核1950A-1950N中的每个子核至少包括第一组执行单元1952A-1952N和媒体/纹理采样器1954A-1954N。在至少一个实施例 中,第二子核1960A-1960N中的每个子核至少包括第二组执行单元 1962A-1962N和采样器1964A-1964N。在至少一个实施例中,每个子核 1950A-1950N、1960A-1960N共享一组共享资源1970A-1970N。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。

[0246] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中,推理和/或训练逻辑615可以在图形处 理器1900中用于至少部分地基于使用在此描述的神经网络训练操作、神经 网络功能和/或架构或神经网络使用情况计算的权重参数来推理或预测操 作。

[0247] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中,此逻辑可与这些图的组件一起使 用,以确定数据异常和适当的响应动作。

[0248] 图20是示出根据至少一个实施例的可以包括执行指令的逻辑电路的 处理器2000的微架构的框图。在至少一个实施例中,处理器2000可以执 行指令,包括x86指令、ARM指令、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2000可以包括用于存储紧缩数据的寄存 器,诸如在利用来自加利福尼亚州圣克拉拉市的英特尔公司的MMX技术 实现的微处理器中的64位宽的MMXTM寄存器。在至少一个实施例中, 以整数和浮点形式两者可用的MMX寄存器可以用伴随单指令、多数据 (“SIMD”)和流SIMD扩展(“SSE”)指令的

紧缩数据元素进行操作。在 至少一个实施例中，涉及SSE2、SSE3、SSE4、AVX或超过(统称为“SSEx”) 技术的128位宽的XMM寄存器可以保持这种紧缩数据操作数。在至少一 个实施例中，处理器2000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。

[0249] 在至少一个实施例中，处理器2000包括有序前端(“前端”)2001，该有序前端用于获取有待执行的指令并且准备有待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2001可以包括若干单元。在至少一 个实施例中，指令预取器2026从存储器获取指令并且将指令馈送到指令解 码器2028，指令解码器2028进而解码或解释指令。例如，在至少一个实 施例中，指令解码器2028将接收到的指令解码成机器可以执行的被称为 “微指令”或“微操作”(也被称为“微操作”或“uop”)的一个或更多个 操作。在至少一个实施例中，指令解码器2028将指令解析为操作码以及相 应的数据和控制字段，微架构可以使用这些字段来执行根据至少一个实施 例的操作。在至少一个实施例中，追踪高速缓存2030可将解码的微操作汇 编成微操作队列2034中的程序有序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2030遇到复杂指令时，微代码ROM2032提供完成 操作所需的uop。

[0250] 在至少一个实施例中，一些指令可被转换成单个微操作，而其他指 令需要若干微操作来完成完整操作。在至少一个实施例中，如果需要多于 四个uop来完成指令，则指令解码器2028可以访问微代码ROM 2032以执 行指令。在至少一个实施例中，指令可被解码成小数量的uop用于在指令 解码器2028处进行处理。在至少一个实施例中，如果需要多个uop来完成 操作，指令可以被存储在微代码ROM2032内。在至少一个实施例中，根 据至少一个实施例，追踪高速缓存2030引用入口点可编程逻辑阵列 (“PLA”)以确定用于读取微代码序列以完成来自微代码ROM 2032的一 个或更多个指令的正确微指令指针。在至少一个实施例中，在微代码 ROM2032完成对指令的uop排序之后，机器的前端2001可以恢复从追踪 高速缓存2030获取uop。

[0251] 在至少一个实施例中，乱序执行引擎(“乱序引擎”)2003可以准备 用于执行的指令。在至少一个实施例中，乱序执行逻辑具有若干缓冲器以 平滑和重新排序指令流，以便在指令流沿着管线下行并且被调度用于执行 时优化性能。在至少一个实施例中，乱序执行引擎2003包括但不限于分配 器/寄存器重命名器2040、存储器微操作队列2042、整数/浮点微操作队列 2044、存储器调度器2046、快速调度器2002、慢速/通用浮点调度器(“慢 速/通用FP调度器”)2004，以及简单浮点调度器(“简单FP调度器”)2006。 在至少一个实施例中，快速调度器2002、慢速/通用浮点调度器2004和简 单浮点调度器2006在此还被统称为“uop调度器2002，2004，2006。”在 至少一个实施例中，分配器/寄存器重命名器2040分配每个uop需要以便 执行的机器缓冲器和资源。在至少一个实施例中，分配器/寄存器重命名器2040将逻辑寄存器重命名到寄存器文件中的条目上。在至少一个实施例中， 在存储器调度器2046和uop调度器2002，2004，2006之前，分配器/寄存 器重命名器2040还为两个uop队列之一中的每个uop分配条目，存储器 uop队列2042用于存储器操作，整数/浮点uop队列2044用于非存储器操 作。在至少一个实施例中，uop调度器2002、2004、2006基于uop的相关 输入寄存器操作数源的准备就绪和执行资源的可用性来确定uop何时准备 好执行，uop需要完成其操作。在至少一个实施例中，至少一个实施例的 快速调度器2002可以在主时钟周期的每一半上进行调度，而慢速/通用浮 点调度器2004和简单浮点调度器2006可以每个主处理器时钟周期调度一 次。在至少一个实施例中，uop调度器2002、2004、2006仲裁分派端口

以 调度uop用于执行。

[0252] 在至少一个实施例中，执行块2011包括但不限于整数寄存器文件/ 旁路网络2008，浮点寄存器文件/旁路网络(“FP寄存器文件/旁路网络”) 2010、地址生成单元(“AGU”)2012和2014，快速算术逻辑单元(ALU) (“快速ALU”)2016和2018、慢速算术逻辑单元(“慢速ALU”)2020， 浮点ALU(“FP”)2022和浮点移动单元(“FP移动”)2024。在至少一个实施例中，整数寄存器文件/旁路网络2008和浮点寄存器文件/旁路网络 2010在此也被称为“寄存器文件2008、2010。”在至少一个实施例中，AGU 2012和2014、快速ALU 2016和2018、慢速ALU 2020、浮点ALU 2022 和浮点移动单元2024在此也被称为“执行单元2012、2014、2016、2018、 2020、2022和2024。”在至少一个实施例中，执行块b11可以包括但不限 于任意数量(包括零)和类型的寄存器文件、旁路网络、地址生成单元和 执行单元以任意组合。

[0253] 在至少一个实施例中，寄存器文件2008、2010可被布置在uop调度 器2002、2004、2006与执行单元2012、2014、2016、2018、2020、2022 和2024之间。在至少一个实施例中，整数寄存器文件/旁路网络2008执行 整数操作。在至少一个实施例中，浮点寄存器文件/旁路网络2010执行浮 点运算。在至少一个实施例中，寄存器文件2008、2010中的每一个可包括但不限于旁路网络，该旁路网络可将尚未写入到寄存器文件中的刚完成的 结果旁路或转发到新的依赖uop。在至少一个实施例中，寄存器文件2008、 2010可以彼此通信数据。在至少一个实施例中，整数寄存器文件/旁路网络 2008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 三十二位数据，并且第二寄存器文件用于高阶三十二位数据。在至少一个 实施例中，浮点寄存器文件/旁路网络2010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从64至128位的操作数。

[0254] 在至少一个实施例中，执行单元2012、2014、2016、2018、2020、 2022、2024可以执行指令。在至少一个实施例中，寄存器文件2008、2010 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2000可以包括但不限于任何数量和组合的执行单元2012、2014、 2016、2018、2020、2022、2024。在至少一个实施例中，浮点ALU2022 和浮点移动单元2024可执行浮点、MMX、SIMD、AVX和SSE或其他操 作，包括专用机器学习指令。在至少一个实施例中，浮点ALU2022可包括 但不限于用于执行除法、平方根和余数microop的64位乘64位浮点除法 器。在至少一个实施例中，涉及浮点值的指令可用浮点硬件来处理。在至 少一个实施例中，ALU操作可以被传递到快速ALU2016、2018。在至少 一个实施例中，快速ALUS 2016、2018可以执行具有半个时钟周期的有效 等待时间的快速操作。在至少一个实施例中，大多数复杂的整数运算去往 慢速ALU2020，因为慢速ALU2020可以包括但不限于用于长等待时间类 型的运算的整数执行硬件，诸如乘法器、移位、标记逻辑和分支处理。在 至少一个实施例中，存储器加载/存储操作可以由AGUS 2012、2014执行。 在至少一个实施例中，快速ALU2016、快速ALU2018和慢速ALU2020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，可以实施快速 ALU2016、快速ALU2018和慢速ALU2020来支持各种数据位大小，包括 16、32、128、256等。在至少一个实施例中，浮点ALU2022和浮点移动 单元2024可以被实现为支持具有不同宽度的位的操作数范围。在至少一个 实施例中，浮点ALU2022和浮点移动单元2024可以结合SIMD和多媒体 指令对128位宽的紧缩数据操作数进行操作。

[0255] 在至少一个实施例中，uop调度器2002、2004、2006在父加载完成 执行之前分派依

赖的操作。在至少一个实施例中，由于uop可被推测性地 调度并在处理器2000中执行，处理器2000还可包括用于处理存储器未命 中的逻辑。在至少一个实施例中，如果数据加载在数据高速缓存中未命中， 则在管线中的飞行中可能存在使调度器具有临时不正确的数据的依赖操作。 在至少一个实施例中，重放机制追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，依赖操作可能需要被重放，并且独立的操作可被允 许完成。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 机制还可以被设计为捕获用于文本串比较操作的指令序列。

[0256] 在至少一个实施例中，术语“寄存器”可以指可以用作用于标识操 作数的指令的一部分的板上处理器存储位置。在至少一个实施例中，寄存 器可以是可从处理器外部(从编程器的角度)使用的那些寄存器。在至少 一个实施例中，寄存器可以不限于特定类型的电路。而是，在至少一个实 施例中，寄存器可以存储数据、提供数据、和执行在此描述的功能。在至 少一个实施例中，本文描述的寄存器可由处理器内的电路使用任何数量的 不同技术来实现，诸如专用物理寄存器、使用寄存器重命名的动态分配的 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 用于打包数据的八个多媒体SIMD寄存器。

[0257] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中，推理和/或训练逻辑615的部分或全部 可以结合到执行块2011和所示或未示出的其他存储器或寄存器中。例如， 在至少一个实施例中，在此描述的训练和/或推理技术可以使用在执行框 2011中示出的ALU中的一个或更多个。此外，可以将权重参数存储在片 上或片外存储器和/或寄存器(示出或未示出)中，这些存储器和/或寄存器 将执行块2011的ALU配置成用于执行在此描述的一个或更多个机器学习 算法、神经网络架构、用例、或训练技术。

[0258] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中，此逻辑可与这些图的组件一起使 用，以确定数据异常和适当的响应动作。

[0259] 图21示出了根据至少一个实施例的深度学习应用处理器2100。在至 少一个实施例中，深度学习应用处理器2100使用指令，所述指令如果由深 度学习应用处理器2100执行，则使得深度学习应用处理器2100执行贯穿 本公开所描述的过程和技术中的一些或全部。在至少一个实施例中，深度 学习应用处理器2100是专用集成电路(ASIC)。在至少一个实施例中，由 于执行一个或更多个指令或两者，应用处理器2100执行矩阵乘法运算“硬 连线”到硬件中。在至少一个实施例中，深度学习应用处理器2100包括但 不限于处理集群2110(1)-2110(12)，芯片间链路(“ICL”)2120(1)-2120 (12)、芯片间控制器(“ICC”)2130(1)-2130(2)，存储器控制器(“Mem Ctrlrs”)2142(1)-2142(4)、高带宽存储器物理层(“HBM PHY”)2144 (1)-2144(4)，管理控制器中央处理单元(“管理控制器CPU”)2150、快速外围组件互连控制器和直接存储器访问块(“PCIe控制器和DMA”) 2170，以及十六通道外围组件互连快速端口(“PCI高速x 16”)2180。

[0260] 在至少一个实施例中，处理集群2110可以执行深度学习操作，包括 基于权重参数的推理或预测操作，所述权重参数计算一个或更多个训练技 术，包括本文中描述的那些。

在至少一个实施例中,每个处理集群2110 可以包括但不限于任何数量和类型的处理器。在至少一个实施例中,深度 学习应用处理器2100可以包括任何数量和类型的处理集群2100。在至少 一个实施例中,芯片间链路2120是双向的。在至少一个实施例中,芯片间 链路2120和芯片间控制器2130使得多个深度学习应用处理器2100能够交 换信息,包括由执行在一个或更多个神经网络中体现的一个或更多个机器 学习算法产生的激活信息。在至少一个实施例中,深度学习应用处理器 2100可以包括任何数量(包括零)和类型的ICL2120和ICC2130。

[0261] 在至少一个实施例中,HBM2 2140提供总共32千兆字节(GB)的 存储器。HBM2 2140(i)与存储器控制器2142(i)和HBM PHY2144(i) 两者相关联。在至少一个实施例中,任何数量的HBM2 2140可以提供任何 类型和总量的高带宽存储器,并且可以与任何数量(包括零)和类型的存 储器控制器2142和HBM PHY 2144相关联。在至少一个实施例中,SPI、I2C、GPIO2160、PCIe控制器和DMA2170和/或PCIe2180可被以任何技 术上可行的方式启用任何数量和类型的通信标准的任何数量和类型的块替 换。

[0262] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中,深度学习应用处理器2100用于训练机 器学习模型(诸如神经网络)以预测或推理提供给深度学习应用处理器 2100的信息。在至少一个实施例中,深度学习应用处理器2100用于基于 已经由另一处理器或系统或由深度学习应用处理器2100训练的经训练的 机器学习模型(例如,神经网络)来推理或预测信息。在至少一个实施例 中,处理器2100可以用于执行本文所描述的一个或更多个神经网络使用情 况。

[0263] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中,此逻辑可与这些图的组件一起使 用,以确定数据异常和适当的响应动作。

[0264] 图22是根据至少一个实施例的神经元形态处理器2200的框图。在 至少一个实施例中,神经元形态处理器2200可从神经元形态处理器2200 外部的源接收一个或更多个输入。在至少一个实施例中,这些输入可被传 送给神经元形态处理器2200内的一个或更多个神经元2202。在至少一个 实施例中,神经元2202及其组件可使用包括一个或更多个算术逻辑单元 (ALU)的电路或逻辑来实现。在至少一个实施例中,神经元形态处理器 2200可包括但不限于数千或数百万个神经元示例2202,但可使用任何合适 数量的神经元2202。在至少一个实施例中,神经元2202的每个示例可包 括神经元输入2204和神经元输出2206。在至少一个实施例中,神经元2202 可生成可被传送给神经元2202的其他示例的输入的输出。例如,在至少一 个实施例中,神经元输入2204和神经元输出2206可以经由突触2208互连。

[0265] 在至少一个实施例中,神经元2202和突触2208可互连,使得神经 元形态处理器2200操作以处理或分析由神经元形态处理器2200接收的信 息。在至少一个实施例中,当通过神经元输入2204接收的输入超过阈值时, 神经元2202可传送输出脉冲(或“激发”或“尖峰”)。在至少一个实施例 中,神经元2202可对在神经元输入2204处接收的信号求和或积分。例如, 在至少一个实施例中,神经元2202可以被实现为漏泄积分激发神经元,其 中如果和(被称为“膜电位”)超过阈值,则神经元2202可以使用传递函 数(诸如S形或阈值函数)来生成输出(或“激发”)。在至少一个实施例 中,泄漏积分激发神经元可以将在神经元输入

2204处接收到的信号求和到 膜电位中，并且还可以应用衰减因子(或泄漏)以减小膜电位。在至少一 个实施例中，如果在神经元输入2204处接收到足够快速地超过阈值的多个 输入信号(即，在膜电位衰减太低而不能激发之前)，则泄漏积分激发神经 元可激发。在至少一个实施例中，可以使用接收输入、将输入集成到膜电 势中和衰减膜电势的电路或逻辑来实现神经元2202。在至少一个实施例中， 可以对输入求平均，或者可以使用任何其他合适的传递函数。此外，在至 少一个实施例中，神经元2202可包括但不限于：在将转移函数应用于神经 元输入2204的结果超过阈值时在神经元输出2206处生成输出尖峰的比较 器电路或逻辑。在至少一个实施例中，一旦神经元2202激发，它可通过例 如将膜电位重置为0或另一合适的默认值来忽略先前接收到的输入信息。 在至少一个实施例中，一旦膜电位被重置为0，神经元2202就可以在合适 的时间段(或不应期)之后恢复正常操作。

[0266] 在至少一个实施例中，神经元2202可以通过突触2208互连。在至 少一个实施例中，突触2208可操作用于将信号从第一神经元2202的输出 传送到第二神经元2202的输入。在至少一个实施例中，神经元2202可在 突触2208的多于一个示例上传送信息。在至少一个实施例中，神经元输出 2206的一个或更多个示例可以经由突触2208的示例连接到同一神经元 2202中的神经元输入2204的示例。在至少一个实施例中，生成要在突触 2208的示例上传送的输出的神经元2202的示例可相对于突触2208的该示 例被称为“突触前神经元”。在至少一个实施例中，接收经由突触2208的 示例传送的输入的神经元2202的示例可以相对于突触2208的该示例被称 为“突触后神经元”。因为神经元2202的示例可以从突触2208的一个或更 多个示例接收输入，并且还可以在突触2208的一个或更多个示例上传送输出，神经元2202的单个示例因此可以是“突触前神经元”和“突触后神经 元”两者，在至少一个实施例中，相对于突触的各个示例2208。

[0267] 在至少一个实施例中，神经元2202可被组织成一个或更多个层。神 经元2202的每个示例可具有一个神经元输出2206，该神经元输出2206可 通过一个或更多个突触2208扇出到一个或更多个神经元输入2204。在至 少一个实施例中，第一层2210中的神经元2202的神经元输出2206可连接 到第二层2212中的神经元2202的神经元输入2204。在至少一个实施例中， 层2210可以被称为“前馈层”。在至少一个实施例中，第一层2210的示例 中的神经元2202的每个示例可以扇出至第二层2212中的神经元2202的每 个示例。在至少一个实施例中，第一层2210可以被称为“全连接前馈层”。 在至少一个实施例中，第二层2212的示例中的神经元2202的每一个示例 可以扇出至第三层2214中的神经元2202的少于所有示例。在至少一个实 施例中，第二层2212可以被称为“稀疏连接的前馈层”。在至少一个实施 例中，第二层2212中的神经元2202可以扇出至多个其他层中的神经元 2202，包括扇出至(相同的)第二层2212中的神经元2202。在至少一个 实施例中，第二层2212可以被称为“回归层”。在至少一个实施例中，神 经元形态处理器2200可包括但不限于递归层和前馈层的任何合适的组合， 包括但不限于稀疏连接的前馈层和完全连接的前馈层两者。

[0268] 在至少一个实施例中，神经元形态处理器2200可包括但不限于用于 将突触2208连接到神经元2202的可重配置互连架构或专用硬连线互连。在至少一个实施例中，神经元形态处理器2200可包括但不限于允许突触基 于神经网络拓扑和神经元扇入/出按需要分配给不同神经元2202的电路或 逻辑。例如，在至少一个实施例中，突触2208可使用互连结构(诸如片上 网络)或具有专用连接来连接到神经元2202。在至少一个实施例中，可使 用

电路或逻辑来实现突触互连及其组件。

[0269] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中,此逻辑可与这些图的组件一起使 用,以确定数据异常和适当的响应动作。

[0270] 图23是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2300包括一个或更多个处理器2302和一个或更多个图形处理器2308,并且可以是单处理器桌面系统、多处理器工作站系统、或具有大量 处理器2302或处理器核2307的服务器系统。在至少一个实施例中,系统 2300是结合在片上系统(SoC)集成电路内以便在移动、手持或嵌入式设 备中使用的处理平台。

[0271] 在至少一个实施例中,系统2300可以包括或结合在基于服务器的游 戏平台中的游戏控制台(包括游戏和媒体控制台)、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2300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2300还可以包括可穿戴设备(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)、与可穿戴设备耦合或集成在可穿 戴设备内。在至少一个实施例中,处理系统2300是电视或机顶盒设备,其 具有一个或更多个处理器2302和由一个或更多个图形处理器2308生成的 图形界面。

[0272] 在至少一个实施例中,一个或更多个处理器2302各自包括用于处理 指令的一个或更多个处理器核2307,这些指令在被执行时执行用于系统和 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2307中的 每一个被配置为处理特定的指令集2309。在至少一个实施例中,指令集 2309可以促进复杂指令集计算(CISC)、精简指令集计算(RISC)或经由 超长指令字(VLIW)的计算。在至少一个实施例中,处理器核2307可以 各自处理不同的指令集2309,其可以包括促进对其他指令集的仿真的指令。 在至少一个实施例中,处理器核2307还可以包括其他处理设备,例如数字 信号处理器(DSP)。

[0273] 在至少一个实施例中,处理器2302包括高速缓冲存储器2304。在至 少一个实施例中,处理器2302可以具有单个内部高速缓存或多级内部高速 缓存。在至少一个实施例中,高速缓存存储器在处理器2302的不同组件之 间共享。在至少一个实施例中,处理器2302还使用外部高速缓存(例如, 3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2307之间共享该外部高速缓存。在至 少一个实施例中,寄存器文件2306附加地包括在处理器2302中,处理器 2302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2306可包括通用寄存器或其他寄存器。

[0274] 在至少一个实施例中,一个或更多个处理器2302与一个或更多个接 口总线2310耦接,以在处理器2302和系统2300中的其他部件之间传输通 信信号,诸如地址、数据或控制信号。在至少一个实施例中,接口总线2310 在一个实施例中可以是处理器总线,诸如直接媒体接口(DMI)总线的版 本。在至少一个实施例中,接口2310不限于DMI总线,并且可以包括一 个或更多个外围部件互连总线(例如,PCI、PCI高速)、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2302包括集成存储器控 制器2316和平台控制器集线器2330。在至少一个实施例中,存储器控制 器2316促进存储器设备与系统2300的其他组件之间的通信,而平台控制 器集线器(PCH)2330经由本地I/O总线提供到I/O设备的连接。

[0275] 在至少一个实施例中,存储器设备2320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、闪存存储器设备、相变存储器设备、或具有适合用作过程存储器的性能的某种其他存储器设 备。在至少一个实施例中,存储器设备2320可以作为系统2300的系统存 储器来操作,以存储数据2322和指令2321,以供在一个或更多个处理器2302执行应用或进程时使用。在至少一个实施例中,存储器控制器2316 还与可选的外部图形处理器2312耦合,该外部图形处理器可以与处理器 2302中的一个或更多个图形处理器2308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2311可以连接到处理器2302。在至少一个 实施例中,显示设备2311可以包括内部显示设备中的一个或更多个,如在 经由显示接口(例如,显示端口等)附接的移动电子设备或膝上型设备或 外部显示设备中。在至少一个实施例中,显示设备2311可以包括头戴式显 示设备(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。

[0276] 在至少一个实施例中,平台控制器集线器2330使得外围设备能够经 由高速I/O总线连接到存储器设备2320和处理器2302。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2346、网络控制器2334、固 件接口2328、无线收发器2326、触摸传感器2325、数据存储设备2324(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2324可 以经由存储接口(例如,SATA)或经由外围总线(诸如外围组件互连总线 (例如,PCI、PCI高速))连接。在至少一个实施例中,触摸传感器2325 可包括触摸屏传感器、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2326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 2328实现与系统固件的通信,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2334可以使能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2310耦合。 在至少一个实施例中,音频控制器2346是多通道高清晰度音频控制器。在 至少一个实施例中,系统2300包括用于将传统(例如,个人系统2(PS/2)) 设备耦合到系统的可选的传统I/O控制器2340。在至少一个实施例中,平 台控制器集线器2330还可连接到一个或更多个通用串行总线(USB)控制 器2342连接输入设备,诸如键盘和鼠标2343组合、相机2344或其他USB 输入设备。

[0277] 在至少一个实施例中,存储器控制器2316和平台控制器集线器2330 的示例可以集成到分立的外部图形处理器(如外部图形处理器2312)中。 在至少一个实施例中,平台控制器集线器2330和/或存储器控制器2316可 以在一个或更多个处理器2302外部。例如,在至少一个实施例中,系统 2300可以包括外部存储器控制器2316和平台控制器集线器2330,其可以 被配置为与处理器2302通信的系统芯片组内的存储器控制器集线器和外 围控制器集线器。

[0278] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中,推理和/或训练逻辑615的部分或全部 可以结合到图形处理器2300中。例如,在至少一个实施例中,在此描述的 训练和/或推理技术可以使用在图形处理器2312中实施的ALU中的一个或 更多个。此外,在至少一个实施例中,在此描述的推理和/或训练操作可以使用除了图6A或6B中所示出的逻辑之外的逻辑来完成。在至少一个实施 例中,权重参数可以存储在片上或片外存储器和/或寄存器(示出或未示出) 中,这些存储器和/或寄存器将

图形处理器2300的ALU配置成用于执行在 此描述的一个或更多个机器学习算法、神经网络架构、用例、或训练技术。

[0279] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中，此逻辑可与这些图的组件一起使 用，以确定数据异常和适当的响应动作。

[0280] 图24是根据至少一个实施例的具有一个或更多个处理器核 2402A-2402N、集成存储器控制器2414和集成图形处理器2408的处理器 2400的框图。在至少一个实施例中，处理器2400可以包括由虚线框表示 的多达并包括附加核2402N的附加核。在至少一个实施例中，处理器核 2402A-2402N中的每一个包括一个或更多个内部高速缓存单元 2404A-2404N。在至少一个实施例中，每个处理器核还可以访问一个或更 多个共享高速缓存单元2406。

[0281] 在至少一个实施例中，内部高速缓存单元2404A-2404N和共享高速 缓存单元2406表示处理器2400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2404A-2404N可以包括每个处理器核内 的至少一级指令和数据高速缓存以及一级或多级共享的中级高速缓存，诸 如2级(L2)、3级(L3)、4级(L4)或其他级别的高速缓存，其中外部 存储器之前的最高级别的缓存被分类为LLC。在至少一个实施例中，高速 缓存一致性逻辑维持不同高速缓存单元2406与2404A-2404N之间的一致 性。

[0282] 在至少一个实施例中，处理器2400还可以包括一个或更多个总线控 制器单元2416和系统代理核2410的集合。在至少一个实施例中，一个或 更多个总线控制器单元2416管理一组外围总线，诸如一个或更多个PCI 或PCI快速总线。在至少一个实施例中，系统代理核2410为不同处理器组 件提供管理功能。在至少一个实施例中，系统代理核2410包括用于管理对 不同外部存储器设备(未示出)的访问的一个或更多个集成存储器控制器 2414。

[0283] 在至少一个实施例中，处理器核2402A-2402N中的一个或更多个包 括支持同时多线程。在至少一个实施例中，系统代理核2410包括用于在多 线程处理期间协调和操作核2402A-2402N的组件。在至少一个实施例中， 系统代理核2410还可以包括功率控制单元(PCU)，其包括用于调节处理 器核2402A-2402N和图形处理器2408的一个或更多个功率状态的逻辑和 组件。

[0284] 在至少一个实施例中，处理器2400另外包括用于执行图形处理操作 的图形处理器2408。在至少一个实施例中，图形处理器2408与共享高速 缓存单元2406和系统代理核2410(包括一个或更多个集成存储器控制器 2414)耦合。在至少一个实施例中，系统代理核2410还包括用于将图形处 理器输出驱动到一个或更多个耦合的显示器的显示控制器2411。在至少一 个实施例中，显示控制器2411还可以是经由至少一个互连与图形处理器2408耦合的单独模块，或者可以集成在图形处理器2408内。

[0285] 在至少一个实施例中，基于环形的互连单元2412用于耦合处理器 2400的内部组件。在至少一个实施例中，可以使用替代的互连单元，诸如 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2408 经由I/O链路2413与环形互连2412耦合。

[0286] 在至少一个实施例中，I/O链路2413表示多种I/O互连中的至少一种， 包括促进不同处理器组件和高性能嵌入式存储器模块2418(诸如eDRAM 模块)之间的通信的封装上I/O互连。在至少一个实施例中，处理器核 2402A-2402N和图形处理器2408中的每一个使用嵌

入式存储器模块2418 作为共享的末级高速缓存。

[0287] 在至少一个实施例中,处理器核2402A-2402N是执行共同指令集架 构的同构核。在至少一个实施例中,处理器核2402A-2402N在指令集架构 (ISA)方面是异构的,其中处理器核2402A-2402N中的一个或更多个执 行共同指令集,而处理器核2402A-2402N中的一个或更多个其他核执行共 同指令集的子集或不同的指令集。在至少一个实施例中,处理器核2402A-2402N在微架构方面是异构的,其中具有相对较高功耗的一个或更 多个核与具有较低功耗的一个或更多个功率核耦合。在至少一个实施例中, 处理器2400可以实现在一个或更多个芯片上或者实现为SoC集成电路。

[0288] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中,推理和/或训练逻辑615的部分或全部 可以结合到处理器2400中。例如,在至少一个实施例中,在此描述的训练 和/或推理技术可以使用在图形处理器2312、一个或更多个图形核 2402A-2402N或图24中的其他部件中具体化的ALU中的一个或更多个。此外,在至少一个实施例中,在此描述的推理和/或训练操作可以使用除了 图6A或6B中所示出的逻辑之外的逻辑来完成。在至少一个实施例中,可 以将权重参数存储在片上或片外存储器和/或寄存器(示出或未示出)中, 这些存储器和/或寄存器将图形处理器2400的ALU配置成用于执行在此描 述的一个或更多个机器学习算法、神经网络架构、用例或训练技术。

[0289] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中,此逻辑可与这些图的组件一起使 用,以确定数据异常和适当的响应动作。

[0290] 图25是根据在此描述的至少一个实施例的图形处理器核2500的硬 件逻辑的框图。在至少一个实施例中,图形处理器核2500包括在图形核阵 列内。在至少一个实施例中,图形处理器核2500(有时被称为核切片)可 以是模块化图形处理器内的一个或更多个图形核。在至少一个实施例中, 图形处理器核2500是一个图形核切片的示例,并且如在此所描述的图形处 理器可以包括基于目标功率和性能包络的多个图形核切片。在至少一个实 施例中,每个图形核2500可以包括与多个子核2501A-2501F(也被称为子 切片)耦合的固定功能块2530,所述多个子核包括具有通用和固定功能逻 辑的模块化块。

[0291] 在至少一个实施例中,固定功能块2530包括可以由图形处理器2500 中的所有子核(例如,在较低性能和/或较低功率的图形处理器实现方式中) 共享的几何结构/固定功能管线2536。在至少一个实施例中,几何结构/固 定功能管线2536包括3D固定功能管线、视频前端单元、线程生成者和线 程分派器、以及管理统一返回缓冲器的统一返回缓冲器管理器。

[0292] 在至少一个实施例中,固定的功能块2530还包括图形SoC接口2537、图形微控制器2538和媒体管线2539。在至少一个实施例中固定,图形SoC 接口2537提供图形核2500与片上系统集成电路内的其他处理器核之间的 接口。在至少一个实施例中,图形微控制器2538是可配置成管理图形处理 器2500的不同功能(包括线程分派、调度和抢占)的可编程子处理器。在 至少一个实施例中,媒体管线2539包括用于促进多媒体数据(包括图像和 视频数据)的解码、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2539通过请求在子核2501-2501F内计算或采样逻辑来实现 媒体操作。

[0293] 在至少一个实施例中，SoC接口2537使得图形核2500能够与通用 应用处理器核(例如，CPU)和/或SoC内的其他部件(包括存储器分层结 构元件，如共享的末级高速缓存存储器、系统RAM、和/或嵌入式片上或 封装上DRAM)进行通信。在至少一个实施例中，SoC接口2537还可以 实现与SoC内的固定功能设备(诸如相机成像管线)的通信，并且实现可 以使用和/或实现可以在图形核2500与SoC内的CPU之间共享的全局存储 器原子学。在至少一个实施例中，SoC接口2537还可以实现针对图形核 2500的功率管理控制，并启用图形核2500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2537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，所述命令流转化器和全局线程分 派器被配置成用于向图形处理器内的一个或更多个图形核中的每个图形核 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 令和指令分派到媒体管线2539，或者当要执行图形处理操作时，可以将命 令和指令分派到几何结构和固定功能管线(例如，几何结构和固定功能管 线2536、几何结构和固定功能管线2514)。

[0294] 在至少一个实施例中，图形微控制器2538可以被配置为执行图形核 2500的不同调度和管理任务。在至少一个实施例中，图形微控制器2538 可以在子核2501A-2501F内的执行单元(EU)阵列2502A-2502F、2504A-2504F内的不同图形并行引擎上执行图形和/或计算工作负荷调度。 在至少一个实施例中，在包括图形核2500的SoC的CPU核上执行的主机软件可以向多个图形处理器门铃中的一个提交工作负荷，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 运行哪个工作负荷；向命令流转化器提交工作负荷；先占运行在引擎上的 现有工作负荷；监测工作负荷的进展；以及当工作负荷完成时通知主机软 件。在至少一个实施例中，图形微控制器2538还可以促进图形核2500的 低功率或空闲状态，从而为图形核2500提供跨低功率状态转换独立于操作 系统和/或系统上的图形驱动器软件来保存和恢复图形核2500内的寄存器 的能力。

[0295] 在至少一个实施例中，图形核2500可以具有多于或少于所示出的子 核2501A-2501F、多达N个模块化子核。在至少一个实施例中，对于N个 子核的每个集合，图形核2500还可以包括共享功能逻辑2510、共享和/或 高速缓存存储器2512、几何/固定功能管线2514以及用于加速不同图形和 计算处理操作的附加固定功能逻辑2516。在至少一个实施例中，共享功能 逻辑2510可以包括可以由图形核2500内的每N个子核共享的逻辑单元(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，固定的、共享的和/或高速缓存存储器2512可以是图形核2500内的N个子核 2501A-2501F的末级高速缓存，并且还可以用作可由多个子核访问的共享 存储器。在至少一个实施例中，几何/固定功能管线2514可代替几何结构/ 固定功能管线2536被包括在固定功能块2530内并且可包括相同或相似的 逻辑单元。

[0296] 在至少一个实施例中，图形核2500包括附加固定功能逻辑2516，附 加固定功能逻辑2516可以包括用于由图形核2500使用的不同固定功能加 速逻辑。在至少一个实施例中，附加的固定功能逻辑2516包括用于仅在位 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，而在几何结构/固定功能管线2516、2536内的完整几何管线和剔除管 线中，剔除管线为可包含在额外固定功能逻辑2516内的额外几何管线。在 至少一个实施例中，剔除管线是完整几何形状管线的经修剪版本。在至少 一个实施例中，完整管线和剔除管线可执行应用程序的不同示例，每一示 例具有单独上下文。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,使得在一些示例中着色能够更早地完成。例如,在 至少一个实施例中,额外固定功能逻辑2516内的剔除管线逻辑可与主应用 程序并行地执行位置着色器,且通常比完整管线更快地产生关键结果,因 为剔除管线提取顶点的位置属性且使顶点的位置属性着色,而不执行光栅 化和将像素再现到帧缓冲器。在至少一个实施例中,剔除管线可使用所产 生的关键结果来计算所有三角形的可见性信息,而不考虑是否剔除那些三 角形。在至少一个实施例中,全管线(其在该示例中可被称为重放管线) 可消耗可见性信息以跳过被剔除的三角形,从而仅对最终被传递到光栅化阶段的可见三角形进行着色。

[0297] 在至少一个实施例中,对于包括用于机器学习训练或推理的优化的 实现,附加的固定功能逻辑2516还可包括机器学习加速逻辑,诸如固定功 能矩阵乘法逻辑。

[0298] 在至少一个实施例中,在每个图形子核2501A-2501F内包括可用于 响应于图形管线、媒体管线或着色器程序的请求而执行图形、媒体和计算 操作的一组执行资源。在至少一个实施例中,图形子核2501A-2501F包括 多个EU阵列2502A-2502F、2504A-2504F、线程调度和线程间通信(TD/IC) 逻辑2503A-2503F、3D(例如,纹理)采样器2505A-2505F、媒体采样器 2506A-2506F、着色器处理器2507A-2507F和共享本地存储器(SLM) 2508A-2508F。EU阵列2502A-2502F、2504A-2504F各自包含多个执行单 元,所述多个执行单元是能够在图形、媒体或计算操作(包含图形、媒体 或计算着色器程序)的服务中执行浮点和整数/定点逻辑操作的通用图形处 理单元。在至少一个实施例中,TD/IC逻辑2503A-2503F为子核内的执行单元执行本地线程调度和线程控制操作,并且促进在子核的执行单元上执 行的线程之间的通信。在至少一个实施例中,3D采样器2505A-2505F可 以将纹理或其他3D图形相关数据读取到存储器中。在至少一个实施例中, 3D采样器可以基于配置的样本状态和与给定纹理相关联的纹理格式来不 同地读取纹理数据。在至少一个实施例中,媒体采样器2506A-2506F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2501A-2501F可以交替地包括统一的3D和媒体 采样器。在至少一个实施例中,在子核2501A-2501F中的每个子核内的执 行单元上执行的线程可以利用每个子核内的共享本地存储器2508A-2508F, 以使得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。

[0299] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑615的细节。在至少一个实施例中,推理和/或训练逻辑615的部分或全部 可以结合到图形处理器2510中。例如,在至少一个实施例中,在此描述的 训练和/或推理技术可以使用在图形处理器2312、图形微控制器2538、几 何结构和固定功能管线2514和2536、或图24中的其他逻辑中实施的ALU 中的一个或更多个。此外,在至少一个实施例中,在此描述的推理和/或训 练操作可以使用除了图6A或6B中所示出的逻辑之外的逻辑来完成。在至 少一个实施例中,可以将权重参数存储在片上或片外存储器和/或寄存器 (示出或未示出)中,这些存储器和/或寄存器将图形处理器2500的ALU 配置成用于执行在此描述的一个或更多个机器学习算法、神经网络架构、用例或训练技术。

[0300] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中,此逻辑可与这些图的组件一起使 用,以确定数据异常和适

当的响应动作。

[0301] 图26A-26B示出了根据至少一个实施例的包括图形处理器核的处理 元件的阵列的线程执行逻辑2600。图26A示出了至少一个实施例，其中使 用线程执行逻辑2600。图26B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。

[0302] 如图26A所示，在至少一个实施例中，线程执行逻辑2600包括着色 器处理器2602、线程分派器2604、指令高速缓存2606、包括多个执行单 元2608A-2608N的可扩展执行单元阵列、一个或更多个采样器2610、数据 高速缓存2612以及数据端口2614。在至少一个实施例中，可扩展执行单 元阵列可以通过基于例如工作负荷的计算要求启用或禁用一个或更多个执 行单元(例如，执行单元2608A、2608B、2608C、2608D至2608N-1和 2608N中的任何一个)来动态地扩展。在至少一个实施例中，可扩展执行 单元经由链接到执行单元中的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2600包括通过指令高速缓存2606、数据端口2614、采样器2610和执行单元2608A-2608N中的一个或更多个到存储器(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，2608A)是能够执行多个同时的硬件线程同时针对 每个线程并行地处理多个数据元素的独立可编程通用计算单元。在至少一 个实施例中，执行单元2608A-2608N的阵列是可扩展的以包括任何数量的 单独执行单元。

[0303] 在至少一个实施例中，执行单元2608A-2608N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2602可以处理不同着色器程序 并且经由线程分派器2604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2604包括用于仲裁来自图形和媒体管线的线程发 起请求并且在执行单元2608A-2608N中的一个或更多个执行单元上示例化 所请求的线程的逻辑。例如，在至少一个实施例中，几何管线可以将顶点、曲面细分或几何着色器分派到线程执行逻辑以供处理。在至少一个实施例 中，线程分派器2604还可以处理来自执行着色器程序的运行时线程生成请 求。

[0304] 在至少一个实施例中，执行单元2608A-2608N支持指令集，所述指 令集包括对许多标准3D图形着色器指令的本地支持，使得以最小的转换 来执行来自图形库(例如，Direct3D和OpenGL)的着色器程序。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，包括一个或更多个 算术逻辑单元(ALU)的执行单元2608A-2608N中的每一个能够进行多发 布单指令多数据(SIMD)执行，并且多线程操作使能高效的执行环境，尽 管存在较高等待时间的存储器访问。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高带宽寄存器文件和相关联的独立线程状态。 在至少一个实施例中，对能够进行整数、单精度和双精度浮点运算、SIMD 分支能力、逻辑运算、先验运算和其他杂项运算的管线而言，执行是每时 钟多发的。在至少一个实施例中，在等待来自存储器或共享函数之一的数 据时，执行单元2608A-2608N内的依赖性逻辑使等待线程休眠，直到所请 求的数据已经被返回。在至少一个实施例中，当等待的线程休眠时，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，执行单元可执行用于像素着色器、片段着色 器或另一类型的着色器程序(包含不同顶点着色器)的操作。

[0305] 在至少一个实施例中，执行单元2608A-2608N中的每个执行单元对 数据元素阵列

进行操作。在至少一个实施例中，数据元素的数量是“执行 大小”或用于指令的通道的数量。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，多个信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2608A-2608N 支持整数和浮点数据类型。

[0306] 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，可以将不同数据元素存储为寄存器中的紧缩数据类型，并且 执行单元将基于元素的数据大小来处理不同元素。例如，在至少一个实施 例中，当对256位宽的向量进行操作时，矢量的256位被存储在寄存器中， 并且执行单元作为四个单独的64位紧缩数据元素(四字(QW)大小的数 据元素)对矢量进行操作，八个单独的32位紧缩数据元素(双字(DW) 大小的数据元素)、十六个单独的16位紧缩数据元素(字(W)大小的数 据元素)，或32个单独的8位数据元素(字节(B)大小数据元素)。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。

[0307] 在至少一个实施例中，一个或更多个执行单元可被组合成具有对融 合EU共同的线程控制逻辑(2607A-2607N)的融合执行单元2609A-2609N。 在至少一个实施例中，多个EU可以融合到一个EU组中。在至少一个实 施例中，融合EU组中的每个EU可以被配置成执行单独的SIMD硬件线 程。融合EU组中的EU数目可以根据不同实施例而变化。在至少一个实 施例中，可以每EU执行不同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2609A-2609N 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2609A 包括第一EU2608A、第二EU2608B和对于第一EU2608A和第二EU2608B 共用的线程控制逻辑2607A。在至少一个实施例中，线程控制逻辑2607A 控制在融合图形执行单元2609A上执行的线程，从而允许融合执行单元 2609A-2609N内的每个EU使用共同指令指针寄存器来执行。

[0308] 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，2606) 包括在线程执行逻辑2600中以高速缓存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，2612)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2610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2610包括用于在将经采样的数据提供给执行单元之前的采样过程 期间处理纹理或媒体数据的专用纹理或媒体采样功能。

[0309] 在至少一个实施例中，在执行期间，图形和媒体管线经由线程生成 和调度逻辑向线程执行逻辑2600发送线程发起请求。在至少一个实施例中， 一旦一组几何对象已经被处理和光栅化成像素数据，调用着色器处理器 2602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) 以进一步计算输出信息，且致使将结果写入到输出表面(例如，色彩缓冲 器、深度缓冲器、模板缓冲器等)。在至少一个实施例中，像素着色器或片 段着色器计算将跨光栅化对象内插的不同顶点属性的值。在至少一个实施 例中，着色器处理器2602内的像素处理器逻辑接着执行应用程序编程接口 (API)供应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2602经由线程分派器2604将线程分派到执行 单元(例如，2608A)。在至少一个实施例中，着色器处理器2602使用采样器2610中的纹理采样逻辑来访问存储在存储器中的纹理图中的纹理数 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计算每 个几何片段的像素颜色数据,或从进一步处理中丢弃一个或更多个像素。

[0310] 在至少一个实施例中,数据端口2614为线程执行逻辑2600提供存 储器访问机制以将经处理数据输出到存储器以供在图形处理器输出管线上 进一步处理。在至少一个实施例中,数据端口2614包括或耦合至一个或更 多个高速缓存存储器(例如,数据高速缓存2612)以便经由数据端口高速 缓存用于存储器访问的数据。

[0311] 如图26B所示,在至少一个实施例中,图形执行单元2608可以包括 指令获取单元2637,通用寄存器文件阵列(GRF)2624、架构寄存器文件 阵列(ARF)2626、线程仲裁器2622,发送单元2630、分支单元2632、一组SIMD浮点单元(FPU)2634,以及,在至少一个实施例中,专用整 数SIMD ALU 2635的集合。在至少一个实施例中,GRF2624和ARF2626 包括与在图形执行单元2608中可以是活动的每个同时的硬件线程相关联 的通用寄存器文件和架构寄存器文件的集合。在至少一个实施例中,在 ARF2626中维护每线程架构状态,而在线程执行期间使用的数据存储在 GRF2624中。在至少一个实施例中,包括每个线程的指令指针的每个线程 的执行状态可以保存在ARF2626中的线程特定寄存器中。

[0312] 在至少一个实施例中,图形执行单元2608具有同时多线程化(SMT) 和细粒度交错多线程化(IMT)的组合的架构。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目标数量的同时线程 和每个执行单元的寄存器数量来微调,其中执行单元资源跨用于执行多个 同时线程的逻辑来划分。

[0313] 在至少一个实施例中,图形执行单元2608可以共同发布多条指令, 每条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 2608的线程仲裁器2622可以将指令分派到发送单元2630、分支单元2642 或SIMD FPU 2634中的一个用于执行。在至少一个实施例中,每个执行线 程可以访问GRF2624内的128个通用寄存器,其中每个寄存器可以存储 32字节,作为32位数据元素的SIMD8元素向量是可访问的。在至少一个 实施例中,每个执行单元线程访问GRF2624内的4K字节,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,多达七个线程可以同时执行,但是每个执行单元的线程数 量也可以根据各实施例而变化。在至少一个实施例中,其中七个线程可以 访问4K字节,GRF2624可以存储总共28K字节。在至少一个实施例中, 灵活寻址模式可允许寄存器被一起寻址以有效地构建更宽的寄存器或表示 条纹矩形块数据结构。

[0314] 在至少一个实施例中,经由消息传递发送单元2630执行的“发送” 指令来分派存储器操作、采样器操作和其他较长等待时间系统通信。在至 少一个实施例中,将分支指令分派到专用分支单元2632以促进SIMD发散 和最终收敛。

[0315] 在至少一个实施例中,图形执行单元2608包括用于执行浮点操作的 一个或更多个SIMD浮点单元(FPU)2634。在至少一个实施例中,FPU 2634 还支持整数计算。在至少一个实施例中,一个或更多个FPU2634可以SIMD 执行多达M个32位浮点(或整数)操作,或者SIMD执行多达2M个16 位整数或16位浮点操作。在至少一个实施例中,一个或更多个FPU中的 至少一个FPU提供扩展的数学能力以支持高吞吐量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 2635, 并且可以对其进行特定优化以执行与机器学习计算相关联的操作。

[0316] 在至少一个实施例中,图形执行单元2608的多个示例的阵列可以在 图形子核分

组(例如,子切片)中示例化。在至少一个实施例中,执行单 元2608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2608上执行的每个线程在不同的信道上执行。

[0317] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中,推理和/或训练逻辑615的部分或全部 可以结合到执行逻辑2600中。此外,在至少一个实施例中,在此描述的推 理和/或训练操作可以使用除了图6A或6B中所示出的逻辑之外的逻辑来 完成。在至少一个实施例中,权重参数可以存储在片上或片外存储器和/ 或寄存器(示出或未示出)中,这些存储器和/或寄存器将执行逻辑2600 的ALU配置成用于执行在此描述的一个或更多个机器学习算法、神经网络 架构、用例、或训练技术。

[0318] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中,此逻辑可与这些图的组件一起使 用,以确定数据异常和适当的响应动作。

[0319] 图27示出了根据至少一个实施例的并行处理单元(“PPU”)2700。在至少一个实施例中,PPU2700配置有机器可读代码,该机器可读代码如 果由PPU2700执行,则使PPU2700执行贯穿本公开所描述的过程和技术 中的一些或全部。在至少一个实施例中,PPU2700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成并行地处 理多个线程上的计算机可读指令(也被称为机器可读指令或简称为指令) 的等待时间隐藏技术。在至少一个实施例中,线程是指执行的线程并且是 被配置为由PPU2700执行的指令集的示例。在至少一个实施例中,PPU 2700是图形处理单元(“GPU”),该图形处理单元被配置成用于实现用于 处理三维(“3D”)图形数据以便生成二维(“2D”)图像数据以便在显示装 置(如液晶显示器(“LCD”)装置)上显示的图形渲染管线。在至少一个 实施例中,PPU2700用于执行诸如线性代数运算和机器学习运算的计算。图27示出了仅用于说明性目的的示例并行处理器,并且应当被解释为在本 公开的范围内考虑的处理器架构的非限制性示例,并且可以采用任何合适 的处理器来补充和/或替代它们。

[0320] 在至少一个实施例中,一个或更多个PPU2700被配置成用于加速高 性能计算(“HPC”)、数据中心、和机器学习应用。在至少一个实施例中, PPU2700被配置为加速深度学习系统和应用,包括以下非限制性示例:自 主车辆平台、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、机器人学、工厂自动化、实时语言翻译、在 线搜索优化和个性化用户推荐等。

[0321] 在至少一个实施例中,PPU 2700包括但不限于:输入/输出(“I/O”) 单元2706、前端单元2710、调度器单元2712、工作分配单元2714、集线 器2716、交叉开关(“Xbar”)2720、一个或更多个通用处理集群(“GPC”) 2718、以及一个或更多个分区单元(“内存分区单元”)2722。在至少一个 实施例中,PPU2700经由一个或更多个高速GPU互连(“GPU互连”)2708 连接到主机处理器或其他PPU2700。在至少一个实施例中,PPU2700经由 互连2702连接到主处理器或其他外围设备。在至少一个实施例中,PPU 2700连接至包括一个或更多个存储器装置(“存储器”)2704的本地存储器。 在至少一个实施例中,存储器装置2704包括但不限于一个或更多个动态随 机存取存储器(“DRAM”)装置。在至少一个实施例中,一个或更多个

DRAM器件被配置和/或可配置为高带宽存储器(“HBM”)子系统,其中 每个器件内堆叠有多个DRAM管芯。

[0322] 在至少一个实施例中,高速GPU互连2708可以指一个基于导线的 多通道通信链路,该基于导线的多通道通信链路被系统用来缩放并且包括 与一个或更多个中央处理单元(“CPU”)组合的一个或更多个PPU2700、支持PPU2700与CPU之间的高速缓存一致性以及CPU掌控。在至少一个 实施例中,数据和/或命令由高速GPU互连2708通过集线器2716传输到 PPU2700的其他单元(如一个或更多个复制引擎、视频编码器、视频解码 器、功率管理单元和图27中未明确示出的其他组件)或从PPU2700的其 他单元(如一个或更多个复制引擎、视频编码器、视频解码器、功率管理 单元和图27中未明确示出的其他组件)传输。

[0323] 在至少一个实施例中,I/O单元2706被配置为通过系统总线2702发 送和接收来自主机处理器(图27中未示出)的通信(例如,命令、数据)。 在至少一个实施例中,I/O单元2706直接经由系统总线2702或通过一个 或更多个中间设备(诸如存储器桥)与主处理器通信。在至少一个实施例 中,I/O单元2706可以通过系统总线2702与一个或更多个其他处理器(如 一个或更多个PPU2700)进行通信。在至少一个实施例中,I/O单元2706 实现用于PCIe总线上的通信的外围组件互连高速(“PCIe”)接口。在至少 一个实施例中,I/O单元2706实现用于与外部设备通信的接口。

[0324] 在至少一个实施例中,I/O单元2706对经由系统总线2702接收的数 据包进行解码。在至少一个实施例中,至少一些数据包表示被配置为使 PPU2700执行不同操作的命令。在至少一个实施例中,I/O单元2706将解 码的命令传输到PPU2700的不同其他单元,如由命令指定的。在至少一个 实施例中,命令被传送到前端单元2710和/或被传送到集线器2716或 PPU2700的其他单元,诸如一个或更多个复制引擎、视频编码器、视频解 码器、功率管理单元等。(图27中未明确示出)。在至少一个实施例中,I/O 单元2706被配置为在PPU2700的不同逻辑单元之间或之中路由通信。

[0325] 在至少一个实施例中,由主机处理器执行的程序将命令流编码在向 PPU2700提供工作负载用于处理的缓冲器中。在至少一个实施例中,工作 负荷包括指令和由这些指令处理的数据。在至少一个实施例中,缓冲器是 存储器中可由主机处理器和PPU2700访问(例如,读/写)的区域--主机接 口单元可被配置为经由I/O单元2706通过系统总线2702发送的存储器请 求来访问连接到系统总线2702的系统存储器中的缓冲器。在至少一个实施 例中,主机处理器将命令流写入到缓冲器,且接着将命令流开始的指针发 射到PPU2700,使得前端单元2710接收指向一个或更多个命令流的指针 且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 PPU2700的各个单元。

[0326] 在至少一个实施例中,前端单元2710耦合到调度器单元2712,调度 器单元2712配置不同GPC2718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2712被配置为追踪与调度器单元2712 管理的不同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 GPC 2718中的哪一个、任务是活动的还是不活动的、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2712管理在GPC2718中的一 个或更多个上的多个任务的执行。

[0327] 在至少一个实施例中,调度器单元2712耦合到工作分配单元2714, 工作分配单元2714被配置为分派用于在GPC2718上执行的任务。在至少 一个实施例中,工作分配单元

2714追踪从调度器单元2712接收的多个被 调度的任务,并且工作分配单元2714管理每个GPC2718的待决任务池和 活动任务池。在至少一个实施例中,待决任务池包括多个时隙(例如,32 个时隙),这些时隙包含被指派由特定GPC2718处理的任务;活动任务池 可包括用于由GPC2718活动地处理的任务的多个槽(例如,4个槽),使 得当GPC2718中的一个完成任务的执行时,该任务从用于GPC 2718的活 动任务池驱逐,并且来自待决任务池的其他任务之一被选择和调度以在 GPC 2718上执行。在至少一个实施例中,如果活跃任务在GPC 2718上是 空闲的,诸如在等待数据相关性被解决时,则活跃任务从GPC 2718被逐 出并返回到待决任务池,同时待决任务池中的另一任务被选择并调度用于 在GPC 2718上执行。

[0328] 在至少一个实施例中,工作分配单元2714经由XBar 2720与一个或 更多个GPC 2718通信。在至少一个实施例中,XBar2720是互连网络,其 将PPU2700的许多单元耦合到PPU2700的其他单元,并且可以被配置为 将工作分配单元2714耦合到特定的GPC2718。在至少一个实施例中, PPU2700的一个或更多个其他单元也可以经由集线器2716连接到XBar2720。

[0329] 在至少一个实施例中,任务由调度器单元2712管理并且由工作分配 单元2714分派给GPC 2718中的一个。GPC 2718被配置为处理任务并且 产生结果。在至少一个实施例中,结果可由GPC 2718内的其他任务消耗, 经由XBar 2720路由至不同的GPC 2718,或者存储在存储器2704中。在 至少一个实施例中,可以经由分区单元2722将结果写入到存储器2704, 分区单元2722实现用于向存储器2704读取数据和向存储器2704写入数据 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2708 传输到另一个PPU2704或CPU。在至少一个实施例中,PPU2700包括但 不限于数量U的分区单元2722,其等于耦接至PPU2700的单独且不同的 存储装置2704的数量。在至少一个实施例中,下面将结合图29更详细地 描述分区单元2722。

[0330] 在至少一个实施例中,主机处理器执行实现应用编程接口(“API”) 的驱动器内核,该应用编程接口使得在主机处理器上执行的一个或更多个 应用能够调度用于在PPU2700上执行的操作。在至少一个实施例中,PPU 2700同时执行多个计算应用,并且PPU2700为多个计算应用提供隔离、服务质量(“QoS”)、和独立的地址空间。在至少一个实施例中,应用生成 使驱动器内核生成用于由PPU2700执行的一个或更多个任务和驱动器内 核向由PPU2700处理的一个或更多个流输出任务的指令(例如,以API 调用的形式)。在至少一个实施例中,每个任务包括一组或多组相关纱线, 其可被称为经线。在至少一个实施例中,经线包括可并行执行的多个相关 线(例如,32个线)。在至少一个实施例中,协作线程可以指代包括用于 执行任务并且通过共享存储器交换数据的指令的多个线程。在至少一个实施例中,结合图29,根据至少一个实施例更详细地描述螺纹和协作螺纹。

[0331] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中,使用深度学习应用处理器来训练机器 学习模型(诸如神经网络)以预测或推理提供给PPU2700的信息。在至少 一个实施例中,PPU2700用于基于已经由另一个处理器或系统或由 PPU2700训练的经训练的机器学习模型(例如,神经网络)来推理或预测 信息。在至少一个实施例中,PPU2700可用于执行本文所述的一个或更多 个神经网络使用情况。

[0332] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训

练操作。在至少一个实施例中,此逻辑可与这些图的组件一起使 用,以确定数据异常和适当的响应动作。

[0333] 图28示出了根据至少一个实施例的通用处理集群(“GPC”)2800。在至少一个实施例中,GPC 2800是图27的GPC 2718。在至少一个实施例 中,每个GPC 2800包括但不限于用于处理任务的多个硬件单元,并且每 个GPC 2800包括,但不限于,管线管理器2802、预光栅操作单元(“PROP”) 2804、光栅引擎2808,工作分配交叉开关(“WDX”)2816、存储器管理 单元(“MMU”)2818、一个或更多个数据处理集群(“DPC”)2806,以及 部件的任何合适组合。

[0334] 在至少一个实施例中,GPC 2800的操作由管线管理器2802控制。在 至少一个实施例中,管线管理器2802管理一个或更多个DPC 2806的配置, 用于处理分配给GPC 2800的任务。在至少一个实施例中,管线管理器2802 配置一个或更多个DPC2806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 2806被配置成用于在可编程流式多处理 器(“SM”)2814上执行顶点着色器程序。在至少一个实施例中,管线管 理器2802被配置为将从工作分配单元接收的数据包路由至GPC 2800内的 适当的逻辑单元,在至少一个实施例中,一些分组可被路由到PROP2804 和/或光栅引擎2808中的固定功能硬件单元,而其他分组可被路由到 DPC2806以供图元引擎2812或SM2814处理。在至少一个实施例中,管 线管理器2802配置DPC2806中的至少一个DPC以实现神经网络模型和/ 或计算管线。

[0335] 在至少一个实施例中,PROP单元2804在至少一个实施例中经配置 以将由光栅引擎2808和DPC2806产生的数据路由到分区单元2722中的光 栅操作(“ROP”)单元,上文结合图27更详细地描述。在至少一个实施例 中,PROP单元2804被配置为执行用于颜色混合、组织像素数据、执行地 址转换等的优化。在至少一个实施例中,光栅引擎2808包括但不限于被配 置为执行不同光栅操作的多个固定功能硬件单元,在至少一个实施例中, 光栅引擎2808包括但不限于设置引擎,粗栅格引擎、剔除引擎、裁剪引擎、 细栅格引擎、分块合并引擎,及其任何合适的组合。在至少一个实施例中, 设置引擎接收经变换的顶点并且生成与由顶点定义的几何图元相关联的平 面方程;将平面方程传输到粗栅格引擎以生成图元的覆盖信息(例如,图 块的x,y覆盖掩码);粗栅格引擎的输出被传输到剔除引擎,在剔除引擎中,剔除与未通过z测试的图元相关联的片段,并且被传输到裁剪引擎, 在裁剪引擎中,裁剪位于视见平截头体外的片段。在至少一个实施例中, 将存在于裁剪和剔除之后的片段传递到精细光栅引擎,以基于由设置引擎 产生的平面方程来产生像素片段的属性。在至少一个实施例中,光栅引擎 2808的输出包括将由任何合适的实体(诸如由DPC2806内实现的片段着 色器)处理的片段。

[0336] 在至少一个实施例中,包括在GPC 2800中的每个DPC 2806包括但 不限于一个M管道控制器(“MPC”)2810;图元引擎2812;一个或更多 个SM2814;及其任何合适的组合。在至少一个实施例中,MPC 2810控制 DPC 2806的操作,将从管线管理器2802接收的分组路由至DPC 2806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2812,图元引擎2812被配置为从存储器中提取与顶点相关联的顶点属 性;相反,可将与着色器程序相关联的包发射到SM2814。

[0337] 在至少一个实施例中,SM2814包括但不限于被配置为处理由多个线 程表示的任务的可编程流处理器。在至少一个实施例中,SM2814是多线 程的并且被配置成用于并发地执行来自特定线程组的多个线程(例如,32 个线程)并实现单指令,多数据(“SIMD”)架构,

其中一组线程(例如，经线)中的每个线程被配置成用于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的指令。在至少一个 实施例中,SM2814实现单指令多线程(“SIMT”)架构,其中一组线程中 的每个线程被配置成用于基于相同的指令集来处理不同的数据集合,但是 其中允许一组线程中的单独线程在执行期间分叉。在至少一个实施例中， 为每一经线维持程序计数器、调用栈和执行状态,从而在经线内的线程发 散时实现经线与经线内的串行执行之间的并发性。在另一实施例中,为每个单独的线程维护程序计数器、调用栈和执行状态,从而实现所有线程之 间、经线内和经线之间的相等并发性。在至少一个实施例中,为每个单独 的线程保持执行状态,并且为了更好的效率,可以会聚并且并行执行执行 相同指令的线程。SM2814的至少一个实施例在下文更详细地描述。

[0338] 在至少一个实施例中,MMU 2818在GPC 2800和存储器分区单元(例 如,图27的分区单元2722)之间提供接口,并且MMU 2818提供虚拟地 址到物理地址的转换、存储器保护和存储器请求的仲裁。在至少一个实施 例中,MMU 2818提供用于执行虚拟地址到存储器中的物理地址的转换的 一个或更多个转换后备缓冲器(“TLB”)。

[0339] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中,使用深度学习应用处理器来训练机器 学习模型(诸如神经网络)以预测或推理提供给GPC2800的信息。在至少 一个实施例中,GPC2800用于基于已经由另一个处理器或系统或由 GPC2800训练的经训练的机器学习模型(例如,神经网络)来推理或预测 信息。在至少一个实施例中,GPC2800可以用来执行在此描述的一个或更 多个神经网络使用情况。

[0340] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中,此逻辑可与这些图的组件一起使 用,以确定数据异常和适当的响应动作。

[0341] 图29示出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2900。在至少一个实施例中,存储器分区单元2900包括但不限 于光栅操作(“ROP”)单元2902;二级(“L2”)高速缓存2904;存储器接 口2906;以及其任何合适的组合。在至少一个实施例中,存储器接口2906 耦合到存储器。在至少一个实施例中,存储器接口2906可以实现32、64、128、1024位数据总线或类似实现方式,用于高速数据传输。在至少一个 实施例中,PPU结合了U个存储器接口2906,每对分区单元2900一个存 储器接口2906,其中每对分区单元2900连接至相应的存储器设备。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,如高带宽存 储器堆栈或图形双倍数据速率、版本5、同步动态无随机存储器(“GDDR5 SDRAM”)。

[0342] 在至少一个实施例中,存储器接口2906实现高带宽存储器第二代 (“HBM2”)存储器接口,并且Y等于半U。在至少一个实施例中,HBM2 存储器栈与PPU位于相同的物理封装上,与传统的GDDR5 SDRAM系统 相比提供了显著的功率和面积节省。在至少一个实施例中,每个HBM2堆 叠包括但不限于四个存储器管芯,并且Y等于4,其中每个HBM2堆叠包 括每管芯两个128位沟道,总共8个沟道和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。在至少一个实施例中,ECC为对数据损坏 敏感的计算应用提供了更高的可靠性。

[0343] 在至少一个实施例中,PPU实现多级存储器分层结构。在至少一个 实施例中,存储器分区单元2900支持统一存储器来为中央处理单元 (“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的频率以确保存储器页面被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2708支持允 许PPU直接访问CPU的页表并且通过PPU提供对CPU存储器的完全访问 的地址转换服务。

[0344] 在至少一个实施例中,复制引擎在多个PPU之间或者在PPU与CPU 之间传输数据。在至少一个实施例中,复制引擎可产生未映射到页表中的 地址的页错误,且存储器分区单元2900接着服务页错误,将地址映射到页 表中,之后复制引擎执行传送。在至少一个实施例中,对于多个处理器之 间的多个复制引擎操作,存储器被钉住(即,不可分页),从而显著减少可 用存储器。在至少一个实施例中,在硬件页面故障的情况下,可将地址传 递到复制引擎,而不考虑存储器页面是否驻留,并且复制过程是透明的。

[0345] 根据至少一个实施例,来自图27的存储器2704或其他系统存储器 的数据由存储器分区单元2900提取并且存储在L2高速缓存2904中,所 述L2高速缓存位于芯片上并且在不同GPC之间共享。在至少一个实施例 中,每个存储器分区单元2900包括但不限于与对应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,较低级高速缓存在GPC 内的不同单元中实现。在至少一个实施例中,每个SM2814可以实现一级 (“L1”)高速缓存,其中L1高速缓存是专用于特定SM2814的私有存储 器,并且取来自L2高速缓存2904的数据并将其存储在每个L1高速缓存 中以便在SM2814的功能单元中进行处理。在至少一个实施例中,L2高速 缓存2904耦合到存储器接口2906和XBar2720。

[0346] 在至少一个实施例中,ROP单元2902执行与像素色彩相关的图形光 栅操作,例如色彩压缩、像素混合和更多。在至少一个实施例中,ROP单 元2902结合光栅引擎2808实施深度测试,从光栅引擎2808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,针对与片 段相关联的样本位置,针对深度缓冲器中的相应深度来测试深度。在至少 一个实施例中,如果片段通过样本位置的深度测试,那么ROP单元2902 更新深度缓冲器且将深度测试的结果发射到光栅引擎2808。将了解,分区 单元2900的数目可不同于GPC的数目,且因此,在至少一个实施例中, 每一ROP单元2902可耦合到GPC中的每一者。在至少一个实施例中,ROP 单元2902追踪从不同GPC接收的包且确定ROP单元2902产生的结果经 由XBar2720路由到哪一者。

[0347] 图30示出了根据至少一个实施例的流式多处理器(“SM”)3000。在 至少一个实施例中,SM 3000是图28的SM 2814。在至少一个实施例中, SM3000包括但不限于指令高速缓存3002;一个或更多个调度器单元3004; 寄存器文件3008;一个或更多个处理单元(“核”)3010;一个或更多个特 殊功能单元(“SFU”)3012;一个或更多个加载/存储单元(“LSU”)3014; 互连网络3016;共享存储器/一级(“L1”)高速缓存3018;及其任何合适 的组合。在至少一个实施例中,工作分配单元分派任务用于在并行处理单 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,则任务被分配给SM3000之一。在至少一个实施例中,调度器单元3004 从工作分配单元接收任务并管理指派给SM3000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3004调度线程块以供作为并 行线程的经线来执行,其中

每个线程块被分配至少一个经线。在至少一个 实施例中，每个经线执行线。在至少一个实施例中，调度器单元3004管理 多个不同的线程块，向不同的线程块分配经线，然后在每个时钟周期期间 将来自多个不同的协作组的指令分派到各个功能单元(例如，处理单元3010、SFU3012和LSU3014)。

[0348] 在至少一个实施例中，协作组可以指用于组织通信的线程的组的编 程模型，其允许开发者表达线程正在通信的粒度，使得能够表达更丰富、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 的同步以用于并行算法的执行。在至少一个实施例中，常规编程模型的应 用提供用于同步协作线程的单个简单构造:跨线程块的所有线程的屏障(例 如，syncthreads()函数)。然而，在至少一个实施例中，程序员可以以小 于线程块粒度来定义线程的组，并在所定义的组内同步，以实现更大的性 能、设计灵活性以及以集体全组功能接口的形式的软件重用。在至少一个 实施例中，协作组使程序员能够在子块(即，小到单个线程)处显式地定 义线程组和多块粒度，并且对协作组中的线程执行诸如同步的集体操作。 在至少一个实施例中，编程模型支持跨软件边界的清晰合成，使得库和实 用函数可以在它们的本地上下文内安全地同步，而不必做出关于收敛的假 设。在至少一个实施例中，协作组原语实现协作并行性的新模式，包括但 不限于生产者-消费者并行性、机会并行性和跨线程块的整个网格的全局同 步。

[0349] 在至少一个实施例中，分派单元3006被配置为向一个或更多个功能 单元发送指令，并且调度器单元3004包括但不限于两个分派单元3006，其使得能够在每个时钟周期期间分派来自相同经线的两个不同指令。在至 少一个实施例中，每个调度器单元3004包括单个分派单元3006或附加分 派单元3006。

[0350] 在至少一个实施例中，在至少一个实施例中，每个SM3000包括但不 限于为SM3000的功能单元提供一组寄存器的寄存器文件3008。在至少一 个实施例中，寄存器文件3008在每个功能单元之间划分，使得每个功能单 元被分配寄存器文件3008的专用部分。在至少一个实施例中，寄存器文件 3008在由SM3000执行的不同经线之间划分，并且寄存器文件3008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SM3000包括但不限于多个L个处理单元3010。在至少一个实施例中， SM3000包括但不限于大量(例如，128个或更多)不同的处理单元3010。 在至少一个实施例中，在至少一个实施例中，每个处理单元3010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施用于浮点算术的IEEE754-2008标准。在至少一个实施例中，处 理单元3010包括但不限于64个单精度(32位)浮点核、64个整数核、32 个双精度(64位)浮点核和8个张量核。

[0351] 根据至少一个实施例，张量芯被配置成执行矩阵运算。在至少一个 实施例中，一个或更多个张量核被包括在处理单元3010中。在至少一个实 施例中，张量核被配置为执行深度学习矩阵运算，诸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，每个张量芯在4x4矩阵上操作， 并且执行矩阵乘法和累加操作D＝A X B+C，其中A、B、C和D是4x4矩 阵。

[0352] 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 量核在具有32位浮点累加的16位浮点输入数据上操作。在至少一个实施 例中，16位浮点乘法使用64个运算并且导致全精度乘

积,然后使用32位 浮点加法将其与其他中间乘积累加以用于4x4x4矩阵乘法。在至少一个实 施例中,张量核用于执行由这些较小元素构建的大得多的二维或更高维的 矩阵运算。在至少一个实施例中,诸如CUDA 9C++API的API暴露专门 的矩阵加载、矩阵相乘和累加以及矩阵存储操作以高效地使用来自 CUDA-C++程序的张量核。在至少一个实施例中,在CUDA级,经线级接 口假设跨越所有32个经线线程的16x16大小矩阵。

[0353] 在至少一个实施例中,每个SM3000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SFU3012。在至少一个实施例中, SFU3012包括但不限于树遍历单元,该树遍历单元被配置成用于遍历分层 树数据结构。在至少一个实施例中,SFU3012包括但不限于被配置为执行 纹理图滤波操作的纹理单元。在至少一个实施例中,纹理单元被配置成从 存储器加载纹理图(例如,纹理元素的2D阵列)并对纹理图进行采样以 产生经采样的纹理值以供SM3000执行的着色器程序中使用。在至少一个 实施例中,纹理图被存储在共享存储器/L1高速缓存3018中。在至少一个 实施例中,根据至少一个实施例,纹理单元实施纹理操作,例如使用mip 图(例如,不同细节等级的纹理图)的过滤操作。在至少一个实施例中, 每个SM3000包括但不限于两个纹理单元。

[0354] 在至少一个实施例中,每个SM3000包括但不限于实现共享存储器 /L1高速缓存3018和寄存器文件3008之间的加载和存储操作的N个 LSU3014。在至少一个实施例中,每个SM3000包括但不限于互连网络3016, 互连网络3016将功能单元中的每一个连接到寄存器文件3008,并且将 LSU3014连接到寄存器文件3008和共享存储器/L1高速缓存3018。在至少一个实施例中,互连网络3016是交叉开关,该交叉开关可被配置成将任何 功能单元连接到寄存器文件3008中的任何寄存器并将LSU3014连接到寄 存器文件3008和共享存储器/L1高速缓存3018中的存储器位置。

[0355] 在至少一个实施例中,共享存储器/L1高速缓存3018是片上存储器 的阵列,在至少一个实施例中,其允许SM3000与图元引擎之间以及 SM3000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3018包括但不限于128KB的存储容量,并且在从SM3000 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3018 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3018、L2高速缓存和存储器中的一个或更多个是后 备存储。

[0356] 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改进的性能。在至少一个实施 例中,例如如果共享存储器被配置为使用容量的一半,则容量被不使用共 享存储器的程序使用或可用作高速缓存,纹理和加载/存储操作可使用剩余 容量。根据至少一个实施例,共享存储器/L1高速缓存3018内的集成使得 共享存储器/L1高速缓存3018能够充当用于流式传输数据的高吞吐量管道,同时提供对频繁重用数据的高带宽和低等待时间访问。在至少一个实施例 中,当被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,从而创建简单 得多的编程模型。在至少一个实施例中,在通用并行计算配置中,工作分 配单元将线程的块直接分配和分配给DPC。在至少一个实施例中,块中的 线程执行相同的程序,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SM3000执行程序和执行计算,共享存储器/L1高速缓存 3018在线程之间通信,以及LSU3014,用于通过共享存储器/L1高速缓存3018和存储器分区单元读取和写入全局存储

器。在至少一个实施例中，当 被配置用于通用并行计算时，SM3000写入调度器单元3004可用于在DPC 上发起新工作的命令。

[0357] 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、平板计算机、服务器、超级计算机、智能电话(例如，无线、手持式设备)、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 备等中或耦合到其上。在至少一个实施例中，PPU被实施在单个半导体衬 底上。在至少一个实施例中，PPU连同一个或更多个其他设备(如附加PPU、存储器、精简指令集计算机(“RISC”)CPU、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。

[0358] 在至少一个实施例中，PPU可以包括在包括一个或更多个存储器设 备的图形卡上。图形卡可被配置成与台式计算机的主板上的PCIe插槽对接。在至少一个实施例中，PPU可以是包括在母板芯片组中的集成图形处理单 元(“iGPU”)。

[0359] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中，深度学习应用处理器被用于训练机器 学习模型(诸如神经网络)以预测或推理提供给SM3000的信息。在至少 一个实施例中，SM3000用于基于已由另一处理器或系统或由SM3000训 练的经训练的机器学习模型(例如，神经网络)来推理或预测信息。在至 少一个实施例中，SM3000可被用于执行本文描述的一个或更多个神经网 络使用情况。

[0360] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中，此逻辑可与这些图的组件一起使 用，以确定数据异常和适当的响应动作。

[0361] 在至少一个实施例中，单个半导体平台可以指唯一的基于单一半导 体的集成电路或芯片。在至少一个实施例中，多芯片模块可以与增强的连 接性一起使用，该增强的连接性模拟片上操作并且相对于利用常规中央处 理单元(“CPU”)和总线实现方式做出实质性改进。在至少一个实施例中， 不同模块也可以根据用户的愿望单独地或在半导体平台的不同组合中定位。

[0362] 在至少一个实施例中，采用机器可读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1004和/或辅助存储器中。如果 计算机程序由一个或更多个处理器执行，则使系统1000能够执行根据至少 一个实施例的不同功能。在至少一个实施例中，存储器1004、存储装置和 /或任何其他存储装置是计算机可读介质的可能示例。在至少一个实施例中， 辅助存储装置可以指任何合适的存储装置或系统，如硬盘驱动器和/或可移 动存储驱动器，代表软盘驱动器、磁带驱动器、致密盘驱动器、数字通用 盘(“DVD”)驱动器、记录装置、通用串行总线(“USB”)闪存等。在至 少一个实施例中，在CPU1002的上下文中实现不同先前附图的架构和/或 功能；并行处理系统1012；集成电路，其能够实现两个CPU1002的能力 的至少一部分；并行处理系统1012；芯片组(例如，被设计为作为执行相 关功能的单元工作和销售的一组集成电路等)；以及一个或更多个集成电路 的任何合适的组合。

[0363] 在至少一个实施例中，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、专用系统和更多系统的上下文中实现不同先前 附图的架构和/或功能。在至少一个实施例中，计算机系统1000可以采取 台式计算机的形式，膝上型计算机、平板计算

机、服务器、超级计算机、智能电话(例如,无线手持式装置),个人数字助理(“PDA”)、数码相机、车辆、头戴式显示器、手持式电子装置,移动电话设备、电视、工作站、游戏控制台、嵌入式系统和/或任何其他类型的逻辑。

[0364] 在至少一个实施例中,并行处理系统1012包括但不限于多个并行处 理单元(“PPU”)1014和相关联的存储器1016。在至少一个实施例中,PPU1014经由一个互连1018和一个开关1020或多路复用器连接到一个主 机处理器或其他外围设备上。在至少一个实施例中,并行处理系统1012 跨PPU 1014分配计算任务,这些计算任务可以是可并行化的——例如,作为跨多个图形处理单元(“GPU”)线程块分配计算任务的一部分。在至 少一个实施例中,存储器跨一些或所有PPU1014被共享和可访问(例如,用于读取和/或写入访问),尽管这样的共享存储器可能引起相对于驻留在PPU1014的本地存储器和寄存器的使用的性能损失。在至少一个实施例中,通过使用诸如__syncthreads()的命令来同步PPU1014的操作,其中(例 如,跨多个PPU1014执行的)块中的所有线程在继续之前到达代码的某个 执行点。

[0365] 虚拟化计算平台

[0366] 公开了涉及用于高级计算的虚拟化计算平台的实施例,高级计算诸 如医疗应用中的图像推理和图像处理。非限制性地,实施例可以包括射线 照相术、磁共振成像(MRI)、核医学、超声、超声检查、弹性成像、光声 成像、断层摄影术、超声心动图、功能性近红外光谱法、以及磁性颗粒成 像、或其组合。在至少一个实施例中,可以另外地或替代地使用在此描述 的虚拟化计算平台和相关联的过程,无限制地,在法医科学分析、表面下 检测和成像(例如,石油勘探、古细菌学、眼科学等)中,地形、海洋学、 地质学、骨科学、气象学、智能区域或对象追踪和监测、传感器数据处理 (例如,RADAR、SONAR、LIDAR等)和/或基因组学和基因测序。

[0367] 参照图31,其是根据至少一个实施例的生成和部署图像处理和推理 管线的过程3100的示例数据流程图。在至少一个实施例中,过程3100可 以被部署用于与成像设备、处理设备、基因组学设备、基因测序设备、放 射学设备和/或一个或更多个设施3102处的其他设备类型一起使用,诸如 医疗设施、医院、医疗机构、诊所、研究或诊断实验室等。在至少一个实 施例中,可以部署过程3100以对测序数据执行基因组学分析和推理。可以 使用在此描述的系统和方法进行的基因组分析的示例包括但不限于变体调 用、突变检测和基因表达定量。过程3100可以在训练系统3104和/或部署 系统3106内执行。在至少一个实施例中,训练系统3104可用于执行机器 学习模型(例如,神经网络、对象检测算法、计算机视觉算法等)的训练、 部署和实现,以供在部署系统3106中使用。在至少一个实施例中,部署系 统3106可以被配置成用于在分布式计算环境之间卸载处理和计算资源以 便降低设施3102处的基础设施要求。在至少一个实施例中,部署系统3106 可以提供用于在设施3102处选择、定制和实现与成像装置(例如,MRI、CT扫描、X射线、超声等)或测序装置一起使用的虚拟仪器的流线型平台。 在至少一个实施例中,虚拟仪器可以包括软件定义的应用,用于关于由成 像装置、测序装置、放射学装置和/或其他装置类型生成的成像数据执行一 个或更多个处理操作。在至少一个实施例中,管线中的一个或更多个应用 程序可在应用程序的执行期间使用或调用部署系统3106的服务(例如,推 理、可视化、计算、AI等)。

[0368] 在至少一个实施例中,用于高级处理和推理管线的应用中的一些应 用可使用机

器学习模型或其他AI来执行一个或更多个处理步骤。在至少一 个实施例中,机器学习模型可以在设施3102处使用在设施3102处生成(并 且存储在设施3102处的一个或更多个图片归档和通信系统(PACS)服务 器上)的数据3108(诸如成像数据)来训练,可以使用来自另一个或更多 个设施(例如,不同的医院、实验室、诊所等)的成像或测序数据3108 来训练,或它们的组合。在至少一个实施例中,训练系统3104可以用于提 供用于为部署系统3106生成工作、可部署机器学习模型的应用、服务和/ 或其他资源。

[0369] 在至少一个实施例中,模型注册表3124可以由可以支持版本化和对 象元数据的对象存储来支持。在至少一个实施例中,对象存储可以通过例 如云存储(例如,图32的云3226)兼容的应用编程接口(API)从云平台 内访问。在至少一个实施例中,模型注册表3124内的机器学习模型可由与 API交互的系统的开发者或合作伙伴上传、列出、修改或删除。在至少一 个实施例中,API可以提供对如下方法的访问,该方法允许用户具有适当 的凭证来将模型与应用相关联,使得可以执行模型作为应用的容器化示例 化的执行的一部分。

[0370] 在至少一个实施例中,训练管线3204(图32)可以包括其中设施3102 正在训练其自己的机器学习模型、或具有需要被优化或更新的现有机器学 习模型的场景。在至少一个实施例中,可以接收由一个或更多个成像装置、测序装置和/或其他装置类型生成的成像数据3108。在至少一个实施例中, 一旦接收到成像数据3108,AI辅助注释3110可以用于帮助生成与成像数 据3108相对应的注释以用作机器学习模型的地面实况数据。在至少一个实 施例中,AI辅助注释3110可以包括一个或更多个机器学习模型(例如, 卷积神经网络(CNN)),所述一个或更多个机器学习模型可以被训练成用 于生成与某些类型的成像数据3108(例如,来自某些设备)和/或成像数据3108中的某些类型的异常相对应的注释。在至少一个实施例中,AI辅助注 释3110然后可以被直接使用,或者可以使用注释工具(例如,由研究人员、临床医生、医生、科学家等)被调整或微调以生成地面实况数据。在至少 一个实施例中,在一些示例中,标记的诊所数据3112(例如,由临床医生、医生、科学家、技师等提供的注释)可以用作用于训练机器学习模型的地 面实况数据。在至少一个实施例中,AI辅助注释3110、标记的诊所数据 3112或其组合可以用作用于训练机器学习模型的地面实况数据。在至少一 个实施例中,经训练的机器学习模型可以被称为输出模型3116,并且可以 由部署系统3106使用,如在此所描述的。

[0371] 在至少一个实施例中,训练管线3204(图32)可以包括其中设施3102 需要机器学习模型以用于执行部署系统3106中的一个或更多个应用的一 个或更多个处理任务的场景,但是设施3102当前可能不具有这样的机器学 习模型(或者可能不具有用于这样的目的的优化的、高效的或有效的模型)。在至少一个实施例中,可以从模型注册表3124中选择现有的机器学习模型。在至少一个实施例中,模型注册表3124可包括经训练以对成像数据执行各 种不同推理任务的机器学习模型。在至少一个实施例中,模型注册表3124 中的机器学习模型可能已经在来自不同于设施3102的设施(例如,位于远 程的设施)的成像数据上被训练。在至少一个实施例中,机器学习模型可 能已经在来自一个位置、两个位置或任意数量的位置的成像数据上训练。在至少一个实施例中,当对来自特定位置的成像数据进行训练时,训练可 以发生在该位置处,或者至少以保护成像数据的保密性或者限制成像数据不在场外传输(例如,以符合HIPAA规定、隐私规定等)的方式进行。在 至少一个实施例中,一旦在一个位置处对模型进行训练或部分训练,就可 以将机器学习模型添加到模型注册

表3124中。在至少一个实施例中，然后 可以在任何数量的其他设施处重新训练或更新机器学习模型，并且可以使 重新训练或更新的模型在模型注册表3124中可用。在至少一个实施例中， 机器学习模型然后可以从模型注册表3124中选择——并且被称为输出模 型3116——并且可以用于部署系统3106中以执行针对部署系统的一个或 更多个应用的一个或更多个处理任务。

[0372] 在至少一个实施例中，训练管线3204(图32)，一种场景可以包括 设施3102，该设施要求机器学习模型用于执行针对部署系统3106中的一 个或更多个应用的一个或更多个处理任务，但是设施3102当前可能不具有 这样的机器学习模型(或者可能不具有用于这样的目的的优化的、高效的 或有效的模型)。在至少一个实施例中，由于种群差异，从模型注册表3124 中选择的机器学习模型可以不针对在设施3102处生成的成像数据3108被 微调或优化，遗传变异、用于训练机器学习模型的训练数据的鲁棒性、训 练数据的异常的多样性，和/或训练数据的其他问题。在至少一个实施例中， AI辅助注释3110可以用于帮助生成对应于成像数据3108的注释，该成像 数据3108被用作用于重新训练或更新机器学习模型的地面实况数据。在至 少一个实施例中，标记的诊所数据3112(例如，由临床医生、医生、科学 家等提供的注释)可以用作用于训练机器学习模型的地面实况数据。在至 少一个实施例中，重新训练或更新机器学习模型可以被称为模型训练3114。 在至少一个实施例中，模型训练3114(例如，AI辅助注释3110、标记的 诊所数据3112或其组合)可以用作用于重新训练或更新机器学习模型的地 面实况数据。在至少一个实施例中，经训练的机器学习模型可以被称为输 出模型3116，并且可以由部署系统3106使用，如在此所描述的。

[0373] 在至少一个实施例中，部署系统3106可以包括软件3118、服务3120、硬件3122和/或其他组件、特征和功能。在至少一个实施例中，部署系统 3106可以包括软件“堆栈”，这样使得软件3118可以构建在服务3120之 上并且可以使用服务3120来执行一些或所有处理任务，并且服务3120和 软件3118可以构建在硬件3122之上并且使用硬件3122来执行处理，部署 系统3106的存储和/或其他计算任务。在至少一个实施例中，软件3118可 以包括任何数量的不同容器，其中每个容器可以执行应用的示例化。在至 少一个实施例中，每个应用可以在高级处理和推理管线中执行一个或更多 个处理任务(例如，推理、对象检测、特征检测、分割、图像增强、校准 等)。在至少一个实施例中，对于每种类型的成像装置(例如，CT、MRI、X射线、超声、超声检查、超声心动图等)，测序装置、放射学装置、基因 组学装置等，可以存在任何数量的容器，这些容器可以相对于由装置生成 的成像数据3108(或其他数据类型，如在此描述的那些)执行数据处理任 务。在至少一个实施例中，可基于对处理成像数据3108所要或所需的不同 容器的选择来定义高级处理和推理管线，除了接收和配置成像数据以供每 个容器使用和/或在通过管线处理之后供设施3102使用的容器之外(例如， 将输出转换回可用数据类型，如医学数字成像和通信(DICOM)数据、放 射学信息系统(RIS)数据、临床信息系统(CIS)数据、远程过程调用(RPC) 数据、基本上符合表示状态转移(REST)接口的数据、基本上符合基于文 件的接口的数据和/或原始数据，用于在设施3102处存储和显示)。在至少 一个实施例中，软件3118内的容器的组合(例如，组成管线的容器的组合)可被称为虚拟工具(如本文更详细地描述的)，并且虚拟工具可利用服务 3120和硬件3122来执行在容器中示例化的应用的一些或所有处理任务。

[0374] 在至少一个实施例中，数据处理管线可以响应于推理请求(例如， 来自部署系统

3106的用户(诸如临床医生、医生、放射科医生等)的请求) 以DICOM、RIS、CIS、REST兼容、RPC、原始和/或其他格式接收输入数 据(例如,成像数据3108)。在至少一个实施例中,输入数据可以表示由 一个或更多个成像装置、测序装置、放射学装置、基因组学装置、和/或其他装置类型生成的一个或更多个图像、视频、和/或其他数据表示。在至少 一个实施例中,数据可经历作为数据处理管线的一部分的预处理以准备数 据以供一个或更多个应用处理。在至少一个实施例中,可以对管线的一个 或更多个推理任务或其他处理任务的输出执行后处理,以便为下一应用准 备输出数据和/或准备输出数据以供用户传输和/或使用(例如,作为对推理 请求的响应)。在至少一个实施例中,推理任务可由一个或更多个机器学习模型来执行,诸如训练的或部署的神经网络,其可包括训练系统3104的输 出模型3116。

[0375] 在至少一个实施例中,数据处理管线的任务可以被封装在一个或更 多个容器中,所述容器各自表示能够引用机器学习模型的应用和虚拟化计 算环境的离散的、全功能的示例化。在至少一个实施例中,容器或应用可 被发布到容器注册表(本文更详细地描述)的私有(例如,受限访问)区 域中,并且训练的或部署的模型可被存储在模型注册表3124中并与一个或 更多个应用相关联。在至少一个实施例中,应用的图像(例如,容器图像) 可在容器注册表中可用,并且一旦由用户从容器注册表中选择以部署在管 线中,图像就可被用于生成容器以供应用的示例化以供用户的系统使用。

[0376] 在至少一个实施例中,开发者(例如,软件开发者、临床医生、医 生等)可开发、发布和存储用于对所提供的数据执行图像处理和/或推理的 应用(例如,作为容器)。在至少一个实施例中,开发、发布和/或存储可 使用与系统相关联的软件开发包(SDK)来执行(例如,以确保所开发的 应用和/或容器与系统兼容或兼容)。在至少一个实施例中,可以利用可以 支持服务3120中的至少一些作为系统(例如,图32的系统3200)的SDK 来本地(例如,在第一设施处,对来自第一设施的数据)测试所开发的应 用。在至少一个实施例中,因为DICOM对象可以包含从一到几百个图像 或其他数据类型的任何地方,并且由于数据的变化,开发者可以负责管理 (例如,设置构造、将预处理构建到应用中等)输入的DICOM数据的提 取和准备。在至少一个实施例中,一旦被系统3200验证(例如,为了准确 性、安全、患者隐私等),应用可以在容器注册表中是可用的,以供用户(例 如,医院、诊所、实验室、医疗保健提供者等)选择和/或实现,以便关于 用户的设施(例如,第二设施)处的数据执行一个或更多个处理任务。

[0377] 在至少一个实施例中,开发者然后可以通过网络共享应用或容器以 供系统(例如,图32的系统3200)的用户访问和使用。在至少一个实施 例中,完成的和经验证的应用或容器可以被存储在容器注册表中,并且相 关联的机器学习模型可以被存储在模型注册表3124中。在至少一个实施例 中,提供推理或图像处理请求的请求实体(例如,医疗设施处的用户)可 以浏览应用的容器注册表和/或模型注册表3124,容器、数据集、机器学习 模型等选择期望的元素组合以包括在数据处理管线中,以及提交成像处理 请求。在至少一个实施例中,请求可以包括执行请求所必需的输入数据(以 及在一些示例中相关联的患者数据),和/或可以包括在处理请求中要执行 的一个或更多个应用和/或机器学习模型的选择。在至少一个实施例中,请 求然后可被传递到部署系统3106的一个或更多个组件(例如,云)以执行 数据处理管线的处理。在至少一个实施例中,由部署系统3106进行的处理 可以包括从容器注册表和/或模型注册表3124引用所选择的元素(例如, 应用、容器、模型等)。在至

少一个实施例中，一旦由管线生成结果，结果 可被返回到用户以供参考(例如，以供在本地、现场工作站或终端上执行 的查看应用套件中查看)。在至少一个实施例中，放射科医生可从包括任何 数量的应用和/或容器的数据处理管线接收结果，其中结果可包括X射线、CT扫描、MRI等中的异常检测。

[0378] 在至少一个实施例中，为了帮助处理或执行管线中的应用或容器，可以利用服务3120。在至少一个实施例中，服务3120可包括计算服务、人工智能(AI)服务、可视化服务、和/或其他服务类型。在至少一个实施 例中，服务3120可以提供对软件3118中的一个或更多个应用共同的功能，因此可以将功能抽象到可以被应用调用或利用的服务。在至少一个实施例 中，由服务3120提供的功能可以动态地和更高效地运行，同时还通过允许 应用并行地处理数据(例如，使用并行计算平台3230(图32))来很好地 缩放。在至少一个实施例中，服务3120可在不同应用之间和之中共享，而 不是要求共享由服务3120提供的相同功能的每个应用具有服务3120的相 应示例。在至少一个实施例中，作为非限制性示例，服务可以包括可以用 于执行检测或分段任务的推理服务器或引擎。在至少一个实施例中，可以包括模型训练服务，其可以提供机器学习模型训练和/或重新训练能力。在 至少一个实施例中，可进一步包括数据增强服务，该数据增强服务可提供 GPU加速数据(例如，DICOM、RIS、CIS、REST兼容、RPC、原始等) 提取、调整大小、缩放和/或其他增强。在至少一个实施例中，可以使用可 视化服务，该可视化服务可以添加图像渲染效果——如光线追踪、光栅化、去噪、锐化等——以向二维(2D)和/或三维(3D)模型添加真实性。在 至少一个实施例中，可以包括为虚拟仪器的管线内的其他应用提供波束成 形、分割、推理、成像和/或支持的虚拟仪器服务。

[0379] 在至少一个实施例中，其中服务3120包括AI服务(例如，推理服 务)，可以通过调用(例如，作为API调用)推理服务(例如，推理服务 器)以执行一个或更多个机器学习模型来执行与用于异常检测的应用(例 如，肿瘤、生长异常、瘢痕形成等)相关联的一个或更多个机器学习模型，或其处理，作为应用执行的一部分。在至少一个实施例中，其中另一个应用包括用于分割任务的一个或更多个机器学习模型，应用可以根据推理服 务调用以执行用于执行与分割任务相关联的处理操作中的一个或更多个的 机器学习模型。在至少一个实施例中，实现包括分割应用和异常检测应用 的高级处理和推理管线的软件3118可被流线化，因为每个应用可调用相同 的推理服务来执行一个或更多个推理任务。

[0380] 在至少一个实施例中，硬件3122可包括GPU、CPU、图形卡、AI/ 深度学习系统(例如，AI超级计算机，诸如NVIDIA的DGX)、云平台或 其组合。在至少一个实施例中，不同类型的硬件3122可以用于为部署系统 3106中的软件3118和服务3120提供有效的、专门构建的支持。在至少一 个实施例中，GPU处理的使用可以实现用于本地处理(例如，在设施3102处)，在AI/深度学习系统内、在云系统中和/或在部署系统3106的其他处 理部件中以改善效率，图像处理、图像重构、分割、MRI检查、中风或心 脏病发作检测的准确度和功效(例如，实时)，渲染中的图像质量等。在至 少一个实施例中，设施可以包括成像装置、基因组学装置、测序装置、和/ 或现场的其他装置类型，这些装置可以利用GPU来生成表示受试者的解剖 结构的成像数据。在至少一个实施例中，作为非限制性示例，软件3118 和/或服务3120可针对GPU处理相对于深度学习、机器学习和/或高性能计 算进行优化。在至少一个实施例中，部署系统3106和/或训练系统3104的 计算环境中的至少一些可以在数据中心中执行一

个或更多个超级计算机或 高性能计算系统,其中GPU优化软件(例如,NVIDIA的DGX系统的硬 件和软件组合)。在至少一个实施例中,数据中心可以符合HIPAA的规定, 使得成像数据和/或其他患者数据的接收、处理和传输相对于患者数据的隐 私被安全地处理。在至少一个实施例中,硬件3122可以包括可以被调用以 并行执行对数据的处理的任何数量的GPU,如在此所描述的。在至少一个 实施例中,云平台还可包括用于深度学习任务、机器学习任务或其他计算 任务的GPU优化执行的GPU处理。在至少一个实施例中,可以使用AI/ 深度学习超级计算机和/或GPU优化的软件(例如,在NVIDIA的DGX系 统上提供的)作为硬件抽象和缩放平台来执行云平台(例如,NVIDIA的 NGC)。在至少一个实施例中,云平台可以将应用容器集群系统或编排系 统(例如,KUBERNETES)集成在多个GPU上以实现无缝缩放和负载平衡。

[0381] 图32是根据至少一个实施例的用于生成和部署成像部署管道的示例 系统3200的系统图。在至少一个实施例中,系统3200可以用于实现图31 的过程3100和/或包括高级处理和推理管线的其他过程。在至少一个实施 例中,系统3200可以包括训练系统3104和部署系统3106。在至少一个实 施例中,训练系统3104和部署系统3106可以使用软件3118、服务3120 和/或硬件3122来实施,如在此所描述的。

[0382] 在至少一个实施例中,系统3200(例如,训练系统3104和/或部署 系统3106)可以在云计算环境中实现(例如,使用云3226)。在至少一个 实施例中,系统3200可以相对于医疗保健服务设施在本地实现,或者实现 为云计算资源和本地计算资源两者的组合。在至少一个实施例中,在实现 云计算的实施例中,患者数据可以与系统3200的将使处理不符合HIPAA 和/或其他数据处理和隐私规定或法律的一个或更多个组件分离或被其处 理。在至少一个实施例中,通过制定的安全措施或协议,对云3226中的 API的访问可被限制到经授权的用户。在至少一个实施例中,安全协议可 包括可由认证(例如,AuthN、AuthZ、Gluecon等)服务签名并且可携带 适当授权的web令牌。在至少一个实施例中,(本文描述的)虚拟工具的 API或系统3200的其他示例化可被限制到已被审查或授权进行交互的一组 公共IP。

[0383] 在至少一个实施例中,系统3200的各种组件可以使用各种不同网络 类型中的任何网络类型在彼此之间和之间进行通信,包括但不限于局域网 (LAN)和/或广域网(WAN),经由有线和/或无线通信协议。在至少一个 实施例中,系统3200的设施和组件之间的通信(例如,用于传输推理请求、 用于接收推理请求的结果等)可通过数据总线、无线数据协议(Wi-Fi)、 有线数据协议(例如,以太网)等传送。

[0384] 在至少一个实施例中,训练系统3104可以执行训练管线3204,类似 于在此关于图31描述的那些。在至少一个实施例中,其中将由部署系统 3106在部署管线3210中使用一个或更多个机器学习模型,训练管线3204 可用于训练或重新训练一个或更多个(例如,预训练的)模型,和/或实现 一个或更多个预训练模型3206(例如,不需要重新训练或更新)。在至少 一个实施例中,作为训练管线3204的结果,可生成一个或更多个输出模型3116。在至少一个实施例中,训练管线3204可包括任何数量的处理步骤, 诸如但不限于成像数据(或其他输入数据)转换或适配(例如,使用DICOM 适配器3202A将DICOM图像转换成适于由相应机器学习模型处理的另一 格式,诸如神经成像信息技术启发(NIfTI)格式),AI辅助注释3110,标 记或注释成像数据3108以生成标记的诊所数据3112,从模型注册表中的 模型选择、模型训练3114、训练、重新训练或更新模型、和/或其他处理步 骤。在至少一个实施例

中,对于由部署系统3106使用的不同的机器学习模 型,可以使用不同的训练管线3204。在至少一个实施例中,类似于参见图 31描述的第一示例的训练管线3204可以用于第一机器学习模型,类似于 参见图31描述的第二示例的训练管线3204可以用于第二机器学习模型,并且类似于关于图31所描述的第三示例的训练管线3204可以用于第三机 器学习模型。在至少一个实施例中,可以根据每个相应的机器学习模型需 要什么来使用训练系统3104内的任务的任何组合。在至少一个实施例中, 机器学习模型中的一个或更多个可能已经被训练并且准备好部署,因此机 器学习模型可以不经历由训练系统3104进行的任何处理,并且可以由部署 系统3106来实现。

[0385] 在至少一个实施例中,取决于实现方式或实施例,一个或更多个输 出模型3116和/或一个或更多个预训练模型3206可以包括任何类型的机器 学习模型。在至少一个实施例中,但不限于,系统3200所使用的机器学习 模型可以包括使用线性回归的一个或更多个机器学习模型,逻辑回归、决 策树、支持向量机(SVM)、原始贝叶斯、k最近邻(Knn)、K均值聚类、 随机森林、维度减少算法、梯度提升算法、神经网络(例如,自动编码器、 卷积、递归、感知器、长/短期记忆(LSTM)、霍普菲尔德、玻尔兹曼、深 度置信、解卷积、生成对手、液态机器等)和/或其他类型的机器学习模型。

[0386] 在至少一个实施例中,训练管线3204可以包括AI辅助注释,如在 此相对于至少图35B更详细地描述的。在至少一个实施例中,可通过任何 数量的技术生成经标记的诊所数据3112(例如,传统注释)。在至少一个 实施例中,标签或其他注释可在绘图程序(例如,注释程序)、计算机辅助 设计(CAD)程序、标签程序、适于生成用于地面实况的注释或标签的另 一类型的程序内生成,和/或在一些示例中可被手绘。在至少一个实施例中, 地面实况数据可以是合成产生的(例如,从计算机模型或渲染中生成),真 实产生(例如,从真实世界数据设计和产生)、机器自动化(例如,使用特 征分析和学习从数据提取特征且接着产生标记),人工注释的(例如,标签 者或注释专家定义标签的位置)和/或其组合。在至少一个实施例中,对于 成像数据3108(或由机器学习模型使用的其他数据类型)的每个示例,可 存在由训练系统3104生成的对应的地面实况数据。在至少一个实施例中, AI辅助注释可以作为部署管线3210的一部分来执行;作为包括在训练管 线3204中的AI辅助注释的补充或替代。在至少一个实施例中,系统3200 可以包括多层平台,该多层平台可以包括可以执行一个或更多个医学成像 和诊断功能的诊断应用(或其他应用类型)的软件层(例如,软件3118)。 在至少一个实施例中,系统3200可通信地耦接至(例如,经由加密链路) 一个或更多个设施的PACS服务器网络。在至少一个实施例中,系统3200 可以被配置为从PACS服务器(例如,经由DICOM适配器3202或者诸如 RIS、CIS、REST兼容、RPC、原始数据等的另一数据类型适配器)访问 和引用数据(例如,DICOM数据、RIS数据、原始数据、CIS数据、REST 兼容数据、RPC数据、原始数据等)以执行操作,诸如训练机器学习模型、 部署机器学习模型、图像处理、推理和/或其他操作。

[0387] 在至少一个实施例中,软件层可以被实现为安全的、加密的和/或经 认证的API,通过该API,可以从一个或更多个外部环境(例如,设施3102) 调用(例如,调用)应用或容器。在至少一个实施例中,应用然后可以调 用或执行一个或更多个服务3120以用于执行与相应应用相关联的计算、 AI或可视化任务,并且软件3118和/或服务3120可以利用硬件3122以有 效和高效的方式执行处理任务。

[0388] 在至少一个实施例中，部署系统3106可以执行部署管线3210。在至 少一个实施例中，部署管线3210可以包括可以顺序地、非顺序地或以其他 方式应用于由成像装置、测序装置、基因组学装置等生成的成像数据(和/ 或其他数据类型)的任何数量的应用——包括如上所述的AI辅助注释。在 至少一个实施例中，如本文所述，用于个别装置的部署管线3210可以被称 为用于装置的虚拟仪器(例如，虚拟超声仪器、虚拟CT扫描仪器、虚拟 测序仪器等)。在至少一个实施例中，对于单个装置，根据装置产生的数据 所需的信息，可以有多于一个部署管线3210。在至少一个实施例中，在期 望从MRI机器检测异常的情况下，可存在第一部署管线3210，并且在期 望从MRI机器的输出进行图像增强的情况下，可存在第二部署管线3210。

[0389] 在至少一个实施例中，可用于部署管线3210的应用程序可包括可用 于对成像数据或来自装置的其他数据执行处理任务的任何应用程序。在至 少一个实施例中，不同的应用可负责图像增强、分割、重构、异常检测、 物体检测、特征检测、治疗计划、剂量测定、射束计划(或其他放射治疗 程序)和/或其他分析、图像处理或推理任务。在至少一个实施例中，部署 系统3106可以定义用于每个应用的构造，使得部署系统3106的用户(例 如，医疗设施、实验室、诊所等)可以理解构造和适配应用以便在它们各 自的设施内实现。在至少一个实施例中，可选择用于图像重构的应用以包 含在部署管线3210中，但是由成像装置生成的数据类型可不同于在应用内 使用的数据类型。在至少一个实施例中，DICOM适配器3202B(和/或 DICOM读取器)或另一数据类型适配器或读取器(例如，RIS、CIS、REST 兼容、RPC、原始等)可在部署管线3210内使用以将数据转换成可由部署 系统3106内的应用使用的形式。在至少一个实施例中，可以累积和预处理 对DICOM、RIS、CIS、REST兼容、RPC、原始和/或其他数据类型库的访 问，包括解码、提取和/或执行任何卷积、颜色校正、锐度、伽马和/或对数 据的其他增强。在至少一个实施例中，DICOM、RIS、CIS、REST兼容、 RPC和/或原始数据可以是无序的，并且可以执行预遍以组织或排序所收集 的数据。在至少一个实施例中，因为不同应用可以共享共同的图像操作， 所以在一些实施例中，数据增强库(例如，作为服务3120之一)可以用于 加速这些操作。在至少一个实施例中，为了避免依赖于CPU处理的常规处 理方法的瓶颈，并行计算平台3230可用于这些处理任务的GPU加速。

[0390] 在至少一个实施例中，图像重构应用可以包括处理任务，该处理任 务包括使用机器学习模型。在至少一个实施例中，用户可能期望使用他们 自己的机器学习模型，或者从模型注册表3124中选择机器学习模型。在至 少一个实施例中，用户可以实现他们自己的机器学习模型或选择机器学习 模型以包括在用于执行处理任务的应用中。在至少一个实施例中，应用可 以是可选择的和可定制的，并且通过定义应用的构造，用于特定用户的应 用的部署和实现被呈现为更无缝的用户体验。在至少一个实施例中，通过 利用系统3200的其他特征——如服务3120和硬件3122——部署管线3210 可以甚至更加用户友好，提供更容易的集成，并且产生更准确、高效和及 时的结果。

[0391] 在至少一个实施例中，部署系统3106可以包括用户界面3214(例如， 图形用户界面、网络界面等)，该用户界面3214可以用于选择用于包括在 部署管线3210中的应用，在建立和/或部署期间安排应用、修改或改变应 用或其参数或构造、使用部署管线3210和与部署管线3210交互，和/或以 其他方式与部署系统3106交互。在至少一个实施例中，尽管未相对于训练 系统3104示出，用户界面3214(或不同的用户界面)可以用于选择在部 署系统

3106中使用的模型、用于选择在训练系统3104中用于训练或重新 训练的模型和/或用于以其他方式与训练系统3104交互。

[0392] 在至少一个实施例中，除了应用编排系统3228之外，可以使用管线 管理器3212来管理部署管道3210的应用或容器与服务3120和/或硬件 3122之间的交互。在至少一个实施例中，管线管理器3212可以被配置为 促进从应用到应用、从应用到服务3120和/或从应用或服务到硬件3122的 交互。在至少一个实施例中，尽管被示为包括在软件3118中，但这不旨在 是限制性的，并且在一些示例中(例如，如图33所示)，管线管理器3212 可被包括在服务3120中。在至少一个实施例中，应用编排系统3228(例 如，Kubernetes、DOCKER等)可以包括容器编排系统，该容器编排系统 可以将应用分组到容器中作为用于协调、管理、扩展和部署的逻辑单元。 在至少一个实施例中，通过将来自部署管线3210的应用(例如，重构应用、 分段应用等)与单独的容器相关联，每个应用可以在自包含环境中(例如， 在内核级)执行以提高速度和效率。

[0393] 在至少一个实施例中，每个应用和/或容器(或其图像)可以被单独 地开发，经修改和部署(例如，第一用户或开发者可开发、修改和部署第 一应用，而第二用户或开发者可开发、修改和部署与第一用户或开发者分 开的第二应用)，这可以允许关注和注意单个应用和/或容器的任务，而不 受另一应用或容器的任务的阻碍。在至少一个实施例中，不同容器或应用 之间的通信和协作可以由管线管理器3212和应用编排系统3228辅助。在 至少一个实施例中，只要系统(例如，基于应用或容器的构造)知道每个 容器或应用的预期输入和/或输出，应用编排系统3228和/或管线管理器 3212可以促进它们之间和之间的通信，以及在每个应用或容器之间和之间 共享资源。在至少一个实施例中，因为部署管线3210中的一个或更多个应 用或容器可共享相同的服务和资源，所以应用编排系统3228可编排、负载平衡和确定在不同应用或容器之间和之中服务或资源的共享。在至少一个 实施例中，调度器可以用于追踪应用或容器的资源需求、这些资源的当前 使用或计划使用和资源可用性。在至少一个实施例中，调度器可以因此向 不同的应用分配资源并且鉴于系统的需求和可用性在应用之间和之中分配 资源。在一些示例中，调度器(和/或应用程序编排系统3228的其他部件) 可以基于强加于系统上的约束(例如，用户约束)确定资源可用性和分布， 如服务质量(QoS)、需要数据输出的紧急性(例如，确定是否执行实时处 理或延迟处理)等。

[0394] 在至少一个实施例中，由部署系统3106中的应用或容器利用和共享 的服务3120可以包括计算服务3216、AI服务3218、可视化服务3220和/ 或其他服务类型。在至少一个实施例中，应用程序可调用(例如，执行) 服务3120中的一个或更多个来执行应用程序的处理操作。在至少一个实施 例中，计算服务3216可由应用利用来执行超级计算或其他高性能计算 (HPC)任务。在至少一个实施例中，一个或更多个计算服务3216可被利 用来执行并行处理(例如，使用并行计算平台3230)以用于通过应用中的 一个或更多个和/或单个应用的一个或更多个任务来基本上同时地处理数 据。在至少一个实施例中，并行计算平台3230(例如，NVIDIA的CUDA) 可以使能GPU(GPGPU)(例如，GPU3222)上的通用计算。在至少一个 实施例中，并行计算平台3230的软件层可以提供对GPU的虚拟指令集和 并行计算元件的访问，以便执行计算内核。在至少一个实施例中，并行计 算平台3230可以包括存储器，并且在一些实施例中，可以在多个容器之间 和之中和/或在单个容器内的不同处理任务之间和之中共享存储器。在至少 一个实施例中，可以针对多个容器和/或针对容器内的多个进程生成

进程间 通信(IPC)调用以便使用来自并行计算平台3230的存储器的共享片段的 相同数据(例如,在应用程序的多个不同阶段或多个应用程序正在处理相 同信息的情况下)。在至少一个实施例中,不是复制数据并将数据移动到存 储器中的不同位置(例如,读/写操作),而是将存储器的相同位置中的相 同数据用于任何数量的处理任务(例如,同时、不同时间等)。在至少一个 实施例中,由于使用数据作为处理的结果来生成新数据,因此可以在不同应用之间存储和共享数据的新位置的该信息。在至少一个实施例中,数据 的位置和更新的或修改的数据的位置可以是如何在容器内理解有效负载的 定义的一部分。

[0395] 在至少一个实施例中,可以利用AI服务3218来执行用于执行与应 用相关联的机器学习模型(例如,被指派执行应用的一个或更多个处理任 务的任务)的推理服务。在至少一个实施例中,AI服务3218可以利用AI 系统3224来执行机器学习模型(例如,神经网络,诸如CNN)以用于分 段、重构、对象检测、特征检测、分类和/或其他推理任务。在至少一个实施例中,一个或更多个部署管线3210的应用可以使用来自训练系统3104 的输出模型3116中的一个或更多个输出模型和/或应用的其他模型来对成 像数据(例如,DICOM数据、RIS数据、CIS数据、REST兼容数据、RPC 数据、原始数据等)执行推理。在至少一个实施例中,使用应用编排系统 3228(例如,调度器)进行推理的两个或更多个示例可以是可用的。在至 少一个实施例中,第一类别可包括高优先级/低等待时间路径,其可实现较 高服务水平协议,诸如用于在紧急情况期间对紧急请求执行推理,或在诊 断期间对放射科医生执行推理。在至少一个实施例中,第二类别可以包括 标准优先级路径,该标准优先级路径可以用于可能非紧急的请求或其中可 以稍后执行分析的请求。在至少一个实施例中,应用编排系统3228可以基 于用于AI服务3218的不同推理任务的优先级路径来分配资源(例如,服 务3120和/或硬件3122)。

[0396] 在至少一个实施例中,共享存储可以安装到系统3200内的AI服务 3218。在至少一个实施例中,共享存储可以作为高速缓存(或其他存储器 类型)操作,并且可以用于处理来自应用的推理请求。在至少一个实施例 中,当提交推理请求时,可由部署系统3106的一组API示例接收请求, 并且可选择一个或更多个示例(例如,为了最佳拟合、为了负载平衡等)来处理请求。在至少一个实施例中,为了处理请求,可以将请求输入到数 据库中,如果尚未在高速缓存中,则可以从模型注册表3124定位机器学习 模型,验证步骤可以确保适当的机器学习模型被加载到高速缓存(例如, 共享存储)中,和/或模型的副本可以被保存到高速缓存。在至少一个实施 例中,(例如,管线管理器3212的)调度器可用于在应用尚未运行或应用 不存在足够的应用示例的情况下启动在请求中引用的应用。在至少一个实 施例中,如果尚未启动推理服务器来执行模型,则可以启动推理服务器。每个模型可以启动任何数量的推理服务器。在至少一个实施例中,在其中 推理服务器被聚类的拉模型中,每当负载平衡有利时,模型可被高速缓存。在至少一个实施例中,推理服务器可以被静态地加载到对应的分布式服务 器中。

[0397] 在至少一个实施例中,可使用在容器中运行的推理服务器来执行推 理。在至少一个实施例中,推理服务器的示例可与模型(以及可选地模型 的多个版本)相关联。在至少一个实施例中,如果在接收到对模型执行推 理的请求时推理服务器的示例不存在,则可加载新示例。在至少一个实施 例中,当启动推理服务器时,可将模型传递给推理服务器,使得只要推理 服务器作为不同示例运行,就可使用同一容器来服务不同模型。

[0398] 在至少一个实施例中,在应用程序执行期间,可以接收针对给定应 用程序的推理请求,并且可以加载容器(例如,托管推理服务器的示例) (如果尚未加载),并且可以调用开始过程。在至少一个实施例中,容器中 的预处理逻辑可加载、解码传入数据和/或对传入数据执行任何附加预处理 (例如,使用CPU和/或GPU)。在至少一个实施例中,一旦数据准备好进 行推理,容器可根据需要对数据执行推理。在至少一个实施例中,这可包 括对一个图像(例如,手X射线)的单个推理调用,或可要求对数百个图 像(例如,胸部CT)的推理。在至少一个实施例中,应用可以在完成之前 总结结果,这可以包括但不限于单个置信度得分、像素级分割、体素级分 割、生成可视化、或生成文本以总结发现。在至少一个实施例中,不同的 模型或应用可被指派不同的优先级。例如,一些模型可具有实时(TAT＜1 分钟)优先级,而其他模型可具有较低优先级(例如,TAT＜10分钟)。在 至少一个实施例中,可以从请求机构或实体测量模型执行时间,并且模型 执行时间可以包括伙伴网络遍历时间,以及在推理服务上的执行。

[0399] 在至少一个实施例中,服务3120与推理应用之间的请求的传送可被 隐藏在软件开发包(SDK)之后,并且可通过队列提供稳健的传输。在至 少一个实施例中,请求将经由单独的应用/租户ID组合的API被放置在队 列中,并且SDK将从队列中拉取请求并将请求给予应用。在至少一个实施 例中,可以在SDK将从其中拾取队列的环境中提供队列的名称。在至少一 个实施例中,通过队列的异步通信可以是有用的,因为它可以允许应用的 任何示例在其变得可用时拾取工作。结果可通过队列传送回来,以确保没 有数据损失。在至少一个实施例中,队列还可以提供分割工作的能力,因 为最高优先级工作可以转到具有与其连接的应用的大多数示例的队列,而 最低优先级工作可以转到具有与其连接的单个示例的队列,其以接收的顺 序处理任务。在至少一个实施例中,应用可以在云3226中生成的GPU加速的示例上运行,并且推理服务可以对GPU执行推理。

[0400] 在至少一个实施例中,可利用可视化服务3220来生成用于查看应用 和/或一个或更多个部署管线3210的输出的可视化。在至少一个实施例中, GPU3222可被可视化服务3220利用来生成可视化。在至少一个实施例中, 呈现效果(诸如光线追踪)可由可视化服务3220实现以生成更高质量的可 视化。在至少一个实施例中,可视化可以包括但不限于2D图像渲染、3D 体积渲染、3D体积重建、2D断层摄影切片、虚拟现实显示器、增强现实 显示器等。在至少一个实施例中,虚拟化环境可以用于生成用于由系统的 用户(例如,医生、护士、放射科医生等)交互的虚拟交互式显示器或环 境(例如,虚拟环境)。在至少一个实施例中,可视化服务3220可以包括 内部可视化器、动态图像和/或其他渲染或图像处理能力或功能(例如,光 线追踪、光栅化、内部光学器件等)。

[0401] 在至少一个实施例中,硬件3122可包括GPU3222、AI系统3224、 云3226和/或用于执行训练系统3104和/或部署系统3106的任何其他硬件。 在至少一个实施例中,GPU3222(例如,NVIDIA的TESLA和/或QUADRO GPU)可包括可用于执行计算服务3216、AI服务3218、可视化服务3220、 其他服务和/或软件3118的特征或功能中的任何一个的处理任务的任何数量的GPU。例如,关于AI服务3218,GPU3222可以用于对成像数据(或 由机器学习模型使用的其他数据类型)执行预处理,对机器学习模型的输 出执行后处理,和/或用于执行推理(例如,用于执行机器学习模型)。在 至少一个实施例中,云3226、AI系统3224和/或系统3200的其他部件可 以使用GPU3222。在至少一个实施例中,云3226可包括用于深度学习任 务的

GPU优化的平台。在至少一个实施例中，AI系统3224可以使用GPU，并且可以使用一个或更多个AI系统3224来执行云3226或任务为深度学习 或推理的至少一部分。照此，尽管硬件3122被示为分立组件，但这不旨在 是限制性的，并且硬件3122的任何组件可以与硬件3122的任何其他组件 组合或被其利用。

[0402] 在至少一个实施例中，AI系统3224可以包括被配置成用于推理、深 度学习、机器学习和/或其他人工智能任务的专用计算系统(例如，超级计 算机或HPC)。在至少一个实施例中，AI系统3224(例如，NVIDIA的DGX) 可以包括除了CPU、RAM、存储装置和/或其他组件、特征或功能性之外 还可以使用多个GPU3222执行的GPU优化的软件(例如，软件栈)。在至少一个实施例中，一个或更多个AI系统3224可以在云3226中(例如，在 数据中心中)实现以用于执行系统3200的基于AI的处理任务中的一些或 全部。

[0403] 在至少一个实施例中，云3226可包括可提供用于执行系统3200的 处理任务的GPU优化的平台的GPU加速的基础结构(例如，NVIDIA的 NGC)。在至少一个实施例中，云3226可以包括用于执行系统3200的基 于AI的任务中的一个或更多个(例如，作为硬件抽象和缩放平台)的AI 系统3224。在至少一个实施例中，云3226可以与利用多个GPU的应用编 排系统3228集成，以使得能够在应用和服务3120之间和之中进行无缝缩 放和负载平衡。在至少一个实施例中，云3226的任务可以是执行系统3200 的服务3120中的至少一些，包括计算服务3216、AI服务3218和/或可视 化服务3220，如本文所描述的。在至少一个实施例中，云3226可执行小 批量推理和大批量推理(例如，执行NVIDIA的TENSORRT)，提供加速 的并行计算API和平台3230(例如，NVIDIA的CUDA)，执行应用编排 系统3228(例如，KUBERNETES)，提供图形渲染API和平台(例如，用 于光线追踪、2D图形、3D图形、和/或其他渲染技术以产生更高质量的影 片学)，和/或可以为系统3200提供其他功能。

[0404] 在至少一个实施例中，为了保护患者机密性(例如，在患者数据或 记录有待不在现场使用的情况下)，云3226可以包括注册表——如深度学 习容器注册表。在至少一个实施例中，注册表可以存储用于对患者数据执 行预处理、后处理或其他处理任务的应用的示例的容器。在至少一个实施 例中，云3226可接收包括容器中的患者数据以及传感器数据的数据，仅对 那些容器中的传感器数据执行所请求的处理，然后将所得的输出和/或可视 化内容转发到适当的各方和/或设备(例如，用于可视化或诊断的现场医疗 设备)，全部不必提取、存储或以其他方式访问患者数据。在至少一个实施 例中，遵照HIPAA和/或其他数据规定来保持患者数据的机密性。

[0405] 图33包括根据至少一个实施例的用于处理成像数据的部署管线 3210A的示例图示。在至少一个实施例中，系统3200——并且具体地部署 系统3106——可以用于定制、更新、和/或将一个或更多个部署管线3210A 集成到一个或更多个生产环境中。在至少一个实施例中，图33的部署管线 3210A包括可由设施处(例如，医院、诊所、实验室、研究环境等处)的 特定用户(或用户组)自定义的部署管线3210A的非限制性示例。在至少 一个实施例中，为了定义用于CT扫描仪3302的部署管线3210A，用户可 以从容器注册表中选择例如关于由CT扫描仪3302生成的成像数据执行特 定功能或任务的一个或更多个应用。在至少一个实施例中，应用可以应用 于部署管线3210A，作为可以利用系统3200的服务3120和/或硬件3122 的容器。此外，部署管线3210A可以包括可以被实施来准备数据以供应用 使用的附加处理任务或应用(例如，DICOM适配器3202B和DICOM读 取器3306可以在部署管线3210A中被

用来准备数据以供CT重构3308、器官分割3310等使用)。在至少一个实施例中,部署管线3210A可以被定 制或选择用于一致的部署、一次使用、或另一频率或间隔。在至少一个实施例中,用户可能希望在特定间隔上针对若干受试者进行CT重构3308和 器官分割3310,并且因此可以在该时间段中部署管线3210A。在至少一个 实施例中,用户可以为来自系统3200的每个请求选择用户想要对该请求的 该数据执行处理的应用。在至少一个实施例中,可以以任何间隔调整部署 管线3210A,并且由于系统3200内的容器结构的适应性和可扩展性,这可 以是无缝过程。

[0406] 在至少一个实施例中,图33的部署管线3210A可包括生成患者或受 试者的成像数据的CT扫描仪3302。在至少一个实施例中,来自CT扫描 仪3302的成像数据可被存储在与容纳CT扫描仪3302的设施相关联的 PACS服务器3304上。PACS服务器3304可包括可直接与设施处的成像模 态(例如,CT扫描仪3302)对接的软件和/或硬件组件。在至少一个实施 例中,DICOM适配器3202B可使用DICOM协议发送和接收DICOM对象。 在至少一个实施例中,DICOM适配器3202B可帮助准备或配置来自PACS 服务器3304的DICOM数据以由部署管路3210A使用。在至少一个实施例 中,一旦通过DICOM适配器3202B处理DICOM数据,管线管理器3212 可以将数据路由到部署管道3210A。在至少一个实施例中,DICOM读取器 3306可以从DICOM数据(例如,原始正弦图像数据,如可视化3316A所 示)中提取图像文件和任何相关联的元数据。在至少一个实施例中,可以 将所提取的工作文件存储在高速缓存中以便由部署管线3210A中的其他应 用进行更快的处理。在至少一个实施例中,一旦DICOM读取器3306完成 提取和/或存储数据,完成的信号就可以被传送到管线管理器3212。在至少 一个实施例中,管线管理器3212然后可以发起或调用部署管线3210A中 的一个或更多个其他应用或容器。

[0407] 在至少一个实施例中,一旦数据(例如,原始正弦图数据)可用于 由CT重构3308应用处理,就可执行CT重构3308应用和/或容器。在至 少一个实施例中,CT重构3308可从高速缓存中读取原始正弦图像数据, 从原始正弦图像数据中重构图像文件(例如,如可视化3316B中所解说的), 并将所得图像文件存储在高速缓存中。在至少一个实施例中,在重构完成 时,可用信号通知管线管理器3212重构任务完成。在至少一个实施例中, 一旦重建完成,并且重建的图像文件可以存储在高速缓存(或其他存储装 置)中,器官分割3310应用和/或容器可以由管线管理器3212触发。在至 少一个实施例中,器官分割3310应用和/或容器可以从高速缓存中读取图 像文件,将图像文件归一化或转换成适合于推理的格式(例如,将图像文 件转换成机器学习模型的输入分辨率),并且针对归一化图像运行推理。在至少一个实施例中,为了对归一化图像运行推理,器官分割3310应用和/ 或容器可以依赖于服务3120,并且管线管理器3212和/或应用编排系统3228可以促进器官分割3310应用和/或容器对服务3120的使用。例如,器 官分割3310应用和/或容器可以利用AI服务3218来对归一化图像执行推 理,并且AI服务3218可以利用硬件3122(例如,AI系统3224)来执行 AI服务3218。在至少一个实施例中,推理的结果可以是可以存储在高速缓 存(或其他存储器)中的掩码文件(例如,如可视化3316C所示)。

[0408] 在至少一个实施例中,一旦处理DICOM数据和/或从DICOM数据提 取的数据的应用已经完成处理,可以为管线管理器3212生成信号。在至少 一个实施例中,管线管理器3212然后可执行DICOM写入器3312以从高 速缓存(或其他存储器)读取结果,将结果封装成

DICOM格式(例如，作为DICOM输出3314)以供生成请求的设施处的用户使用。在至少一个实施例中，然后可将DICOM输出3314传输至DICOM适配器3202B以准 备DICOM输出3314用于存储在PACS服务器3304上(例如，用于由设 施处的DICOM查看器查看)。在至少一个实施例中，响应于重构和分割的 请求，可生成可视化3316B和3316C，并且可可视化3316B和3316C对于 用户可用于诊断、研究和/或其他目的。

[0409] 虽然在至少一个实施例中被示出为在部署管线3210A中的连续应用，但是可以并行地处理CT重构3308和器官分割3310应用。在至少一个实 施例中，在应用程序彼此不具有依赖性并且数据可用于每个应用程序(例 如，在DICOM读取器3306提取数据之后)的情况下，应用程序可同时、基本上同时或以一些重叠执行。在至少一个实施例中，在两个或更多个应 用需要类似的服务3120的情况下，系统3200的调度器可用于在不同应用 之间和之中负载平衡和分配计算或处理资源。在至少一个实施例中，在一 些实施例中，并行计算平台3230可以用于为应用执行并行处理以减少部署 管线3210A的运行时间从而提供实时结果。

[0410] 在至少一个实施例中，并且参见图34A-34B，部署系统3106可以被 实施为用于执行不同功能(如图像处理)的一个或更多个虚拟仪器——例 如图像处理、分割、增强、AI、可视化、和推理——使用成像设备(例如，CT扫描仪、X射线机、MRI机等)，测序设备、基因组学设备和/或其他设 备类型。在至少一个实施例中，系统3200可以允许创建和提供虚拟工具，所述虚拟工具可以包括软件定义的部署管线3210，所述部署管线可以接收 由一个或更多个设备生成的原始/未经处理的输入数据和输出经处理/重构 的数据。在至少一个实施例中，表示虚拟仪器的部署管线3210(例如，3210A 和3210B)可以诸如通过利用机器学习模型将智能实施到管线中，以向系 统提供容器化的推理支持。在至少一个实施例中，虚拟工具可执行任何数 量的容器，每个容器包括应用的示例，诸如在期望实时处理的情况下，表 示虚拟工具的部署管线3210可以是静态的(例如，可以设置容器和/或应 用)，而在其他示例中，可以(例如，在每个请求的基础上)从应用或资源 池(例如，在容器注册表内)选择用于虚拟工具的容器和/或应用。

[0411] 在至少一个实施例中，系统3200可以被示例化或执行为例如部署在 设施处的放射科机器、成像装置和/或另一装置类型旁边或以其他方式与其 通信的计算系统中的设施处的设施处的一个或更多个现场虚拟仪器。然而， 在至少一个实施例中，场所内安装可在设备自身的计算系统(例如，集成 到成像设备的计算系统)内、在本地数据中心(例如，场所内数据中心) 中和/或在云环境(例如，云3226)中示例化或执行。在至少一个实施例中， 在一些示例中，作为虚拟仪器操作的部署系统3106可以由超级计算机或其 他HPC系统来示例化。在至少一个实施例中，场所内安装可允许用于实时 处理的高带宽使用(例如，经由较高吞吐量的本地通信接口，诸如以太网 上的RF)。在至少一个实施例中，在虚拟仪器支持超声设备或其他成像模 态的情况下，实时或近实时处理可能是特别有用的，在所述其他成像模态 中，对于精确的诊断和分析预期或需要即时可视化。在至少一个实施例中， 当本地需求超过场所内容量或能力时，云计算架构可能能够向云计算服务 提供商或其他计算集群动态突发。在至少一个实施例中，云架构在被实现 时可被调谐用于训练神经网络或其他机器学习模型，如本文中相对于训练 系统3104所描述的。在至少一个实施例中，在训练管线就位的情况下，机 器学习模型可以在它们处理来自它们支持的设备的附加数据时被连续地学 习和改进。在至少一个实施例中，可以使用附加数据、新数据、现有机器 学习模型和/或

新的或更新的机器学习模型来连续地改进虚拟工具。

[0412] 在至少一个实施例中,计算系统可以包括在此描述的硬件3122中的 一些或全部,并且硬件3122可以以多种方式中的任何方式分布,包括在设 备内,作为耦合到设备并位于设备附近的计算设备的一部分,在设施处的 本地数据中心中和/或在云3226中。在至少一个实施例中,因为部署系统 3106和相关联的应用或容器是在软件中创建的(例如,作为应用的离散的 容器化示例),虚拟仪器的行为、操作和配置以及由虚拟仪器生成的输出, 可以根据需要修改或定制,而不必改变或改变虚拟仪器支持的设备的原始 输出。

[0413] 图34A包括根据至少一个实施例的支持超声设备的虚拟仪器的示例 数据流程图。在至少一个实施例中,部署管线3210B可利用系统3200的 服务3120中的一个或更多个。在至少一个实施例中,部署管线3210B和 服务3120可以在本地或在云3226中利用系统的硬件3122。在至少一个实 施例中,虽然未示出,但是过程3400可以由管线管理器3212、应用编排系统3228和/或并行计算平台3230来促进。

[0414] 在至少一个实施例中,过程3400可包括从超声设备3402接收成像 数据。在至少一个实施例中,成像数据可以DICOM格式(或其他格式, 诸如,RIS、CIS、REST兼容、RPC、原始等)存储在PACS服务器上, 并且可由系统3200接收用于通过部署管线3210进行处理,该部署管线被 选择或定制为用于超声设备3402的虚拟仪器(例如,虚拟超声)。在至少 一个实施例中,成像数据可直接从成像装置(例如,超声设备3402)接收 并由虚拟仪器处理。在至少一个实施例中,通信地耦合在成像设备与虚拟 仪器之间的换能器或其他信号转换器可以将由成像设备生成的信号数据转 换成可以由虚拟仪器处理的图像数据。在至少一个实施例中,原始数据和/ 或图像数据可以应用于DICOM读取器3306以提取数据以供部署管线3210B的应用程序或容器使用。在至少一个实施例中,DICOM读取器3306 可利用数据扩充库3414(例如,NVIDIA的DALI)作为服务3120(例如, 作为计算服务3216中的一个)来提取、调整大小、重新缩放和/或以其他 方式准备数据以供应用或容器使用。

[0415] 在至少一个实施例中,一旦准备好数据,就可执行重构3406应用程 序和/或容器以将来自超声设备3402的数据重构为图像文件。在至少一个 实施例中,在重构3406之后或与重构3406同时,可执行检测3408应用程 序和/或容器以用于异常检测、对象检测、特征检测和/或与数据相关的其他 检测任务。在至少一个实施例中,重构3406期间生成的图像文件可在检测3408期间用于标识异常、对象、特征等。在至少一个实施例中,检测3408 应用可以利用推理引擎3416(例如,作为AI服务3218中的一个)来对数 据执行推理以生成检测。在至少一个实施例中,检测3408应用可以执行或 调用一个或更多个机器学习模型(例如,来自训练系统3104)。

[0416] 在至少一个实施例中,一旦重构3406和/或检测3408完成,就可以 使用来自这些应用和/或容器的数据输出来生成可视化3410,诸如显示在工 作站或显示终端上的可视化3412(例如,灰度输出)。在至少一个实施例 中,可视化可允许技术人员或其他用户相对于超声设备3402可视化部署管 线3210B的结果。在至少一个实施例中,可以通过利用系统3200的渲染 组件3418(例如,一个或更多个可视化服务3220中的一个)来执行可视 化3410。在至少一个实施例中,渲染组件3418可执行2D、OpenGL或光 线追踪服务以生成可视化3412。

[0417] 图34B包括根据至少一个实施例的支持CT扫描仪的虚拟仪器的示例 数据流程图。

在至少一个实施例中，部署管线3210C可利用系统3200的 服务3120中的一个或更多个。在至少一个实施例中，部署管线3210C和 服务3120可以在本地或在云3226中利用系统的硬件3122。在至少一个实 施例中，虽然未示出，但是可以通过管线管理器3212、应用编排系统3228 和/或并行计算平台3230来促进过程3420。

[0418] 在至少一个实施例中，过程3420可包括CT扫描仪3422生成可由 DICOM读取器3306(例如，直接，经由PACS服务器3304，在处理之后 等)接收的原始数据。在至少一个实施例中，虚拟CT(由部署管线3210C 示例化)可包括用于监测患者(例如，患者移动检测AI3426)和/或用于 调整或优化CT扫描仪3422的曝光(例如，使用曝光控制AI3424)的第 一实时管线。在至少一个实施例中，一个或更多个应用(例如，3424和3426) 可以利用服务3120，诸如AI服务3218。在至少一个实施例中，暴光控制 AI3424应用(或容器)和/或患者移动检测AI3426应用(或容器)的输出 可用作对CT扫描仪3422和/或技师的反馈，以调整暴露(或CT扫描仪 3422的其他设置)和/或通知患者移动较少。

[0419] 在至少一个实施例中，部署管线3210C可包括用于分析由CT扫描仪 3422生成的数据的非实时管线。在至少一个实施例中，第二管线可包括CT重构3308应用和/或容器，粗略检测AI 3428应用和/或容器、精细检测 AI 3432应用和/或容器(例如，其中某些结果由粗略检测AI 3428检测)， 可视化3430应用和/或容器，以及DICOM写入器3312(和/或其他数据类 型写入器，例如RIS、CIS、REST兼容、RPC、原始等)应用和/或容器。 在至少一个实施例中，由CT扫描仪3422生成的原始数据可通过部署管线 3210C的管线(被示例化为虚拟CT仪器)被传递以生成结果。在至少一 个实施例中，来自DICOM写入器3312的结果可被传输用于显示和/或可 存储在PACS服务器3304上以用于由技术人员、从业者或其他用户以后检 索、分析或显示。

[0420] 图35A示出了根据至少一个实施例的用于训练、重新训练或更新机 器学习模型的过程3500的数据流程图。在至少一个实施例中，可以使用图 32的系统3200作为非限制性示例来执行过程3500。在至少一个实施例中， 过程3500可以利用系统3200的服务3120和/或硬件3122，如在此所描述 的。在至少一个实施例中，由过程3500生成的精炼模型3512可以由部署 系统3106针对部署管线3210中的一个或更多个容器化应用来执行。

[0421] 在至少一个实施例中，模型训练3114可以包括使用新的训练数据(例 如，新的输入数据，诸如顾客数据集3506和/或与输入数据相关联的新的 地面实况数据)来重新训练或更新初始模型3504(例如，预训练模型)。 在至少一个实施例中，为了重新训练或更新初始模型3504，初始模型3504 的输出或损失层可被重置、或删除、和/或用更新的或新的输出或损失层替 换。在至少一个实施例中，初始模型3504可以具有先前微调的参数(例如， 权重和/或偏置)，这些参数保持来自先前训练，因此训练或重新训练3114 可以不花费与从头训练模型一样长的时间或需要尽可能多的处理。在至少 一个实施例中，在模型训练3114期间，通过重置或替换初始模型3504的 输出或损失层，在生成关于新的预测时，可基于与输出或一个或更多个损 失层的准确度相关联的损失计算来为新的数据集更新和重新调谐参数，顾 客数据集3506(例如，图31的成像数据3108)。

[0422] 在至少一个实施例中，预训练模型3206可被存储在数据存储或注册 表(例如，图31的模型注册表3124)中。在至少一个实施例中，预训练 模型3206可以已经在除了执行过程3500的设施之外的一个或更多个设施 处至少部分地被训练。在至少一个实施例中，为了

保护不同设施的患者、受试者或客户端的隐私和权限，可能已经使用场所内生成的客户或患者数 据在场所内训练了预训练模型3206。在至少一个实施例中，可以使用云 3226和/或其他硬件3122来训练预训练模型3206，但是机密的、隐私保护 的患者数据可以不被传输到云3226的任何组件(或其他房屋外硬件)、不 被云3226的任何组件使用或不被云3226的任何组件访问。在至少一个实 施例中，在使用来自一个以上设施的患者数据来训练预训练模型3206时， 可以在对来自另一个设施的患者或客户数据进行训练之前针对每个设施单 独地训练预训练模型3206。在至少一个实施例中，诸如其中顾客或患者数 据已被发布隐私问题(例如，通过等待、供实验使用等)，或者在客户或患 者数据包括在公共数据集中的情况下，来自任何数量的设施的客户或患者 数据可被用来训练场所内和/或场所外的预训练模型3206，诸如在数据中心 或其他云计算基础设施中。

[0423] 在至少一个实施例中，当选择在部署管线3210中使用的应用时，用 户还可以选择将用于特定应用的机器学习模型。在至少一个实施例中，用 户可以不具有供使用的模型，因此用户可以选择预先训练的模型3206来与 应用一起使用。在至少一个实施例中，预先训练的模型3206可以不被优化 用于在用户的设施的顾客数据集3506上生成准确的结果(例如，基于患者 多样性、人口统计、所使用的医学成像设备的类型等)。在至少一个实施例中，在将预训练模型3206部署到部署管线3210中以供与应用一起使用之 前，预训练模型3206可以被更新、重新训练和/或微调以供在相应设施处 使用。

[0424] 在至少一个实施例中，用户可以选择要被更新、重新训练和/或精细 调谐的预训练模型3206，并且预训练模型3206可以被称为用于过程3500 内的训练系统3104的初始模型3504。在至少一个实施例中，顾客数据集 3506(例如，成像数据、基因组学数据、测序数据或由设施处的设备生成 的其他数据类型)可以用于对初始模型3504执行模型训练3114(其可以 包括但不限于转移学习)以生成经细化的模型3512。在至少一个实施例中， 对应于顾客数据集3506的地面实况数据可由训练系统3104生成。在至少 一个实施例中，地面真实数据可以至少部分地由设施处的临床医生、科学 家、医生、执业医生生成(例如，如图31的标记的临床数据3112)。

[0425] 在至少一个实施例中，AI辅助注释3110可以在一些示例中用于生成 地面实况数据。在至少一个实施例中，AI辅助注释3110(例如，使用AI 辅助注释SDK来实现)可以利用机器学习模型(例如，神经网络)来生成 针对顾客数据集的建议的或预测的地面实况数据。在至少一个实施例中， 用户3510可使用计算设备3508上的用户界面(图形用户界面(GUI))内的注释工具。

[0426] 在至少一个实施例中，用户3510可经由计算设备3508与GUI交互 以编辑或微调(自动)注释。在至少一个实施例中，多边形编辑特征可用 于将多边形的顶点移动到更准确或精细调整的位置。

[0427] 在至少一个实施例中，一旦顾客数据集3506具有相关联的地面实况 数据，地面实况数据(例如，来自AI辅助注释、手动标记等)就可在模型 训练3114期间被用于生成经细化的模型3512。在至少一个实施例中，顾 客数据集3506可以被应用到初始模型3504任意次数，并且地面实况数据 可以用于更新初始模型3504的参数，直到针对精化模型3512获得可接受 的准确度水平。在至少一个实施例中，一旦生成精化模型3512，精化模型 3512可以部署在用于执行关于医学成像数据的一个或更多个处理任务的 设施处的一个或更多个部署

管线3210内。

[0428] 在至少一个实施例中，精炼模型3512可以被上传到模型注册表3124 中的预训练模型3206以由另一设施选择。在至少一个实施例中，他的过程 可以在任何数量的设施处完成，使得经细化的模型3512可以在新的数据集 上被进一步细化任何次数以生成更通用的模型。

[0429] 图35B是根据至少一个实施例的用于利用预先训练的注释模型来增 强注释工具的客户端-服务器架构3532的示例图示。在至少一个实施例中， AI辅助注释工具3536可以基于客户端-服务器架构3532被示例化。在至 少一个实施例中，成像应用中的注释工具3536可以帮助放射科医生例如识 别器官和异常。在至少一个实施例中，成像应用可以包括软件工具，作为 非限制性示例，该软件工具帮助用户3510识别原始图像3534中(例如， 在3D MRI或CT扫描中)感兴趣的特定器官上的几个极端点并且接收特 定器官的所有2D切片的自动注释的结果。在至少一个实施例中，结果可 以作为训练数据3538存储在数据存储中并且用作(例如但不限于)用于训 练的地面实况数据。在至少一个实施例中，当计算设备3508发送AI辅助 注释3110的极端点时，深度学习模型例如可以接收该数据作为输入并且返 回分段器官或异常的推理结果。在至少一个实施例中，预示例化的注释工 具(诸如图35B中的AI辅助注释工具3536B)可通过向服务器(诸如可 包括例如存储在注释模型注册表中的预训练模型3542的集合的注释助理 服务器3540)进行API调用(例如，API调用3544)来增强。在至少一个 实施例中，注释模型注册表可以存储预训练模型3542(例如，机器学习模型，诸如深度学习模型)，其被预先训练以对特定器官或异常执行AI辅助 注释。这些模型可通过使用训练管线3204进一步更新。在至少一个实施例 中，当添加新的经标记的诊所数据3112时，可随时间改进预安装的注释工 具。

[0430] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中，此逻辑可与这些图的组件一起使 用，以确定数据异常和适当的响应动作。

[0431] 自主车辆

[0432] 图36A示出了根据至少一个实施例的自主车辆3600的示例。在至少 一个实施例中，自主车辆3600(可替代地在此被称为“车辆3600”)可以 是但不限于客运车辆，如轿车、卡车、公共汽车和/或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1a00可以是用于拖运 货物的半拖拉机-拖车卡车。在至少一个实施例中，交通工具1a00可以是 飞机、机器人交通工具或其他类型的交通工具。

[0433] 可以就自动化水平来描述自主车辆，该自动化水平由国家公路交通 安全管理局(“NHTSA”)、美国运输部的分部以及汽车工程师协会(“SAE”) “与用于公路机动车辆驾驶自动化系统相关的术语的分类和定义”(例如， 标准号J3016-201806，公开于2018年6月15日，标准号J3016-201609， 公开于2016年9月30日，以及此标准的先前和未来版本)。在一个或更多 个实施例中，车辆3600可以能够具有根据自主驾驶水平的水平1-水平5 中的一个或更多个的功能。例如，在至少一个实施例中，取决于实施例， 运载工具3600可以能够进行有条件自动化(3级)、高自动化(4级)和/ 或全自动化(5级)。

[0434] 在至少一个实施例中，车辆3600可包括但不限于诸如底盘、车身、 车轮(例如，2、4、6、8、18等)、轮胎、车轴和车辆的其他部件的部件。 在至少一个实施例中，车辆3600可包

括但不限于推进系统3650,诸如内 燃机、混合动力电站、全电力发动机和/或另一推进系统类型。在至少一个 实施例中,推进系统3650可连接至车辆3600的传动系,该传动系可包括但不限于变速器,以使车辆3600能够推进。在至少一个实施例中,推进系 统3650可响应于从节气门/加速器3652接收信号进行控制。

[0435] 在至少一个实施例中,可包括但不限于方向盘的转向系统3654用于 当推进系统3650操作时(例如,当车辆在运动中时)转向车辆3600(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3654可接 收来自转向致动器3656的信号。方向盘对于全自动化(5级)功能可以是 可选的。在至少一个实施例中,制动传感器系统3646可以用于响应于从制 动致动器3648和/或制动传感器接收信号来操作车辆制动器。

[0436] 在至少一个实施例中,控制器3636(其可以包括但不限于一个或更 多个片上系统(“SoC”)(图36A中未示出)和/或图形处理单元(“GPU”)) 向车辆3600的一个或更多个部件和/或系统提供信号(例如,表示命令)。 例如,在至少一个实施例中,控制器3636可以发送信号以经由制动致动器 3648操作车辆制动、经由转向致动器3656操作转向系统3654、和/或经由 节气门/加速器3652操作推进系统3650。控制器3636可包括一个或更多个 车载(例如,集成)计算装置(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)以使能自主驾驶和/或辅助人类 驾驶员驾驶车辆3600。在至少一个实施例中,一个或更多个控制器3636 可以包括用于自主驾驶功能的第一控制器3636、用于功能安全功能的第二 控制器3636、用于人工智能功能(例如,计算机视觉)的第三控制器3636、 用于信息娱乐功能的第四控制器3636、用于在紧急状况下冗余的第五控制 器3636和/或其他控制器。在至少一个实施例中,单个控制器3636可以处 理上述功能中的两个或更多个,两个或更多个控制器3636可以处理单个功 能,和/或其任何组合。

[0437] 在至少一个实施例中,控制器3636响应于从一个或更多个传感器接 收的传感器数据(例如,传感器输入)提供用于控制车辆3600的一个或更 多个部件和/或系统的信号。在至少一个实施例中,可从例如但不限于以下 装置接收传感器数据,一个或更多个全球导航卫星系统(“GNSS”)传感器 3658(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3660、一个或更多个超声传感器3662、一个或更多个LIDAR传感 器3664、一个或更多个惯性测量单元(“IMU”)传感器3666(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3696、一个或更多个立体相机3668、 一个或更多个宽视角相机3670(例如,鱼眼相机)、一个或更多个红外相 机3672、一个或更多个环绕相机3674(例如,360度相机)、远距离相机 (图36A中未示出)、中距照相机(图36A中未示出)、速度传感器3644 (例如,用于测量车辆3600的速度)、振动传感器3642、转向传感器3640、 制动传感器(例如,作为制动传感器系统3646的部分)和/或其他传感器 类型。

[0438] 在至少一个实施例中,一个或更多个控制器3636中的一个或更多个 可以接收来自车辆3600的仪表板3632的输入(例如,由输入数据表示) 并且经由人机界面(“HMI”)显示器3634提供输出(例如,由输出数据、显示数据等表示),听觉信号器、扬声器、和/或通过车辆3600的其他部件。 在至少一个实施例中,输出可包括诸如车辆速度、速度、时间的信息,地 图数据(例如,高清晰度地图(图36A中未示出))、位置数据(例如,车 辆3600的位置,如在地图上)、方向、其他车辆的位置(例如,占用网格)、 关于如由控制器3636所感知的对象

和对象的状态的信息等。例如,在至少 一个实施例中,HMI显示器3634可以显示关于一个或更多个对象(例如, 街道标志、注意标志、交通灯改变等)的存在的信息,和/或关于车辆已经 做出、正在做出或将要做出的驾驶操纵(例如,现在改变车道、在两英里 内驶出口34B等)的信息。

[0439] 在至少一个实施例中,车辆3600进一步包括网络接口3624,网络接 口3624可使用无线天线3626和/或调制解调器通过一个或更多个网络通信。例如,在至少一个实施例中,网络接口3624可以能够通过长期演进(“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”)等进行通信。在至少一个实施例中,一个或更多个无线天线3626还可以使用一个或更多 个局域网(如蓝牙、蓝牙低功耗(“LE”)、Z-波、ZigBee等)和/或一个或 更多个低功率广域网(“LPWAN”),如LoRaWAN、SigFox等来使能环境 中的对象(例如,车辆、移动设备等)之间的通信。

[0440] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中,推理和/或训练逻辑615可以在系统图 36A中用于至少部分地基于使用在此描述的神经网络训练操作、神经网络 功能和/或架构或神经网络使用情况计算的权重参数来推理或预测操作。

[0441] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中,此逻辑可与这些图的组件一起使 用,以确定数据异常和适当的响应动作。

[0442] 图36B示出了根据至少一个实施例的用于图36A的自主车辆3600 的相机位置和视野的示例。在至少一个实施例中,相机和相应的视场是一 个举例实施例,并且不旨在是限制性的。例如,在至少一个实施例中,可 以包括附加的和/或替代的摄像机和/或摄像机可以位于车辆3600上的不同 位置。

[0443] 在至少一个实施例中,相机的相机类型可包括但不限于可适于与车 辆3600的组件和/或系统一起使用的数字相机。在至少一个实施例中,一 个或更多个相机可以在汽车安全完整性水平(“ASIL”)B和/或在另一个 ASIL下操作。在至少一个实施例中,取决于实施例,相机类型可以具有任 何图像捕获速率,诸如每秒60帧(fps)、120fps、240fps等。在至少一个 实施例中,相机可以能够使用卷帘式快门、全局快门、另一类型的快门或 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,一个红色澄清透明蓝色(“RCCB”)滤色器阵列、一个红色蓝色绿色澄清(“RBGC”)滤色器阵列,Foveon X3滤色器阵列、拜耳传感器(“RGGB”)滤色器阵列、单色传感器滤色器阵列,和/或另一 种类型的滤色器阵列。在至少一个实施例中,可以使用透明像素相机(如 具有RCCC、RCCB和/或RBGC滤色器阵列的相机)来努力增加光敏度。

[0444] 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的驾驶 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,可以安装多功能单摄像头以提供包括车道偏 离警告、交通标志辅助和智能前照灯控制的功能。在至少一个实施例中, 一个或更多个相机中的一个或更多个(例如,所有相机)可以同时记录和 提供图像数据(例如,视频)。

[0445] 在至少一个实施例中,相机中的一个或更多个相机可以安装在安装 组件中,如定

制设计的(三维(“3D”)打印的)组件中,以便削减可能干 扰相机的图像数据捕获能力的来自车内的杂散光和反射(例如,来自仪表 板的反射在挡风玻璃镜子中反射)。参考翼镜安装组件,在至少一个实施例 中,翼镜组件可以被定制3D打印,使得相机安装板匹配翼镜的形状。在 至少一个实施例中,一个或更多个相机可以被集成到翼镜中。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集成在驾驶室的每个拐 角处的四个柱内。

[0446] 在至少一个实施例中,具有包括车辆3600前方环境的部分的视野的 摄像机(例如,前置摄像机)可用于环绕视野,在控制器3636和/或控制 SoC中的一个或更多个的帮助下,帮助标识面向前的路径和障碍物,以及 帮助,提供对生成占用网格和/或确定优选车辆路径至关重要的信息。在至 少一个实施例中,前置摄像机可以用于执行与LIDAR相同的ADAS功能 中的许多功能,包括但不限于紧急制动、行人检测和碰撞避免。在至少一 个实施例中,前置相机还可以用于ADAS功能和系统,包括但不限于车道 偏离警告(“LDW”)、自主巡航控制(“ACC”)、和/或其他功能(如交通 标志识别)。

[0447] 在至少一个实施例中,多种相机可以用于正面配置,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,宽视角相机3670可用于感知从周边进入视野的对象(例如, 行人、交叉交通或自行车)。尽管图36B中仅示出一个宽视角照相机3670, 但在其他实施例中,车辆3600上可存在任意数量(包括零)的宽视角照相 机3670。在至少一个实施例中,任何数量的一个或更多个长距相机3698 (例如,长视图立体相机对)可被用于基于深度的对象检测,尤其对于神 经网络尚未被训练的对象。在至少一个实施例中,一个或更多个长距相机 3698还可用于对象检测和分类,以及基本对象追踪。

[0448] 在至少一个实施例中,在前置配置中还可以包括任何数量的立体相 机3668。在至少一个实施例中,一个或更多个立体相机3668可以包括集 成控制单元,该集成控制单元包括可扩展处理单元,该可扩展处理单元可 以提供可编程逻辑(“FPGA”)和在单芯片上具有集成控制器局域网 (“CAN”)或以太网接口的多核微处理器。在至少一个实施例中,这种单元可用于生成车辆3600的环境的3D图,包括图像中的所有点的距离估计。 在至少一个实施例中,一个或更多个立体相机3668中的一个或更多个可包 括,非限制性地,一个或更多个紧凑立体视觉传感器可以包括但不限于两 个摄像机镜头(左和右各一个)和图像处理芯片,所述图像处理芯片可以 测量从车辆3600到目标对象的距离并且使用生成的信息(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 文描述的立体相机之外或替代本文描述的立体相机,可以使用其他类型的 立体相机3668。

[0449] 在至少一个实施例中,具有包括车辆3600侧面环境的部分的视野的 摄像机(例如,侧视摄像机)可用于周围视野、提供用于创建和更新占用 网格的信息以及生成侧面碰撞碰撞警告。例如,在至少一个实施例中,一 个或更多个环绕相机3674(例如,如图36B所示的四个环绕相机3674) 可以定位在车辆3600上。在至少一个实施例中,一个或更多个环绕相机 3674可以包括但不限于一个或更多个宽视角相机3670、一个或更多个鱼眼 相机、一个或更多个360度相机和/或类似物的任何数量和组合。例如,在 至少一个实施例中,四个鱼眼摄像头可以位于车辆3600的前面、后面和侧 面。在至少一个实施例中,车辆3600可使用三个环绕相机3674(例如, 左、右和后),并且可将一个或更多个其他摄像机(例如,朝前的摄像机) 用作第四环绕视角摄像机。

[0450] 在至少一个实施例中，具有包括车辆3600后方环境的部分的视野的 摄像机(例如，后视摄像机)可用于停车辅助、环绕视野、后方碰撞警告 以及创建和更新占用网格。在至少一个实施例中，可以使用各种各样的相 机，包括但不限于也适合作为一个或更多个前置相机的相机(例如，一个 或更多个长距相机3698和/或一个或更多个中距相机3676、一个或更多个 立体相机3668、一个或更多个红外相机3672等)，如在此所描述的。

[0451] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中，推理和/或训练逻辑615可以在图36B 的系统中用于至少部分地基于使用在此描述的神经网络训练操作、神经网 络功能和/或架构或神经网络使用情况计算的权重参数来推理或预测操作。

[0452] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中，此逻辑可与这些图的组件一起使 用，以确定数据异常和适当的响应动作。

[0453] 图36C是示出根据至少一个实施例的用于图36A的自主车辆3600 的示例系统架构的框图。在至少一个实施例中，图36C中的车辆3600的 每个部件、特征和系统被示出为经由总线3602连接。在至少一个实施例中， 总线3602可以包括但不限于CAN数据接口(可替代地在此被称为“CAN 总线”)。在至少一个实施例中，CAN总线可以是车辆3600内部的网络， 用于帮助控制车辆3600的各种特征和功能，诸如制动器的致动、加速、制 动、转向、风挡刮水器等。在至少一个实施例中，总线3602可被配置为具 有几十或甚至几百个节点，每个节点具有其自身的唯一标识符(例如，CAN ID)。在至少一个实施例中，可以读取总线3602以便找到方向盘角度、地 面速度、发动机每分钟转数(“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3602可以是ASILB兼容的CAN总线。

[0454] 在至少一个实施例中，作为CAN的补充或替代，可以使用FlexRay 和/或以太网。在至少一个实施例中，可以存在任何数量的总线3602，这些 总线可以包括但不限于零个或多个CAN总线、零个或多个FlexRay总线、 零个或多个以太网总线、和/或零个或多个使用不同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3602可用于执行不同的功能，和/或可用于冗余。例如，第一总线3602可用于碰撞避免功能，而第二总线 3602可用于致动控制。在至少一个实施例中，每个总线3602可与车辆3600 的任何部件通信，并且两个或多个总线3602可与相同的部件通信。在至少 一个实施例中，任何数目的片上系统(“SoC”)3604中的每个、控制器3636 中的每个、和/或车辆内的每个计算机可以访问相同的输入数据(例如，来 自车辆3600的传感器的输入)，并且可以连接至公共总线，如CAN总线。

[0455] 在至少一个实施例中，车辆3600可包括一个或更多个控制器3636， 诸如本文中参照图36A所描述的那些。控制器3636可以用于各种功能。 在至少一个实施例中，一个或更多个控制器3636可以耦合至车辆3600的 不同其他部件和系统中的任何部件和系统，并且可以用于控制车辆3600、 车辆3600的人工智能、车辆3600的信息娱乐等。

[0456] 在至少一个实施例中，车辆3600可包括任何数量的SoC3604。SoC 3604中的每个SoC可以包括但不限于中央处理单元(“CPU”)3606、图形 处理单元(“GPU”)3608、一个或更多个处理器3610、一个或更多个高速 缓存3612、一个或更多个加速器3614、一个或更多个数据存储3616、和/ 或未示出的其他部件和特征。在至少一个实施例中，一个或更多个

SoC3604 可以用于控制各种平台和系统中的车辆3600。例如,在至少一个实施例中, 一个或更多个SoC 3604可以在具有高清晰度(“HD”)地图3622的系统(例 如,车辆3600的系统)中组合,所述高清晰度(“HD”)地图可以经由网 络接口3624从一个或更多个服务器(图36C中未示出)获得地图刷新和/ 或更新。

[0457] 在至少一个实施例中,一个或更多个CPU3606可以包括CPU集群或 CPU复合体(可替代地在本文中被称为“CCPLEX”)。在至少一个实施例 中,一个或更多个CPU 3606可以包括多个核和/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CPU3606可以在一致多处理器 配置中包括八个核。在至少一个实施例中,一个或更多个CPU3606可包括 四个双核集群,其中每个集群具有专用L2高速缓存(例如,2MBL2高速 缓存)。在至少一个实施例中,一个或更多个CPU3606(例如,CCPLEX) 可以被配置为支持同时的群集操作,从而使得一个或更多个CPU3606的群 集的任何组合能够在任何给定时间活动。

[0458] 在至少一个实施例中,一个或更多个CPU3606中的一个或更多个可 实现功率管理能力,包括但不限于以下特征中的一个或更多个:各个硬件 块可在空闲时被自动时钟门控以节省动态功率;当核由于等待中断(“WFI”) /等待事件(“WFE”)指令的执行而未主动执行指令时,可门控每一核时钟; 每个核可以被独立地功率门控;当所有核被时钟门控或功率门控时,每个 核簇可以被独立地时钟门控;和/或每个核心群集可在所有核心都被功率门控时独立地被功率门控。在至少一个实施例中,一个或更多个CPU3606 还可以实现用于管理功率状态的增强算法,其中规定了允许的功率状态和 期望的唤醒时间,并且硬件/微代码确定对于核、群集和CCPLEX要进入 的最佳功率状态。在至少一个实施例中,处理单元可以在工作卸载到微代 码的软件中支持简化的功率状态进入序列。

[0459] 在至少一个实施例中,一个或更多个GPU3608可以包括集成GPU(可 替代地在此被称为“iGPU”)。在至少一个实施例中,一个或更多个GPU3608 可以是可编程的并且对于并行工作负荷可以是高效的。在至少一个实施例 中,一个或更多个GPU3608在至少一个实施例中可以使用增强的张量指令 集。在至少一个实施例中,一个或更多个GPU3608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存储容量的L1高速缓存),并且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GPU3608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GPU3608可以使用一个 或更多个计算应用编程接口(API)。在至少一个实施例中,一个或更多个GPU3608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)。

[0460] 在至少一个实施例中,GPU3608中的一个或更多个可经电力优化以 用于汽车和嵌入式使用情况中的最佳性能。例如,在一个实施例中,可以 在鳍式场效应晶体管(“FinFET”)上制造GPU 3608。在至少一个实施例 中,每个流式微处理器可以合并被分成多个块的多个混合精度处理单元。 例如,但不限于,64个PF32核和32个PF64核可被划分成四个处理块。 在至少一个实施例中,每个处理块可以被分配16FP32核、8FP64核、16 INT32核、用于深度学习矩阵运算的两个混合精度NVIDIA TENSOR CORE、零级(“L0”)指令高速缓存、经线调度器、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,以便利用计算和寻址计算的混合来提供工作负荷的有效执行。 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 现并

行线程之间的更细粒度的同步和协作。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 编程的同时改善性能。

[0461] 在至少一个实施例中,一个或更多个GPU3608中的一个或更多个可 以包括高带宽存储器(“HBM”)和/或16GB的HBM2存储器子系统,以 便在一些示例中提供约900GB/秒的峰值存储器带宽。在至少一个实施例中, 除了HBM存储器之外或替代HBM存储器,可以使用同步图形随机存取存 储器(“SGRAM”),如图形双倍数据速率类型五同步随机存取存储器(“GDDR5”)。

[0462] 在至少一个实施例中,一个或更多个GPU3608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 GPU 3608直接访问CPU 3606页表。在至少一个实施例中,当GPU 3608 存储器管理单元(“MMU”)经历未命中时,地址转换请求可以被传输到 CPU 3606。作为响应,在至少一个实施例中,CPU3606可在其页表中查找地址的虚拟到物理映射,且将翻译传送回到GPU3608。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 CPU3606和一个或更多个GPU3608两者的存储器,从而简化一个或更多 个GPU3608编程和将应用移植到一个或更多个GPU3608。

[0463] 在至少一个实施例中,GPU3608可包含可追踪GPU3608对其他处理 器的存储器的存取频率的任何数目个存取计数器。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存储器页面被移动到最频繁地访问 页面的处理器的物理存储器,由此提高在处理器之间共享的存储器范围的 效率。

[0464] 在至少一个实施例中,一个或更多个SoC3604中的一个或更多个SoC 可以包括任何数量的一个或更多个高速缓存3612,包括在此描述的那些。例如,在至少一个实施例中,一个或更多个高速缓存3612可以包括一个或 更多个CPU3606和一个或更多个GPU3608均可用的(例如,连接一个或 更多个CPU3606和一个或更多个GPU3608两者的)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3612可以包括可以例如 通过使用高速缓存一致性协议(例如,MEI、MESI、MSI等)来追踪行的 状态的回写高速缓存。在至少一个实施例中,L3高速缓存可以包括4MB 或更多,这取决于实施例,尽管可以使用较小的高速缓存大小。

[0465] 在至少一个实施例中,一个或更多个SoC3604中的一个或更多个可 以包括一个或更多个加速器3614(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SoC3604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)可以使硬件加速簇能够加速 神经网络和其他计算。在至少一个实施例中,硬件加速群集可以用于补充 GPU3608和卸载GPU3608的一些任务(例如,释放GPU3608的更多周期 用于执行其他任务)。在至少一个实施例中,一个或更多个加速器3614可 以用于足够稳定以易于进行加速的目标工作负荷(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 可以包括基于区域的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。

[0466] 在至少一个实施例中,一个或更多个加速器3614(例如,硬件加速 簇)可以包括一个或更多个深度学习加速器(“DLA”)。DLA可以包括但 不限于一个或更多个张量器处理单元(“TPU”),该一个或更多个张量器处 理单元可以被配置成用于深度学习应用和推理提供

额外的每秒十万亿次操 作。在至少一个实施例中，一个或更多个TPU可以是配置成执行图像处理 功能(例如，针对CNN、RCNN等)和被优化用于执行图像处理功能(例 如，针对CNN、RCNN等)的加速器。DLA可进一步针对神经网络类型 和浮点运算的特定集合以及推理来优化。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每毫米更多的性能，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可以 执行若干功能，包括单示例卷积功能，支持例如针对特征和权重两者的INT8、INT16和FP16数据类型以及后处理器功能。在至少一个实施例中， 一个或更多个DLA可以对用于多种功能中的任何功能的经处理或未经处 理的数据快速且高效地执行神经网络，尤其是CNN，包括例如但不限于： CNN，用于使用来自相机传感器的数据进行对象标识和检测；CNN，用于 使用来自照相机传感器的数据进行距离估计；CNN，用于使用来自麦克风 3696的数据进行急救车辆检测和识别和检测；CNN，用于使用来自相机传 感器的数据进行面部识别和车辆所有者识别；和/或CNN，用于安全和/或 安全相关事件。

[0467] 在至少一个实施例中，一个或更多个DLA可以执行一个或更多个 GPU3608的任何功能，并且通过使用例如推理加速器，设计者可以将一个 或更多个DLA或一个或更多个GPU3608作为目标用于任何功能。例如， 在至少一个实施例中，设计者可将对CNN的处理和浮点操作集中在DLA 上，并将其他功能留给GPU3608和/或其他加速器3614。

[0468] 在至少一个实施例中，一个或更多个加速器3614(例如，硬件加速 簇)可以包括一个或更多个可编程视觉加速器(“PVA”)，所述可编程视觉 加速器在此可以可替代地被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可以被设计和配置成用于加速用于高级驾驶员辅助系 统(“ADAS”)3638、自主驾驶、增强现实(“AR”)应用、和/或虚拟现实 (“VR”)应用的计算机视觉算法。PVA可提供性能和柔性之间的平衡。例 如，在至少一个实施例中，每个PVA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、直接存储器访问(“DMA”)、和/或任何数 量的向量处理器。

[0469] 在至少一个实施例中，RISC核可以与图像传感器(例如，在此描述 的任何相机的图像传感器)、图像信号处理器等交互。在至少一个实施例中， RISC核中的每一个可以包括任何量的存储器。在至少一个实施例中，取决 于实施例，RISC核可以使用多个协议中的任何一个。在至少一个实施例中， RISC核可以执行实时操作系统(“RTOS”)。在至少一个实施例中，RISC 核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 存储器设备来实现。例如，在至少一个实施例中，RISC核可以包括指令高 速缓存和/或紧密耦合的RAM。

[0470] 在至少一个实施例中，DMA可以使得PVA的组件能够独立于 CPU3606来访问系统存储器。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可以支持高达六个或更多的寻址维 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。

[0471] 在至少一个实施例中，向量处理器可以是可编程处理器，所述可编 程处理器可以被设计成高效地和灵活地执行针对计算机视觉算法的编程并 且提供信号处理能力。在至少一个实施例中，PVA可以包括PVA核和两个 向量处理子系统分区。在至少一个实施例中，PVA核可包括处理器子系统、一个或更多个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操作为PVA的主处理引擎,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可以包括数字信号处理器, 例如像单指令、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的组合可以增强吞吐量和速度。

[0472] 在至少一个实施例中,向量处理器中的每一个可以包括指令高速缓 存并且可以耦合到专用存储器。因此,在至少一个实施例中,向量处理器 中的每一个可以被配置成独立于其他向量处理器执行。在至少一个实施例 中,包括在特定PVA中的向量处理器可以被配置成采用数据并行性。例如, 在至少一个实施例中,包括在单个PVA中的多个矢量处理器可以执行相同 的计算机视觉算法,但是在图像的不同区域上。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同一图像上同时执行不同的计算机 视觉算法,或者甚至在顺序图像或图像的部分上执行不同的算法。在至少 一个实施例中,除其他之外,任何数量的PVA可被包括在硬件加速集群中, 并且任何数量的向量处理器可被包括在每个PVA中。在至少一个实施例中, 一个或更多个PVA可以包括附加纠错码(“ECC”)存储器,以增强整体系 统安全性。

[0473] 在至少一个实施例中,一个或更多个加速器3614(例如,硬件加速 集群)可以包括计算机视觉片上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3614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的SRAM,其由例如但不限 于八个场可配置存储器块组成,其可以由PVA和DLA两者访问。在至少 一个实施例中,每对存储器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可经由向PVA和DLA提 供对存储器的高速访问的主干来访问存储器。在至少一个实施例中,主干 可以包括将PVA和DLA互连到存储器(例如,使用APB)的计算机视觉 片上网络。

[0474] 在至少一个实施例中,计算机视觉片上网络可以包括在传输任何控 制信号/地址/数据之前确定PVA和DLA两者提供就绪信号和有效信号的接 口。在至少一个实施例中,接口可以提供用于传送控制信号/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,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,尽管可以使用其他标准和协议。

[0475] 在至少一个实施例中,一个或更多个SoC3604中的一个或更多个可 以包括实时光线追踪硬件加速器。在至少一个实施例中,实时光线追踪硬 件加速器可用于快速且高效地确定对象的位置和范围(例如,在世界模型 内),生成用于RADAR信号解释、用于声音传播合成和/或分析的实时可 视化模拟,用于SONAR系统的模拟、用于一般波传播模拟、用于出于定 位和/或其他功能的目的与LIDAR数据进行比较,和/或用于其他用途。

[0476] 在至少一个实施例中,一个或更多个加速器3614(例如,硬件加速 器群集)具有用于自主驾驶的广泛的使用阵列。在至少一个实施例中,PVA 可以是可编程视觉加速器,其可用于ADAS和自动交通工具中的关键处理 阶段。在至少一个实施例中,PVA的能力在低功率和低等待时间下对于需 要可预测处理的算法域是良好匹配的。换言之,PVA对半密集或密集的规 则计算表现良好,甚至对需要具有低等待时间和低功率的可预测运行时间 的小数据集合也表现良好。在至少一个实施例中,自主车辆(如车辆3600、PVA)被设计成用于运行

经典计算机视觉算法,因为它们在对象检测和以 整数数学运算时是有效的。

[0477] 例如,根据技术的至少一个实施例,PVA用于执行计算机立体视觉。在至少一个实施例中,在一些示例中,可使用基于半全局匹配的算法,但 这不意欲为限制性的。在至少一个实施例中,用于3-5级自主驾驶的应用 在运行中使用运动估计/立体匹配(例如,来自运动的结构、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。

[0478] 在至少一个实施例中,PVA可用于执行密集光流。例如,在至少一 个实施例中,PVA可处理原始RADAR数据(例如,使用4D快速傅里叶 变换)以提供经处理的RADAR数据。在至少一个实施例中,PVA用于飞 行时间深度处理,例如通过处理原始飞行时间数据以提供经处理的飞行时 间数据。

[0479] 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,包括例如但不限于输出每个对象检测的置信度测量的神经 网络。在至少一个实施例中,置信度可以表示或解释为概率,或提供每个 检测与其他检测相比的相对“权重”。在至少一个实施例中,置信度使得系 统能够做出关于哪些检测应当被认为是真肯定检测而不是假肯定检测的进 一步决定。例如,在至少一个实施例中,系统可以设置置信度的阈值,并且仅将超过阈值的检测视为真的肯定检测。在使用自动紧急制动(“AEB”) 系统的实施例中,误肯定检测将使车辆自动执行紧急制动,这显然是不期 望的。在至少一个实施例中,高度置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,诸如 边界框尺寸、(例如从另一子系统获得的)接地平面估计、来自与车辆3600 取向相关的IMU传感器3666的输出,距离、从神经网络和/或其他传感器 (例如,一个或更多个LIDAR传感器3664或一个或更多个RADAR传感 器3660)获得的对象的3D位置估计,除其他之外。

[0480] 在至少一个实施例中,一个或更多个SoC3604中的一个或更多个可 以包括一个或更多个数据存储3616(例如,存储器)。在至少一个实施例 中,一个或更多个数据存储3616可以是一个或更多个SoC3604的片上存 储器,其可以存储要在一个或更多个GPU3608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3616的容量可足够大以 存储神经网络的多个示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3616可以包括一个或更多个L2或L3高速缓存。以及.

[0481] 在至少一个实施例中,一个或更多个SoC3604中的一个或更多个SoC 可以包括任意数量的一个或更多个处理器3610(例如,嵌入式处理器)。 在至少一个实施例中,一个或更多个处理器3610可以包括引导和功率管理 处理器,其可以是用于处理引导功率和管理功能和相关安全实施的专用处 理器和子系统。在至少一个实施例中,引导和功率管理处理器可以是SoC 3604引导序列的一部分,并且可以提供运行时间功率管理服务。在至少一 个实施例中,启动电源和管理处理器可以提供时钟和电压编程、系统低功 率状态转变的辅助、SoC 3604热和温度传感器的管理、和/或SoC 3604功 率状态的管理。在至少一个实施例中,每个温度传感器可以被实现为其输 出频率与温度成比例的环形振荡器,并且SoC3604可以使用环形振荡器来 检测CPU3606、GPU3608和/或加速器3614的温度。在至少一个实施例中, 如果确定温度超过阈值,则启动和电力管理处理器可以进入温度故障例程 并将SoC

3604置于较低电力状态和/或将车辆3600置于排队到安全停止模 式(例如,将车辆3600带到安全停止)。

[0482] 在至少一个实施例中,一个或更多个处理器3610可以进一步包括可 以用作音频处理引擎的一组嵌入式处理器。在至少一个实施例中,音频处 理引擎可以是音频子系统,其使得能够在多个接口和广泛和灵活范围的音 频I/O接口上对多通道音频进行全硬件支持。在至少一个实施例中,音频 处理引擎是具有专用RAM的数字信号处理器的专用处理器核。

[0483] 在至少一个实施例中,一个或更多个处理器3610还可以包括常通处 理器引擎,其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使 用情况。在至少一个实施例中,常通处理器引擎可以包括但不限于处理器 核、紧密耦合的RAM、支持外围设备(例如,定时器和中断控制器)、不 同I/O控制器外围设备和路由逻辑。

[0484] 在至少一个实施例中,一个或更多个处理器3610还可以包括安全集 群引擎,其包括但不限于用于处理汽车应用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。在安全模式中,在至少一个实施例中,两个或更多 个核可在锁步模式中操作,并且充当具有比较逻辑的单个核以检测其操作 之间的任何差异。在至少一个实施例中,一个或更多个处理器3610可以进 一步包括实时相机引擎,所述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3610还可以包括高动态范围信号处理器,所述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,所述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。

[0485] 在至少一个实施例中,一个或更多个处理器3610可以包括视频图像 合成器,所述视频图像合成器可以是(例如,在微处理器上实现的)处理 块,所述处理块实现视频回放应用所需的视频后处理功能,以产生播放器 窗口的最终图像。在至少一个实施例中,视频图像合成器可对宽视角相机 3670、环绕相机3674和/或舱内监视相机传感器执行透镜失真校正。在至 少一个实施例中,一个或更多个机舱内监视相机传感器优选地由在一个或 更多个SoC3604的另一示例上运行的神经网络监视,其被配置成标识机舱 内事件并相应地响应。在至少一个实施例中,机舱内系统可以执行(但不 限于)嘴唇阅读以激活蜂窝服务并发起电话呼叫、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、或提供语音激活的 网络冲浪。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驾 驶员是可用的并且另外被禁用。

[0486] 在至少一个实施例中,视频图像合成器可包括用于空间和时间降噪 的增强的时间降噪。例如,在至少一个实施例中,在视频中发生运动的情 况下,降噪适当地对空间信息进行加权,减小由相邻帧提供的信息的权重。 在至少一个实施例中,在图像或图像的部分不包括运动的情况下,由视频 图像合成器执行的时间噪声降低可使用来自先前图像的信息来降低当前图 像中的噪声。

[0487] 在至少一个实施例中,视频图像合成器还可以被配置为对输入的立 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用于用户界面合成,并且不需要GPU3608来连续地渲 染新的表面。在至少一个实施例中,当

GPU3608通电并主动进行3D渲染 时,视频图像合成器可用于卸载GPU3608以改善性能和响应性。

[0488] 在至少一个实施例中,一个或更多个SoC 3604中的一个或更多个可 以进一步包括用于从相机接收视频和输入的移动行业处理器接口(“MIPI”) 相机串行接口、高速接口、和/或可以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SoC3604中的一个或更多个 还可以包括输入/输出控制器,一个或更多个输入/输出控制器可以由软件控 制并且可以用于接收未提交到特定角色的I/O信号。

[0489] 在至少一个实施例中,一个或更多个SoC 3604中的一个或更多个可 以进一步包括广泛范围的外围接口,以使得能够与外围设备、音频编码器/ 解码器(“编解码器”)、功率管理、和/或其他设备进行通信。SoC3604可 用于处理来自相机(例如,通过千兆位多媒体串行链路和以太网连接)的 数据,传感器(例如,可以通过以太网连接的LIDAR传感器3664、RADAR 传感器3660等),来自总线3602的数据(例如,车辆3600的速度、方向 盘位置等)、来自GNSS传感器3658的数据(例如,通过以太网或CAN 总线连接),等等。在至少一个实施例中,一个或更多个SoC3604中的一 个或更多个还可以包括专用高性能大容量存储控制器,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CPU3606从例程数据管理任务释放。

[0490] 在至少一个实施例中,一个或更多个SoC3604可以是具有跨越自动 化水平3-5的灵活架构的端到端平台,由此提供全面的功能安全架构,其 充分利用和有效地利用计算机视觉和ADAS技术以用于多样性和冗余性, 为灵活、可靠的驱动软件栈以及深度学习工具提供平台。在至少一个实施 例中,一个或更多个SoC3604可以比常规系统更快、更可靠、甚至更节能 和空间高效。例如,在至少一个实施例中,一个或更多个加速器3614当与 一个或更多个CPU 3606、一个或更多个GPU 3608、和一个或更多个数据 存储装置3616组合时可以提供用于3-5级自主车辆的快速、高效的平台。

[0491] 在至少一个实施例中,计算机视觉算法可在CPU上执行,所述CPU 可使用高级编程语言(例如C编程语言)配置以跨各种各样的视觉数据执 行各种各样的处理算法。然而,在至少一个实施例中,CPU经常不能满足 许多计算机视觉应用的性能要求,例如像与执行时间和功耗相关的那些。 在至少一个实施例中,许多CPU不能实时执行复杂的目标检测算法,其在 车载ADAS应用中和在实际3-5级自主车辆中使用。

[0492] 在此所描述的实施例允许同时和/或顺序地执行多个神经网络,并且 允许将结果组合在一起以实现3级至5级自主驾驶功能。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GPU3620)上执 行的CNN可以包括文本和词语识别,从而允许超级计算机读取和理解流 量标志,包括神经网络尚未被具体训练的标志。在至少一个实施例中,DLA可以进一步包括神经网络,该神经网络能够标识、解释和提供对符号的语 义理解,并且将该语义理解传递给在CPU复合体上运行的路径规划模块。

[0493] 在至少一个实施例中,对于级别3、4或5驱动,多个神经网络可以 同时运行。例如,在至少一个实施例中,由“警告:闪光指示结冰状况” 组成的警告标志连同电灯可以独立地或共同地被若干神经网络解释。在至 少一个实施例中,标志本身可以由第一部署的神经网络(例如,已经训练 的神经网络)标识为交通标志,并且文本“闪光灯指示结冰状况”可以由第二部署的神经网络解释,其通知交通工具的路径规划软件(优选在CPU 复合体上执行)何

时检测到闪光灯,存在结冰条件。在至少一个实施例中，闪光灯可通过在多个帧上操作第三部署的神经网络来识别,通知交通工具 的路径规划软件闪光灯的存在(或不存在)。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GPU3608上运行。

[0494] 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可使 用来自相机传感器的数据来识别授权驾驶员和/或车辆3600的所有者的存 在。在至少一个实施例中,常开传感器处理引擎可以用于在所有者接近驾 驶员门并打开灯时解锁车辆,并且在安全模式中用于在所有者离开车辆时 禁用车辆。以此方式,一个或更多个SoC3604提供防止盗窃和/或劫持的安 全性。

[0495] 在至少一个实施例中,用于急救车辆检测和识别的CNN可使用来自 麦克风3696的数据来检测和识别急救车辆警报器。在至少一个实施例中， 一个或更多个SoC3604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,在DLA上运行的CNN被训练 为识别急救车辆的相对关闭速度(例如,通过使用多普勒效应)。在至少一 个实施例中,CNN还可被训练为识别特定于车辆正在操作的局部区域的应 急车辆,如由GNSS传感器3658识别的。在至少一个实施例中,当在欧洲 操作时,CNN将寻求检测欧洲汽笛,并且当在美国时,CNN将寻求仅识 别北美汽笛。在至少一个实施例中,一旦检测到紧急车辆,控制程序就可 用于在超声传感器3662的帮助下执行紧急车辆安全例程、减慢车辆、向道 路侧拉、停车车辆、和/或空转车辆,直到紧急车辆经过。

[0496] 在至少一个实施例中,车辆3600可包括可经由高速互连(例如,PCIe) 耦合到SoC3604的CPU3618(例如,离散CPU或dCPU)。在至少一个实 施例中,一个或更多个CPU3618可以包括例如X86处理器。例如,一个 或更多个CPU 3618可以用于执行各种功能中的任何功能,包括仲裁ADAS 传感器与一个或更多个SoC 3604之间的潜在不一致结果,和/或监测一个 或更多个控制器3636和/或片上信息娱乐系统(“信息娱乐SoC”)3630的 状态和健康。

[0497] 在至少一个实施例中,车辆3600可包括可经由高速互连(例如， NVIDIA的NVLINK)耦合到SoC3604的GPU3620(例如,离散GPU或 dGPU)。在至少一个实施例中,一个或更多个GPU 3620可以诸如通过执 行冗余和/或不同的神经网络来提供附加的人工智能功能,并且可以用于至 少部分地基于来自车辆3600的传感器的输入(例如,传感器数据)来训练 和/或更新神经网络。

[0498] 在至少一个实施例中,车辆3600还可包括网络接口3624,网络接口 3624可包括但不限于无线天线3626(例如,用于不同通信协议的一个或更 多个无线天线3626,诸如蜂窝天线、蓝牙天线等)。在至少一个实施例中， 网络接口3624可以用于使得能够通过互联网与云(例如,与服务器和/或 其他网络设备)、与其他车辆、和/或与计算设备(例如,乘客的客户端设 备)进行无线连接。在至少一个实施例中,为了与其他车辆通信,可以在 车辆360与其他车辆之间建立直接链路和/或可以(例如,跨网络和通过互 联网)建立间接链路。在至少一个实施例中,可以使用车辆到车辆通信链 路来提供直接链路。车辆到车辆通信链路可提供关于车辆3600附近的车辆 (例如,车辆3600前面、侧面和/或后面的车辆)的车辆3600信息。在至 少一个实施例中,上述功能可以是车辆3600的协作式自适应巡航控制功能 的一部分。

[0499] 在至少一个实施例中,网络接口3624可以包括提供调制和解调功能 并且使得控制器3636能够通过无线网络通信的SoC。在至少一个实施例中， 网络接口3624可以包括用

于从基带到射频的上转换和从射频到基带的下 转换的射频前端。在至少一个实施例中，频率转换可以以任何技术上可行 的方式执行。例如，频率转换可通过众所周知的过程和/或使用超外差过程 来执行。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-波、ZigBee、LoRaWAN 和/或其他无线协议进行通信的无线功能。

[0500] 在至少一个实施例中，车辆3600还可以包括数据存储3628，数据存 储3628可以包括但不限于片外(例如，SoC3604外)存储装置。在至少一 个实施例中，一个或更多个数据存储3628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/或可以存储至少一个数据 位的其他部件和/或设备。

[0501] 在至少一个实施例中，车辆3600还可以包括GNSS传感器3658(例 如，GPS和/或辅助GPS传感器)，以帮助映射、感知、占用网格生成和/ 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3658，包括例如但不限于使用具有以太网到串行(例如，RS-232)桥的 USB连接器的GPS。

[0502] 在至少一个实施例中，车辆3600还可包括RADAR传感器3660。一 个或更多个RADAR传感器3660可由车辆3600用于远距离车辆检测，甚 至在黑暗和/或恶劣天气条件下。在至少一个实施例中，RADAR功能性安 全水平可以是ASIL B。在一些示例中，一个或更多个RADAR传感器3660 可使用CAN和/或总线3602(例如，用于传输由一个或更多个RADAR传 感器3660生成的数据)来控制和访问对象追踪数据，访问以太网来访问原 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，一个或更多个RADAR传感器3660可以适合于前RADAR 使用、后RADAR使用和侧RADAR使用。在至少一个实施例中，一个或 更多个RADAR传感器3660中的一个或更多个是脉冲多普勒RADAR传感 器。

[0503] 在至少一个实施例中，一个或更多个RADAR传感器3660可包括不 同的配置，诸如具有窄视野的长范围、具有宽视角的短范围、短范围侧覆 盖等。在至少一个实施例中，长距离RADAR可以用于自适应巡航控制功 能。在至少一个实施例中，长距离RADAR系统可提供由两个或更多个独 立扫描实现的宽视角，诸如在250m范围内。在至少一个实施例中，一个 或更多个RADAR传感器3660可帮助区分静止和移动物体，并且可由 ADAS系统3638用于紧急制动辅助和前方碰撞警告。包括在长距离 RADAR系统中的传感器3660可包括但不限于具有多个(例如，六个或更 多个)固定RADAR天线和高速CAN和FlexRay接口的单站多模RADAR。 在至少一个实施例中，通过六个天线，中央的四个天线可以创建聚焦波束 图案，其被设计为以来自相邻车道中的交通的最小干扰以更高速度记录车 辆3600的周围环境。在至少一个实施例中，其他两个天线可以扩大视野， 使得可以快速检测进入或离开车辆3600的车道的车辆。

[0504] 在至少一个实施例中，中范围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)的视场。在至少 一个实施例中，短程RADAR系统可包括但不限于被设计成安装在后保险 杠的两端处的任何数量的RADAR传感器3660。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梁，这两个 梁持续地监测车辆后方和旁边的盲点。在至少一个实施例中，短程RADAR 系统可在ADAS系统3638中用于盲点检测和/或车道变换辅助。

[0505] 在至少一个实施例中，车辆3600还可以包括超声传感器3662。可位 于车辆3600的前面、后面和/或侧面的超声传感器3662可用于停车辅助和 /或创建和更新占用网格。在至少一个实施例中，可使用各种各样的超声传 感器3662，并且不同的超声传感器3662可用于不同的检测范围(例如， 2.5m，4m)。在至少一个实施例中，一个或更多个超声传感器3662可在 ASILB的功能安全水平下操作。

[0506] 在至少一个实施例中，车辆3600可包括LIDAR传感器3664。LIDAR 传感器3664可用于对象和行人检测、紧急制动、碰撞避免和/或其他功能。 在至少一个实施例中，一个或更多个LIDAR传感器3664可以是功能安全 水平ASIL B。在至少一个实施例中，车辆3600可以包括可以使用以太网 (例如，向千兆位以太网交换机提供数据)的多个LIDAR传感器3664(例 如，两个、四个、六个等)。

[0507] 在至少一个实施例中，一个或更多个LIDAR传感器3664可以能够 提供用于360度视场的对象及其距离的列表。在至少一个实施例中，可商 购的LIDAR传感器3664可具有约100m的广告范围，具有2cm-3cm的准 确度，并且支持例如100Mbps以太网连接。在至少一个实施例中，可以使 用一个或更多个不突出的LIDAR传感器3664。在这种实施例中，一个或更多个LIDAR传感器3664可以被实现为可以嵌入到车辆3600的前部、后 部、侧部和/或拐角中的小型设备。在至少一个实施例中，在这种实施例中， 一个或更多个LIDAR传感器3664可以提供高达120度水平和35度垂直视 场，甚至对于低反射率物体也具有200m范围。在至少一个实施例中，一 个或更多个前置LIDAR传感器3664可被配置为用于45度和135度之间的水平视场。

[0508] 在至少一个实施例中，还可使用诸如3D闪光LIDAR的LIDAR技术。3D闪光LIDAR使用激光器的闪光灯作为传输源，以照亮车辆3600的周围 环境高达约200m。在至少一个实施例中，闪光LIDAR单元包括但不限于 接收器，该接收器记录每个像素上的激光脉冲渡越时间和反射光，每个像 素进而对应于从车辆3600到物体的距离。在至少一个实施例中，闪光LIDAR可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，车辆3600 的每一侧各一个。在至少一个实施例中，3D闪存LIDAR系统包括但不限 于除了风扇之外没有移动部件的固态3D启动阵列LIDAR相机(例如，非 扫描LIDAR设备)。在至少一个实施例中，一个或更多个闪光LIDAR设 备可以每帧使用5纳秒I类(眼睛安全)激光脉冲，并且可以以3D范围点 云和共同配准的强度数据的形式捕获反射的激光。

[0509] 在至少一个实施例中，车辆可进一步包括IMU传感器3666。在至少 一个实施例中，在至少一个实施例中，一个或更多个IMU传感器3666可 以位于车辆3600的后轴的中心。在至少一个实施例中，一个或更多个IMU 传感器3666可以包括例如但不限于加速度计、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，诸如在六轴应用中，一个或更多个IMU传感器3666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，诸如在九轴应用中， 一个或更多个IMU传感器3666可以包括但不限于加速度计、陀螺仪和磁 力计。

[0510] 在至少一个实施例中，一个或更多个IMU传感器3666可以被实现为 小型高性能GPS辅助式惯性导航系统(“GPS/INS”)，该GPS辅助式惯性 导航系统将微机电系统(“MEMS”)惯性传感器、高灵敏度GPS接收器和 高级卡尔曼滤波算法进行组合以提供位置、速度和姿

态的估计。在至少一 个实施例中，一个或更多个IMU传感器3666可以使得车辆3600能够通过 直接观察和关联从GPS到一个或更多个IMU传感器3666的速度变化来估 计航向而不需要来自磁传感器的输入。在至少一个实施例中，IMU传感器 3666和GNSS传感器3658可组合在单个集成单元中。

[0511] 在至少一个实施例中，车辆3600可包括放置在车辆3600中和/或周 围的麦克风3696。在至少一个实施例中，一个或更多个麦克风3696可用 于紧急车辆检测和识别，等等。

[0512] 在至少一个实施例中，车辆3600还可包括任意数量的相机类型，包 括立体相机3668、宽视角相机3670、红外相机3672、环绕相机3674、长 距相机3698、中距相机3676和/或其他相机类型。在至少一个实施例中， 相机可以用于捕获车辆3600的整个周边周围的图像数据。在至少一个实施 例中，所使用的摄像头的类型取决于车辆3600。在至少一个实施例中，相 机类型的任何组合可用于提供车辆3600周围的必要覆盖。在至少一个实施 例中，相机的数量可以取决于实施例而不同。例如，在至少一个实施例中， 车辆3600可包括六个摄像机、七个摄像机、十个摄像机、十二个摄像机或 另一数量的摄像机。作为示例而非限制，相机可以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，本文中先前 参照图36A和图36B更详细地描述了每个照相机。

[0513] 在至少一个实施例中，车辆3600还可包括振动传感器3642。在至少 一个实施例中，一个或更多个振动传感器3642可以测量车辆3600的部件 (诸如，一个或更多个车轴)的振动。例如，在至少一个实施例中，振动 的变化可以表示路面的变化。在至少一个实施例中，当使用两个或更多个 振动传感器3642时，振动之间的差异可以用于确定路面的摩擦或滑移(例 如，当振动差异在动力驱动轴和自由旋转轴之间时)。

[0514] 在至少一个实施例中，车辆3600可包括ADAS系统3638。在一些 示例中，ADAS系统3638可包括但不限于SoC。在至少一个实施例中， ADAS系统3638可以包括但不限于任意数量和组合的自主/自适应/自动巡 航控制(“ACC”)系统，协作式自适应巡航控制(“CACC”)系统、前向 碰撞预警(“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、盲点警告(“BSW”)系 统，后方交叉路口预警(“RCTW”)系统、碰撞预警(“CW”)系统、车道 对中(“LC”)系统，和/或其他系统、特征和/或功能。

[0515] 在至少一个实施例中，ACC系统可使用RADAR传感器3660、LIDAR 传感器3664和/或任何数量的相机。在至少一个实施例中，ACC系统可包 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 系统监视和控制距车辆3600正前方的车辆的距离，并自动调整车辆3600 的速度以保持距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，并在必要时建议车辆3600改变车道。在至少一个实施 例中，横向ACC与诸如LC和CW的其他ADAS应用相关。

[0516] 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 息可以经由无线链路或经由网络连接(例如，通过互联网)间接地从其他 车辆经由网络接口3624和/或无线天线3626接收。在至少一个实施例中， 直接链路可以由车辆到车辆(“V2V”)通信链路提供，而间接链路可以由 基础设施到车辆(“I2V”)通信链路提供。通常，V2V通信概念提供关于紧前面的车辆(例如，紧前面的车辆和与车辆3600相同车道的车辆)的信 息，而I2V通信概念提供关于更前面的交通的信息。在至少一个实施例中， CACC系统可包括I2V和V2V信息源中的任一个或两者。在至少一个实施 例中，给定车辆3600之前的车辆的信息，CACC系统可

以更可靠,并且其 具有提高交通流平滑度和减少道路上的拥堵的潜力。

[0517] 在至少一个实施例中,FCW系统被设计成用于警告驾驶员危险,从 而使得驾驶员可以采取校正动作。在至少一个实施例中,FCW系统使用耦 合到专用处理器、DSP、FPGA和/或ASIC的前置相机和/或RADAR传感 器3660,前置相机和/或RADAR传感器3660电耦合到驱动器反馈,诸如 显示器、扬声器和/或振动组件。在至少一个实施例中,FCW系统可提供 警告,诸如声音、视觉警告、振动和/或快速制动脉冲的形式。

[0518] 在至少一个实施例中,AEB系统检测到与另一车辆或其他物体的即 将发生的前方碰撞,并且如果驾驶员在指定时间或距离参数内没有采取校 正动作,则AEB系统可自动应用制动。在至少一个实施例中,AEB系统 可使用耦合到专用处理器、DSP、FPGA和/或ASIC的前置相机和/或 RADAR传感器3660。在至少一个实施例中,当AEB系统检测到危险时, AEB系统通常首先提醒驾驶员采取纠正措施以避免碰撞,并且如果驾驶员 不采取纠正措施,AEB系统可自动施加刹车以努力防止或至少减轻预测的 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支持 和/或碰撞即将来临的制动之类的技术。

[0519] 在至少一个实施例中,LDW系统提供视觉、听觉和/或触觉警告,诸 如方向盘或座椅振动,以在车辆3600越过车道标志时警告驾驶员。在至少 一个实施例中,当驾驶员通过激活转向信号指示故意车道偏离时,LDW系 统不激活。在至少一个实施例中,LDW系统可使用耦合到专用处理器、 DSP、FPGA和/或ASIC的面向前侧的相机,其电耦合到驱动器反馈,诸如显示器、扬声器和/或振动组件。在至少一个实施例中,LKA系统是LDW 系统的变形。如果车辆3600开始离开车道,LKA系统提供转向输入或制 动以校正车辆3600。

[0520] 在至少一个实施例中,BSW系统检测并警告驾驶员汽车的盲点中的 车辆。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 告以指示合并或改变车道是不安全的。在至少一个实施例中,BSW系统可 以在驾驶员使用转向信号时提供额外的警告。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、电耦合到驱动 器反馈(诸如显示器、扬声器和/或振动组件)的面向后侧的相机和/或 RADAR传感器3660。

[0521] 在至少一个实施例中,当车辆3600备份时,当在后置摄像头范围之 外检测到物体时,RCTW系统可提供视觉、听觉和/或触觉通知。在至少一 个实施例中,RCTW系统包括AEB系统以确保应用车辆制动以避免碰撞。 在至少一个实施例中,RCTW系统可使用耦合到专用处理器、DSP、FPGA 和/或ASIC的一个或更多个后置RADAR传感器3660,该传感器电耦合到 驱动器反馈,诸如显示器、扬声器和/或振动组件。

[0522] 在至少一个实施例中,常规ADAS系统可能易于出现假阳性结果, 这可能是烦人的和对驾驶员分心,但通常不是灾难性的,因为常规ADAS 系统警告驾驶员并允许驾驶员决定安全状况是否真正存在并相应地行动。 在至少一个实施例中,在冲突结果的情况下,车辆3600本身决定是从主计 算机还是从计算机(例如,第一控制器3636或第二控制器3636)得到结 果。例如,在至少一个实施例中,ADAS系统3638可以是用于向备份计算 机合理化模块提供感知信息的备份和/或辅助计算机。在至少一个实施例中, 备份计算机合理化监控器可以在硬件部件上运行冗余多样的软件以检测感 知和动态驱动任务中的故障。在至少一个实施例中,来自ADAS系统3638 的输出可提供给监督MCU。在至少一个实施例中,如果来自主计算机和辅 助计算机的输出冲突,则监督MCU确定如何协调冲突以确保安全操作。

[0523] 在至少一个实施例中，主计算机可被配置为给监督MCU提供置信度 得分，指示主计算机在选择的结果中的置信度。在至少一个实施例中，如 果置信度分数超过阈值，则监督MCU可遵循主计算机的方向，而不管副 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，在置信度分 数不满足阈值的情况下，以及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可在计算机之间仲裁以确定适当的结果。

[0524] 在至少一个实施例中，监督MCU可被配置为运行神经网络，该神经 网络被训练并被配置为至少部分地基于来自主计算机和辅计算机的输出来 确定辅计算机提供假警报的条件。在至少一个实施例中，监督MCU中的 一个或更多个神经网络可以学习副计算机的输出何时可以被信任以及它何 时不能信任。例如，在至少一个实施例中，当副计算机是基于RADAR的 FCW系统时，监督MCU中的一个或更多个神经网络可以学习FCW系统 何时识别实际上不是危险的金属对象，诸如触发警报的排水栅或检修孔盖。 在至少一个实施例中，当副计算机是基于摄像机的LDW系统时，监督MCU 中的神经网络可以学习当自行车者或行人存在并且车道偏离实际上是最安 全的操纵时超越LDW。在至少一个实施例中，监督MCU可包括适于运行 具有相关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可包括和/或包括为SoC3604的部件。

[0525] 在至少一个实施例中，ADAS系统3638可包括使用计算机视觉的传 统规则执行ADAS功能的辅助计算机。在至少一个实施例中，副计算机可 使用经典计算机视觉规则(如果-那么)，并且监督MCU中神经网络的存 在可改善可靠性、安全性和性能。例如，在至少一个实施例中，不同的实 现方式和有意的非身份使得整个系统更容错，尤其是对由软件(或软件- 硬件接口)功能引起的故障更容错。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错误或错误，并且运行在副计算机上的 不相同的软件代码提供相同的总体结果，则监督MCU可具有总体结果是 正确的更大的置信度，并且主计算机上的软件或硬件中的错误不引起材料 错误。

[0526] 在至少一个实施例中，ADAS系统3638的输出可以被馈送到主计算 机的感知块和/或主计算机的动态驱动任务块中。例如，在至少一个实施例 中，如果ADAS系统3638指示由于直接在前的对象引起的向前碰撞警告， 则感知块可在识别对象时使用该信息。在至少一个实施例中，副计算机可 以具有其自己的神经网络，该神经网络被训练并且因此降低误报的风险， 如在此所描述的。

[0527] 在至少一个实施例中，车辆3600可以进一步包括信息娱乐SoC 3630 (例如，车载信息娱乐系统(IVI))。尽管被示出和描述为SoC，在至少一 个实施例中，信息娱乐系统3630可以不是SoC，并且可以包括但不限于两 个或更多个离散部件。在至少一个实施例中，信息娱乐SoC3630可以包括 但不限于可以用于提供音频(例如，音乐、个人数字助理、导航指令、新 闻、无线电等)的硬件和软件的组合，视频(例如，TV、电影、流等)、电话(例如，免提呼叫)、网络连接(例如，LTE、WiFi等)，和/或到车辆 3600的信息服务(例如，导航系统、后方停车辅助、无线电数据系统、车 辆相关信息，诸如燃料水平、总行驶距离、制动燃料水平、油水平、门打 开/关闭、空气过滤器信息等)。例如，信息娱乐SoC3630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，车库、车载娱乐、WiFi、方向盘音频控制、免提语音控制、平视显示器(“HUD”)，HMI显 示器3634、远程信息处理设备、控制面板(例如，用于控制各种部件、特 征和/或系统和/或与各种部件、特征和/或系统交互)，和/或其他部

件。在 至少一个实施例中，信息娱乐SoC 3630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)，如来自ADAS系统3638的信息、自主驾驶 信息(如计划车辆操控)、轨迹、周围环境信息(例如，交叉口信息、车辆 信息、道路信息等)、和/或其他信息。

[0528] 在至少一个实施例中，信息娱乐SoC3630可以包括任何量和类型的 GPU功能。在至少一个实施例中，信息娱乐SoC3630可以通过总线3602 (例如，CAN总线、以太网等)与车辆3600的其他设备、系统和/或部件 通信。在至少一个实施例中，信息娱乐SoC 3630可以耦合至监督MCU，从而使得信息娱乐系统的GPU可以在主控制器3636(例如，车辆3600的 主和/或备份计算机)故障的情况下执行一些自驱动功能。在至少一个实施 例中，信息娱乐SoC3630可以将车辆3600置于排队到安全停止模式，如 本文所述。

[0529] 在至少一个实施例中，车辆3600还可包括仪表板3632(例如，数字 仪表板、电子仪表板、数字仪表板等)。在至少一个实施例中，组合仪表 3632可以包括但不限于控制器和/或超级计算机(例如，分立的控制器或超 级计算机)。在至少一个实施例中，组合仪表3632可以包括但不限于任意 数量和组合的一组仪表，例如速度计，燃料液位、油压、转速表、里程表、 转向指示器、换档位置指示器、安全带警告灯，驻车制动警告灯、发动机 故障灯、辅助约束系统(例如，气囊)信息，照明控制、安全系统控制、 导航信息等。在一些示例中，可以在信息娱乐SoC3630和仪表板3632之 间显示和/或共享信息。在至少一个实施例中，组合仪表3632可以被包括 作为信息娱乐SoC3630的一部分，反之亦然。

[0530] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。下面结合图6A和/或图6B提供关于推理和/或训练逻辑 615的细节。在至少一个实施例中，推理和/或训练逻辑615可以在图36C 的系统中用于至少部分地基于使用在此描述的神经网络训练操作、神经网 络功能和/或架构或神经网络使用情况计算的权重参数来推理或预测操作。

[0531] 推理和/或训练逻辑615用于执行与一个或更多个实施例相关联的推 理和/或训练操作。在至少一个实施例中，此逻辑可与这些图的组件一起使 用，以确定数据异常和适当的响应动作。

[0532] 图36D是根据至少一个实施例的用于在基于云的服务器与图36A的 自主车辆3600之间通信的系统3676的图。在至少一个实施例中，系统3676 可以包括但不限于服务器3678、网络3690和任意数量和类型的车辆，包 括车辆3600。在至少一个实施例中，一个或更多个服务器3678可以包括 但不限于多个GPU3684(A)-3684(H)(在此统称为GPU3684)、PCIe交换机3682(A)-3682(D)(在此统称为PCIe交换机3682)和/或CPU3680 (A)-3680(B)(在此统称为CPU3680)。GPU3684、CPU3680和PCIe 交换机3682可以与高速互连(例如但不限于由NVIDIA开发的NVLink接 口3688和/或PCIe连接3686)互连。在至少一个实施例中，GPU3684经 由NV链路和/或NV交换机SoC连接，并且GPU3684和PCIe交换机3682 经由PCIe互连连接。在至少一个实施例中，尽管示出了八个GPU3684、 两个CPU3680和四个PCIe交换机3682，但这不旨在是限制性的。在至少 一个实施例中，一个或更多个服务器3678中的每一个可包括但不限于任何 组合的任何数量的GPU3684、CPU3680和/或PCIe交换机3682。例如，在 至少一个实施例中，一个或更多个服务器3678可以各自包括八个、十六个、 三十二个和/或更多个GPU3684。

[0533] 在至少一个实施例中，一个或更多个服务器3678可以通过一个或更 多个网络

3690并且从车辆接收表示示出未预期的或改变的道路状况(如最 近开始的道路工作)的图像的图像数据。在至少一个实施例中,一个或更 多个服务器3678可以通过一个或更多个网络3690向车辆传输神经网络3692、经更新的神经网络3692、和/或地图信息3694,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3694的更新可 以包括但不限于对HD地图3622的更新,诸如关于建筑工地、坑洼、绕路、 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3692、经更 新的神经网络3692、和/或地图信息3694可能已经由在从环境中的任何数 量的车辆接收的数据中表示的新的训练和/或经验产生,和/或至少部分地基 于在数据中心处执行的训练(例如,使用一个或更多个服务器3678和/或 其他服务器)。

[0534] 在至少一个实施例中,一个或更多个服务器3678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)。在至少一个实施 例中,训练数据可以由车辆生成,和/或可以在模拟中生成(例如,使用游 戏引擎)。在至少一个实施例中,任何量的训练数据被标记(例如,其中相 关联的神经网络受益于受监督学习)和/或经历其他预处理。在至少一个实 施例中,不标记和/或预处理任何量的训练数据(例如,在相关联的神经网络不需要受监督学习的情况下)。在至少一个实施例中,一旦训练了机器学 习模型,车辆就可以使用机器学习模型(例如,通过网络3690传输至车辆), 和/或服务器3678可以使用机器学习模型来远程监测车辆。

[0535] 在至少一个实施例中,一个或更多个服务器3678可以从车辆接收数 据,并且将数据应用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3678可以包括由一个或更多个 GPU3684供电的深度学习超级计算机和/或专用AI计算机,诸如由NVIDIA 开发的DGX和DGX站机器。然而,在至少一个实施例中,一个或更多个服务器3678可以包括使用由CPU供电的数据中心的深度学习基础设施。

[0536] 在至少一个实施例中,一个或更多个服务器3678的深度学习基础设 施可以能够快速、实时推理,并且可以使用该能力来评估和验证车辆3600 中的处理器、软件和/或相关联的硬件的健康。例如,在至少一个实施例中, 深度学习基础设施可从车辆3600接收周期性更新,诸如车辆3600已位于 该图像序列中的图像序列和/或对象(例如,经由计算机视觉和/或其他机器 学习对象分类技术)。在至少一个实施例中,深度学习基础结构可以运行其自己的神经网络以识别对象,并将它们与车辆3600识别的对象进行比较, 如果结果不匹配并且深度学习基础设施推理车辆3600中的AI发生故障, 然后服务器3678可以向车辆3600发送指示车辆3600的故障保护计算机采 取控制的信号,通知乘客,完成安全的停车操作。

[0537] 在至少一个实施例中,一个或更多个服务器3678可以包括一个或更 多个GPU3684和一个或更多个可编程推理加速器(例如,NVIDIA的 TensorRT3)。在至少一个实施例中,GPU供电的服务器和推理加速的组合 可以使得实时响应成为可能。在至少一个实施例中,诸如在性能不太关键 的情况下,由CPU、FPGA和其他处理器供电的服务器可以用于推理。在至少一个实施例中,推理和/或训练逻辑615用于执行一个或更多个实施例。下面结合图6A和/或图6B提供关于推理和/或训练逻辑615的细节。

[0538] 其他变化在本公开的精神内。由此,虽然所公开的技术易受不同修 改和替代构造的影响,但其某些所解说的实施例在附图中示出并且已在上 文详细地描述。然而,应当理解,并非旨在将本公开限制为所公开的一种 或多种特定形式,而是相反,旨在覆盖落在如

所附权利要求所限定的本公 开的精神和范围内的所有修改、替代构造和等同物。

[0539] 在描述所披露的实施例的背景下(尤其在以下权利要求的背景下)，术语“一个/一种(a/an)”和“该(the)”以及类似指称词的使用应被解释 为覆盖单数和复数两者,除非在此另外指明或与上下文明显矛盾,并且不 作为术语的定义。除非另外指出,术语“包含”、“具有”、“包括”和“含 有”应被解释为开放式术语(意指“包括但不限于”)。术语“连接”在未修改和提及物理连接时应被解释为部分地或完全地包含在一起、附接至其 上、或连接在一起,即使存在介入的东西。除非本文另有说明,本文列举 的数值范围仅旨在用作单独提及落入范围内的每个单独数值的简略方法，并且每个单独数值并入说明书中,就好像其在本文单独列举一样。除非上 下文另外指出或矛盾,否则术语“集合”(例如,“项目集合”)或“子集” 的使用应被解释为包括一个或更多个成员的非空集合。进一步,除非上下 文另外指出或矛盾,否则对应集合的术语“子集”不一定表示对应集合的 适当子集,但是子集和对应集合可以是相等的。

[0540] 连接语言,如“A、B和C中的至少一项”形式的短语,“或“A、B 和C中的至少一个”,除非另有明确说明或者与上下文明显矛盾,用上下 文另外理解,所述上下文通常用于呈现项目、术语,等可以是A或B或C，或者A和B和C的集合的任何非空子集。例如,在具有三个成员的集合的 说明性示例中,连接性短语“A、B和C中的至少一个”和“A、B和C 中的至少一个”是指以下集合中的任何集合:{A}、{B}、{C}、{A、B}、 {A、C}、{B、C}、{A、B、C}。因此,此类连接性语言一般不旨在暗示 某些实施例要求各自存在A中的至少一个、B中的至少一个、以及C中的 至少一个。此外,除非上下文另外指出或矛盾,否则术语“多个”指示为 复数的状态(例如,“多个项目”指示多个项目)。多个是至少两个项目，但当明确地或通过上下文如此指示时可能更多。进一步地,除非另外指明 或另外从上下文清楚,短语“基于”是指“至少部分地基于”而不是“仅 基于。”

[0541] 在此描述的过程的操作可以按任何适合的顺序进行,除非在此另外 指明或另外与上下文明显矛盾。在至少一个实施例中,诸如本文中描述的 那些过程(或其变型和/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,并被实现为在一个或更多个处理器上共同执 行的代码(例如,可执行指令、一个或更多个计算机程序或一个或更多个 应用),通过硬件或其组合。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非瞬态计算机 可读存储介质,该非瞬态计算机可读存储介质排除瞬态信号(例如,传播 的瞬态电或电磁传输)但包括瞬态信号的收发机内的非瞬态数据存储电路 (例如,缓冲器、高速缓存和队列)。在至少一个实施例中,代码(例如，可执行代码或源代码)被存储在其上存储有可执行指令(或用于存储可执 行指令的其他存储器)的一个或更多个非瞬态计算机可读存储介质的集合 上,当由计算机系统的一个或更多个处理器执行时(即,由于被执行),使 计算机系统执行本文描述的操作。在至少一个实施例中,一组非暂态计算 机可读存储介质包括多个非暂态计算机可读存储介质和多个非暂态计算机 可读存储介质中的各个非暂态计算机可读存储介质中的一个或更多个缺少 全部代码,而多个非暂态计算机可读存储介质共同地存储全部代码。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,非瞬态计算机可读存储介质存储指令,并且主中央处理单元 (“CPU”)执行指令中的一些指令,而图形

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指令的不同子集。

[0542] 因此,在至少一个实施例中,计算机系统被配置为实现单独地或共 同地执行本文中描述的过程的操作的一个或更多个服务,并且此类计算机 系统配置有能够实现操作的执行的可应用的硬件和/或软件。进一步,实现 本公开的至少一个实施例的计算机系统是单个设备,并且在另一实施例中, 是包括不同地操作的多个设备的分布式计算机系统,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。

[0543] 在此提供的任何和所有示例或示例性语言(例如,“如”)的使用仅 旨在更好地阐明本披露的实施例并且不对本披露的范围构成限制,除非另 外要求。说明书中的语言不应被解释为指示任何未要求保护的元素对于实 践本公开是必不可少的。

[0544] 在此引用的所有参考文献(包括出版物、专利申请、以及专利)均 通过引用结合在此,其程度如同每个参考文献被单独地并且具体地指示为 通过引用结合并且在此以其全文列出一样。

[0545] 在说明书和权利要求书中,可以使用术语“耦合”和“连接”及其 派生词。应当理解,这些术语可能不旨在作为彼此的同义词。而是,在具 体示例中,“连接”或“耦合”可以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”还可以指两个或更多个元件彼此不直接接触, 但仍彼此合作或交互。

[0546] 除非另有明确说明,可以理解的是贯穿说明书术语,如“处理,“计 算”、“计算”、“确定”等指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,其操纵和/或变换被表示为物理的(诸如电子的)数据,计 算系统的寄存器和/或存储器内的量到类似地表示为计算系统的存储器内 的物理量的其他数据中,寄存器或其他此类信息存储、传输或显示装置。

[0547] 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 的电子数据并且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部分。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。如在此使 用的,“软件”过程可以包括例如随着时间执行工作的软件和/或硬件实体, 如任务、线程和智能代理。而且,每一过程可指代用于循序或并行地、连 续地或间歇地执行指令的多个过程。术语“系统”和“方法”在此可互换 地使用,只要系统可以体现一种或多种方法并且方法可以被认为是系统。

[0548] 在本文件中,可以参考获得、获取、接收、或输入模拟或数字数据 到子系统、计算机系统、或计算机实现的机器中。获得、获取、接收或输 入模拟和数字数据可以以各种方式完成,如通过接收作为函数调用或对应 用程序编程接口的调用的参数的数据。在一些实现中,获得、获取、接收 或输入模拟或数字数据的过程可通过经由串行或并行接口传送数据来完成。 在另一实现中,获得、获取、接收或输入模拟或数字数据的过程可通过经 由计算机网络从提供实体向获取实体传送数据来完成。还可以参考提供、 输出、传输、发送或呈现模拟或数字数据。在不同示例中,提供、输出、 传送、发送或呈现模拟或数字数据的过程可通过作为函数调用的输入或输 出参数、应用编程接口的参数或进程间通信机制来传送数据来完成。

[0549] 虽然以上论述陈述所描述的技术的举例实施例,但其他架构可用以 实施所描述

的功能性,且既定处于本发明的范围内。此外,尽管以上出于 讨论的目的定义了职责的特定分布,但是取决于情况,可以以不同的方式 分布和划分各种功能和职责。

[0550] 此外,尽管已经用结构特征和/或方法动作专用的语言描述了本主题,但应当理解的是,所附权利要求书中所要求保护的主题不必限于所描述的 具体特征或动作。相反,特定特征和动作被公开为实现权利要求的示范性 形式。

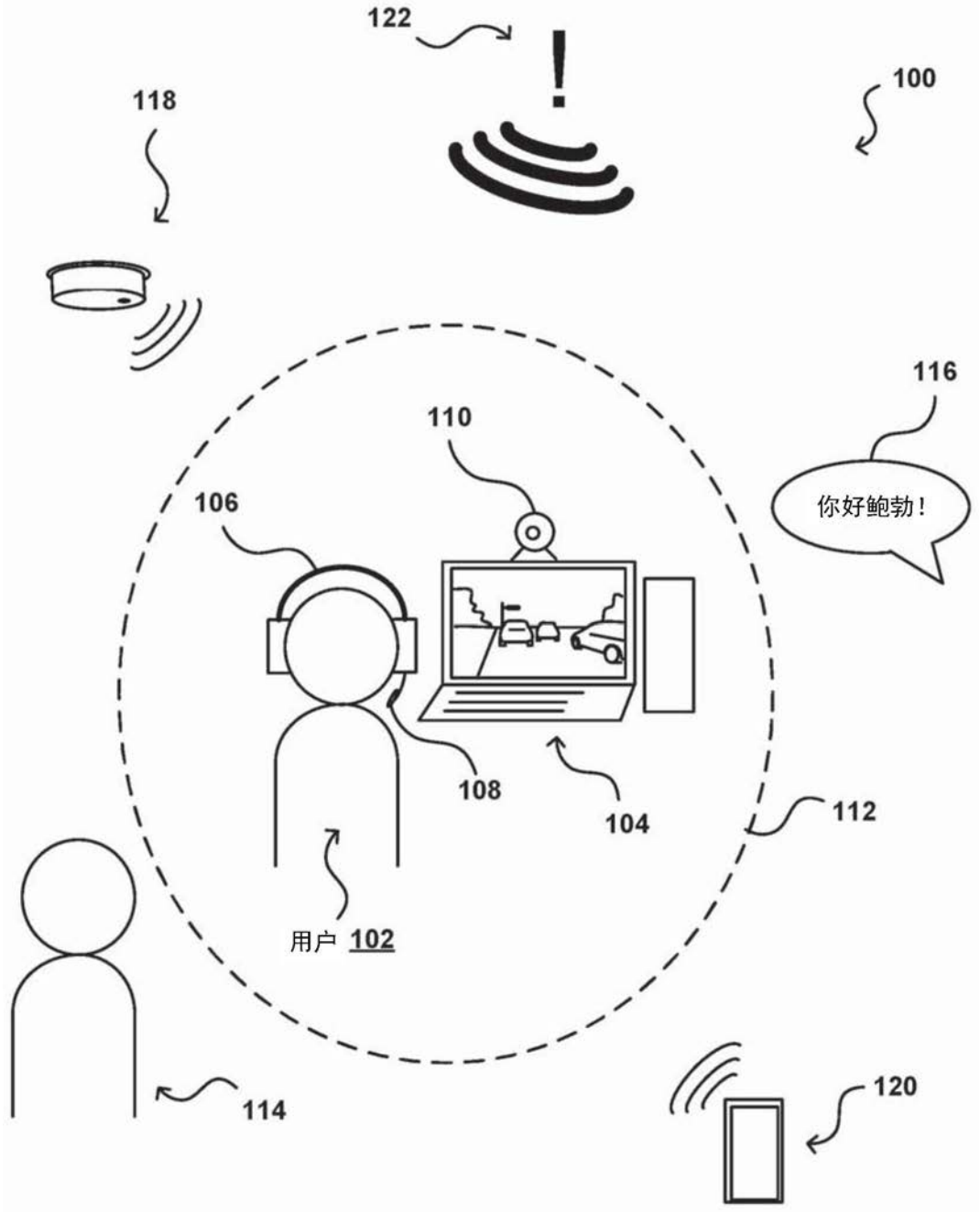


图1

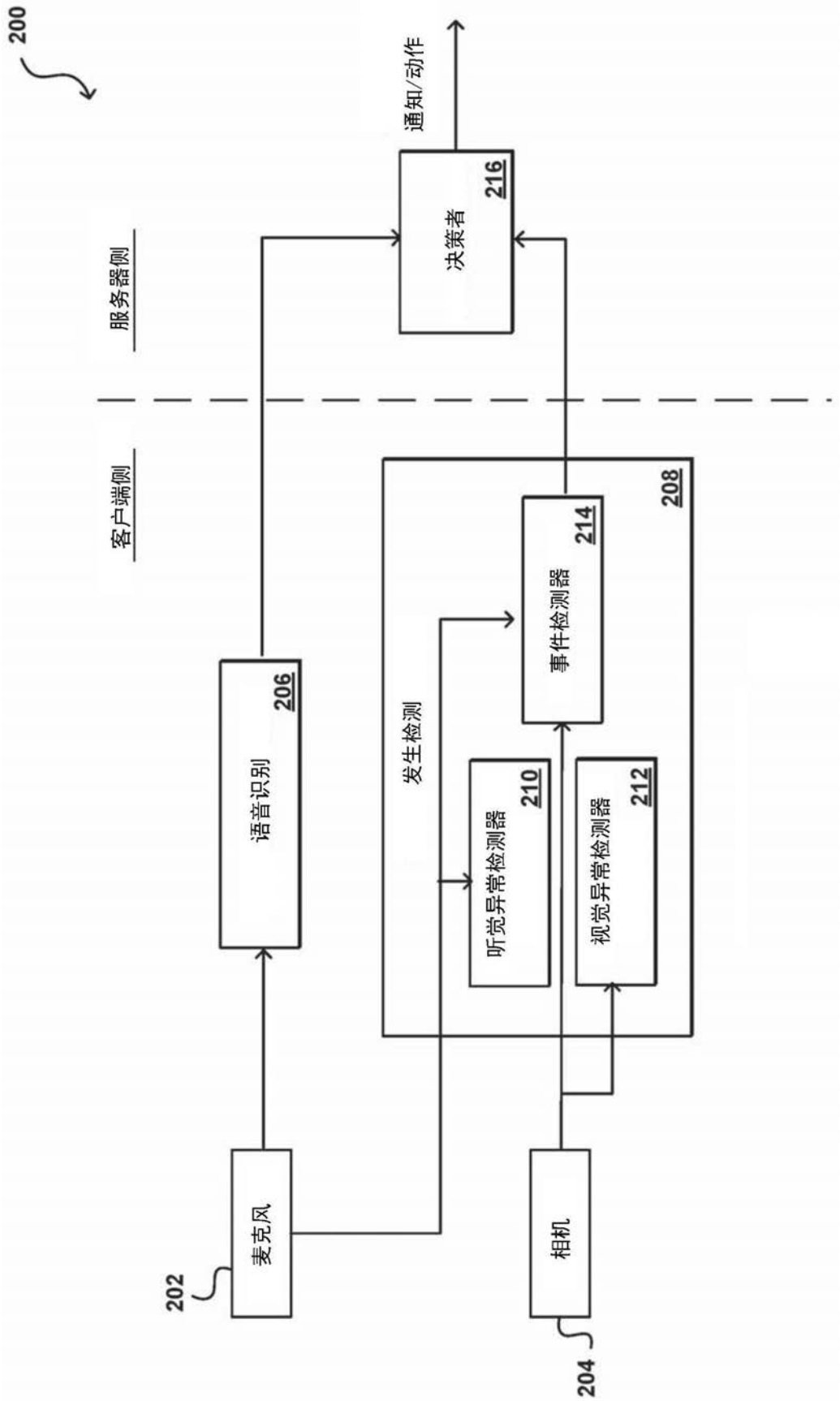

200
通知/动作
服务器侧
决策者
216
客户端侧
208
事件检测器
214
语音识别
206
发生检测
听觉异常检测器
210
视觉异常检测器
212
麦克风
202
相机
204

图2

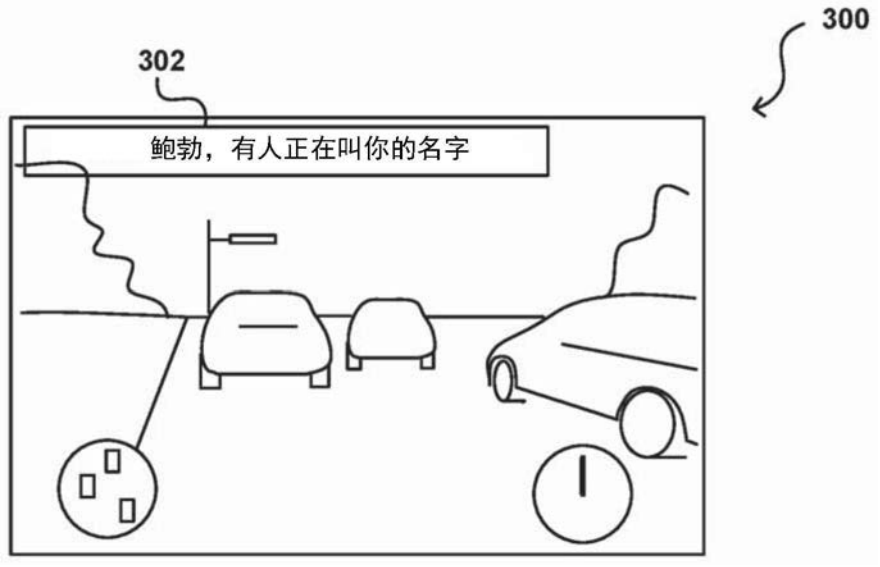

300
302
鲍勃，有人正在叫你的名字

图3A

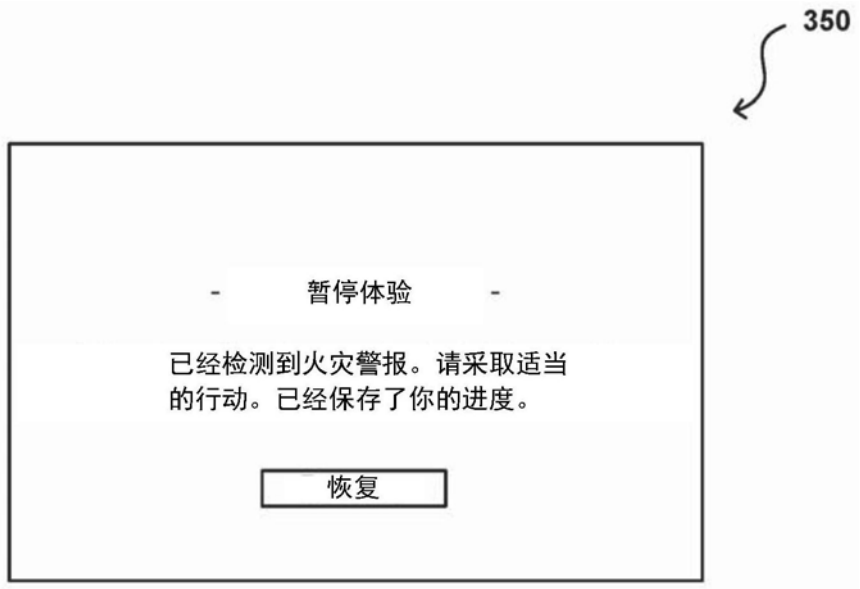

350
暂停体验
已经检测到火灾警报。请采取适当的行动。已经保存了你的进度。
恢复

图3B

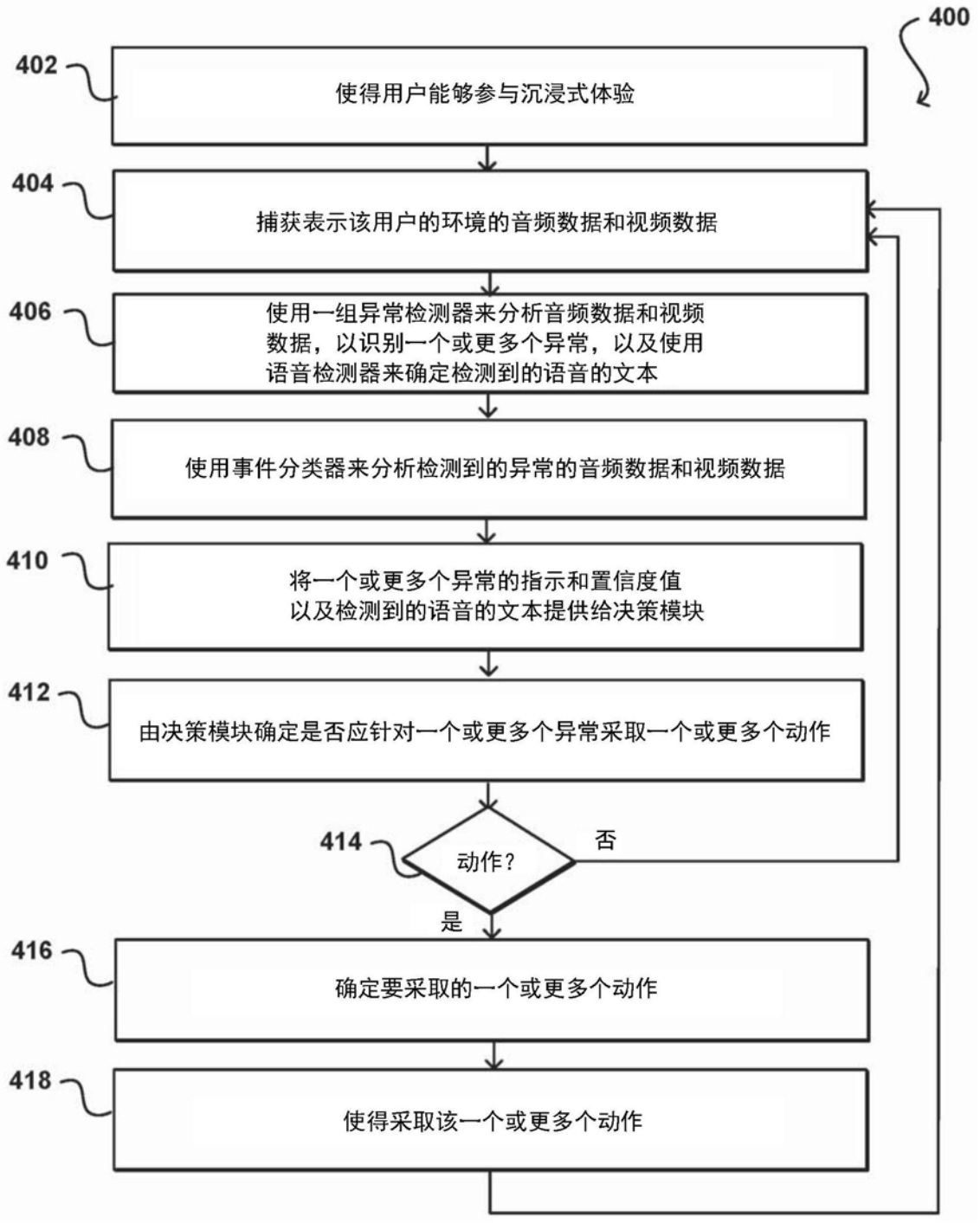

400
402
使得用户能够参与沉浸式体验
404
捕获表示该用户的环境的音频数据和视频数据
406
使用一组异常检测器来分析音频数据和视频数据，以识别一个或更多个异常，以及使用语音检测器来确定检测到的语音的文本
408
使用事件分类器来分析检测到的异常的音频数据和视频数据
410
将一个或更多个异常的指示和置信度值以及检测到的语音的文本提供给决策模块
412
由决策模块确定是否应针对一个或更多个异常采取一个或更多个动作
414
动作？
否
是
416
确定要采取的一个或更多个动作
418
使得采取该一个或更多个动作

图4

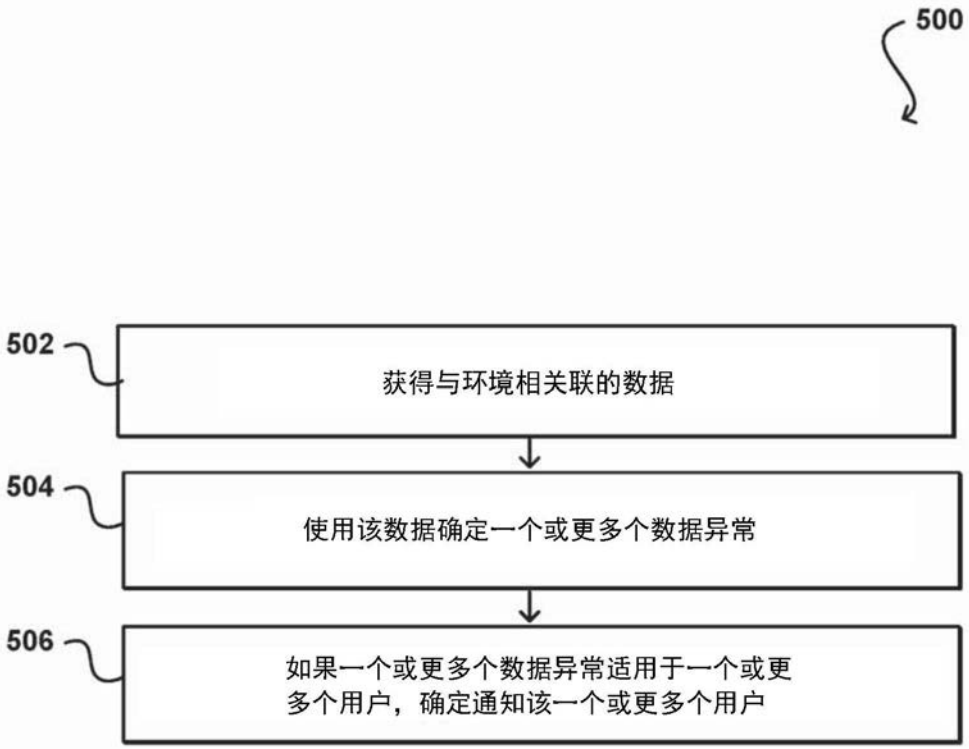

500
502
获得与环境相关联的数据
504
使用该数据确定一个或更多个数据异常
506
如果一个或更多个数据异常适用于一个或更多个用户，确定通知该一个或更多个用户

图5

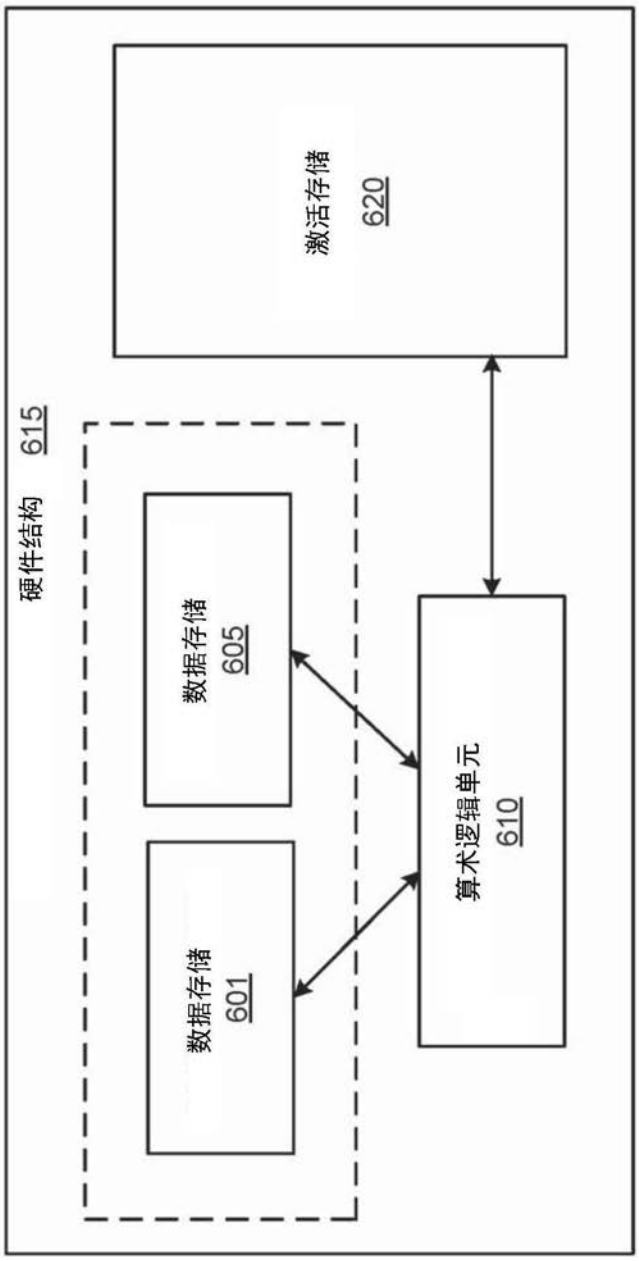


图6A

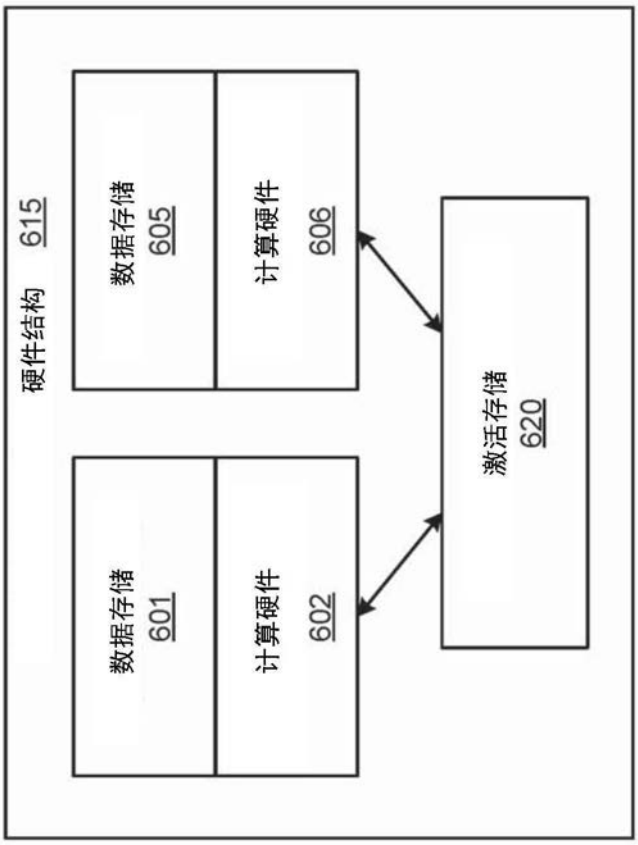


图6B

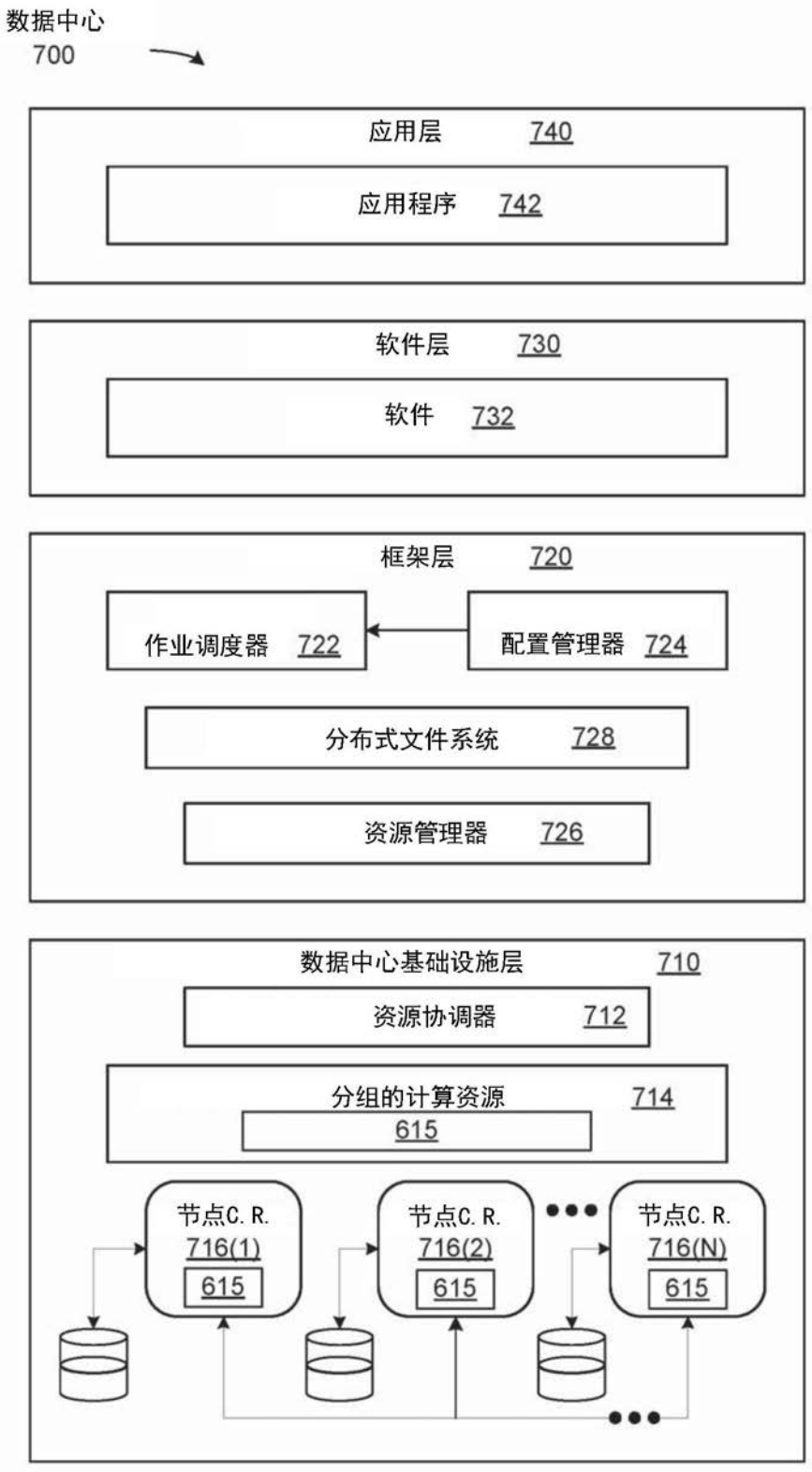

数据中心
700
应用层 740
应用程序 742
软件层 730
软件 732
框架层 720
作业调度器 722
配置管理器 724
分布式文件系统 728
资源管理器 726
数据中心基础设施层 710
资源协调器 712
分组的计算资源 714
615
节点C. R.
716(1)
615
节点C. R.
716(2)
615
节点C. R.
716(N)
615

图7

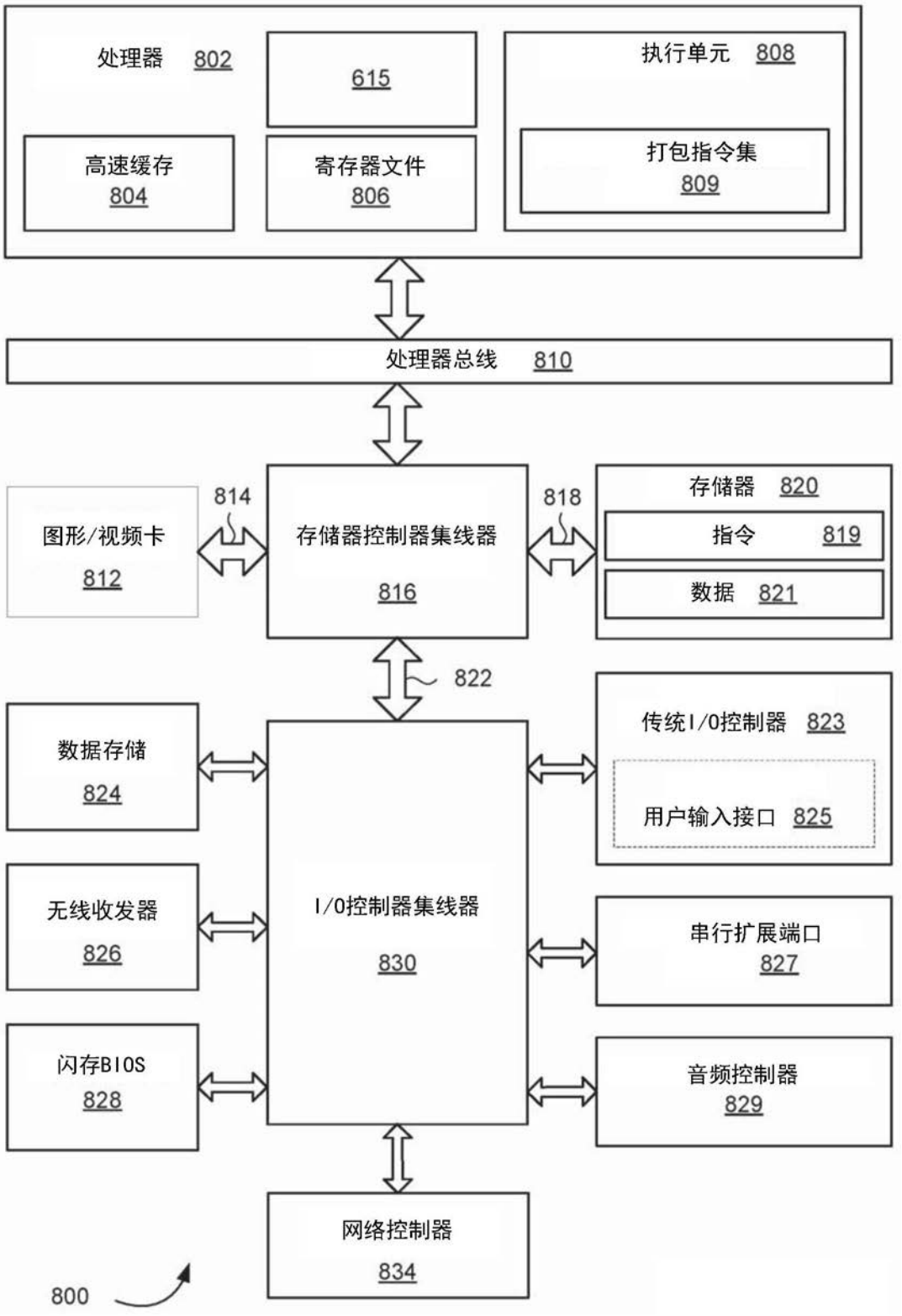

处理器 802
615
执行单元 808
高速缓存 804
寄存器文件 806
打包指令集 809
处理器总线 810
图形/视频卡 812
814
存储器控制器集线器 816
818
存储器 820
指令 819
数据 821
822
数据存储 824
传统I/O控制器 823
用户输入接口 825
无线收发器 826
I/O控制器集线器 830
串行扩展端口 827
闪存BIOS 828
音频控制器 829
网络控制器 834
800

图8

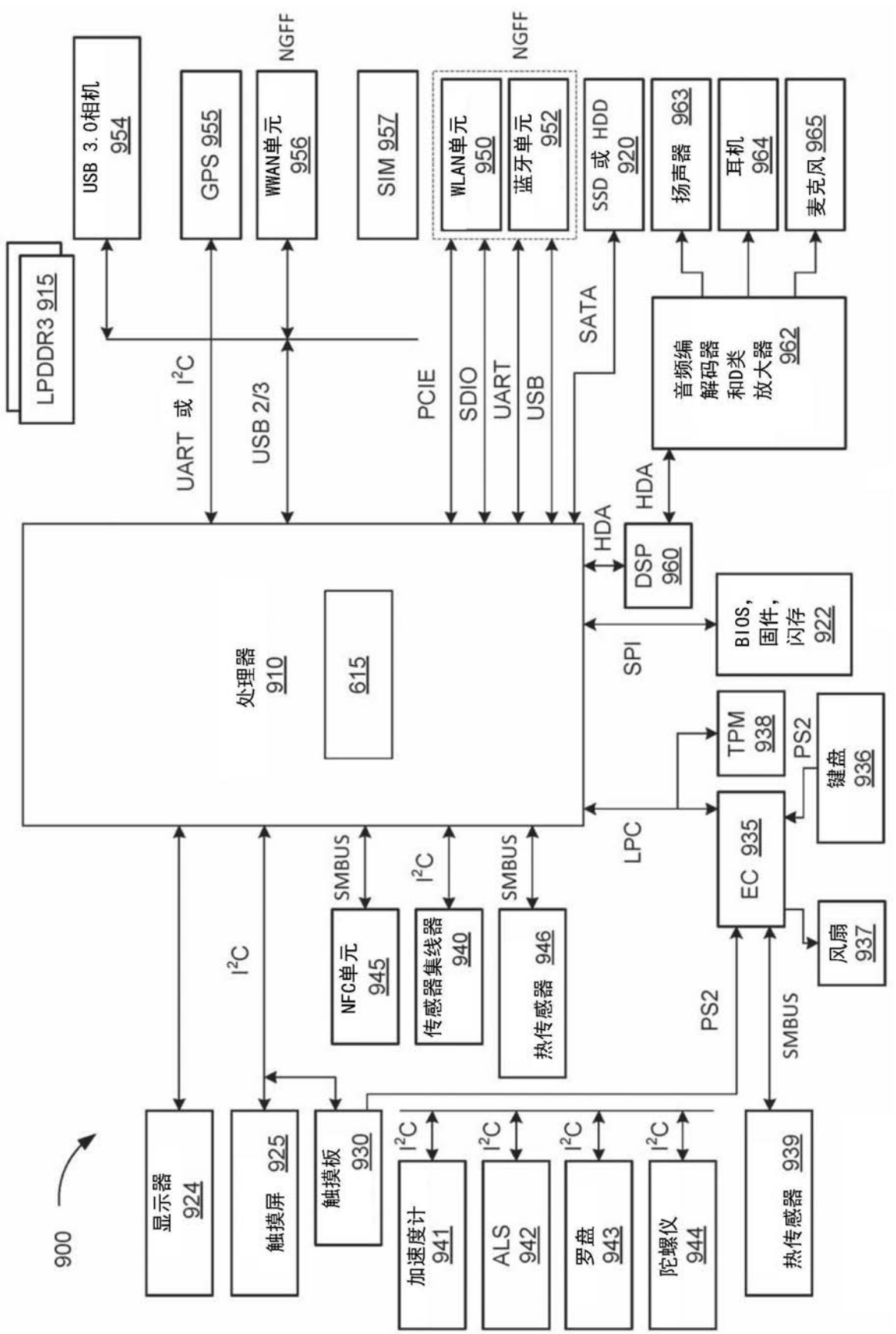

USB 3.0相机 954
GPS 955
NGFF
WWAN单元 956
SIM 957
NGFF
WLAN单元 950
蓝牙单元 952
SSD 或 HDD 920
扬声器 963
耳机 964
麦克风 965
LPDDR3 915
UART 或 I2C
USB 2/3
PCIE
SDIO
UART
USB
SATA
音频编解码器和D类放大器 962
HDA
HDA
DSP 960
处理器 910
615
SPI
BIOS，固件，闪存 922
TPM 938
PS2
键盘 936
LPC
EC 935
SMBUS
I2C
SMBUS
NFC单元 945
传感器集线器 940
热传感器 946
I2C
PS2
SMBUS
风扇 937
900
显示器 924
触摸屏 925
触摸板 930
I2C
I2C
I2C
I2C
加速度计 941
ALS 942
罗盘 943
陀螺仪 944
热传感器 939

图9

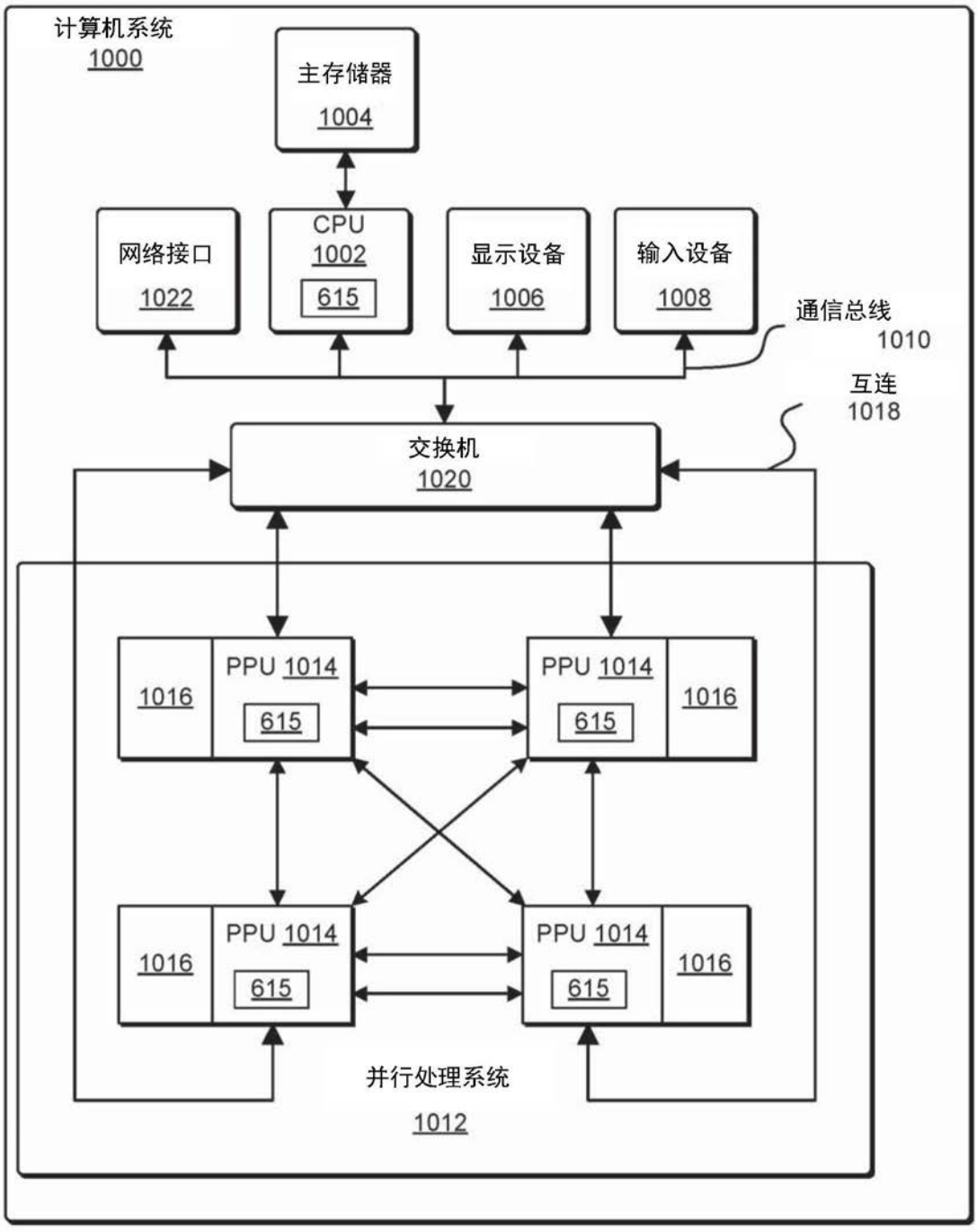

计算机系统
1000
主存储器
1004
网络接口
1022
CPU
1002
615
显示设备
1006
输入设备
1008
通信总线
1010
互连
1018
交换机
1020
1016
PPU 1014
615
PPU 1014
615
1016
1016
PPU 1014
615
PPU 1014
615
1016
并行处理系统
1012

图10

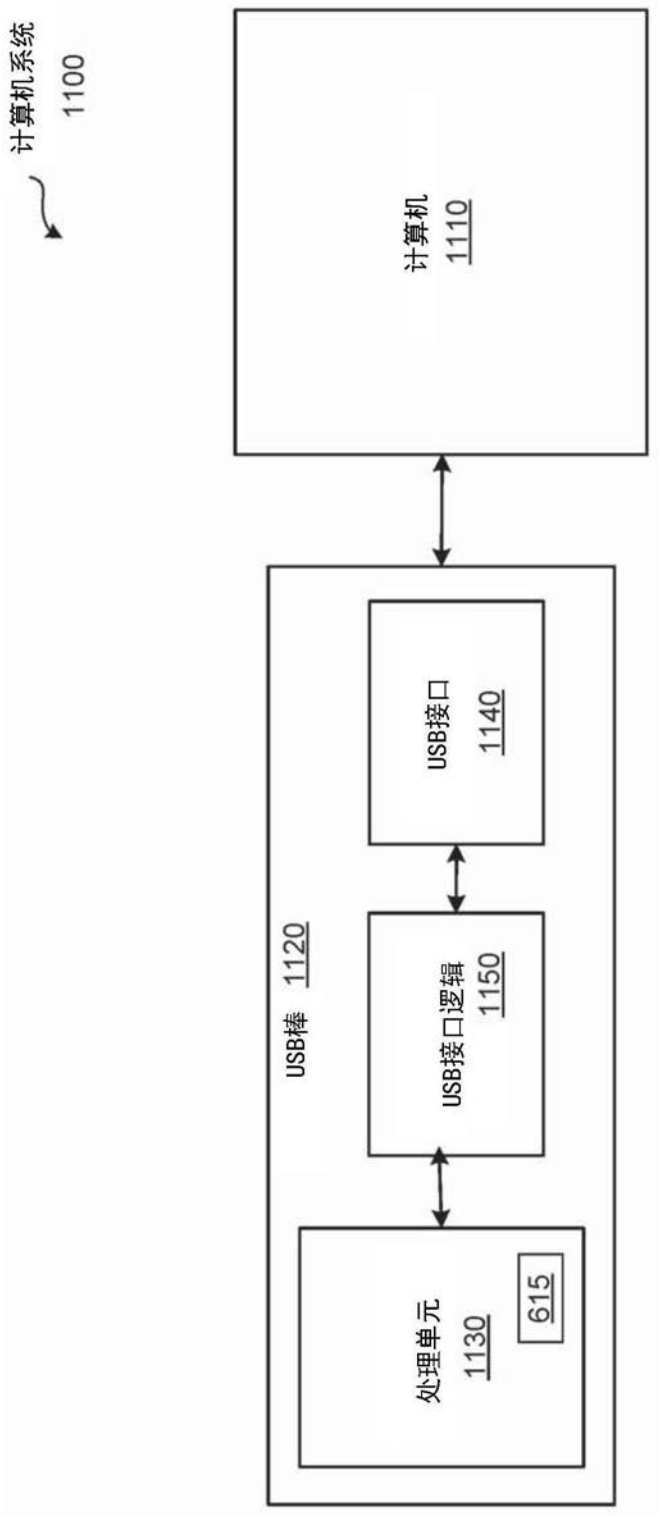

计算机系统
1100
计算机
1110
USB棒 1120
USB接口
1140
USB接口逻辑
1150
处理单元
1130
615

图11

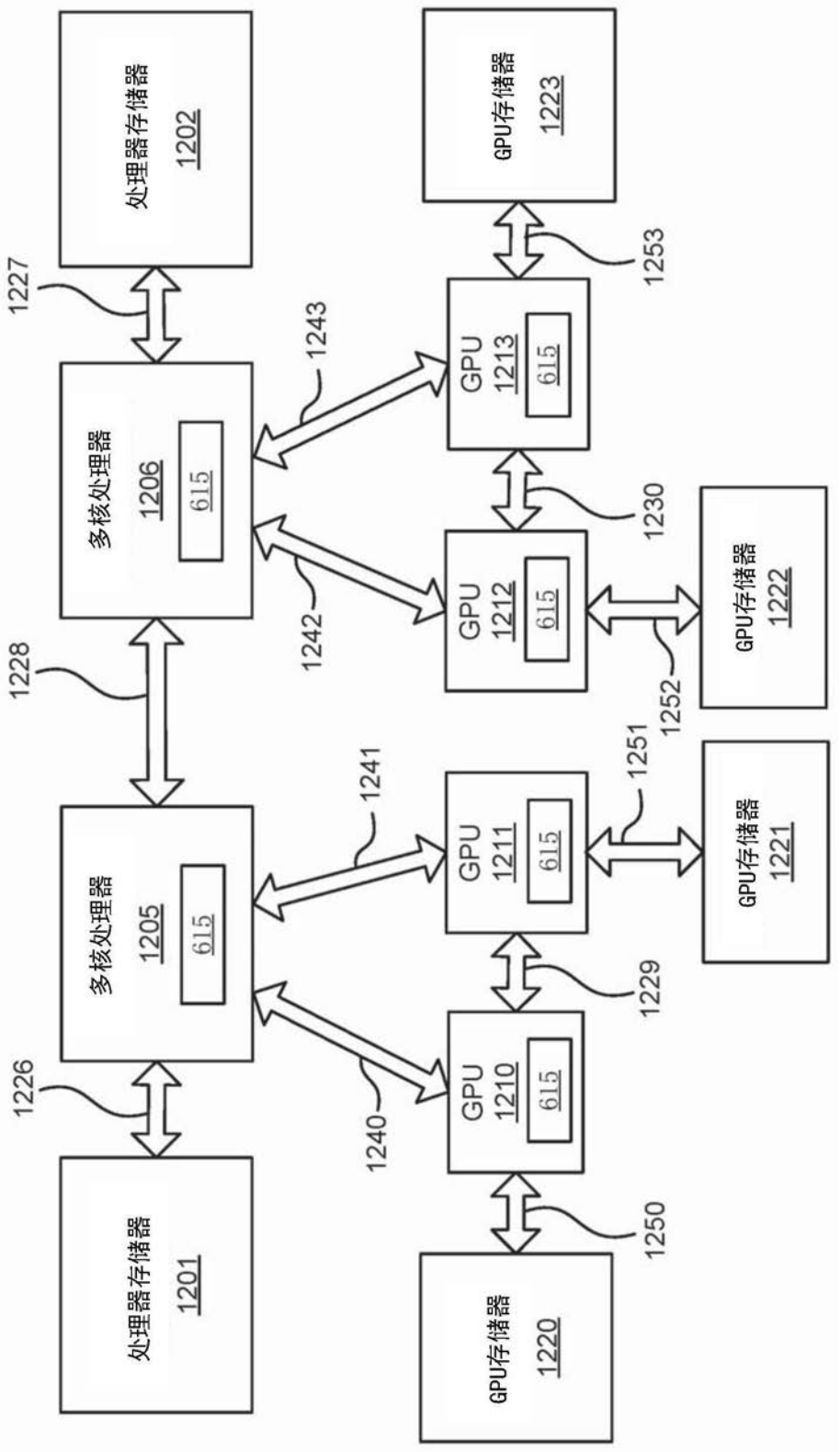

处理器存储器
1202
GPU存储器
1223
1227
1253
1243
多核处理器
1206
615
GPU
1213
615
1230
1242
GPU
1212
615
GPU存储器
1222
1228
1252
1251
1241
多核处理器
1205
615
GPU
1211
615
GPU存储器
1221
1229
1226
1240
GPU
1210
615
处理器存储器
1201
1250
GPU存储器
1220

图12A

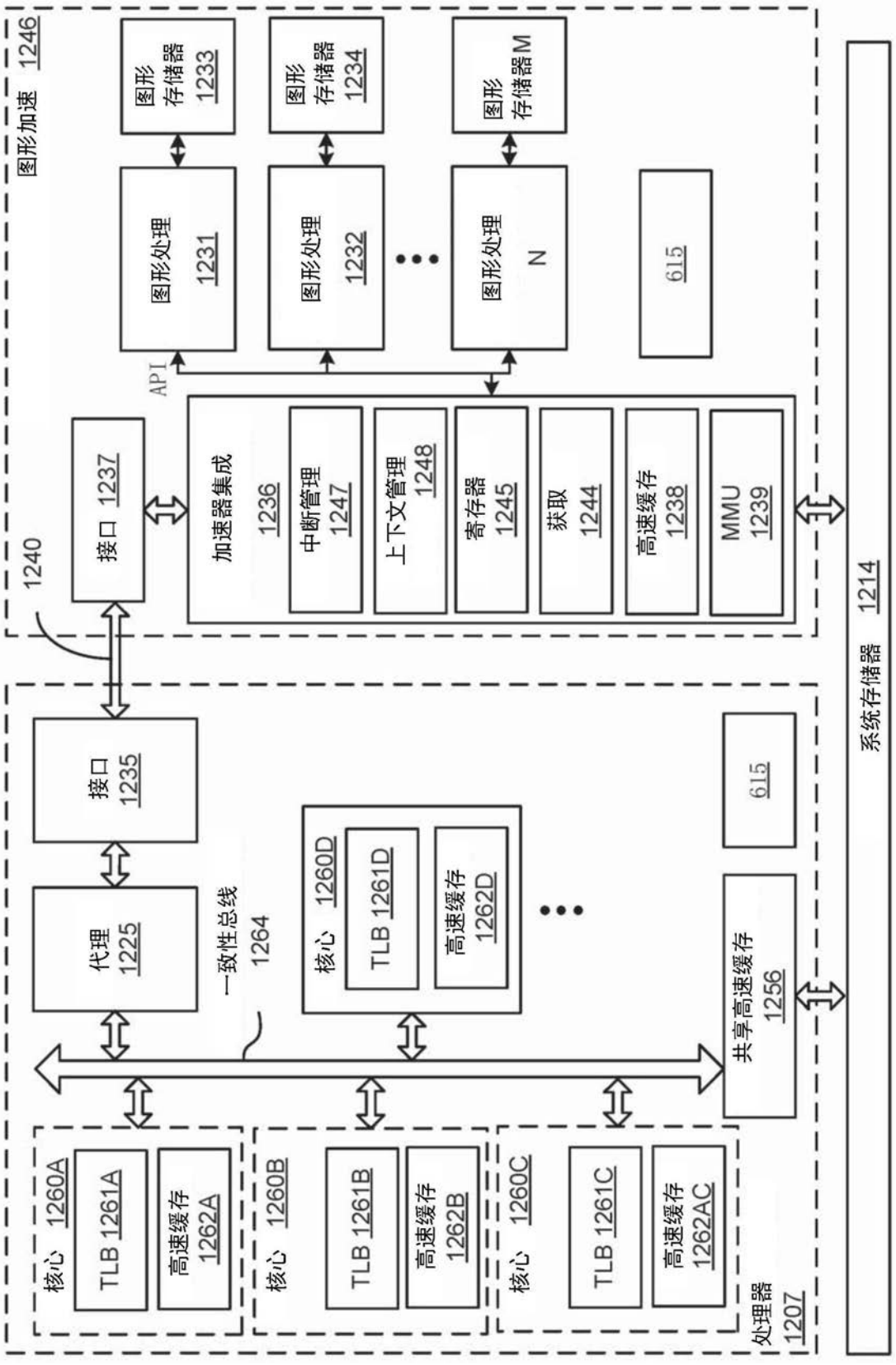

图形加速 1246
图形存储器 1233
图形存储器 1234
图形存储器M
图形处理 1231
图形处理 1232
图形处理 N
615
API
1240
接口 1237
加速器集成 1236
中断管理 1247
上下文管理 1248
寄存器 1245
获取 1244
高速缓存 1238
MMU 1239
系统存储器 1214
接口 1235
代理 1225
一致性总线 1264
核心 1260D
TLB 1261D
高速缓存 1262D
615
共享高速缓存 1256
核心 1260A
TLB 1261A
高速缓存 1262A
核心 1260B
TLB 1261B
高速缓存 1262B
核心 1260C
TLB 1261C
高速缓存 1262AC
处理器 1207

图12B

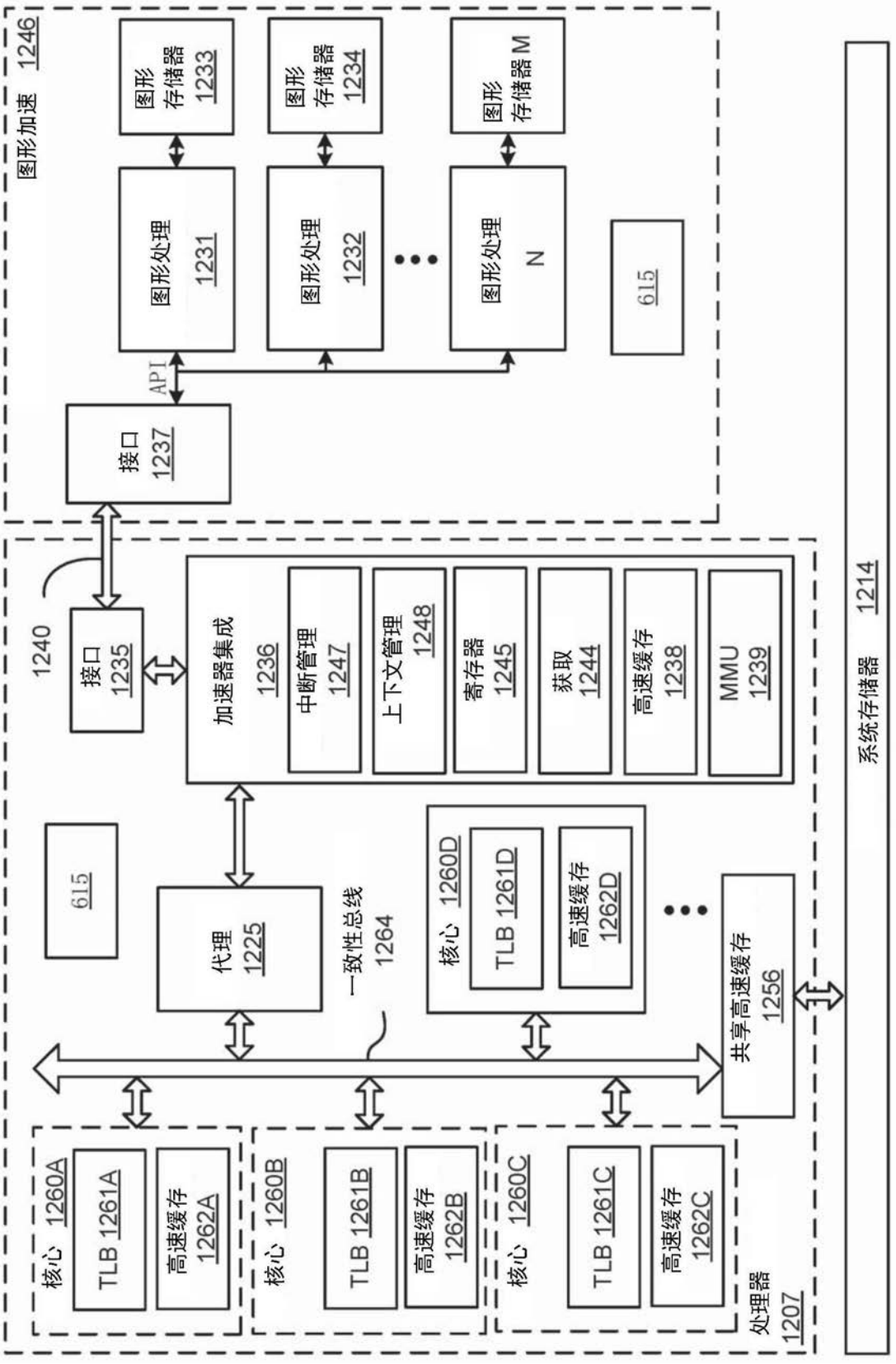

图形加速 1246
图形存储器 1233
图形存储器 1234
图形存储器 M
图形处理 1231
图形处理 1232
图形处理 N
615
API
接口 1237
1240
接口 1235
加速器集成 1236
中断管理 1247
上下文管理 1248
寄存器 1245
获取 1244
高速缓存 1238
MMU 1239
系统存储器 1214
615
代理 1225
一致性总线 1264
核心 1260D
TLB 1261D
高速缓存 1262D
共享高速缓存 1256
核心 1260A
TLB 1261A
高速缓存 1262A
核心 1260B
TLB 1261B
高速缓存 1262B
核心 1260C
TLB 1261C
高速缓存 1262C
处理器 1207

图12C

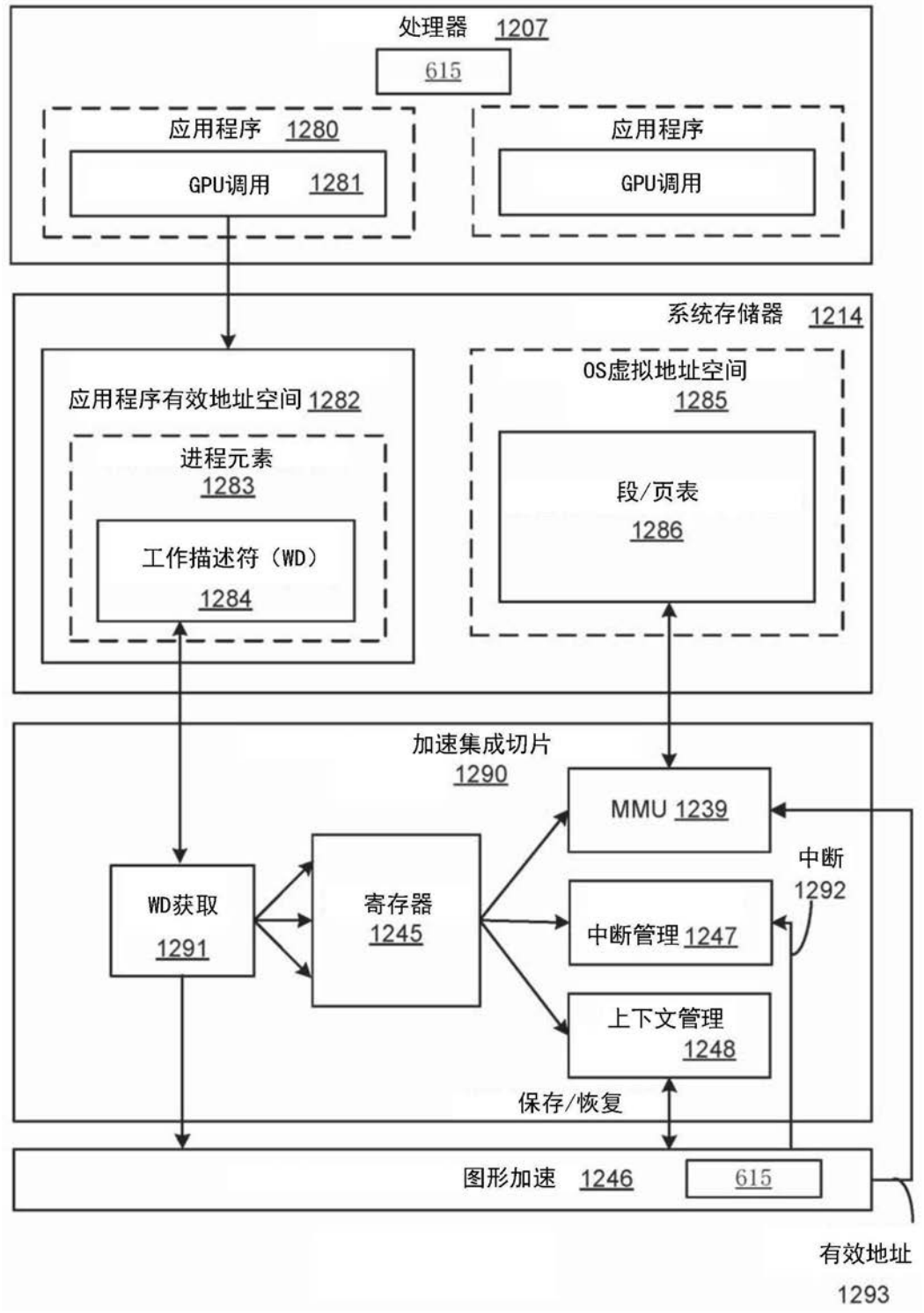

处理器 1207
615
应用程序 1280
GPU调用 1281
应用程序
GPU调用
系统存储器 1214
应用程序有效地址空间 1282
进程元素
1283
工作描述符（WD）
1284
OS虚拟地址空间
1285
段/页表
1286
加速集成切片
1290
MMU 1239
中断
1292
WD获取
1291
寄存器
1245
中断管理 1247
上下文管理
1248
保存/恢复
图形加速 1246
615
有效地址
1293

图12D

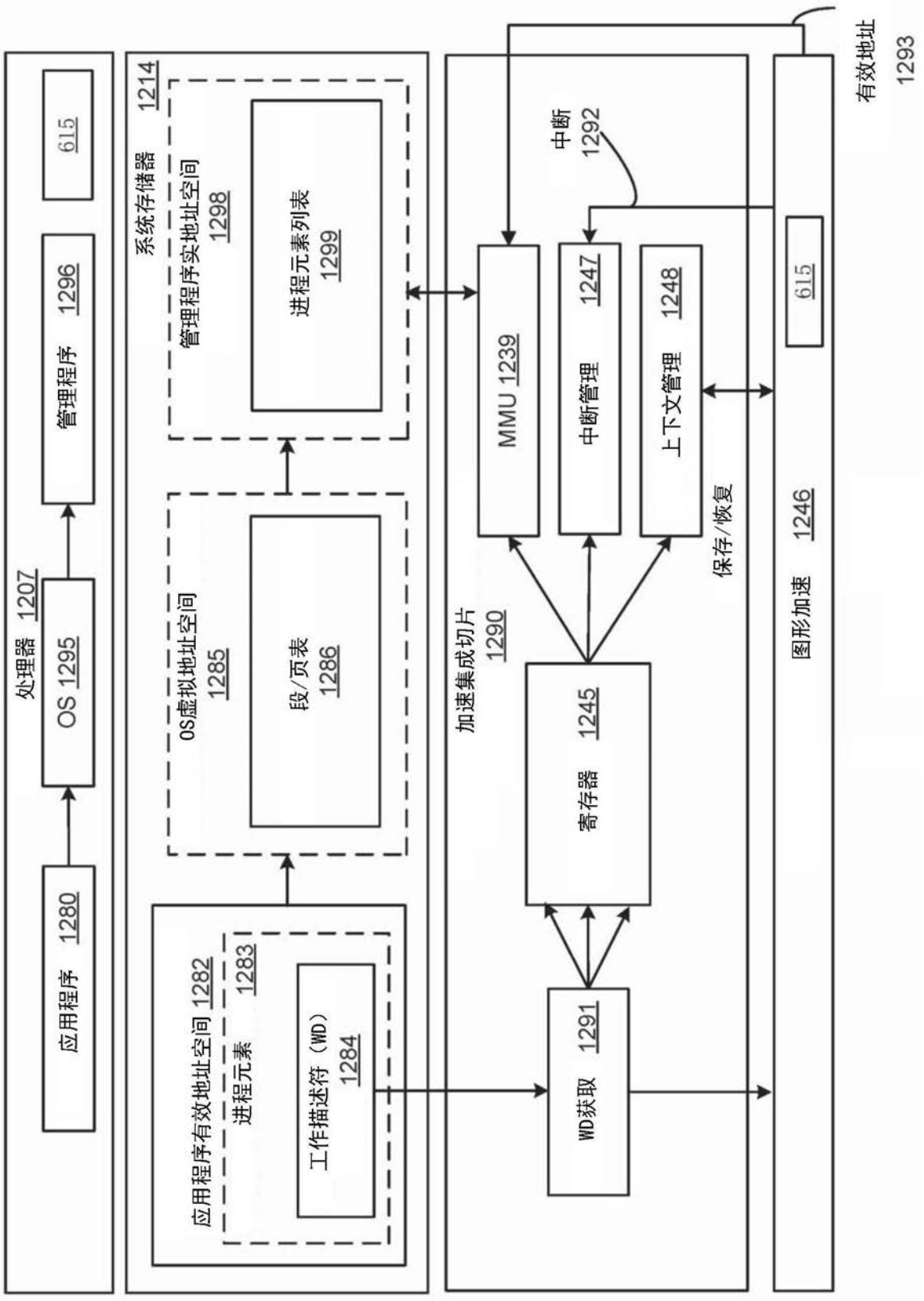

处理器 1207
应用程序 1280
OS 1295
管理程序 1296
615
系统存储器 1214
管理程序实地址空间 1298
进程元素列表 1299
OS虚拟地址空间 1285
段/页表 1286
应用程序有效地址空间 1282
进程元素 1283
工作描述符（WD） 1284
加速集成切片 1290
MMU 1239
中断管理 1247
上下文管理 1248
寄存器 1245
WD获取 1291
保存/恢复
中断 1292
有效地址 1293
图形加速 1246
615

图12E

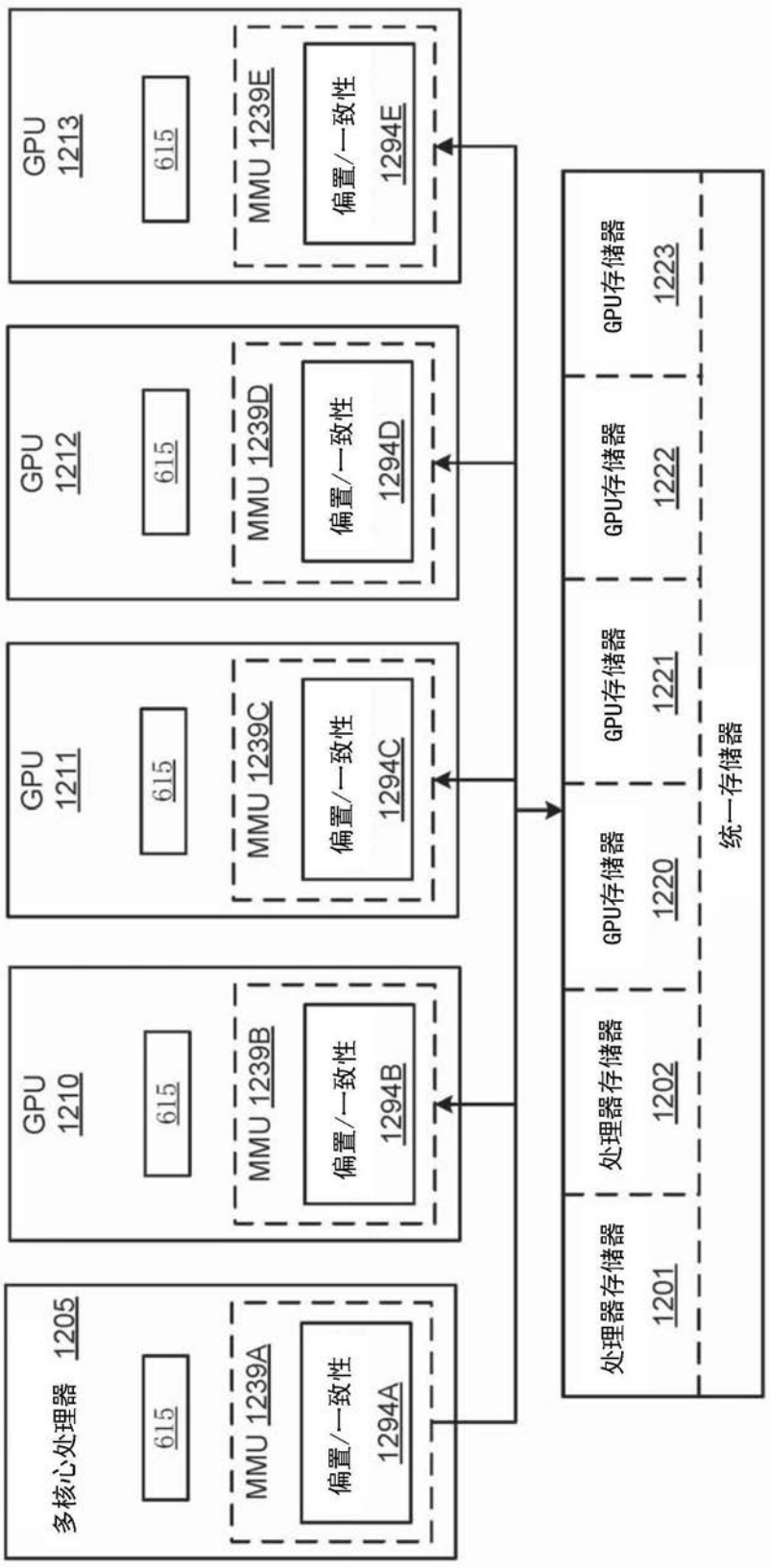

多核心处理器 1205
615
MMU 1239A
偏置/一致性 1294A
GPU 1210
615
MMU 1239B
偏置/一致性 1294B
GPU 1211
615
MMU 1239C
偏置/一致性 1294C
GPU 1212
615
MMU 1239D
偏置/一致性 1294D
GPU 1213
615
MMU 1239E
偏置/一致性 1294E
处理器存储器 1201
处理器存储器 1202
GPU存储器 1220
GPU存储器 1221
GPU存储器 1222
GPU存储器 1223
统一存储器

图12F

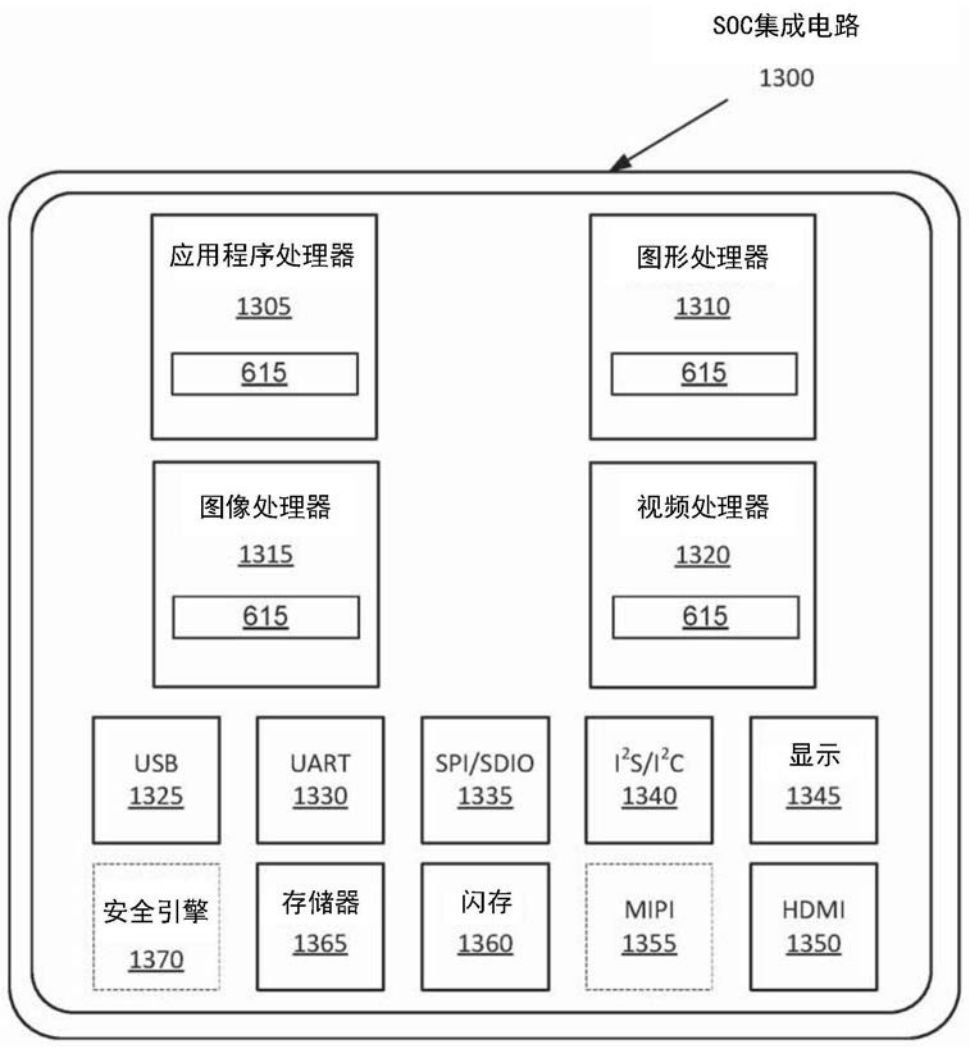

SOC集成电路
1300
应用程序处理器
1305
615
图形处理器
1310
615
图像处理器
1315
615
视频处理器
1320
615
USB
1325
UART
1330
SPI/SDIO
1335
I²S/I²C
1340
显示
1345
安全引擎
1370
存储器
1365
闪存
1360
MIPI
1355
HDMI
1350

图13

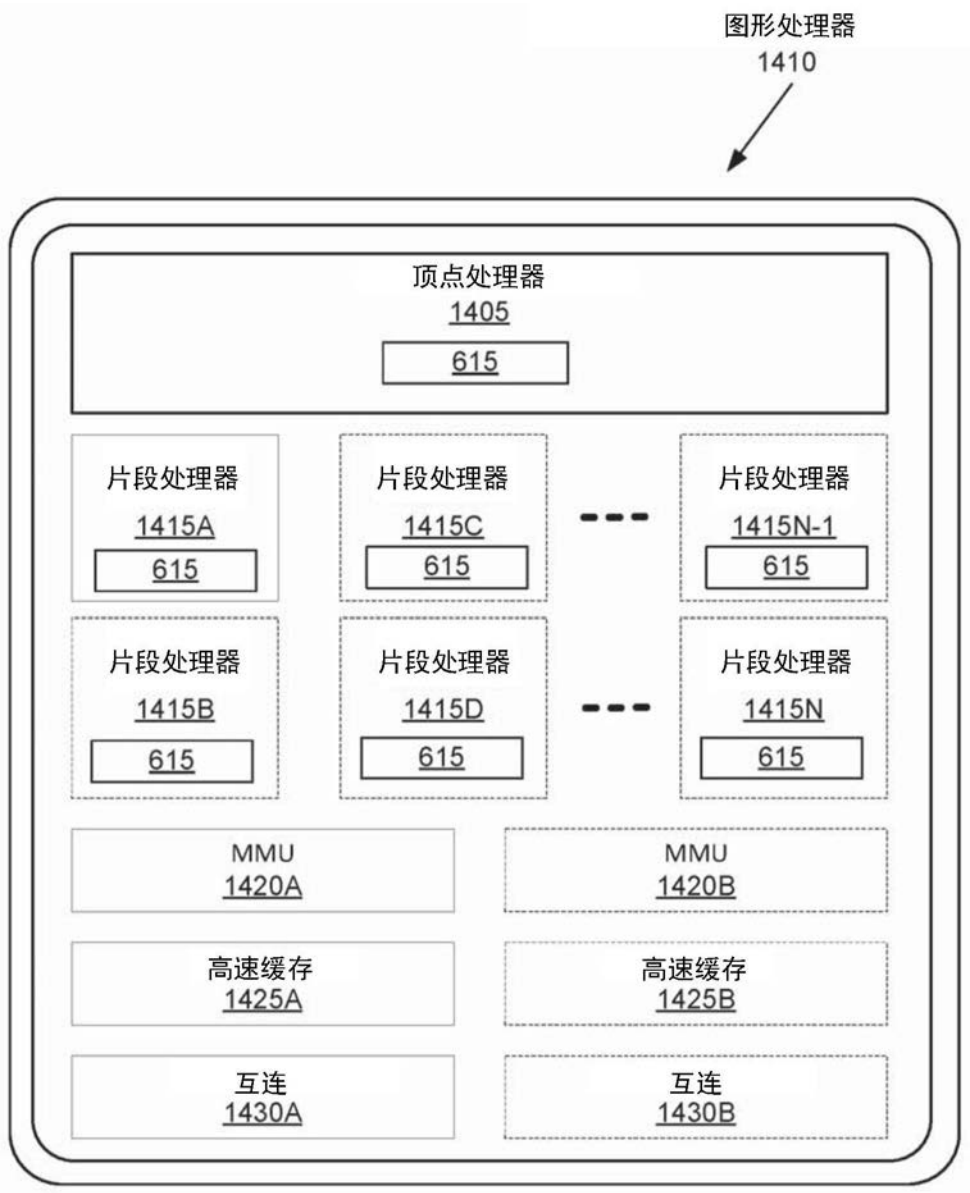

图形处理器
1410
顶点处理器
1405
615
片段处理器
1415A
615
片段处理器
1415C
615
片段处理器
1415N-1
615
片段处理器
1415B
615
片段处理器
1415D
615
片段处理器
1415N
615
MMU
1420A
MMU
1420B
高速缓存
1425A
高速缓存
1425B
互连
1430A
互连
1430B

图14A

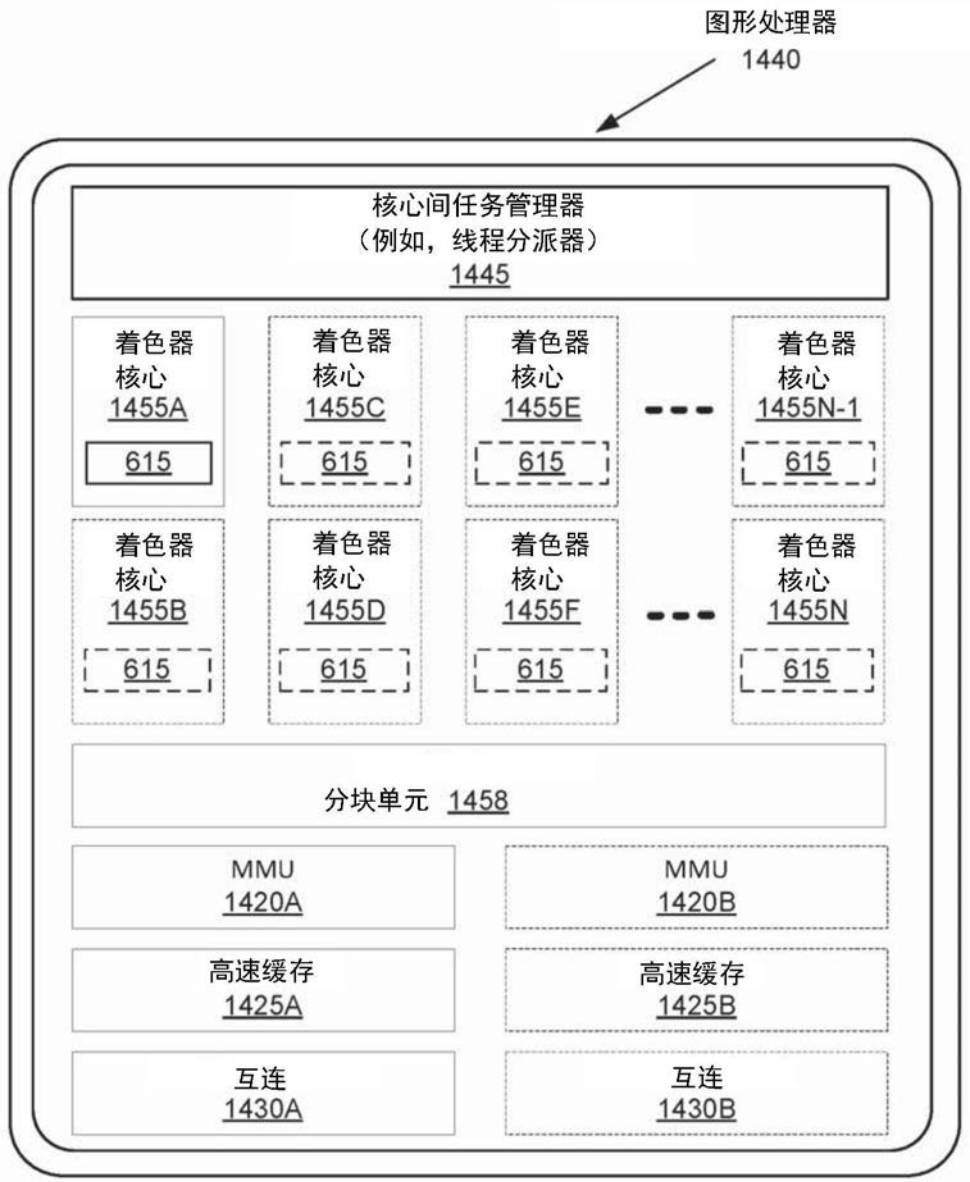

图形处理器
1440
核心间任务管理器
（例如，线程分派器）
1445
着色器
核心
1455A
615
着色器
核心
1455C
615
着色器
核心
1455E
615
着色器
核心
1455N-1
615
着色器
核心
1455B
615
着色器
核心
1455D
615
着色器
核心
1455F
615
着色器
核心
1455N
615
分块单元 1458
MMU
1420A
MMU
1420B
高速缓存
1425A
高速缓存
1425B
互连
1430A
互连
1430B

图14B

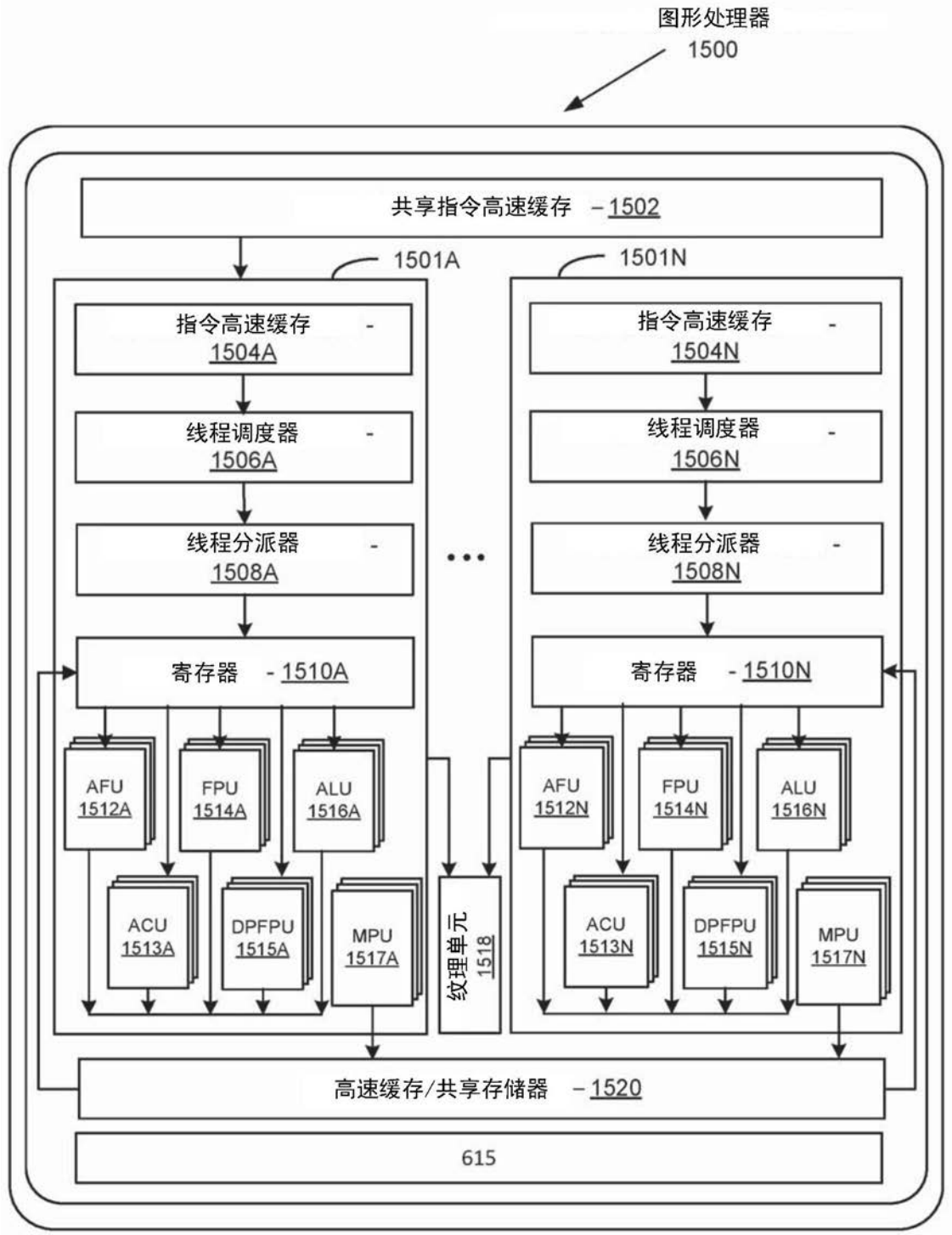

图形处理器
1500
共享指令高速缓存 -1502
1501A
1501N
指令高速缓存 - 1504A
线程调度器 - 1506A
线程分派器 - 1508A
寄存器 - 1510A
AFU 1512A
FPU 1514A
ALU 1516A
ACU 1513A
DPFPU 1515A
MPU 1517A
纹理单元 1518
指令高速缓存 - 1504N
线程调度器 - 1506N
线程分派器 - 1508N
寄存器 - 1510N
AFU 1512N
FPU 1514N
ALU 1516N
ACU 1513N
DPFPU 1515N
MPU 1517N
高速缓存/共享存储器 -1520
615

图15A

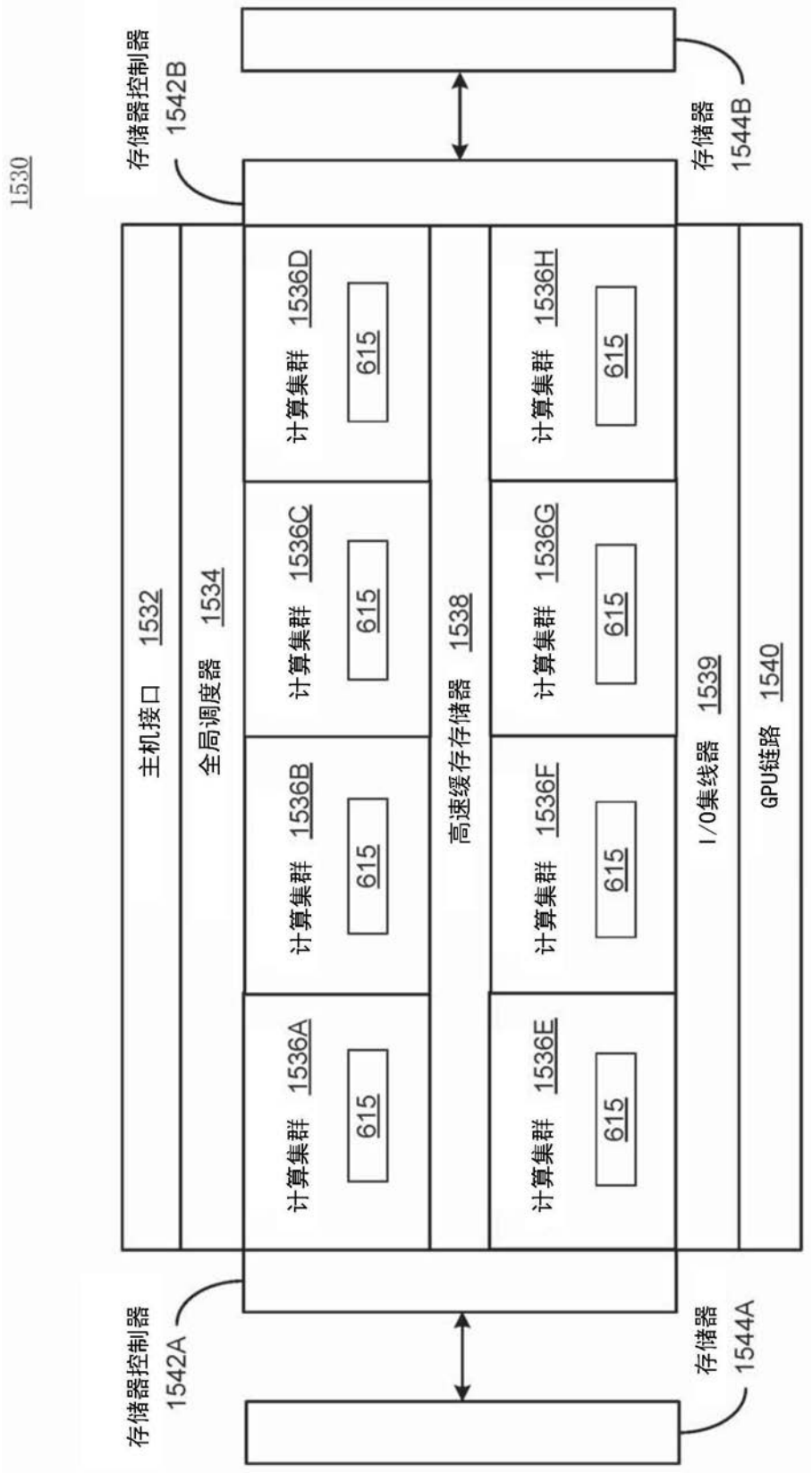

1530
存储器控制器 1542B
存储器 1544B
主机接口 1532
全局调度器 1534
计算集群 1536D
615
计算集群 1536H
615
计算集群 1536C
615
计算集群 1536G
615
高速缓存存储器 1538
计算集群 1536B
615
计算集群 1536F
615
I/O集线器 1539
计算集群 1536A
615
计算集群 1536E
615
GPU链路 1540
存储器控制器 1542A
存储器 1544A

图15B

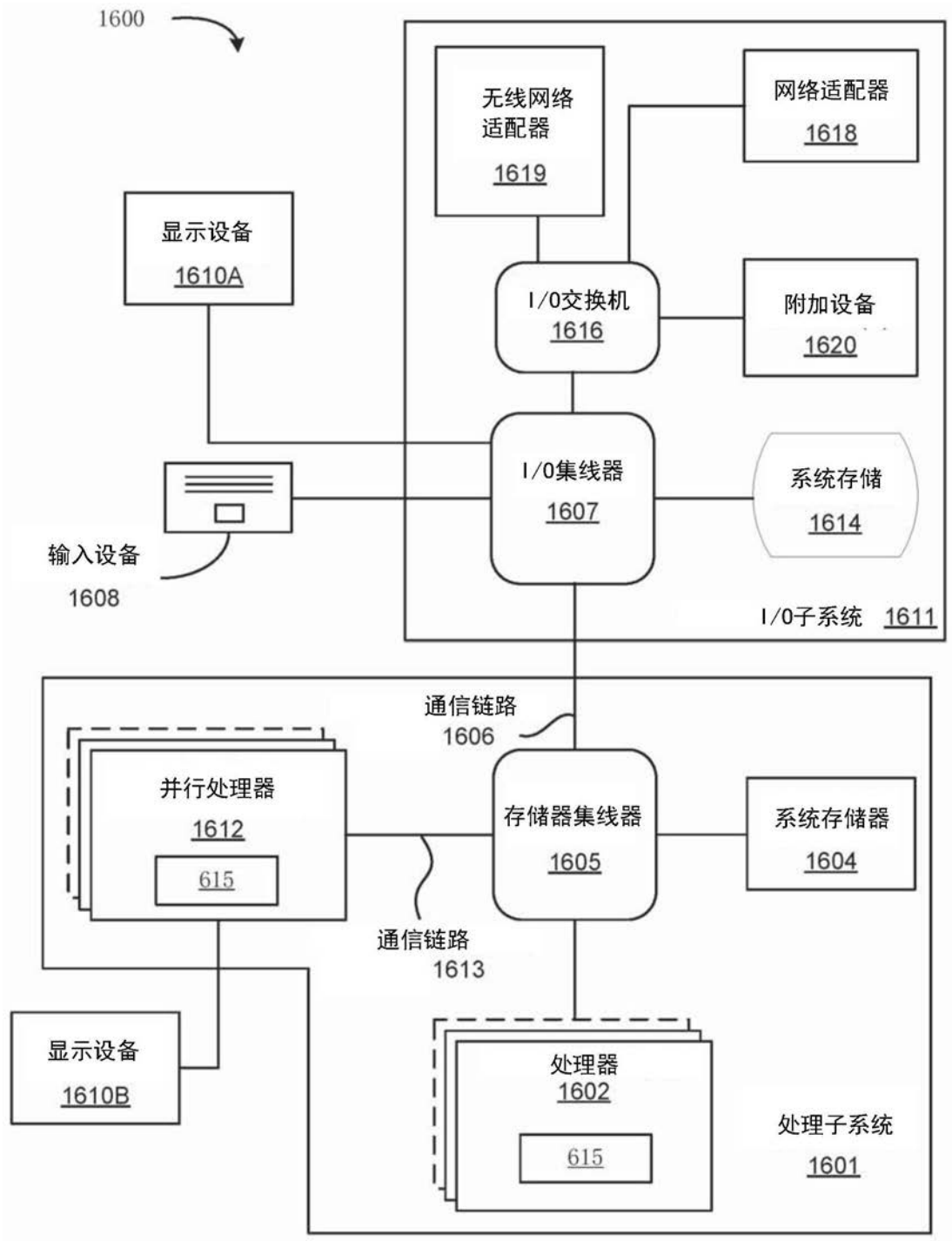

1600
无线网络适配器
1619
网络适配器
1618
显示设备
1610A
I/0交换机
1616
附加设备
1620
I/0集线器
1607
系统存储
1614
输入设备
1608
I/0子系统 1611
通信链路
1606
并行处理器
1612
615
存储器集线器
1605
系统存储器
1604
通信链路
1613
显示设备
1610B
处理器
1602
615
处理子系统
1601

图16

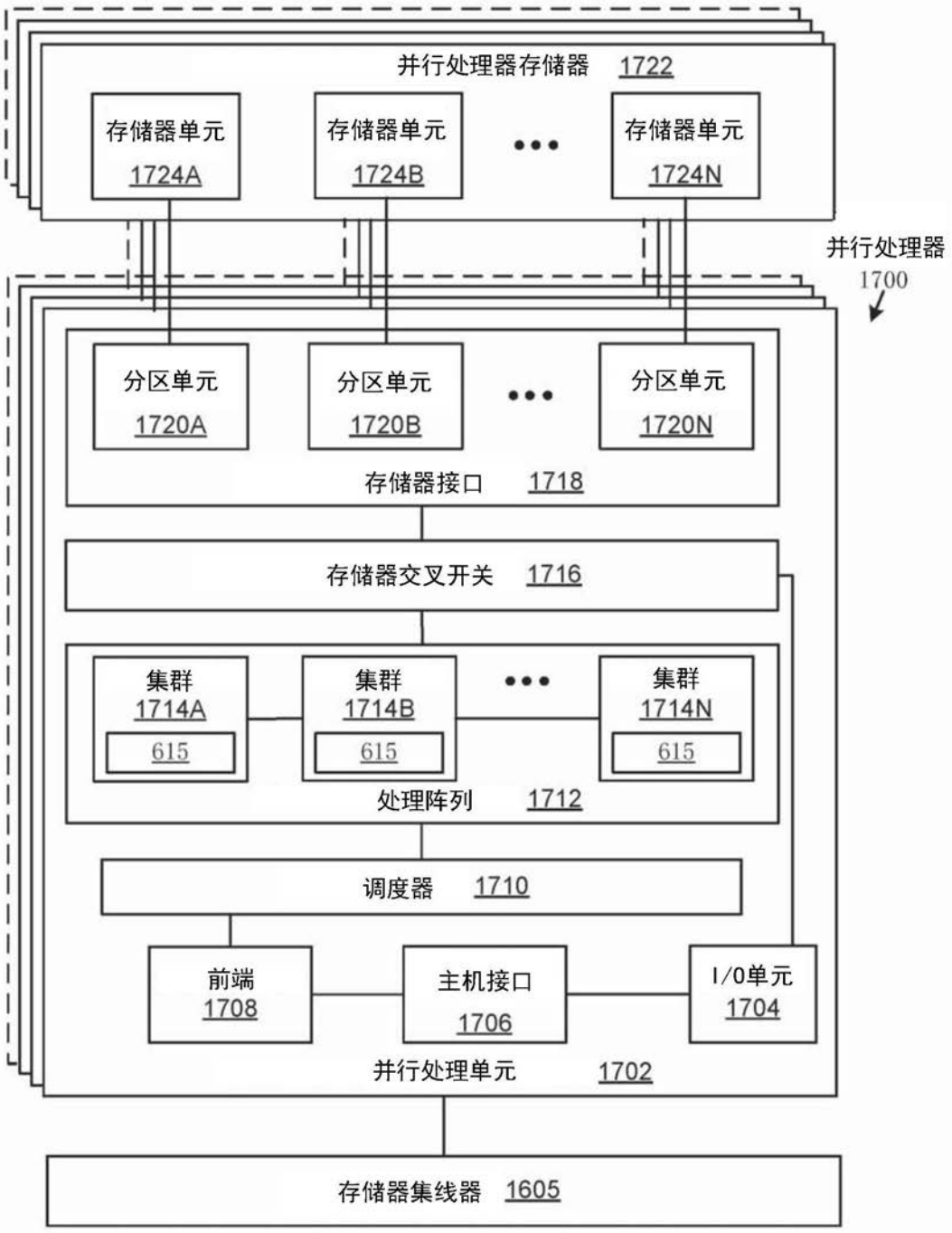

并行处理器存储器 1722
存储器单元 1724A
存储器单元 1724B
存储器单元 1724N
并行处理器 1700
分区单元 1720A
分区单元 1720B
分区单元 1720N
存储器接口 1718
存储器交叉开关 1716
集群 1714A
615
集群 1714B
615
集群 1714N
615
处理阵列 1712
调度器 1710
前端 1708
主机接口 1706
I/O单元 1704
并行处理单元 1702
存储器集线器 1605

图17A

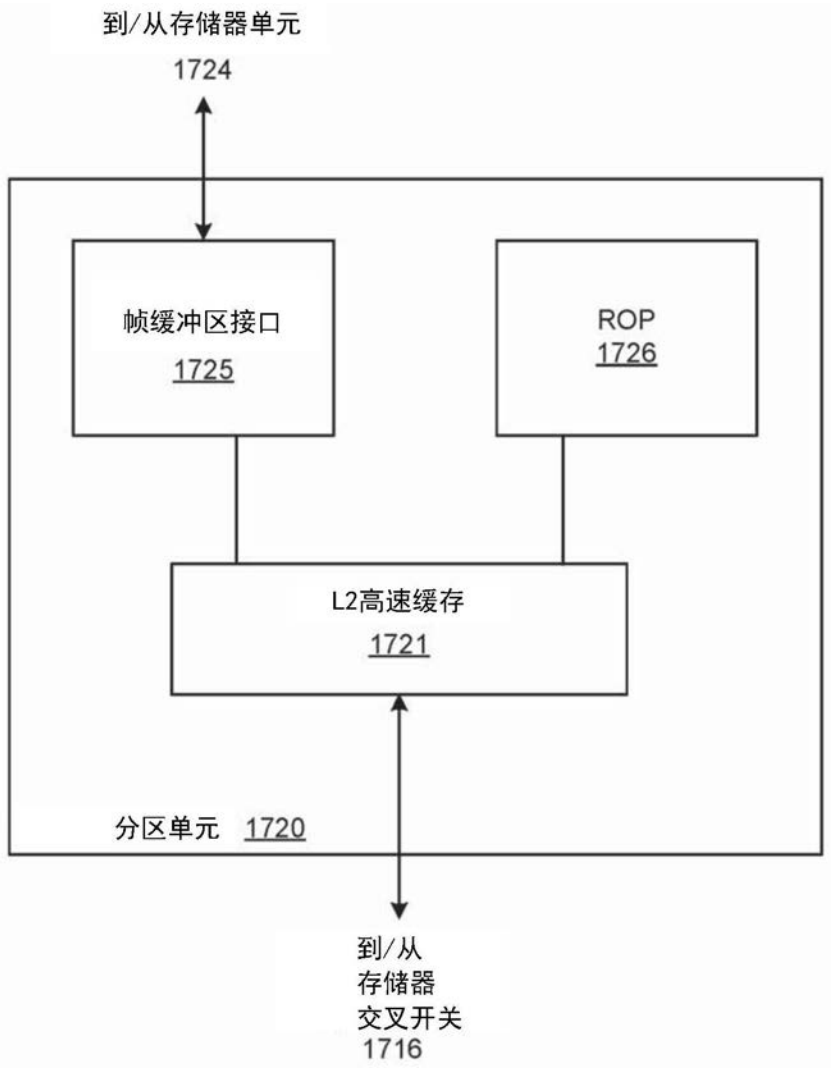

到/从存储器单元
1724
帧缓冲区接口
1725
ROP
1726
L2高速缓存
1721
分区单元 1720
到/从
存储器
交叉开关
1716

图17B

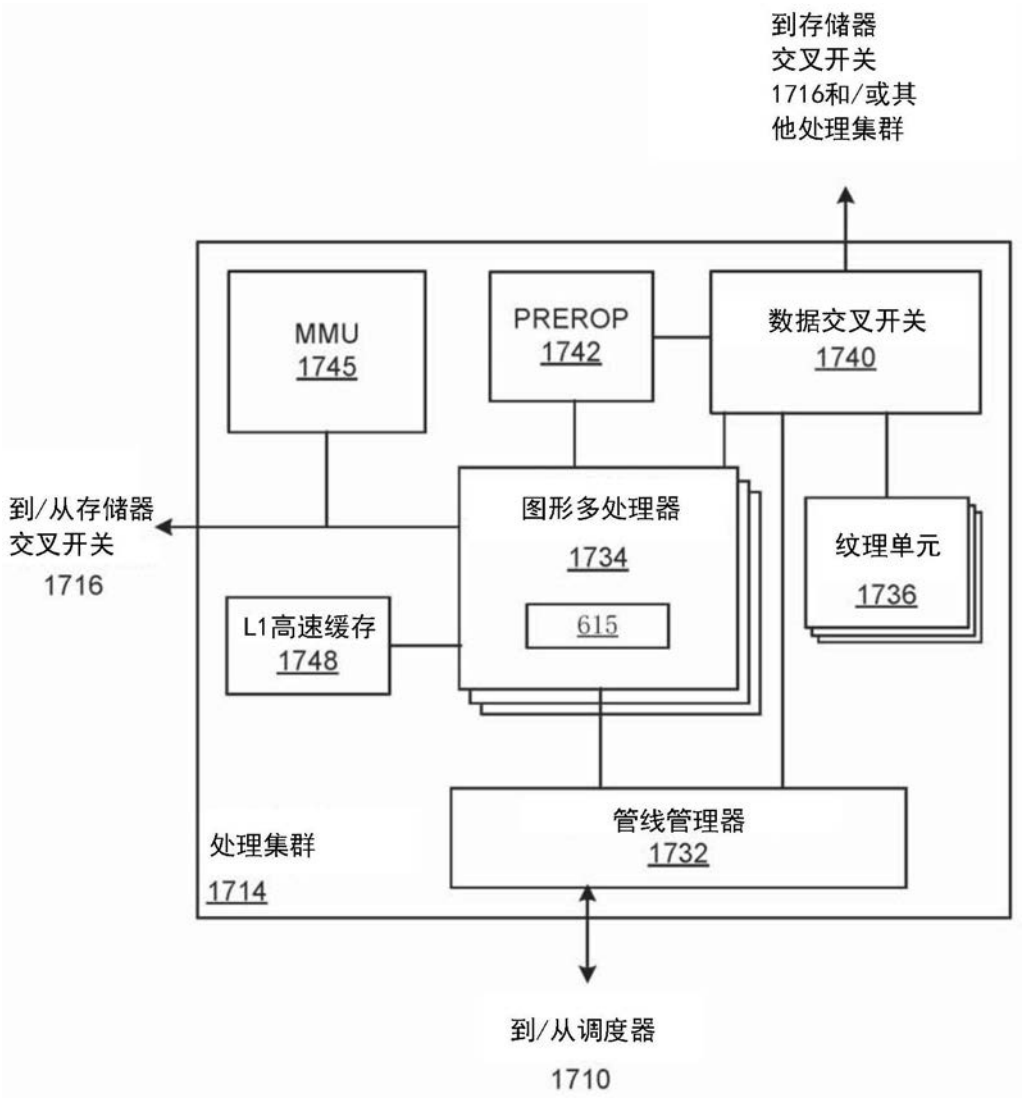

到存储器
交叉开关
1716和/或其
他处理集群
MMU
1745
PREROP
1742
数据交叉开关
1740
到/从存储器
交叉开关
1716
图形多处理器
1734
615
纹理单元
1736
L1高速缓存
1748
管线管理器
1732
处理集群
1714
到/从调度器
1710

图17C

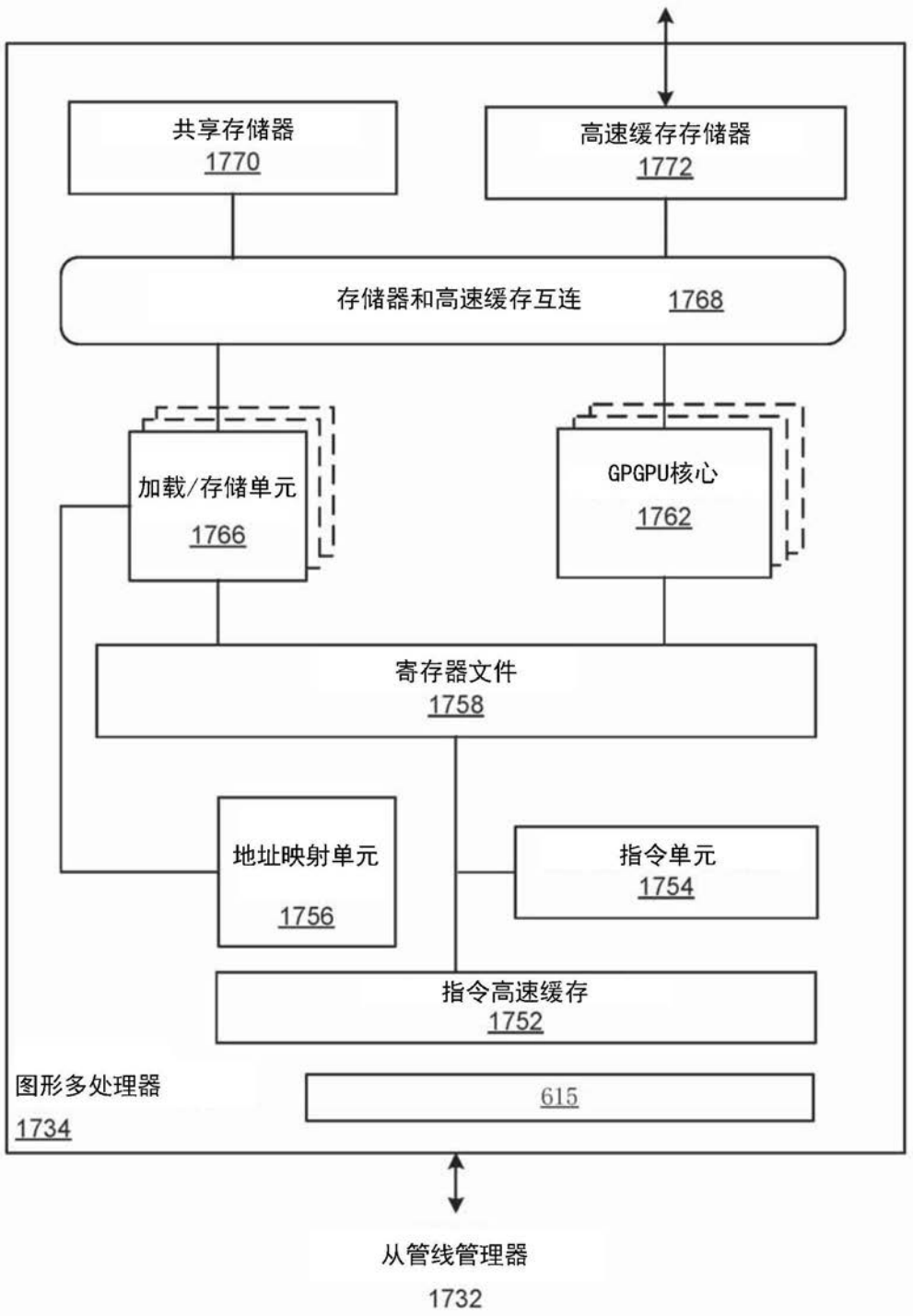

共享存储器
1770
高速缓存存储器
1772
存储器和高速缓存互连
1768
加载/存储单元
1766
GPGPU核心
1762
寄存器文件
1758
地址映射单元
1756
指令单元
1754
指令高速缓存
1752
图形多处理器
1734
615
从管线管理器
1732

图17D

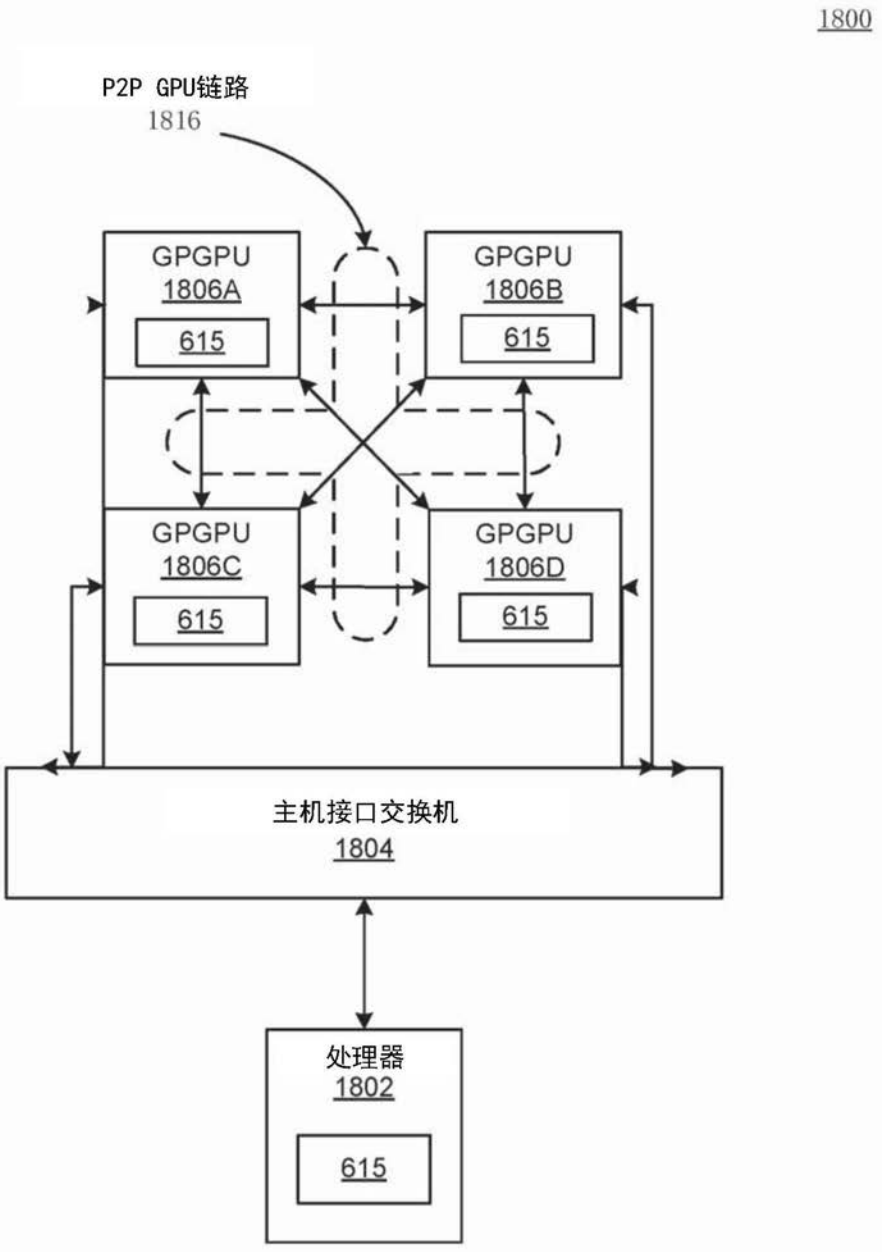

1800
P2P GPU链路
1816
GPGPU
1806A
615
GPGPU
1806B
615
GPGPU
1806C
615
GPGPU
1806D
615
主机接口交换机
1804
处理器
1802
615

图18

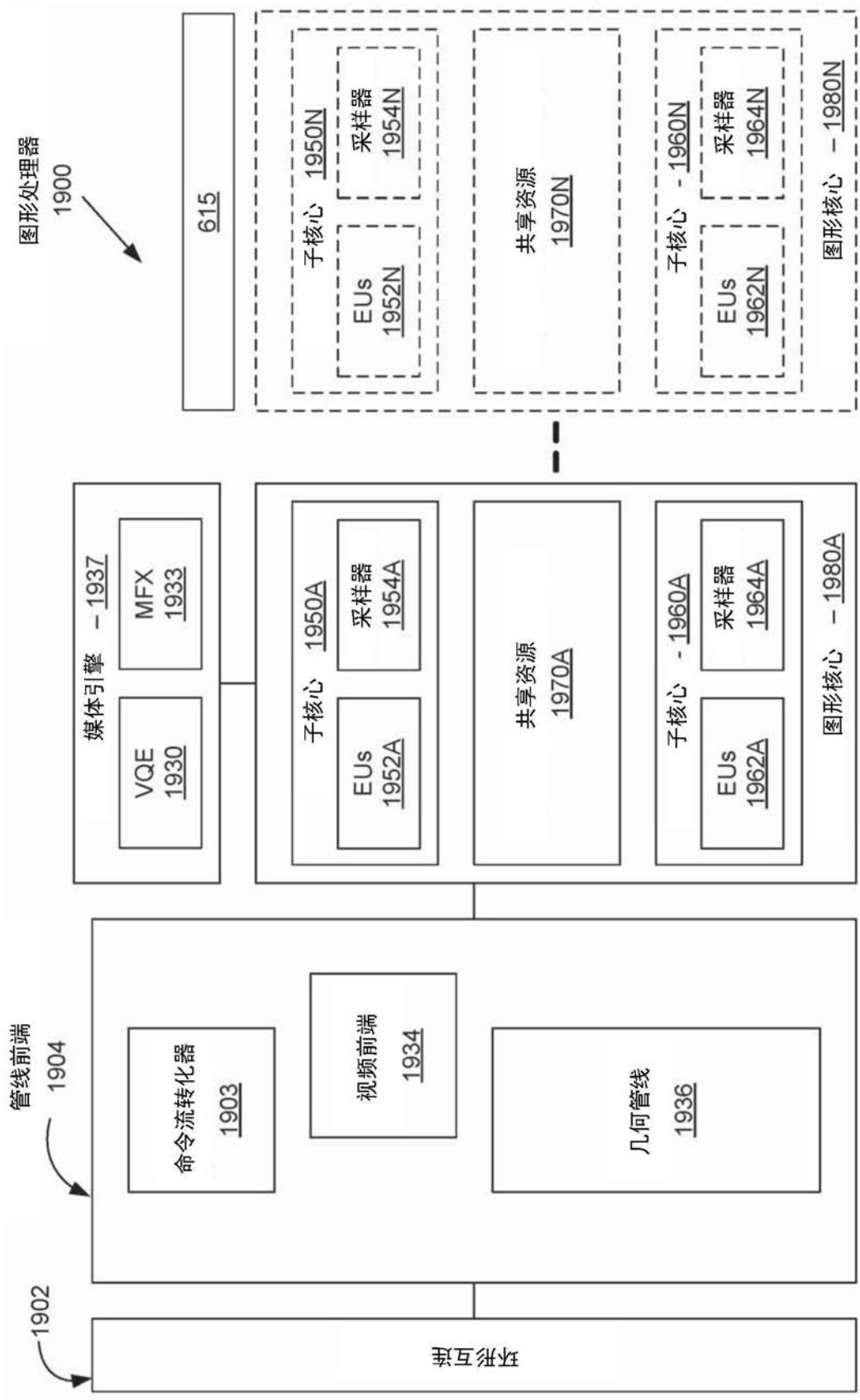

图形处理器
1900
615
子核心 1950N
EUs 1952N
采样器 1954N
共享资源 1970N
子核心 - 1960N
EUs 1962N
采样器 1964N
图形核心 - 1980N
媒体引擎 - 1937
VQE 1930
MFX 1933
子核心 1950A
EUs 1952A
采样器 1954A
共享资源 1970A
子核心 - 1960A
EUs 1962A
采样器 1964A
图形核心 - 1980A
管线前端 1904
命令流转化器 1903
视频前端 1934
几何管线 1936
1902
环形互连

图19

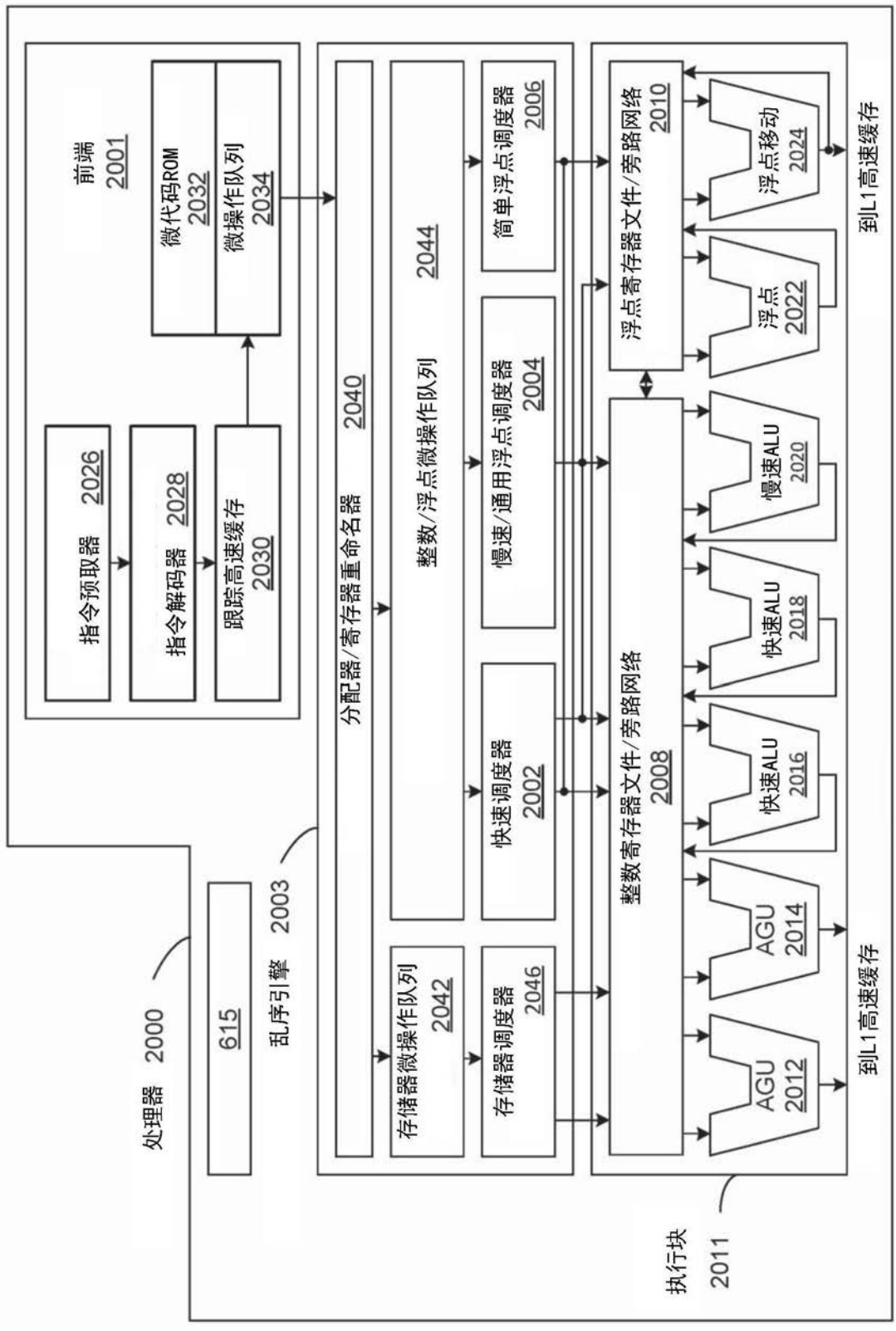

处理器 2000
615
乱序引擎 2003
前端 2001
指令预取器 2026
指令解码器 2028
跟踪高速缓存 2030
微代码ROM 2032
微操作队列 2034
分配器/寄存器重命名器 2040
存储器微操作队列 2042
存储器调度器 2046
整数/浮点微操作队列 2044
快速调度器 2002
慢速/通用浮点调度器 2004
简单浮点调度器 2006
整数寄存器文件/旁路网络 2008
浮点寄存器文件/旁路网络 2010
执行块 2011
AGU 2012
AGU 2014
快速ALU 2016
快速ALU 2018
慢速ALU 2020
浮点 2022
浮点移动 2024
到L1高速缓存
到L1高速缓存

图20

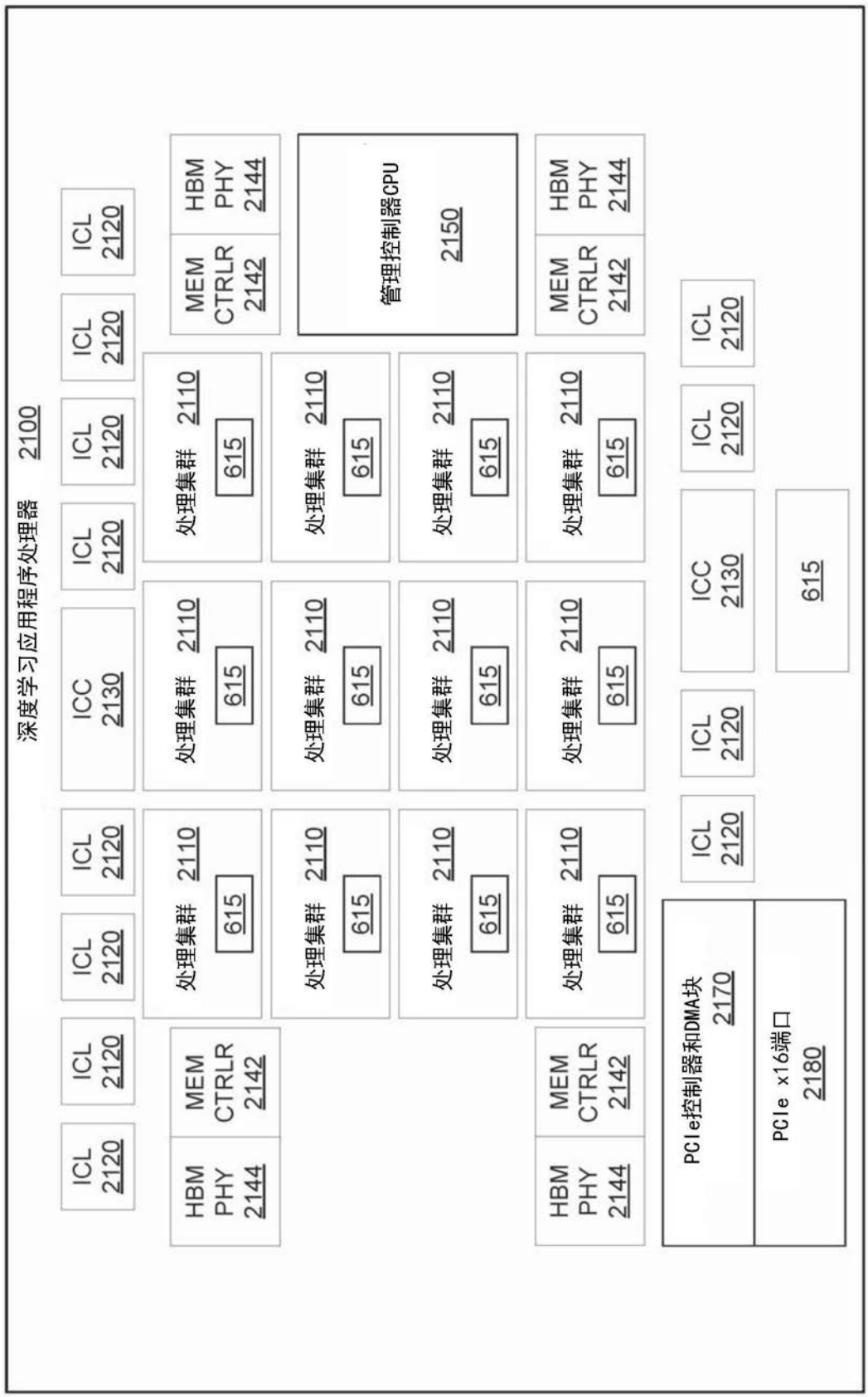

深度学习应用程序处理器 2100
ICL 2120
ICC 2130
处理集群 2110
615
MEM CTRLR 2142
HBM PHY 2144
管理控制器CPU 2150
PCIe控制器和DMA块 2170
PCIe x16端口 2180

图21

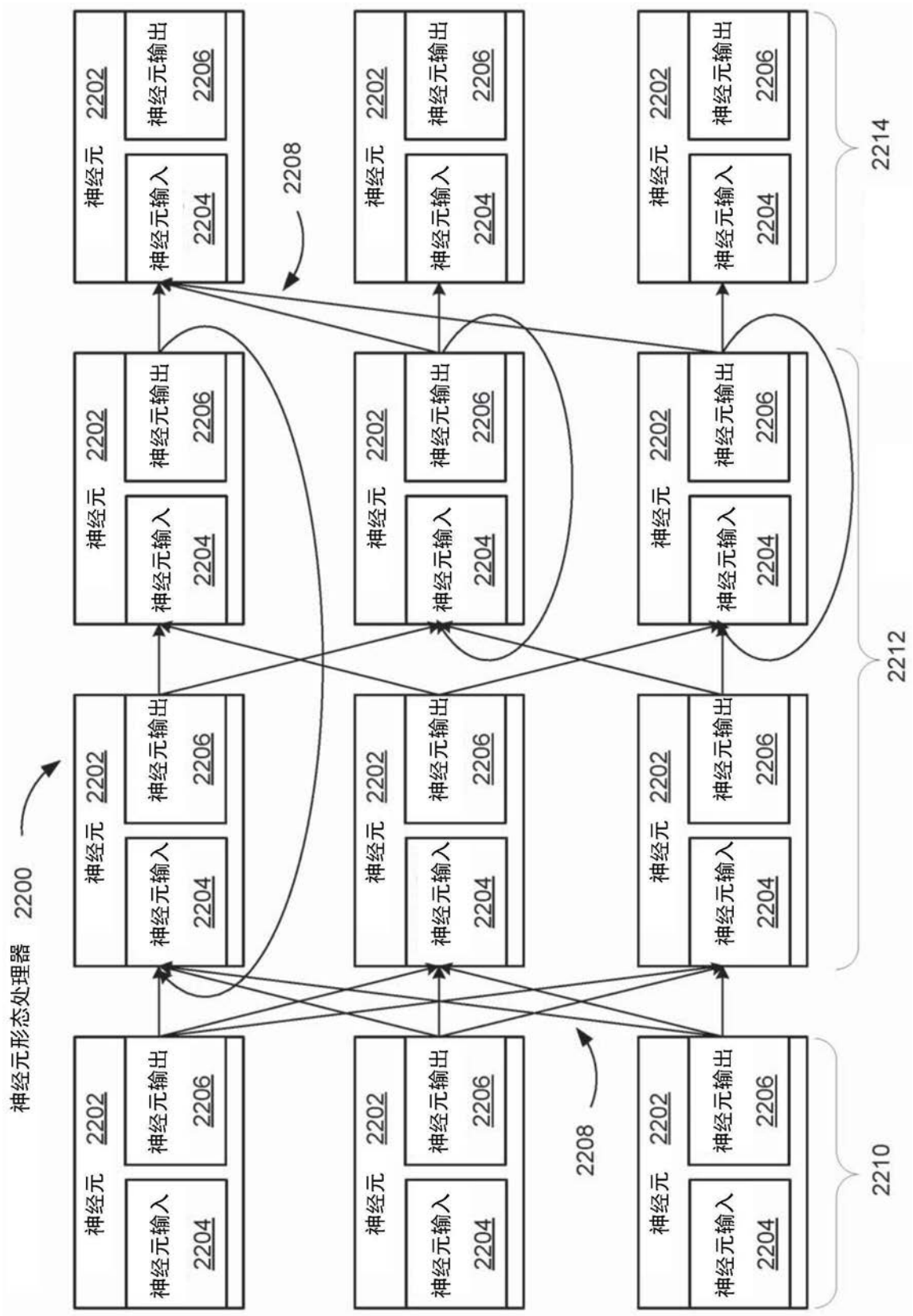

神经元形态处理器 2200
神经元 2202
神经元输入 2204
神经元输出 2206
2208
2208
2210
2212
2214

图22

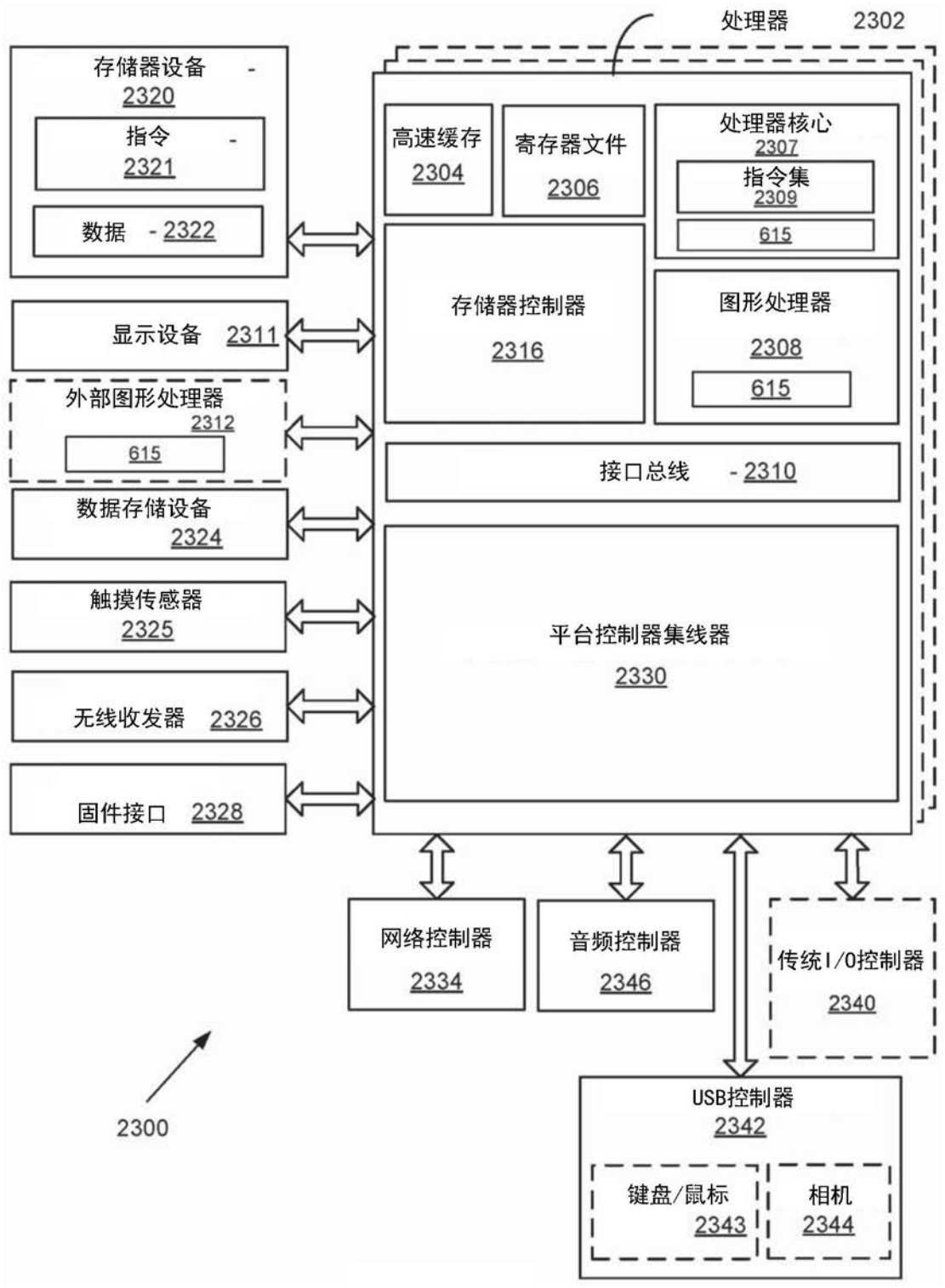

处理器 2302
存储器设备 2320
指令 2321
数据 2322
显示设备 2311
外部图形处理器 2312
615
数据存储设备 2324
触摸传感器 2325
无线收发器 2326
固件接口 2328
高速缓存 2304
寄存器文件 2306
处理器核心 2307
指令集 2309
615
存储器控制器 2316
图形处理器 2308
615
接口总线 2310
平台控制器集线器 2330
网络控制器 2334
音频控制器 2346
传统I/O控制器 2340
USB控制器 2342
键盘/鼠标 2343
相机 2344
2300

图23

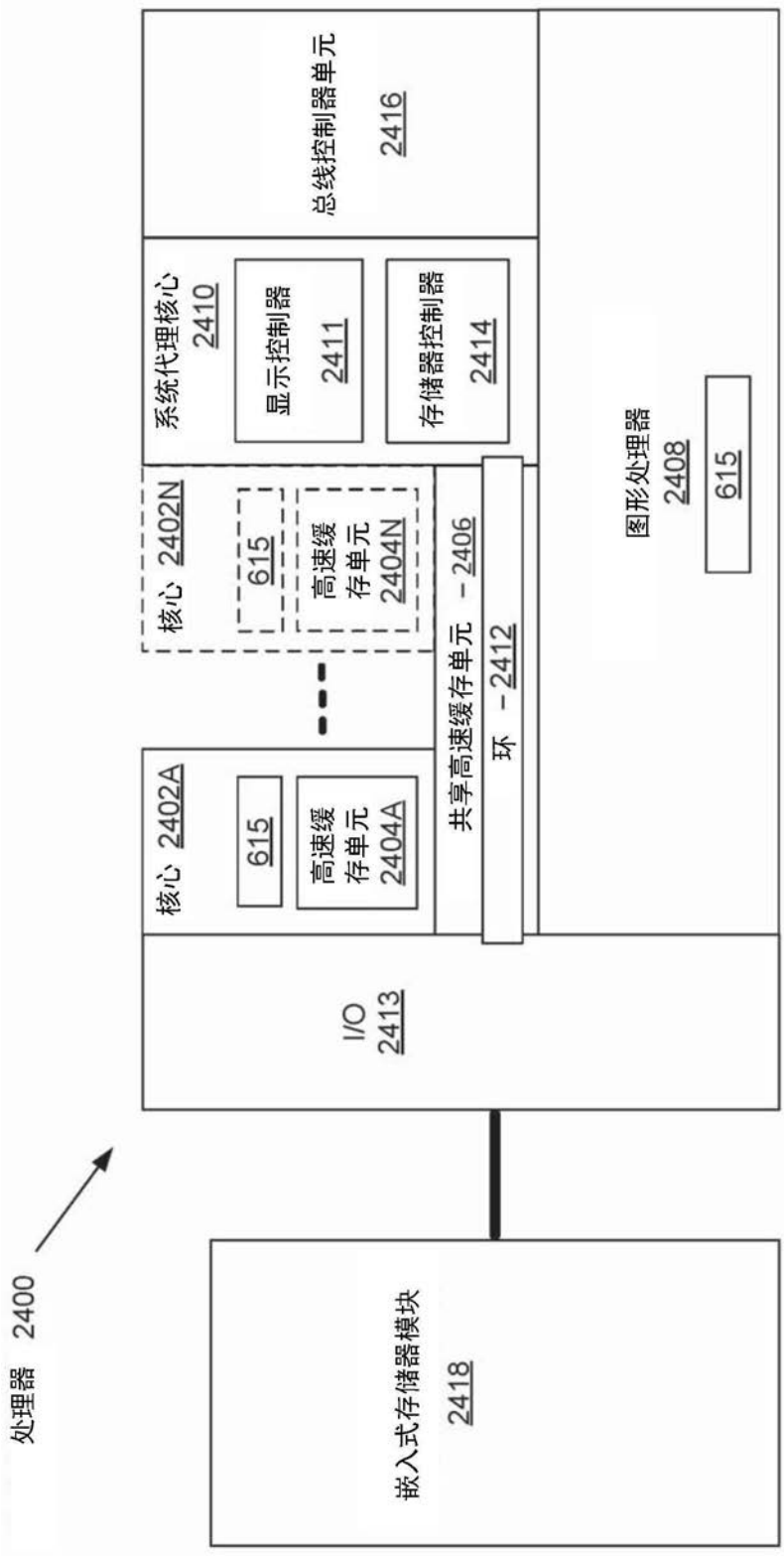

处理器 2400
嵌入式存储器模块 2418
I/O 2413
核心 2402A
615
高速缓存单元 2404A
核心 2402N
615
高速缓存单元 2404N
共享高速缓存单元 - 2406
环 - 2412
图形处理器 2408
615
系统代理核心 2410
显示控制器 2411
存储器控制器 2414
总线控制器单元 2416

图24

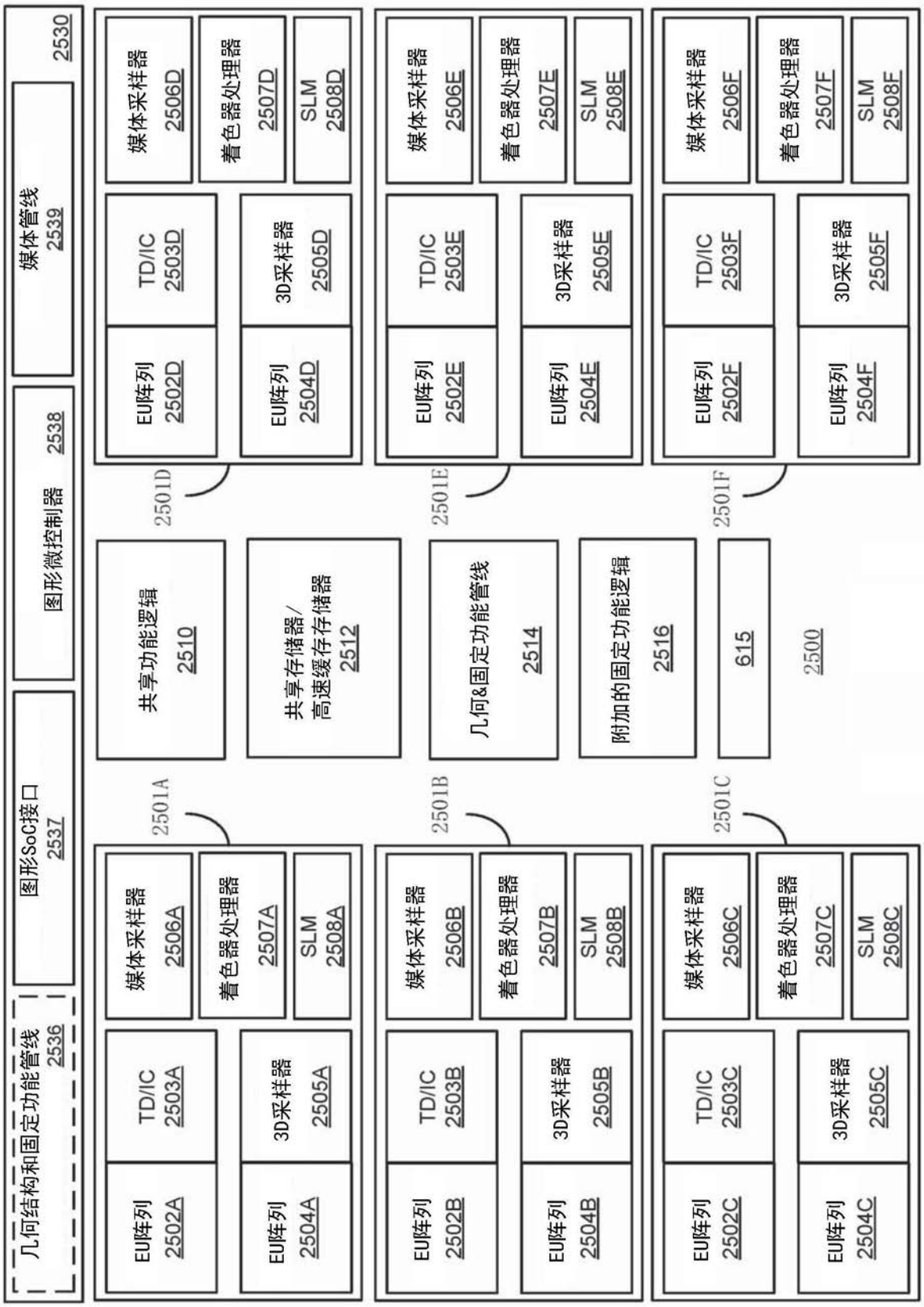

几何结构和固定功能管线 2536
图形SoC接口 2537
图形微控制器 2538
媒体管线 2539
2530
2500
2501A
2501B
2501C
2501D
2501E
2501F
EU阵列 2502A
TD/IC 2503A
媒体采样器 2506A
着色器处理器 2507A
SLM 2508A
EU阵列 2504A
3D采样器 2505A
EU阵列 2502B
TD/IC 2503B
媒体采样器 2506B
着色器处理器 2507B
SLM 2508B
EU阵列 2504B
3D采样器 2505B
EU阵列 2502C
TD/IC 2503C
媒体采样器 2506C
着色器处理器 2507C
SLM 2508C
EU阵列 2504C
3D采样器 2505C
EU阵列 2502D
TD/IC 2503D
媒体采样器 2506D
着色器处理器 2507D
SLM 2508D
EU阵列 2504D
3D采样器 2505D
EU阵列 2502E
TD/IC 2503E
媒体采样器 2506E
着色器处理器 2507E
SLM 2508E
EU阵列 2504E
3D采样器 2505E
EU阵列 2502F
TD/IC 2503F
媒体采样器 2506F
着色器处理器 2507F
SLM 2508F
EU阵列 2504F
3D采样器 2505F
共享功能逻辑 2510
共享存储器/ 高速缓存存储器 2512
几何&固定功能管线 2514
附加的固定功能逻辑 2516
615

图25

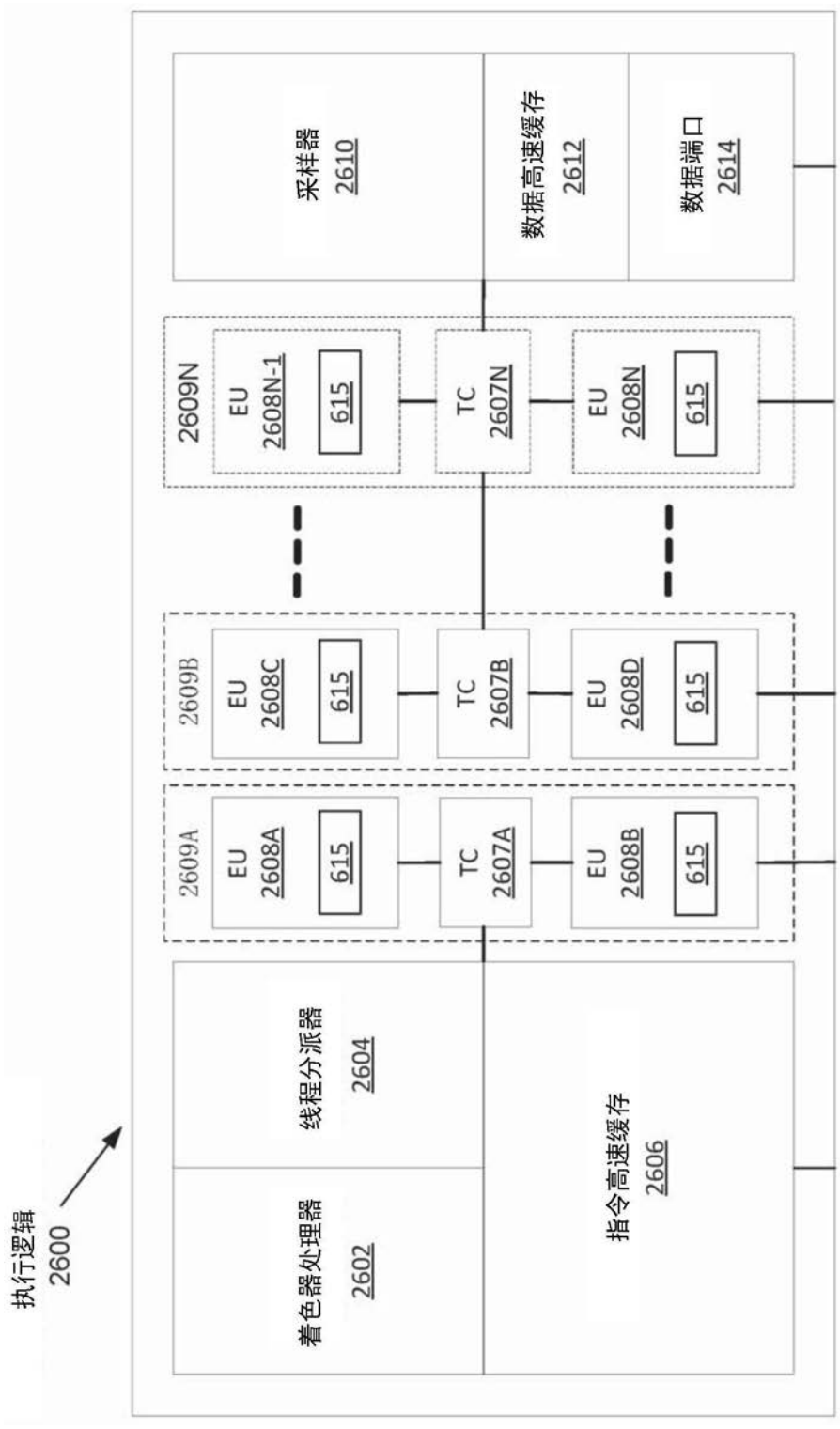

执行逻辑
2600
着色器处理器
2602
线程分派器
2604
指令高速缓存
2606
采样器
2610
数据高速缓存
2612
数据端口
2614
2609A
EU 2608A
615
TC 2607A
EU 2608B
615
2609B
EU 2608C
615
TC 2607B
EU 2608D
615
2609N
EU 2608N-1
615
TC 2607N
EU 2608N
615

图26A

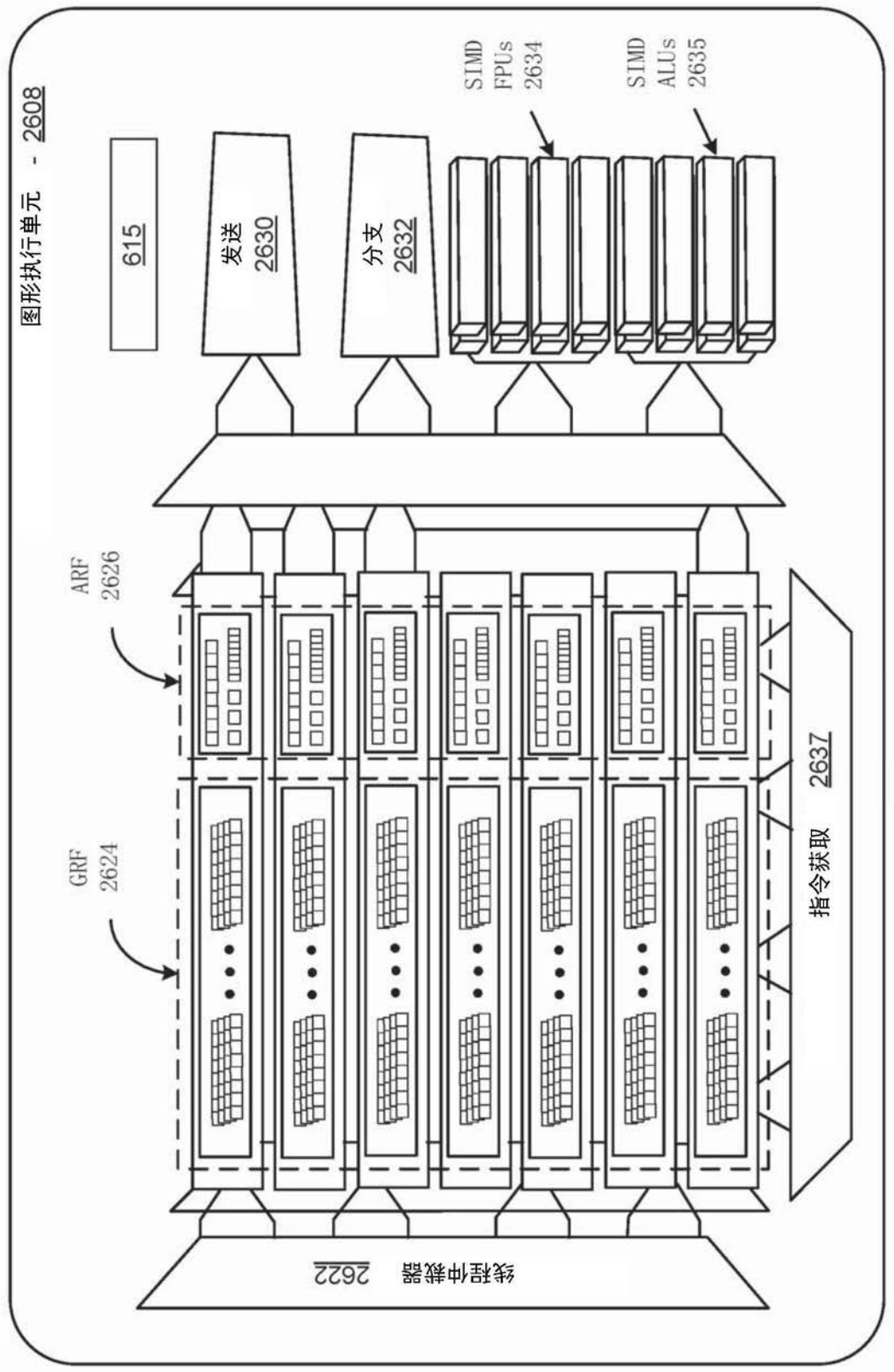

图形执行单元 - 2608
615
发送 2630
分支 2632
SIMD FPUs 2634
SIMD ALUs 2635
ARF 2626
GRF 2624
指令获取 2637
线程仲裁器 2622

图26B

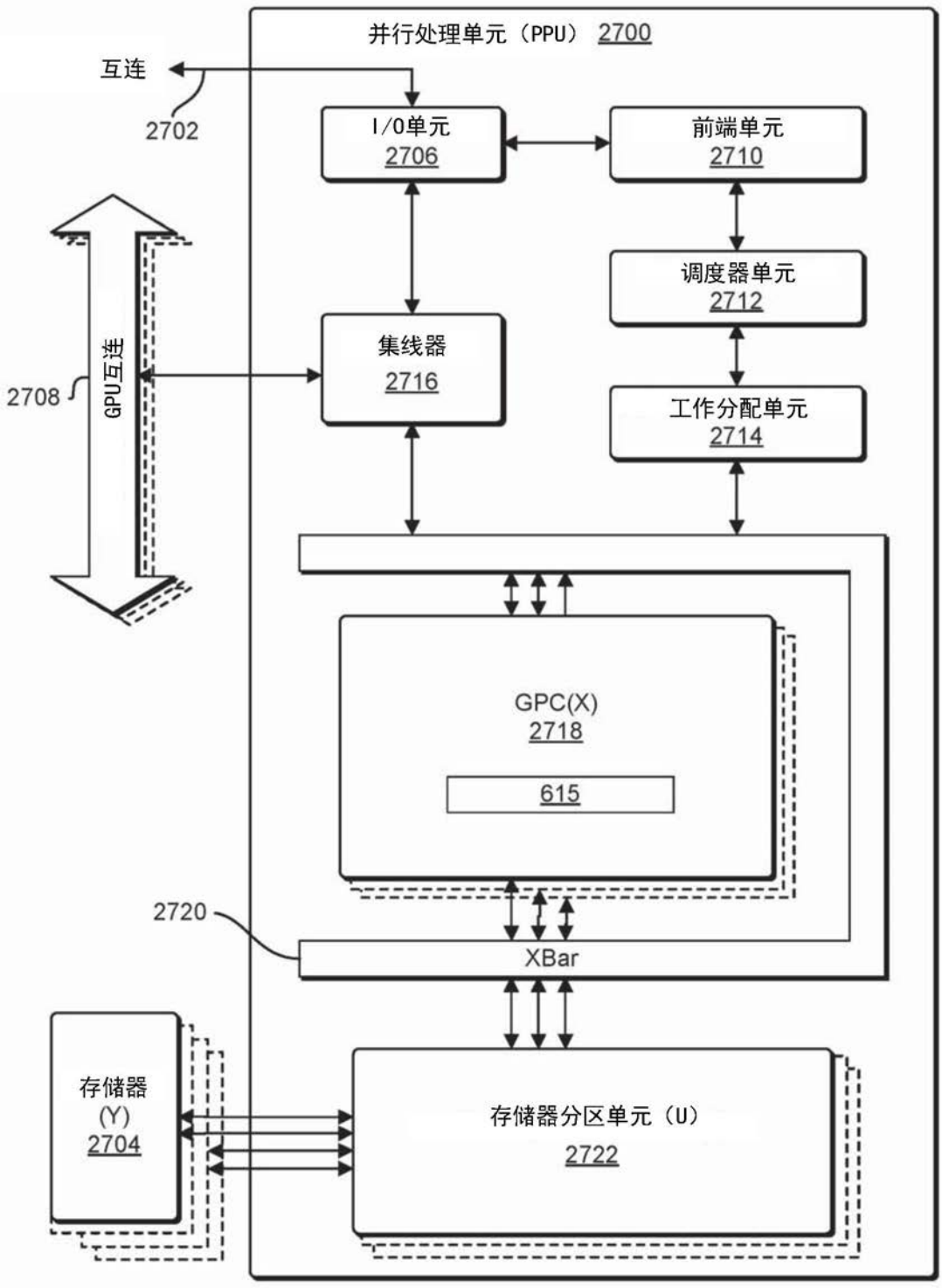

并行处理单元（PPU） 2700
互连
2702
I/O单元
2706
前端单元
2710
调度器单元
2712
集线器
2716
工作分配单元
2714
2708
GPU互连
GPC(X)
2718
615
2720
XBar
存储器
(Y)
2704
存储器分区单元（U）
2722

图27

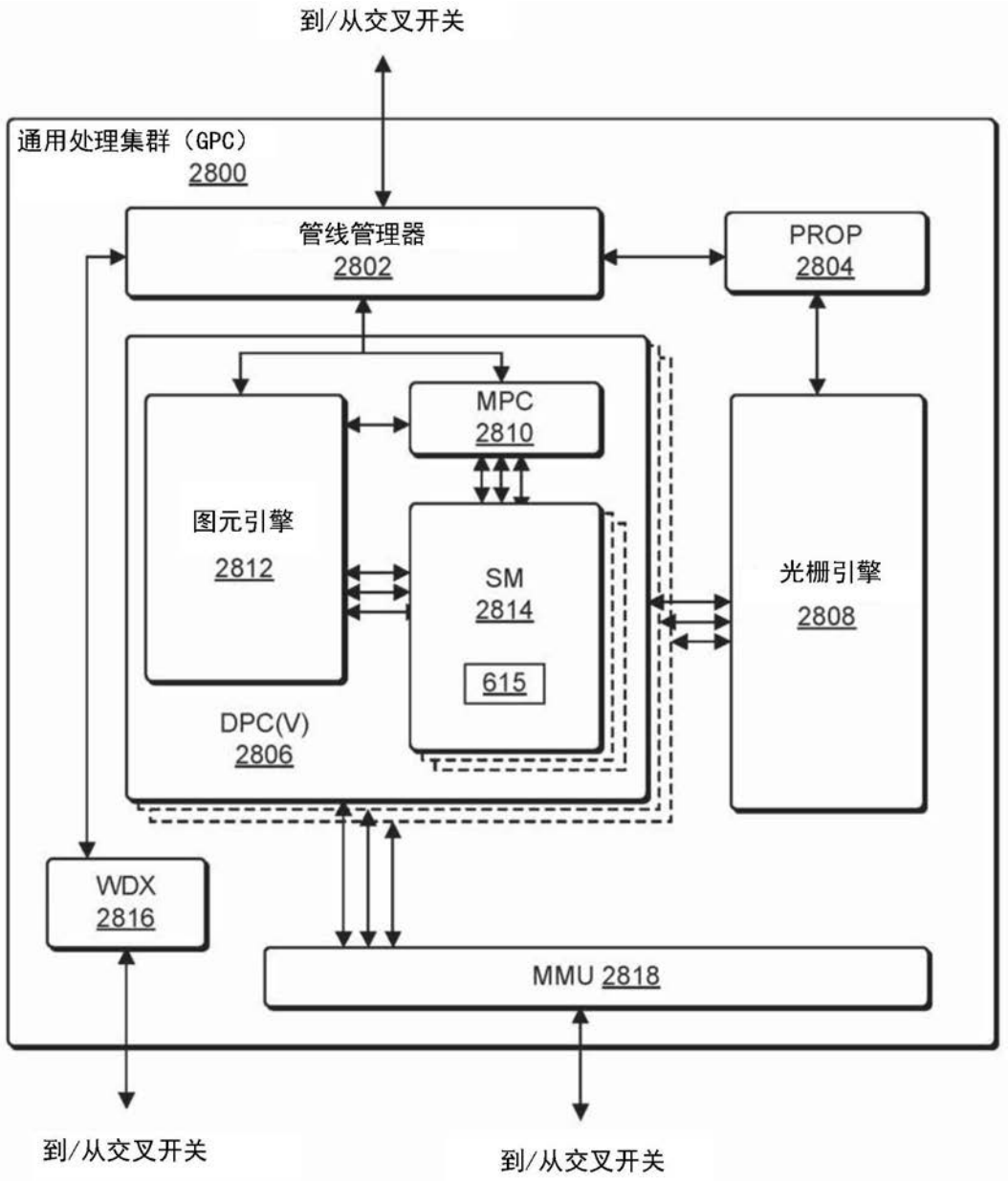

到/从交叉开关
通用处理集群（GPC）
2800
管线管理器
2802
PROP
2804
MPC
2810
图元引擎
2812
SM
2814
615
光栅引擎
2808
DPC(V)
2806
WDX
2816
MMU 2818
到/从交叉开关
到/从交叉开关

图28

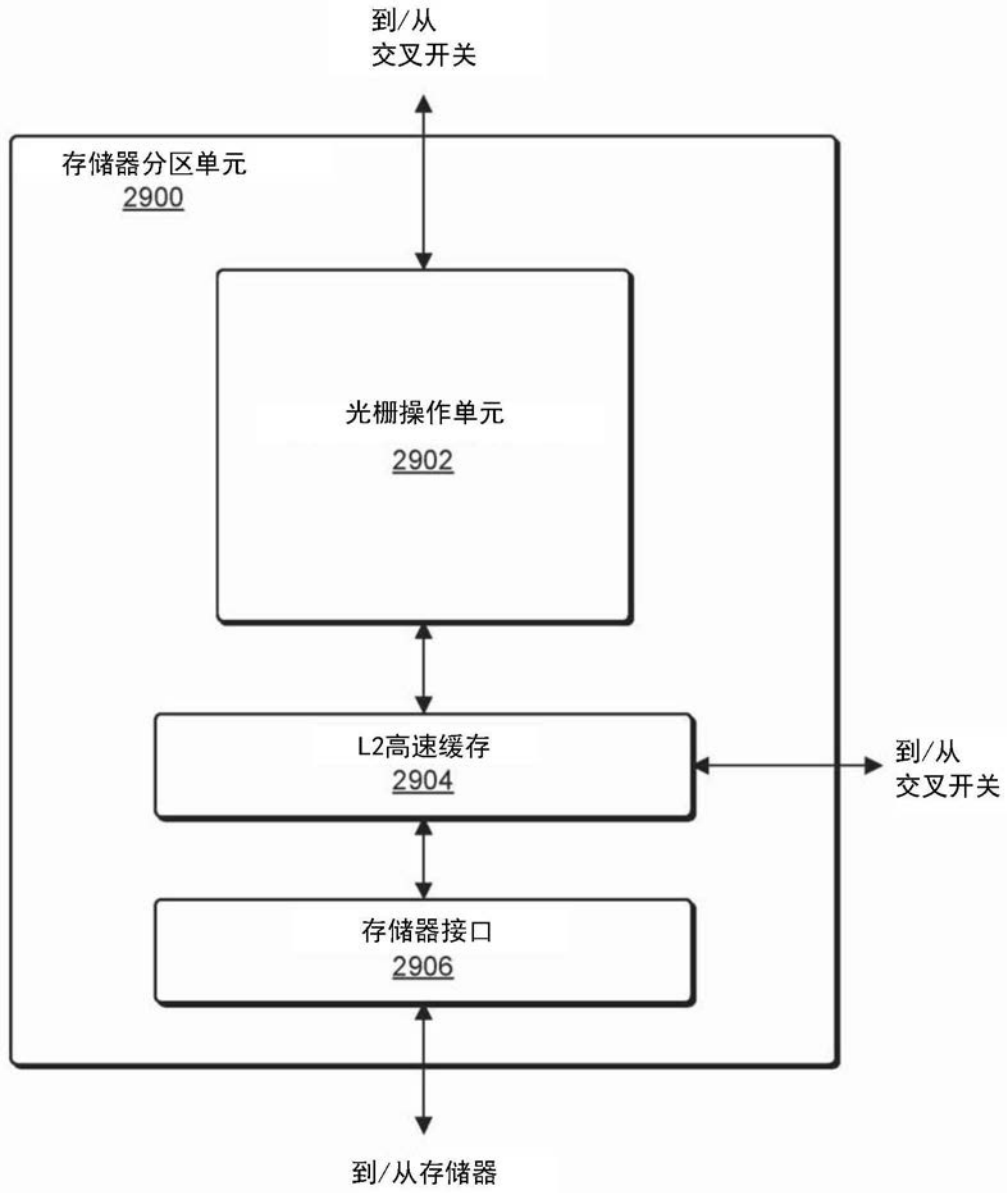

到/从
交叉开关
存储器分区单元
2900
光栅操作单元
2902
L2高速缓存
2904
到/从
交叉开关
存储器接口
2906
到/从存储器

图29

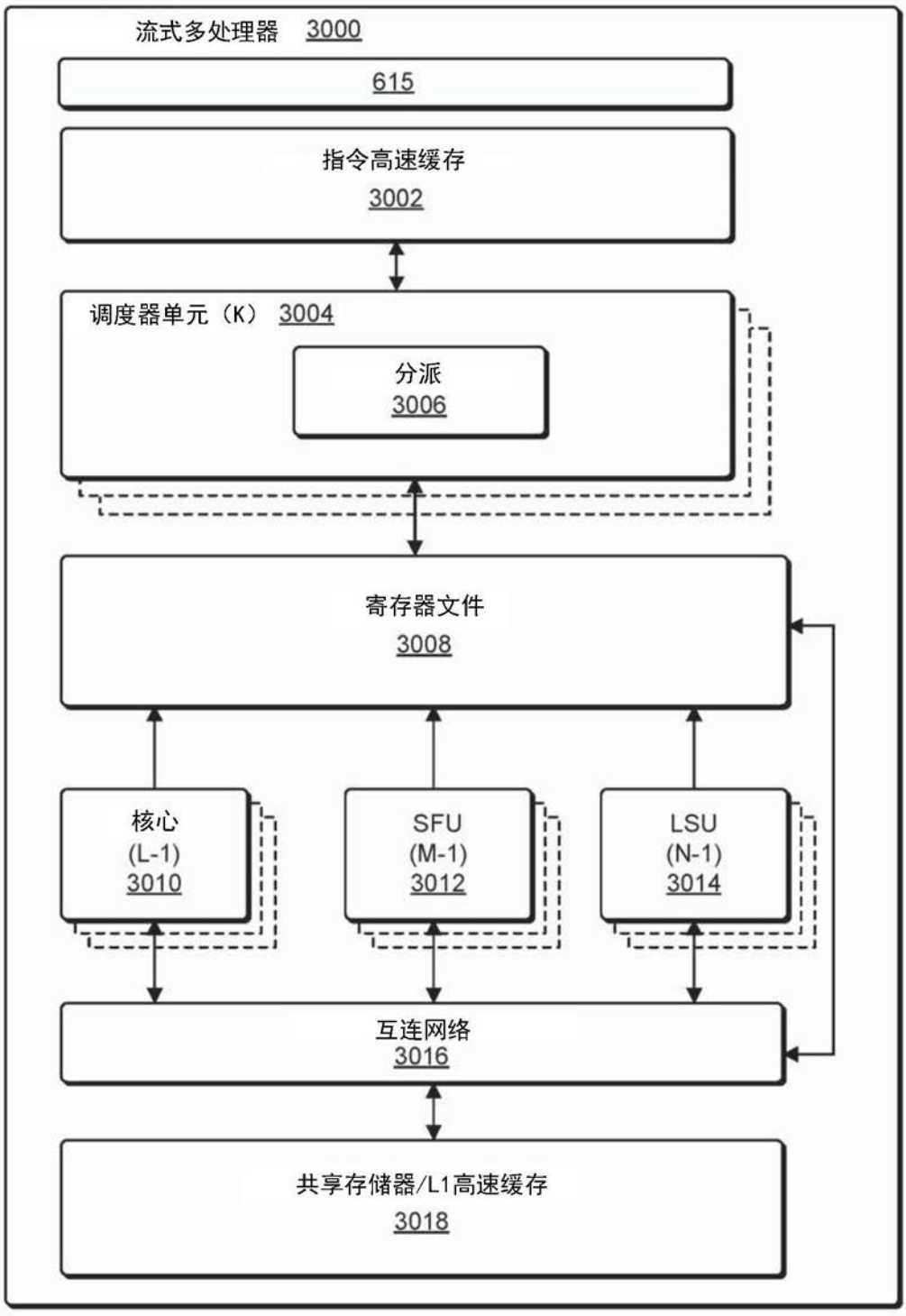

流式多处理器 3000
615
指令高速缓存
3002
调度器单元（K） 3004
分派
3006
寄存器文件
3008
核心
(L-1)
3010
SFU
(M-1)
3012
LSU
(N-1)
3014
互连网络
3016
共享存储器/L1高速缓存
3018

图30

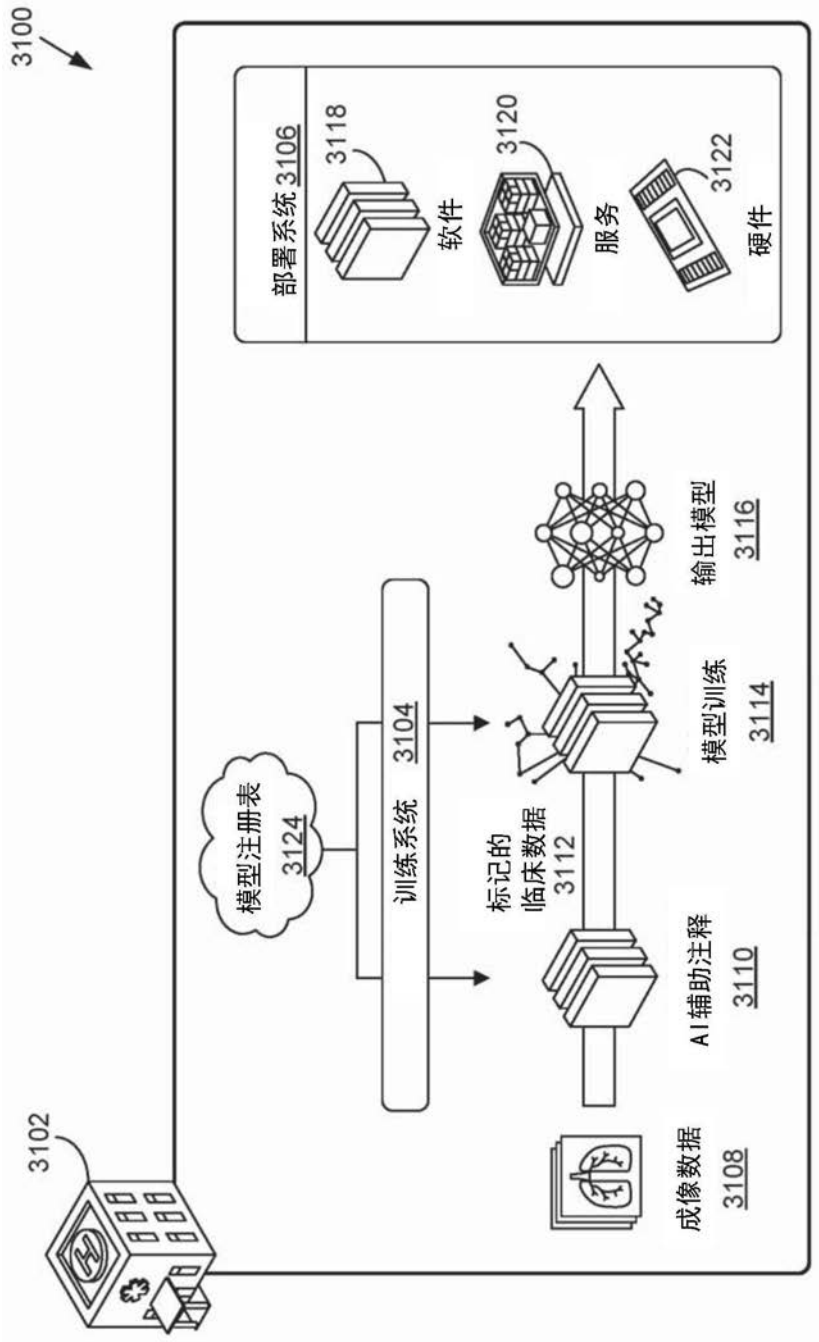

3100
部署系统 3106
3118
软件
3120
服务
3122
硬件
输出模型
3116
模型训练
3114
模型注册表
3124
训练系统 3104
标记的
临床数据
3112
AI辅助注释
3110
成像数据
3108
3102

图31

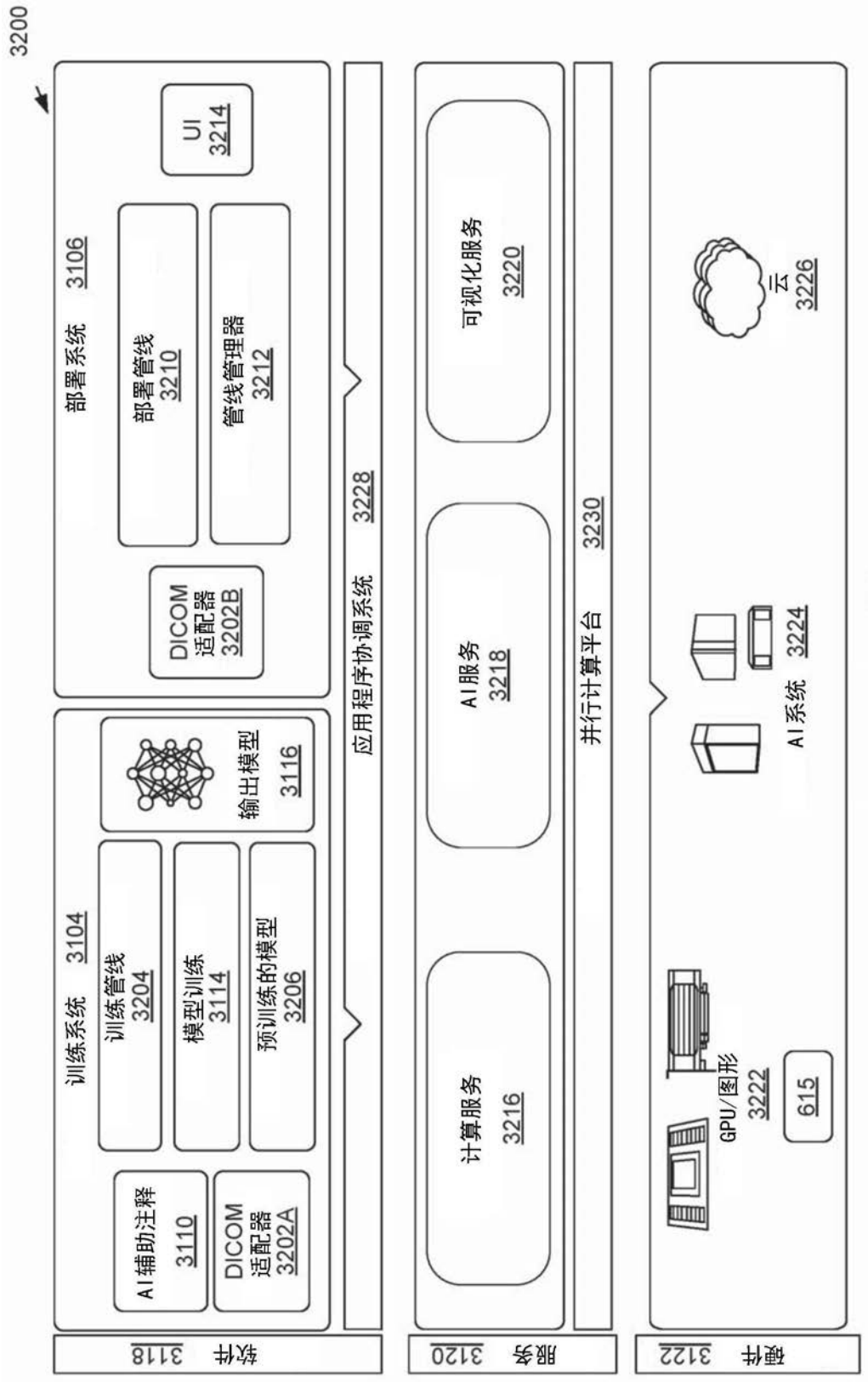

3200
部署系统 3106
部署管线 3210
管线管理器 3212
UI 3214
DICOM 适配器 3202B
应用程序协调系统 3228
训练系统 3104
训练管线 3204
模型训练 3114
预训练的模型 3206
输出模型 3116
AI辅助注释 3110
DICOM 适配器 3202A
软件 3118
计算服务 3216
AI服务 3218
可视化服务 3220
并行计算平台 3230
服务器 3120
云 3226
AI系统 3224
GPU/图形 3222
615
硬件 3122

图32

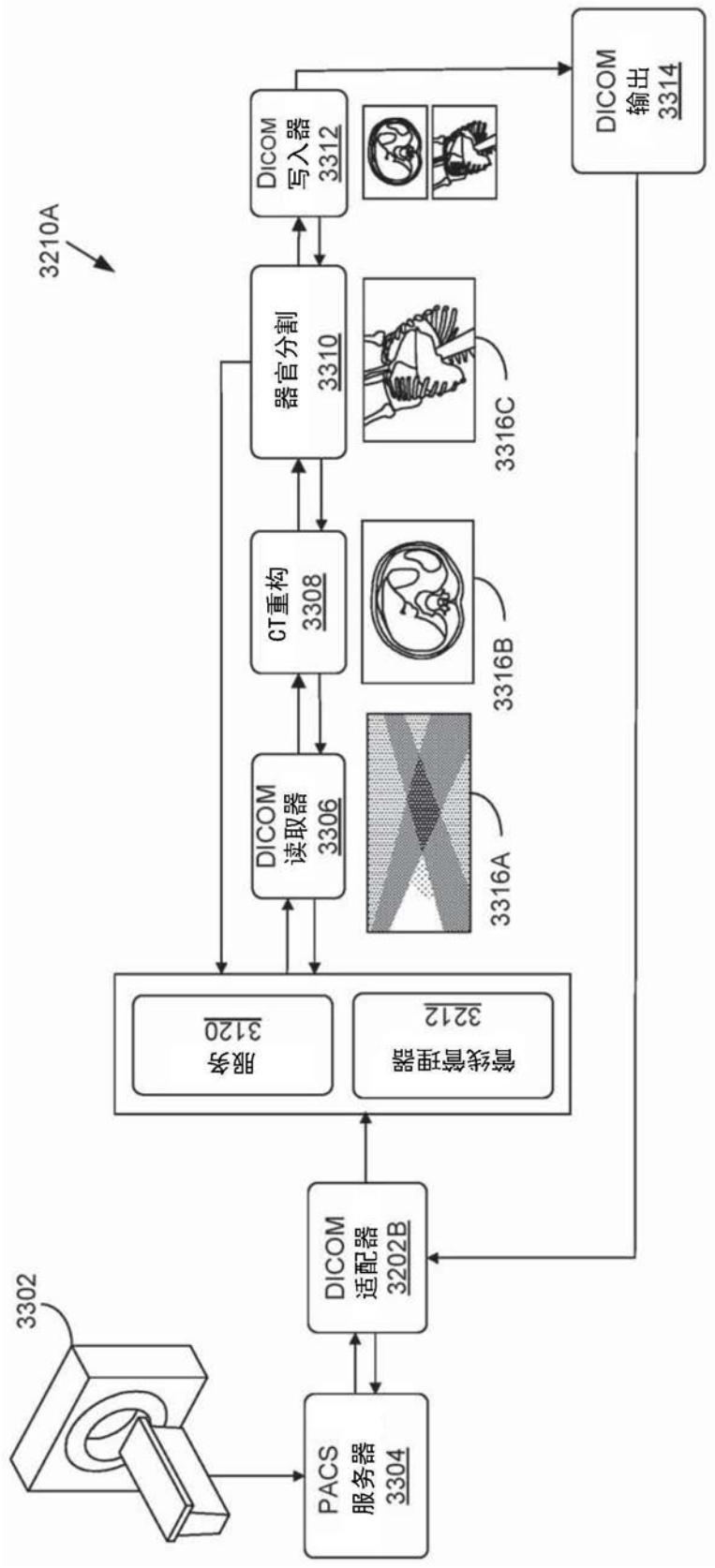

3210A
DICOM 写入器 3312
DICOM 输出 3314
器官分割 3310
CT重构 3308
DICOM 读取器 3306
3316A
3316B
3316C
服务 3120
管线管理器 3212
DICOM 适配器 3202B
PACS 服务器 3304
3302

图33

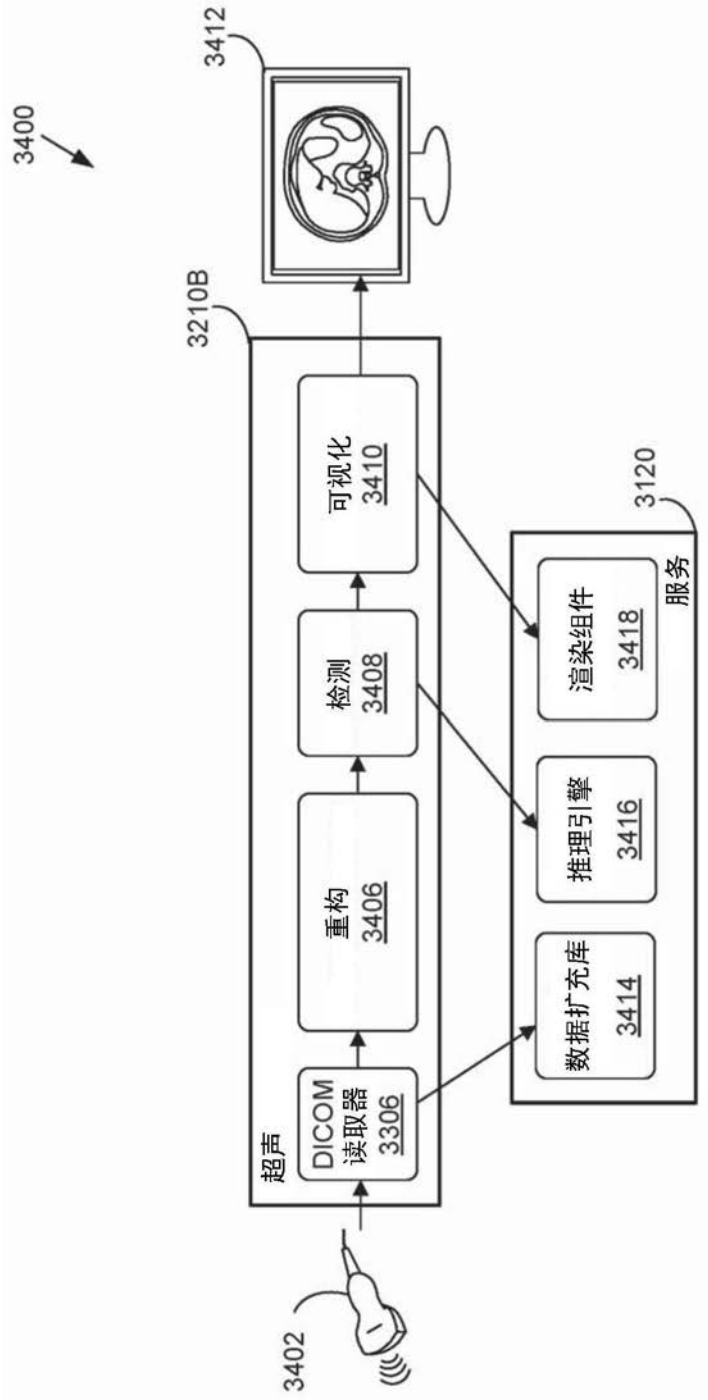

3400
3412
3210B
超声
DICOM
读取器
3306
重构
3406
检测
3408
可视化
3410
3120
服务
渲染组件
3418
推理引擎
3416
数据扩充库
3414
3402

图34A

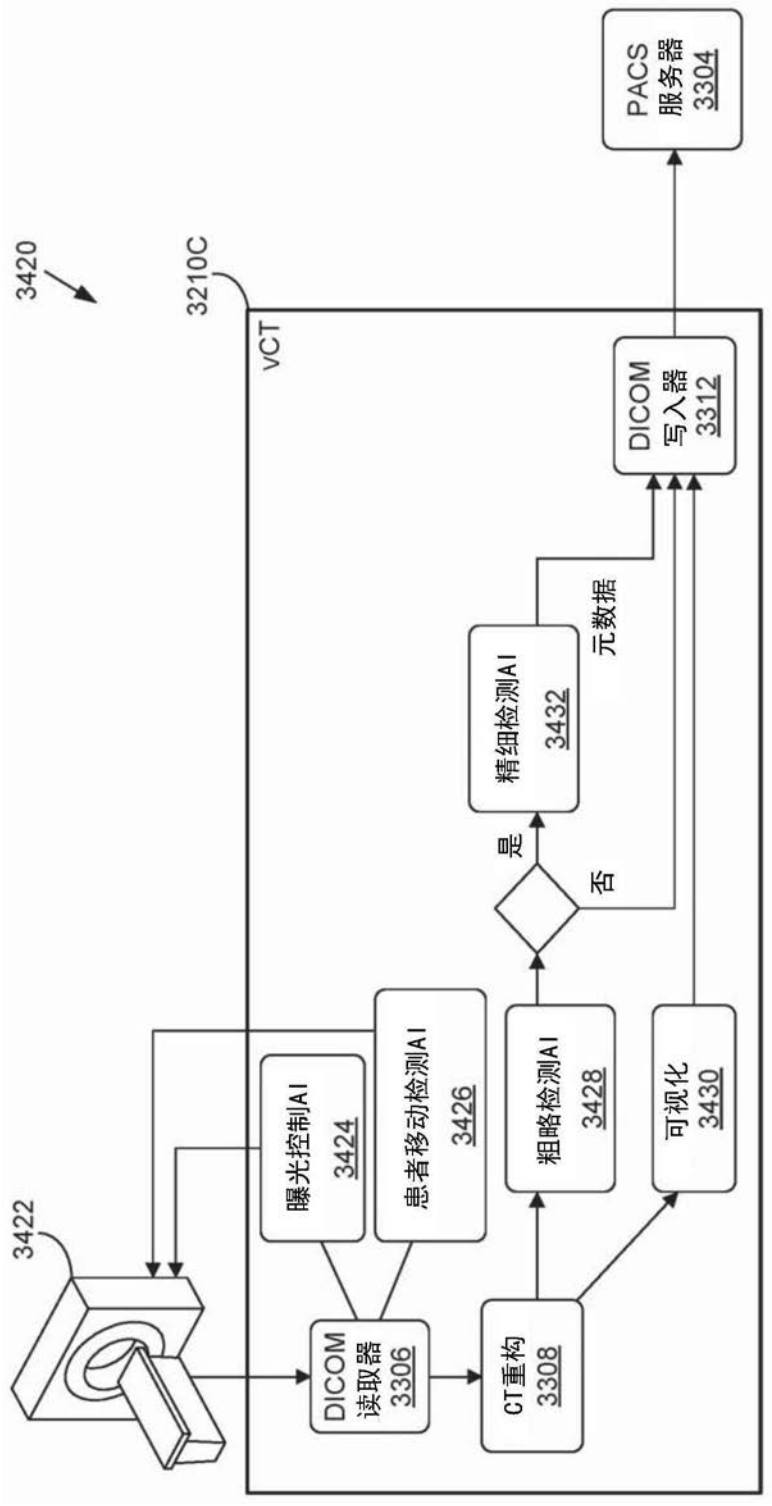

3420
3210C
vCT
PACS
服务器
3304
DICOM
写入器
3312
元数据
精细检测AI
3432
是
否
曝光控制AI
3424
患者移动检测AI
3426
粗略检测AI
3428
可视化
3430
3422
DICOM
读取器
3306
CT重构
3308

图34B

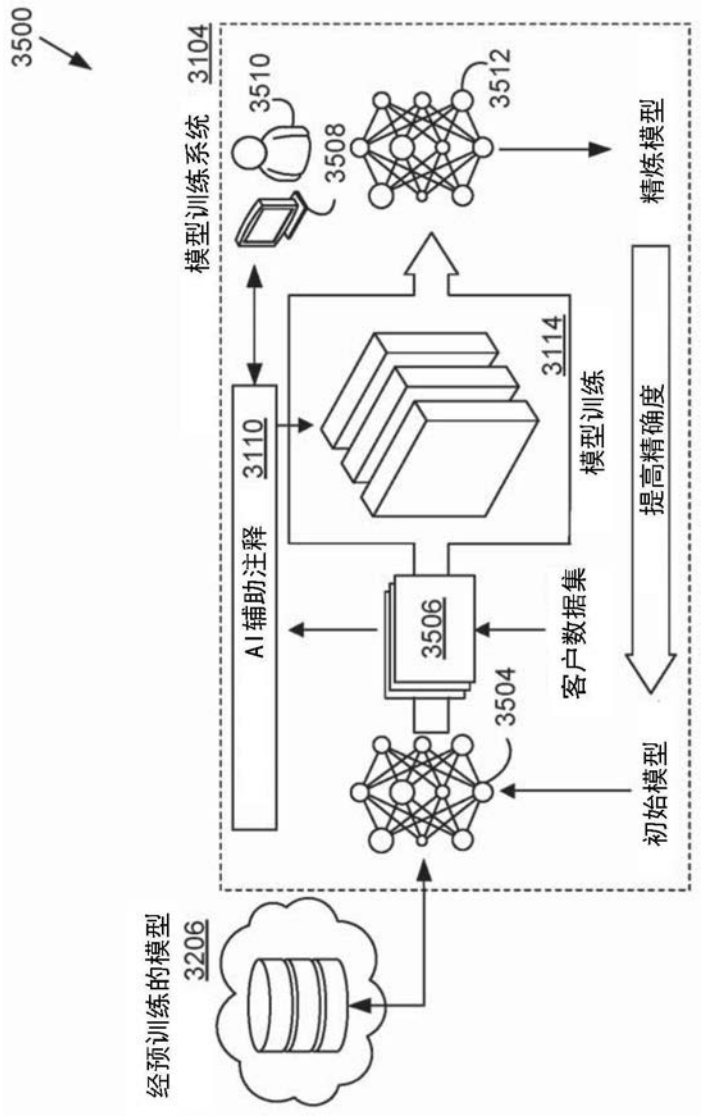

3500
模型训练系统 3104
3510
3508
3512
精炼模型
3114
模型训练
AI辅助注释 3110
提高精确度
3506
客户数据集
3504
初始模型
经预训练的模型
3206

图35A

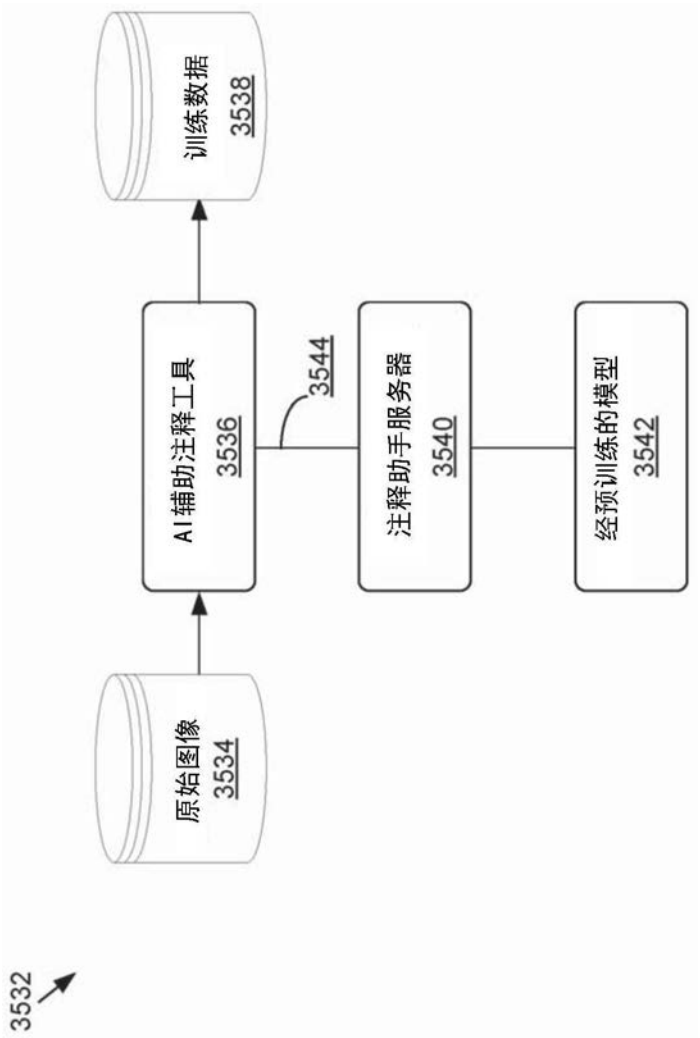

训练数据
3538
AI辅助注释工具
3536
3544
注释助手服务器
3540
经预训练的模型
3542
原始图像
3534
3532

图35B

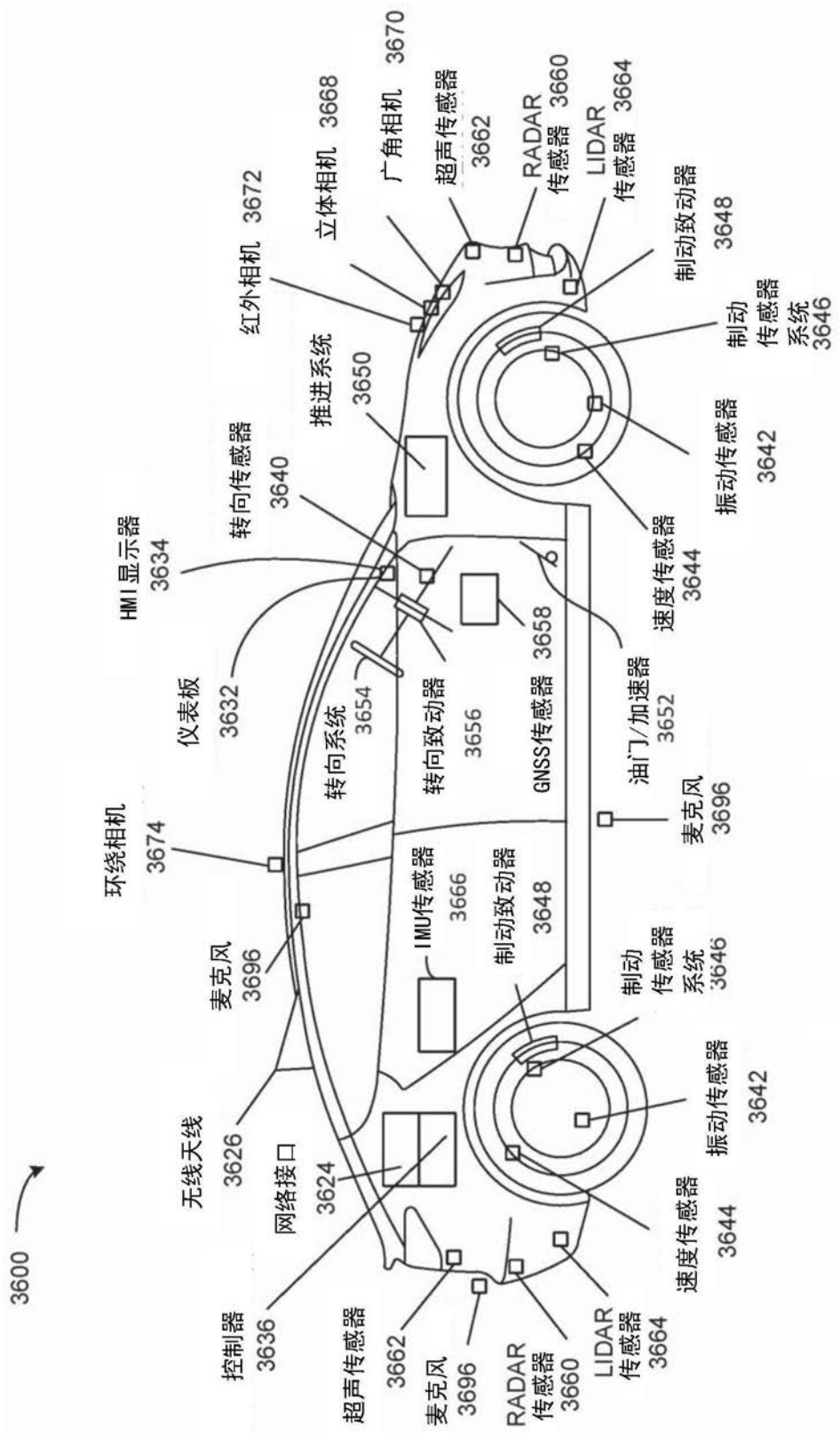

3600
控制器 3636
超声传感器 3662
麦克风 3696
RADAR 传感器 3660
LIDAR 传感器 3664
速度传感器 3644
振动传感器 3642
制动传感器系统 3646
制动致动器 3648
无线天线 3626
网络接口 3624
麦克风 3696
IMU传感器 3666
制动致动器 3648
环绕相机 3674
仪表板 3632
转向系统 3654
转向致动器 3656
GNSS传感器 3658
油门/加速器 3652
麦克风 3696
速度传感器 3644
振动传感器 3642
制动传感器系统 3646
制动致动器 3648
HMI显示器 3634
转向传感器 3640
推进系统 3650
红外相机 3672
立体相机 3668
广角相机 3670
超声传感器 3662
RADAR 传感器 3660
LIDAR 传感器 3664

图36A

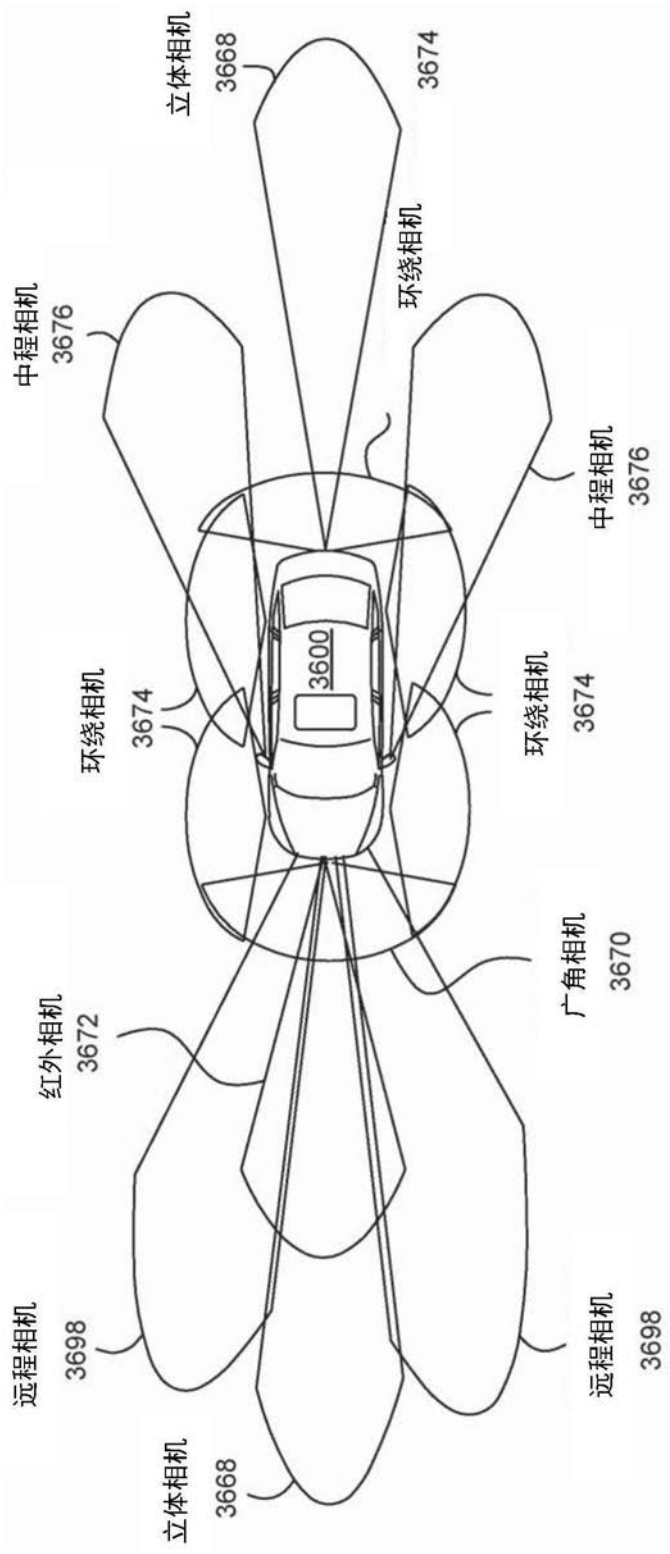

立体相机
3668
3674
环绕相机
中程相机
3676
中程相机
3676
3600
环绕相机
3674
环绕相机
3674
广角相机
3670
红外相机
3672
远程相机
3698
远程相机
3698
立体相机
3668

图36B

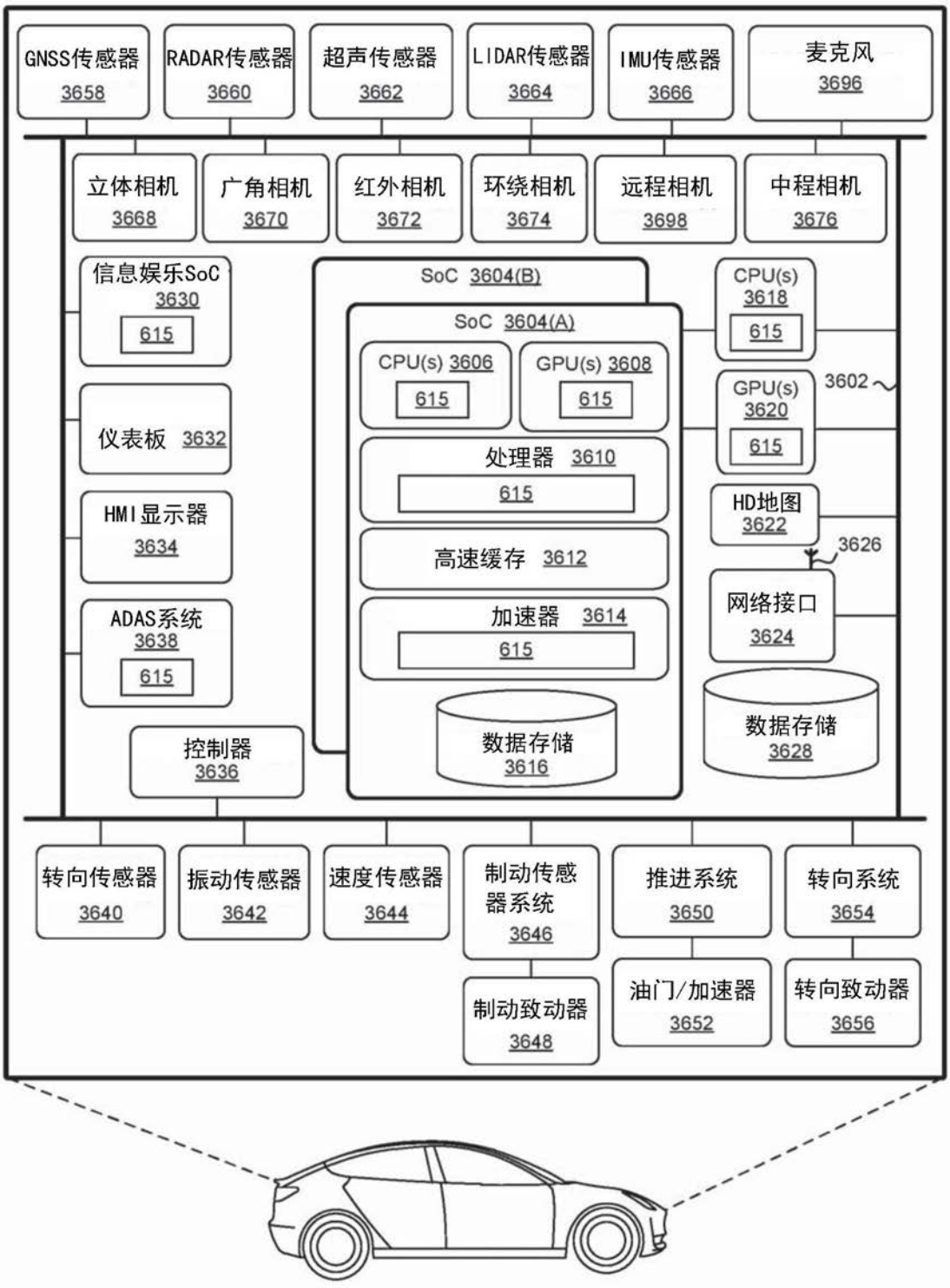

GNSS传感器 3658
RADAR传感器 3660
超声传感器 3662
LIDAR传感器 3664
IMU传感器 3666
麦克风 3696
立体相机 3668
广角相机 3670
红外相机 3672
环绕相机 3674
远程相机 3698
中程相机 3676
信息娱乐SoC 3630
615
仪表板 3632
HMI显示器 3634
ADAS系统 3638
615
控制器 3636
SoC 3604(B)
SoC 3604(A)
CPU(s) 3606
615
GPU(s) 3608
615
处理器 3610
615
高速缓存 3612
加速器 3614
615
数据存储 3616
CPU(s) 3618
615
GPU(s) 3620
615
3602
HD地图 3622
3626
网络接口 3624
数据存储 3628
转向传感器 3640
振动传感器 3642
速度传感器 3644
制动传感器系统 3646
制动致动器 3648
推进系统 3650
油门/加速器 3652
转向系统 3654
转向致动器 3656

图36C

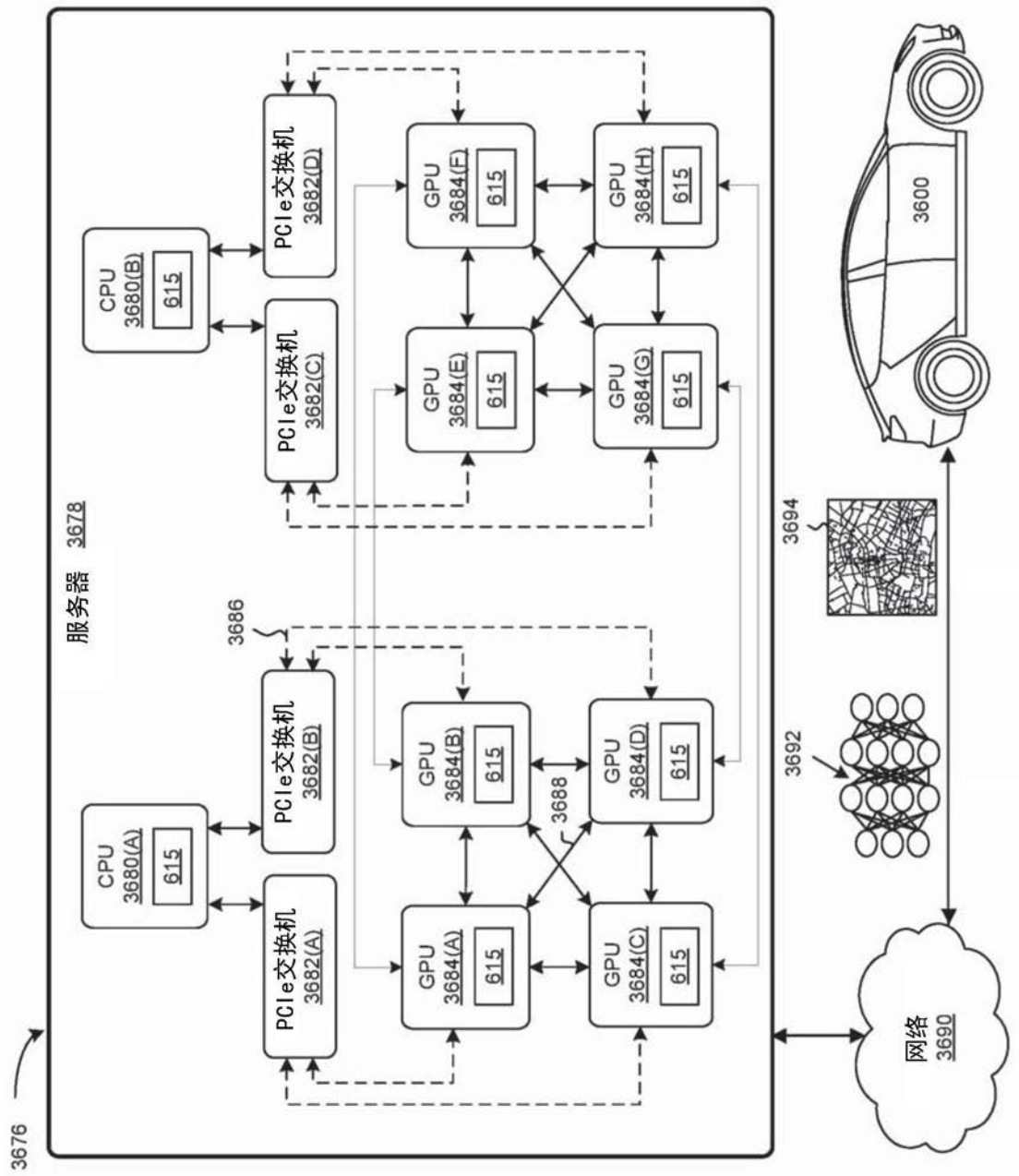

3676
服务器 3678
CPU 3680(A)
615
CPU 3680(B)
615
PCIe交换机 3682(A)
PCIe交换机 3682(B)
PCIe交换机 3682(C)
PCIe交换机 3682(D)
3686
3688
GPU 3684(A)
GPU 3684(B)
GPU 3684(C)
GPU 3684(D)
GPU 3684(E)
GPU 3684(F)
GPU 3684(G)
GPU 3684(H)
3600
3694
3692
网络 3690

图36D