一种采用检查点机制的动态主动GPU抢占机制

介绍

GPU由于其大规模并行处理能力，已经在高性能计算，机器学习和科学计算等领域得到广泛应用。这些方面的计算如今以服务的形式存在于数据中心或云上，而GPU则可作为共享基础硬件资源提供给不同的用户。多任务在GPU中为支持并行服务和任务已经变得必不可少。已经有些主要的硬件特性支持多任务处理，例如Nvidia Kepler体系结构提供的HYPER-Q, AMD支持的命令处理机制。虽然已经有这些机制支持多任务处理，但还需要更多的工作来支持真正的多任务处理机制。

上下文切换是一种在CPU中常用的支持并行性的技术，如今已经被应用到了GPU来支持多任务。CPU的进程相对来说非常轻量化，所以在上下文切换和任务时分应用上非常快速高效。但是，一个CUDA的上下文相比于CPU却是非常巨大的。比如说在NVIDIA GTX980 GPU上，上下文包括每个流多核处理器的256KB的寄存器和96KB的共享内存，对于一个含16个流多核处理器的GPU，其上下文大小达到了5664KB，存储这样大的上下文需要花费大量的存储带宽，并带来严重的性能损失。

之前已经有许多尝试来降低GPU的上下文切换的开销。最早出现的技术是让上下文切换仅在一部分SM上进行，这样其他的SM就能保持执行。被切换的SM回完全停止执行指令，以完成上下文的存取，这些操作对于存储带宽的要求依然非常高。之后出现的方法是，让一部分线程块继续执行直到完成，而只上下文切换一部分线程块。这样可以最大化的利用程序访存和上下文访存的重叠并女性。这个技术被进一步加强扩展，允许在需要抢占时每一个SM内部的不同线程块同时进行执行（直至线程块完成）、丢弃（满足幂等性）、和上下文切换。这些选择均取决于每个线程块对于抢占的终止时间的要求。除了这些工作，一种轻量级的上下文切换技术被设计来降低需要保存在片外存储器的上下文大小。所有这些方法都是通过被动的方法来实现抢占，即只有当抢占请求到来以后才激活所有的操作。因此，如果没有丢弃操作，抢占延迟依然是对性能的一大挑战。

在本章，我们提出了一种动态主动的抢占机制，PEP。这种抢占机制能够大大降低抢占延迟和开销。通过观察kernel从CPU到GPU的发射过程，我们发现kernel的实际执行总是在kernel发射之后。从一个kernel在CPU发射到开始在GPU执行，大约需要几十个毫秒的数量级，我们可以通过预计抢占请求的到达时间来主动准备上下文切换。当抢占的kernel真正到达的时候，需要等待完成的上下文切换工作将变得非常小。因此，需要等待的抢占时间将变得非常短。准备上下文切换的工作我们采用了检查点（checkpoint）的概念。第一个检查点，我们在抢占被预计发生时备份当前的上下文。当真正的抢占请求到达GPU后，仅需再备份变化的上下文。备份变化的上下文相比于完整的kernel的上下文能够节省大量的时间，减少抢占kernel的等待时间。平均来看，总的需要备份的上下文不大于所有的上下文。我们还观察到分配的上下文在线程块的生命周期里并不是完全激活的。所以，我们为寄存器设置脏位，来表明该寄存器是否是激活的。只有有效的寄存器才会被备份，这大大减少了需要存储的上下文大小。此外，我们设计了一个动态实时调度策略来确定抢占方法。小的kernel将要继续执行直到结束，而大的kernel需要采用checkpoint（上下文切换）来进行抢占。这个算法可以达到最小延迟和开销。

我们的贡献主要包括：

1. 我们研究了kernel发射的过程，观察到抢占的事件可以被预测。
2. 我们引入了一种主动的抢占技术来减少抢占的kernel等待上下文切换的时间。采用主动checkpoint技术，当真正的抢占请求到来时，只有一小部分的脏上下文需要被存储。
3. 我们使用了一种相对简单的脏数据存储技术来减少上下文大小，这可以减少不必要的上下文存储。
4. 我们开发了一种相对更加精确的线程块执行时间和上下文切换时间估算方法，设计了实时动态选择算法来确定采用的抢占方法。我们可以完成小kernel和大kernel的抢占，并使之达到最短延迟和最小开销。

我们实验评测了PEP，并与之前最好的抢占工作Chimera在集中不同类型的测试集里比较。我们的实验结果显示，相比之前的工作Chimera，我们可以将平均抢占延迟从8.9us降低到3.6us。我们采用的简单的上下文大小减少技术，将需要存储的上下文从完整的上下文大小减少了16.1%。PEP的总开销，即平均线程块切换延迟相比Chimera减少了6.3%。

背景和动机

在这一节，我们首先简单介绍了GPU的基本结构及其工作模式。我们的基准结构模拟的是一款NVIDIA的GPU体系结构。因此，我们在本章主要使用NVIDIA/CUDA的术语。但是，本章的想法也可以应用到其他厂商的GPU。我们还介绍了checkpointing的方法，该方法在我们的设计中起到了关键作用。

基准结构

1. GPU程序执行：典型的GPU程序包括两个部分的代码：在CPU上运行的主机部分的代码，以及在GPU上运行的设备代码（kernels）。Kernels是以SIMT的模式执行（单条指令，多条线程）。一个kernel的执行意味着无数线程同时在GPU上并行执行。线程会被程序员组合成线程块。

NVIDIA GPU的CUDA编程模型以CUDA C的形式展示给程序员，即从C语言和实时库中扩展而来。图1是一段CUDA程序的例子。一段典型的CUDA C程序的操作序列包括：

1. 声明和分配主机和设备的内存（8-13行）
2. 从主机内存向设备内存迁移数据（14行）
3. 发射kernel。在这个例子里，程序员发射N/256个线程块，每个线程块包括了256条线程（15行）
4. 从设备内存向主机内存迁移数据（16行）
5. 释放内存空间（18-19行）

线程块之间是互相独立的，他们被分别发送到SM上。每个SM上能够并行的线程块数量受限于设备的资源（包括寄存器，共享内存和线程的数量），这个信息在编译时可以知道。大多数之前的抢占策略设计的工作是以线程块的粒度完成的，也会采用可用资源的信息帮助抢占策略的选择。

1. GPU体系结构：图2是GPU基准体系结构，我们本章所描述的GPU体系结构均基于此。当一个GPU程序收到主机CPU执行时发送的操作指令，用户空间实时引擎将API调用指令发送来控制数据操作和kernel的发射。GPU设备驱动发送这些操作指令到流控制管理器的队列里。流控制管理器通过软件队列来管理多条不同的流；每一条流里的指令将被串行执行。一般来说，CPU会先声明并分配存储空间，然后调用cudaMalloc来分配GPU上的全局内存。之后，一个cudaMemcpy（H2D）API的调用将数据从主机内存移动到设备内存。一旦所有的数据被传输完成，流控制管理器可以发射内核，即传输kernel信息（例如维度配置和每个条目的PC地址）到内核函数管理器单元（KMU）。当所有的信息准备完毕，kernel会请求SM资源。如果SM没有足够的资源，kernel需要等待kernel等待池。如果正在等待的kernel的优先级高于正在执行的kernels，则等待kernel有可能需要抢占SM中正在占用资源的线程块，否则，需要等待之前的kernel执行完毕再接着执行。

一旦kernel准备好开始执行，它将被传输到kernel分发单元（KDU）。线程块调度器将相应的线程块分发到不同的SM。每个SM能够处理的线程块的最大数量取决于资源限制，包括可以执行的线程块数量、线程数量、寄存器数量和共享内存空间。在每个SM中的kernel执行过程中，线程块将被分成warps，每个warps包含最多32个线程。每个SM包含一个或多个warp调度器，来选择发出哪一个warp。在nvidia gtx980 gpu体系结构中，每个warp调度器控制32个流处理单元（SP），每个流处理单元处理一个线程。当一个warp因为访存或其他耗时较长的操作而停滞时，调度器会切换执行其他的warp。切换warp没有任何开销，因为所有warps的上下文已经被放入寄存器和共享内存中。因此，GPU通过掩藏停滞warp的延迟，大大提升了并行性能。

B. 先前的抢占方法

当抢占发生时，每个SM的操作可以是独立的，这意味着有一些SM可能是执行抢占，而另一些SM可以继续执行直到结束。被抢占的SM需要将SM内的上下文存储备份到全局内存。一个SM的上下文就是其执行的状态，包括SIMT栈、寄存器和共享内存。SIMT栈存储的是线程执行信息，例如程序计数器和激活掩码（用于分支处理）。相比于寄存器和共享内存的大小，SIMT栈的大小可以忽略不计，因此在本章我们暂不考虑SIMT栈。一个线程块在其执行的时候占用SM的资源；线程块保持活动状态直到其执行完毕。但是，在其执行的时候，有可能有新的kernel会被发射。当该新的kernel需要满足严格的延迟要求，如果等待上一个kernel执行完毕再开始执行会打破该延迟要求。因此我们需要去抢占一些活动的线程块来为新的kernel的线程块提供资源。但是，线程块的上下文相对来说非常大，将他们存到全局内存将引入相当大的开销，抢占延迟会非常大。如表1所示，抢占延迟（平均上下文切换时间）有可能超过20us。这些延迟给需要满足延迟要求的新kernel造成很大隐患。