GPU培训

思腾合力 SITONHOLY

NVIDIA ELITE PARTNER

GPU 服务器 | 集群 | 个性化定制

主题

1 为什么选择GPU？

2 GPU适合哪些运算？

3 GPU如何实现并行？

4 多GPU如何通信？

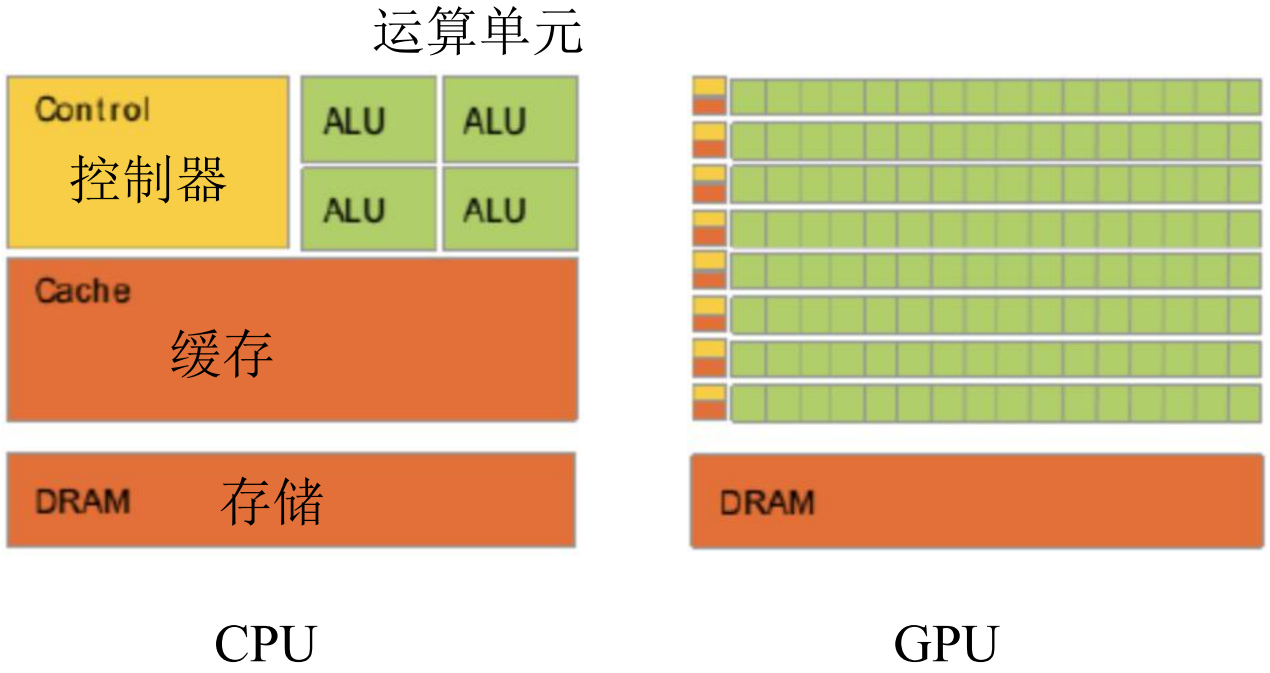
5 GPU计算方案配置

01 为什么选择GPU

1.为什么选择GPU

GPU (graphics processing unit)，图形处理器，专门为了渲染设计的。

CPU (graphics processing unit)，中央处理器，通用型强，专用性低。



渲染：电脑显示器上显示的图像，在显示在显示器之前，要经过一系列处理，这个过程就叫“渲染”

1.为什么选择GPU

- 渲染：几何点位置和颜色的计算，在数学上用四维向量和变换矩阵的乘法，GPU被设计为专门适合做类似运算的专用处理器。
- 现在游戏、3D设计对渲染的要求越来越高，GPU的性能越做越强。论纯理论计算性能，要比CPU高出几十上百倍。
- 卷积神经网络CNN数学上就是许多卷积运算和矩阵运算的组合，其中卷积运算也可以通过矩阵运算完成。因此深度学习就可以非常恰当地用GPU进行加速了。
- GP(general purpose)GPU，通用GPU，不再局限于图形领域，将其能力扩展到其他计算密集的领域。

02

GPU适合哪些运算

2. GPU适合哪些运算

适合：

- 大量的轻量级运算
- 高度并行
- 计算密集型
- 控制简单
- 多任务执行
- 浮点型运算

不适合：

- 并行度小的应用
- 不规则的任务并行
- 频繁的全局同步
- 在线程之间，会出现随机的点对点同步的应用

03

GPU如何实现并行

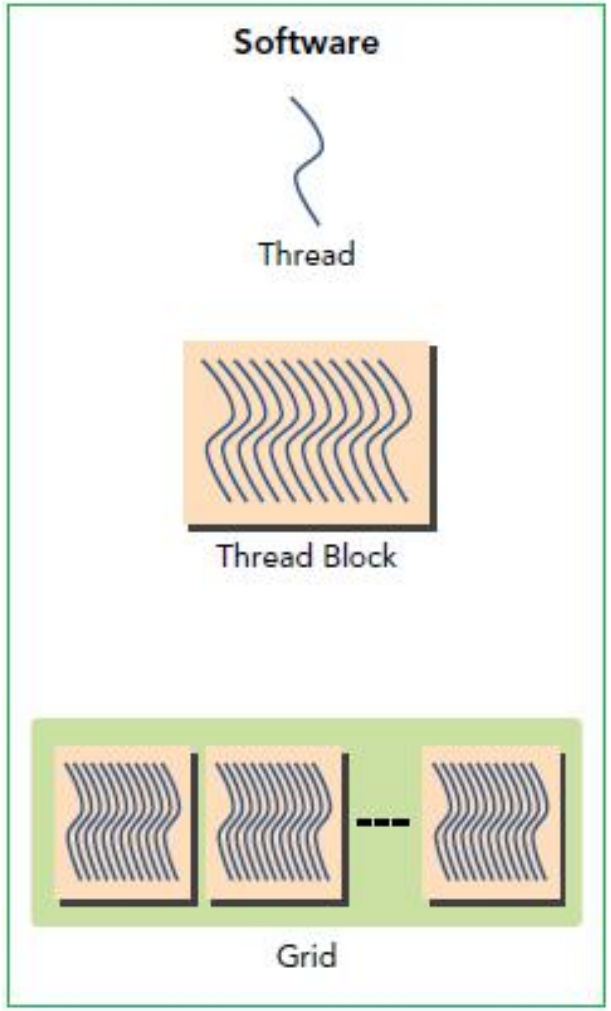
3. GPU如何实现并行

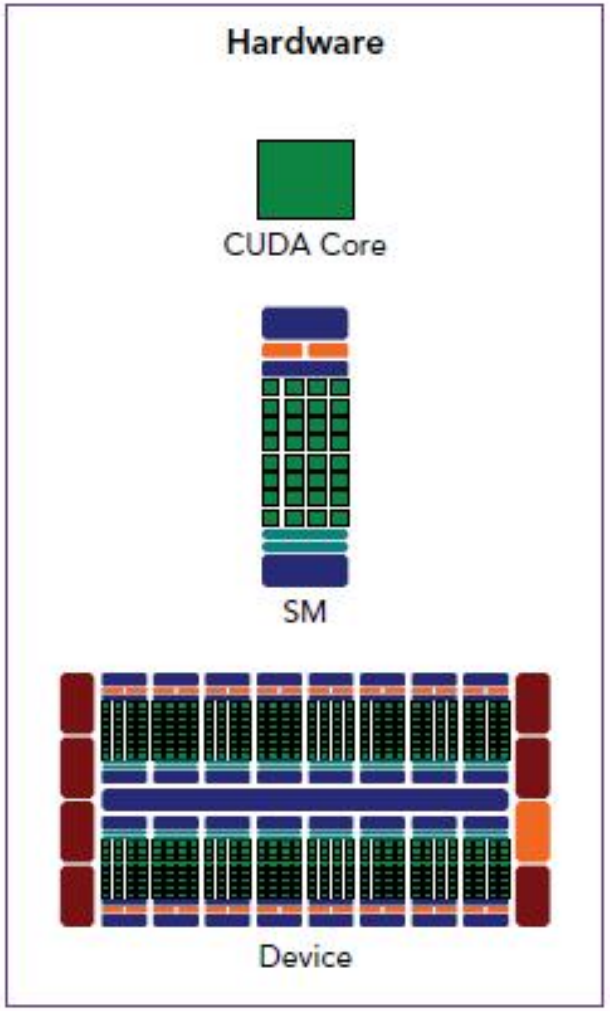
GPU架构：

3. GPU如何实现并行

软件--硬件：

1. 线程：是GPU运算中的最小执行单元，线程能够完成一个最小的　逻辑意义操作。



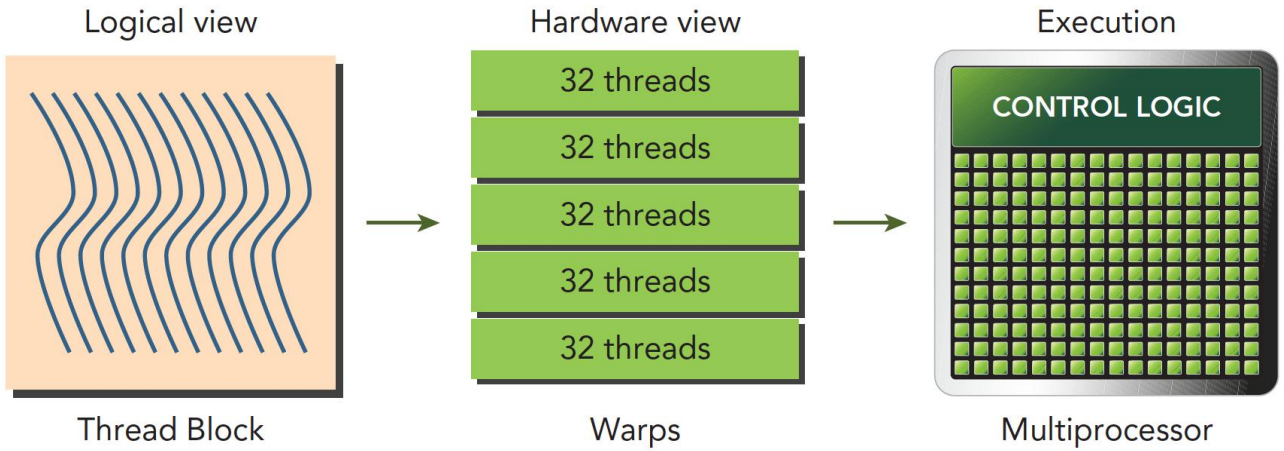


3. GPU如何实现并行

2. **线程束：**是GPU中的基本执行单元。GPU是一组SIMD处理器的集合，每个线程束中的线程是同时执行的。

3. **线程块：**一个线程块包含多个线程束，在一个线程块内的所有线程，都可以使用共享内存来进行通信、同步。

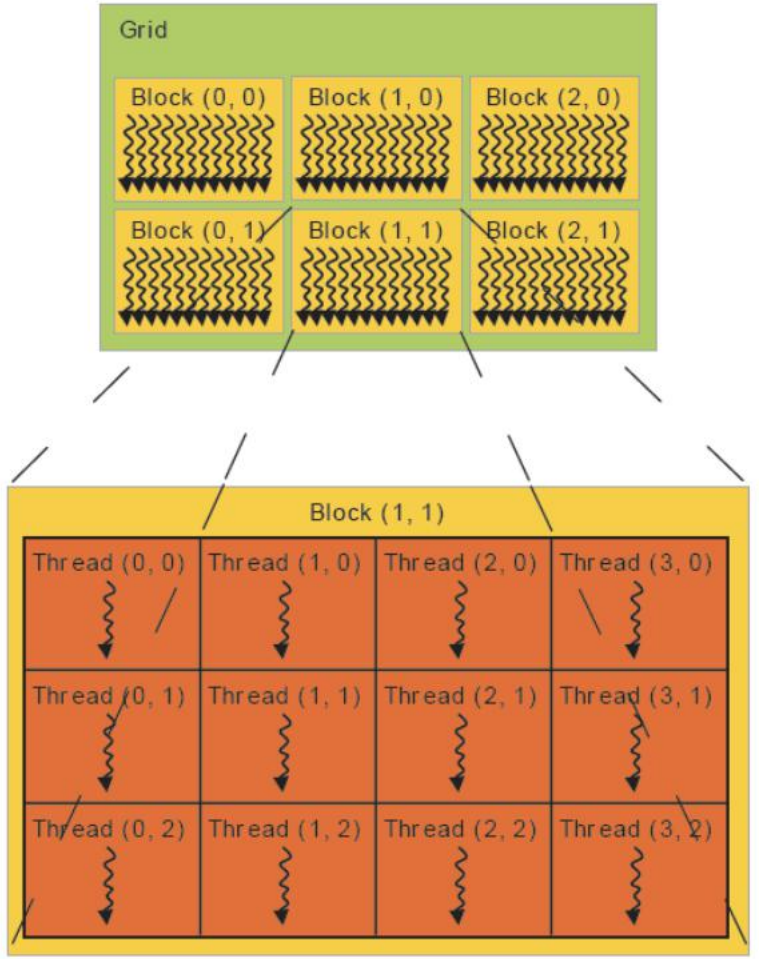
4. **流多处理器：**相当于CPU中的核，负责线程束的调度执行



3. GPU如何实现并行

GPU 并行编程的核心在于线程，一个线程就是程序中的一个单一指令流，一个个线程组合在一起就构成了并行计算网格，成为了并行的程序。

CUDA 并行编程架构可以用网格 (GRID) 来形容：一个网格好比一只军队。块好比军队的每个部门 (后勤部，指挥部，通信部等)。每个块又分成好多个线程束，这些线程束好比部门内部的小分队。

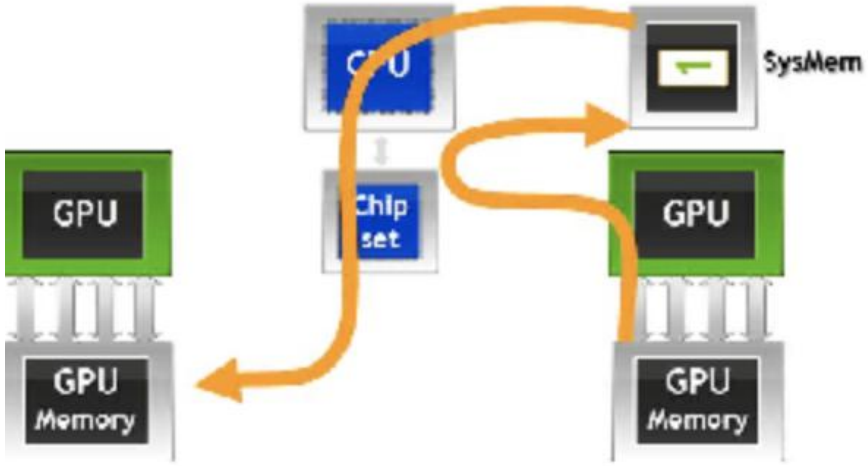


04

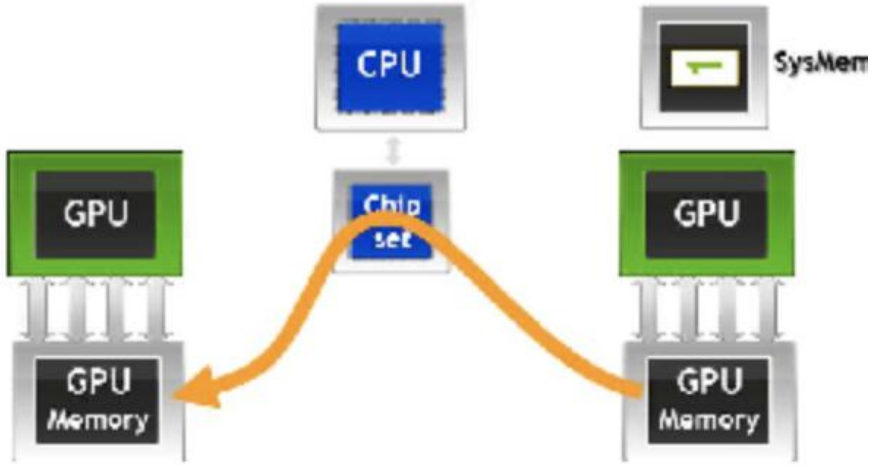
多GPU如何实现通信

4. 多GPU如何实现通信：单机多卡

- 2011年，GPUDirect Peer-to-Peer(P2P)使GPU可以通过PCIe直接访问目标GPU的显存。



No GPUDirect P2P



GPUDirect P2P

4. 多GPU如何实现通信：单机多卡

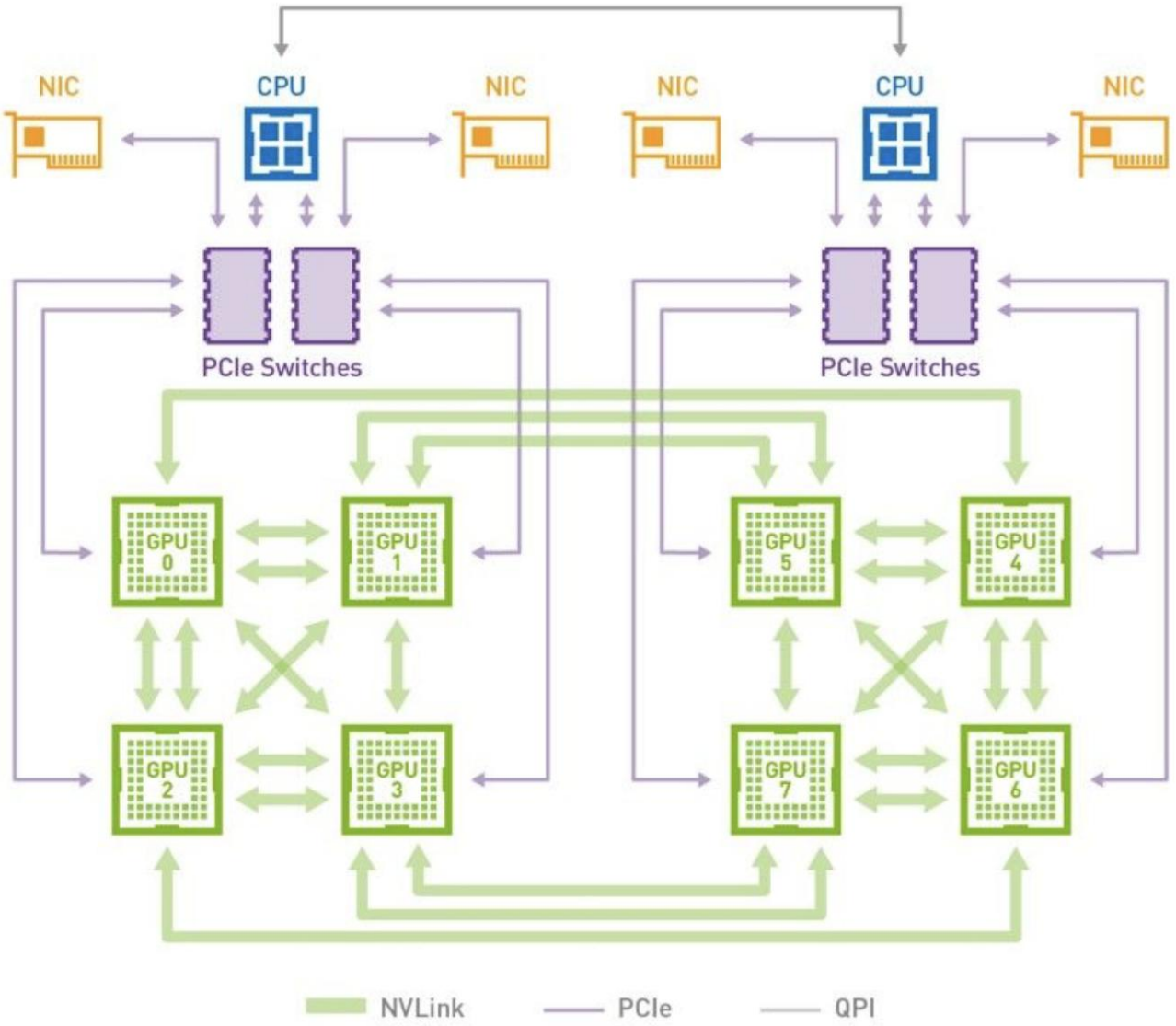
- 2014年3月，发布NVLink技术，能在多GPU之间实现非凡的连接带宽。

P100搭载NVLink 1.0，有4个NVLink通道，每个40GB/s的双向带宽，P100可以最大达到160GB/s带宽。
V100搭载NVLink 2.0，有6个通道，每个通道达到50G的双向带宽，V100可以最大达到300GB/s的带宽。

PCIe Gen3 32GB/s

4. 多GPU如何实现通信：单机多卡

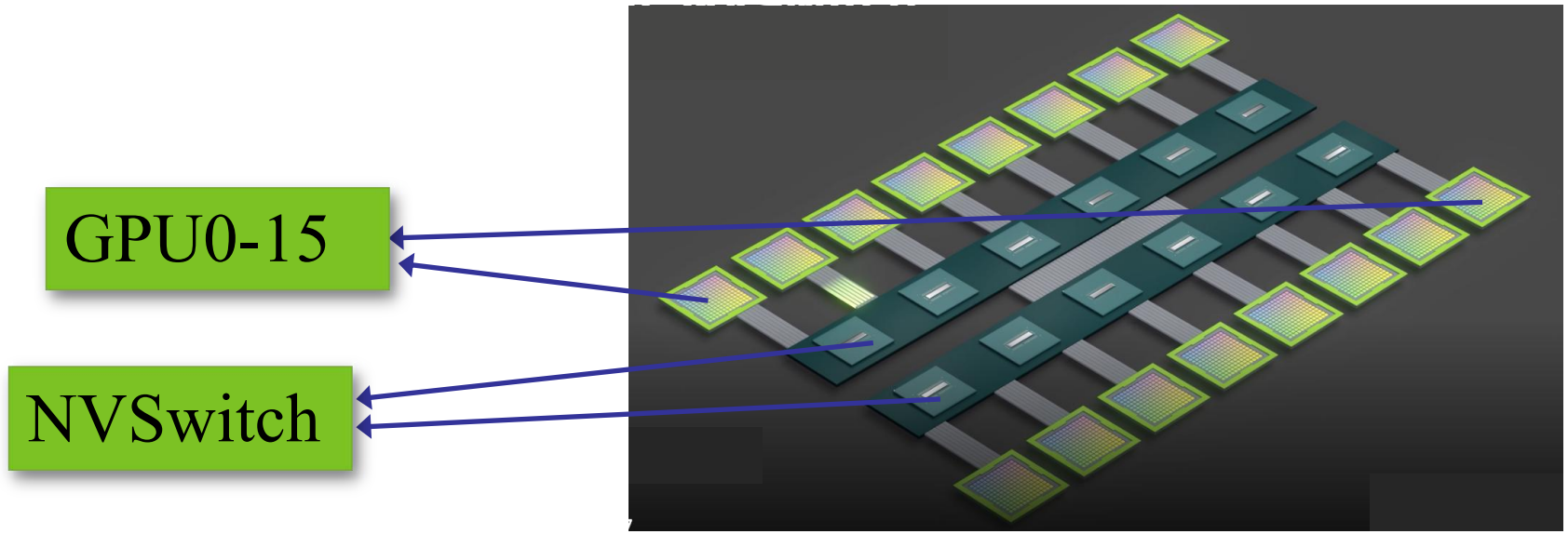
虽然V100有6个NVlink通道，但是实际上因为无法做到全连接，2个GPU间最多只能有2个NVLink通道100G/s的双向带宽。而GPU与CPU间通信仍然使用PCIe总线。CPU间通信使用QPI总线。



HGX-1/DGX-1使用的8个V100的混合立方网格拓扑结构

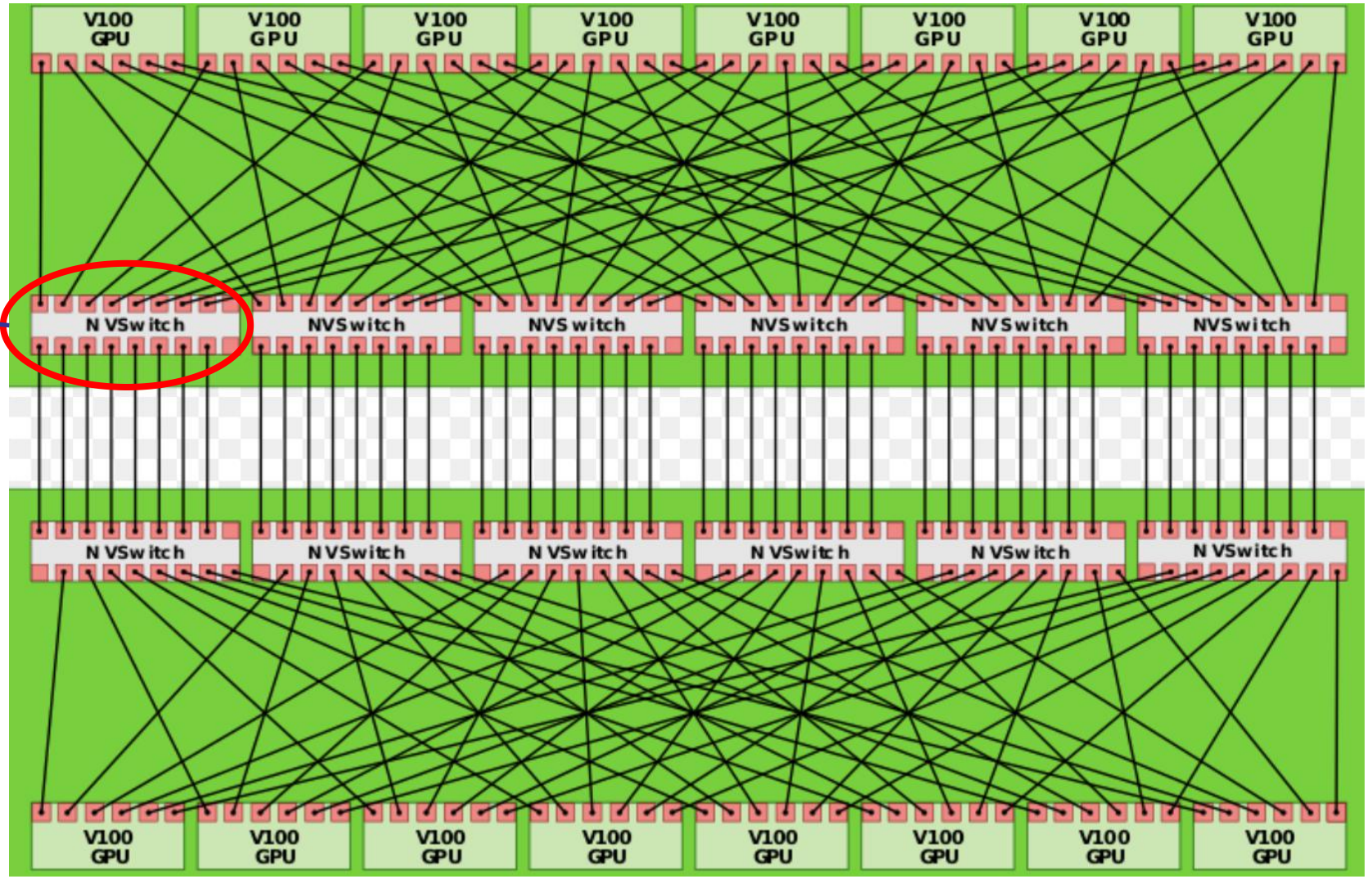
4. 多GPU如何实现通信：单机多卡

- NVSwitch交换机通过一个或多个交换机路由GPU来实现更大GPU的集群。支持单个服务器节点中 16 个全互联的 GPU，并可使全部 8 个 GPU 对分别以 300 GB/s 的惊人速度进行同时通信。



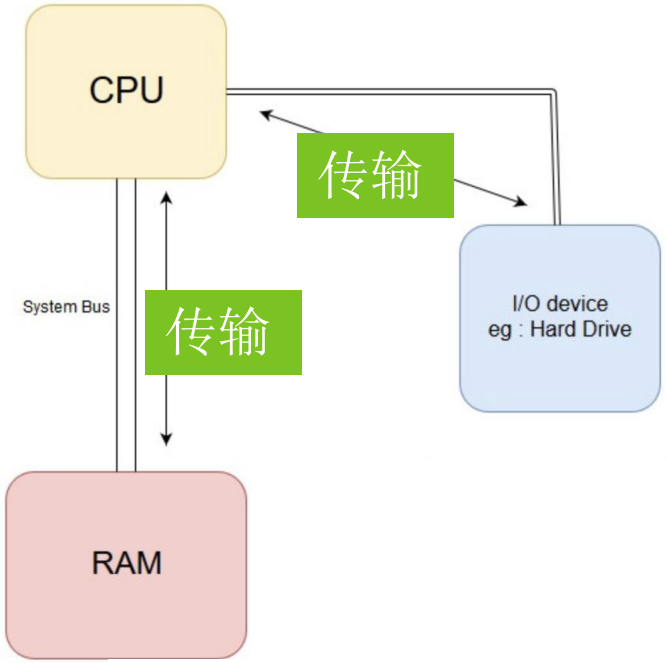
4. 多GPU如何实现通信：单机多卡

支持单个服务器节点中 16 个全互联的 GPU，并可使全部 8 个 GPU 对分别以 300 GB/s 的惊人速度进行同时通信。



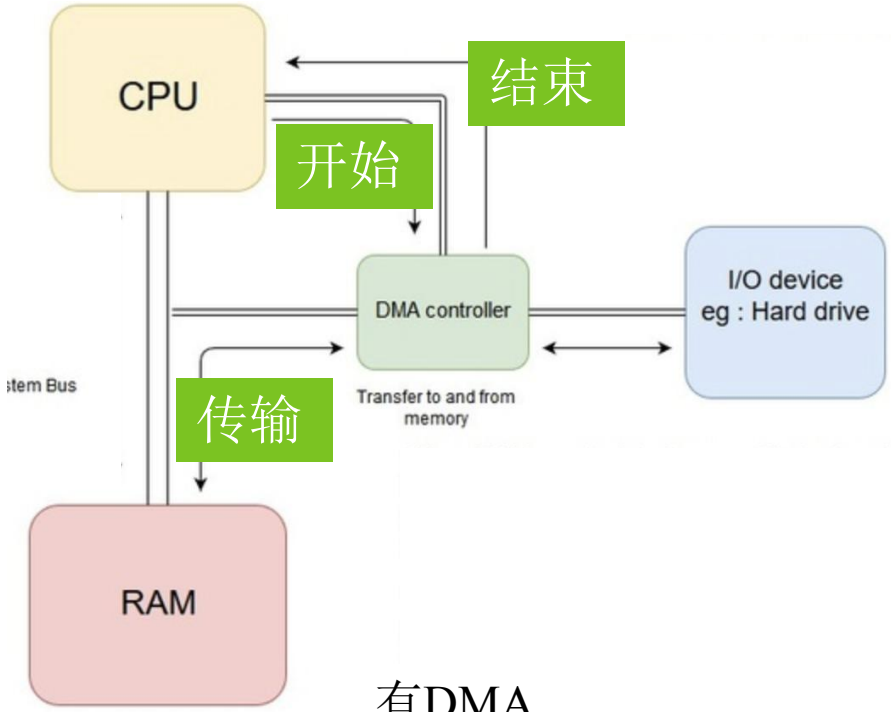
4. 多GPU如何实现通信：多机多卡

- DMA直接内存访问:完全由硬件执行I/O交换的工作方式。



没有DMA

设备内存与系统内存的数据交换必须要CPU参与

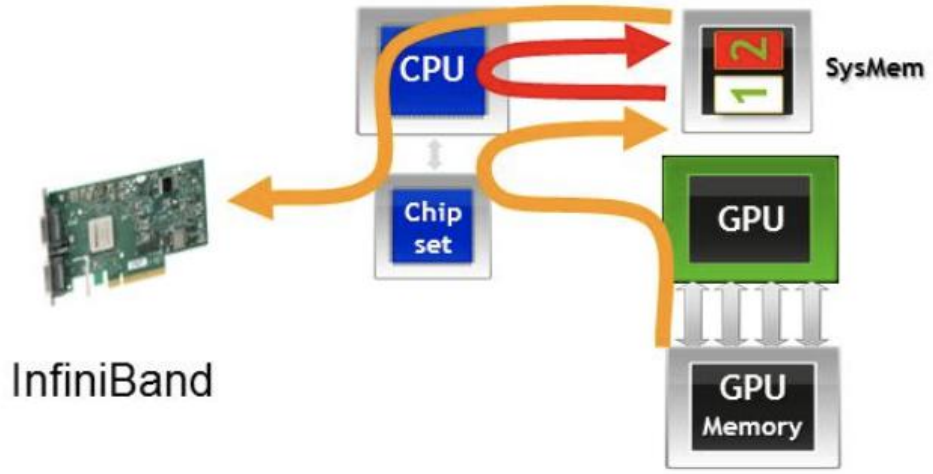


有DMA

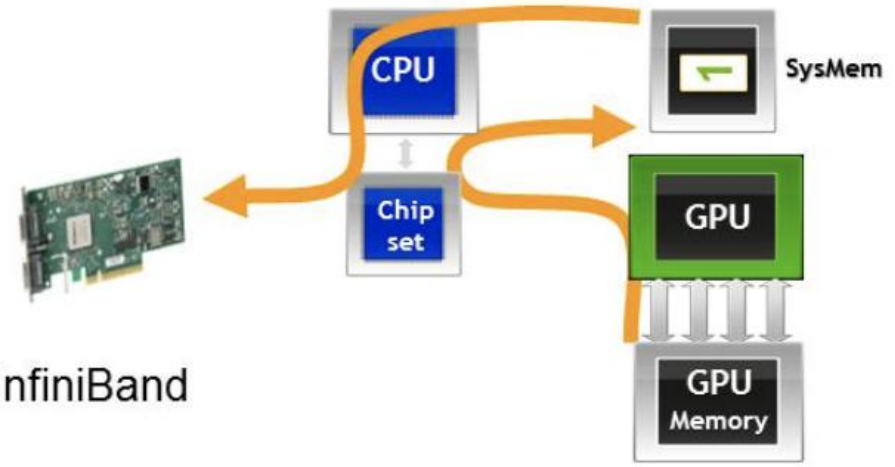
DMA控制器从CPU完全接管对总线的控制，数据交换不经过CPU，而直接在内存和IO设备之间进行

4. 多GPU如何实现通信：多机多卡

- GPUDirect :多个GPU，第三方网络适配器，固态驱动器（SSD）和其他设备可以直接读写CUDA主机和设备内存。



没有GPUDirect



有GPUDirect

4. 多GPU如何实现通信：多机多卡

- RDMA 远程直接内存访问：在计算机之间网络数据传输时Offload CPU负载的高吞吐、低延时通信技术。

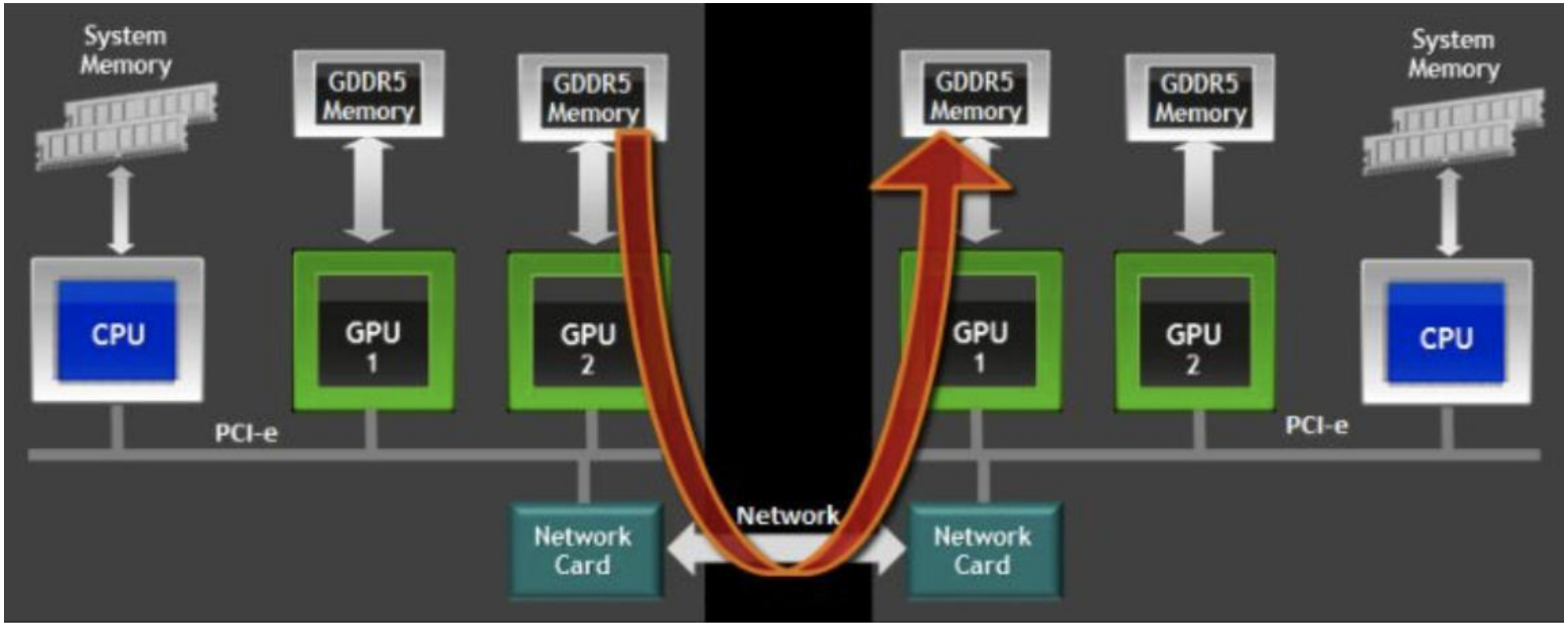
  在网卡上将RDMA协议固化于硬件，以及支持零复制网络技术和内核内存旁路技术来达到其高性能的远程直接数据存取的目标。

- GPUDirect RDMA：加速多机间GPU通信的技术。

  计算机1的GPU可以直接访问计算机2的GPU内存。

4. 多GPU如何实现通信：多机多卡

GPUDirect RDMA:



减少了GPU通信的数据复制次数，通信延迟进一步降低。

05

GPU计算方案配置

5. GPU计算方案配置

考虑因素：

1. **计算比例：**通常应用程序的执行需要GPU与CPU协同完成，可根据GPU计算部分所占比重，配置节点GPU卡密度；

2. **计算规模：**根据不同应用数据规模及计算类型，可以选择单机单GPU卡、单机多GPU卡和GPU集群应用模式；

3. **内存容量：**GPU计算节点内存容量建议配置为：GPU个数*GPU显存容量+32GB；

5. GPU计算方案配置

4. **数据通信：**在单机多卡模式下，可使用 GPU Direct P2P技术加速GPU之间数据传输速度；在GPU集群模式下，可使用GPU Direct RDMA功能，加速数据通信提升程序的执行效率，同时可根据应用程序对集群通信带宽及延迟的需求，选择高速Infiniband网络或万兆网络；

5. **存储系统：**单节点应用模式下一般数据量比较小，对存储系统性能要求不高，一般采用本地存储；集群环境下，应用数据量比较大，一般配置大容量、统一、高速的并行文件系统，另外对一些特殊应用，如石油、天然气应用，可以在每个GPU计算节点内部配置SSD硬盘，作为分级存储使用，加速节点内部数据交换；

5. GPU计算方案配置

6. **管理调度：**合理选择GPU集群的作业调度和监控系统，可以提升集群的使用效率，降低维护成本。

单机单卡模式：

适合小数据规模应用或初级用户测试、实验使用，方案设计需要同时兼顾GPU与CPU的计算性能。适合应用类型为只支持单GPU加速应用，程序执行过程中通过任务划分，由GPU和CPU共同完成计算任务，或程序中只有部分模块采用了单GPU加速功能。

5. GPU计算方案配置

单机多卡模式：

单机多卡模式下，应用对单节点计算性能和密度要求高，程序可以同时调用多个GPU使用，大部分计算任务也由GPU来承担，而CPU负责复杂指令处理及调度部分。

GPU服务器插多个GPU卡的情况下，建议多个GPU插在同一个CPU端，这样可以使用GPU Direct P2P，避免在节点内部GPU之间跨QPI通信，加速程序在多GPU运行效率。

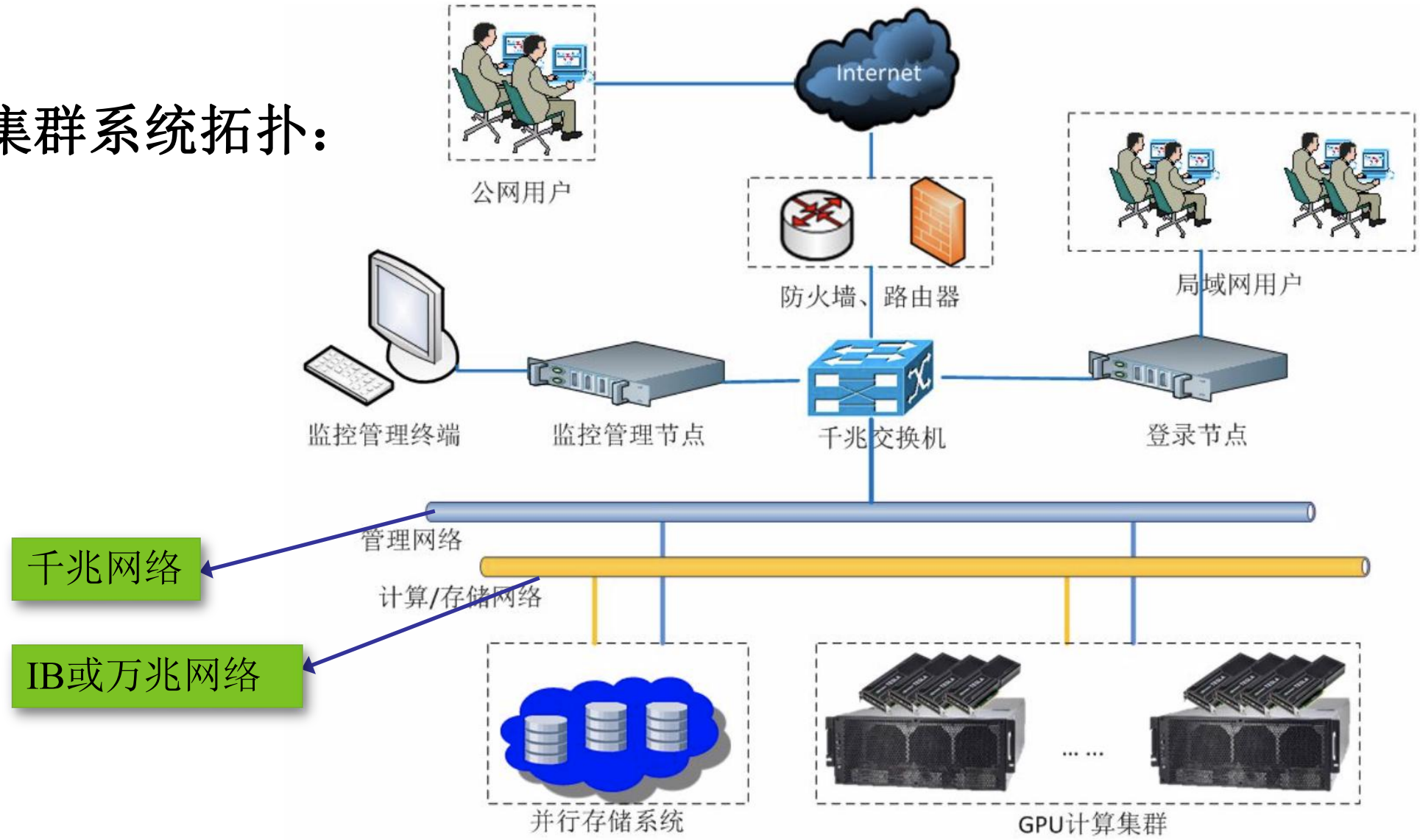
5. GPU计算方案配置

GPU集群模式：

GPU集群根据应用类型每个节点配置一个（兼顾GPU和CPU计算能力）或多个GPU卡（GPU作为节点内部主要计算单元），集群内部各节点配置相同；计算节点之间使用高速Infiniband或万兆网络作为集群的计算和互联网络；采用并行文件系统为整个集群提供高速、稳定的数据存储服务；千兆网络作为管理网络用做整个集群的监控和管理，用户可通过GPU集群管理、调度系统在外网或局域网内提交作业到GPU集群，并可实时监测到GPU集群的运行状况。

5. GPU计算方案配置

GPU集群系统拓扑：



参考文献链接：

[1] https://blog.csdn.net/xihuanyuye/article/details/81178352

[2] https://www.cnblogs.com/muchen/p/6138691.html

[3] https://yq.aliyun.com/articles/591403?spm=a2c4e.11153940.blogcont599183.6.42d5496fvLUe4T

[4] https://yq.aliyun.com/articles/599183?spm=a2c4e.11153940.blogcont599183.7.77da496f2DxNtb

[5] https://yq.aliyun.com/articles/603617?spm=a2c4e.11153940.blogcont591403.15.2bbd1cb8EICsaR

[6] https://www.nvidia.cn/data-center/nvlink/

[7] https://en.wikichip.org/wiki/nvidia/nvswitch

[8] https://www.sohu.com/a/150642645_632967

[9] https://www.quora.com/What-is-the-function-of-DMA-in-a-computer

[10] https://developer.nvidia.com/gpudirect

思腾合力 SITONHOLY | 为AI提供澎湃动力

CONTACT US

400-012-9522
010-86460505

思腾合力
"siton-aiserver"

天津地址： 天津市武清区逸仙园中山大路东D楼一层
北京地址： 北京市海淀区安宁庄西路9号院金泰富地大厦3层317室

www.aiserver.cn

contact@aiserver.cn

视觉计算推动者

[服务器 • DGX • DLI • 集群 • 个性化定制]

欢迎关注思腾合力
官方微信公众号

Thanks !