AI 芯片 – GPU 详解

NVSwitch基础

ZOMI

Talk Overview

1. **硬件基础**
   - GPU 工作原理
   - GPU AI编程本质
2. **英伟达 GPU 架构**
   - GPU基础概念
   - 从 Fermi 到 Volta 架构
   - Turing 到 Hopper 架构
   - Tensor Code 和 NVLink 详解
3. **GPU 图形处理**
   - GPU 逻辑模块划分
   - 算法到 GPU 硬件
   - GPU 的软件栈
   - 图形流水线基础
   - 流水线不可编译单元
   - 光线跟踪流水线

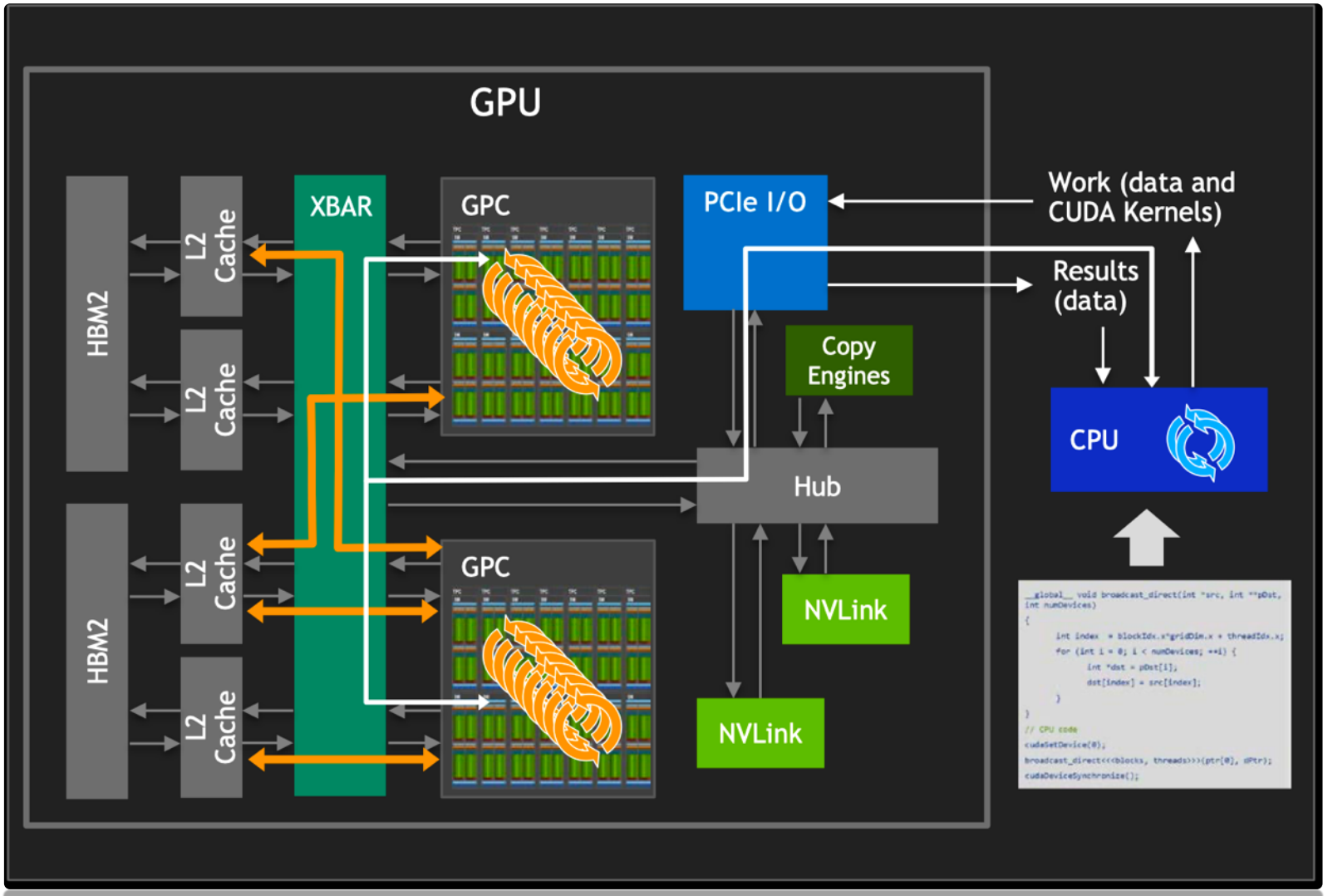
Talk Overview

I. 基本内容

- Appearance – NVSwitch 出现
- Details – NVSwitch 详解
- DGX – DGX 服务器介绍

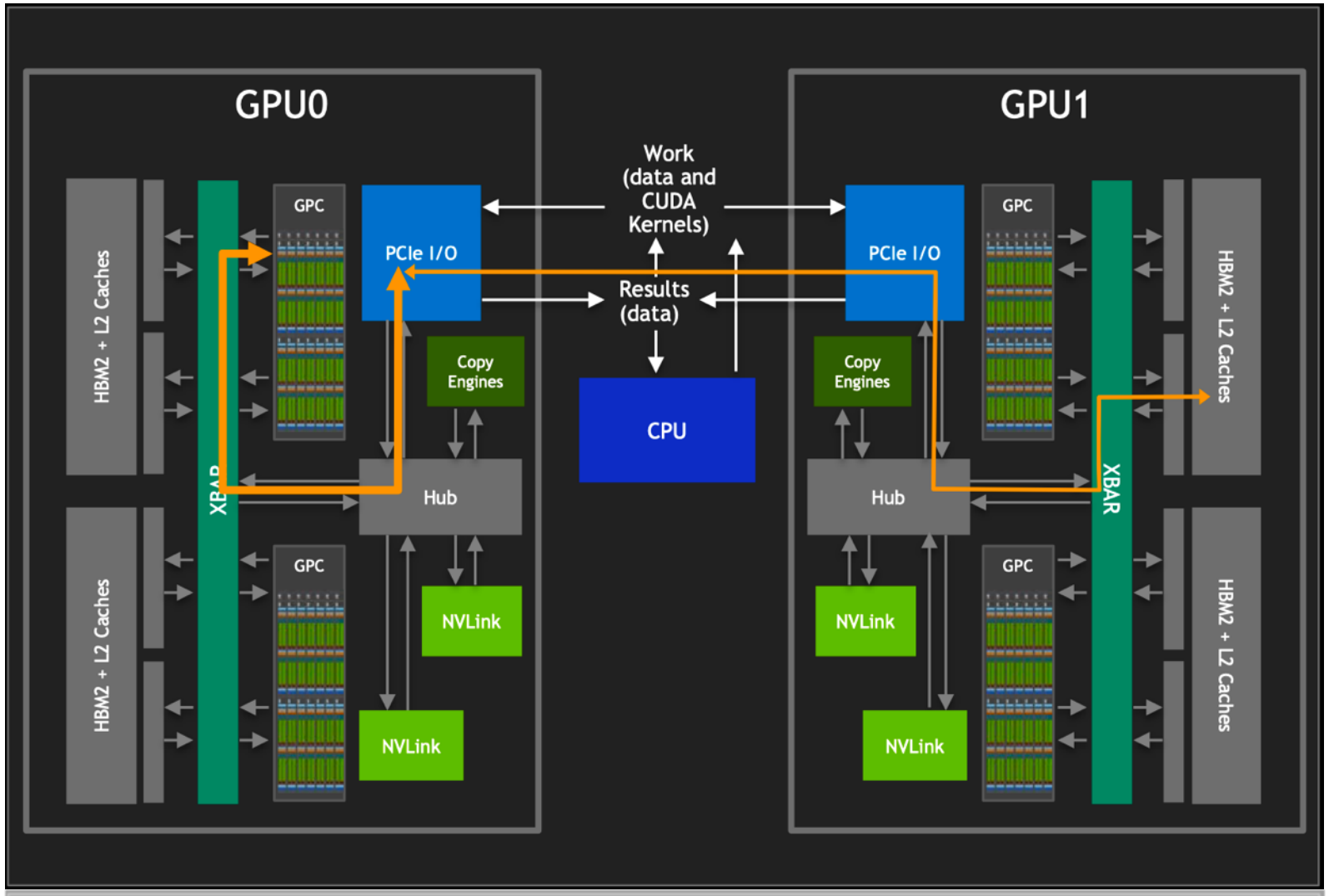
单 GPU 里面多个 SM 核心

- 使用 CUDA 来驱动硬件并行执行真正的计算；
- GPU 把线程工作分配给每个 GPC/SM cores；
- GPC/SM cores 利用 HBM2 中的数据进行计算；
- GPC/SM cores 之间可以共享 HBM2 中的数据；



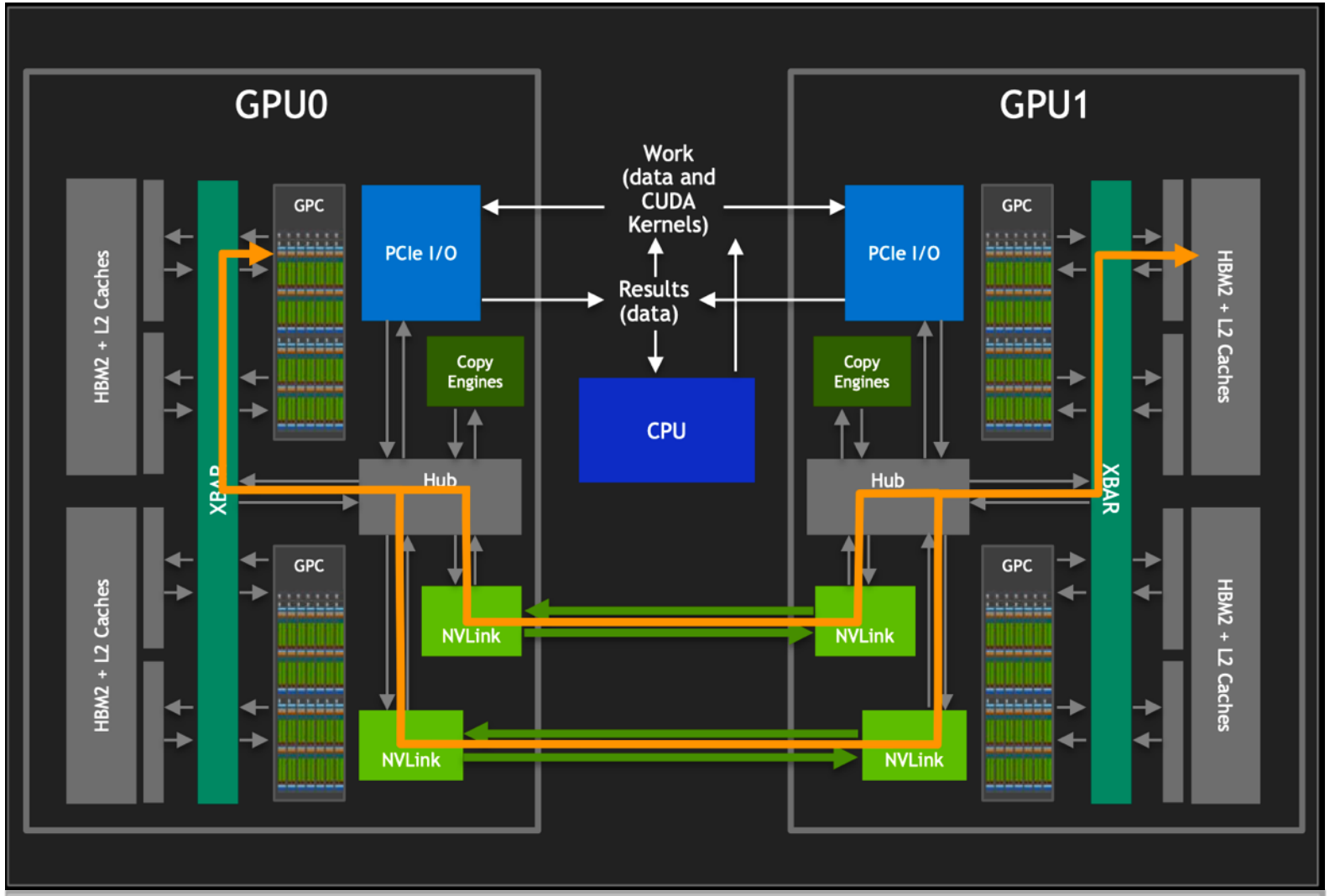
GPU 间通过 PCIe 通信

- 如果要对其他 GPU 的 HBM 2 进行访问，需要走 PCIe；
- GPU-to-GPU 之间的交互需要通过 CPU 进行分配调度；
- PCIe 的带宽限制了 GPU-to-GPU 的速率；



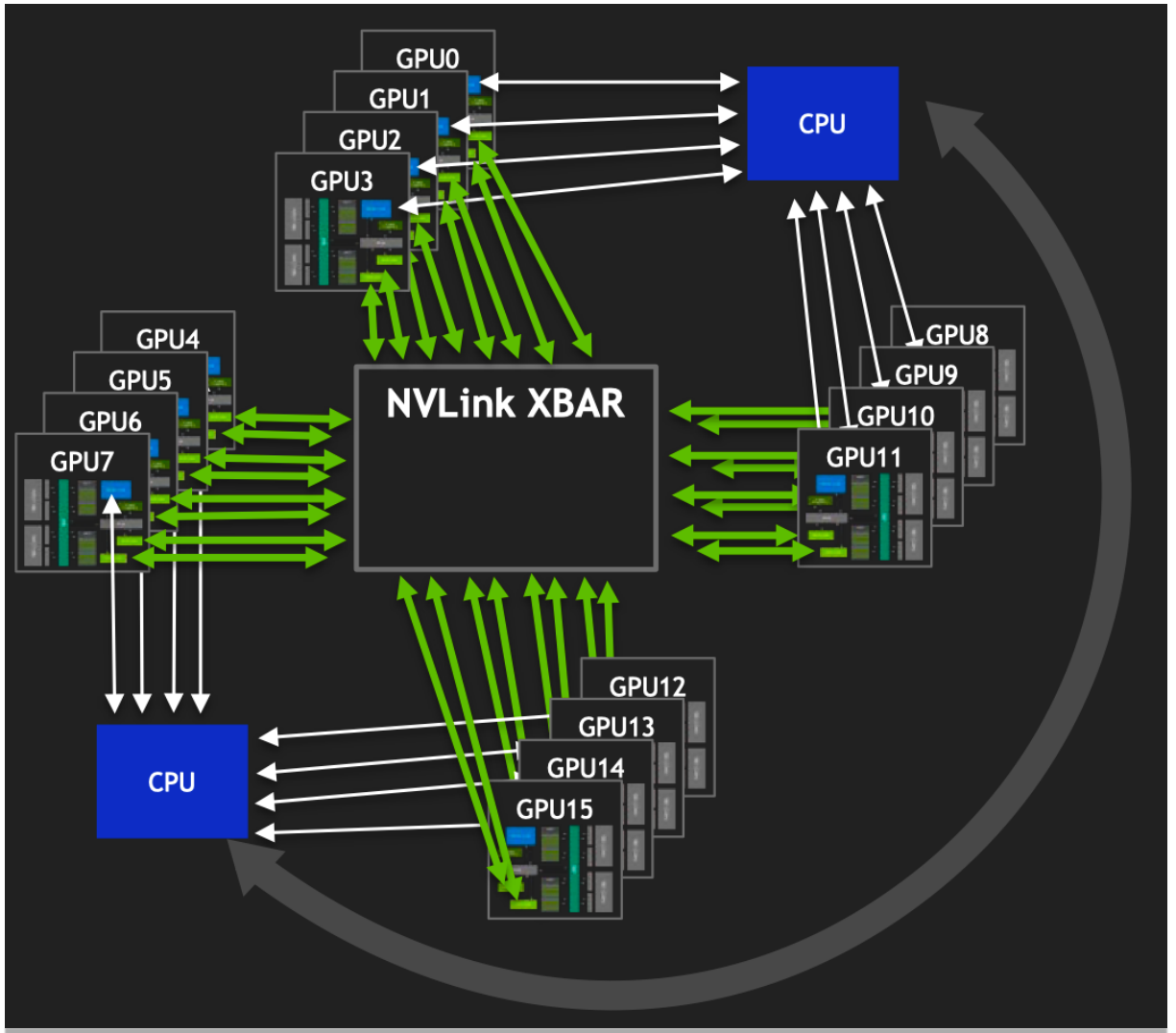
GPU 间使用 NVLink

- GPCs 可以访问卡间 HBM2 内存数据；
- 通过多条 NVLink 来对其他 GPU 内的 HBM2 数据进行访问；
- 成为了 XBARs 的桥梁，并且与 PCIe 不冲突，作为互补方案；



GPU 间互联

- NVLink 可以使得更多的GPU间进行互联；
- 实现单个 GPU 驱动进程可以控制所有GPU的计算任务；
- HBM2 可以在不受其他进程干扰下下访问（LD/ST指令、RDMA）；
- XBAR 作为桥接器可以独立演进发展，提供更高的带宽；



NVSwitch的出现

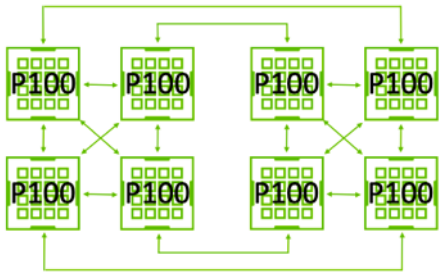
NVLink 发展

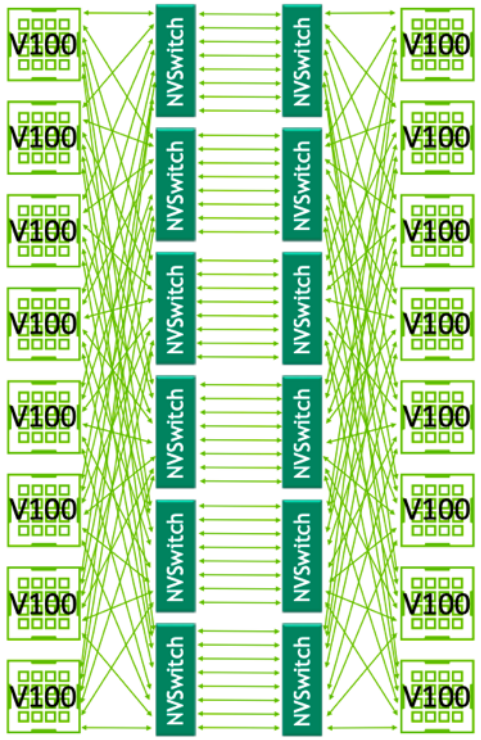
	First Generation	Second Generation	Third Generation	Fourth Generation
NVLink bandwidth per GPU	160GB/s	300GB/s	600GB/s	900GB/s
Maximum Number of Links per GPU	4	6	12	18
Supported NVIDIA Architectures	NVIDIA Pascal architecture	NVIDIA Volta architecture	NVIDIA Ampere Architecture	NVIDIA Hoppe Architecture

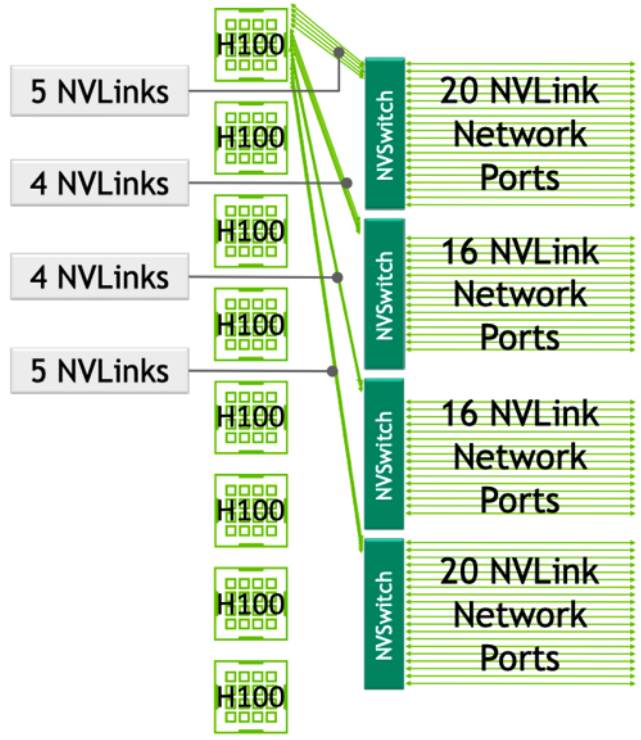
NVSwitch发展

	First Generation	Second Generation	Third Generation
Number of GPUs with direct connection / node	Up to 8	Up to 8	Up to 8
NVSwitch GPU-to-GPU bandwidth	300GB/s	600GB/s	900GB/s
Total aggregate bandwidth	2.4TB/s	4.8TB/s	7.2TB/s
Supported NVIDIA architectures	NVIDIA Volta architecture	NVIDIA Ampere architecture	NVIDIA Hopper architecture

NVLink 与 NVSwitch 相关的服务器





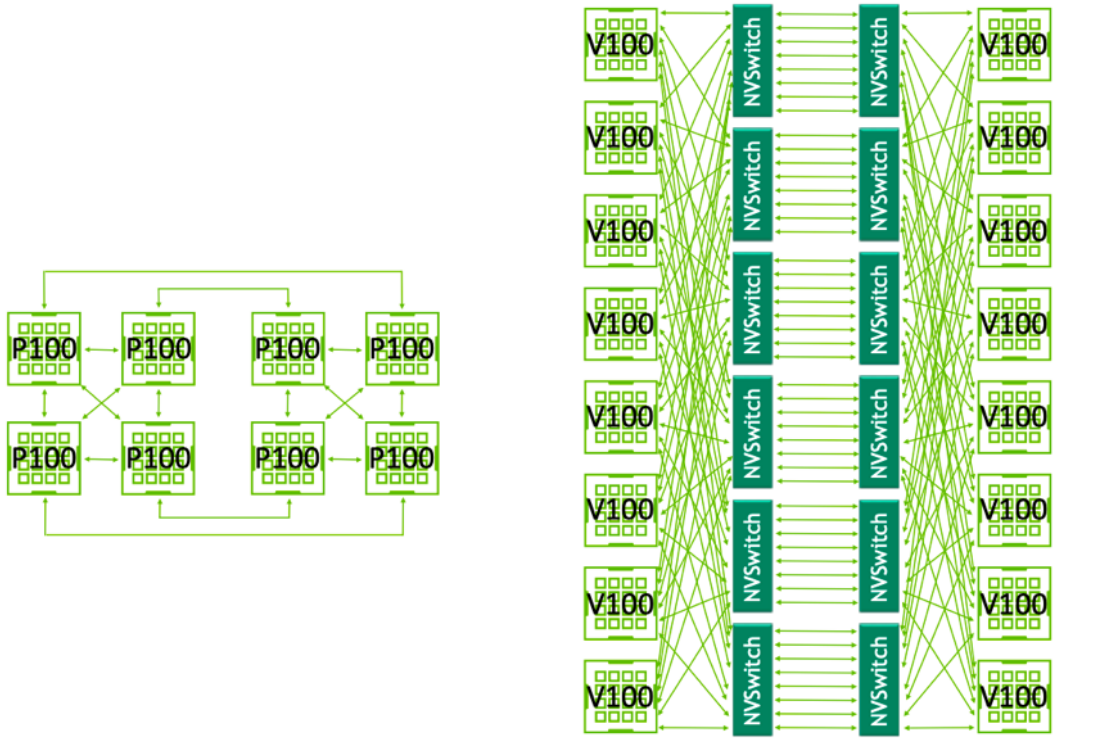


2016	2018	2020	2022
DGX-1 (P100)	DGX-2 (V100)	DGX A100	DGX H100
140GB/s Bisection BW 40GB/s AllReduce BW	2.4TB/s Bisection BW 75GB/s AllReduce BW	2.4TB/s Bisection BW 150GB/s AllReduce BW	3.6TB/s Bisection BW 450GB/s AllReduce BW

NVLink 与 NVSwitch 相关的服务器

- NVLink 2.0 技术使得单服务器中 8 个 GPU 无法达到全连接为解决该问题，NVIDIA 在 2018 年发布了 NVSwitch，实现了 NVLink 的全连接。

- NVIDIA NVSwitch 是首款节点交换架构，可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 对分别达到 300GB/s 的速度同时进行通信。



2016

DGX-1 (P100)

140GB/s Bisection BW
40GB/s AllReduce BW

2018

DGX-2 (V100)

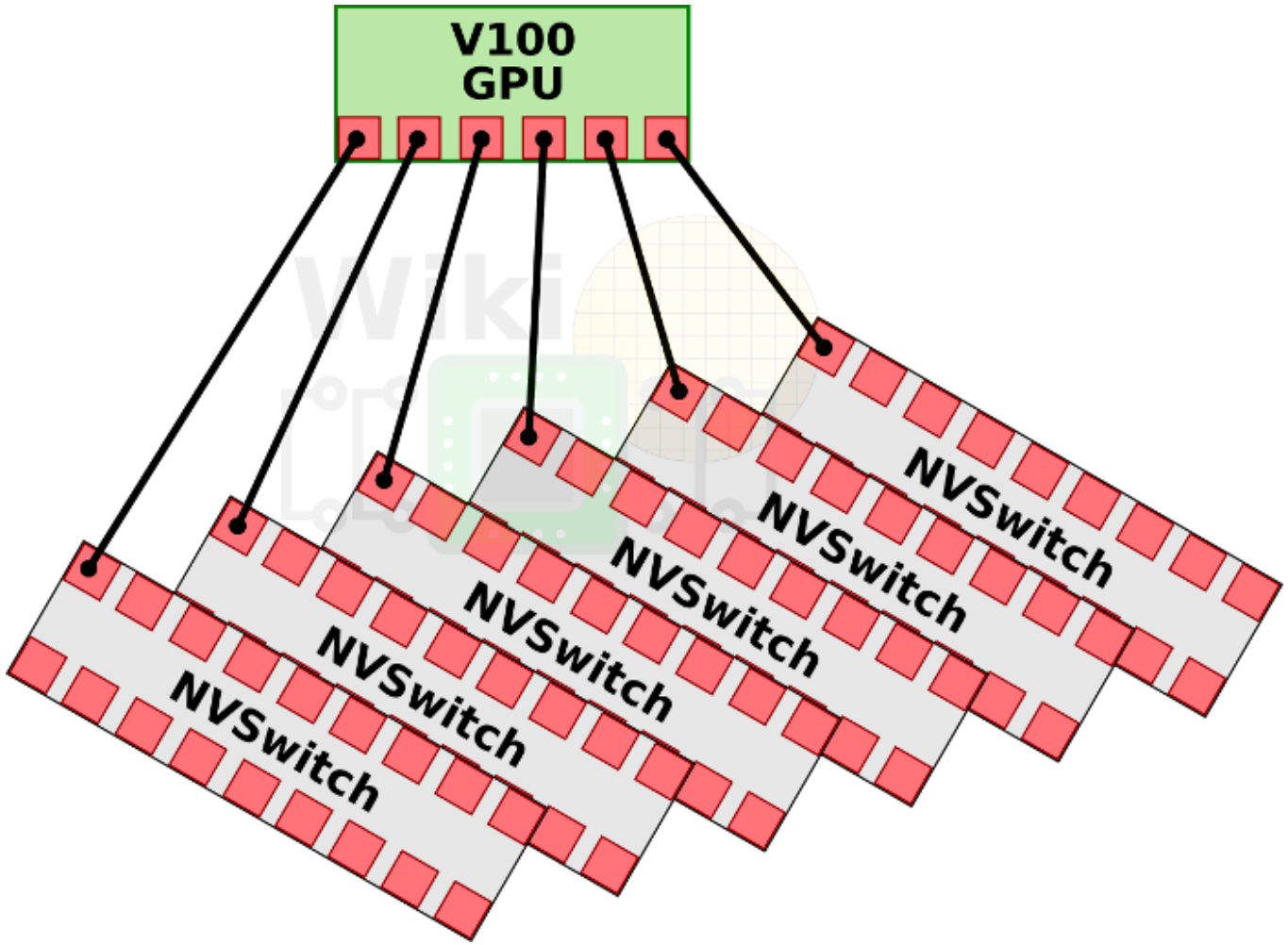
2.4TB/s Bisection BW
75GB/s AllReduce BW

NVSwitch详解

第一代 NVSwitch

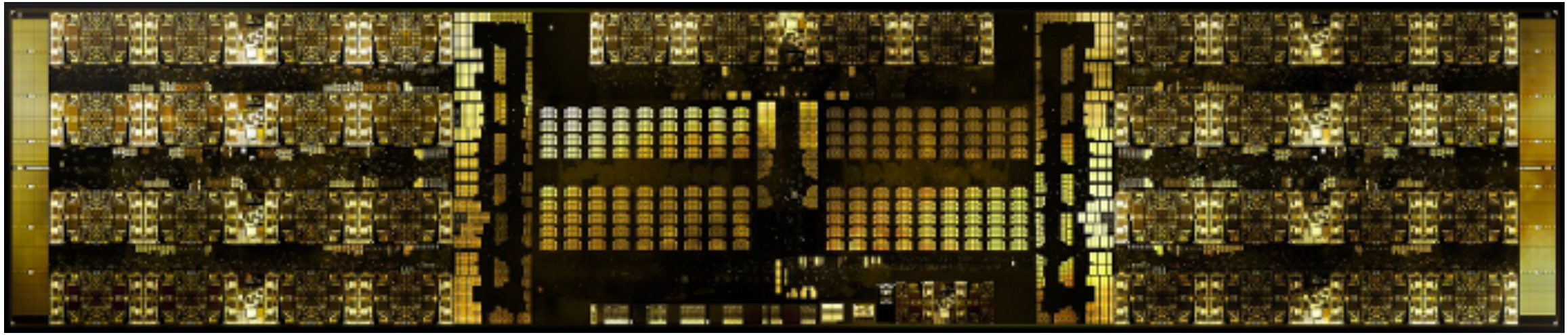
- 通过 NVSwitch 提供的18路接口， NVSwitch 能够让 nvidia设计出完全无阻塞的全互联16路GPU系统。

- 每块v100中的6路 NVLink 将分别连接到6块 NVSwitch 上面。这样8块v100与6块 NVSwitch 完全连接，构成一个基板。



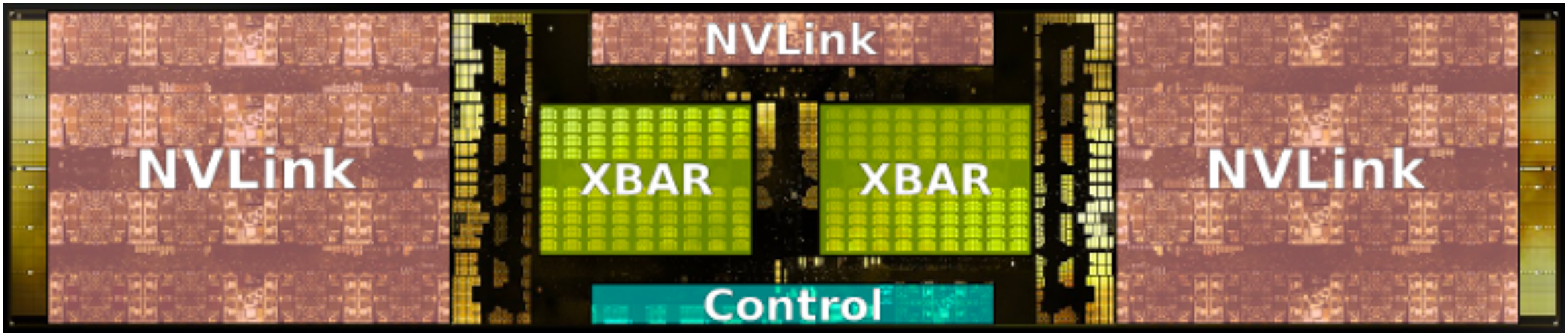
第一代 NVSwitch

- NVSwitch 是一块独立的 NVLink 芯片，其提供了高达18路 NVLink 的接口。支持 NVLink 2.0，每个接口均能提供双信道高达50GB/s 带宽，总计能够提供900GB/s的带宽。功率100w，基于台积电12nm FinFet FFN 工艺，拥有2b个晶体管。

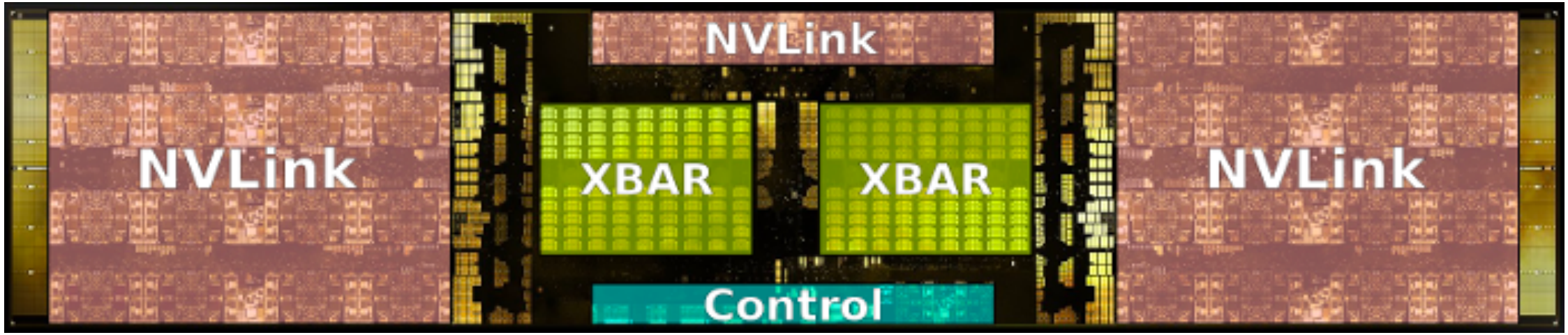


第一代 NVSwitch

- NVSwitch 封装在1940 个 pin 大小为 4cm2 的 BGA芯片中，其中 576 个针脚专门服务于 18 路的 NVLink，剩下的针脚用于电源和 I/O 接口，如管理端口x4 pcie、I2c、GPIO等。



NVLink2 NVSwitch1 特点



Parameter	Spec
每NVLink双向带宽	51.5 GBps
NRZ Lane Rate (x8 per NVLink)	25.78125 Gbps
Transistors	2 Billion
Process	TSMC 12FFN
Die Size	106 mm^2

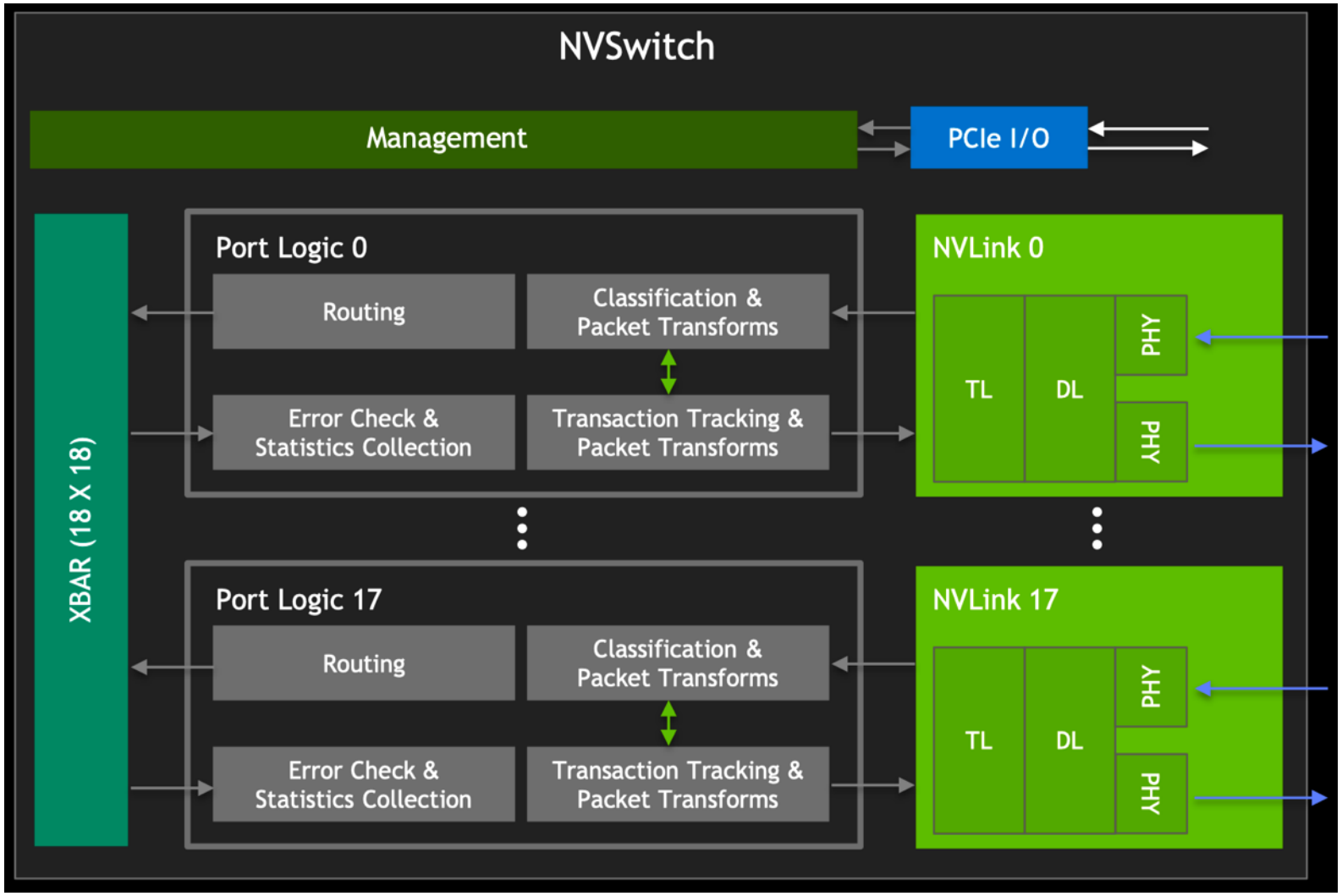
Parameter	Spec
Bidirectional Aggregate Bandwidth	928 GBps
NVLink Ports	25.78125 Gbps
Mgmt Port (config, maintenance, errors)	PCIe
LD/ST BW Efficiency (128B pkts)	80.0%
Copy Engine BW Efficiency (256B pkts)	88.9%

NVSwitch：DIE 电路

- **电路设计紧凑**：在电路上数据包的处理和交换（XBAR）逻辑非常紧凑
- **提升IO面积**：I/O块占据了50%以上的面积，最大限度地提高物理链接电路的面积比
- **简化封装衬底布线**：所有NVLink端口在电路上都是并行的

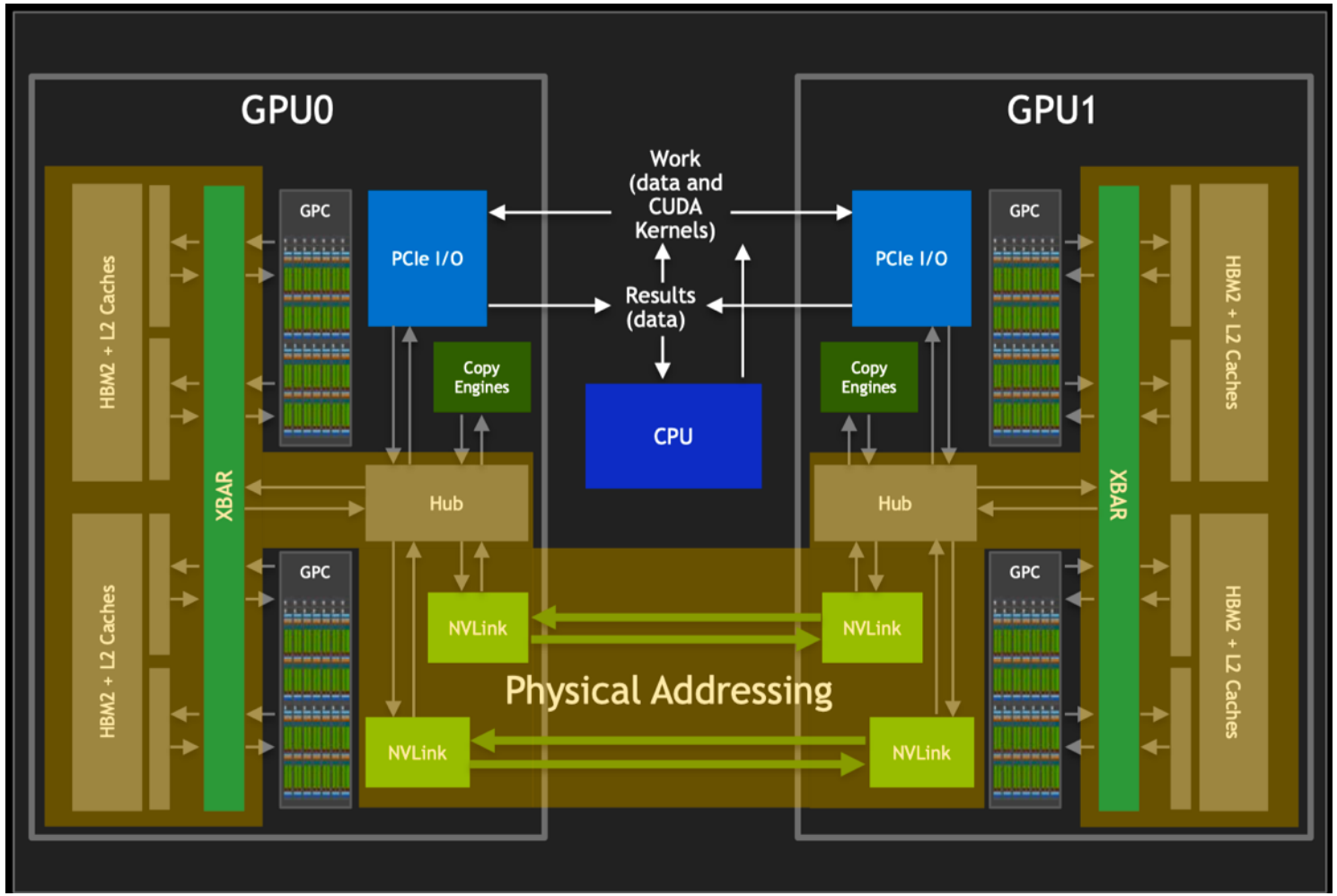
第一代 NVSwitch Block

1. GPU XBAR作为NVLink的桥接设备，非一般网络设备；
2. 数据包在多个GPU上流动/交换，客户面感知单个GPU；
3. 基于 SRAM 缓冲，XBAR 非阻塞；
4. 从V100重新使用NVLink IP块和XBAR设计/验证能力；



NVSwitch：物理共享内存

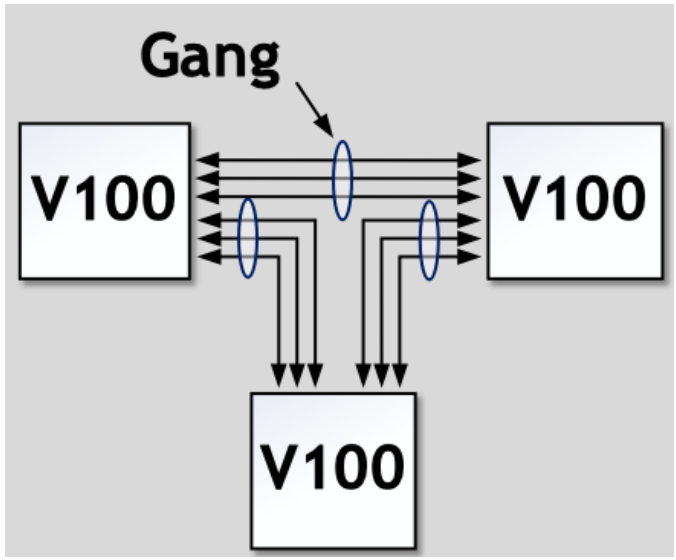
- 虚拟到物理地址转换GPC完成
- 物理地址在NVLink数据包中传输



简化互联交换方式

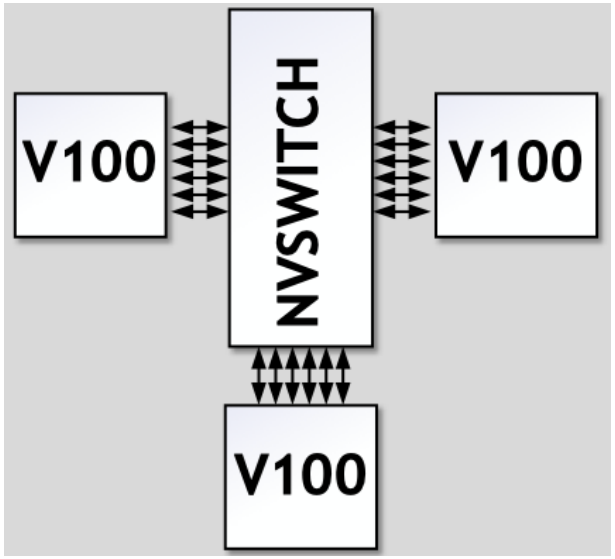
Without NVSwitch

- GPU 间直接连接
- 将NVLinks聚合为多个组 Gang
- GPU-GPU 间最大带宽受限于 Gang



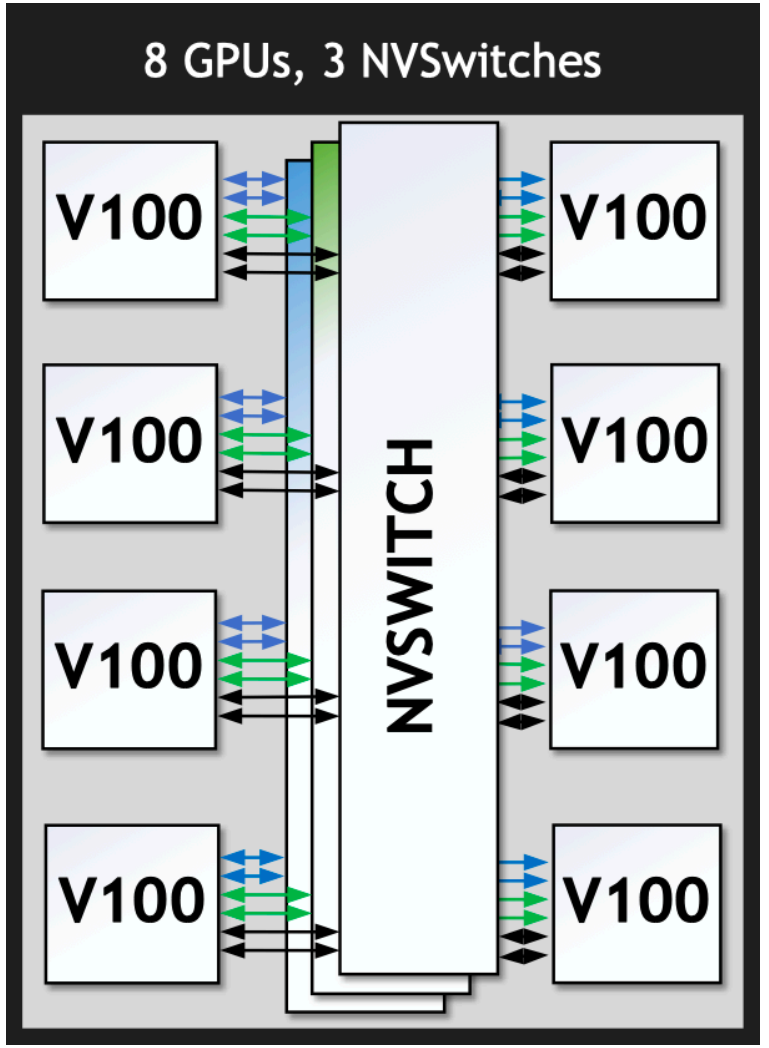
With NVSwitch

- 所有链路上数据可交互，任何一对GPU互联
- 只要不超过六个NVLink的总带宽，单GPU流量非阻塞

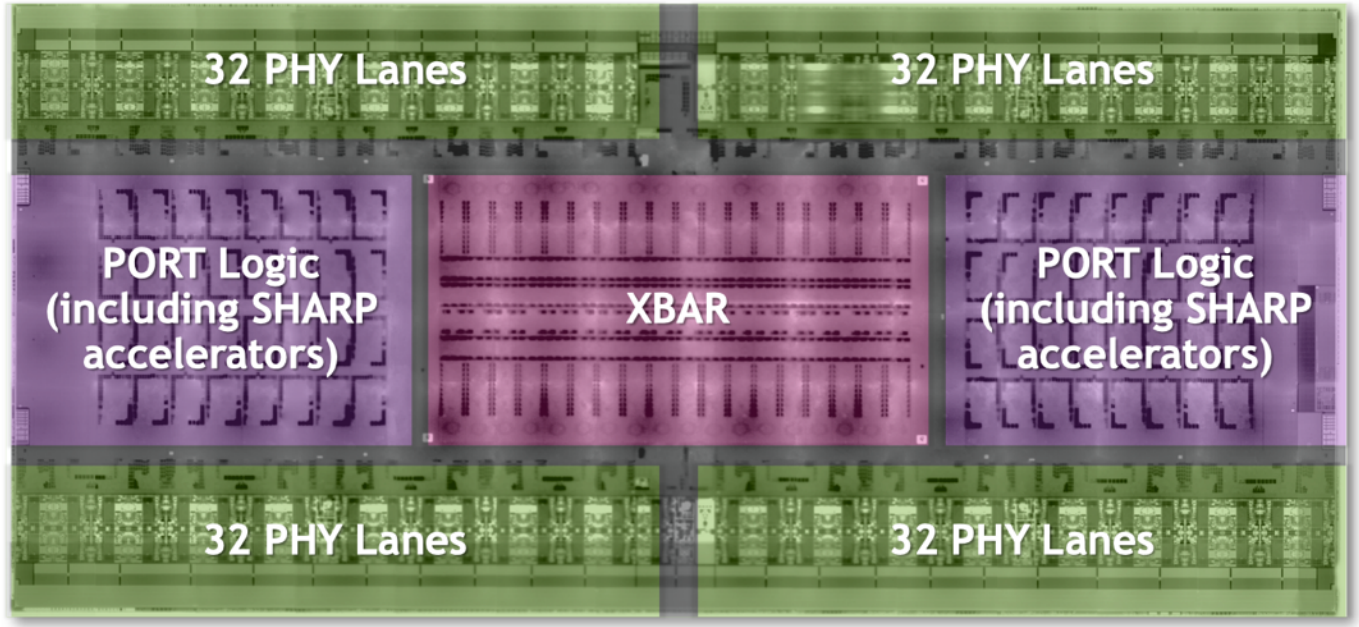


提供多个GPU间互联系统

1. 可以添加更多NVSwitch来支持更多GPU
2. 八个GPU可以用三个NVSwitch构建
3. 数据可以在所有GPU链路上交互
4. 任何一对GPU可以使用完整300GBps双向带宽通信
5. NVSwitch XBAR提供从A-B点唯一路径，实现通信无阻塞、无干扰



NVLink4 NVSwitch3 特点



最大 NVSwitch 模块

- TSMC 4N process
- 25.1B transistors
- 294mm2
- 50mmX50mm package (2645 balls)

最高的带宽

- 64 NVLink4 ports (x2 per NVLink)
- 3.2TB/s full-duplex bandwidth
- 50Gbaud PAM4 diff-pair signaling
- All ports NVLink Network capable

新能力

- 400GFLOPS of FP32 SHARP
- NVLink Network management, Securit y and telemetry engines

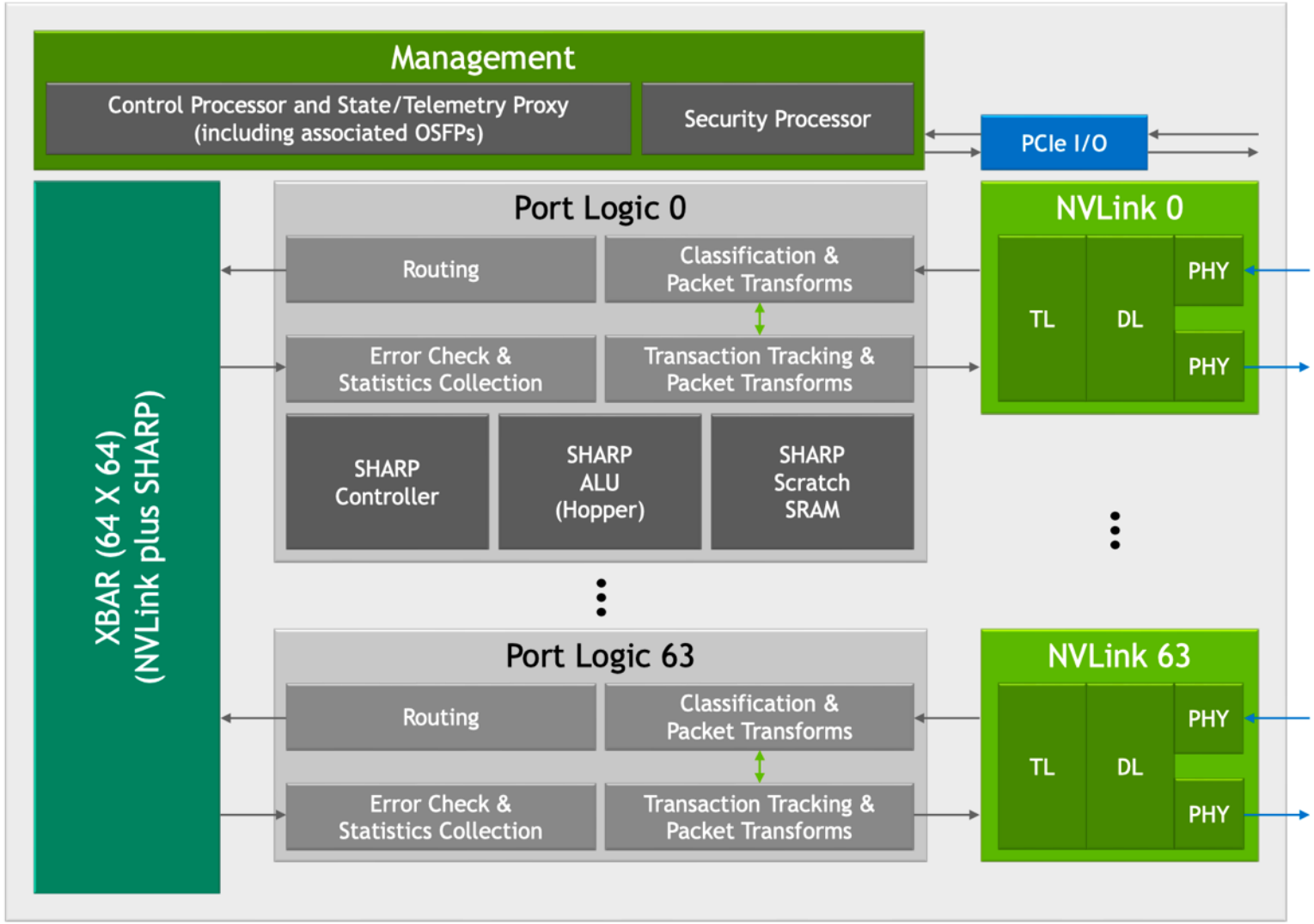
第三代 NVSwitch Block

新SHARP模块

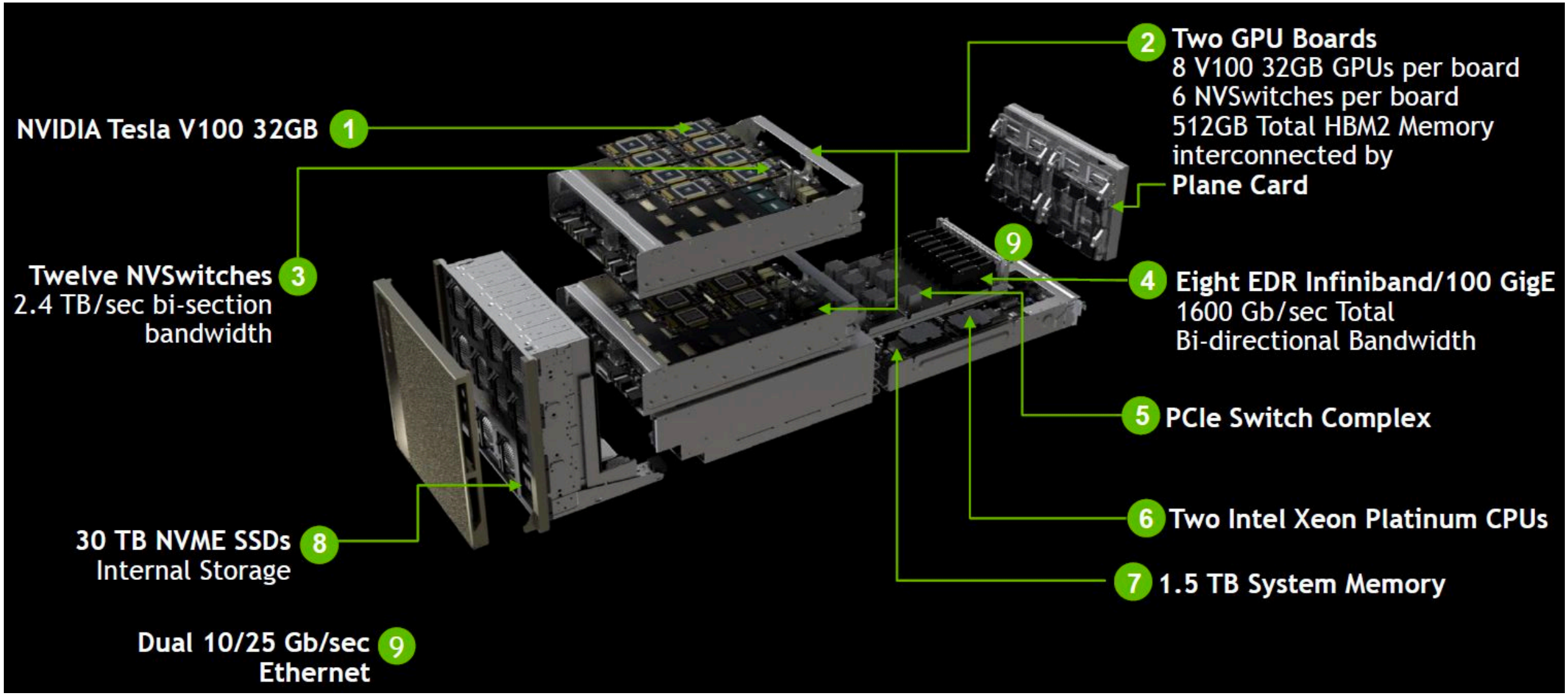
- 支持多种运算符（逻辑、加减等）和数据格式（FP16/BF16）
- SHARP控制器可以并行管理多达128个SHARP组
- XBAR的带宽调整到能够跟匹配SHARP相关的数据传输

新NVLink模块

- 安全处理器保护数据和芯片
- 分区功能将端口隔离到NVLink网络
- 控制器支持下一代OSFP电缆



DGX 服務器


NVIDIA Tesla V100 32GB 1
2 Two GPU Boards
8 V100 32GB GPUs per board
6 NVSwitches per board
512GB Total HBM2 Memory
interconnected by
Plane Card
Twelve NVSwitches 3
2.4 TB/sec bi-section
bandwidth
9
4 Eight EDR Infiniband/100 GigE
1600 Gb/sec Total
Bi-directional Bandwidth
5 PCIe Switch Complex
6 Two Intel Xeon Platinum CPUs
30 TB NVME SSDs 8
Internal Storage
7 1.5 TB System Memory
Dual 10/25 Gb/sec 9
Ethernet

Reference 引用&参考

1. https://fuse.wikichip.org/news/1224/a-look-at-nvidias-nvlink-interconnect-and-the-nvswitch/
2. https://blog.csdn.net/BtB5e6Nsu1g511Eg5XEg/article/details/86762135
3. https://developer.nvidia.com/blog/upgrading-multi-gpu-interconnectivity-with-the-third-generation-nvidia-nvswitch/
4. https://www.nextplatform.com/2016/05/04/nvlink-takes-gpu-acceleration-next-level/
5. https://www.servethehome.com/nvidia-nvlink4-nvswitch-at-hot-chips-34/
6. https://www.servethehome.com/nvidia-nvswitch-details-at-hot-chips-30/

   https://hc34.hotchips.org/
7. https://www.infoq.cn/article/3d4msrvs8zotgcj7*krt
8. https://zhuanlan.zhihu.com/p/399405214
9. https://www.zhihu.com/question/63219175
10. https://developer.aliyun.com/article/591403
11. https://developer.aliyun.com/article/603617
12. https://developer.aliyun.com/article/599183
13. https://blog.csdn.net/tony_vip/article/details/117131380