中山大學
SUN YAT-SEN UNIVERSITY

国家超级计算广州中心
NATIONAL SUPERCOMPUTER CENTER IN GUANGZHOU

Advanced Computer Architecture
高级计算机体系结构

第8讲：DLP and GPU (3)

GPU虚拟化

邓志涛

DCS5367, 11/23/2021

Self Introduction

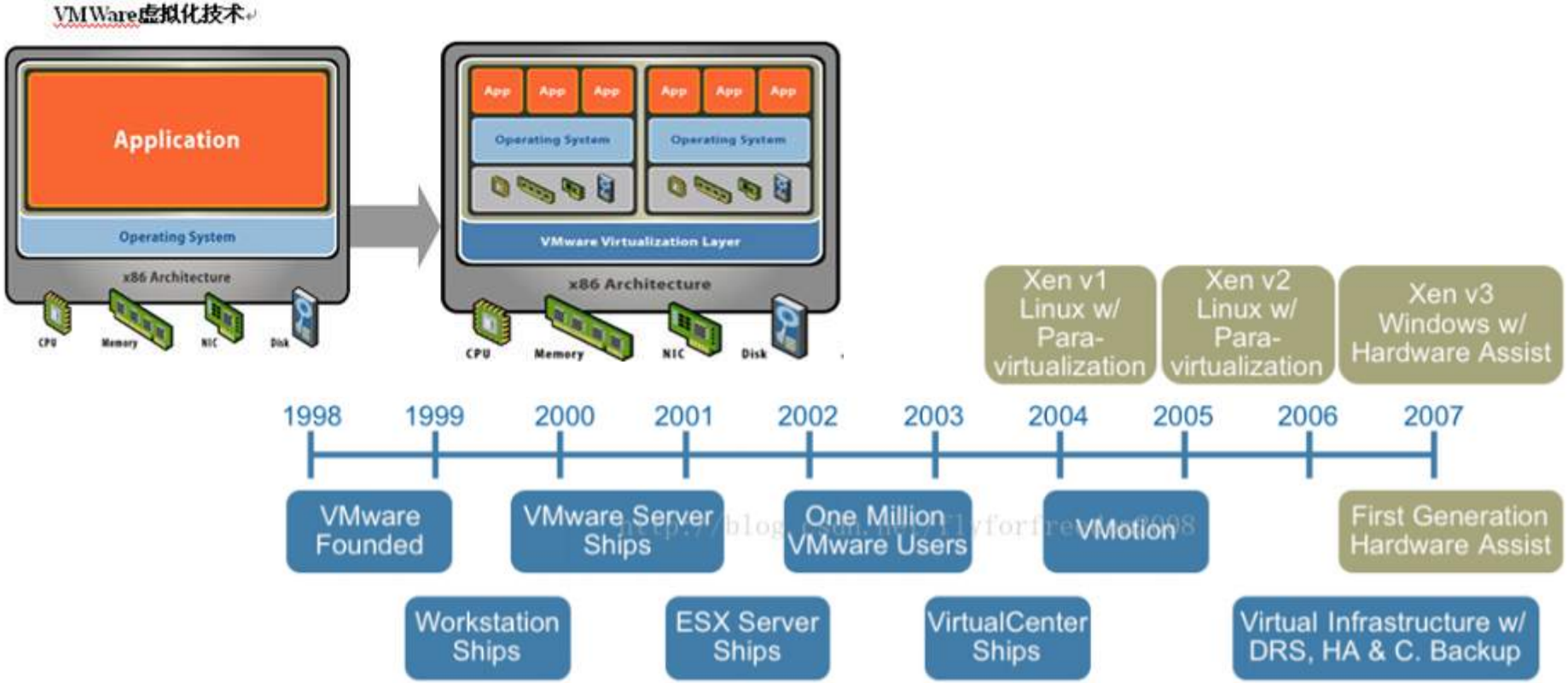
- ZHITAO DENG(邓志涛)
- Educational background
  - Bachelor degree： Northeastern University 东北大学
  - Master degree： Aizu University(Japan) 会津大学
  - Doctor： SUN YAT-SEN University 中山大学
- Work experience：
  - SOLEKIA Limited(Fujitsu Partner)
  - NSCC-GZ

Agenda

- Background
- Motivation
- GPU Virtualization
- Conclusion

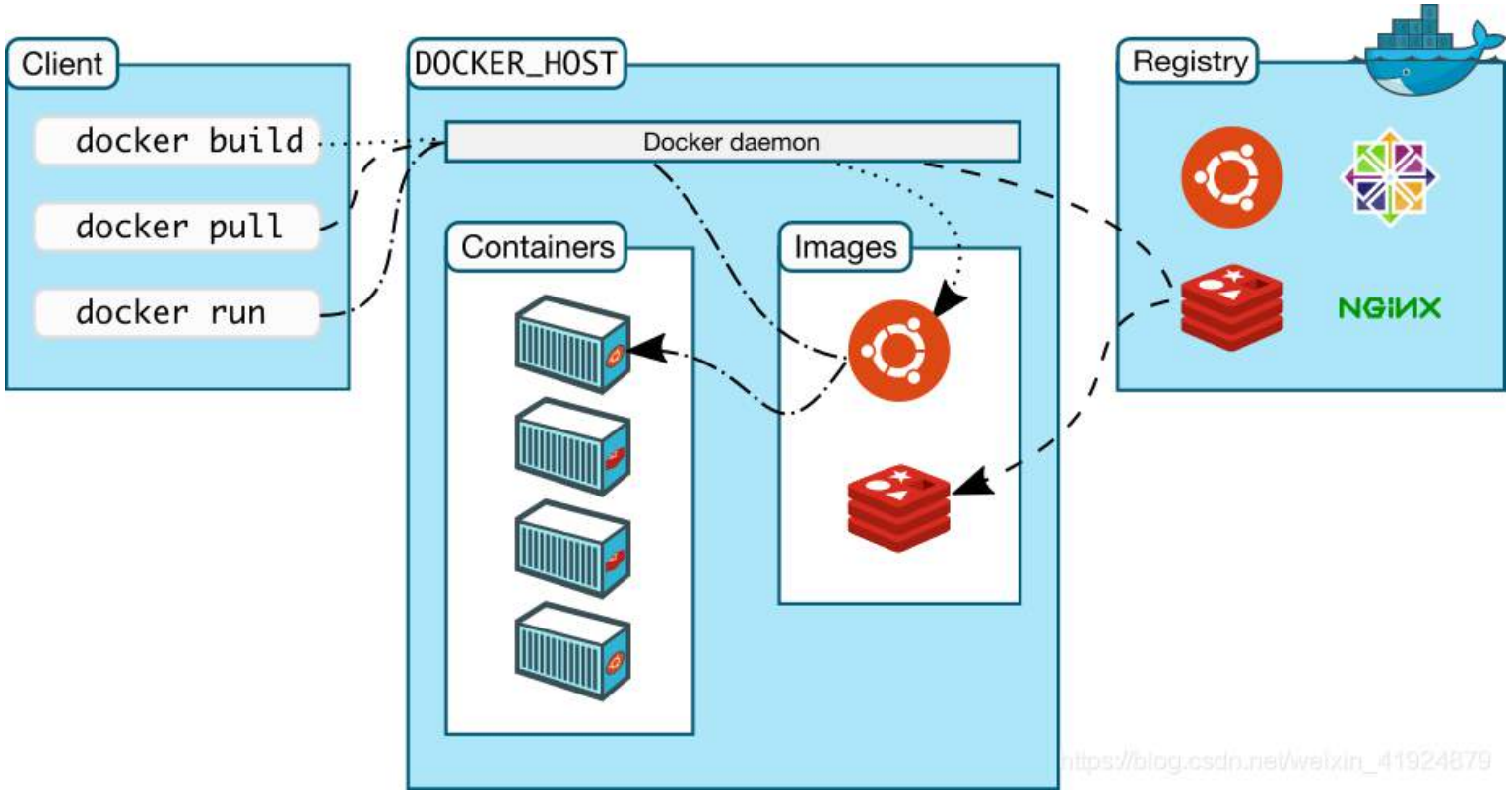
Background

- VM (Virtual Machine，虚拟机，虚机)
  - Combine binary translation and direct instruction execution.(调用转化)
  - Pooling Tech(池化技术)



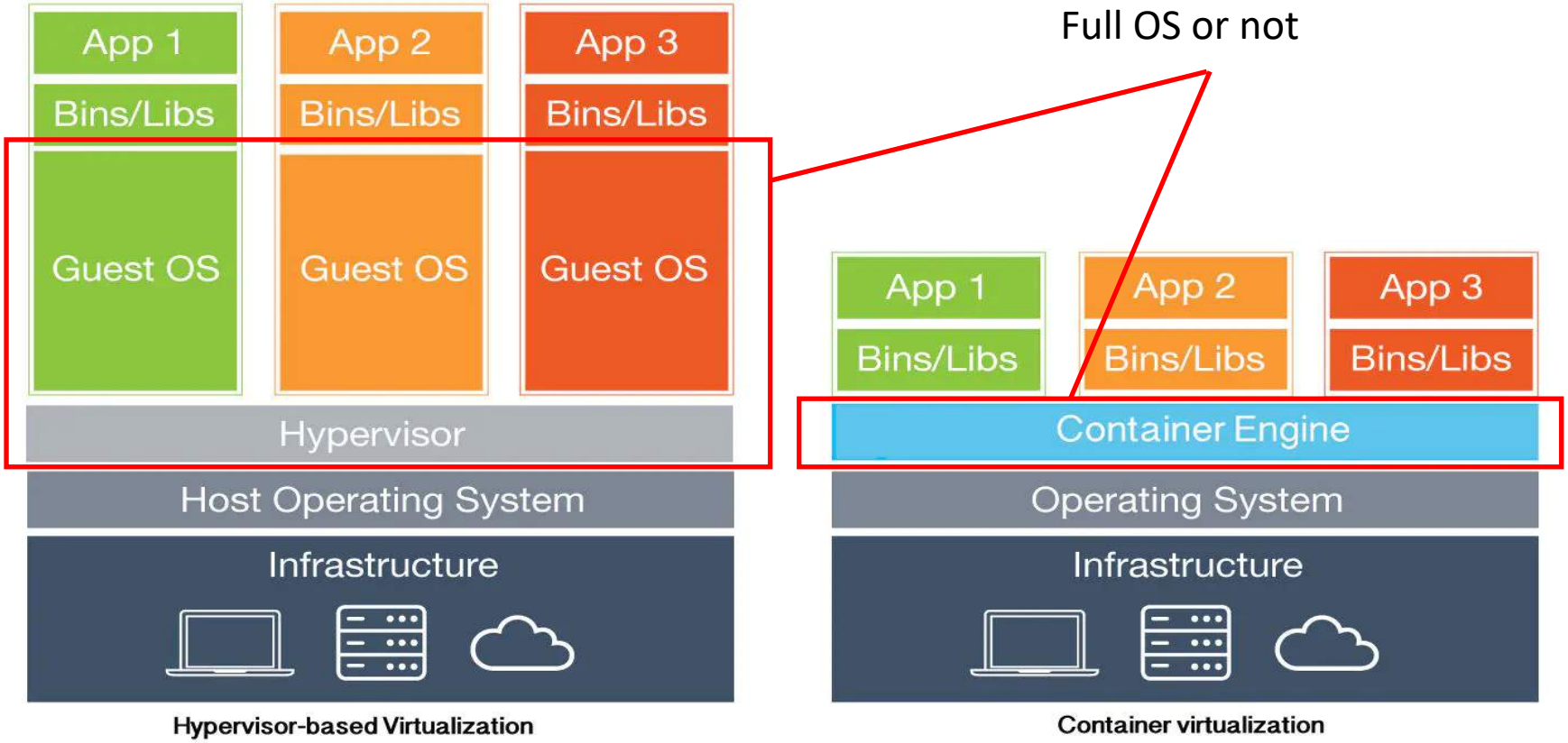
Background

- Container(容器)
  - Lightweight（轻量级）
  - Easy for starting up



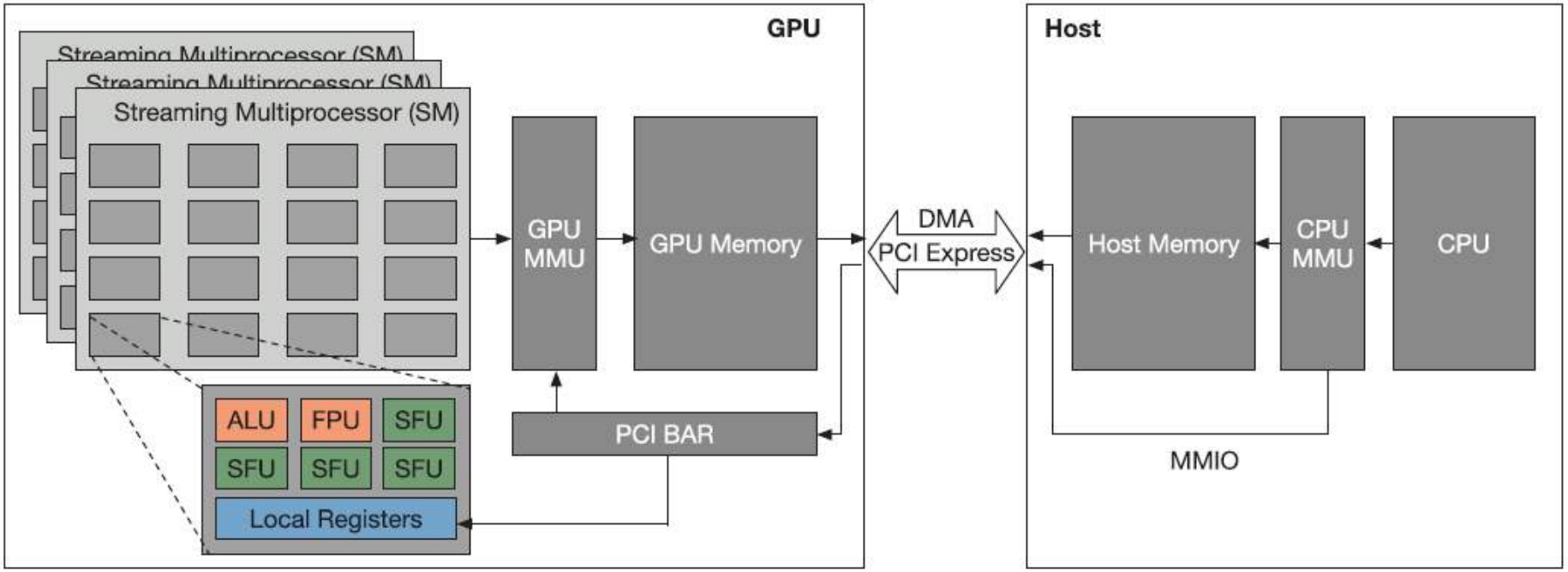
Background

- Different between VM and Container.(虚机与容器)



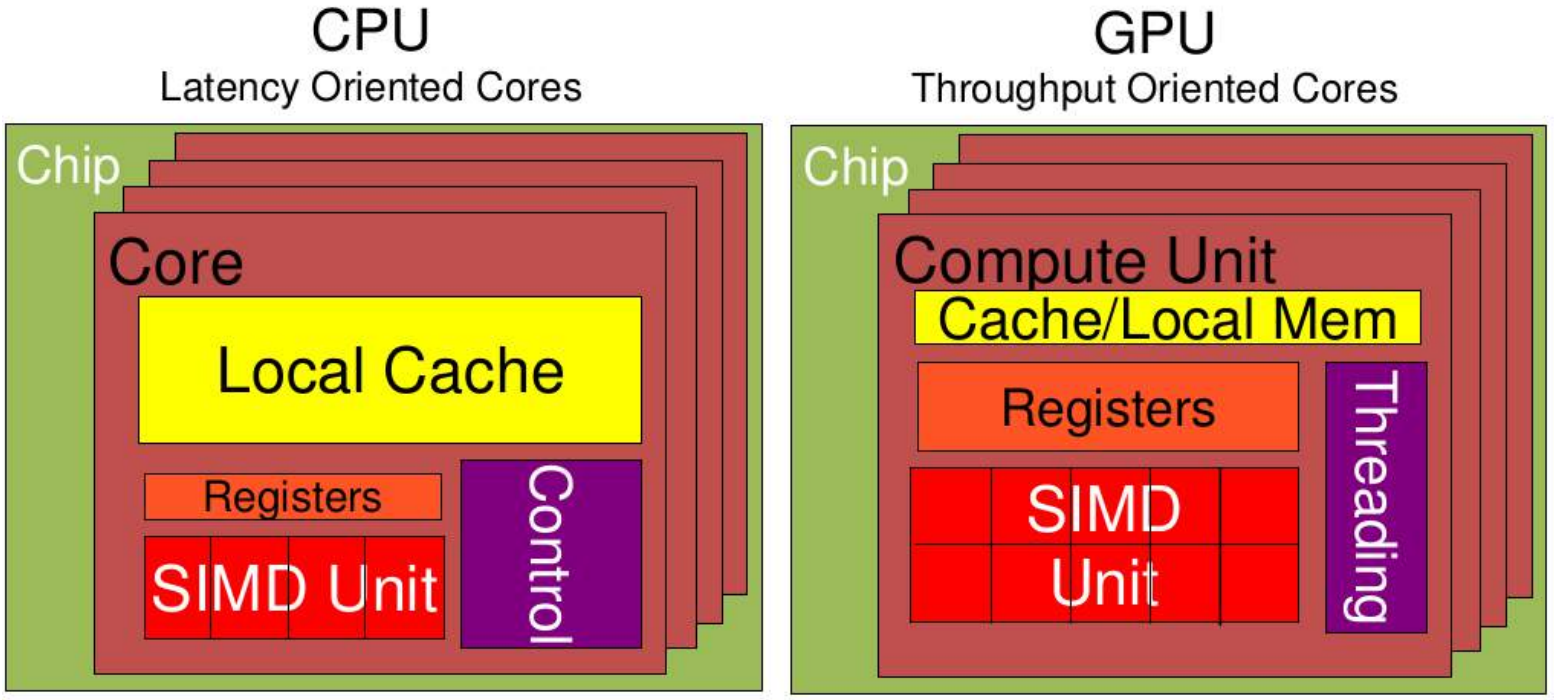
Background

- Architecture of a heterogeneous system



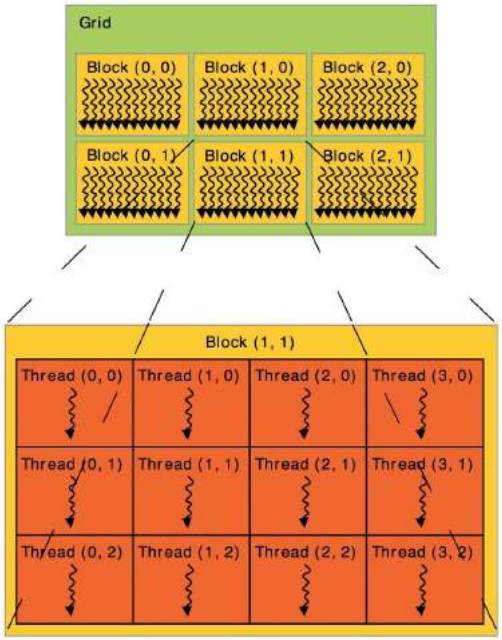
Background

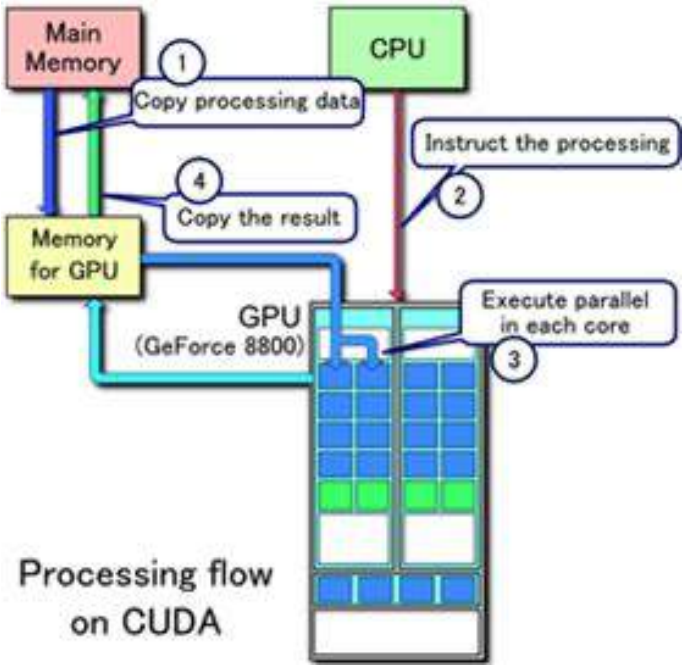
- GPU(Graphics Processing Unit)
  - Throughput Oriented(面向吞吐量)
  - Low Local Cache(缓存较低)



Background

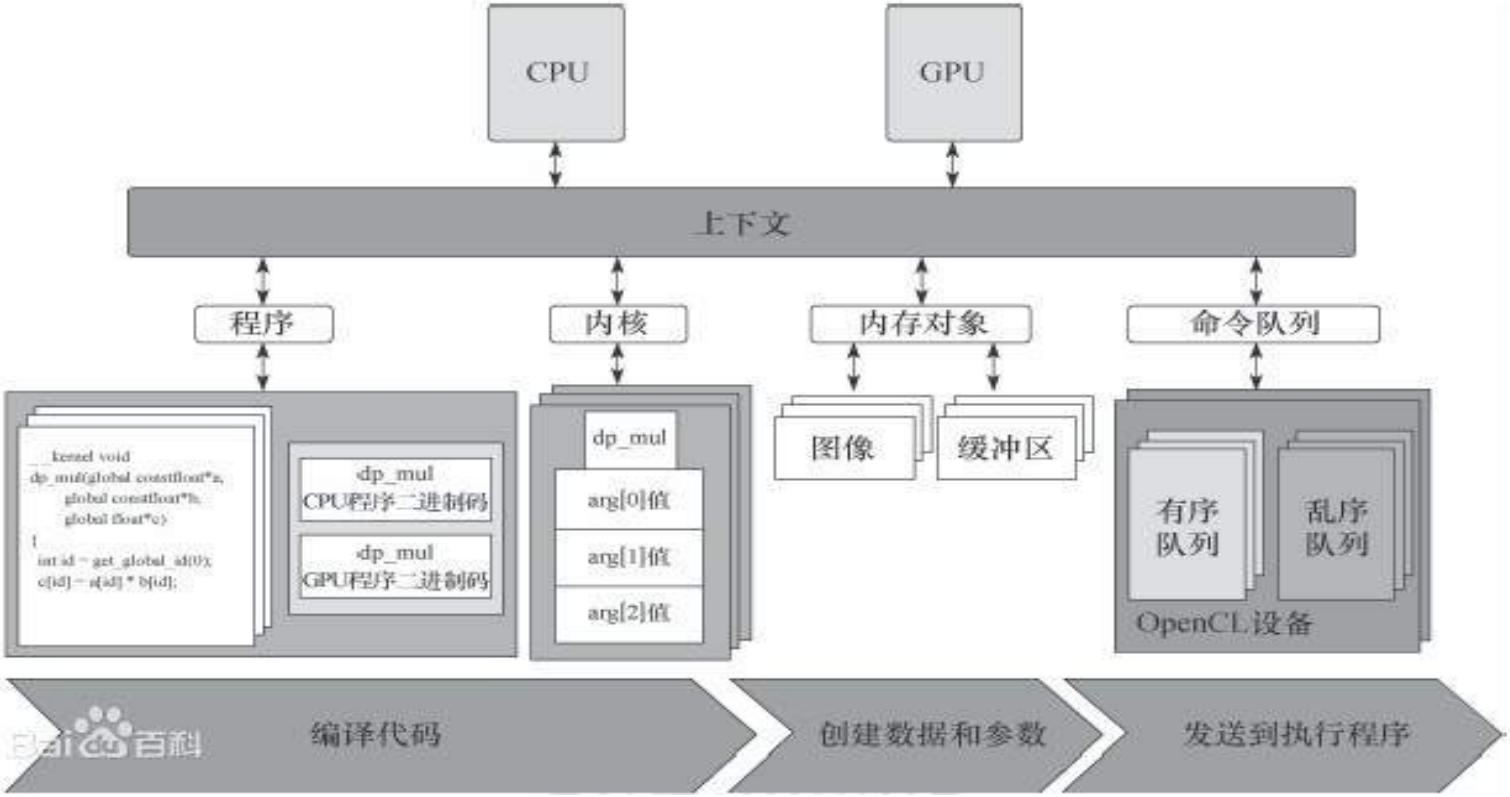
- CUDA(Compute Unified Device Architecture)
  - Design for Nvidia GPU.(面向NVIDIA)
  - Host code and Device code





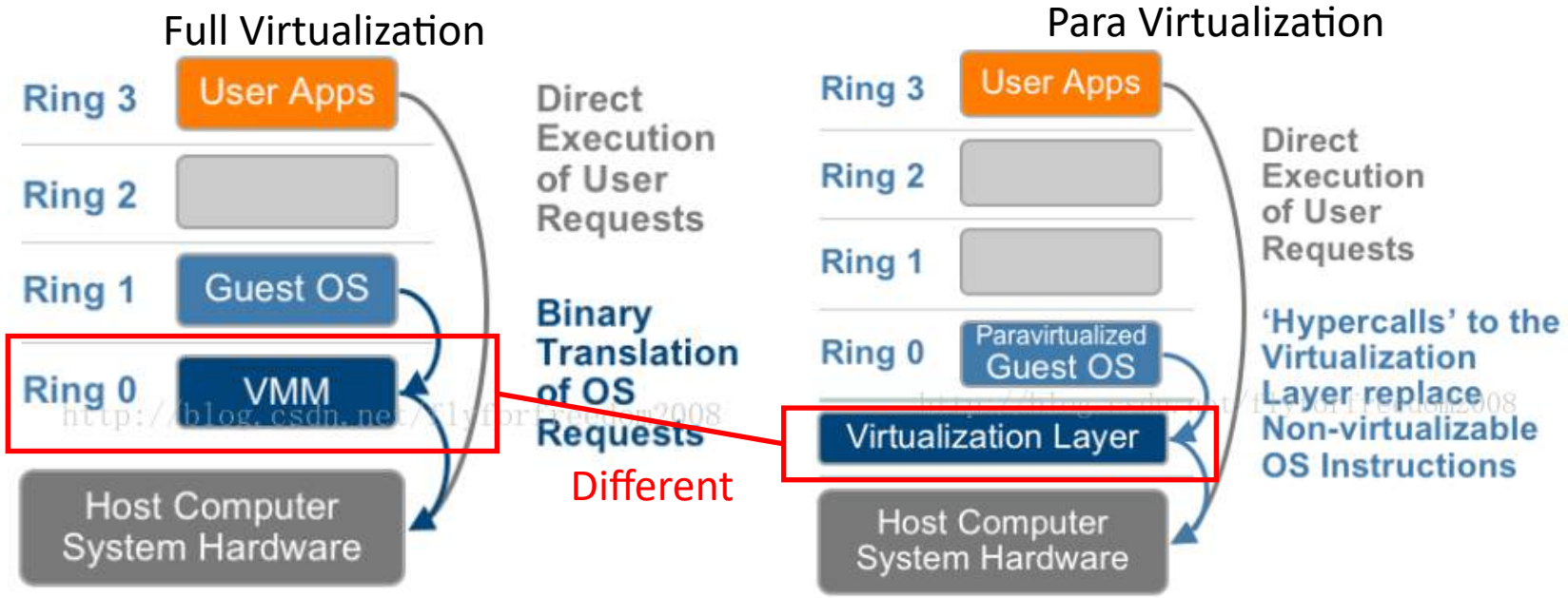
Background

- OpenCL（Open Computing Language）
  - Design by Apple, for all accelerator(Motivation).
  - Context, kernel Function ,global function



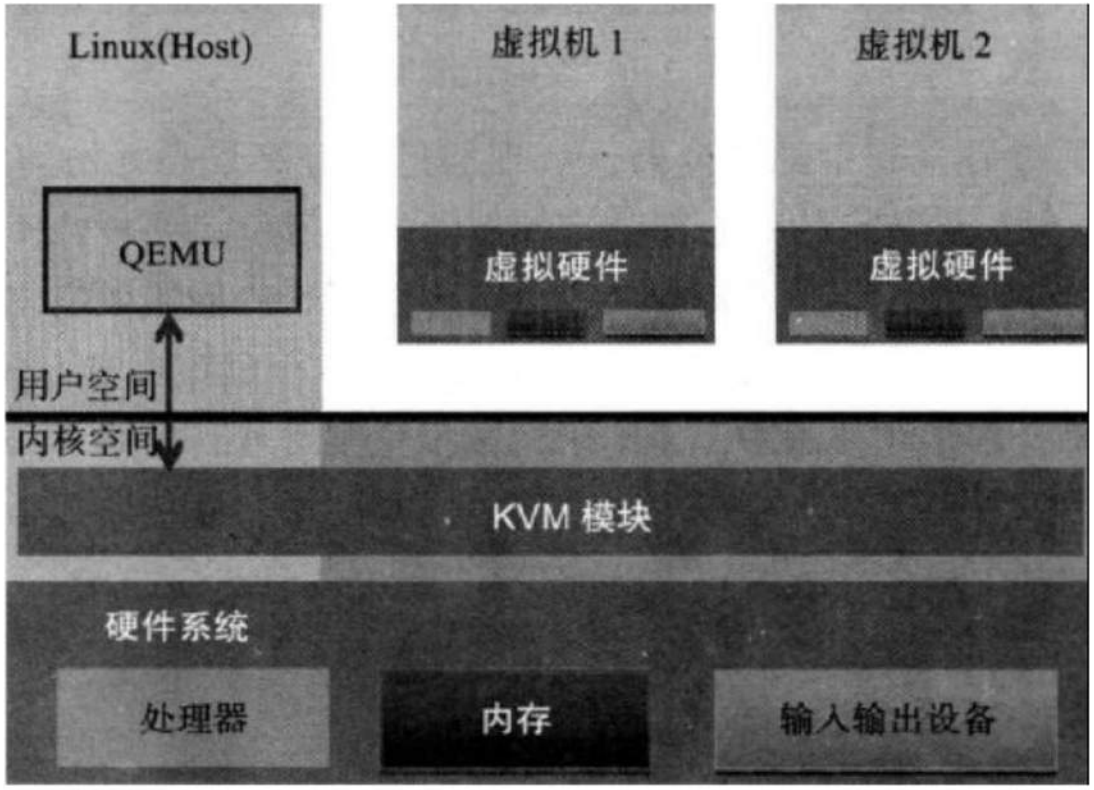
Background

- Full & Para Virtualization(全&半虚拟化)
  - Decouple with the HW.(与物理硬件解耦)
  - Instruction translation.
  - Instruction hook.



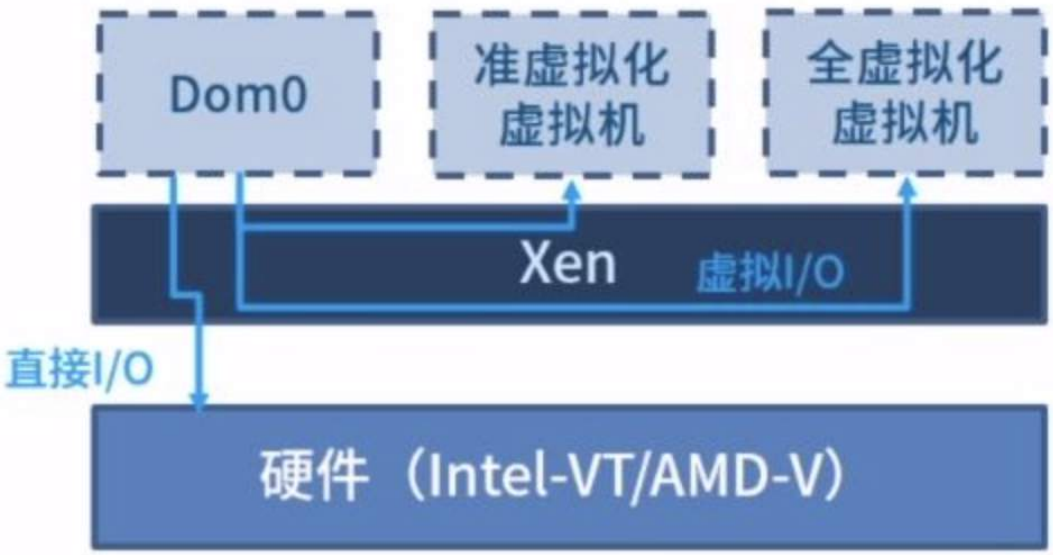
Background

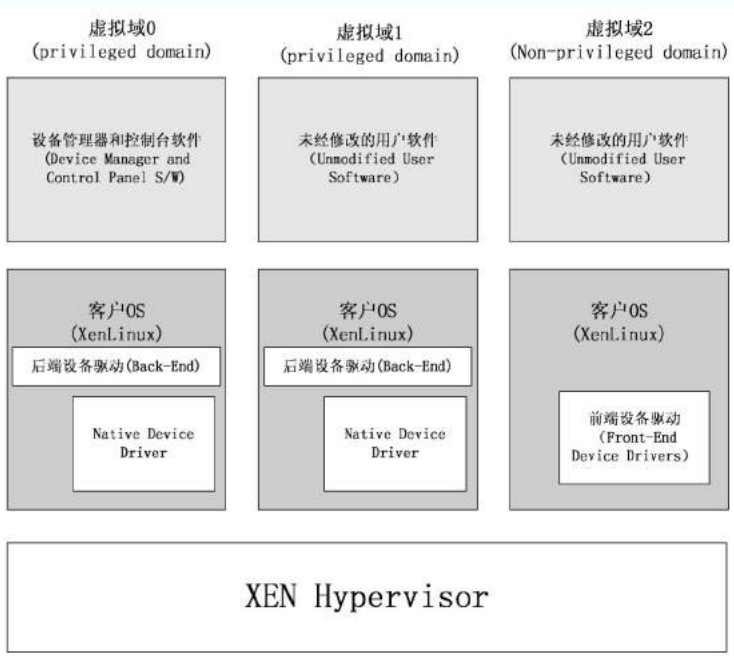
- KVM(Kernel-based Virtual Machine)
  - Open source. VM Module.
  - Full-Virtualization



Background

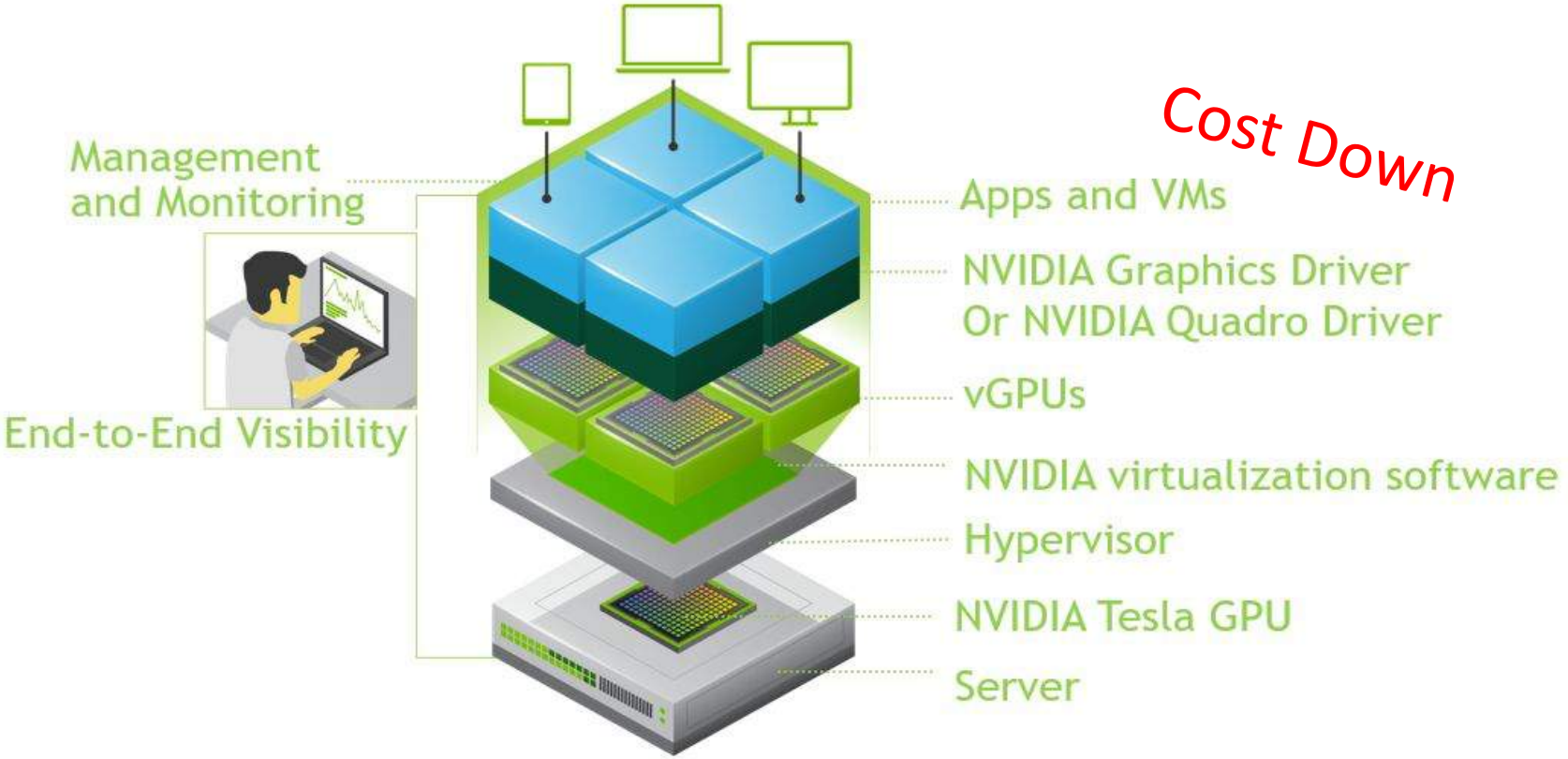
- XEN
  - Virtual Machine(VM) monitor.
  - Para-Virtualization
  - Operation System(OS) must explicitly modify.





Motivation

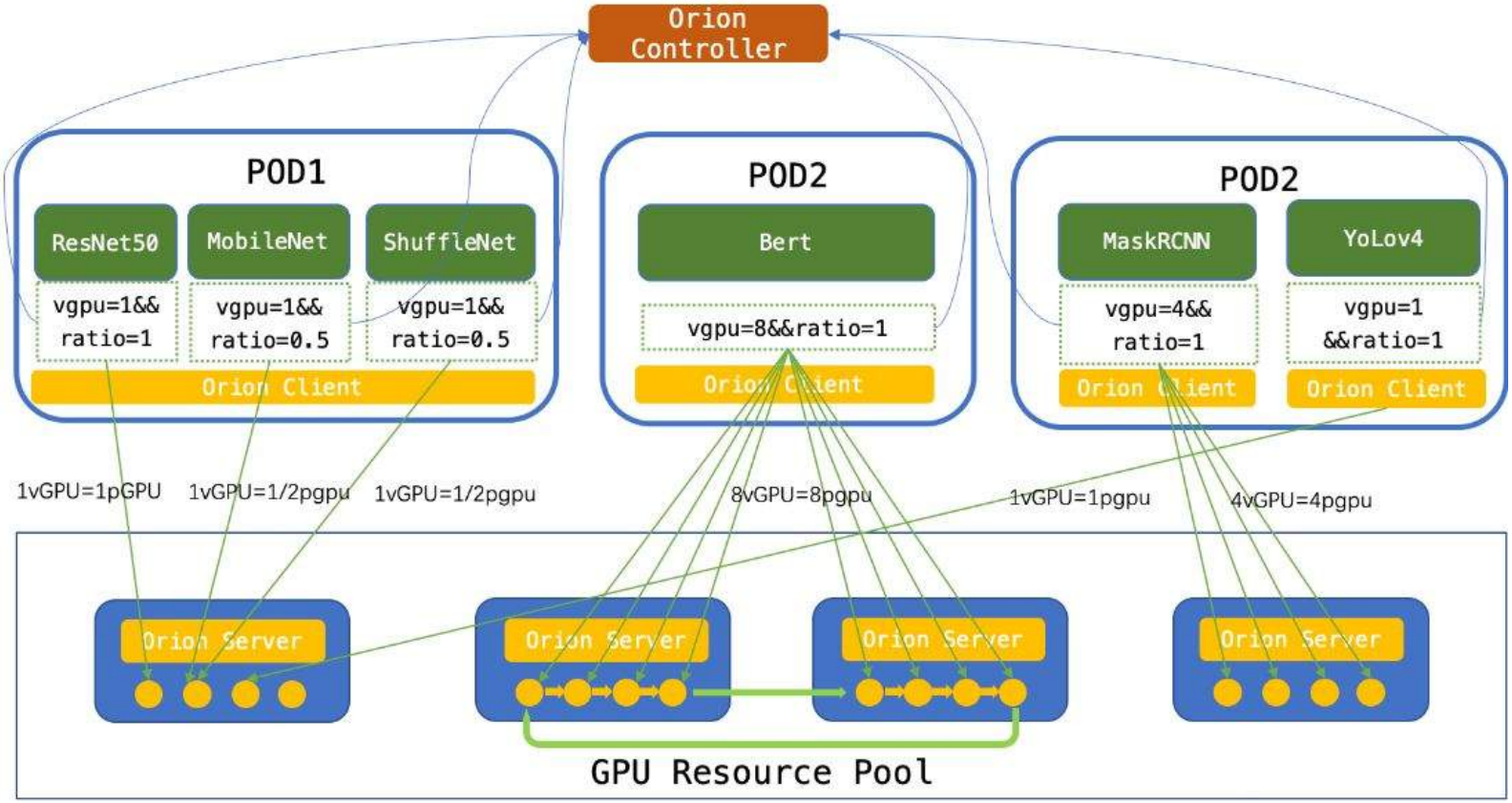
- Why vGpu is needed？
  - Improve the GPU resource utilization(资源利用率高)
  - Improve the service performance(用户体验提升)



More info: https://www.nvidia.cn/data-center/virtual-solutions/

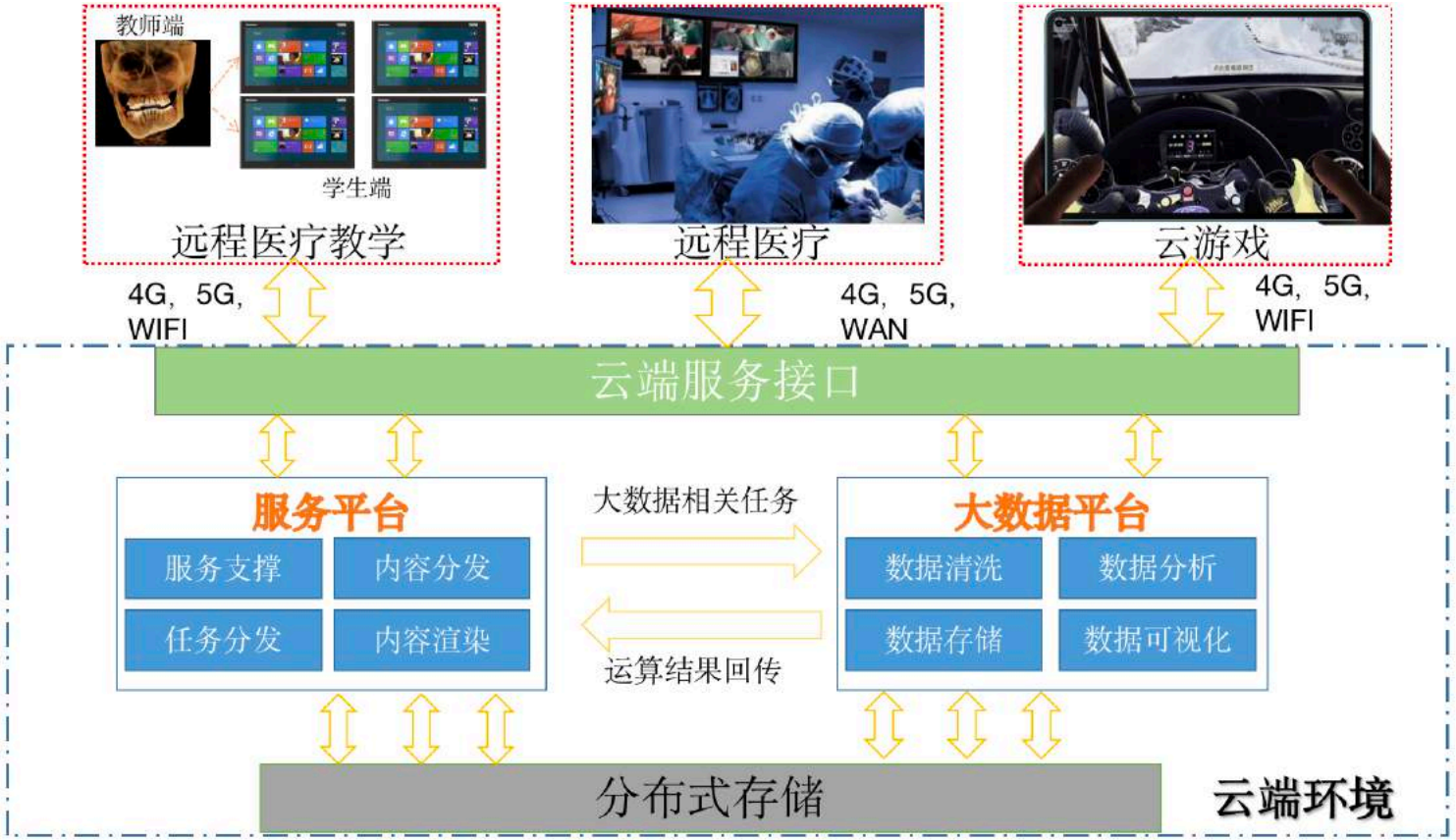
Motivation

- Resource pooling is an useful way



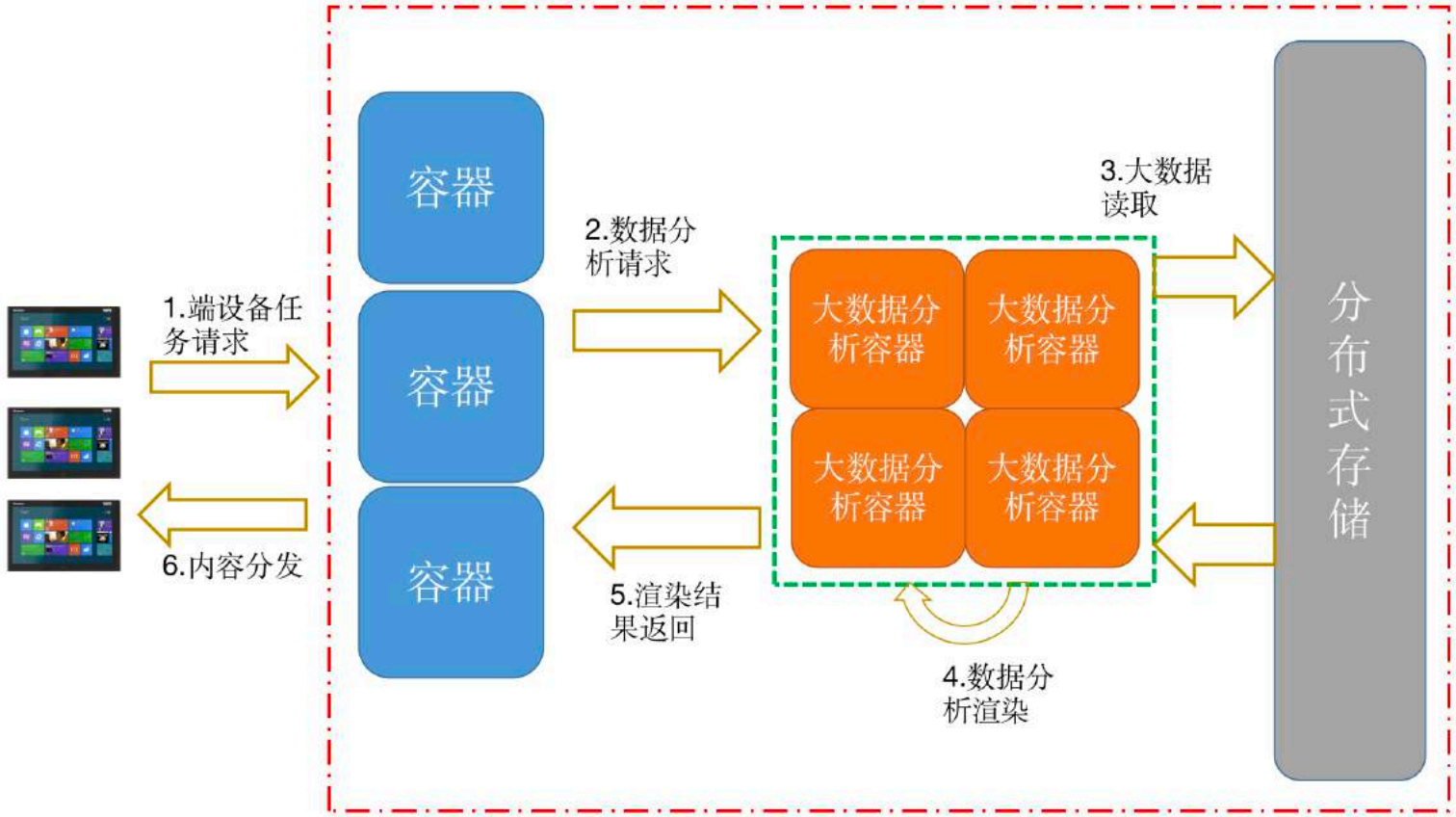
Motivation

- More services need GPU for improving user performance.
- But how to deal with GPU request?



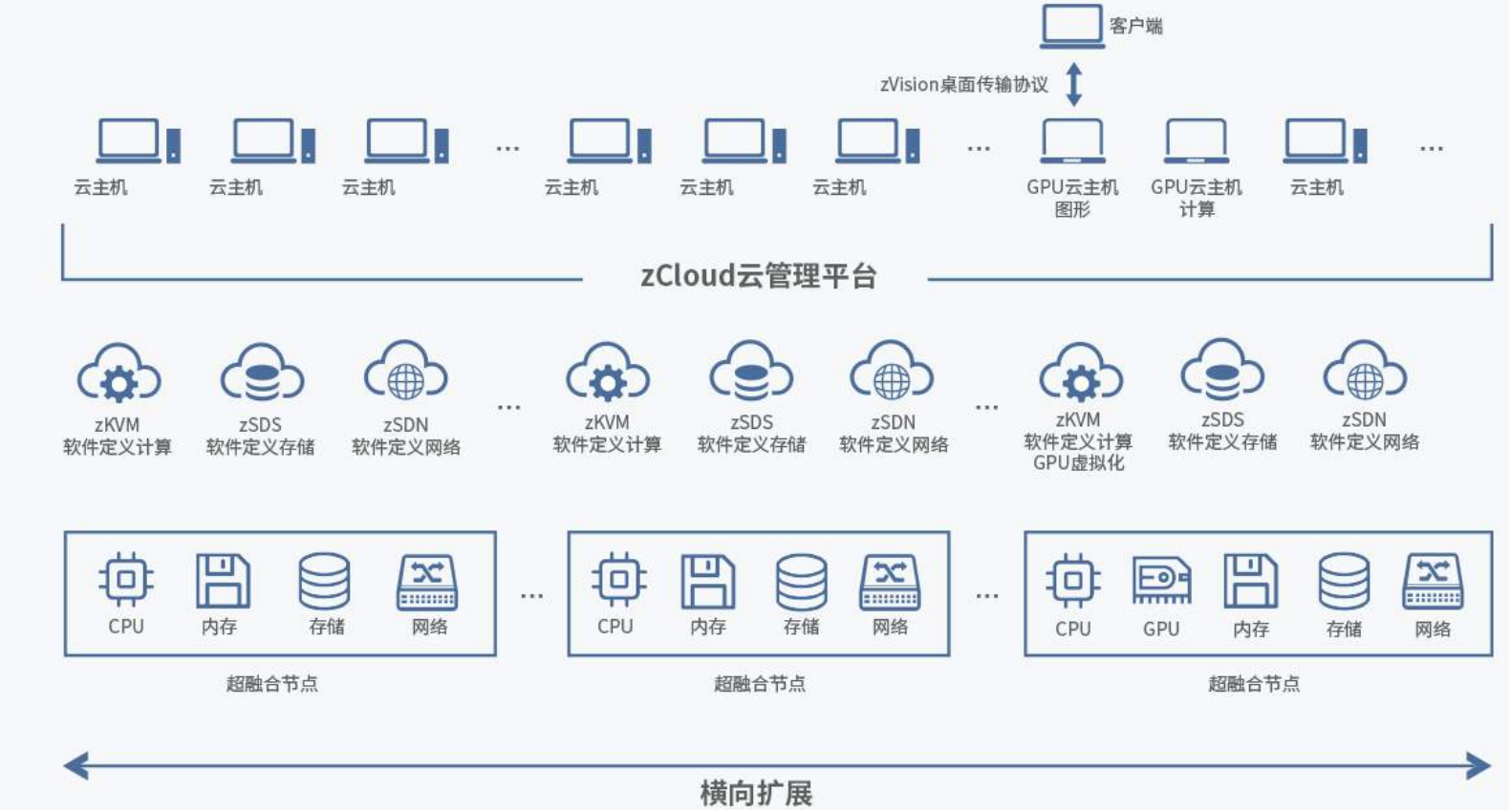
Motivation

- Some solution of GPU
- Solution1: C-S model



Motivation

- Solution2: Virtualization(虚拟化)



More info: http://www.zettakit.com/index/product/index/s/zCloud.html

Motivation

- Solution 3: GPU Passthrough(透传)

More info: http://www.zettakit.com/index/product/index/s/zVision.html

Motivation

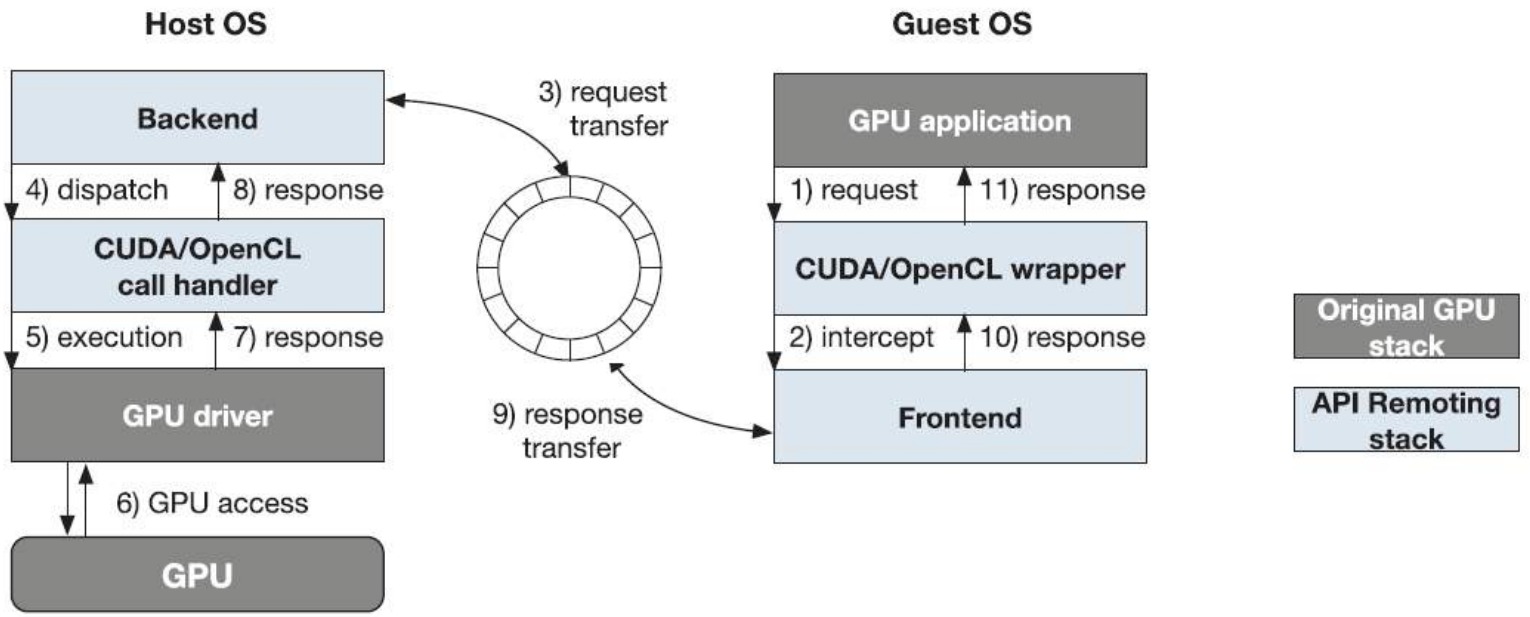
- Advantages
  - GPU resource utilization(资源利用率)
  - Service performance(用户体验)
  - Resource multiplexing(资源复用)
  - Green(节能)
- Disadvantages
  - Additional management cost(额外的管理成本)
  - Additional development cost(额外的开发成本)
  - Additional Hardware cost (额外的硬件成本)

GPU Virtualization

- API Remoting(远程API)
  - rCUDA, vCUDA, gRemote
- Para & Full Virtualization(半&全虚拟化)
  - gVirt, GPUvm, VMCG
- Hardware-Based GPU Virtualization(基于硬件的虚拟化)
  - GRID, VT-d，

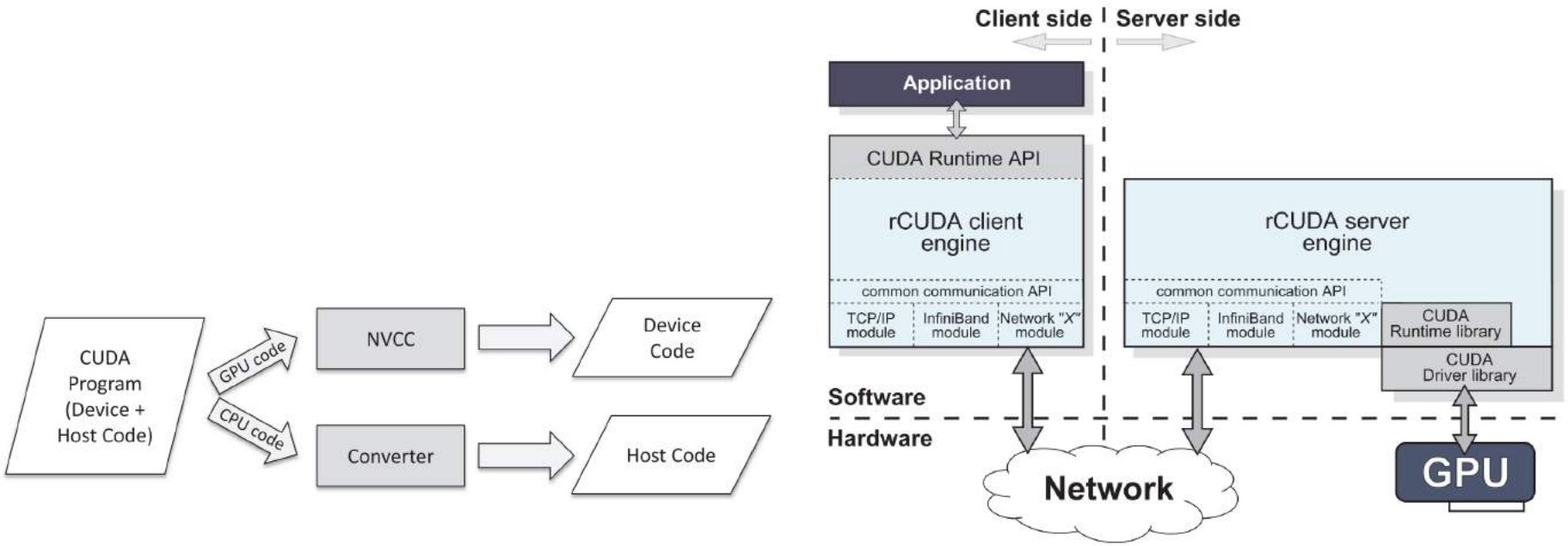
API-Remoting(远程API)

- API-Remoting is forwards GPUcalls in the guest to the host.
- Process:
  - 拦截-转发-执行-回复



API-Remoting(远程API)

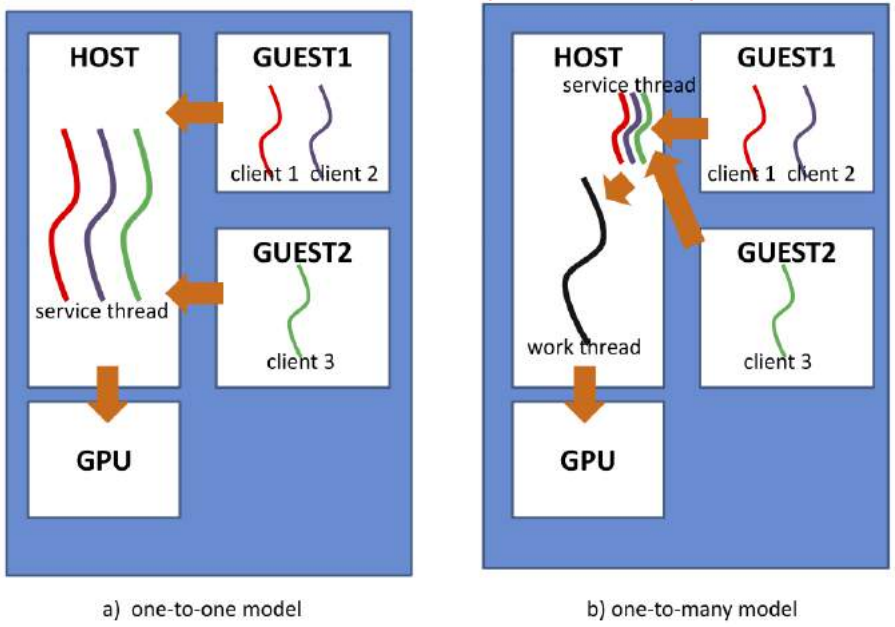
- rCUDA（2010-Now）
  - C-S Architecture
  - Standalone Compiler(独立编译器)
  - GPU call translate to GPU request.



API-Remoting(远程API)

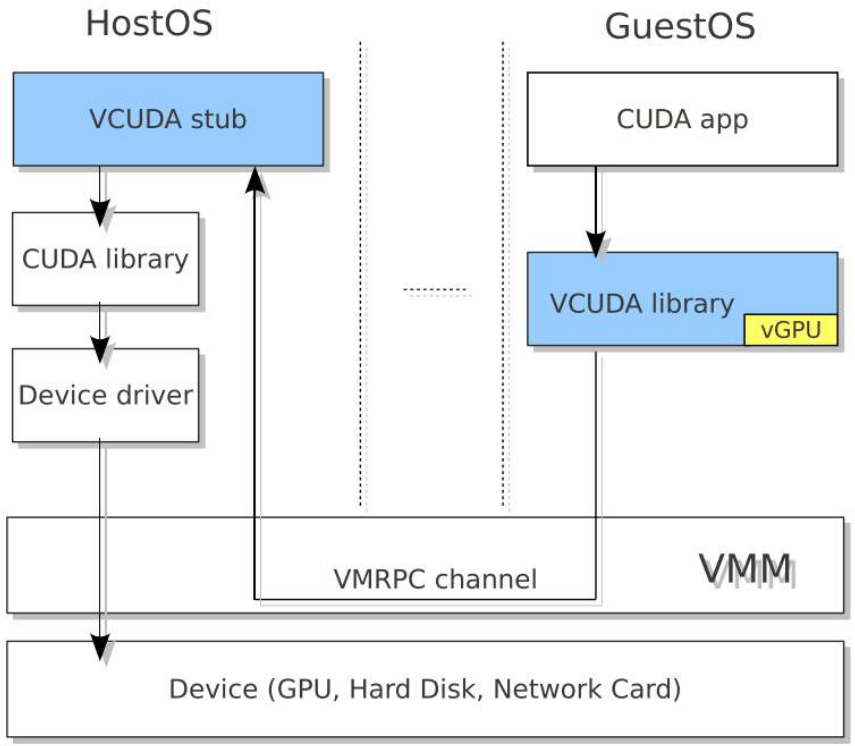
- vCUDA（TOC12）
  - Base on KVM, CUDA remote process call.
  - Build a multi-channel multiplexing and S & R

Batch Allocate(批量执行)



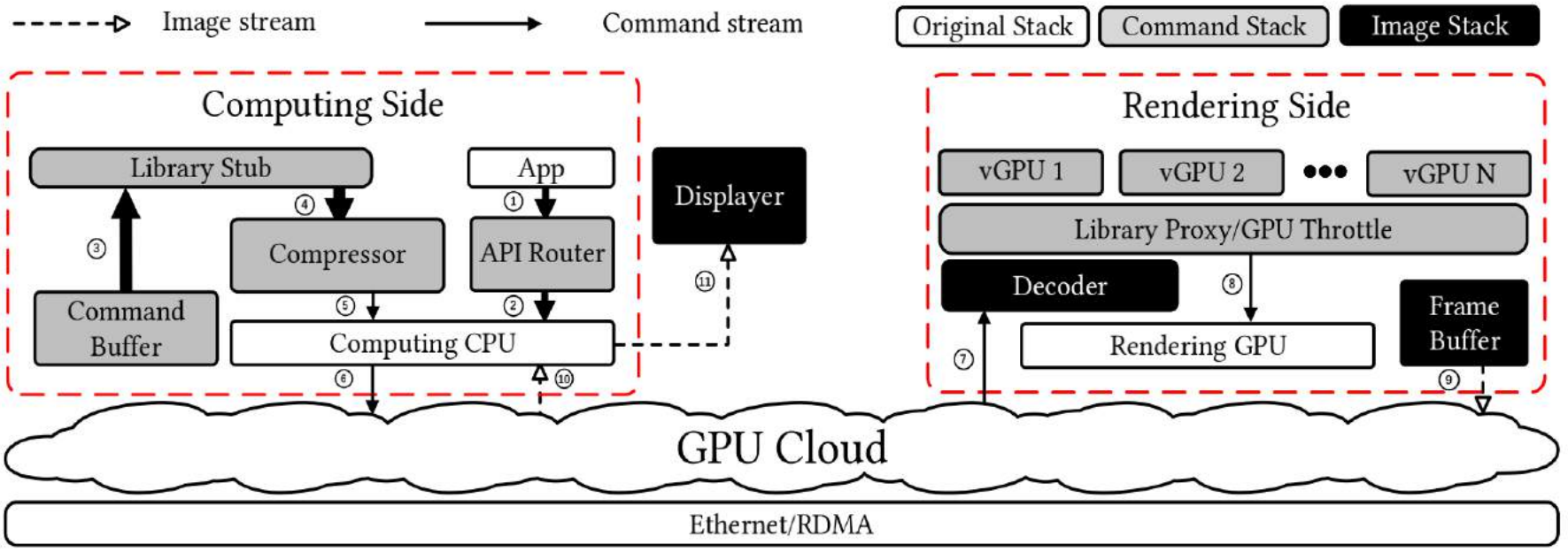
a) one-to-one model

b) one-to-many model



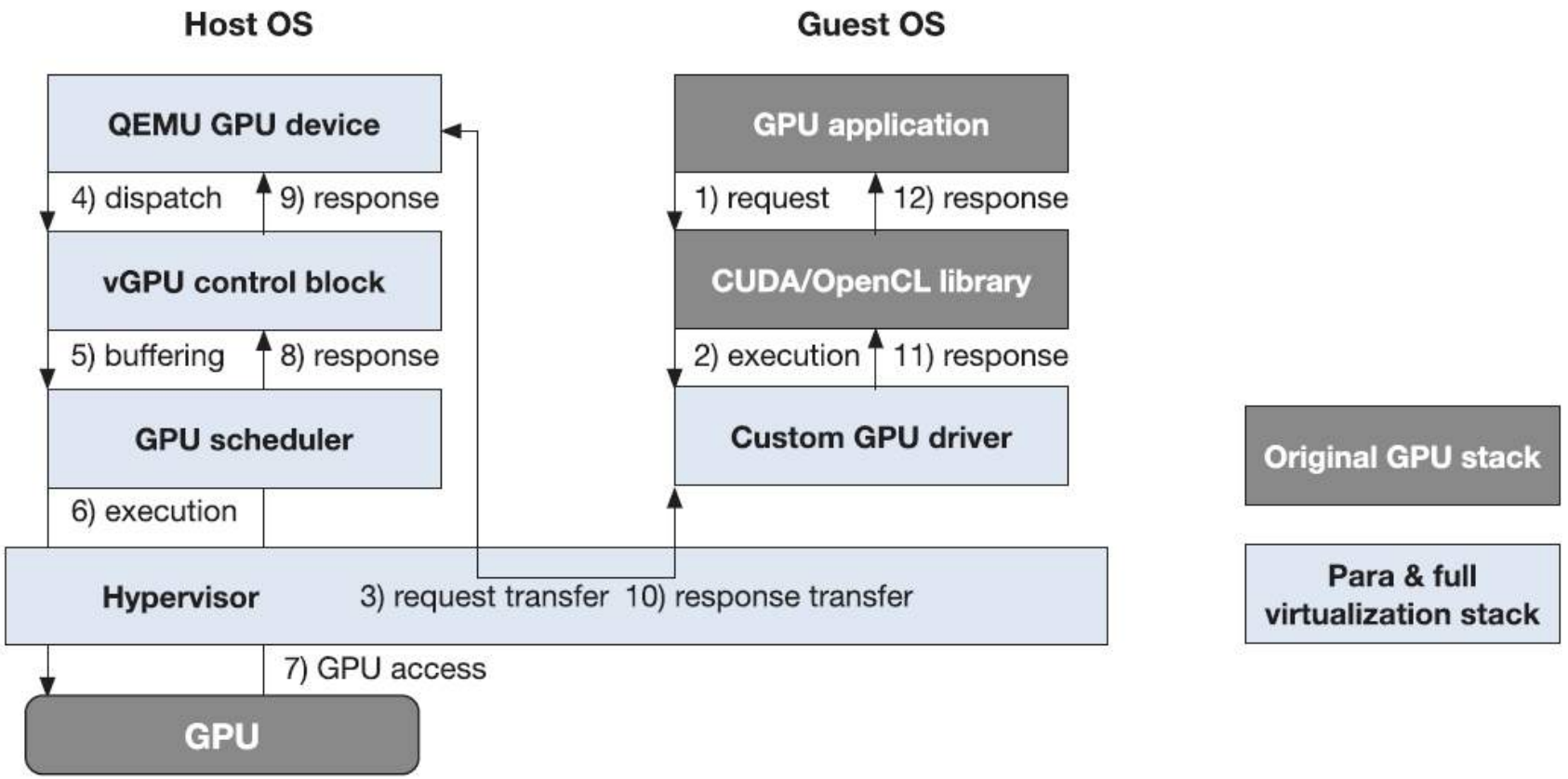
API-Remoting(远程API)

- gRemote(HPDC20)
  - Propose GPU throttling technology
  - Transfer GPU-related API only



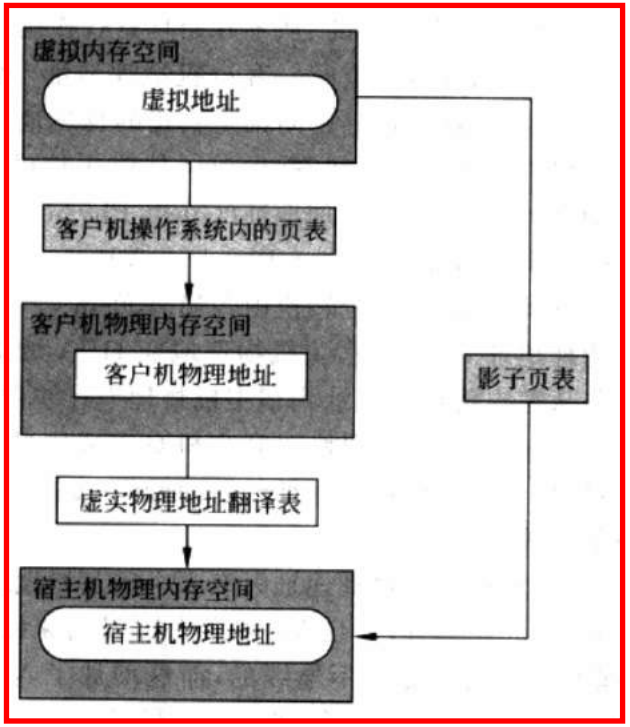
Para & Full Virtualization(半&全虚拟化)

- API remoting solutions cannot be used currently on new graphics hardware and the most recent GPU libraries

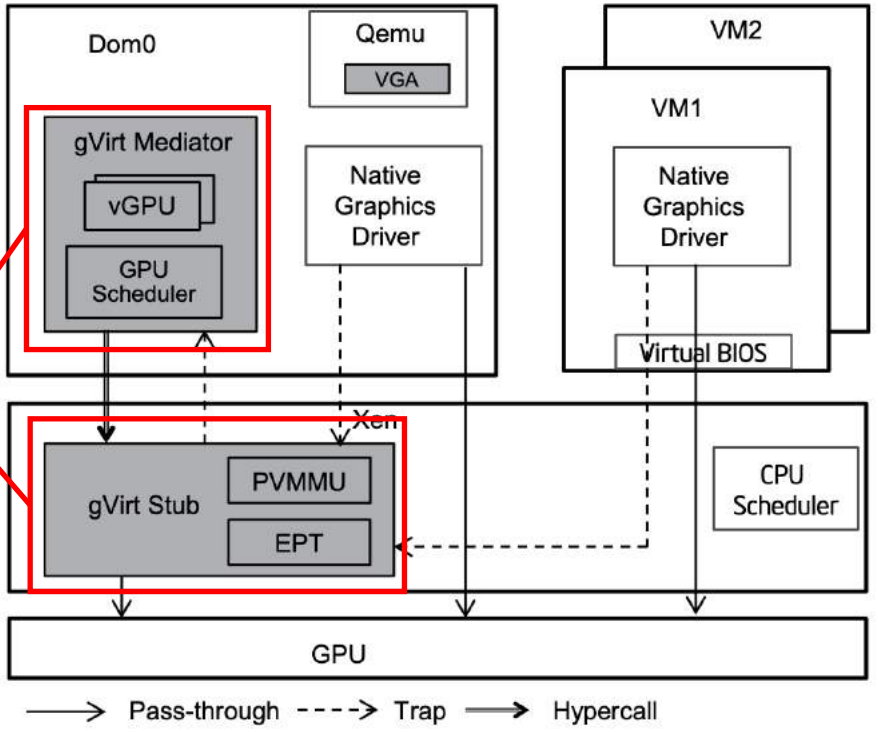


Para & Full Virtualization(半&全虚拟化)

- gVirt（ATC14）
  - Use the memory resource partition to ensure the performance of the VM
  - Use scheduler and shadow GTT to improve performance

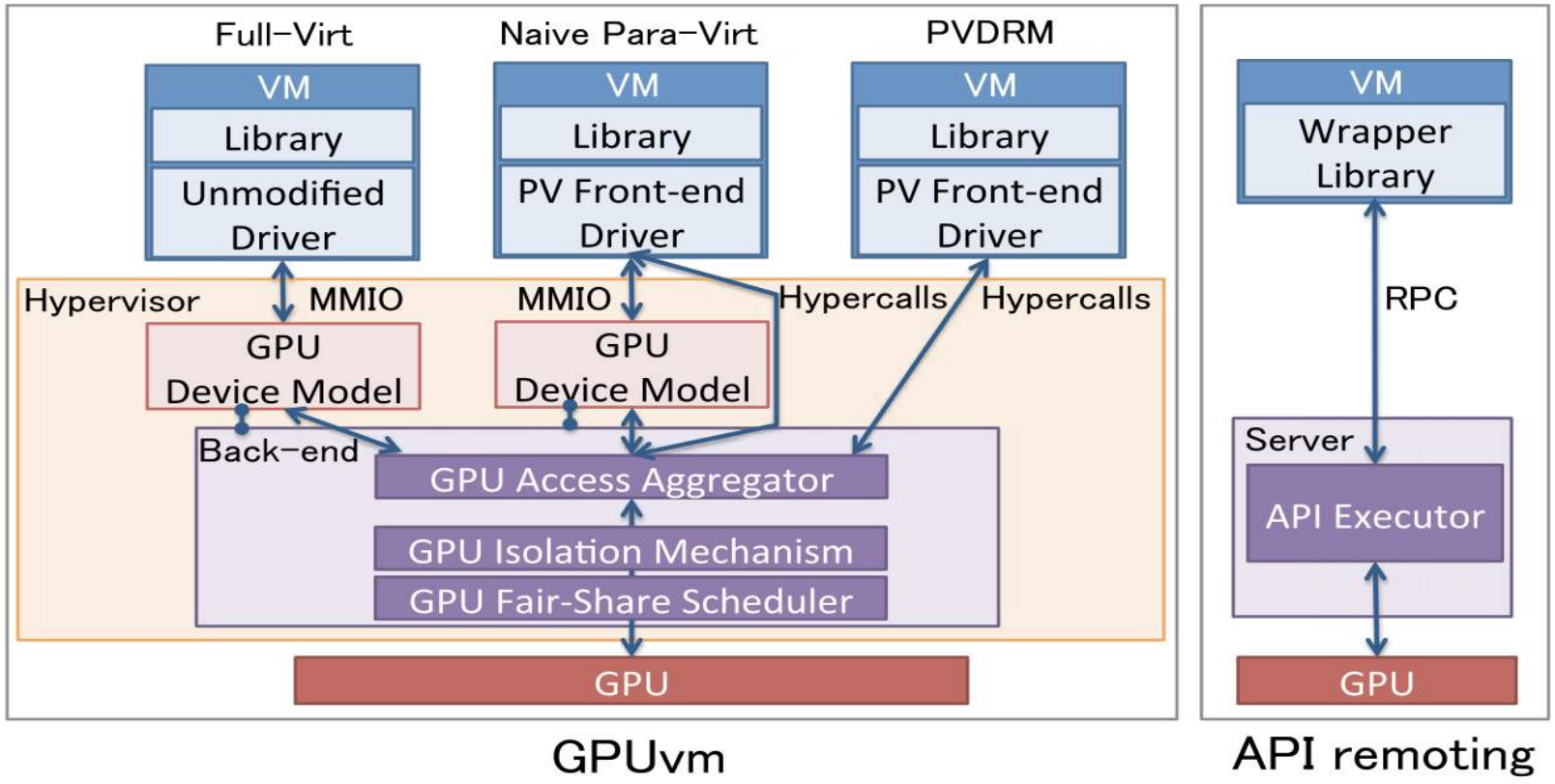


内存显存管理



Para & Full Virtualization(半&全虚拟化)

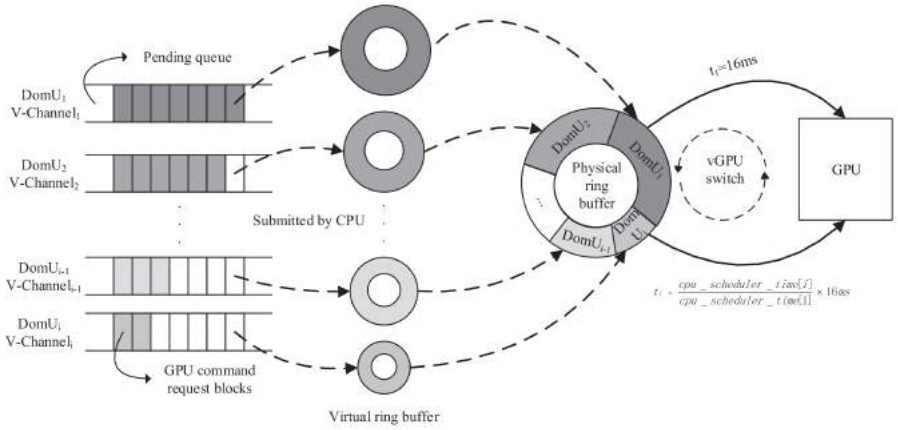
- GPUvm（ToC16）
  - Contains a variety of virtualization methods(多方式虚拟化)
  - By management of GPU Page to achieve a reuse of GPU

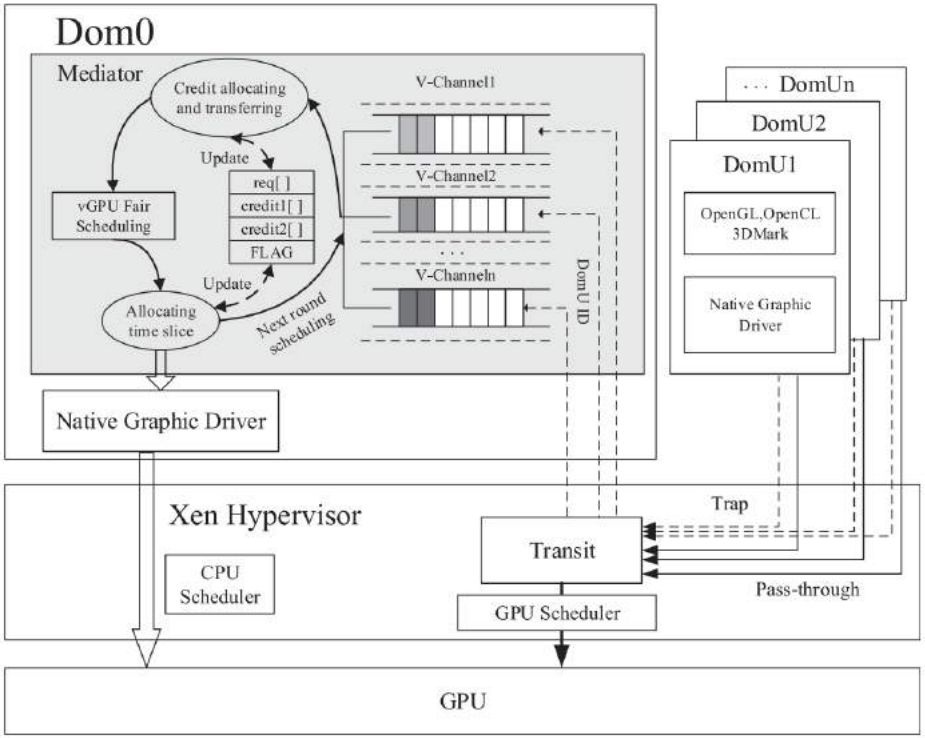


Para & Full Virtualization(半&全虚拟化)

- VMCG（TPDS19）
  - Store GPU requests with separate queues
  - Using fair VGPU scheduling algorithm

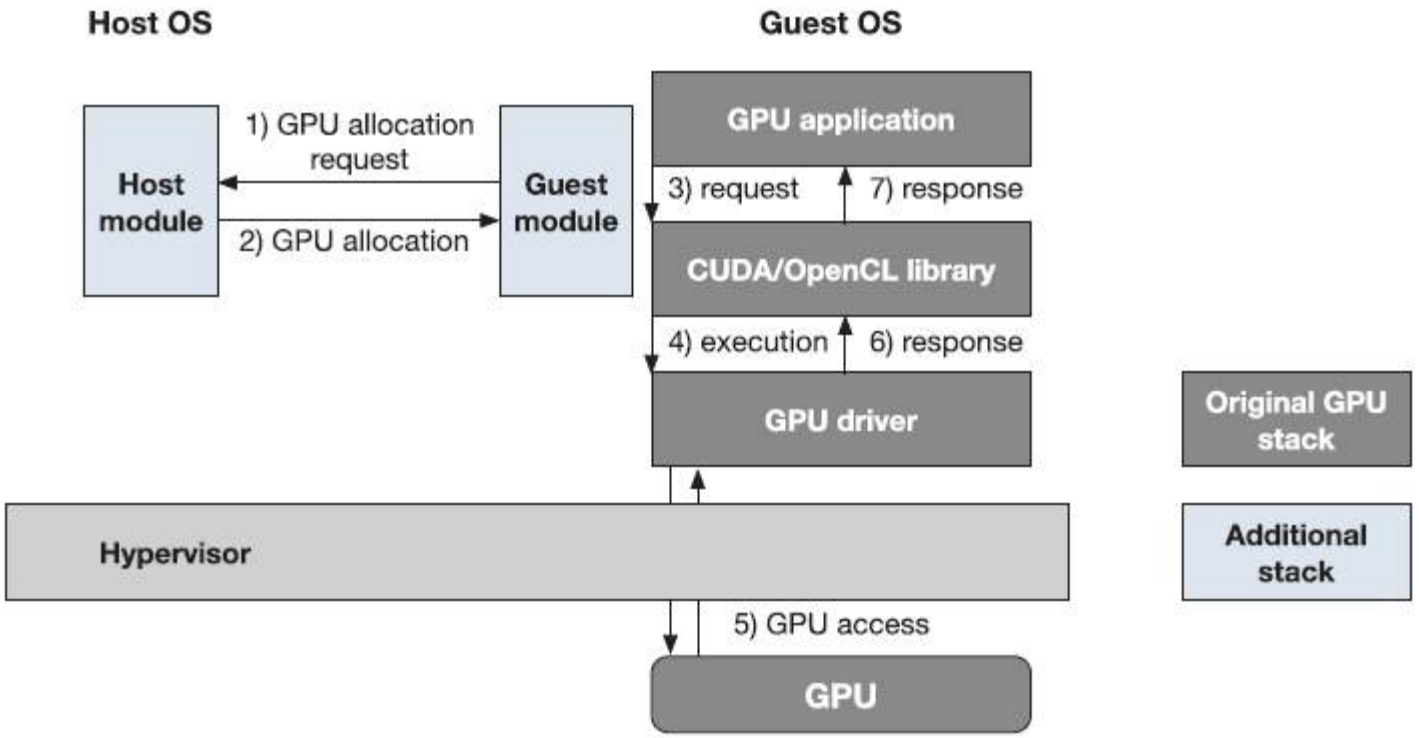
虚拟化的Channel来存储请求





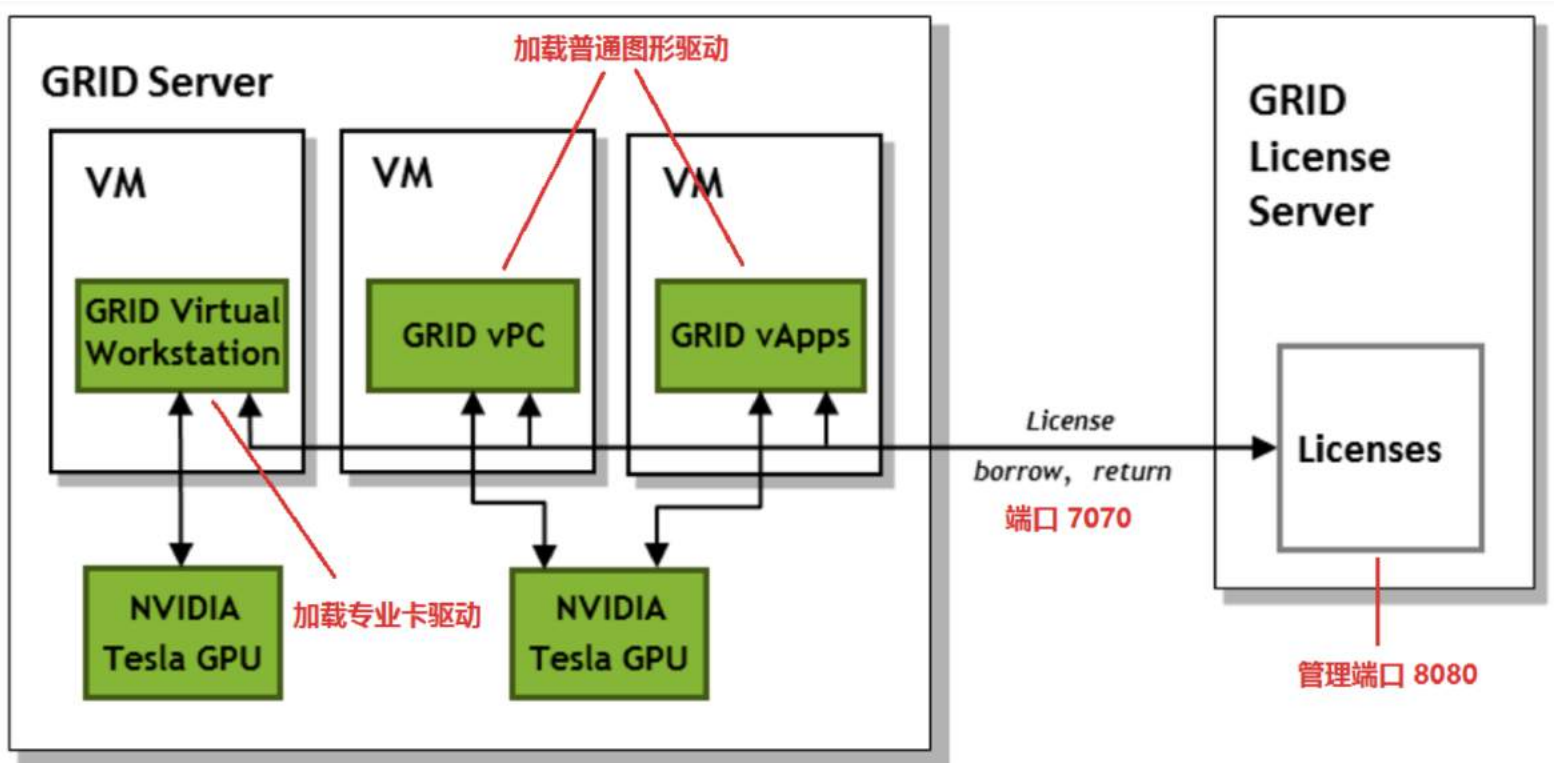
Hardware-Based GPU Virtualization

- Have a Hardware support (需要硬件支持)



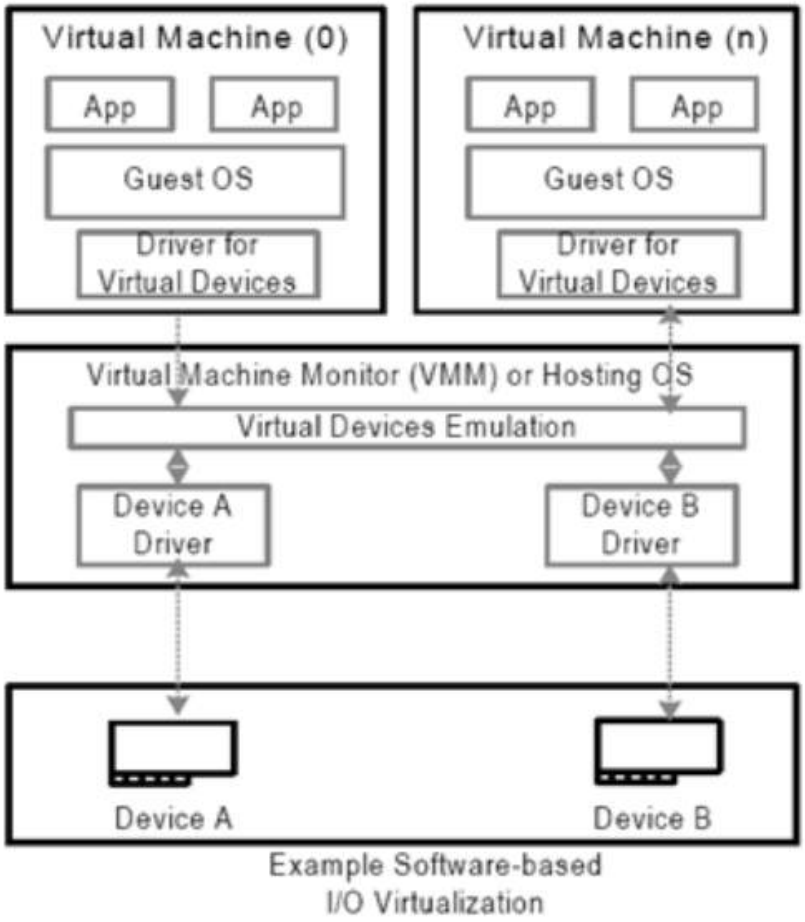
Hardware-Based Gpgpu Virtualization

- NVIDIA GRID

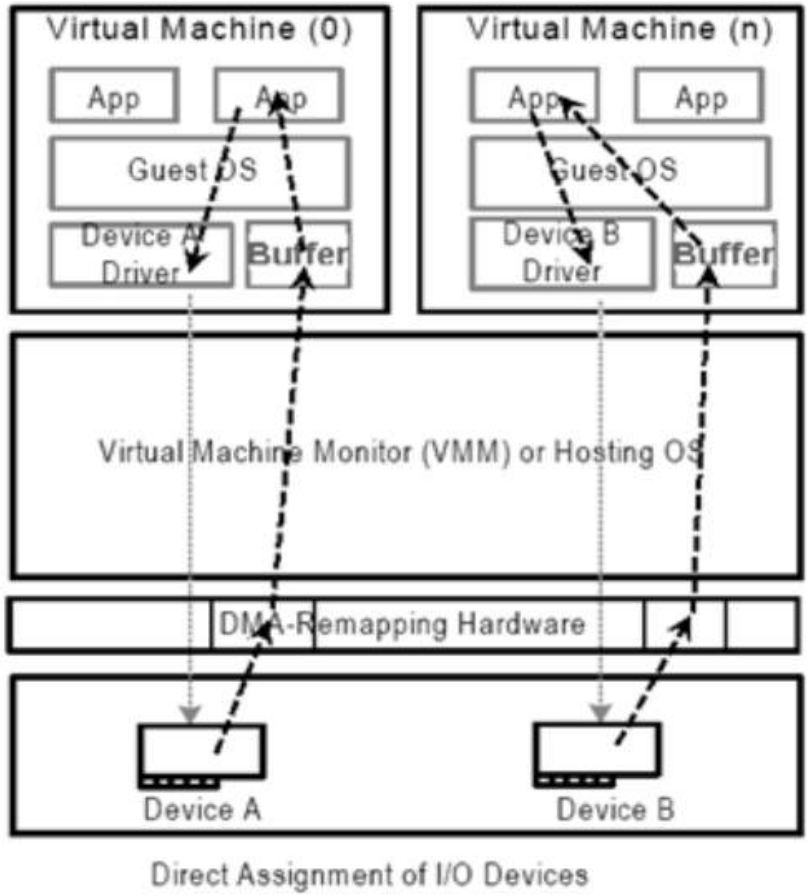


Hardware-Based GPU Virtualization

- Intel VT-d



Example Software-based I/O Virtualization



Direct Assignment of I/O Devices

Conclusion

- GPU 虚拟化允许在多个异构虚拟机之间共享一个物理 GPU，以节省成本，确保 GPU 设备的最佳使用，并为其客户提供高性能平台。
- 异构计算作为一种新的范式，需要更多的关注CPU-GPU协同
- 为了实现多租户的GPU虚拟化技术，其基础是一个多目标优化的调度技术。

Ref:

- José Duato, Antonio J. Peña, Federico Silla, Rafael Mayo, Enrique S. Quintana-Ortí. rCUDA: Reducing the number of GPU-based accelerators in high performance clusters. HPCS 2010: 224-231
- Carlos Reaño, Federico Silla, Adrián Castelló, Antonio J. Peña, Rafael Mayo, Enrique S. Quintana-Ortí, José Duato. Improving the user experience of the rCUDA remote GPU virtualization framework. Concurr. Comput. Pract. Exp. 27(14): 3746-3770 (2015)
- Giulio Giunta, Raffaele Montella, Giuseppe Agrillo, Giuseppe Coviello. A GPGPU Transparent Virtualization Component for High Performance Computing Clouds. Euro-Par (1) 2010: 379-391
- Vishakha Gupta, Karsten Schwan, Niraj Tolia, Vanish Talwar, Parthasarathy Ranganathan.Pegasus:Coordinated Scheduling for Virtualized Accelerator-based Systems. USENIX Annual Technical Conference 2011
- Lin Shi, Hao Chen, Jianhua Sun, Kenli Li,vCUDA. GPU-Accelerated High-Performance Computing in Virtual Machines. IEEE Trans. Computers 61(6): 804-816 (2012)
- Chao Yu, Yuebin Bai, Hailong Yang, Kun Cheng, Yuhao Gu, Zhongzhi Luan, Depei Qian. SMGuard: A Flexible and Fine-Grained Resource Management Framework for GPUs. IEEE Trans. Parallel Distributed Syst. 29(12): 2849-2862 (2018)
- Hangchen Yu, Arthur Michener Peters, Amogh Akshintala, Christopher J. Rossbach. AvA: Accelerated Virtualization of Accelerators. ASPLOS 2020: 807-825
- Dongjie Tang, Yun Wang, Linsheng Li, Jiacheng Ma, Xue Liu, Zhengwei Qi, Haibing Guan. gRemote: API-Forwarding Powered Cloud Rendering. HPDC 2020: 197-201

Ref:

- Mochi Xue, Kun Tian, Yaozu Dong, Jiacheng Ma, Jiajun Wang, Zhengwei Qi, Bingsheng He, Haibing Guan. gScale: Scaling up GPU Virtualization with Dynamic Sharing of Graphics Memory Space. USENIX Annual Technical Conference 2016: 579-590
- Cheol-Ho Hong, Ivor T. A. Spence, Dimitrios S. Nikolopoulos. FairGV: Fair and Fast GPU Virtualization. IEEE Trans. Parallel Distributed Syst. 28(12): 3472-3485 (2017)
- Huailiang Tan, Yanjie Tan, Xiaofei He, Kenli Li, Keqin Li. A Virtual Multi-Channel GPU Fair Scheduling Method for Virtual Machines. IEEE Trans. Parallel Distributed Syst. 30(2): 257-270 (2019)
- Qiumin Lu, Jianguo Yao, Haibing Guan, Ping Gao:gQoS: A QoS-Oriented GPU Virtualization with Adaptive Capacity Sharing. IEEE Trans. Parallel Distributed Syst. 31(4): 843-855 (2020)

Thank you for your listening