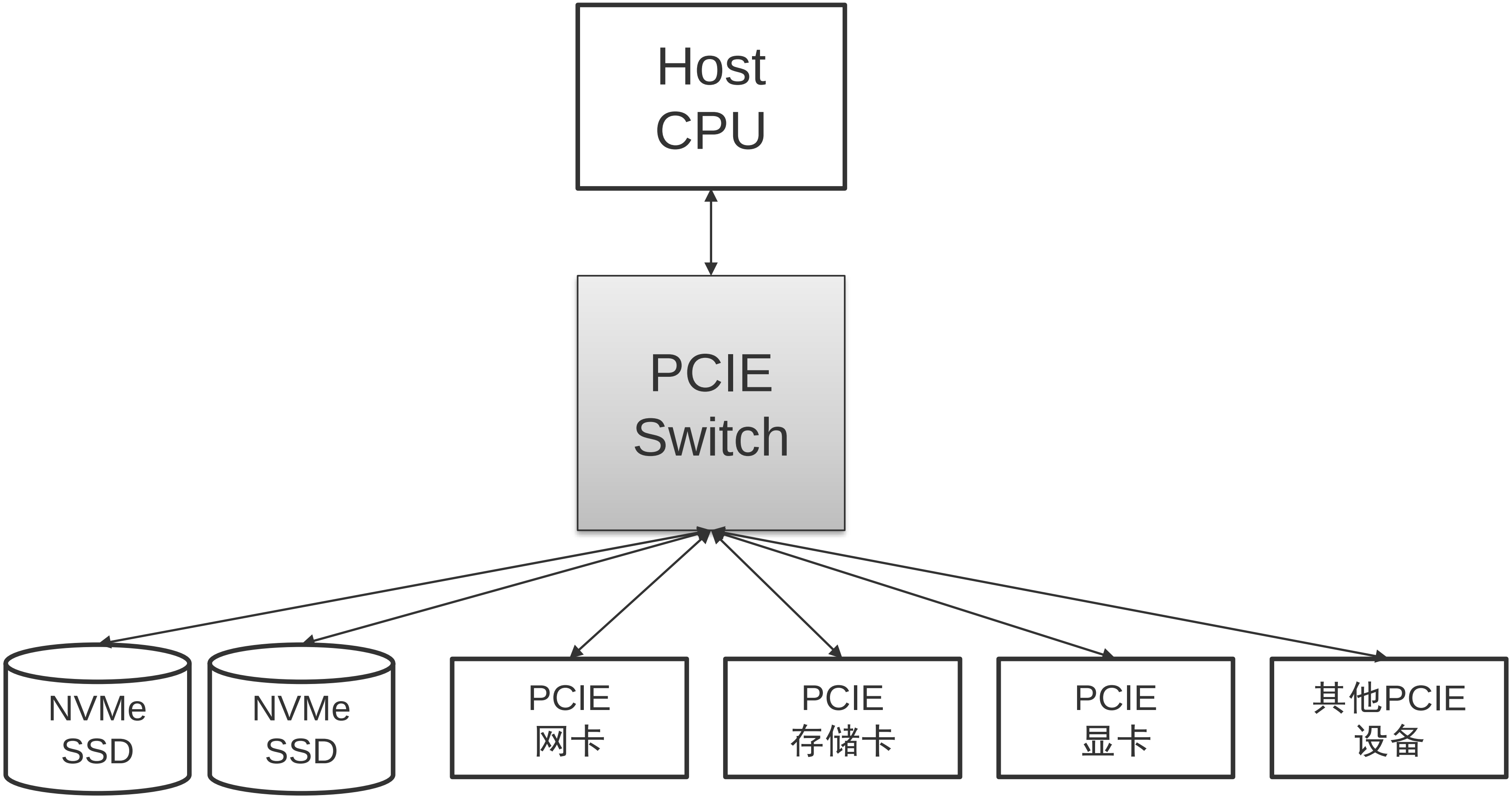
Power Matters.TM

利用PCIE Switch+RDMA技术加速数据中心IO流量

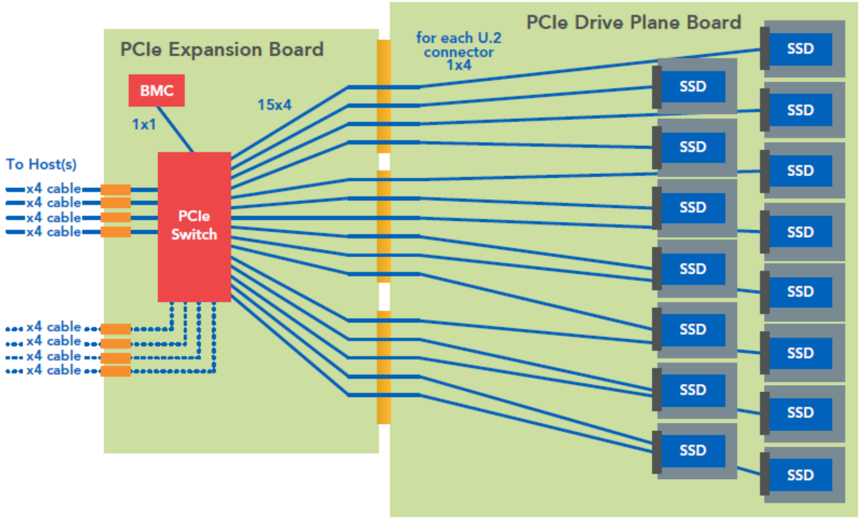
张冬 高级资深数据中心架构师
dong.zhang@Microsemi.com

PCIE Switch能做什么——FANOUT

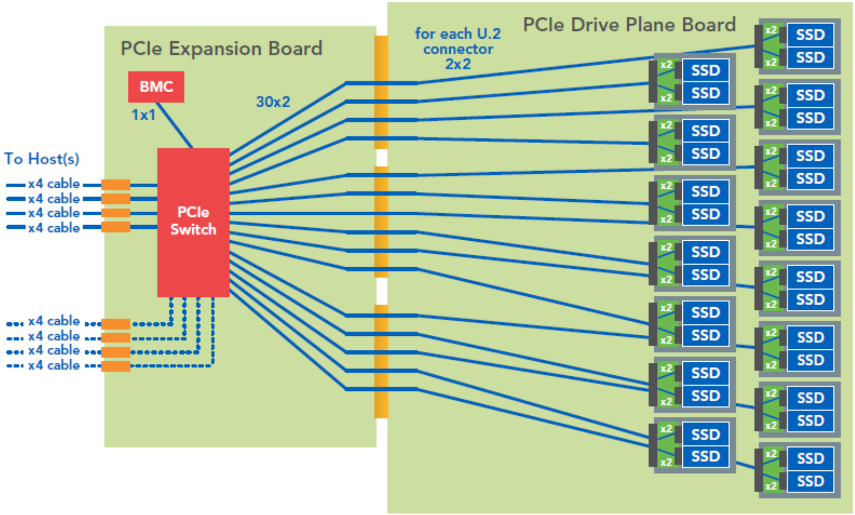


单芯片最大支持48个Port——Microsemi特有

PCIe switch configurations



15 x4 SSDs

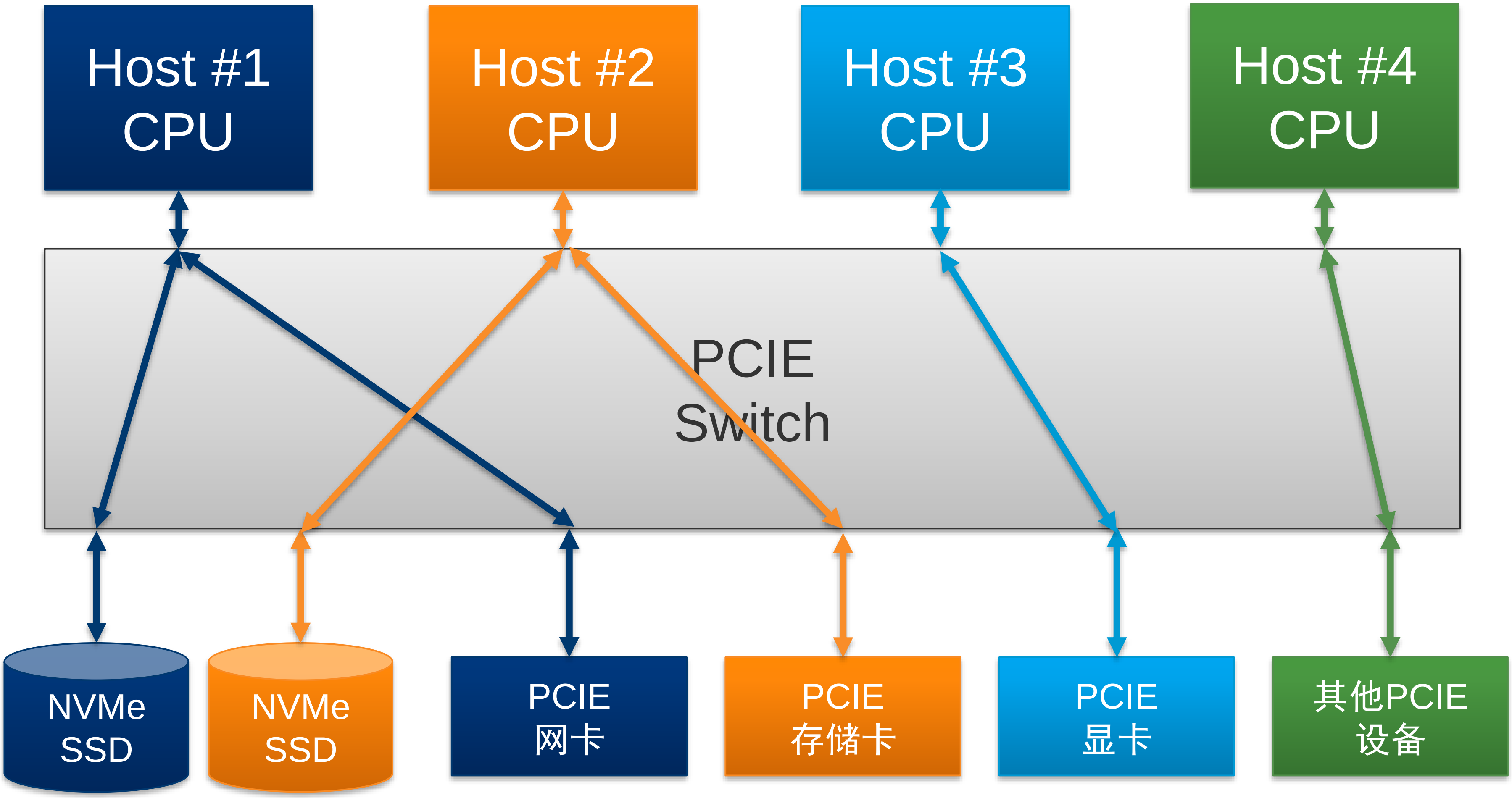


30 x2 SSDs

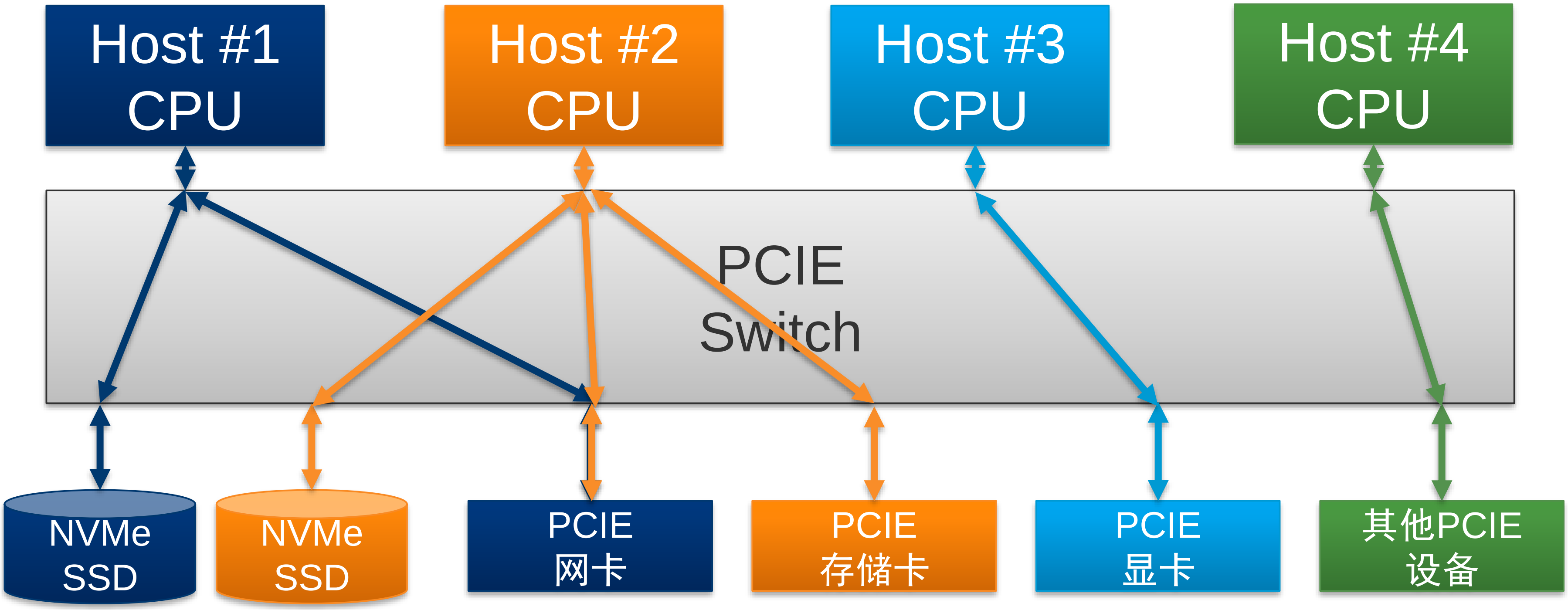
Open Compute (Facebook) Reference Design

http://www.opencompute.org/wiki/Storage *(search for Lightning)*

PCIE Switch能做什么——分区

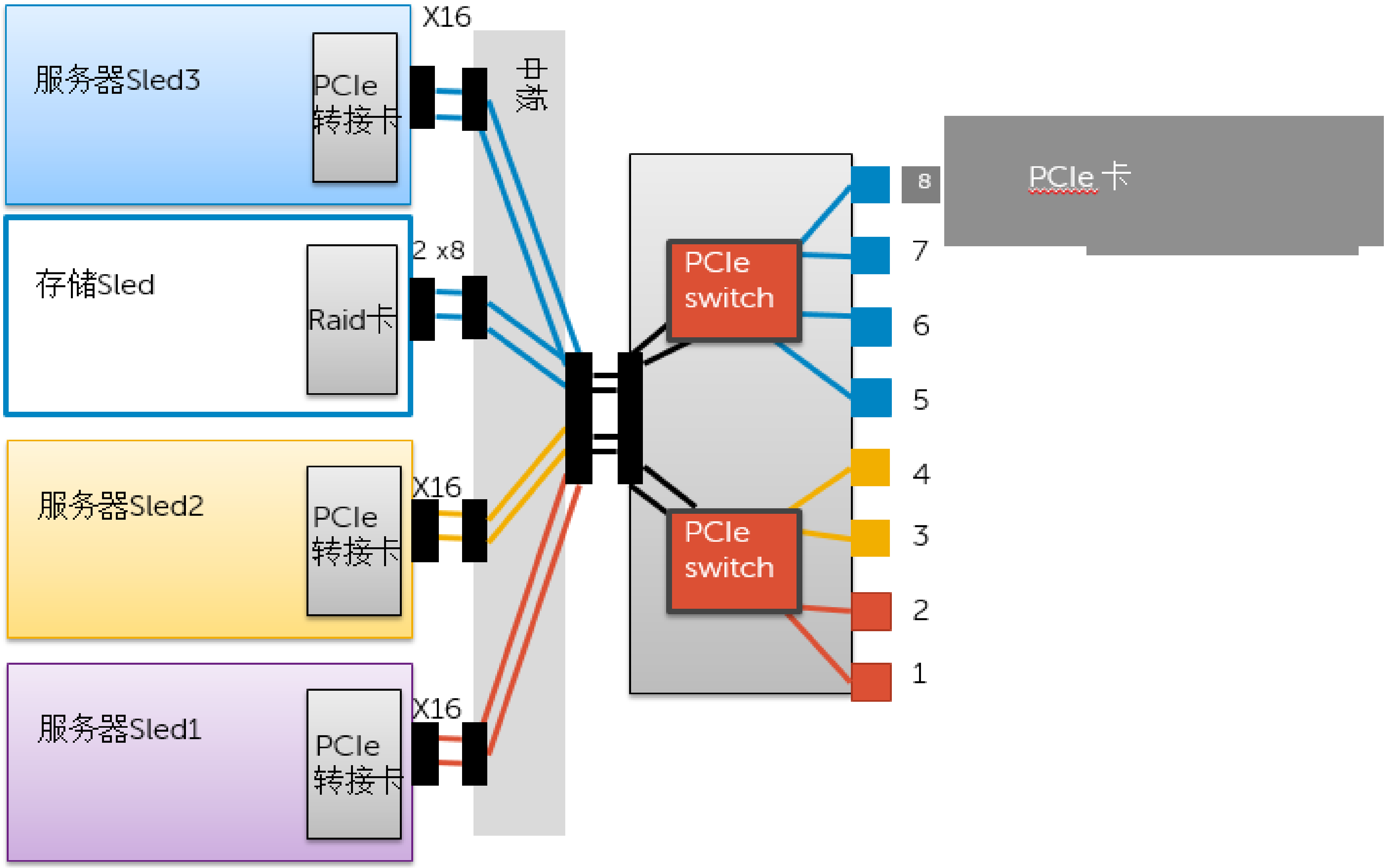


动态分区——Microsemi PCIE Switch特有

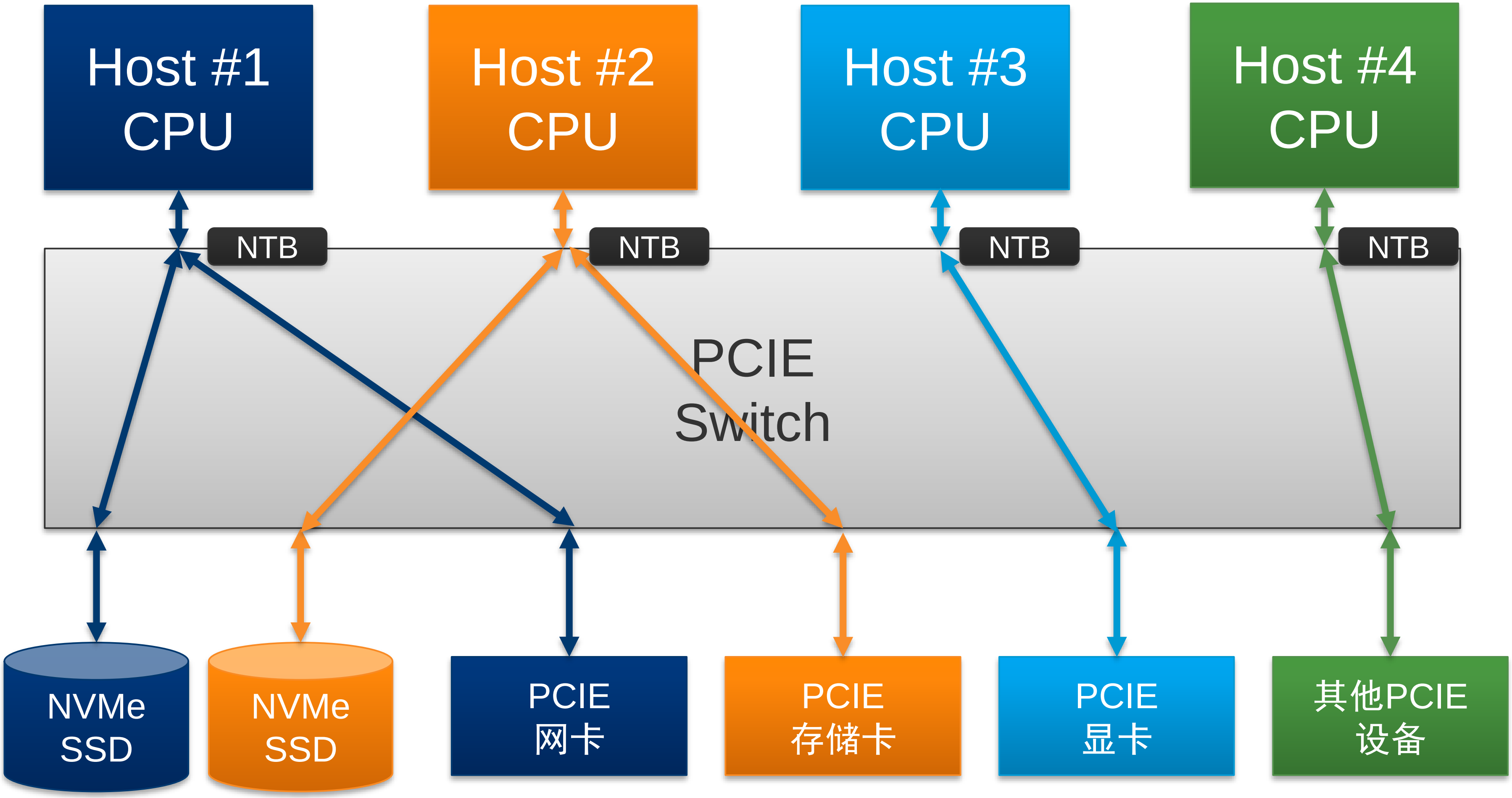


分区的重新配置不需要主机端重启，不中断主机端原有的IO访问。新添加设备动态发现并理解可用。

基于PCIE Switch分区的服务器设计

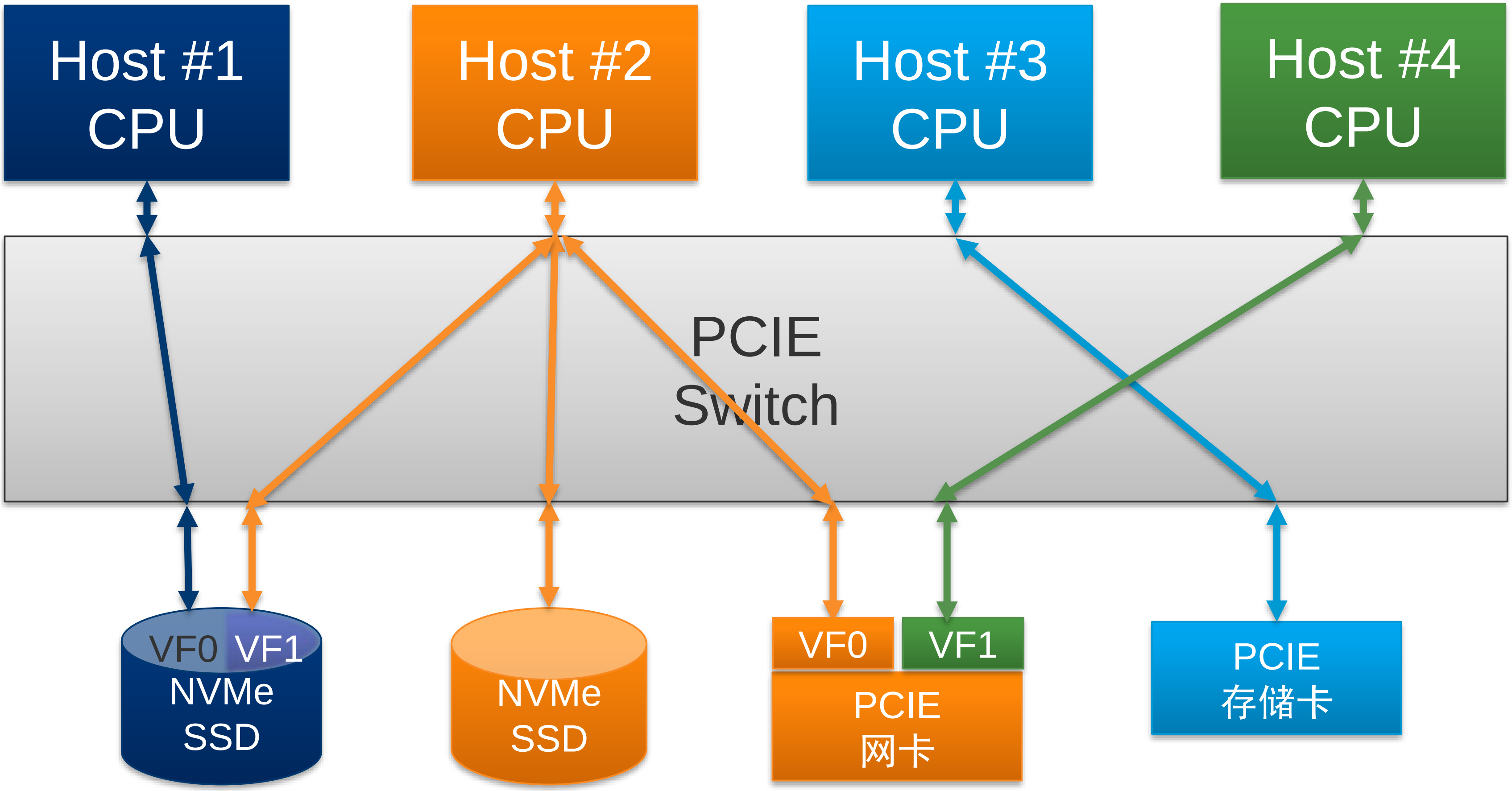


PCIE Switch能做什么——多机互联



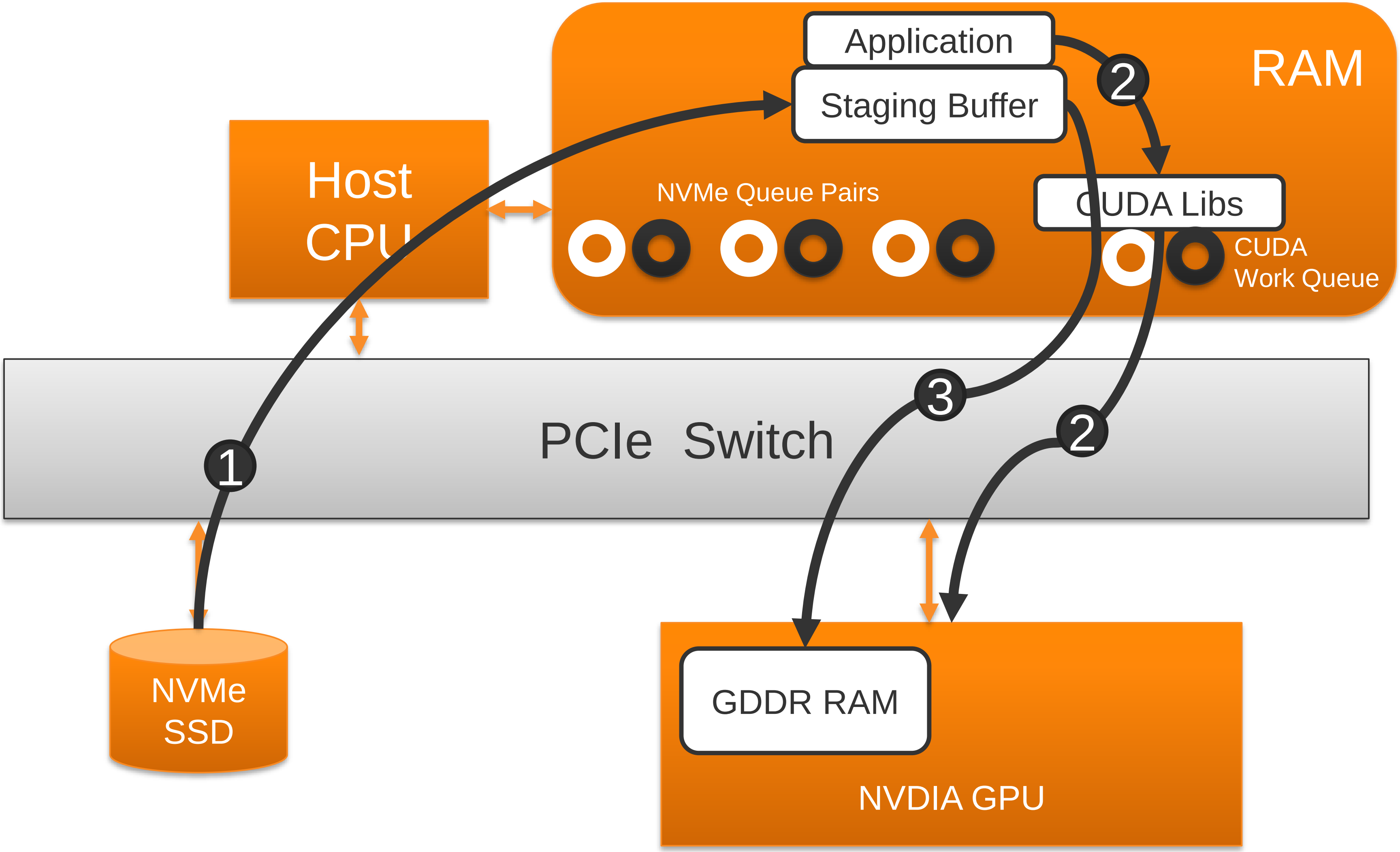
Microsemi PCIE Switch最大支持48台主机互联

PCIE Switch能做什么——共享IO

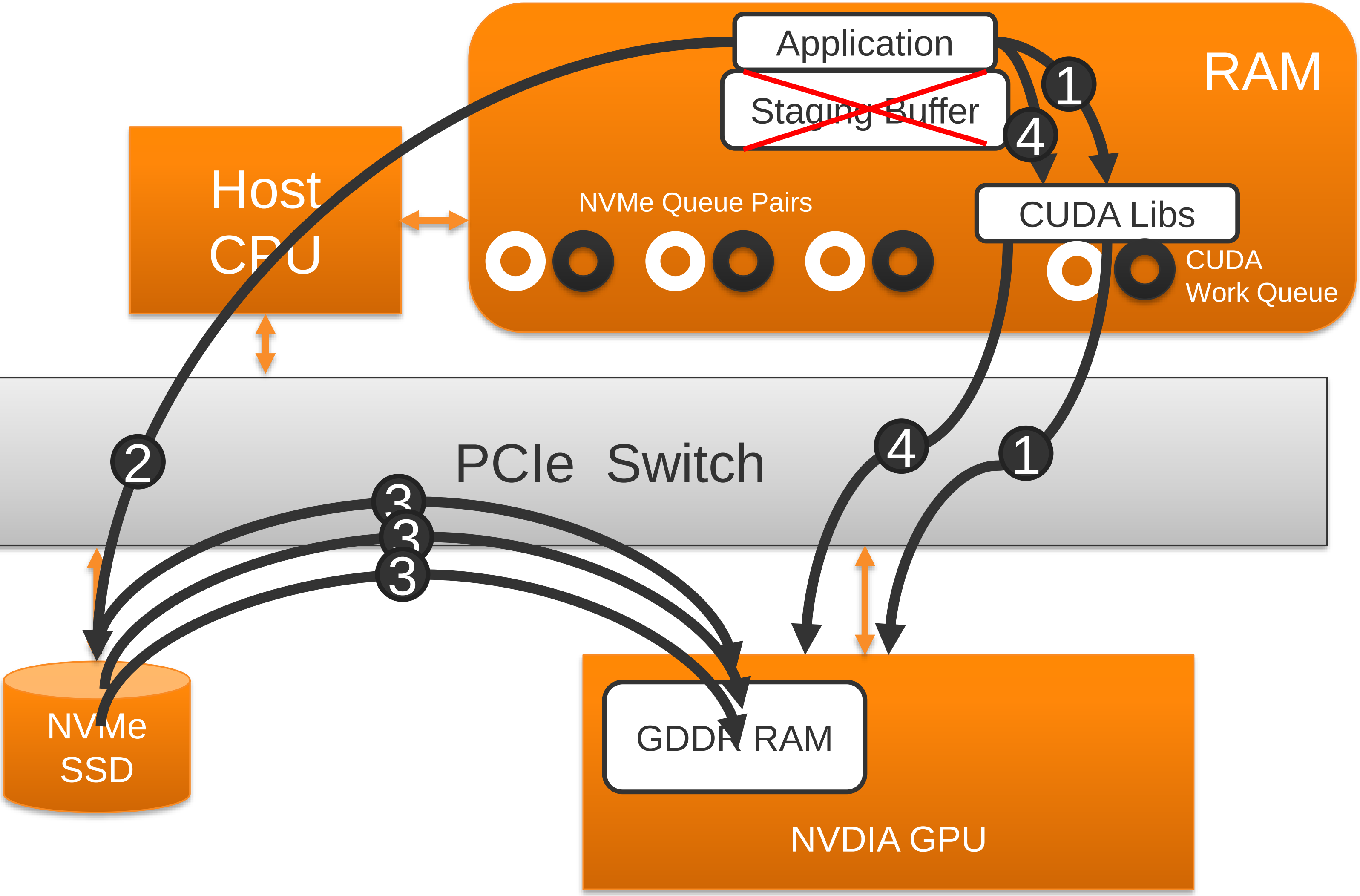


将支持SRIOV的PCIE设备透明转换为MRIOV模式

PCIE Switch能做什么——P2P传输



PCIE Switch能做什么——P2P传输



利用P2P GPU Direct加速图像搜索过程

数据读取速度	带宽(GB/s)	主机RAM相对使用率
传统做法	1.90	5230
P2P	**2.50**	**1**

图像处理速度	HDD	SDD	
	兆像素/秒	兆像素/秒	瓶颈点
CPU	77.0	122.8	CPU
CUDA	95.1	312.5	DRAM
P2P	**N/A**	**534.2**	**GPU**

实现P2P传输的条件

- 支持将MMIO BAR注册为DAX块设备的Linux内核。Microsemi开发了参考代码，对Linux内核有88行的变更。从而可支持将 DAX设备地址空间作为DMA的源端。

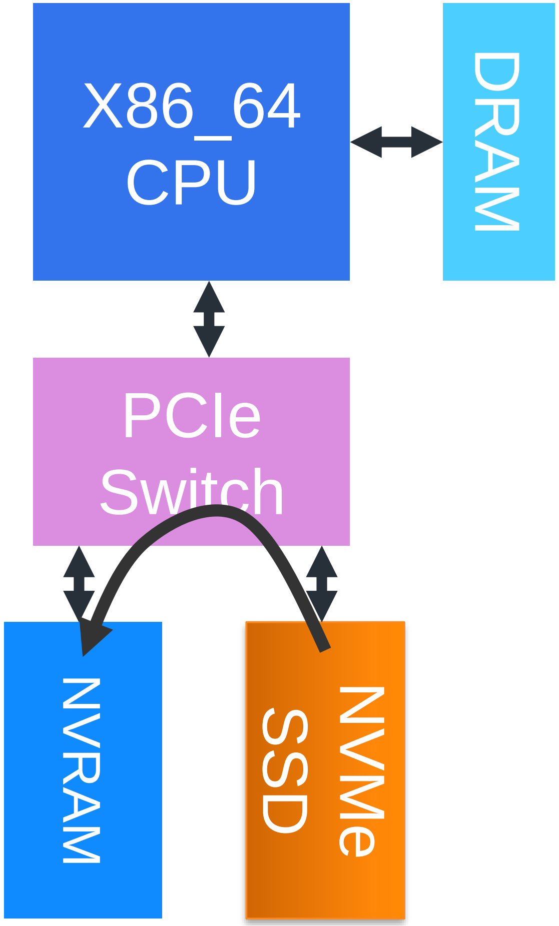
- CUDA 6.0 以上版本， 支持GPU Direct P2P传输

- 将DMA过程中的get_user_pages()， 以及put_pages()下游代码进行变更。通过对应的DAX设备的IOCTRL 下发P2P DMA传输。#define NVME_IOCTL_SUBMIT_GPU_IO _IOW('N', 0x45, struct nvme_gpu_io)

应用场景: NVMe SSD <-> NVRAM

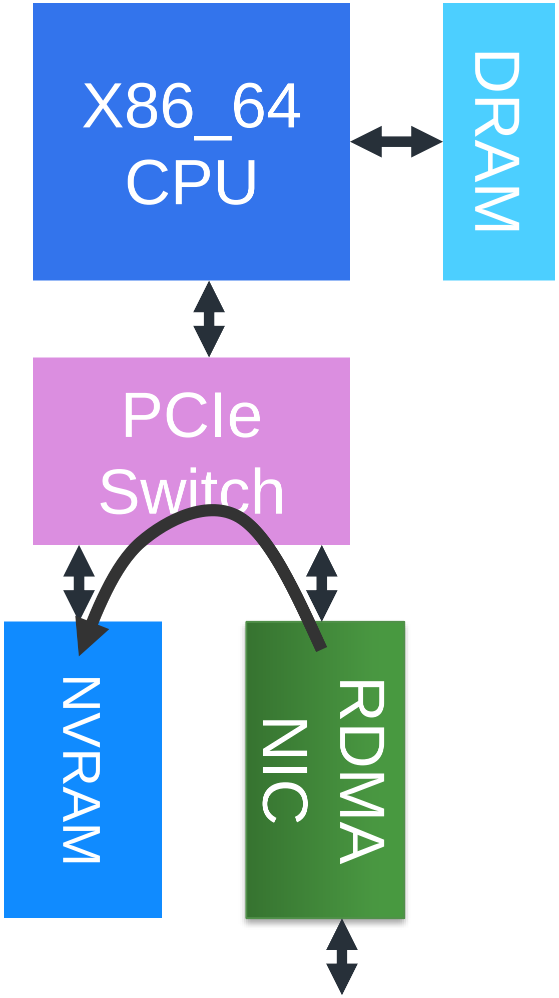
- NVRAM作为写缓存，NVMe SSD直接从NVRAM中读取数据并写入Flash， bypass本地CPU

- 可以使用NVRAM DAX设备提供的mmap()将文件映射到NVRAM锁对应的内存区域， NVMe SSD直接读写该区域。



应用场景: RDMA NIC <-> NVRAM

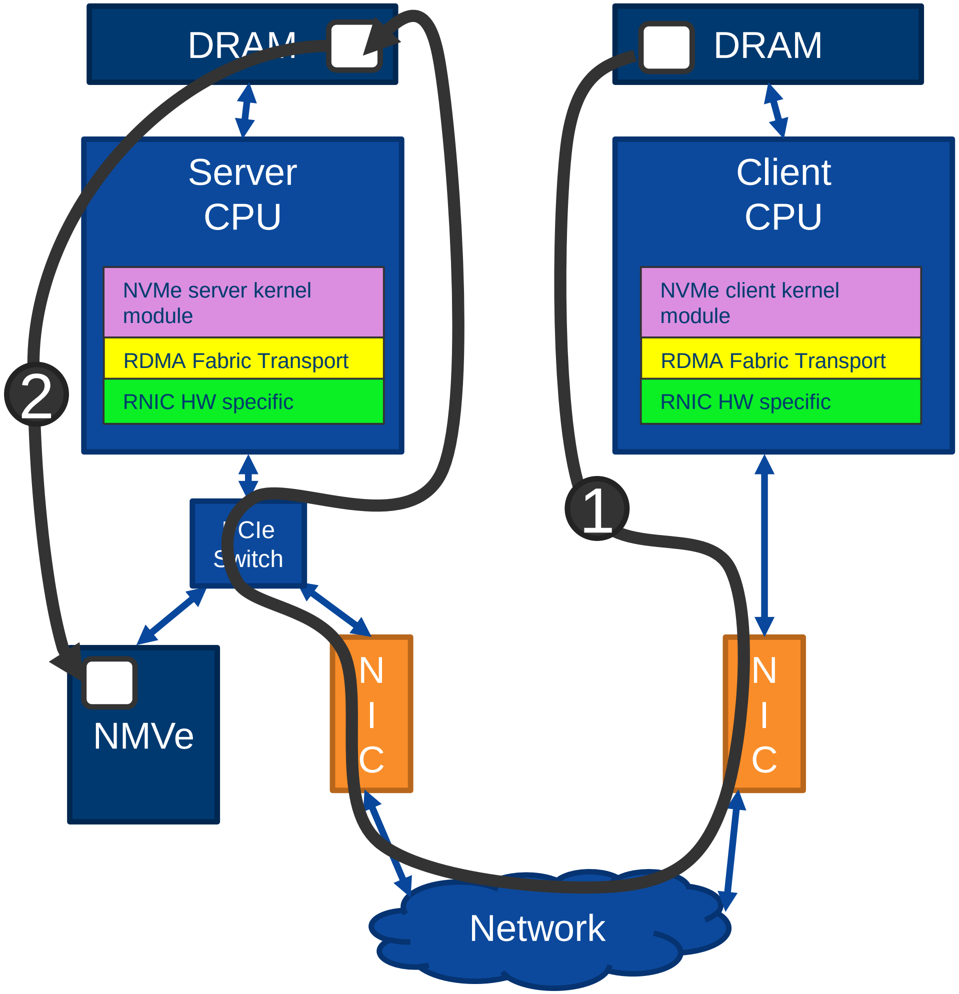
- 对端主机向本端NVRAM直接推送数据， 数据传送过程bypass两端的CPU
- 可以使用NVRAM DAX设备提供的mmap()将文件映射到对应的内存区域， 对端RDMA访问本端文件



应用场景: NVMe over Fabric with RDMA

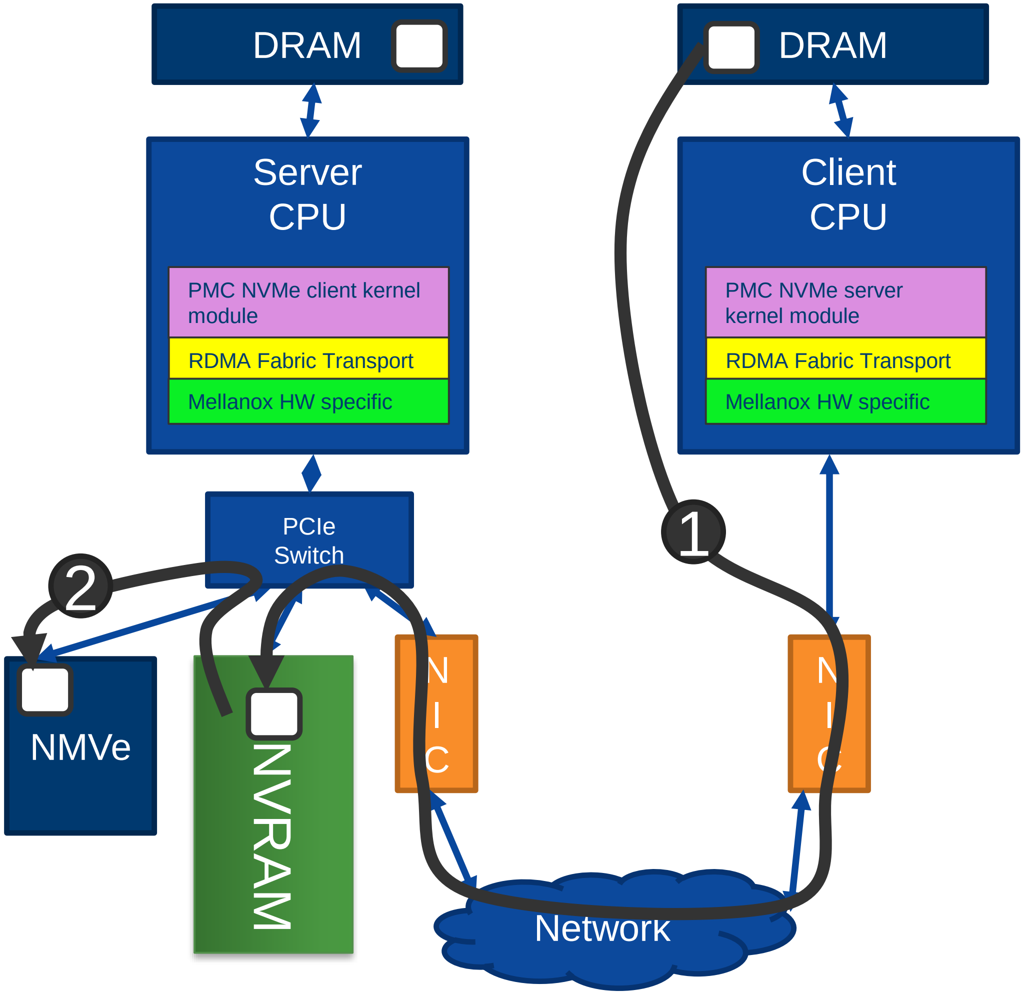
- NVMe Initiator端与Target端预先注册好双方各自的Memory Region，将对应的NVMe Queue映射到RNIC的Queue Pair，采用RDMA交互指令及数据，

- NVMe Target端程序接收到IO及数据之后，经过处理，向本地NVMe SSD发起IO读写请求。



应用场景: NVMEoF with RDMA & Cache & P2P

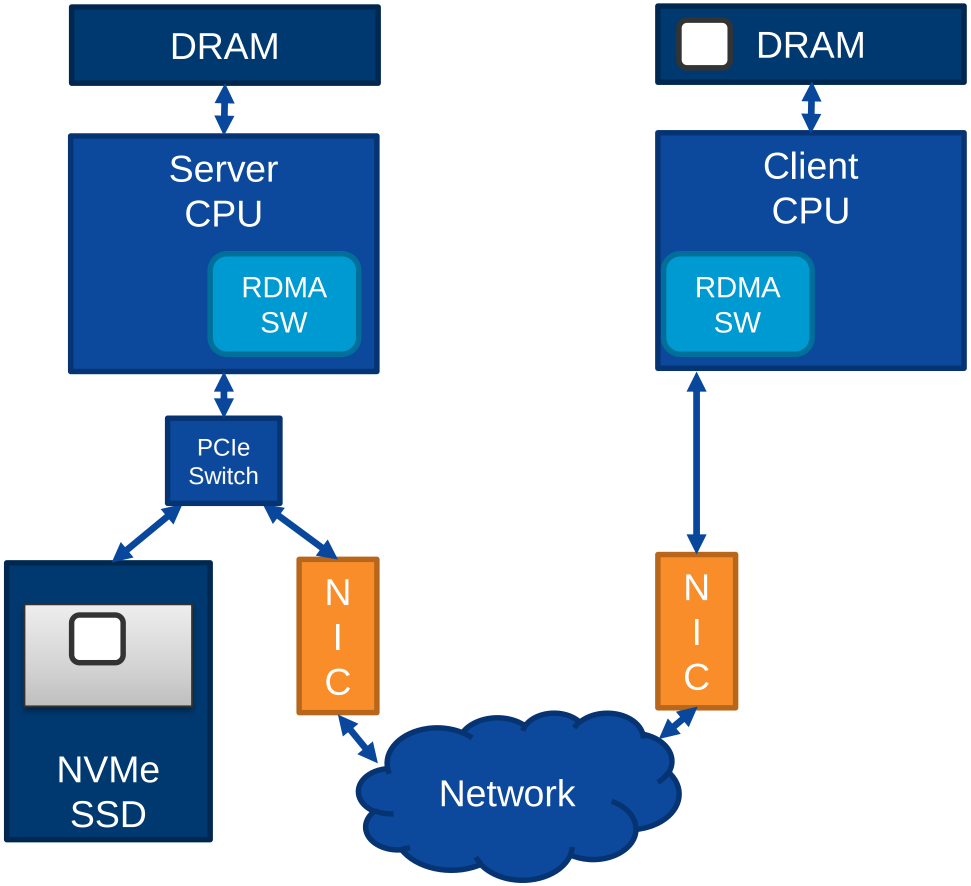
- NVMe Initiator端与Target端预先注册好双方各自的Memory Region, 其中一方落入NVRAM空间。将对应的NVMe Queue映射到RNIC的Queue Pair, 采用RDMA交互指令及数据。
- NVRAM作为写缓存，可降低全固态阵列的时延抖动。
- NVMe Target端程序接收到IO及数据之后, 向本地NVMe SSD发起IO读写请求，在数据传输时，NVMe SSD直接从NVRAM中将数据读出，bypass CPU。



应用场景: NVMEoF with RDMA & CMB & P2P

- NVMe CMB作为RNIC的MR，远端将指令和数据直接通过RDMA写入CMB。远端直接读写本地NVMe盘。

- NVMe IO completion消息依然需要本地代码处理。可以在NVMe盘固件中做开发支持RDMA verb，从而可bypass本地CPU。



Thank YOU!

Microsemi Corporation (Nasdaq: MSCC) offers a comprehensive portfolio of semiconductor and system solutions for aerospace & defense, communications, data center and industrial markets. Products include high-performance and radiation-hardened analog mixed-signal integrated circuits, FPGAs, SoCs and ASICs; power management products; timing and synchronization devices and precise time solutions, setting the world's standard for time; voice processing devices; RF solutions; discrete components; enterprise storage and communication solutions, security technologies and scalable anti-tamper products; Ethernet solutions; Power-over-Ethernet ICs and midspans; as well as custom design capabilities and services. Microsemi is headquartered in Aliso Viejo, Calif., and has approximately 4,800 employees globally. Learn more at www.microsemi.com

Power Matters.™

Microsemi Corporate Headquarters
One Enterprise, Aliso Viejo, CA 92656 USA
Within the USA: +1 (800) 713-4113
Outside the USA: +1 (949) 380-6100
Sales: +1 (949) 380-6136
Fax: +1 (949) 215-4996
email: sales.support@microsemi.com
www.microsemi.com

Microsemi makes no warranty, representation, or guarantee regarding the information contained herein or the suitability of its products and services for any particular purpose, nor does Microsemi assume any liability whatsoever arising out of the application or use of any product or circuit. The products sold hereunder and any other products sold by Microsemi have been subject to limited testing and should not be used in conjunction with mission-critical equipment or applications. Any performance specifications are believed to be reliable but are not verified, and Buyer must conduct and complete all performance and other testing of the products, alone and together with, or installed in, any end-products. Buyer shall not rely on any data and performance specifications or parameters provided by Microsemi. It is the Buyer's responsibility to independently determine suitability of any products and to test and verify the same. The information provided by Microsemi hereunder is provided "as is, where is" and with all faults, and the entire risk associated with such information is entirely with the Buyer. Microsemi does not grant, explicitly or implicitly, to any party any patent rights, licenses, or any other IP rights, whether with regard to such information itself or anything described by such information. Information provided in this document is proprietary to Microsemi, and Microsemi reserves the right to make any changes to the information in this document or to any products and services at any time without notice.