



Flash Memory Summit

Using PCIe® Fabrics to Push Disaggregation and Composability into the New Era

Chetana Kaushik, Senior Applications Engineer
Microchip Technology

Agenda



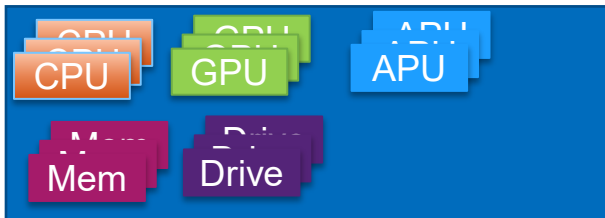
Flash Memory Summit

- Disaggregation and Composability
- PCIe Switching Limitations
- Switch Fabric Solution
- Switch Fabric Examples

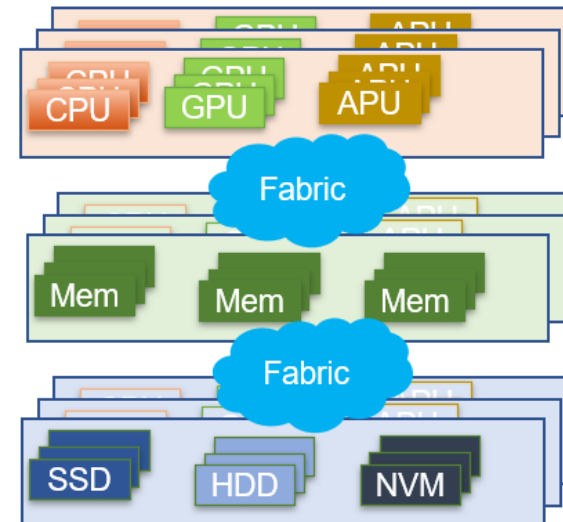


Disaggregation and Composability

- Increased market demand for Compute, Network, Memory and Storage devices using PCIe fabrics
- System designers need:
 - Efficient resource deployment, High BW, low latency interconnect
 - Flexible, composable architectures



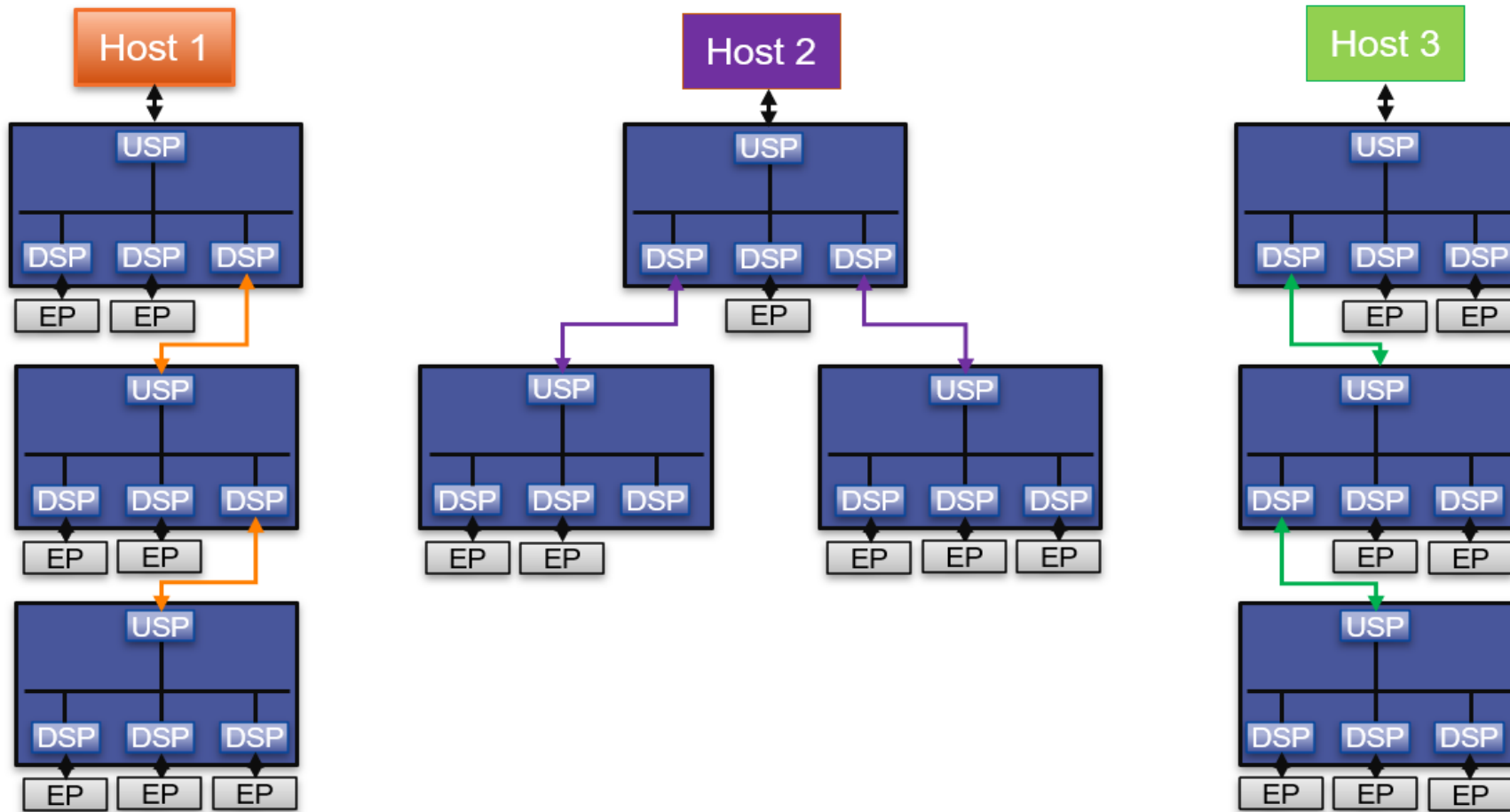
Hyper Converged Infrastructure



Composable Disaggregate Infrastructure

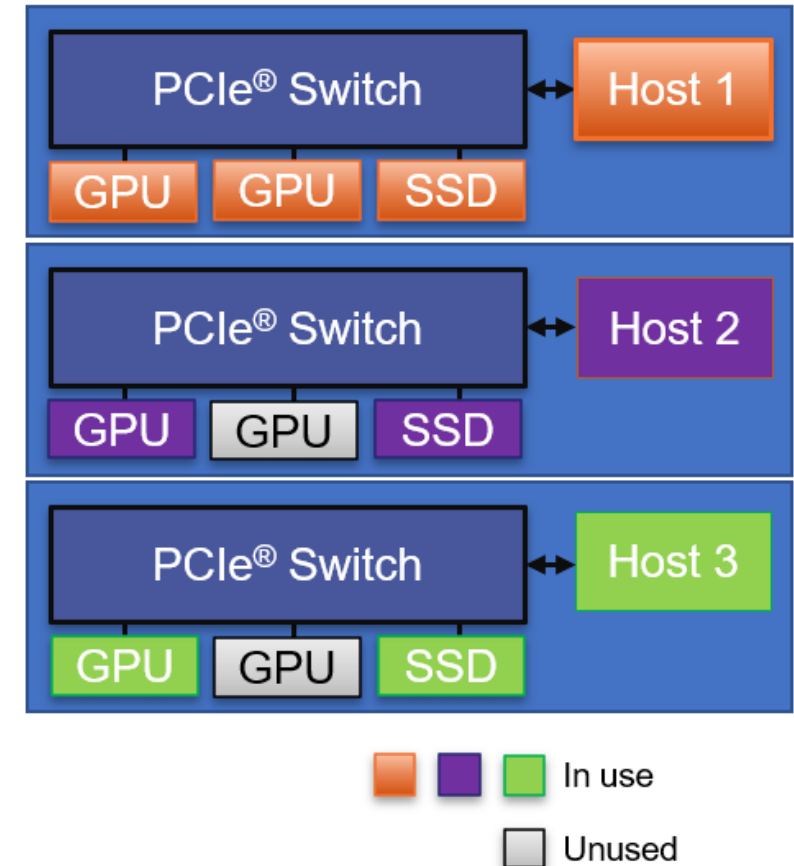
Standard PCIe[®] Hierarchy Restriction

- Standard PCIe hierarchy is restrictive, making scale out challenging



Standard PCIe® Single Domain Restriction

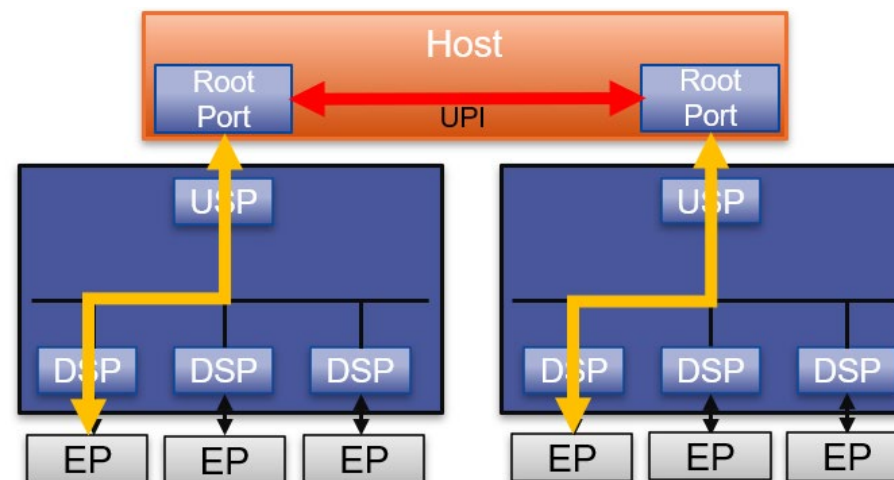
- PCIe is single domain
 - Unused EPs are stranded
 - Complicated, non-standard NT drivers required for sharing
- Multi-function limitations
 - All EP functions must belong to a single Root Complex
 - Under utilized EPs can't be shared





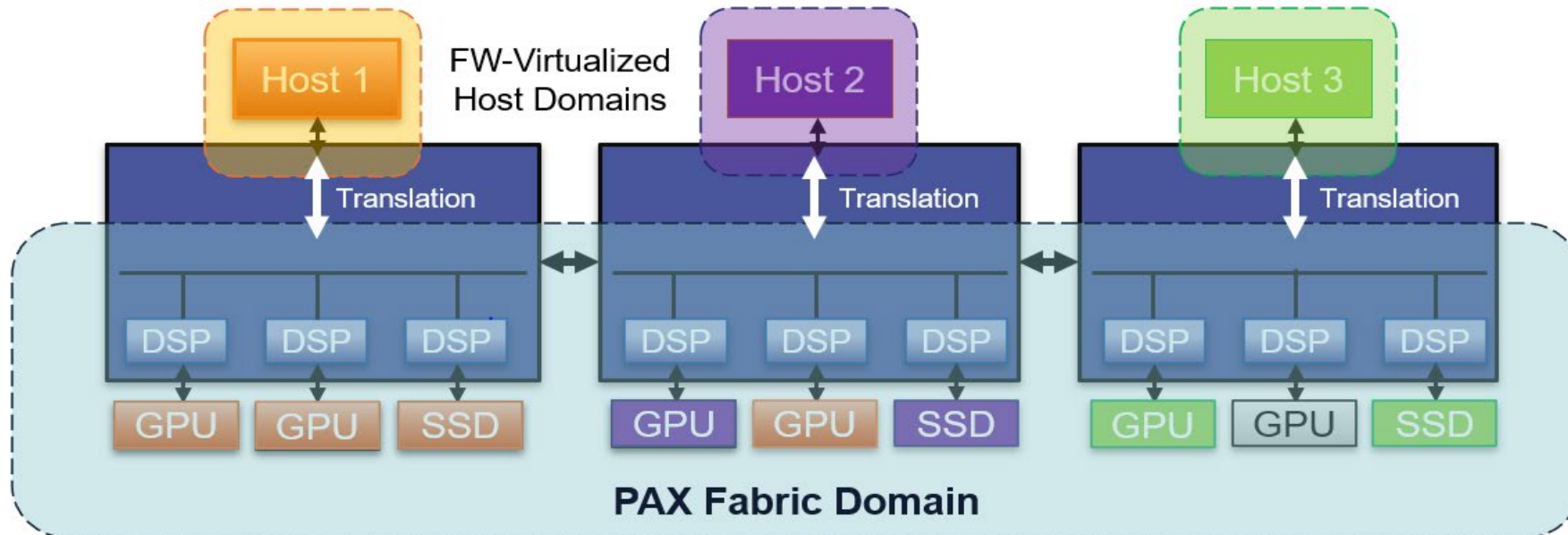
Standard PCIe® Routing Restrictions

- Limitation with standard PCIe: traffic must flow up through the Root Complex to reach another switch tree; loops and redundant paths are not supported
- Tree structure for routing
- Performance on multi-core systems bottlenecked by UPI



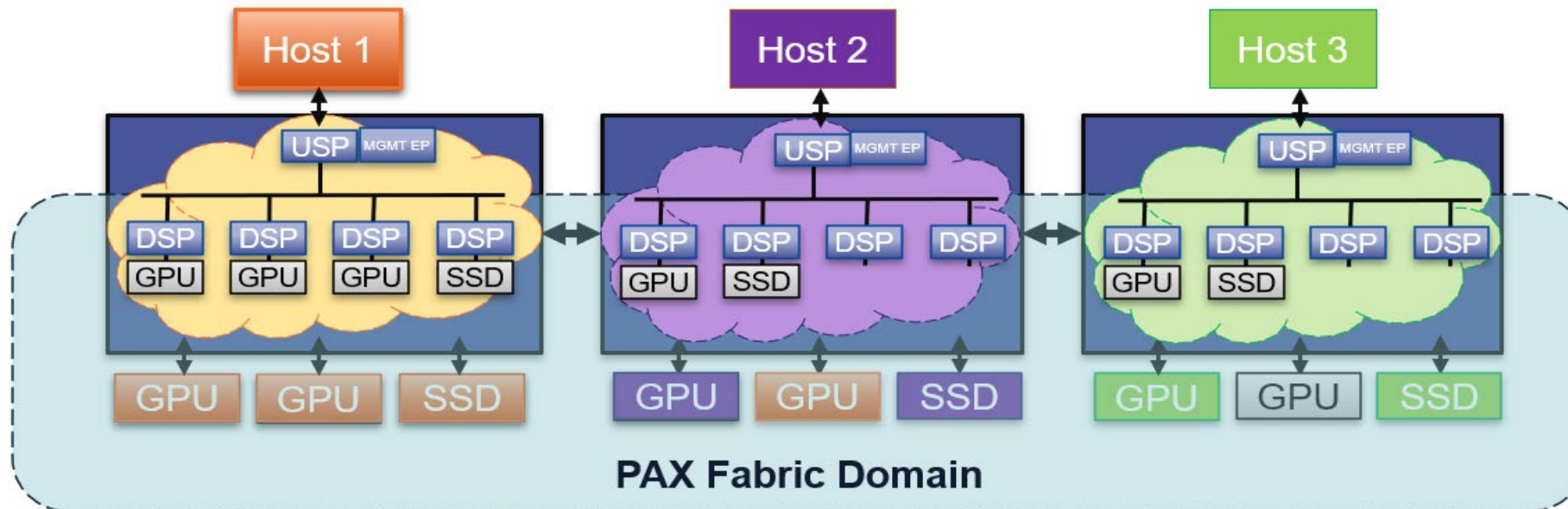
Switch Fabric - Enumeration

- FW on each switch enumerates EPs with address, ID from fabric domain
- Host enumerates a virtual EP, and translations are applied to MemRd/Wrs



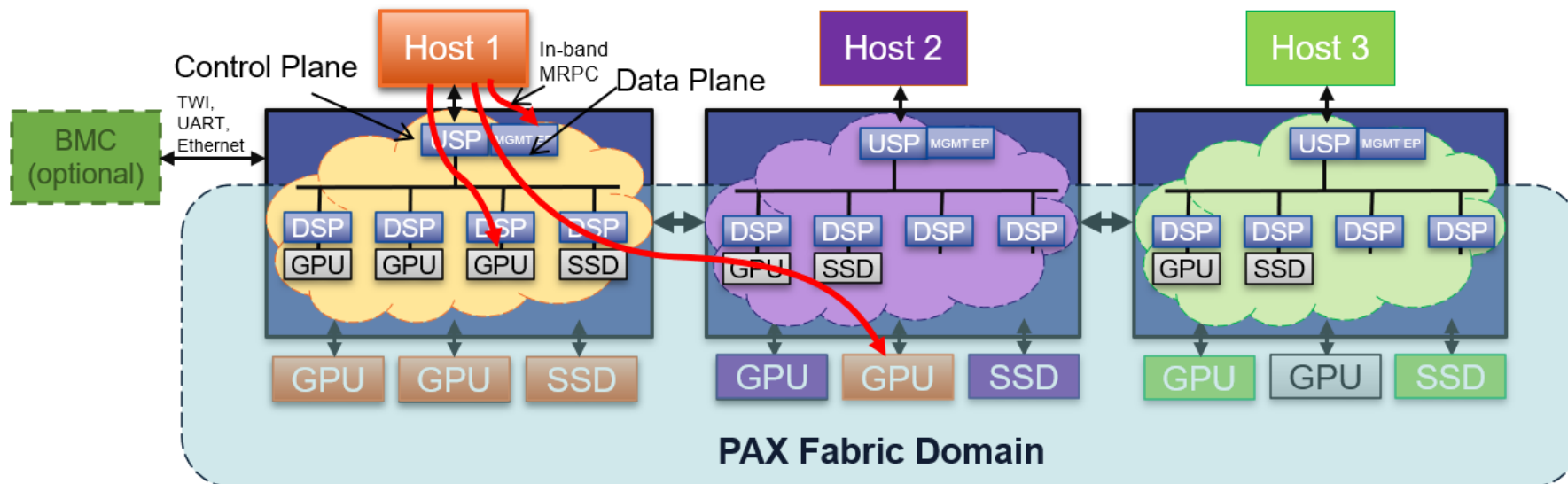
Switch Fabric - Routing

- Fabric routing is innovative, non-hierarchical
- Fabric links are shared among hosts
- Embedded FW virtualizes simple, PCIe spec-compliant switch
- Fabric details abstracted; EPs appear directly connected to switch



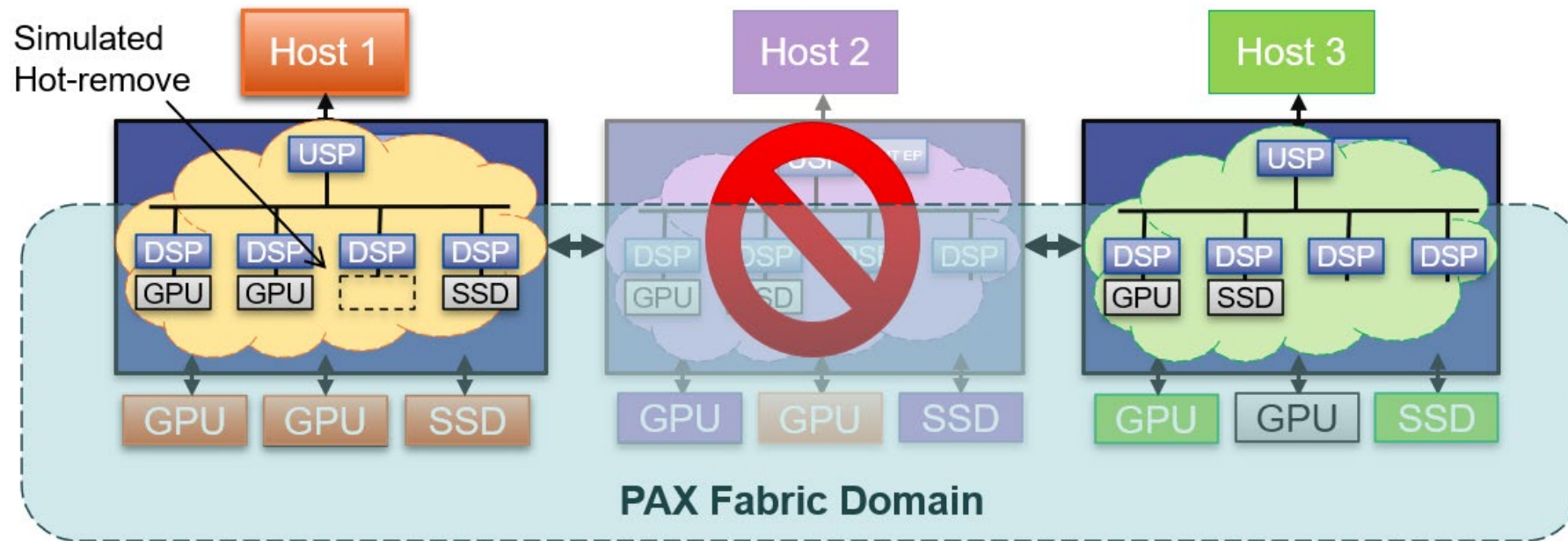
Switch Fabric - Management

- Embedded CPU handles control plane, data is routed by switch HW
- Fabric can be managed via PCIe, TWI, UART, Ethernet or FW SDK
- Fabric is managed through simple MRPC interface (bindings, debug)
- Kernel driver and user-space management tool to be made available



Switch Fabric – Fault Isolation

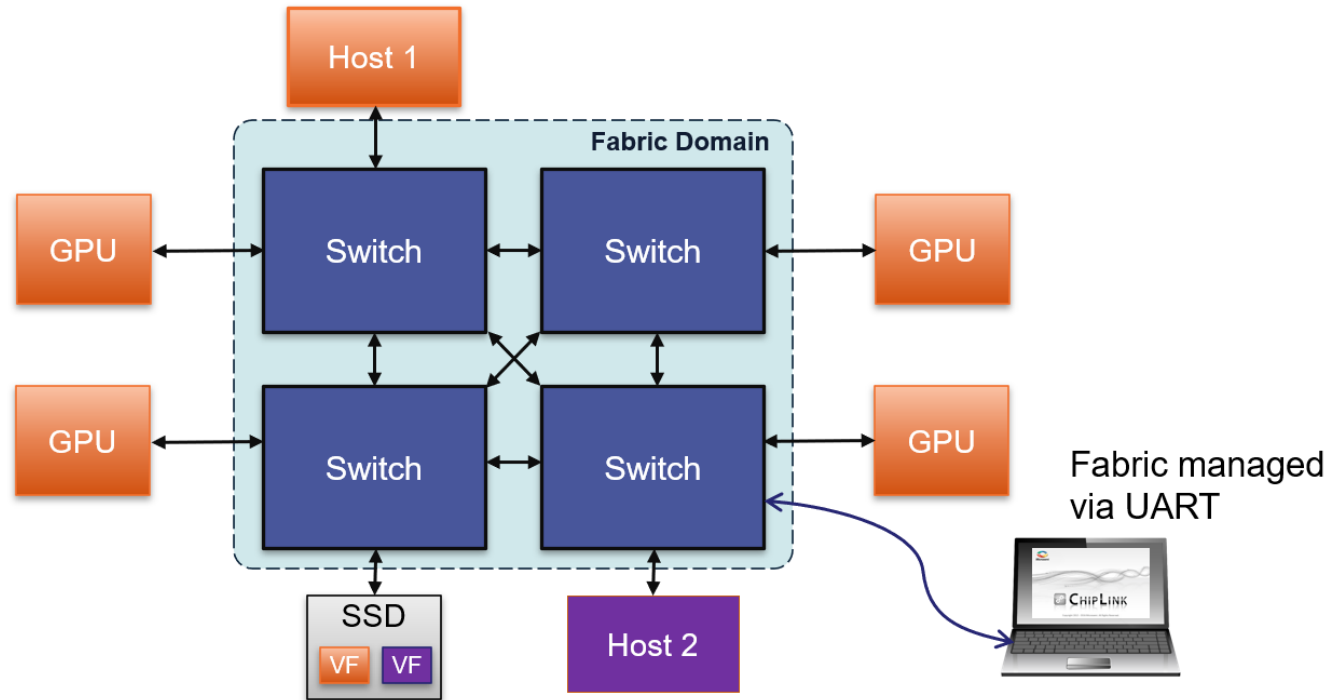
- Host domain virtualization isolates host from errors in fabric (AER, DPC, etc.)
- Host only sees spec-compliant hot-remove, simulated by fabric FW



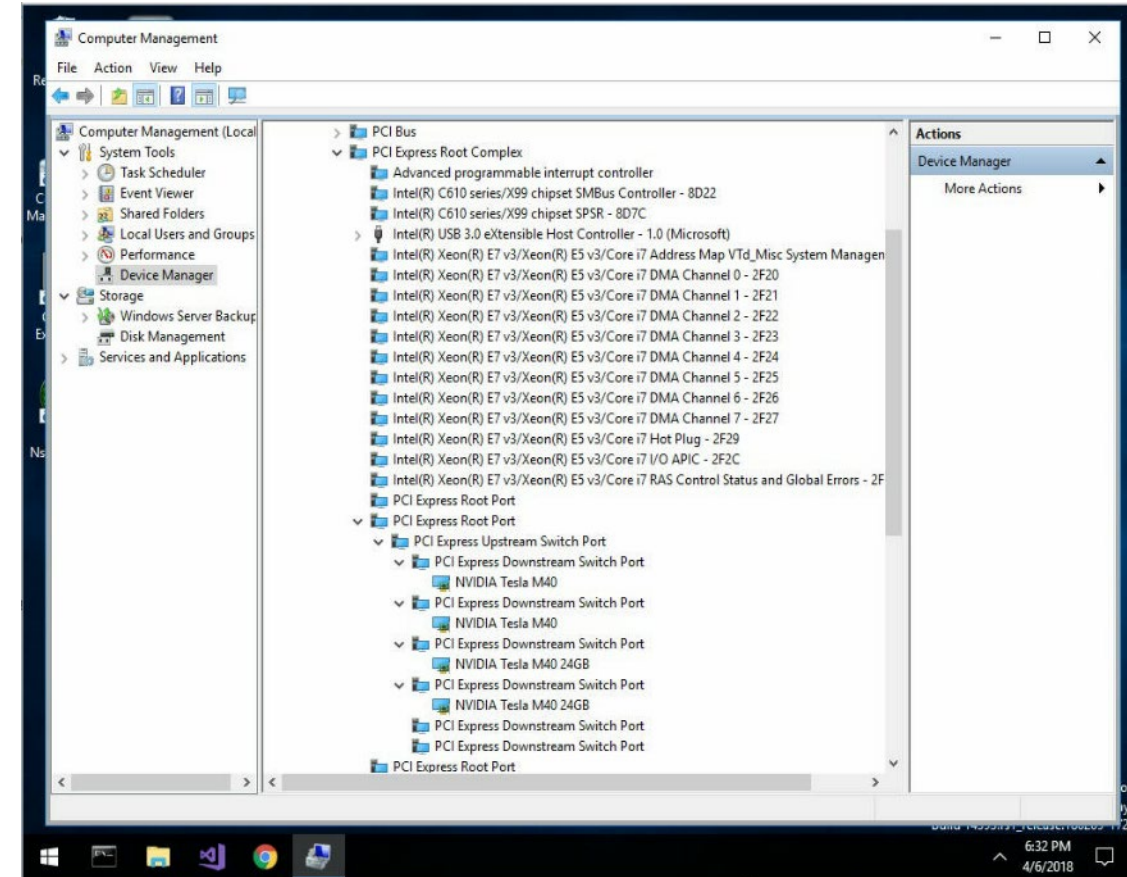
Examples of Disaggregation and Composability

- Dynamic Resource Allocation
- Multi-host sharing

Dynamic Assignment of Pooled GPUs



All GPUs are assigned to Host 1
to increase performance





Dynamic Assignment of Pooled GPUs

CUDA P2P Bandwidth

P2P Connectivity Matrix

D\D	0	1	2	3
0	1	1	1	1
1	1	1	1	1
2	1	1	1	1
3	1	1	1	1

Unidirectional P2P=Enabled Bandwidth Matrix (GB/s)

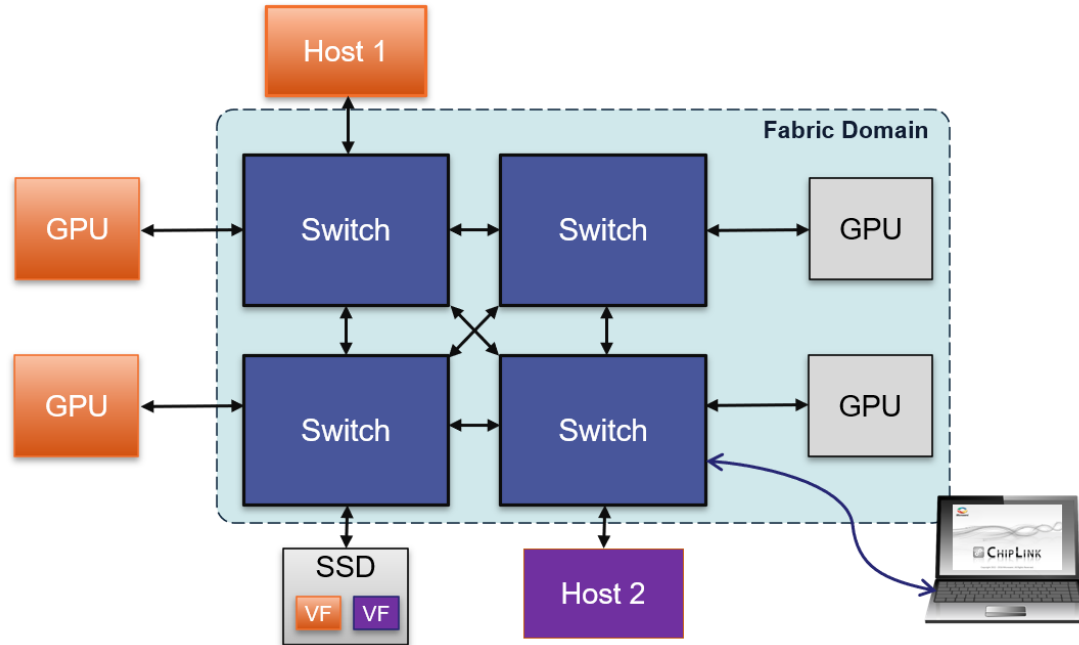
D\D	0	1	2	3
0	210.61	12.96	12.54	12.53
1	12.52	211.35	13.08	13.06
2	12.52	12.52	212.61	13.05
3	13.06	13.06	12.54	211.36

Bidirectional P2P=Enabled Bandwidth Matrix (GB/s)

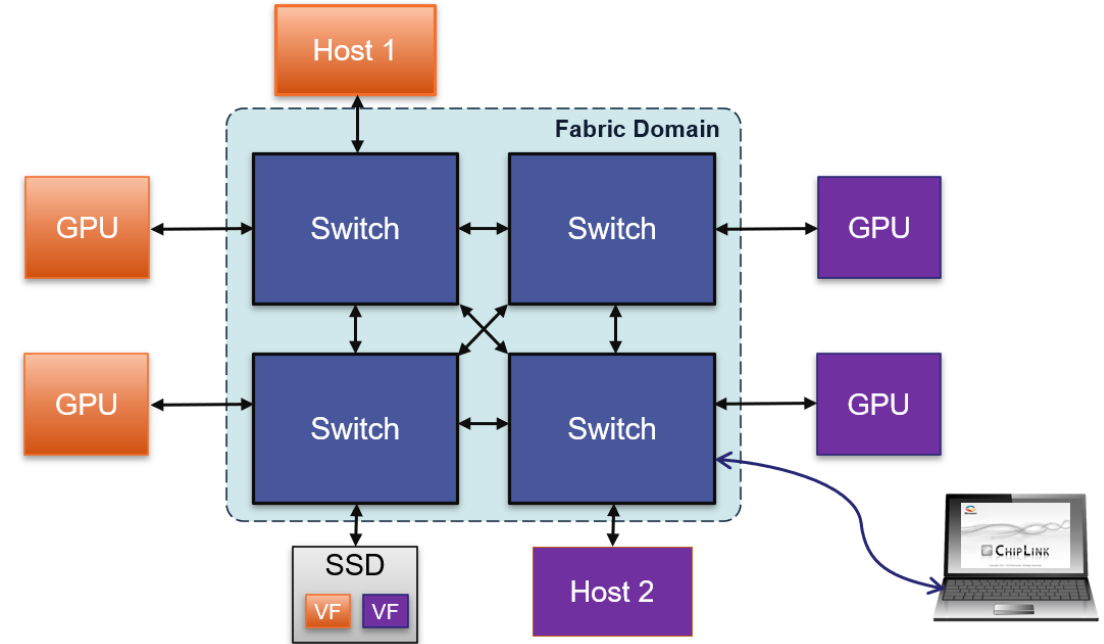
D\D	0	1	2	3
0	213.51	24.81	24.77	24.72
1	24.73	213.55	24.73	25.74
2	24.53	24.57	214.73	24.86
3	24.83	25.72	24.73	214.58

Dynamic Assignment of Pooled GPUs

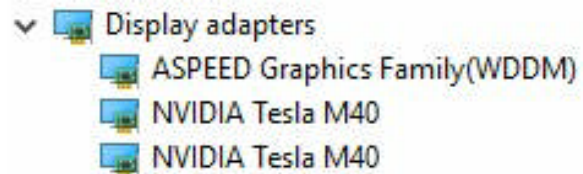
Host 1 workload completes, and GPUs are released back into fabric pool



Spare GPUs are assigned to Host 2

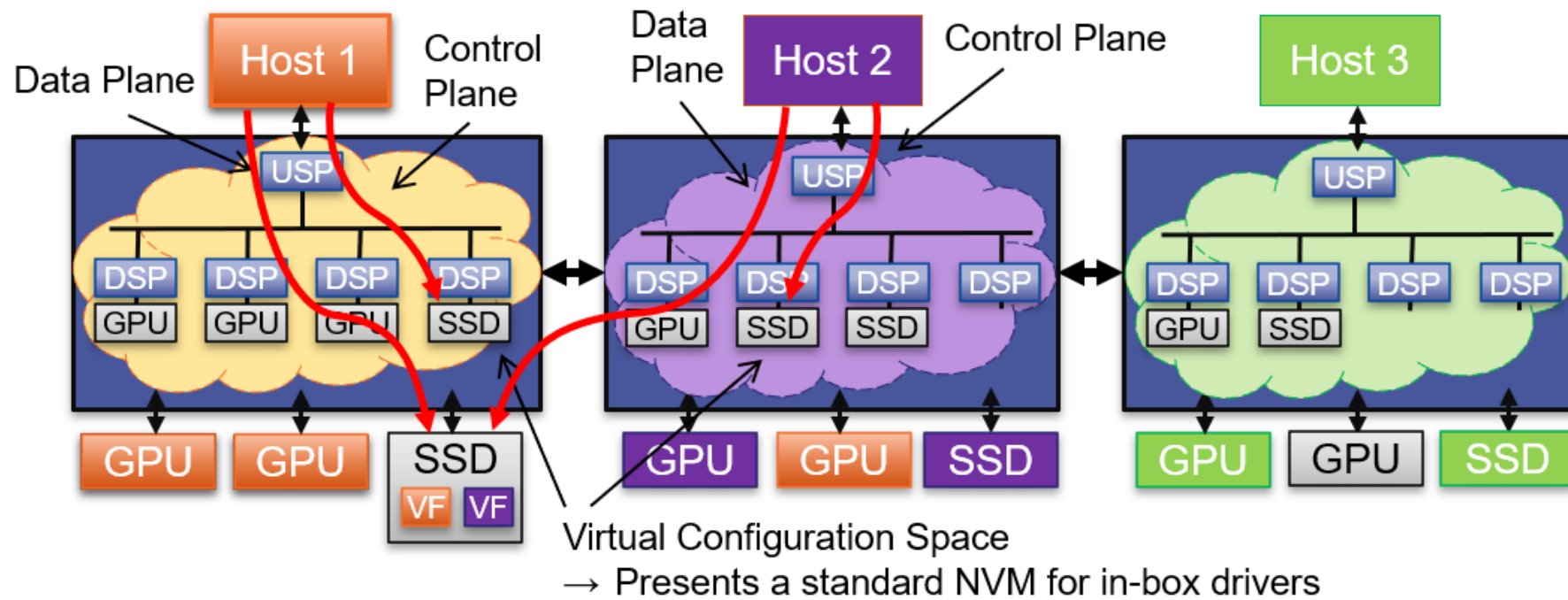


Windows Host Still Running During Dynamic Reassignment



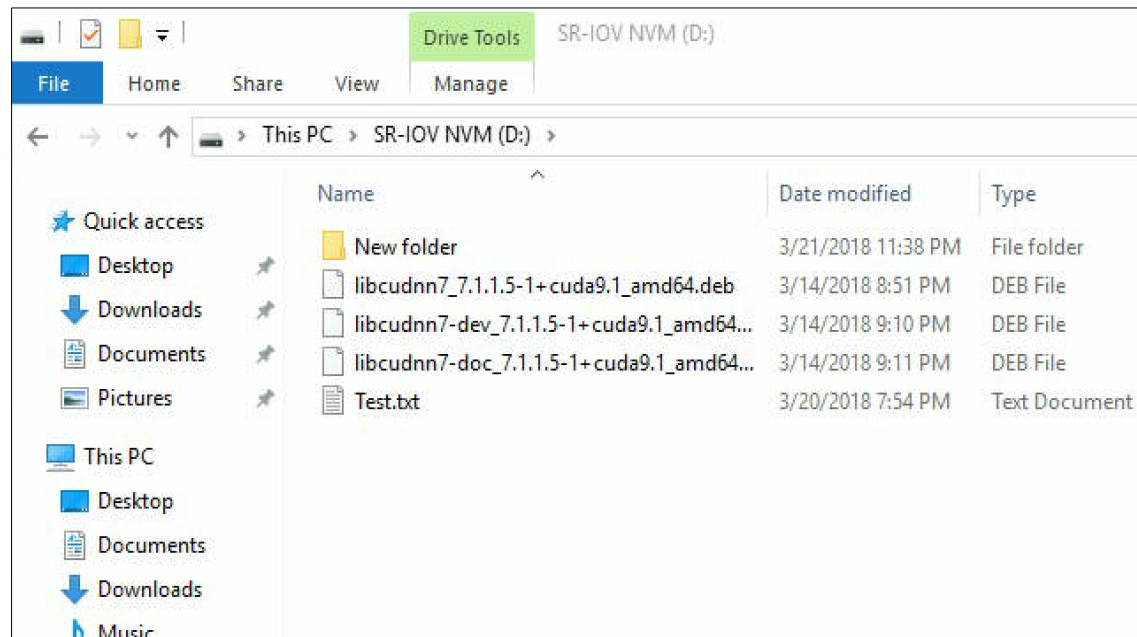
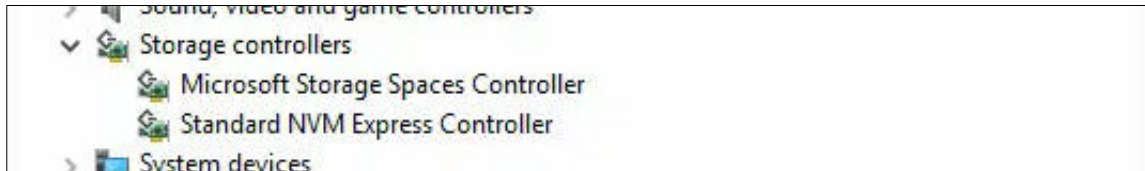
Switch Fabrics - Disaggregation/Sharing

- SR-IOV: EP appears as multiple functions
- Fabric resources assigned by function to multiple hosts



Multi-host Sharing of NVMe

NVM VF Appears as Standard NVM Device



```
00.0-[03-07]--+-00.0-[04]----00.0 NVIDIA Corporation GM200GL [Tesla M40]
+-01.0-[05]----00.0 NVIDIA Corporation GM200GL [Tesla M40]
+-02.0-[06]----00.0 Samsung Electronics Co Ltd Device a822
\-03.0-[07]--
```

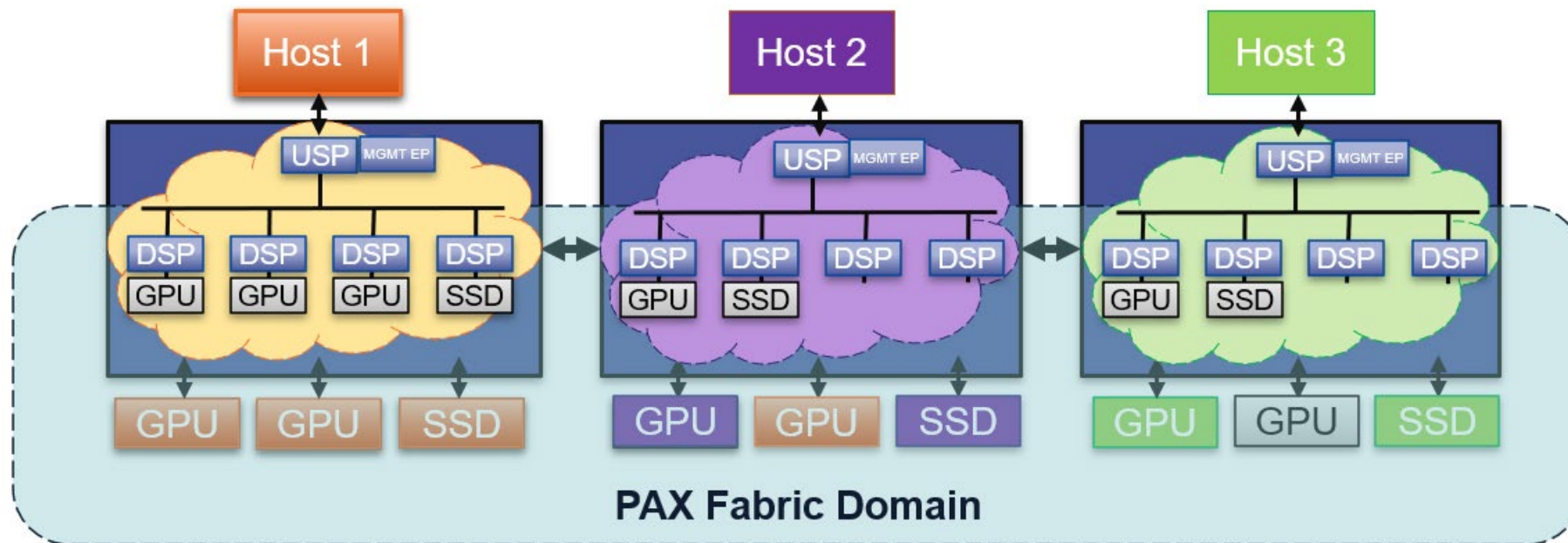
```
ubuntu@bbyapps-ubuntu1604-se:~$ ls /dev
autofs          hidraw1         loop3           nvme0n1p2      sg0             tty2            tty4            tty6            ttyS20         userio
bbyapps-ubuntu1604-se-vg hidraw2         loop4           port           sg1             tty20           tty40           tty60           ttyS21         vcs
block           hidraw3         loop5           ppp            sg2             tty21           tty41           tty61           ttyS22         vcs1
log             hpet           loop6           psaux          shm             tty22           tty42           tty62           ttyS23         vcs2
btrfs-control   hugepages      loop7           ptmx           snapshot        tty23           tty43           tty63           ttyS24         vcs3
bus             hwrng          loop-control    ptp0           sd              tty24           tty44           tty7            ttyS25         vcs4
char            i2c-0          mapper          ptp1           stderr          tty25           tty45           tty8            ttyS26         vcs5
console         i2c-1          mcelog          pts            stdin           tty26           tty46           tty9            ttyS27         vcs6
core            i2c-2          mem             random          stdout          tty27           tty47           ttyprintk       ttyS28         vcsa
cpu             i2c-3          memory_bandwidth rfcill         tty             tty28           tty48           ttyS0           ttyS29         vcsa1
cpu_dma_latency i2c-4          queue           rtc            tty0            tty29           tty49           ttyS1           ttyS3           vcsa2
cuse            i2c-5          wet            rtc0           tty1            tty3            tty5            ttyS10          ttyS30         vcsa3
disk           i2c-6          network_latency sda            tty10           tty30           tty50           ttyS11          ttyS31         vcsa4
dn-0            initctl        network_throughput sda1           tty11           tty31           tty51           ttyS12          ttyS4           vcsa5
dn-1            input          null            sda2           tty12           tty32           tty52           ttyS13          ttyS5           vcsa6
dri             kmsg           nvidia0         sda5           tty13           tty33           tty53           ttyS14          ttyS6           vfi
ecryptfs        kvm            nvidia1         sdb            tty14           tty34           tty54           ttyS15          ttyS7           vga_arbiter
fb0             lightnvm       nvidiaact1      sdb1           tty15           tty35           tty55           ttyS16          ttyS8           vhci
fd              log            nvidia-uvm      sdb2           tty16           tty36           tty56           ttyS17          ttyS9           vhost-net
full            loop0          nvme0            sdb3           tty17           tty37           tty57           ttyS18          uhid           zero
fuse            loop1          nvme0n1          sdc            tty18           tty38           tty58           ttyS19          uinput
hidraw0         loop2          nvme0n1p1        sdc1           tty19           tty39           tty59           ttyS2           urandom
```

```
ubuntu@bbyapps-ubuntu1604-se:~$ sudo mount /dev/nvme0n1p2 /mnt
ubuntu@bbyapps-ubuntu1604-se:~$ ls /mnt
lib cudnn7_7.1.1.5-1+cuda9.1_amd64.deb  lib cudnn7-doc_7.1.1.5-1+cuda9.1_amd64.deb  $RECYCLE.BIN  Test.txt
lib cudnn7-dev_7.1.1.5-1+cuda9.1_amd64.deb  New folder  System Volume Information
```

```
ubuntu@bbyapps-ubuntu1604-se:~$
```


Switch Fabric - Advanced Solutions

- Virtual Host Domain offers complete control over switch and EP attributes
- SDK allows customers to modify EP CSR contents, customize bindings, implement enclosure management application and more



Using PAX Fabrics to Push Disaggregation and Composability

- Increased market demand for GPUs and NVMe® drives using PCIe fabrics
- System designers need:
 - Efficient resource deployment, High BW, low latency interconnect
 - Flexible, composable architectures
- Benefits of PCIe fabrics with PAX:
 - Scalable, low-latency, cost-effective
 - Simple Management (PCIe, UART, TWI, Ethernet)
 - Multi-host sharing of SR-IOV NVMe devices
- Additional new features for Gen5
 - Enhanced on-chip PCIe analyzer for debugging TLPs per port and Ordered Sets per lane
 - Enhanced LTSSM monitor for advanced triggering
 - Automatic port bifurcation



Switchtec™
PAX NXG5



Switchtec™
PAX NXG4



Thank you !
Visit Microchip Booth # 613