Flash Memory Summit

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

Chetana Kaushik, Senior Applications Engineer

Microchip Technology

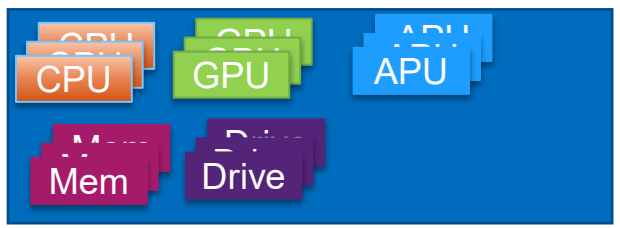
Agenda

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

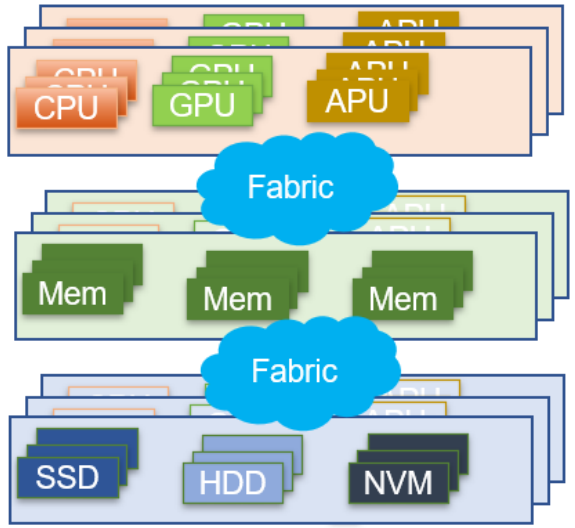
Disaggregation and Composability

- 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



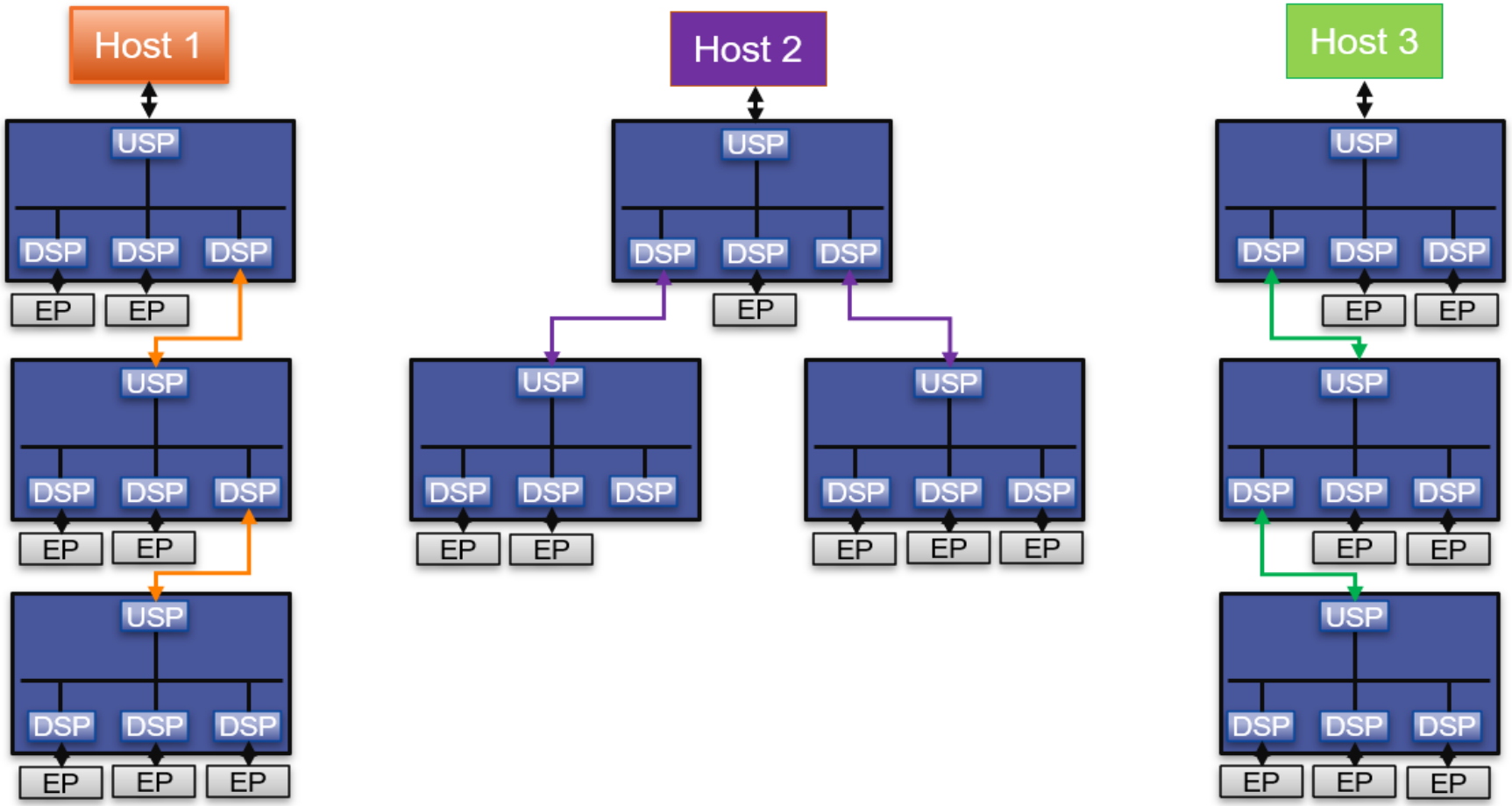
Hyper Converged Infrastructure



Composable Disaggregate Infrastructure

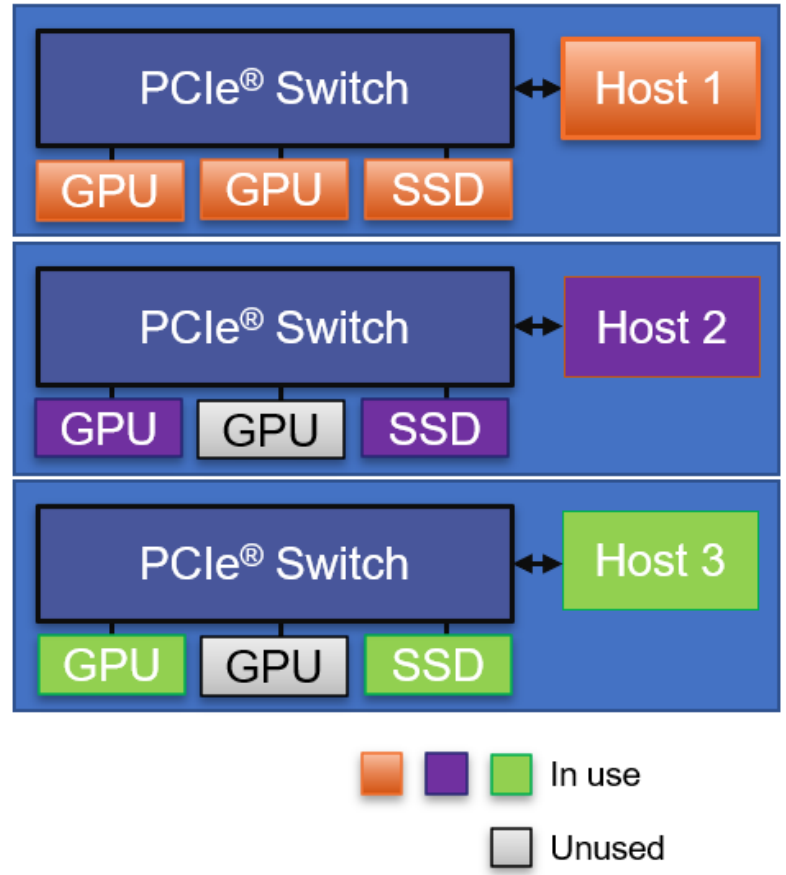
Standard PCIe® Hierarchy Restriction

- Standard PCIe hierarchy is restrictive, making scale out challenging



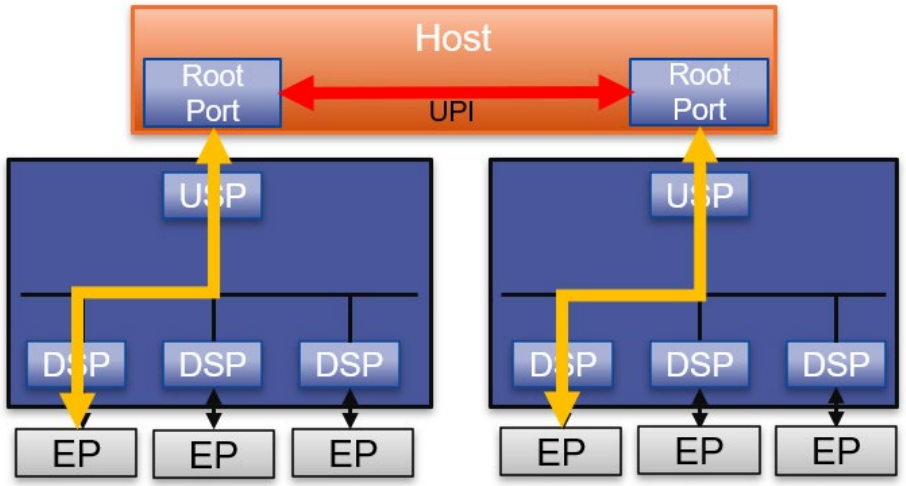
Standard PCIe® Single Domain Restriction

- 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



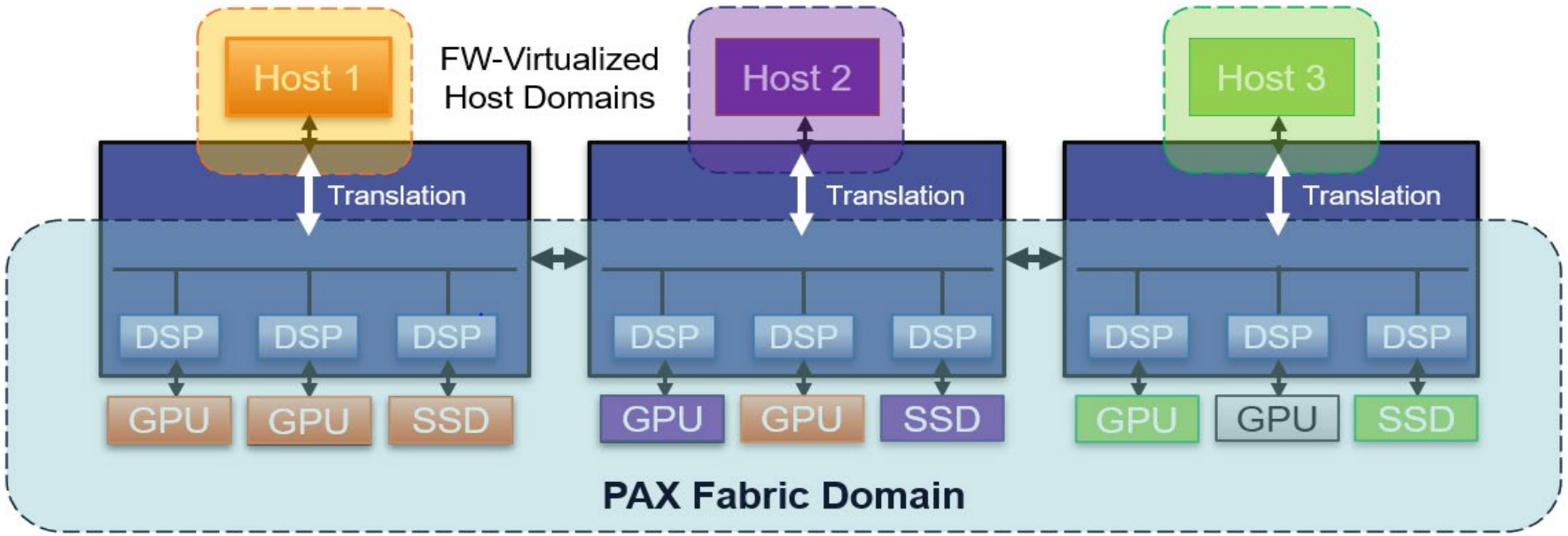
Standard PCIe® Routing Restrictions

- 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



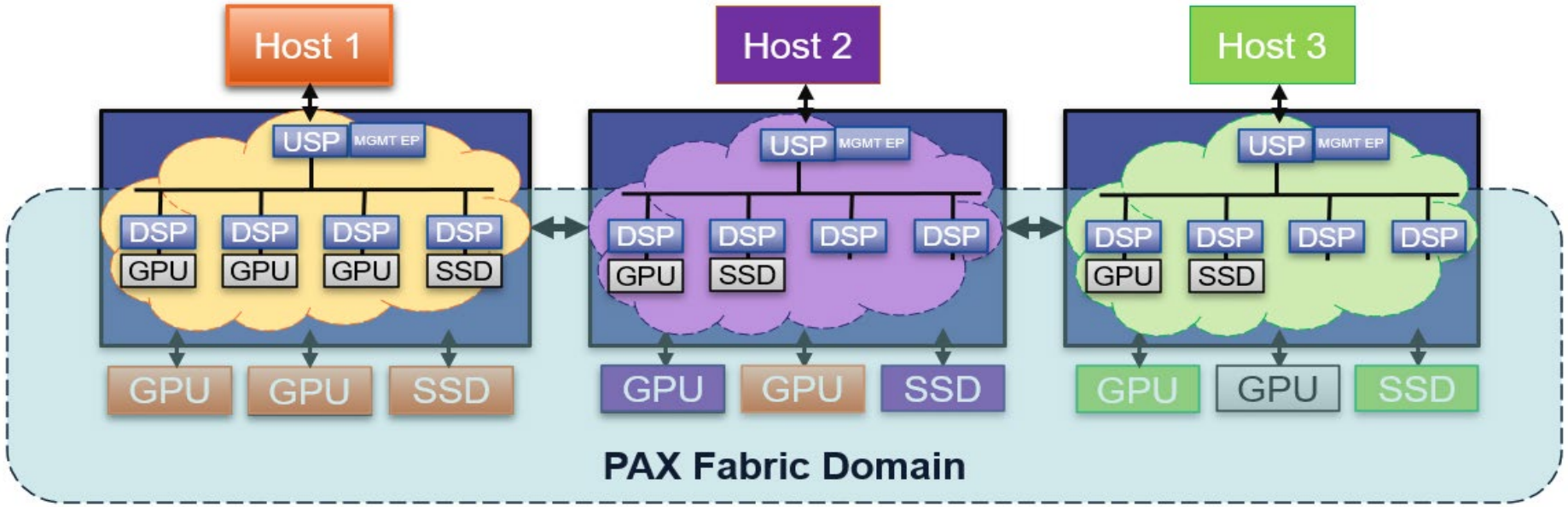
Switch Fabric - Enumeration

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



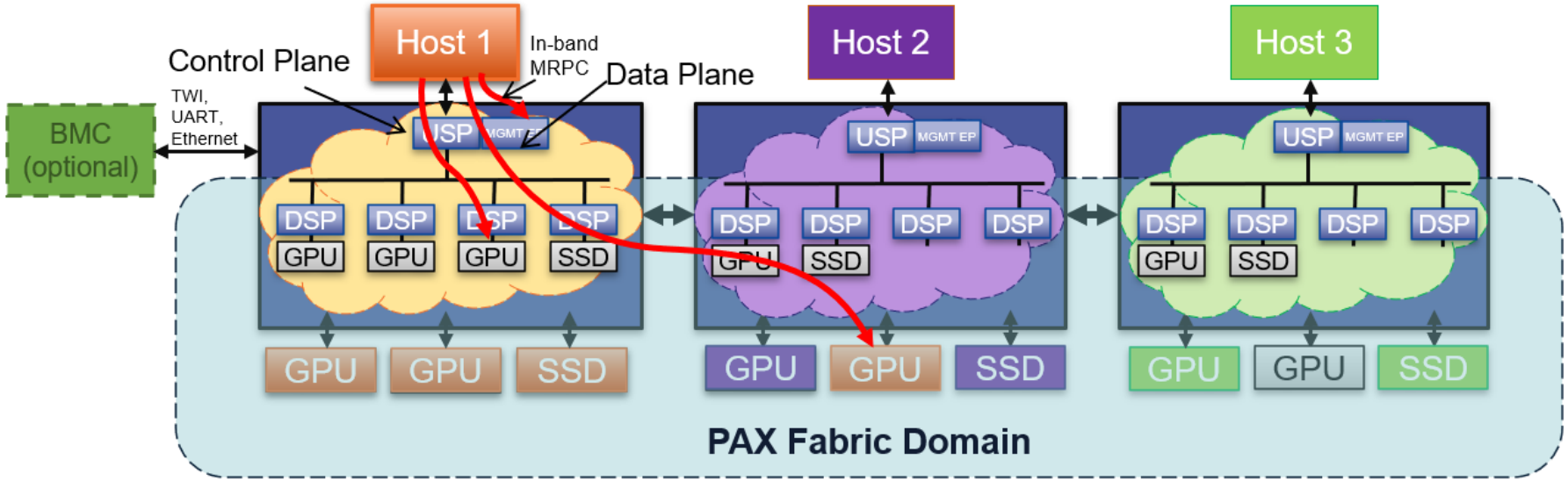
Switch Fabric - Routing

- 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



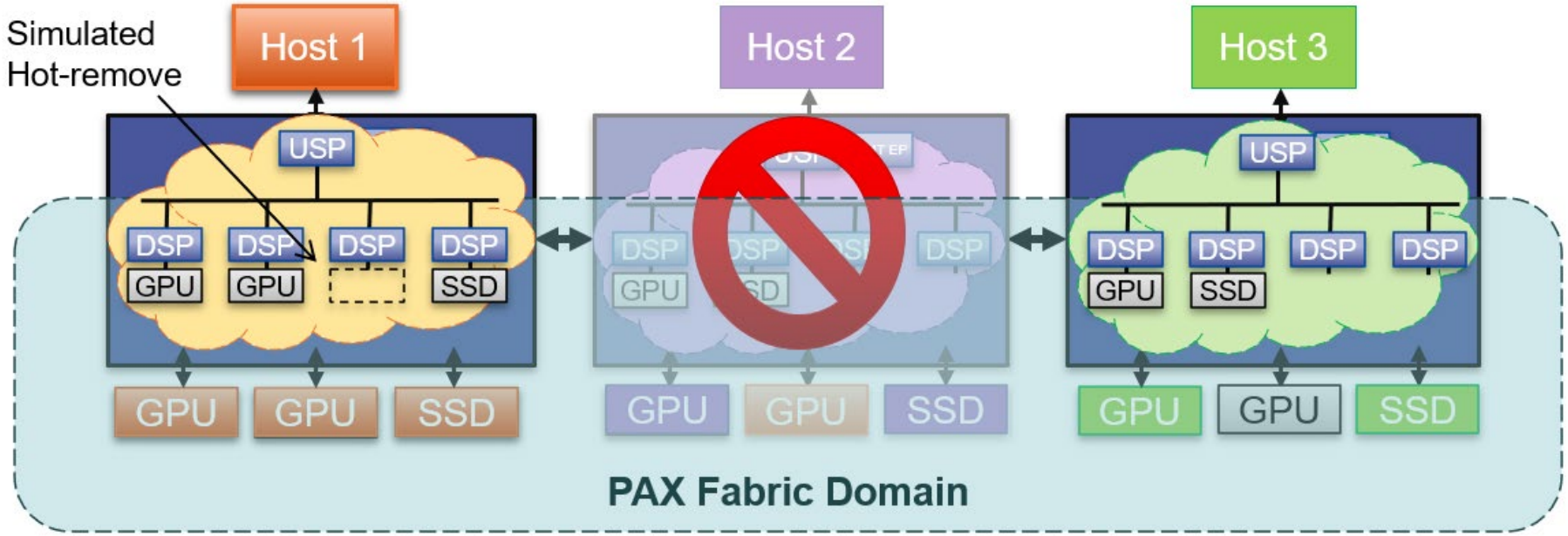
Switch Fabric - Management

- 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

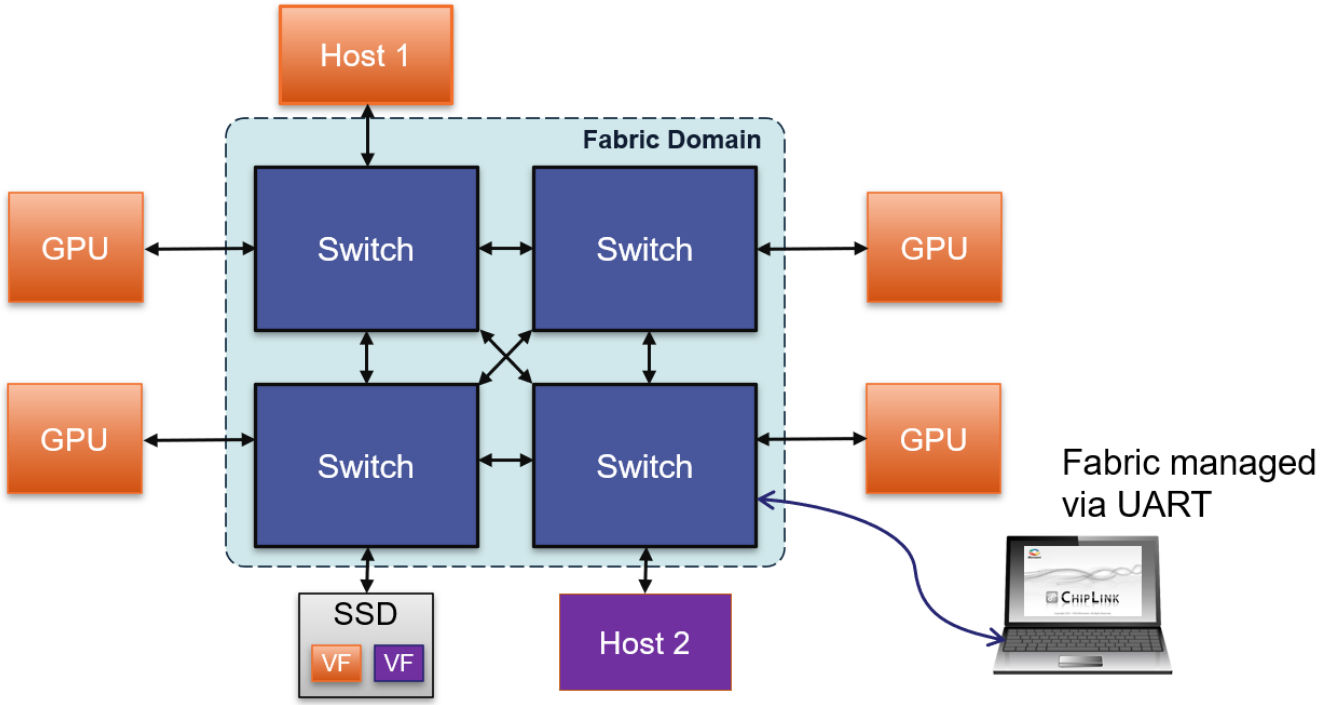
- Host domain virtualization isolates host from errors in fabric (AER, DPC, etc.)
- Host only sees spec-compliant hot-remove, simulated by fabric FW

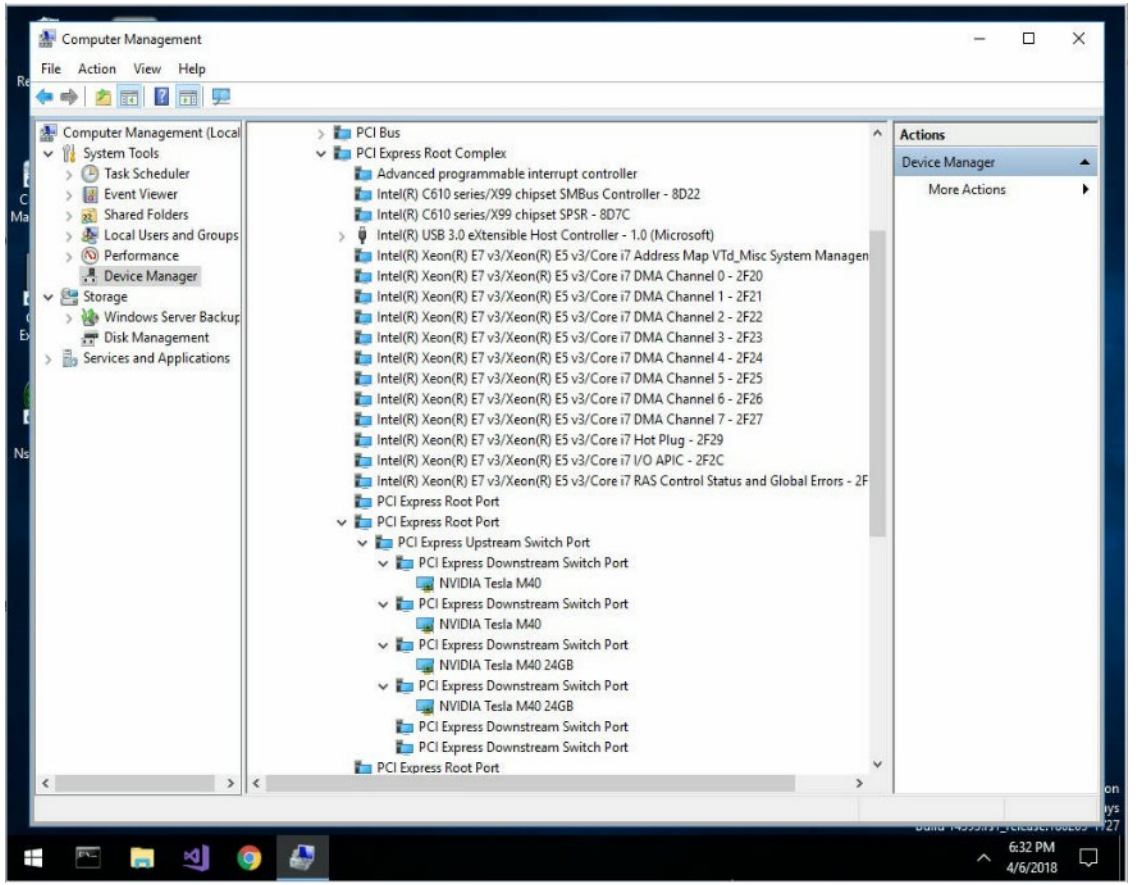


Examples of Disaggregation and Composability

- 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

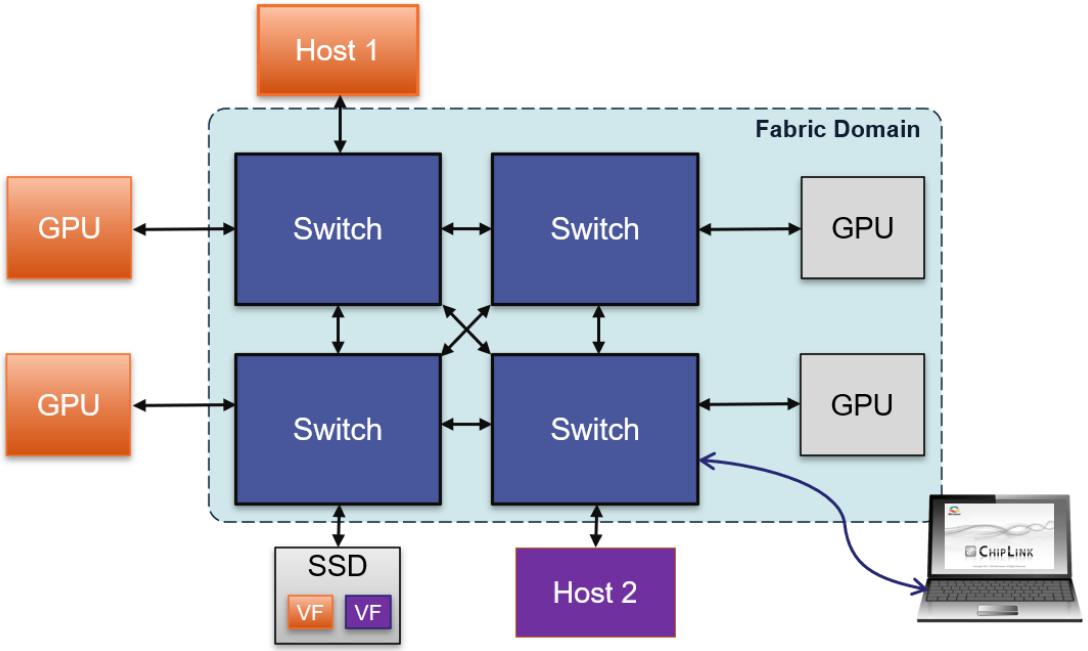
```
P2P Connectivity Matrix
     D\D     0     1     2     3
     0       1     1     1     1
     1       1     1     1     1
     2       1     1     1     1
     3       1     1     1     1
```

```
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
```

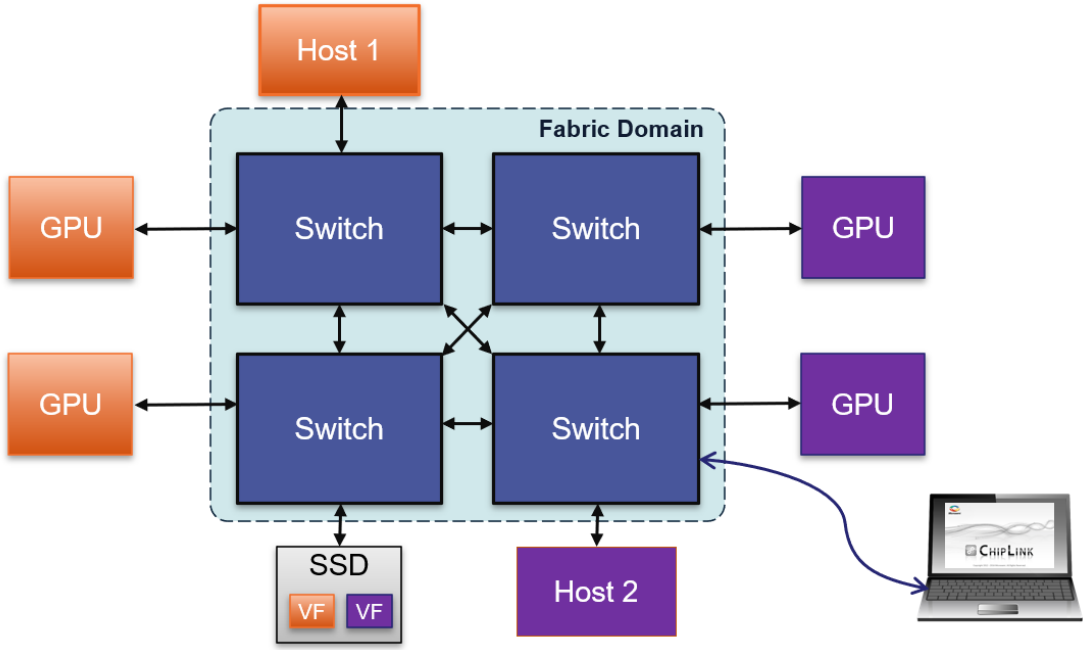
```
Bidirectional P2P=Enabled Bandwidth Matrix (GB/s)
   D\D     0      1      2      3
     0 213.51  24.81  24.77  24.72
     1  24.73 213.55  24.73  25.74
     2  24.53  24.57 214.73  24.86
     3  24.83  25.72  24.73 214.58
```


Dynamic Assignment of Pooled GPUs

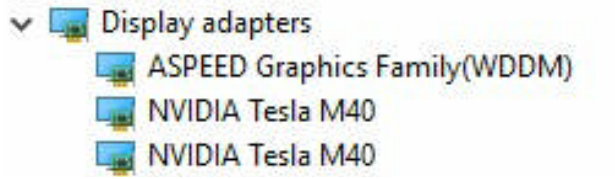
Host 1 workload completes, and GPUs are released back into fabric pool



Spare GPUs are assigned to Host 2

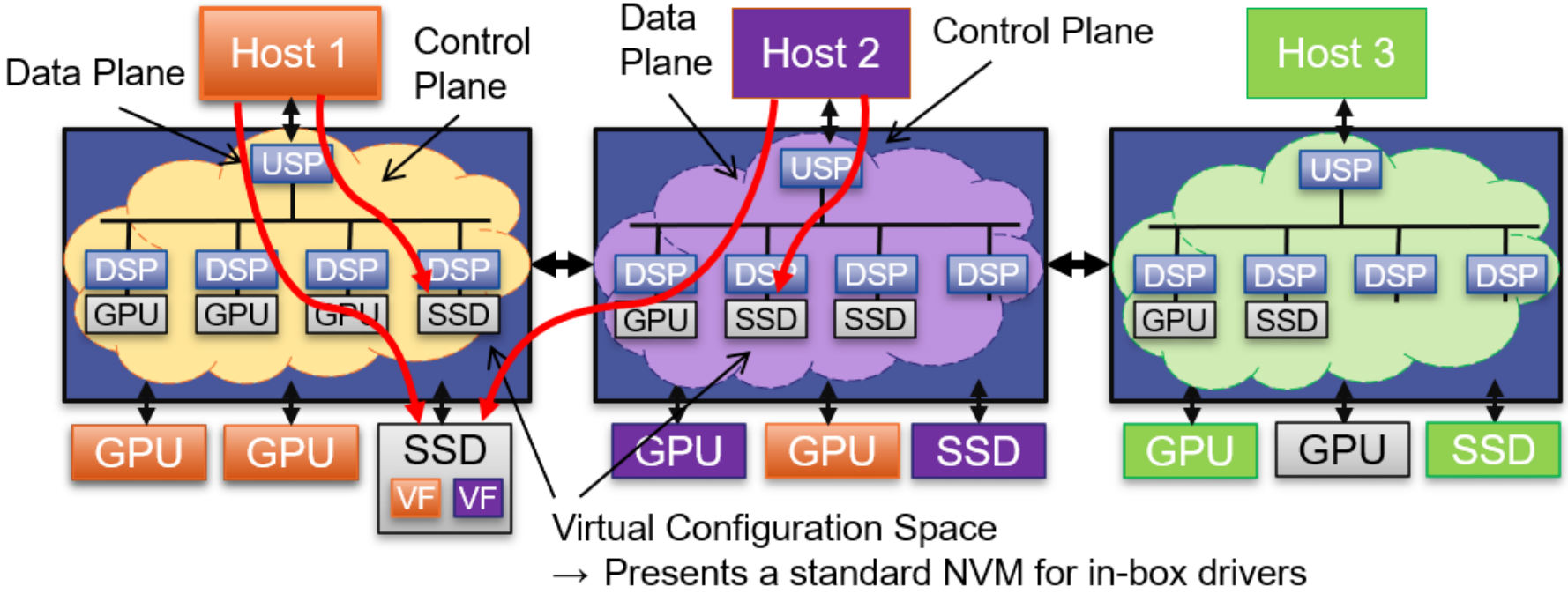


Windows Host Still Running During Dynamic Reassignment



Switch Fabrics - Disaggregation/Sharing

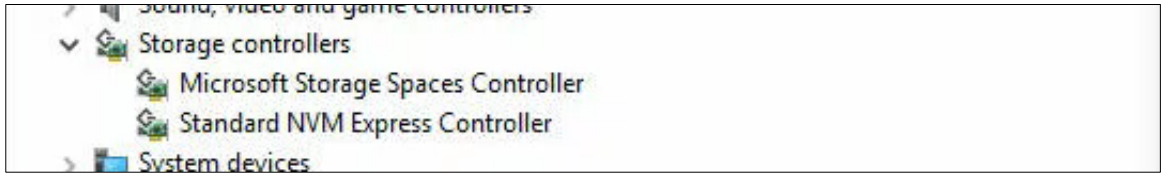
- SR-IOV: EP appears as multiple functions
- Fabric resources assigned by function to multiple hosts

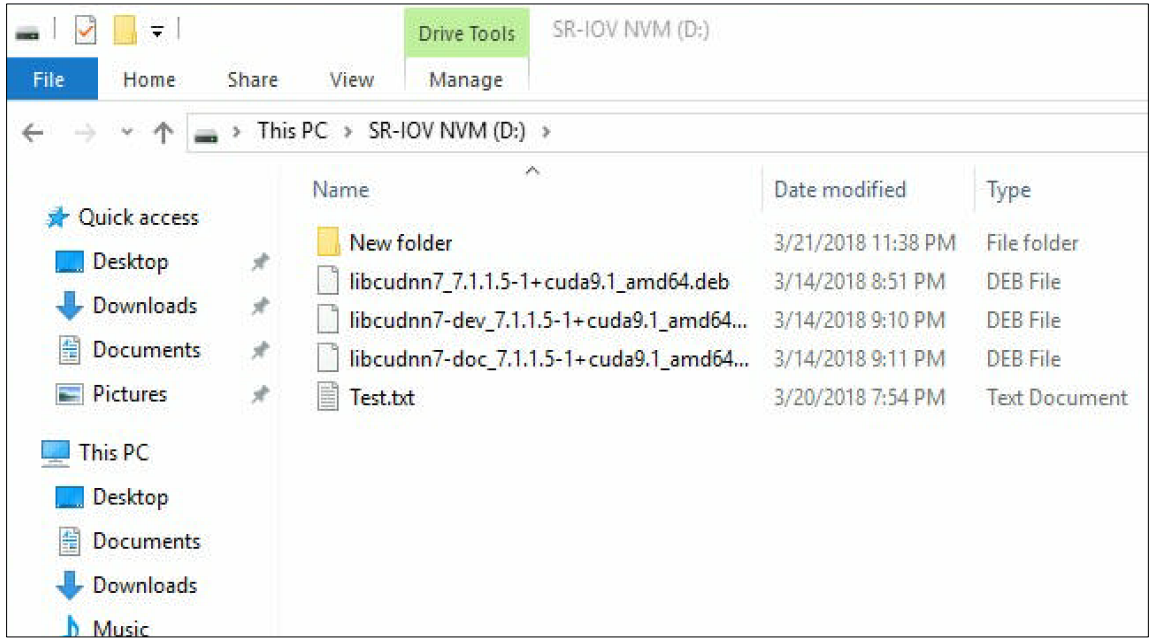


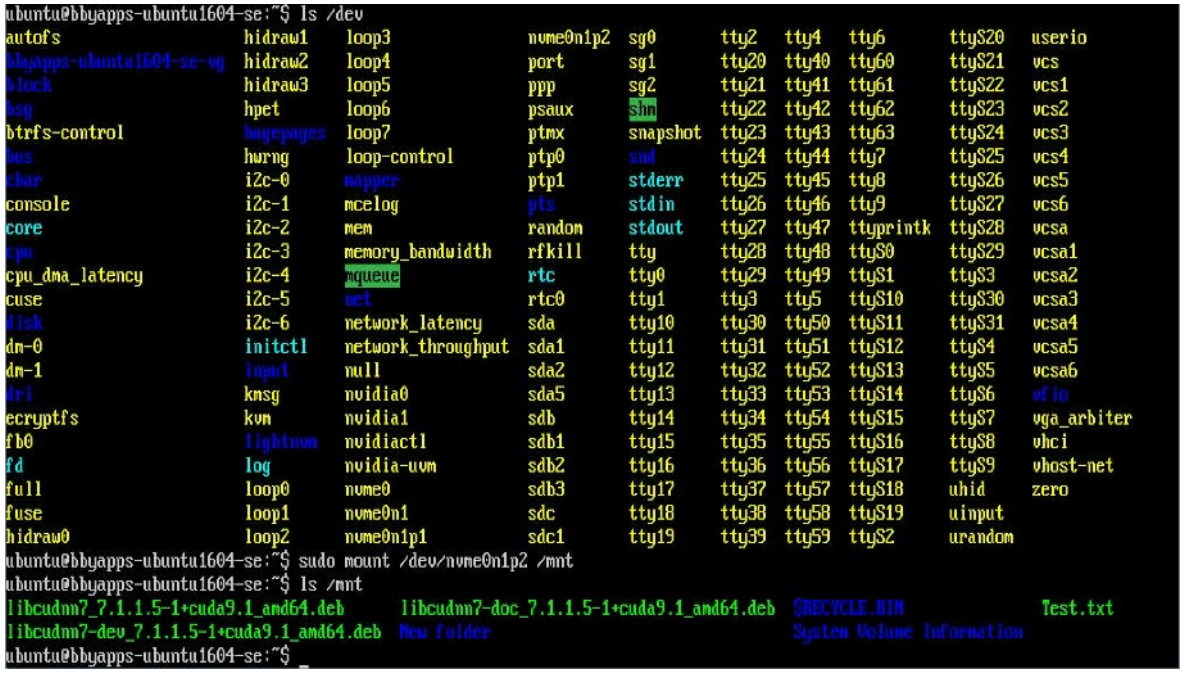
Multi-host Sharing of NVMe

Flash Memory Summit

NVM VF Appears as Standard NVM Device



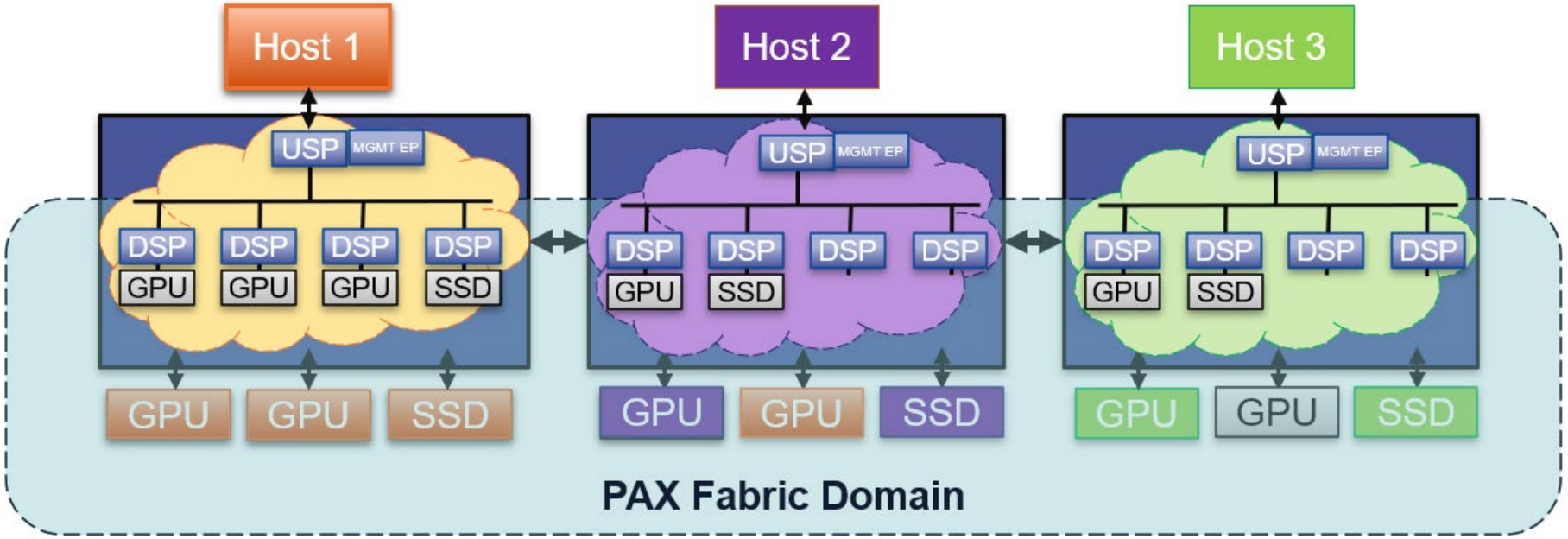




35 YEARS OF NAND FLASH MEMORY INNOVATION

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

