Flash Memory Summit

PCIe® Fabrics Advanced Solutions

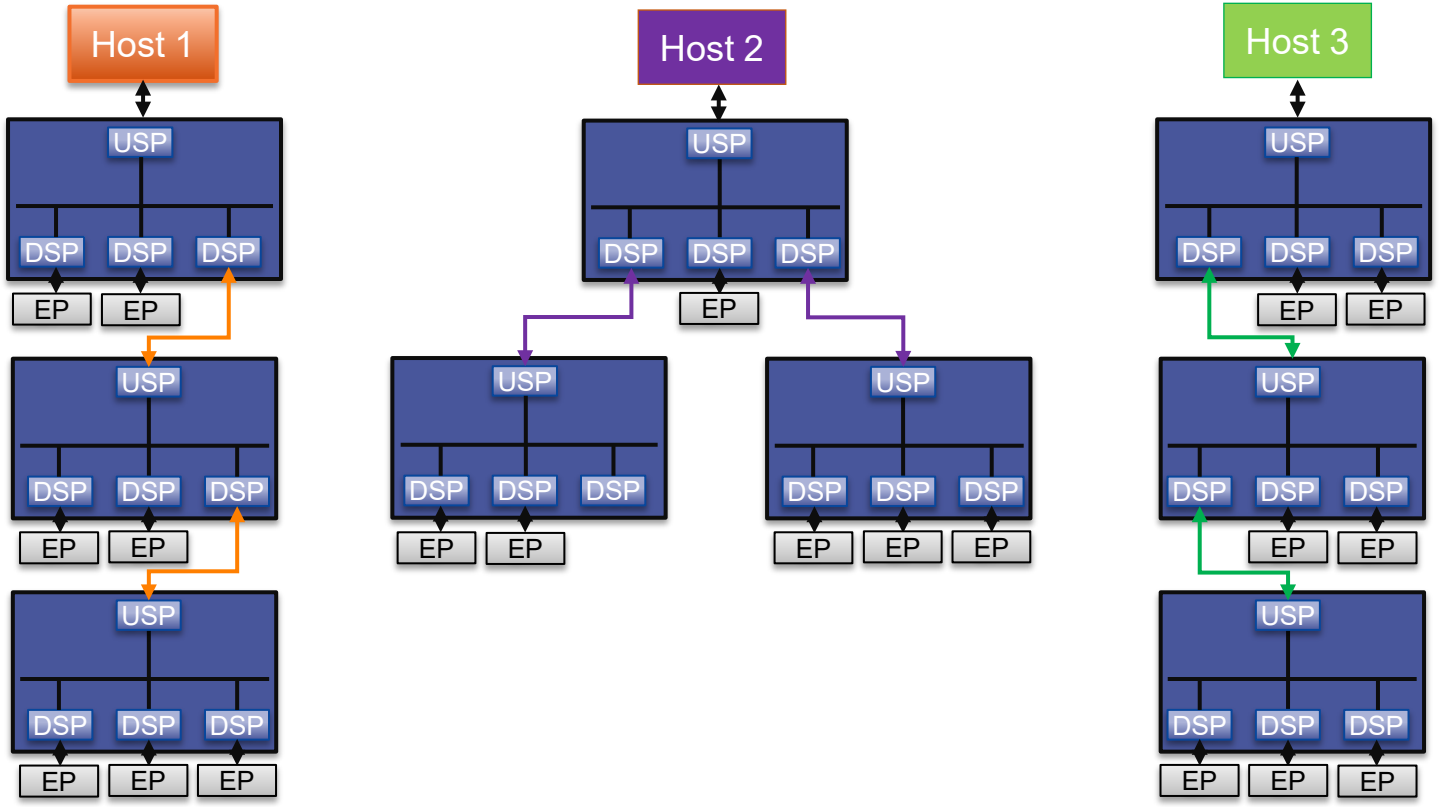
Presenter: Chetana Kaushik, Principal Applications Engineer

Microchip Technology Inc.

MICROCHIP

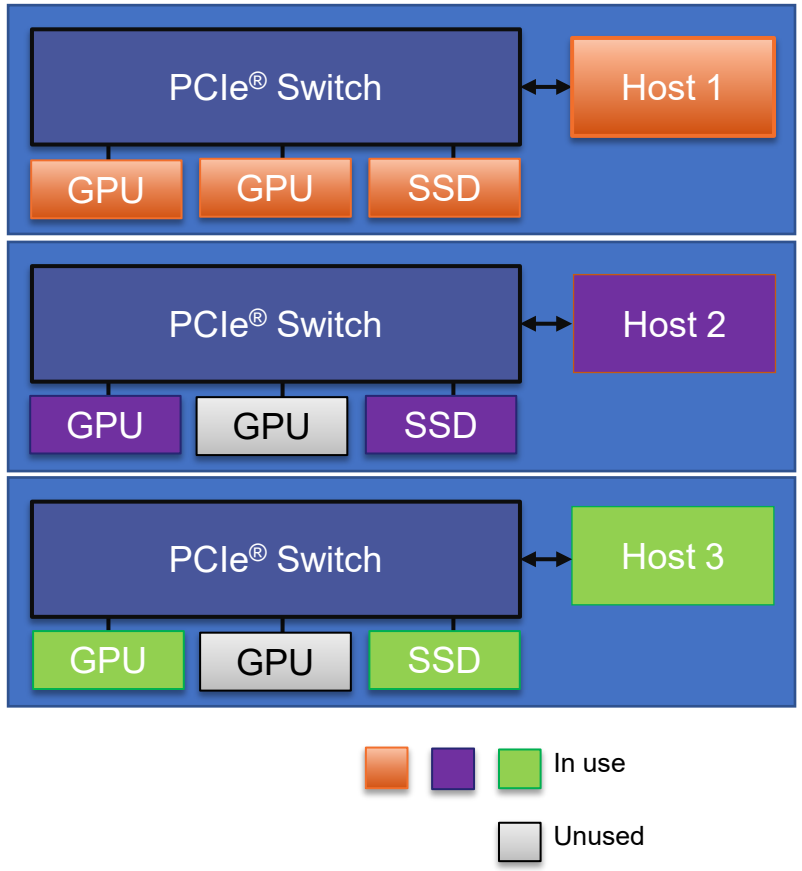
Standard PCIe® Hierarchy Restriction

- Standard PCIe hierarchy is restrictive, making scale out challenging



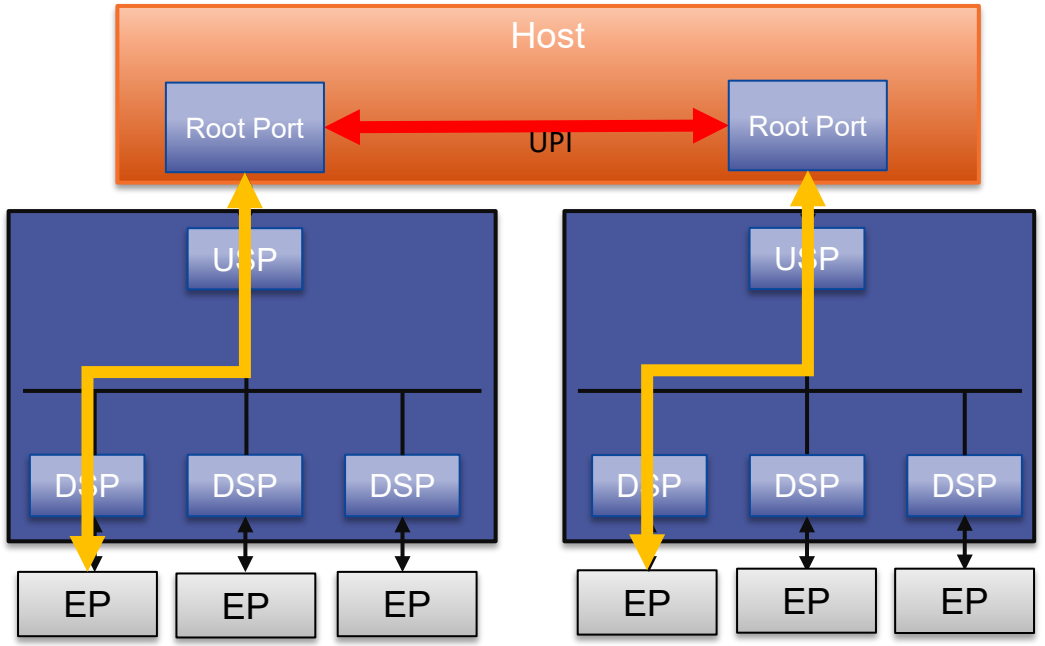
Standard PCIe® Single Domain Restriction

- 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
  - Underutilized EPs can't be shared



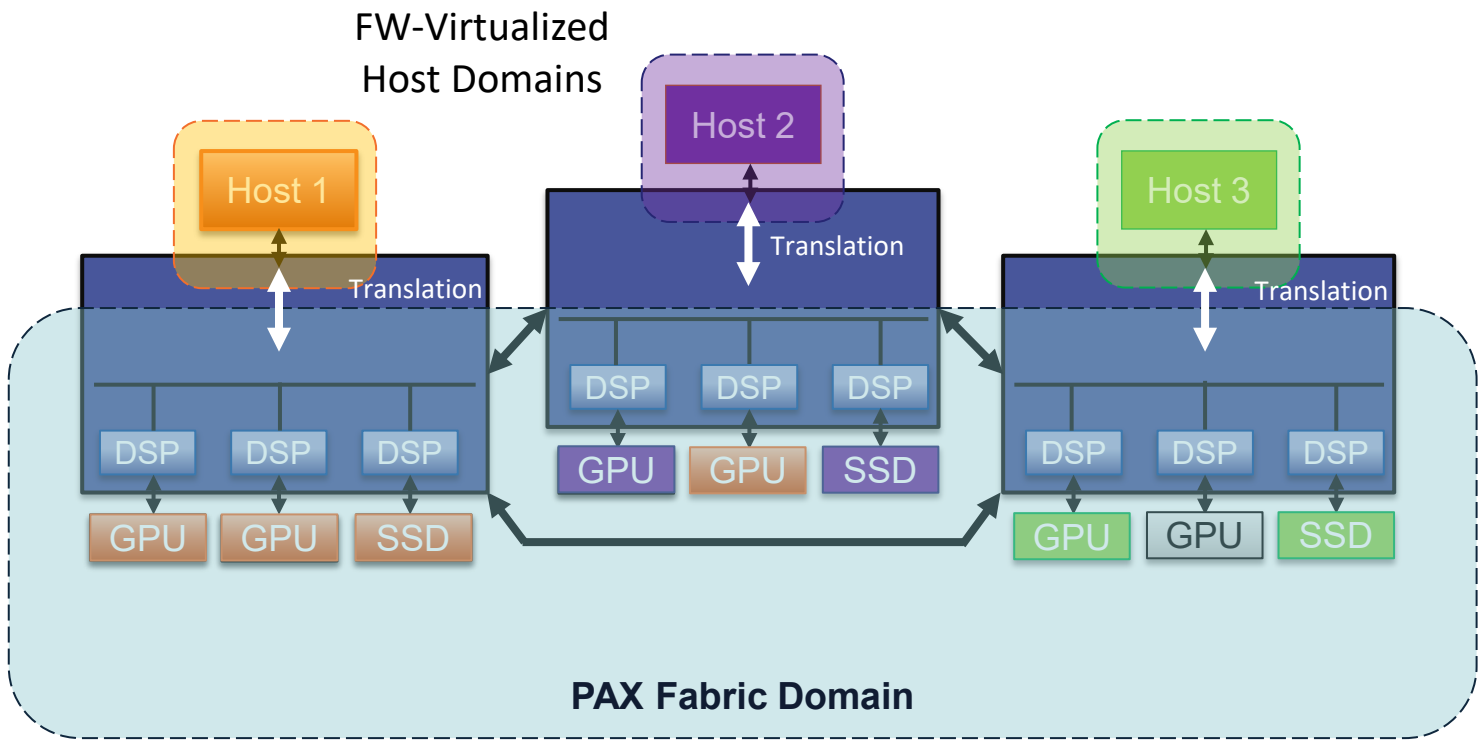
Standard PCIe® Routing Restrictions

- 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 multicore systems bottlenecked by UPI



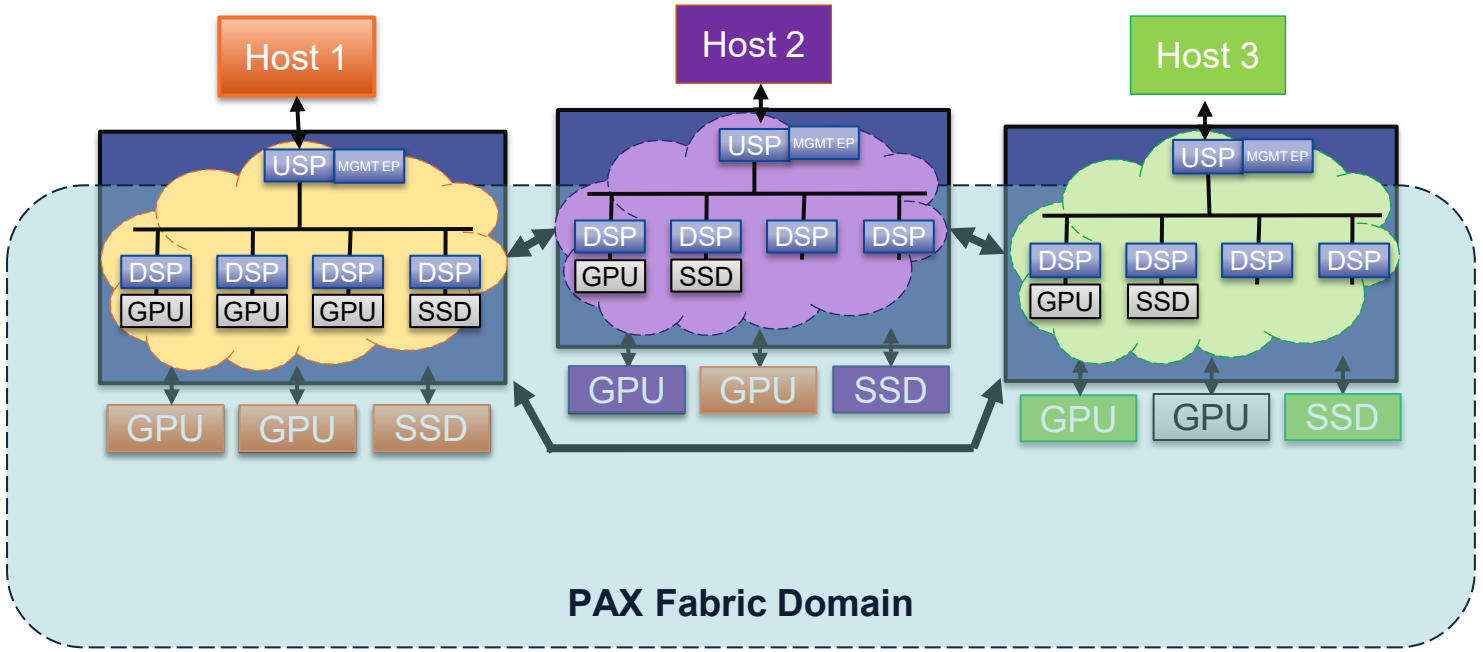
PCIe® Fabrics for Scaling

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



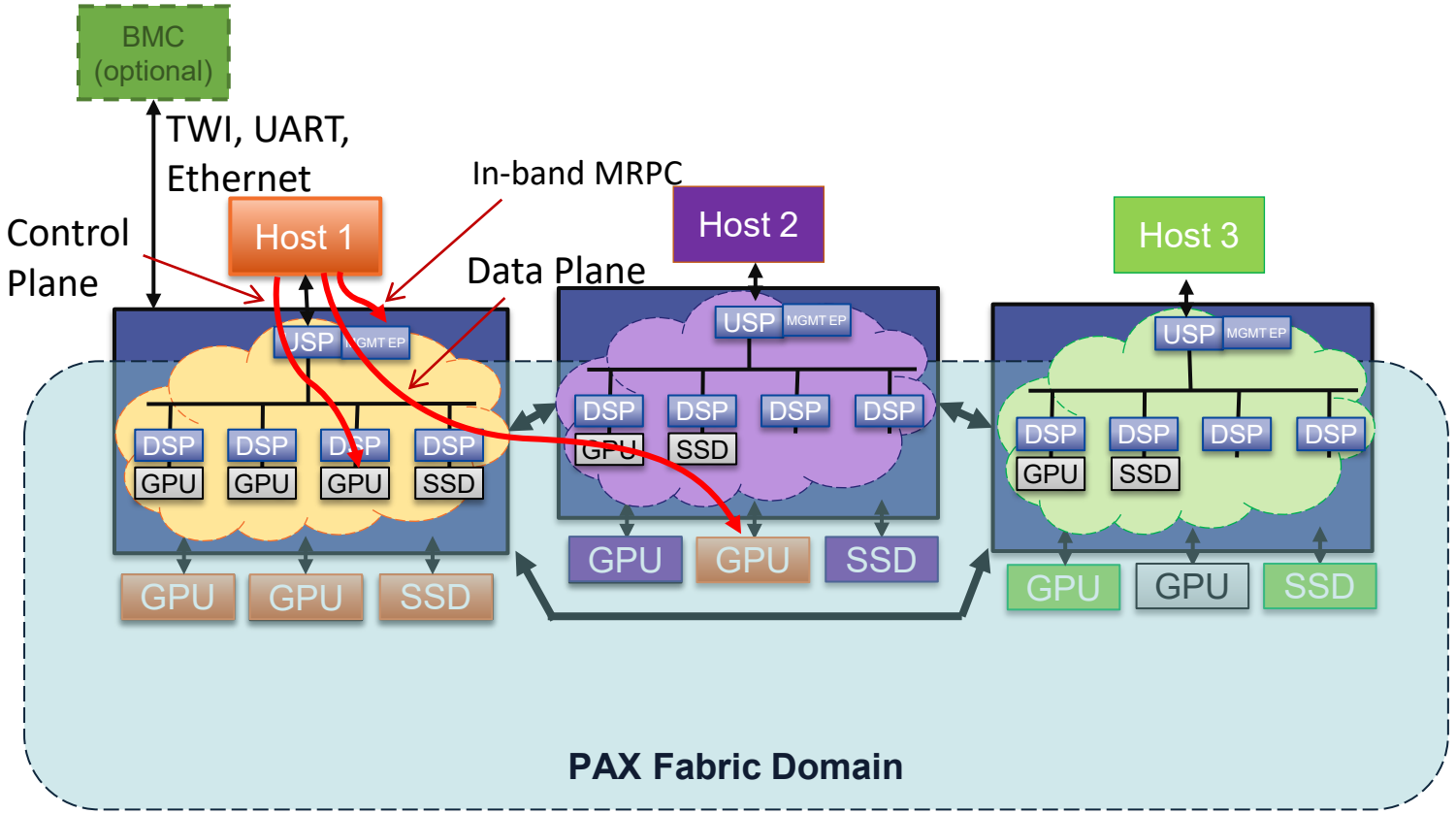
PCIe® Fabrics for Composability/Scaling

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



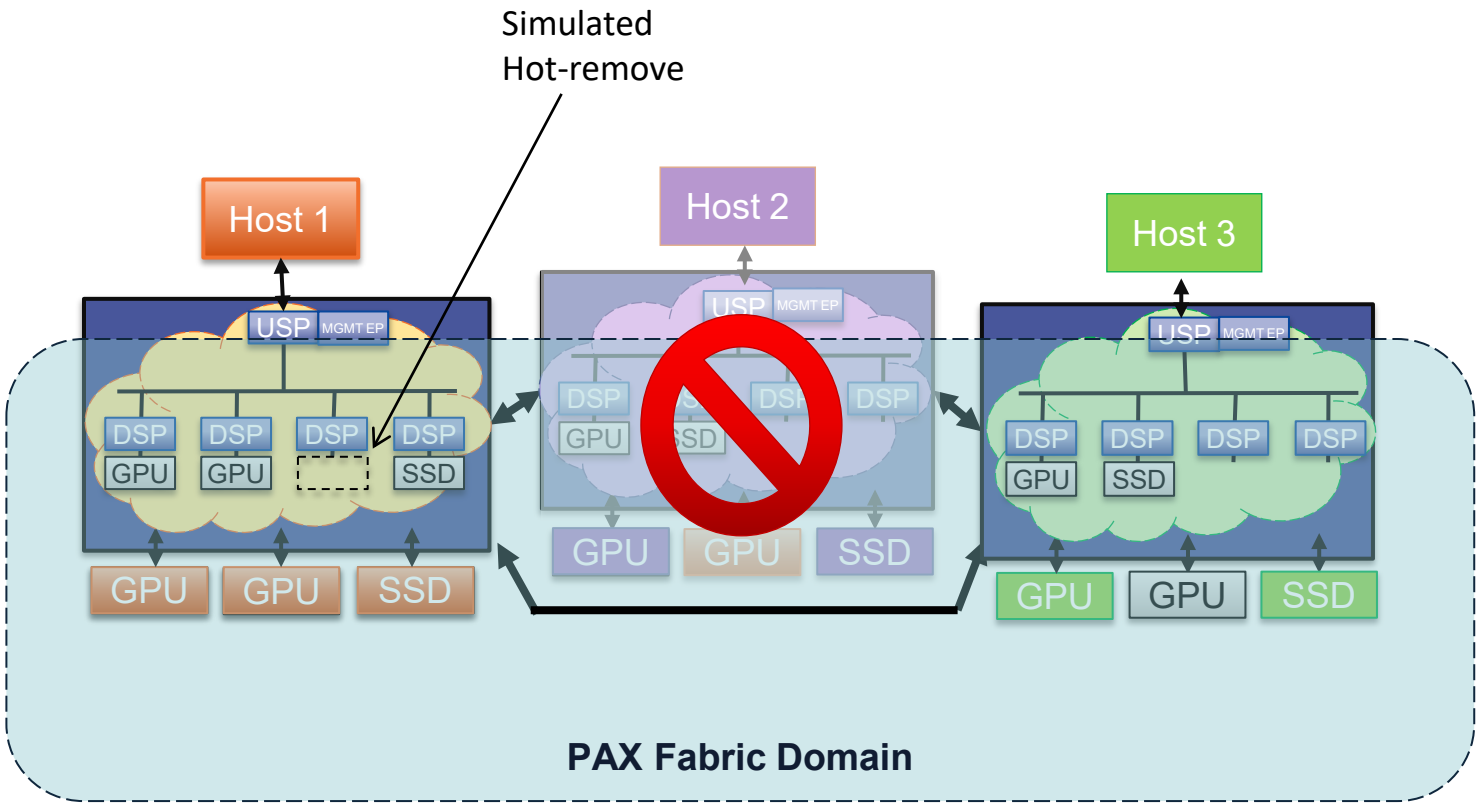
PCIe® Fabrics for Composability/Scaling

- Embedded CPU handles control plane
  - Data is routed directly by switch HW
- Fabric can be managed via PCIe, TWI, UART, Ethernet or FW SDK, MIPS
- Fabric is managed through simple Memory-Mapped Procedure Call (MRPC) interface (bindings, debug, etc.)
- Kernel driver and user-space management tool provided

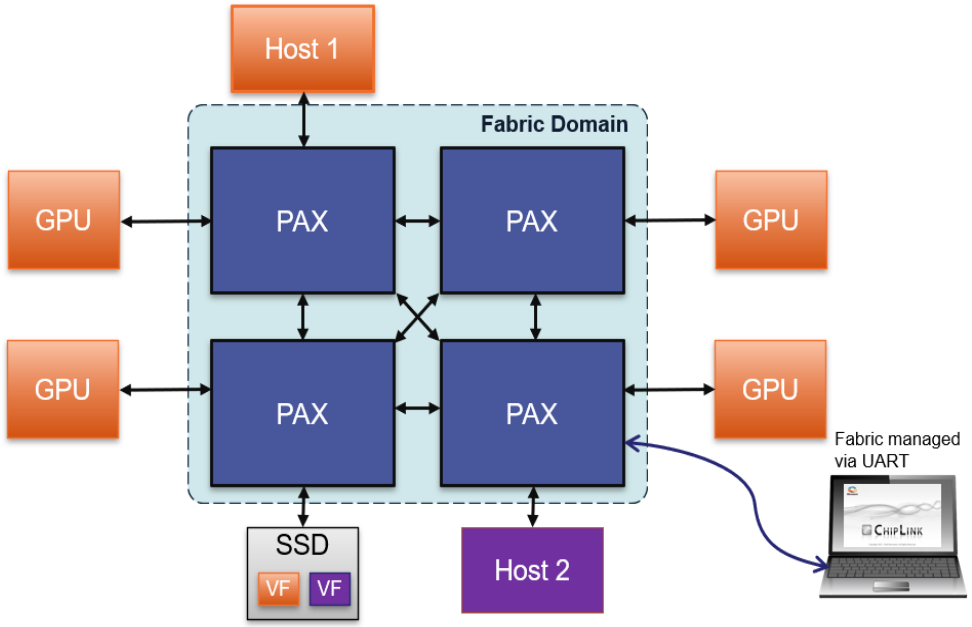


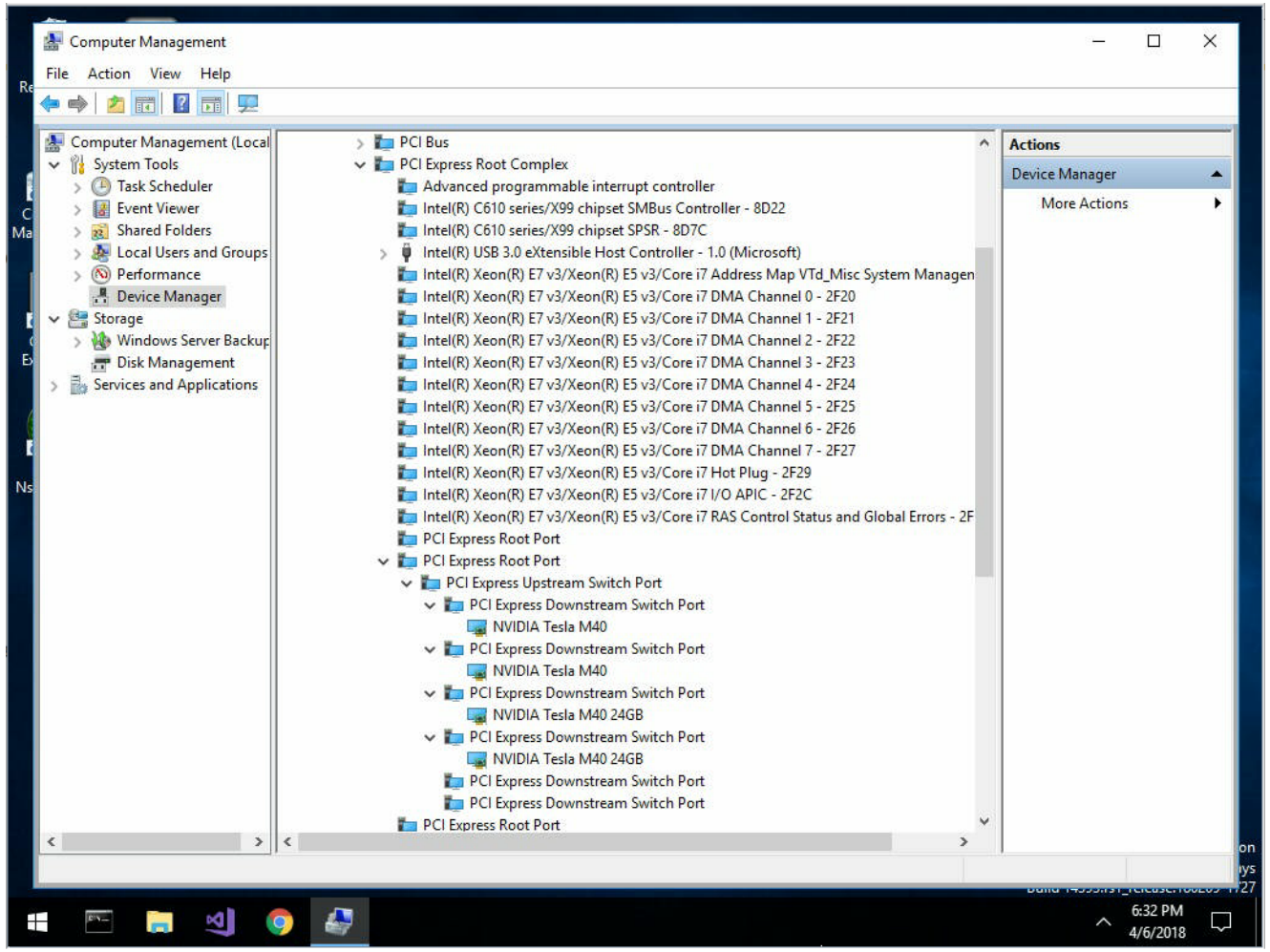
PCIe® Fabrics for Composability/Scaling

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



Dynamic Assignment of Pooled GPUs

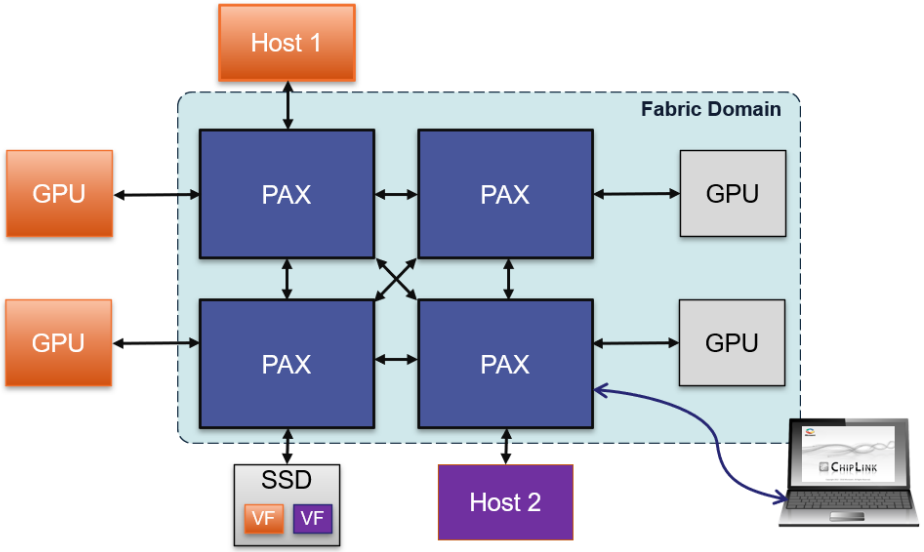


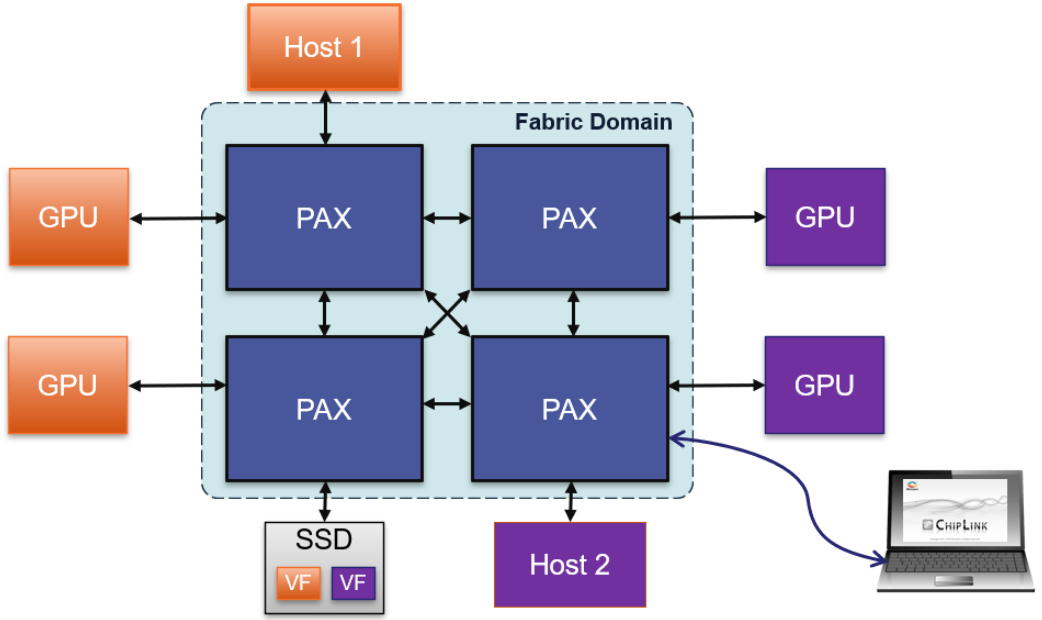


All GPUs are assigned to Host 1 to increase performance

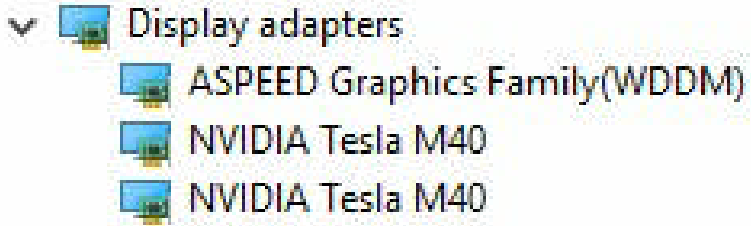
Dynamic Assignment of Pooled GPUs

- Host 1 workload completes, and GPUs are released back into fabric pool
- Spare GPUs are assigned to Host 2



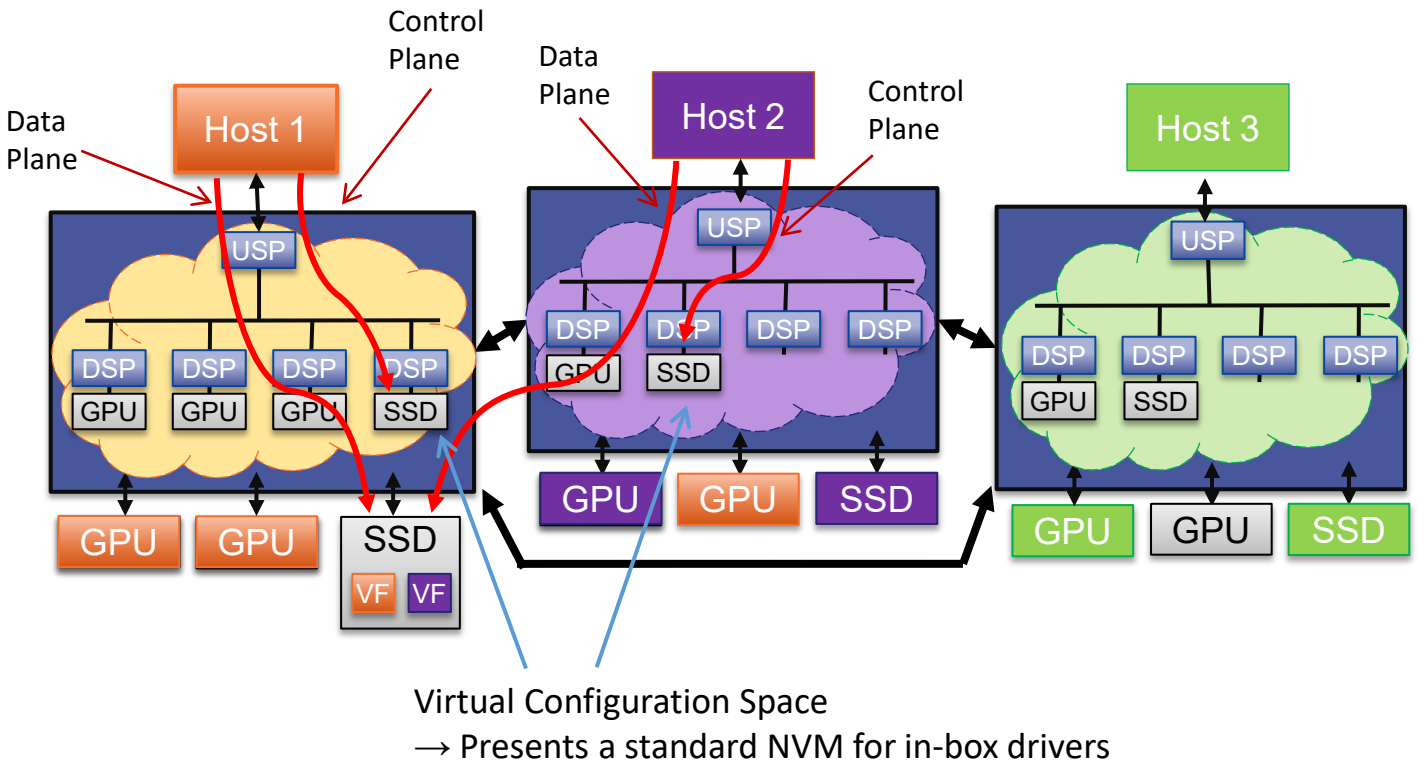


Windows Host Still Running During Dynamic Reassignment



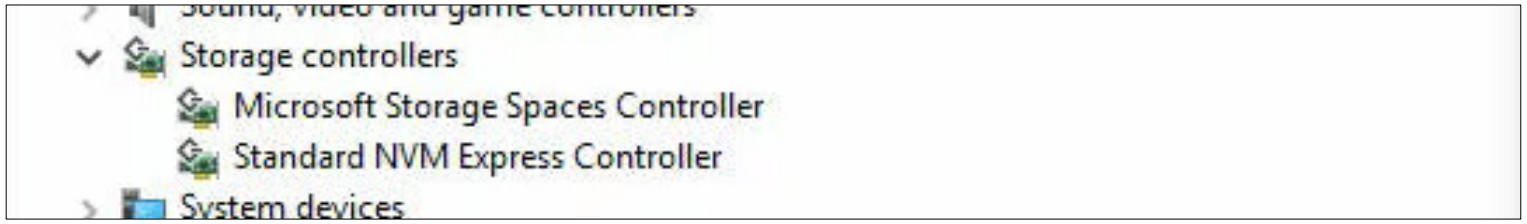
PCIe® Fabrics for Disaggregation/Sharing

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



Multi-Host Sharing of NVMe®

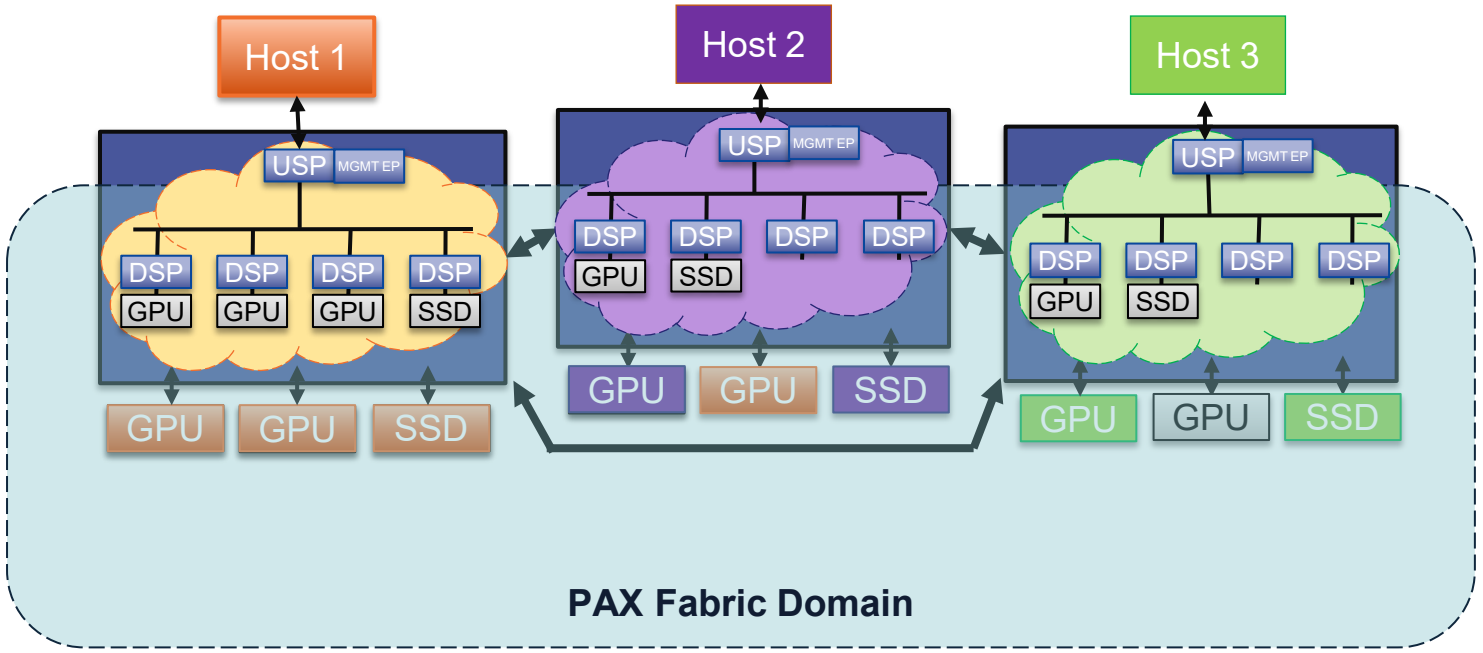
- NVM VF Appears as a Standard NVM Device




```
|        \-02.0-[51-64]--+-[0000:54]---00.0  Samsung Electronics Co Ltd
NVMe SSD Controller PM173X
|                  +-[0000:53]---00.0  Samsung Electronics Co Ltd NVMe SSD
Controller PM173X
```

PCIe® Fabrics for Advanced Solutions

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



Flash Memory Summit

Using PCIe® 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
  - Standard host drivers

MICROCHIP

Thank you