PCIe Switch Applications in AI System

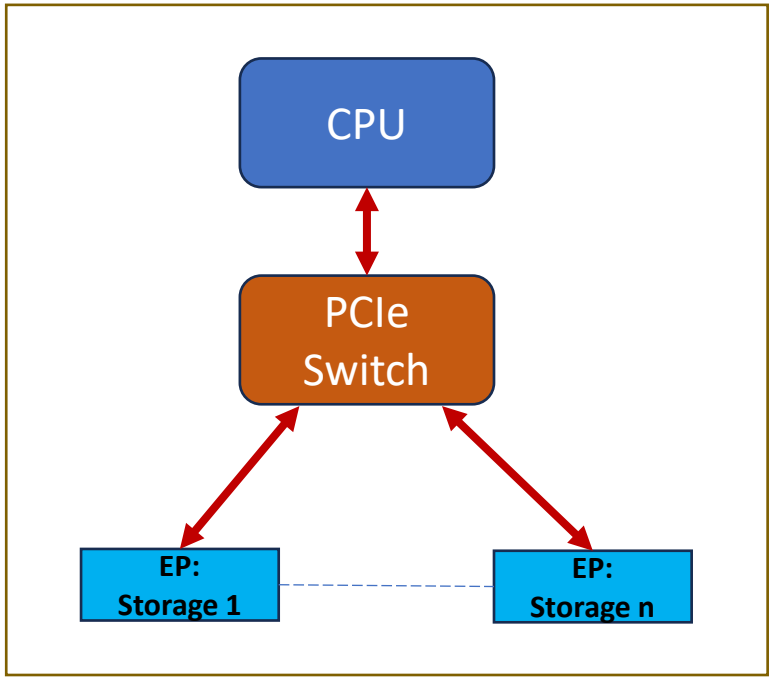
Tam Do
Microchip

Agenda

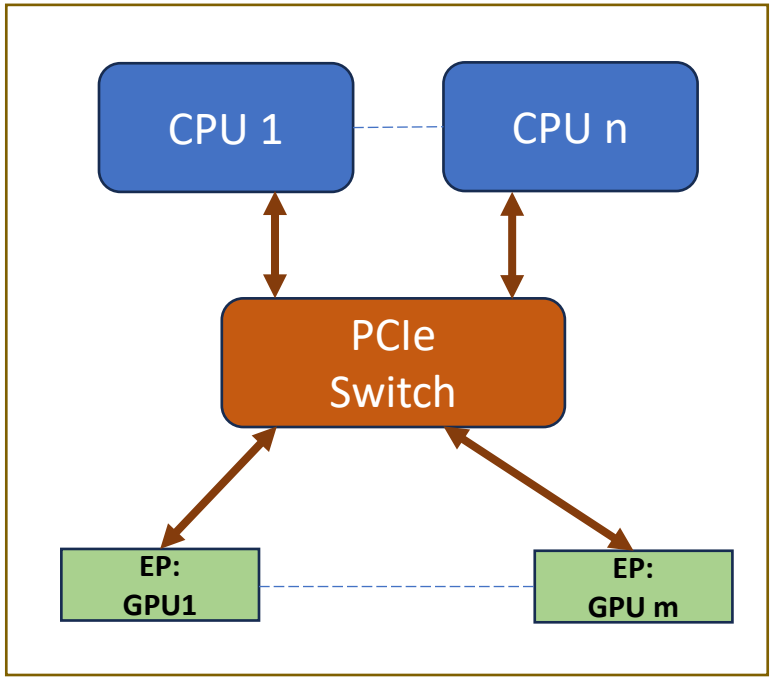
- PCIe Switch: from general fan out to AI application
- CPU, GPU and PCIe Switch
- CPU to GPU Ratio
- Application Example
- Heterogeneous Accelerators
- Low Cost Accelerators
- Summary

PCIe Switch Transition to AI Applications

- Traditional PCIe Switch Use Case:

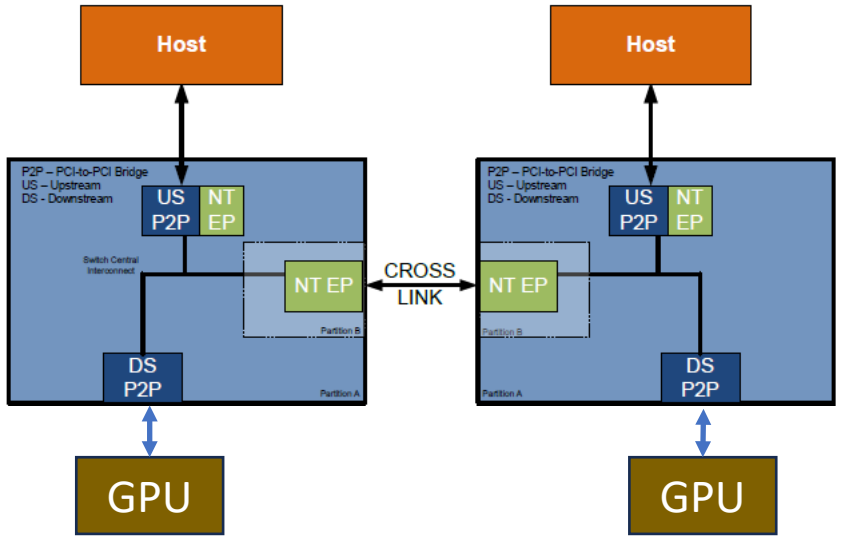


- PCIe Switch in AI Use case:



PCIe Switch Feature Set Review

- Number of PCIe lanes (from 28 to 160)
- Ports bifurcate (from x2 to x16)
- Non-transparent bridging (NTB) assignable to any port
- PCIe multicast
- Crosslink between multiple PCIe switch
- Dynamic port bifurcation and partitioning
- Error containment with hot plug controllers on all ports
- Virtual endpoints (synthetic, management, DMA & NVMe-MI
- Diagnostics, analysis and telemetry
- Power management (ASPM)



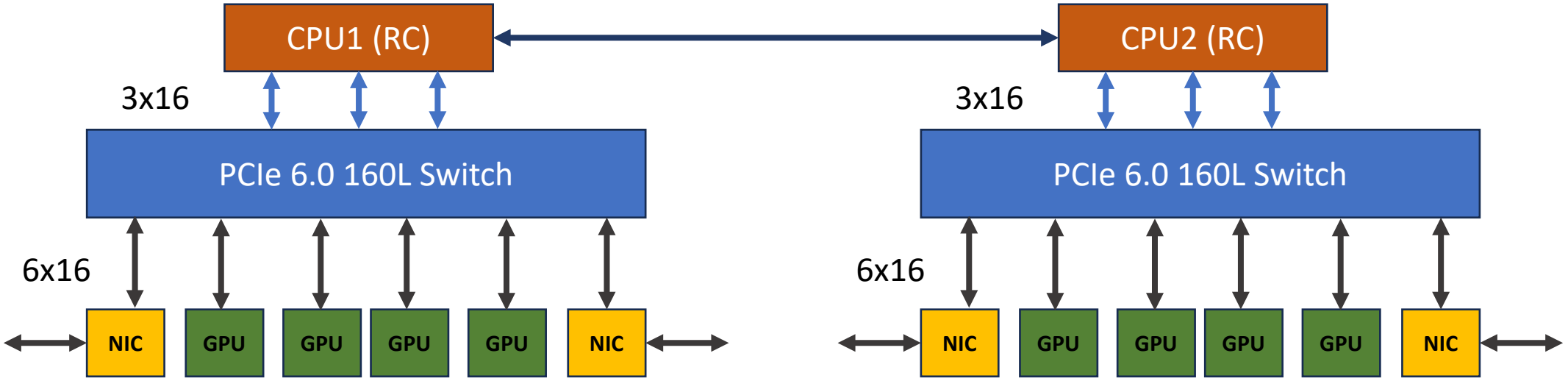
Ex: NT Crosslink connecting 2 host domains

CPU (cores) to GPU (EP) Ratio

- The ideal CPU cores to GPU ratio depends on the specific workload
- For most AI servers, the rule of thumb:
  - Training ratio: 1GPU: 16-32 CPU cores
  - Inference ratio : 1GPU: 4-8 CPU cores

Use case	Ratio (CPU cores per GPU)	Notes
Model Training (CNNs, NLP, Vision)	8–32 cores per GPU	Depends on data loading and GPU power
LLM Training (e.g. GPT, BERT)	16–32 cores per GPU	Tokenization and parallelism require more CPU
Inference (Batch)	4–8 cores per GPU	Balanced between performance and efficiency
Inference (Real-Time)	2–4 cores per GPU	Latency-sensitive, fewer cores needed
Data Preprocessing / ML Pipelines	CPU-heavy	GPUs optional or minimal
HPC / Simulation	8–64 cores per GPU	Depends on simulation and compute split

Higher PCIe Lane Count Switch for AI Application

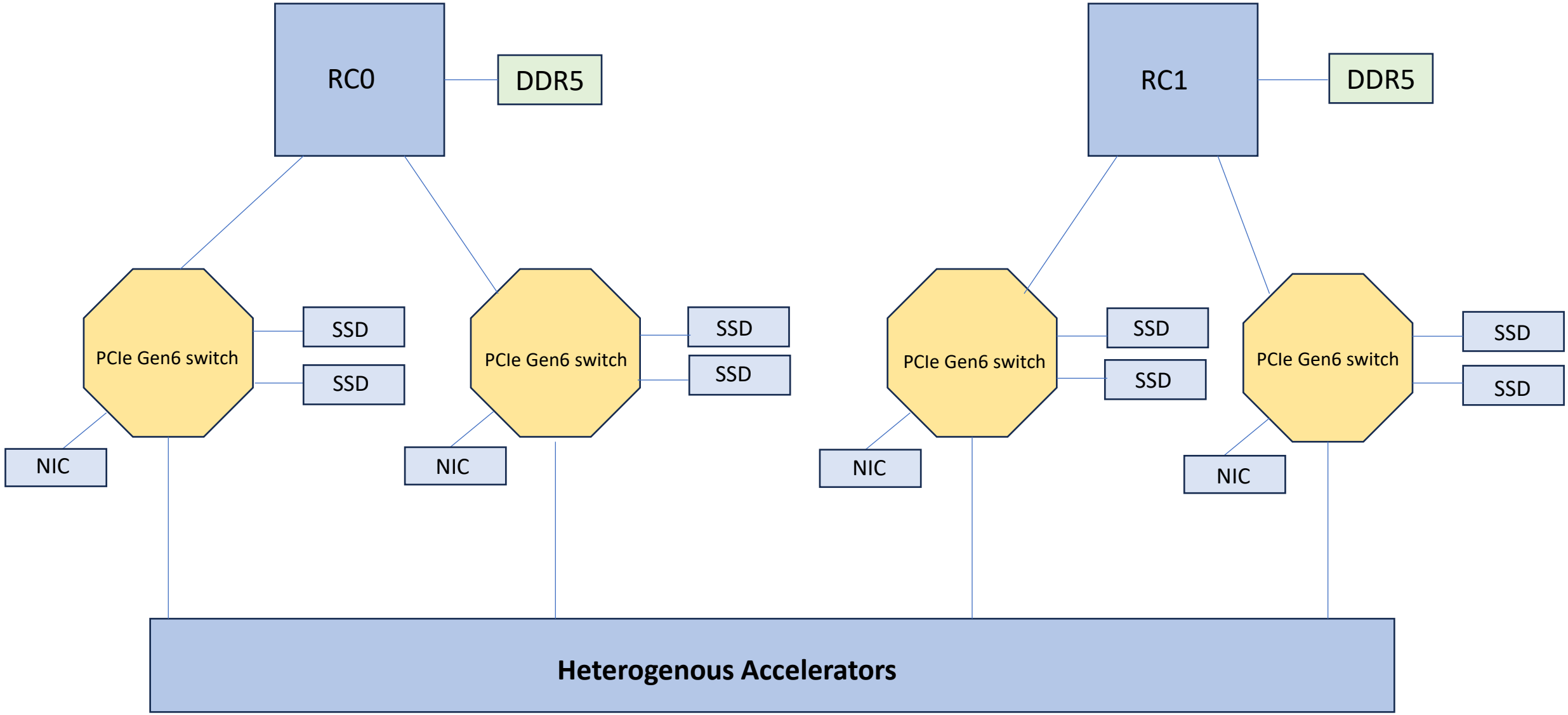


High lane count PCIe switches allows multiple GPUs to be connected with full PCIe bandwidth, avoiding bottlenecks. It also allow GPUs to communicate directly with each other, bypassing the CPU when possible.

Heterogenous Accelerators

- Combining various types of CPUs, GPUs, TPUs or FPGA to improve performance
- AI and HPC Platform Scalability and Performance
- Mixed Traffic and Multi-Device Connectivity
- Extending Reach and Signal Integrity in Complex Topologies
- Enabling Next-Generation AI and Cloud-Scale Deployments

Heterogenous Accelerators Example



Low-Cost Accelerators

- Building more open and cost-effective AI and HPC systems.
- Well suited for connecting flash storage and NICs to accelerators and CPUs in AI servers, enabling system builders to create cheaper and more open AI and HPC platforms.
- This flexibility helps reduce the overall system cost by consolidating connectivity and simplifying design
- PAM-4 and error correction features ensure reliable high-speed data transfers even in mixed-device environments, supporting a broad range of accelerator types without compromising performance or signal integrity

Summary

- AI/ML changes the traditional use cases for PCIe fan out switches
- High data rate and bandwidth require a higher lane count
- CPU to GPU ratio play a key role on how the PCIe switch is utilized
- Accelerators either Heterogeneous or Low cost play an important role in the data center AI application
- PCIe switch with high lane count will continue to play a key role in AI