The Impact of the Power Architecture on Summit

Christopher Zimmer PhD
HPC Systems Engineer
Transitioning from Titan to Summit

• **Titan**
  – 27 Petaflop
  – AMD Interlagos (16 Cores)
  – 32 GB DDR3
  – Tesla K20X

• **Summit**
  – 200 Petaflop
  – 2 IBM Power 9 (44 cores)
  – 512 GB DDR4
  – 6 Volta V100
Oak Ridge Leadership Computing Facility

- Open Science: Government / Academia / Industry
- Access is proposal based Incite / ALCC
- Capability Computing: The scientific applications requiring the largest systems
Titan Node Layout

- DRAM 32 GB
- x86
- GDDR 6 GB
- GPU 1.4 TF

Bandwidth:
- PCIe Gen2
- GDDR/DRAM Bus (aggregate B/W)
- 55 GB/s
- 5 GB/s
- 250 GB/s
Power Provides More Bandwidth

• 55 to 270 GB/s DRAM
• 5 to 32 GB/s I/O to HCA
• 600 GB/s of NVLINK
Power 9 Memory Bandwidth

• 8 Channels of Memory Bandwidth from Each Power 9
  – 135 GB/s per P9 (Expected STREAM Performance)
  – Peak 170 GB/s

• 64 GB/s X-Bus Bandwidth
  – Reduces NUMA impacts

• User Visible Results
  – On Titan main memory accesses were so costly most applications only ran applications out of GPU memory. Summit reduces this pain.
PCle Gen 4

• Not supported in other server class architectures yet.

• User Visible Results
  – Mellanox HCA Gen 4 X16
    • 2 - X8 Bifurcated Slot
    • 16 GB/s of bandwidth from either Power 9
  – Samsung PM1725A X8 Slot
    • 6 GB/s Read
    • 2.2 GB/s Write
Power Provides CAPI 2

- Capi 2 Protocol over PCIe Gen 4
- Reduces latency between host and HCA
- Adds address translation services
  - Enables CUDA Managed Memory for use in On Demand Paging
    - CUDA UVM buffers access from HCA
    - Uses host memory management unit for translations

User Visible Results:
- Reduced costs for inter-processor communication initiated from GPUs
  - (GPU must still ring host doorbell to initiate transfer)
Power Provides NVLink2 to Power 9

- NVLink2 between Host CPU and NV GPUs
- 150 GB/s to GPUs from each Power 9
  - On Titan for a GPU: Host Memory Access was 50:1
  - On Summit with NVLink2 GPU: Host Memory Access is 20:1

- User Visible Results
  - CUDA Managed Memory
    - A pointer allocated on the host, referenced on the GPU will not crash application
Conclusion

• It’s a commonly held belief that bandwidth gates HPC application performance
  – Memory Bandwidth
  – Interconnect Bandwidth
  – I/O Bandwidth

• The Power Architecture is providing Summit with more bandwidth:
  – 270 GB/s to DRAM
  – 600 GB/s of Aggregate NVLINK / 300 GB/s from Hosts to GPUs
  – 32 GB/s to Interconnect (25 Usable)