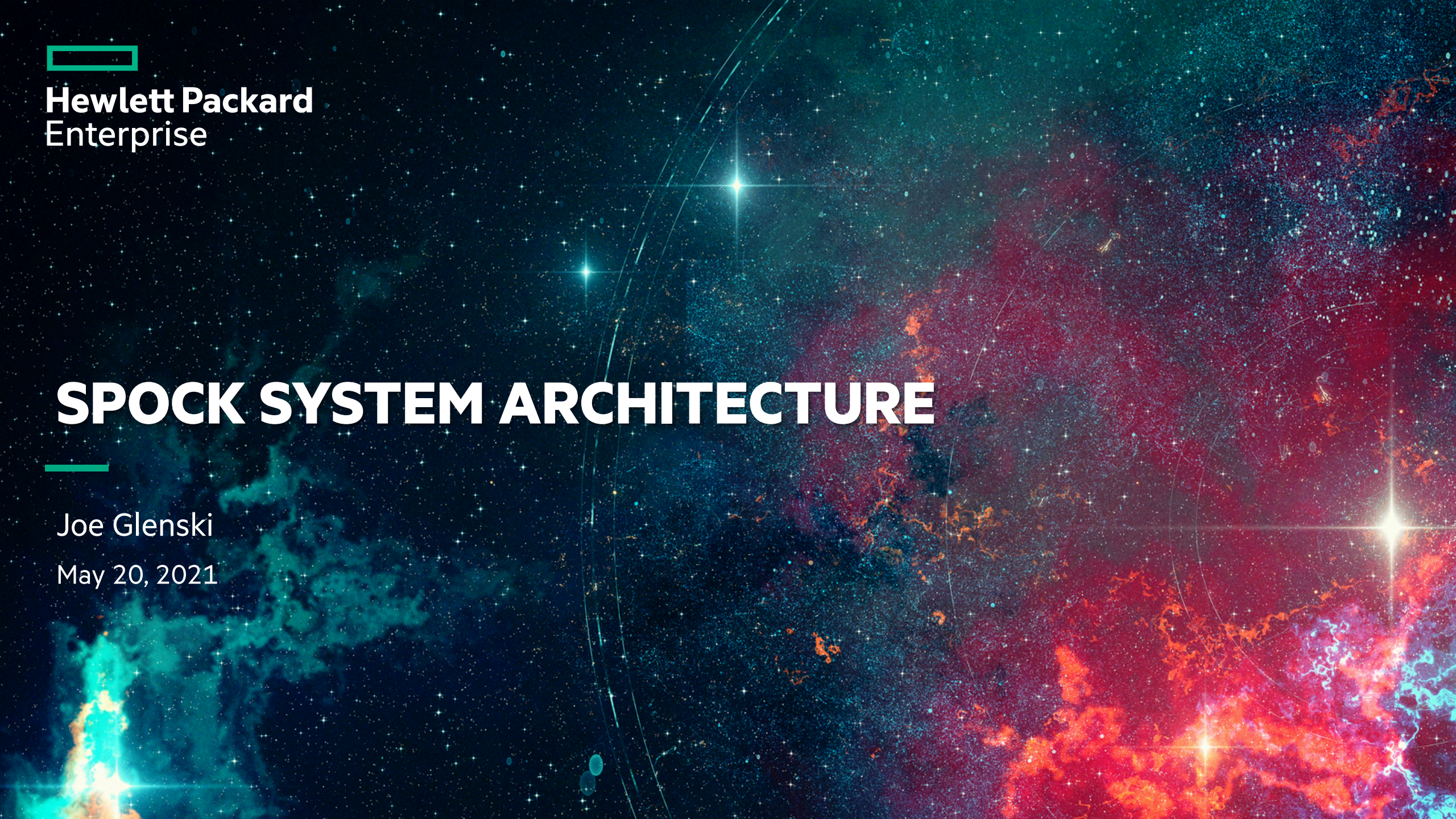


SPOCK SYSTEM ARCHITECTURE

Joe Glenski

May 20, 2021

TOPICS

- Spock System Overview
- Node Design
- Slingshot Interconnect
- User Access Nodes
- Storage
- Application Software Stack

SPOCK SYSTEM OVERVIEW

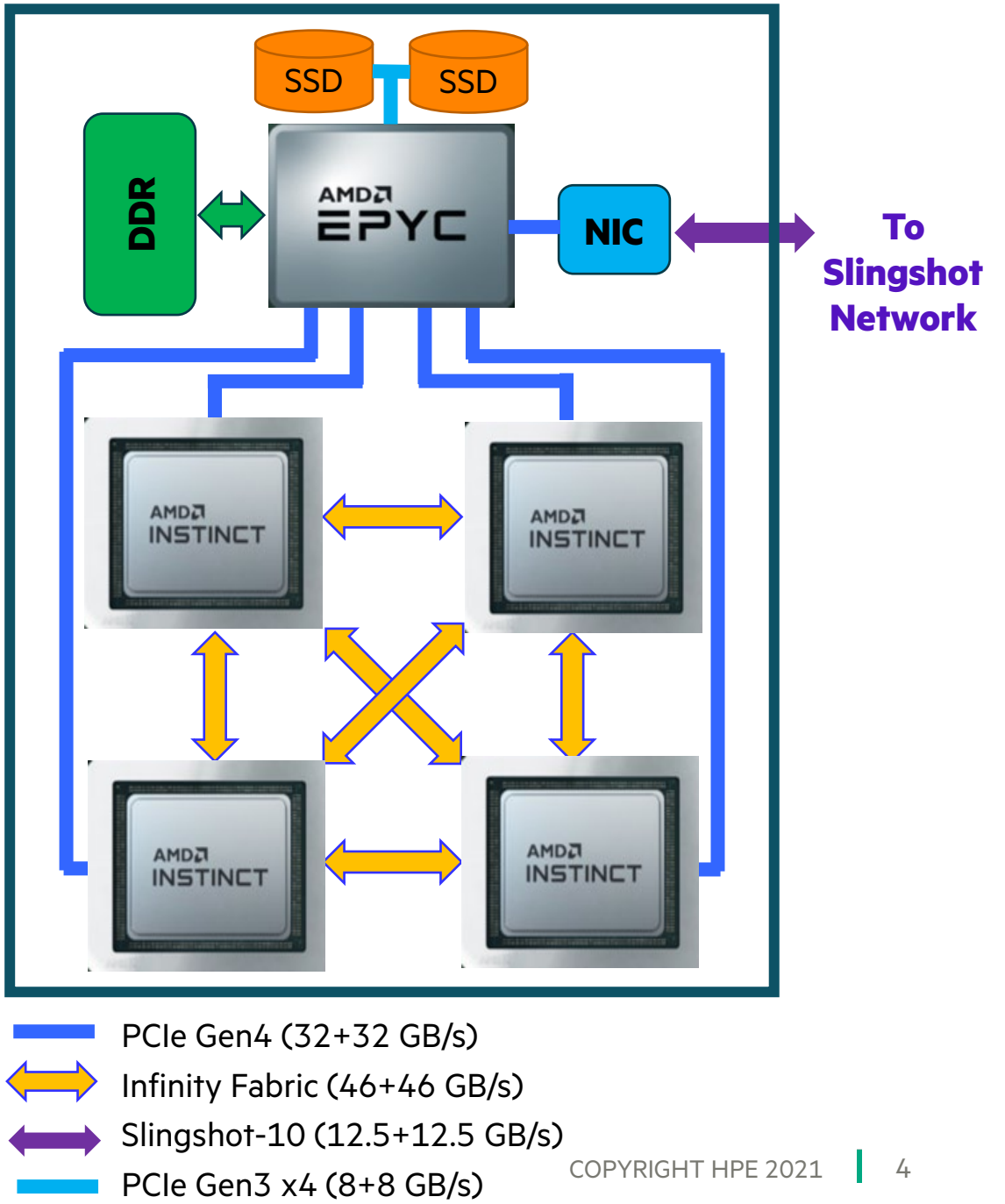
HPE Cray Supercomputer architecture

- Early-access testbed for Frontier
- 3 Racks holding 36 compute nodes
- 1 Rack holding support and management nodes
- HPE Slingshot interconnect - “SS10”
- AMD EPYC CPUs - “Rome”
- AMD Instinct GPUs - “MI100”
- HPE Cray Software stack
  - Essentially the same software stack that will run on Frontier, just earlier versions
- OLCF Provided Storage

Spock Rack elevation

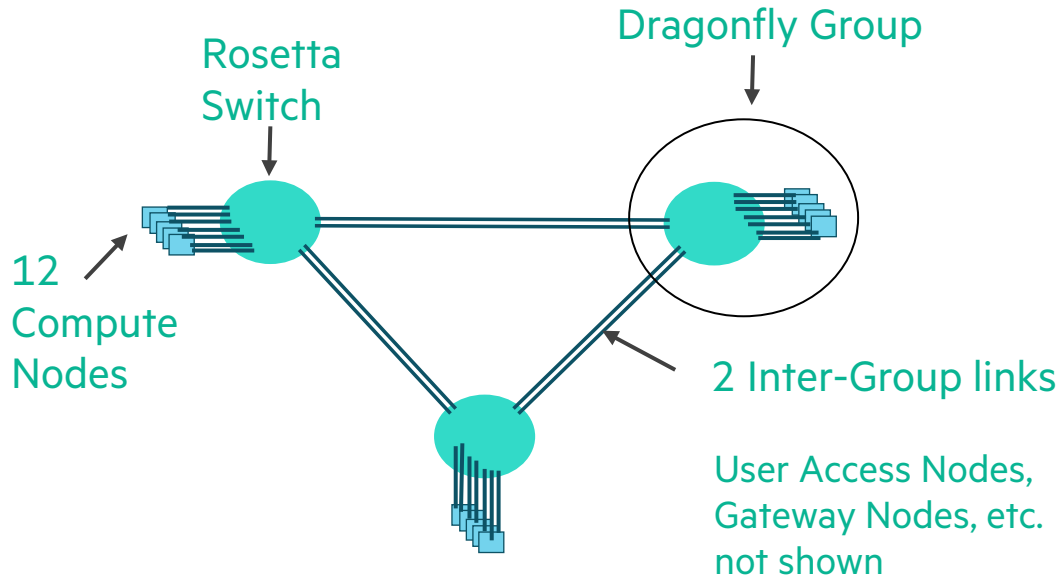
SPOCK COMPUTE NODE DESIGN

- 1x AMD EPYC 7662 “Rome” 64 core processor
  - 2 hardware threads per physical core, base clock 2.0 GHz
- 256 GB DDR4 memory with 205 GB/s peak bandwidth
- 2x NVMe 3TB SSDs
- 4x AMD “MI100” Instinct GPUs
  - 32 GB High-Bandwidth Memory (HBM)
  - 1.2 TB/s peak bandwidth
  - 11.5 TFLOPS double-precision peak for modeling & simulation
  - 184.6 TFLOPS in half-precision peak for machine learning and data analytics.
- PCIe Gen4 connections between CPU and GPUS
  - Peak host-to-device (H2D) and device-to-host (D2H) data transfers of 32+32 GB/s
- AMD Infinity Fabric between GPUS
  - Peak device-to-device bandwidth of 46+46 GB/s, low latency
- 1x HPE Slingshot-10 interconnect port
  - Provides 12.5+12.5 GB/s to other nodes



SPOCK SLINGSHOT INTERCONNECT

- High speed, low latency network architecture
- Uses proven Dragonfly topology
- Single port Node Injection – 1 link from node to switch
  - Bi-directional bandwidth of 12.5 GB/s
- “Class 1” topology with 3 HPE Rosetta switches
  - High radix, 64-port, 12.8 Tb/s bandwidth switch
- 3 Groups, each with 12 compute nodes in the group
- All to All connections between groups
  - 2x links to each other group
  - Bi-directional bandwidth of 25 GB/s per link
- Advanced flow control features designed to explicitly address congestion and bottlenecks
  - Adaptive Routing, Quality of Service, Congestion Control
  - Ensure consistent, predictable, reliable performance



J. Kim, W. J. Dally, S. Scott, and D. Abts. Technology-driven, highly-scalable dragonfly topology. ACM SIGARCH, 2008.

Kim, J., Dally, W., Scott, S., Abts, D.: Cost-Efficient Dragonfly Topology for Large-Scale Systems. IEEE Micro. 29(1), 33–40 (2009)

D. De Sensi, S. Di Girolamo, K. H. McMahon, D. Roweth and T. Hoefler, An In-Depth Analysis of the Slingshot Interconnect, SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, 2020, pp. 1-14,

SPOCK USER ACCESS NODES

- Spock has 2 User Access Nodes
- These are for user compiles, job launches, etc.
- Each user access node contains:
  - 2x AMD EPYC “Rome” 64 core processor
  - 512 GB DDR4 memory (256 per CPU)
  - 1x NVMe 2TB SSD
  - 1x AMD “MI100” Instinct GPU
  - 1x HPE Slingshot-10 interconnect port
  - 2x 10 GbE Ethernet NICs for user access
  - 2x 480 GB SSDs
- The processors are the same as on the compute node, but the internal node architecture is different

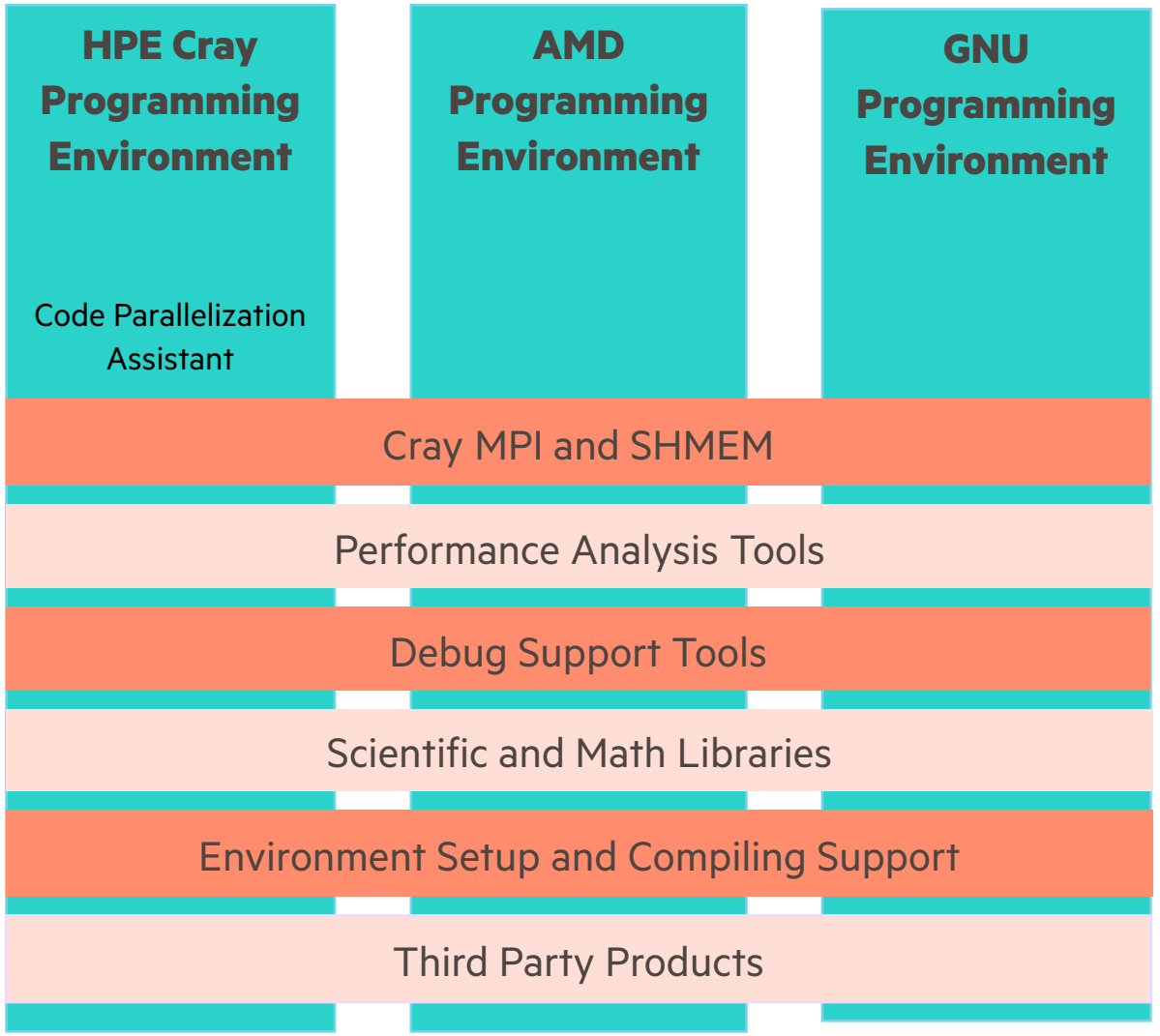
STORAGE FOR SPOCK

- Spock is connected to the “Alpine” IBM Spectrum Scale™ parallel filesystem
  - Provides 250 PB of storage capacity (/gpfs/alpine/...)
  - Peak write speed of the filesystem is ~2.5 TB/s
  - Usable bandwidth to Spock will be ~20 GB/s
- Spock also has access to the center-wide NFS-based filesystem
  - Provides user (/ccs/home/...) and project (/ccs/project/...) areas

APPLICATION SOFTWARE STACK FOR SPOCK

- ORNL, LLNL, HPE, and AMD are working together to deliver a full software stack targeted at Frontier
  - Will provide compiler and library choice, performance, and programmability
  - Includes:
    - Multiple programming environments
    - Performance and correctness tools
    - Optimizations such as:
      - MPI GPU-to-GPU data movement
      - libsci_acc
      - DL Plugin
    - Compiler interoperability
  - This software is a work in progress

- Spock will get updated versions of the software as they become available



THANK YOU

glenski@hpe.com