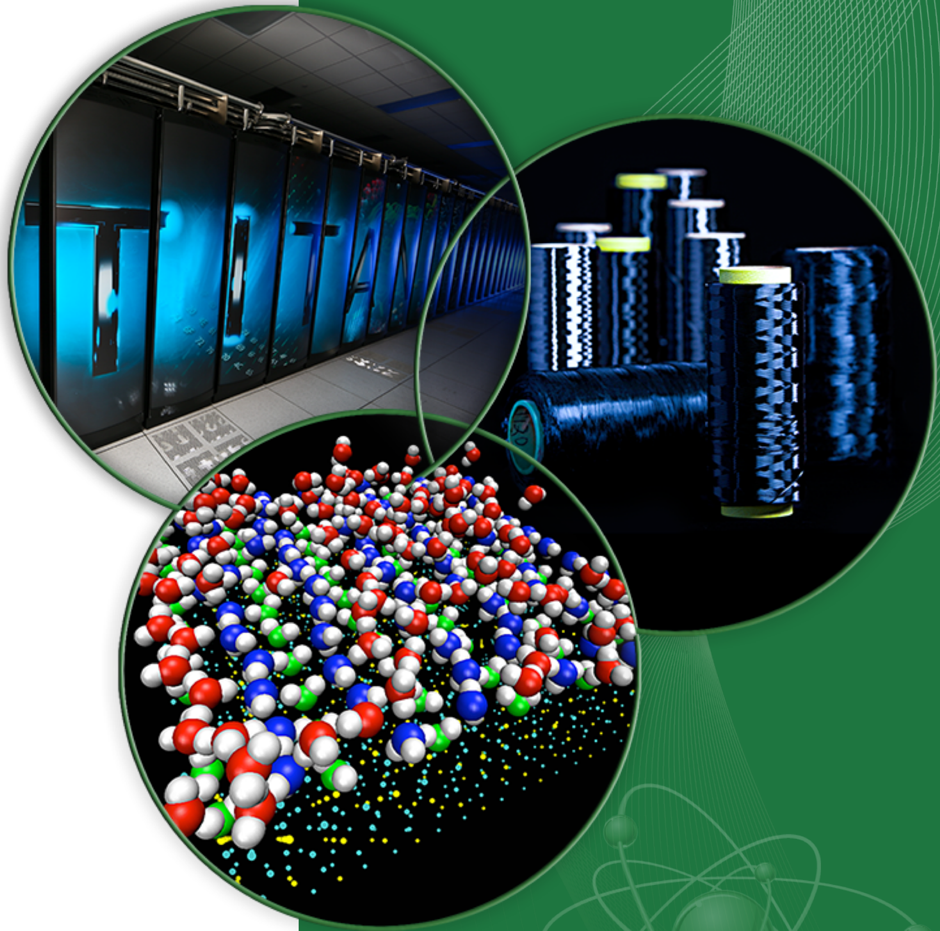
SHARP and AR



ORNL is managed by UT-Battelle
for the US Department of Energy

OAK RIDGE
National Laboratory

Summit Data Network

- Mellanox EDR Network
  - Non-blocking fat-tree
  - Bisection BW 115 TB/s
  - 2 Physical ports (4 Virtual) per node 25 GB/s
  - Advanced features:
    - Adaptive Routing
    - SHARP

Adaptive Routing

- Infiniband is traditionally statically routed
  - This leads to higher congestion
  - Under traditional congested scenario a 1:1 fat-tree is only expected to achieve 55% of its available bandwidth

- Summit EDR Introduces Adaptive Routing
  - Enables out of order packets on the network
  - Packets are load balanced at each switch to better distribute the network workload

Enabling Adaptive Routing

- Using Spectrum MPI set environment variables
  - PAMI_IBV_ENABLE_OOO_AR=1
  - PAMI_IBV_QP_SERVICE_LEVEL=8

- When should you use AR
  - Always

- It will be enabled by default in the future.

SHARP

- Scalable Hierarchical Aggregation (and) Reduction Protocol
  - Means: Our network builds fancy trees in switches to accelerate some collective operations
  - Supported Collectives (Small <= 2048)
    - Barrier
    - Broadcast
    - Reduce
    - Allreduce

SHARP Performance Measurements

- Barrier
  - 6us@512 nodes vs 21-23 for software
- Allreduce
  - 18us@2048 nodes vs 85 - 139

Things to know

- It’s a shared resource
  - You may request it and not get it, we’re imposing allocation policies that favor jobs > 1% of the machine.
  - If you use a lot (a lot) of sub-communicators
    - It creates an OST tree for every communicator group
- Small collectives
- Bitwise reproducibility
  - OST locations are dynamic and change

How to use it

- ENABLE_SHARP="-E HCOLL_ENABLE_SHARP=2 -E HCOLL_SHARP_NP=2 -E SHARP_COLL_LOG_LEVEL=3 -E HCOLL_BCOL_P2P_ALLREDUCE_SHARP_MAX=2048 -E SHARP_COLL_JOB_QUOTA_OSTS=64 -E SHARP_COLL_POLL_BATCH=1 -E SHARP_COLL_SHARP_ENABLE_MCAST_TARGET=0 -E SHARP_COLL_ENABLE_MCAST_TARGET=0 -E SHARP_COLL_JOB_QUOTA_PAYLOAD_PER_OST=256”

- ENABLE_HCOLL="-mca coll_hcoll_enable 1 -mca coll_hcoll_np 0 -mca coll ^basic -mca coll ^ibm -HCOLL -FCA”

- jsrun -n … -r … $ENABLE_SHARP --smpiargs="$ENABLE_HCOLL”

- Most important option
  - HCOLL_ENABLE_SHARP=
    - 1 (Probe and use it) Falls back to HCOLL if unsuccessful
    - 2 (Force use it) Falls back to application failure if unsuccessful
    - 3 & 4 Various nuances on 2

More info

- http://www.mellanox.com/related-docs/prod_acceleration_software/Mellanox_SHARP_SW_Deployment_Guide_v5.0.pdf
  - Details the 23 environment variables that can be used to tune SHARP