HPC ADVISORY COUNCIL

SNAP Performance Benchmark and Profiling

April 2014

Note

- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: HP, Mellanox

- **For more information on the supporting vendors solutions please refer to:**
  - www.mellanox.com, http://www.hp.com/go/hpc

- **For more information on the application:**
  - https://asc.llnl.gov/CORAL-benchmarks/#snap

- **SNAP**
  - Stands for: SN (Discrete Ordinates) Application Proxy
  - Serves as a proxy application to model the performance of a modern discrete ordinates neutral particle transport application.
  - May be considered an update to Sweep3D, which Is intended for hybrid computing architectures
  - Mimics the computational workload, memory requirements, and communication patterns of PARTISN
  - SNAP Is modeled off the LANL code PARTISN
    - PARTISN solves the linear Boltzmann transport equation (TE)
    - A governing equation for determining the number of neutral particles in a multi-dimensional phase space

Objectives

- **The presented research was done to provide best practices**
  - SNAP performance benchmarking
  - Interconnect performance comparisons
  - MPI performance comparison
  - Understanding SNAP communication patterns

- **The presented results will demonstrate**
  - The scalability of the compute environment to provide nearly linear application scalability

Test Cluster Configuration

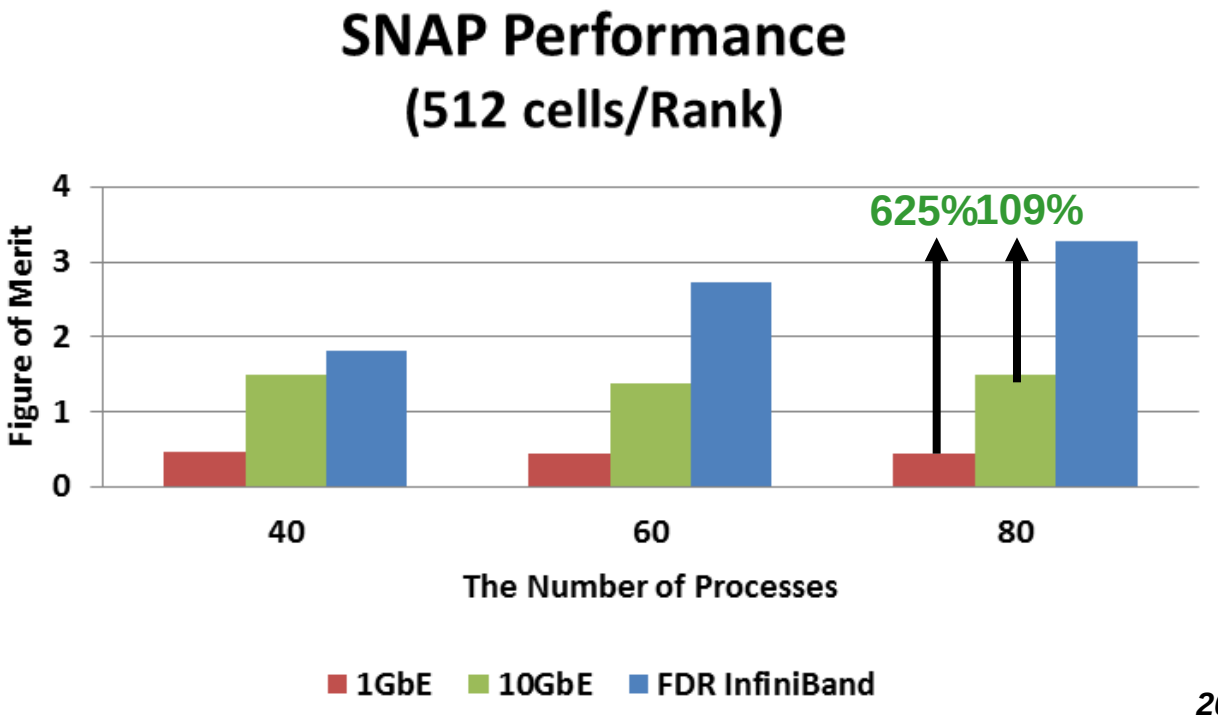
- **HP ProLiant SL230s Gen8 4-node "Athena" cluster**
  - Processors: Dual-Socket 10-core Intel Xeon E5-2680v2 @ 2.8 GHz CPUs
  - Memory: 32GB per node, 1600MHz DDR3 Dual-Ranked DIMMs
  - OS: RHEL 6 Update 2, OFED 2.0-3.0.0 InfiniBand SW stack
- **Mellanox Connect-IB FDR InfiniBand adapters**
- **Mellanox ConnectX-3 VPI Ethernet adapters**
- **Mellanox SwitchX SX6036 56Gb/s FDR InfiniBand and Ethernet VPI Switch**
- **MPI: Platform MPI 8.3, Open MPI 1.6.5**
- **Compiler: GNU Compilers**
- **Application: SNAP 1.03**
- **Benchmark Workload:**
- **Input dataset:**
  - 512 cells per rank

About HP ProLiant SL230s Gen8

Item	HP ProLiant SL230s Gen8 Server
Processor	Two Intel® Xeon® E5-2600 v2 Series, 4/6/8/10/12 Cores,
Chipset	Intel® Xeon E5-2600 v2 product family
Memory	(256 GB), 16 DIMM slots, DDR3 up to 1600MHz, ECC
Max Memory	256 GB
Internal Storage	Two LFF non-hot plug SAS, SATA bays or Four SFF non-hot plug SAS, SATA, SSD bays Two Hot Plug SFF Drives (Option)
Max Internal Storage	8TB
Networking	Dual port 1GbE NIC/ Single 10G Nic
I/O Slots	One PCIe Gen3 x16 LP slot 1Gb and 10Gb Ethernet, IB, and FlexF abric options
Ports	Front: (1) Management, (2) 1GbE, (1) Serial, (1) S.U.V port, (2) PCIe, and Internal Micro SD card & Active Health
Power Supplies	750, 1200W (92% or 94%), high power chassis
Integrated Management	iLO4 hardware-based power capping via SL Advanced Power Manager
Additional Features	Shared Power & Cooling and up to 8 nodes per 4U chassis, single GPU support, Fusion I/O support
Form Factor	16P/8GPUs/4U chassis

SNAP Performance - Interconnect

- **FDR InfiniBand is the most efficient inter-node communication for SNAP**
  - Outperforms 10GbE by 109% at 80 MPI processes
  - Outperforms 1GbE by 625% at 80 MPI processes
  - Performance benefit of InfiniBand expects to grow at larger CPU core counts

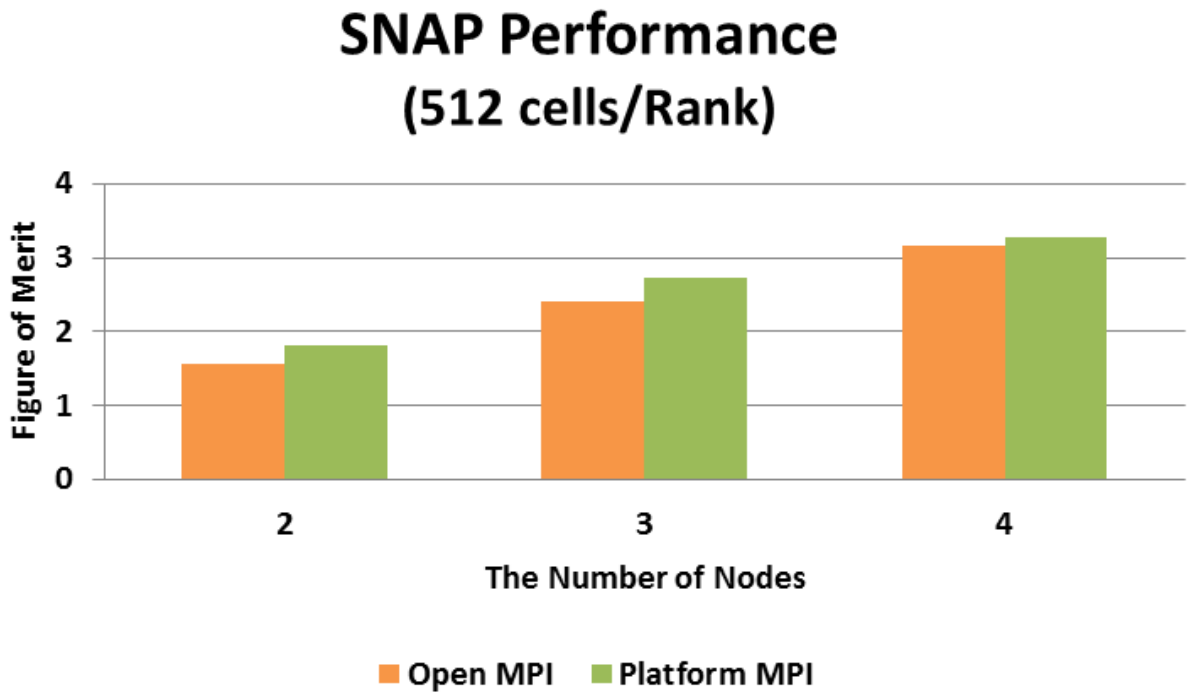


Higher is better

20 Processes/Node

SNAP Performance - MPI

- **Platform MPI shows higher performance at small node count**
  - The gap between Open MPI and Platform closes in at higher core counts
  - Processor binding and same compiler flags are used for both cases

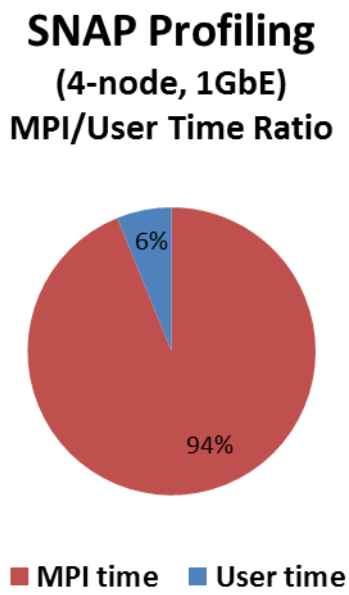


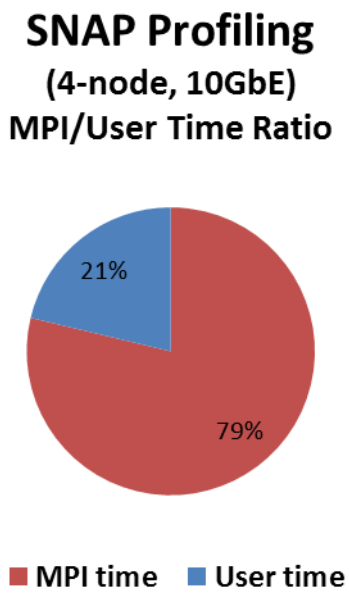
Higher is better

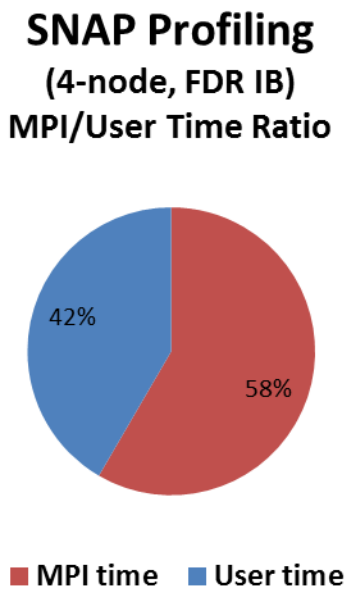
20 Processes/Node

SNAP Profiling – MPI Time Ratio

- **FDR InfiniBand reduces the communication time at scale**
  - FDR InfiniBand consumes about 58% of total runtime
  - 1GbE consumes 94% of total time
  - 10GbE consumes about 79% of total runtime



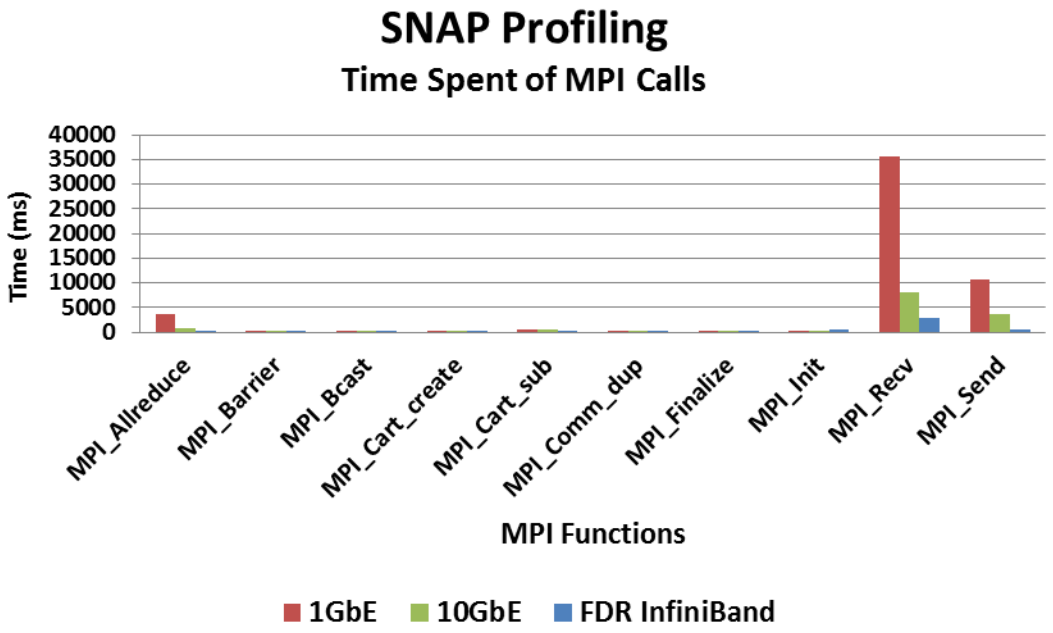


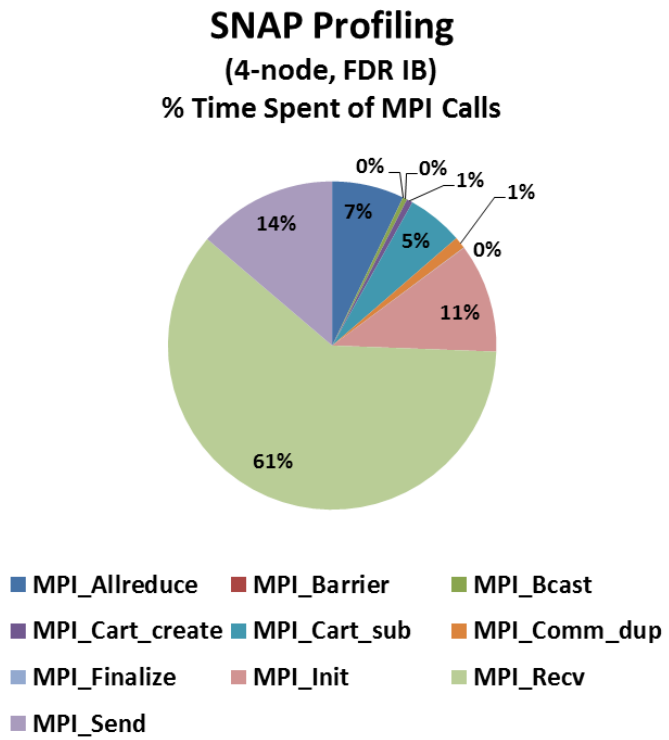


20 Processes/Node

SNAP Profiling – MPI Functions

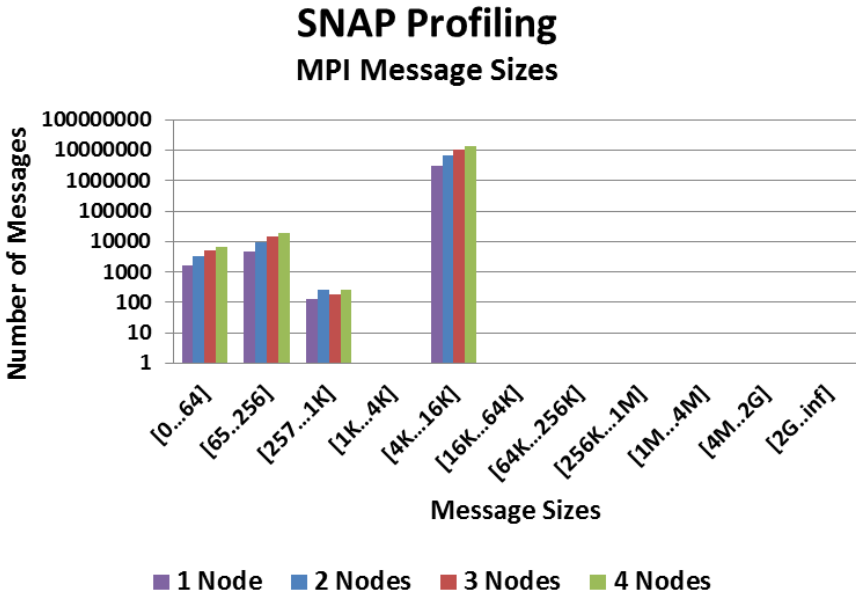
- **The most time consuming MPI functions:**
  - MPI_Recv (61%), MPI_Send (14%)
- **Time spent on network bandwidth differentiates among interconnects**
  - 1GbE and 10GbE spent more time in MPI_Recv/MPI_Send
  - Demonstrated that InfiniBand performs better than Ethernet networks

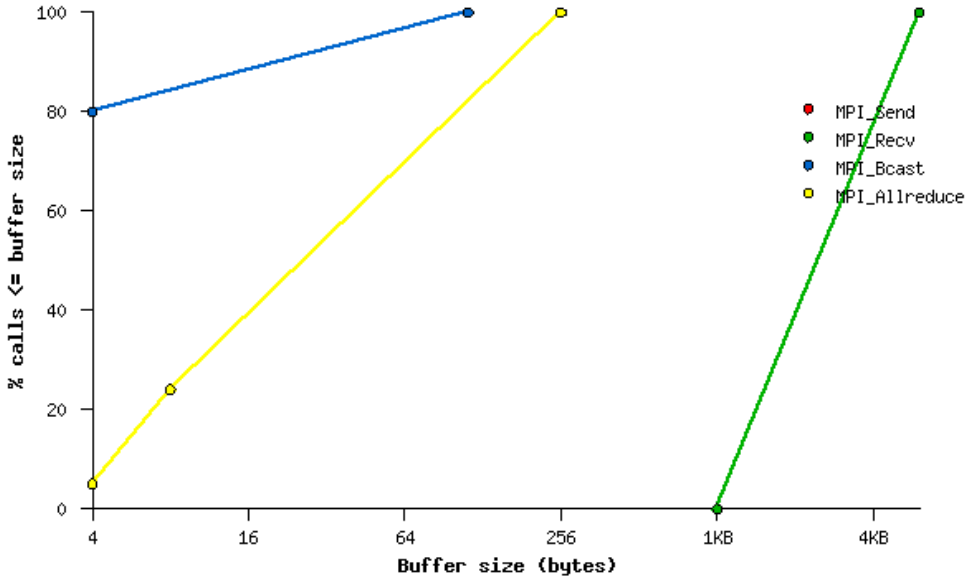




SNAP Profiling – MPI Message Sizes

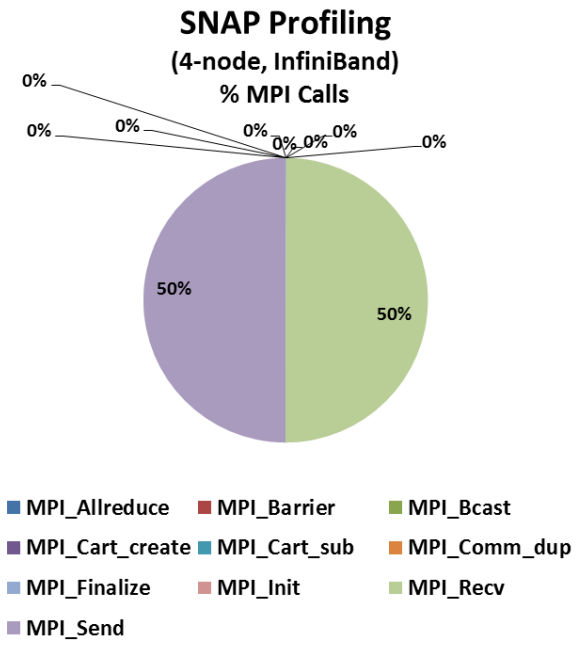
- **Messages for MPI_Recv are concentrated around 4KB**
  - Most of the blocking send and receive happened at those messages
- **The rest of the messages are less by comparison**
  - The MPI collective operations appeared at smaller sizes between below 1KB
  - MPI_Bcast at ~4B to 64B
  - MPI_Allreduce at < 256B
- **The messaging behavior stays as job scales**

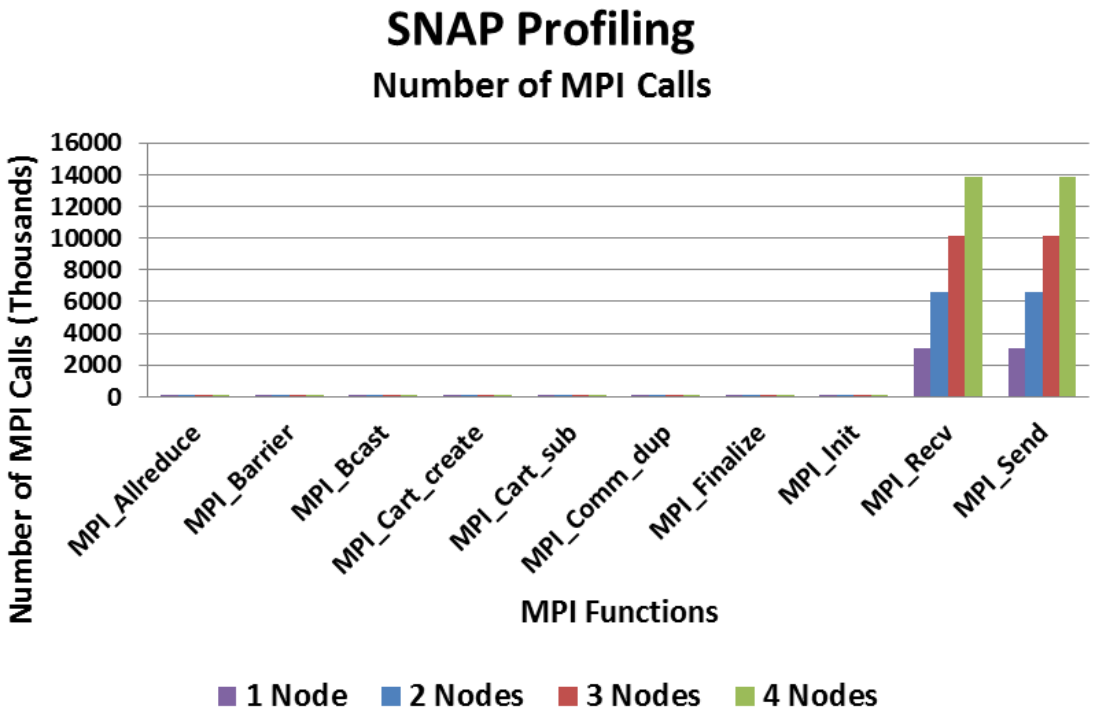




SNAP Profiling – MPI Functions

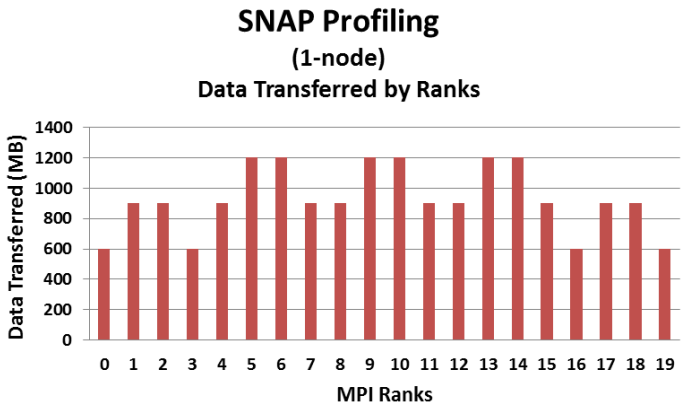
- **The number of MPI calls used is split between the two**
  - MPI_Recv (~50%) and MPI_Send (~50%)

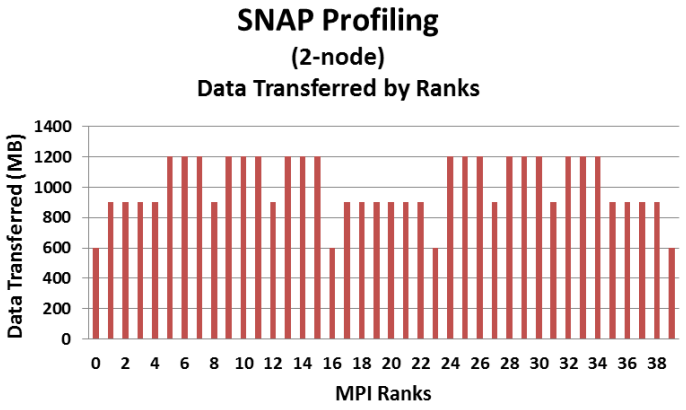


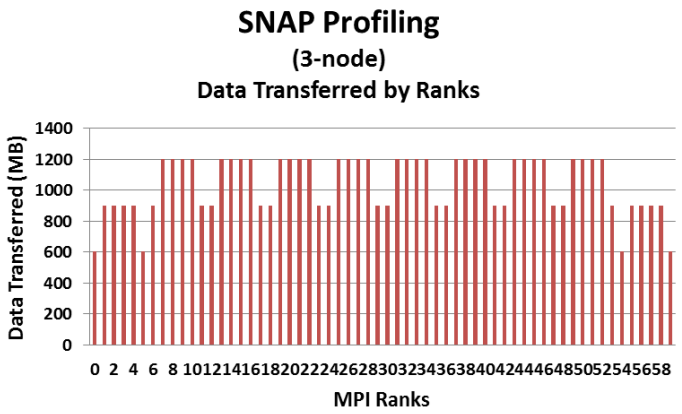


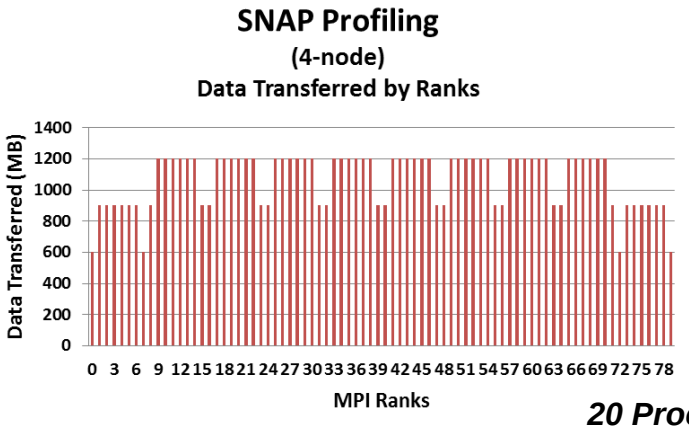
SNAP Profiling – Network Transfers

- **The application shows some imbalances in communication per rank**
  - Some ranks conducts more communications than others
  - The level/size of messages stays the same as node count increases





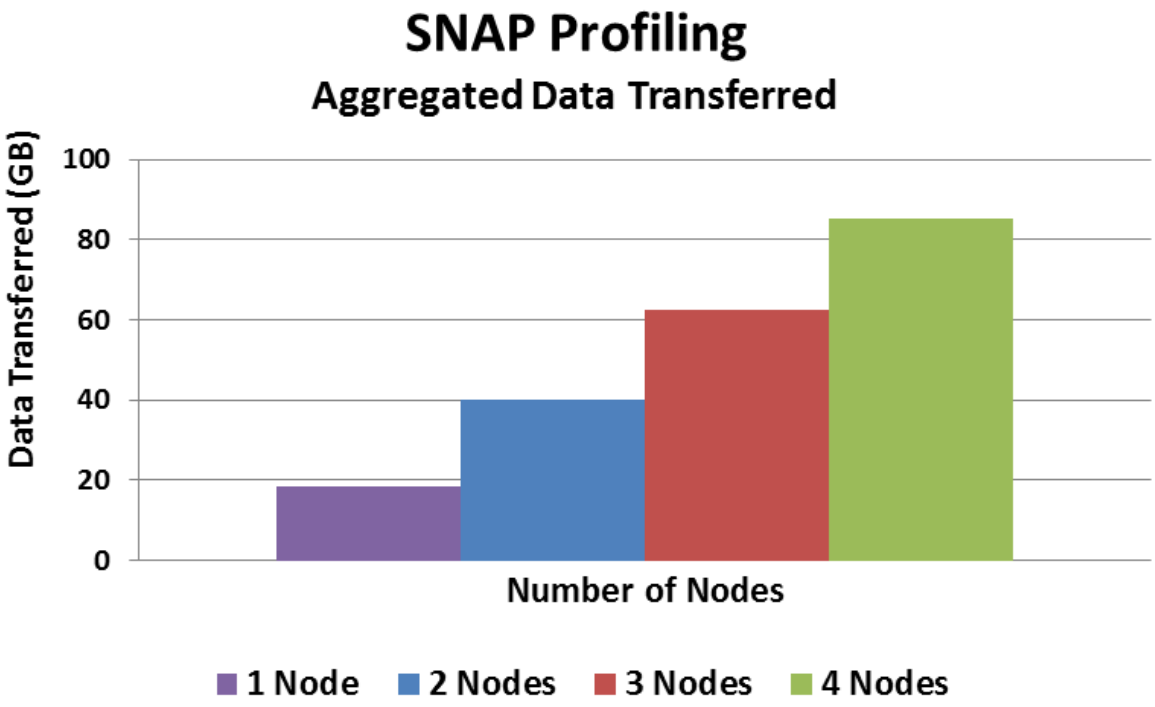




20 Processes/Node

SNAP Profiling – Aggregated Transfer

- **Aggregated data transfer refers to:**
  - Amount of data being transferred in the network between all MPI ranks collectively
- **Data transfer grows as more ranks involved**
  - For weak scaling problem, the problem size grows together with cluster size
  - As problem size grows, the amount of data transferred grows simultaneously



20 MPI Processes

SNAP Summary

- **Performance of SNAP is directly affected by the network throughput**
- **FDR InfiniBand delivers the most efficient network communication for SNAP**
  - Outperforms 10GbE up to 109% at 4 nodes (or 80 MPI processes)
  - Outperforms 1GbE up to 625% at 4 nodes (or 80 MPI processes)
- **MPI Profiling**
  - FDR InfiniBand reduces communication time; leave more time for computation
    - FDR InfiniBand consumes 58% of total time
    - Versus 79% against 10GbE and a whopping 94% against 1GbE
  - Point-to-point communications (MPI_Recv) is the most time-consumed operation
  - Blocking communications are seen:
    - Time spent: MPI_Recv(61%)
    - Most used are split between MPI_Send and MPI_Recv, each accounts for 50%

Thank You

HPC Advisory Council