The Impact of Inter-node Latency versus Intra-node Latency on HPC Applications

The 23rd IASTED International Conference on PDCS 2011

HPC|Scale Working Group, Dec 2011

Gilad Shainer, Pak Lui, Tong Liu, Todd Wilde, Jeff Layton

HPC Advisory Council, USA

HPC Advisory Council Mission

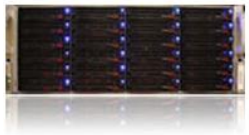
- **World-wide HPC organization (300+ members)**
- **Bridges the gap between HPC usage and its potential**
- **Provides best practices and a support/development center**
- **Explores future technologies and future developments**
- **Explores advanced topics – HPC|Cloud, HPC|Scale, HPC|GPU, HPC|Storage**
- **Leading edge solutions and technology demonstrations**
- **For more info: http://www.hpcadvisorycouncil.com**

HPC Advisory Council Members

HPC Advisory Council HPC Center

HPC ADVISORY COUNCIL

InfiniBand-based Storage (Lustre)



- Two Intel Core i7 920 CPUs (2.67GHz)
- DDR3-1333MHz memory (6GB total)
- Seagate Cheetah 15K 450GB SAS Hard Disk
- OS: RHEL5.2
- Mellanox ConnectX-2 40Gb/s QDR InfiniBand adapter

Lustre

Apply for System Access

intel Mellanox TECHNOLOGIES

Maia

- Dell™ PowerEdge™ C6100 4-node cluster
- Dell™ PowerEdge™ C410x PCIe Expansion Chassis
- Six-Core Intel® Xeon® processor X5670 @ 2.93 GHz
- 4 NVIDIA® Tesla C2050 (Fermi) GPU
- Mellanox ConnectX®-2 VPI 40Gb/s InfiniBand mezzanine card
- Mellanox 36-Port 40Gb/s InfiniBand switch
- Memory: 24GB memory per node

GPU Cluster

Apply for System Access

DELL intel Mellanox TECHNOLOGIES

Venus

- SUN 2200 8-node cluster
- Intel Xeon Quad-core X5472 CPUs
- Mellanox InfiniBand ConnectX® 20Gb/s InfiniBand adapter
- Mellanox 20Gb/s InfiniBand Switch
- Memory: 32GB

Apply for System Access

intel Mellanox TECHNOLOGIES Sun

Plutus

- HP Cluster Platform 3000SL
- 16 nodes HP ProLiant SL2x170z scalable servers
- Six-Core Intel® Xeon® processor X5670@ 2.93 GHz
- Memory: 24GB memory per node
- Mellanox Technology-based 40Gb/s InfiniBand adapters and switch

192 cores

Apply for System Access

hp intel Mellanox TECHNOLOGIES

Dodecas

- AMD 6000 Series platform, 8-node cluster (24-cores per node)
- AMD Opteron™ Model 6172, 45nm technology
- Mellanox ConnectX®-2 40Gb/s InfiniBand Adapters
- Mellanox M3601Q 36-Port 40Gb/InfiniBand Switch
- Memory: 32GB memory per node

Apply for System Access

AMD Mellanox TECHNOLOGIES

Vesta

- Dell™ PowerEdge™ R815 11-node cluster
- Four processors AMD Opteron 6174 (Magny-Cours), 48 Cores per node
- Dual Mellanox ConnectX®-2 40Gb/s InfiniBand adapters per node
- Mellanox 36-Port 40Gb/s InfiniBand Switch
- Memory 128 GB, 1333 MHz memory per node

Apply for System Access

704 cores

AMD DELL Mellanox TECHNOLOGIES Microsoft

Janus

- Dell™ PowerEdge™ M610 38-node cluster
- Six-Core Intel® Xeon® processor X5670 @ 2.93 GHz
- Intel Cluster Ready certified cluster
- Mellanox ConnectX®-2 40Gb/s InfiniBand mezzanine card
- Mellanox M3601Q 36-Port 40Gb/s InfiniBand Switch
- Memory: 24GB memory per node

intel Cluster Ready

Apply for System Access

456 cores

DELL intel Mellanox TECHNOLOGIES Microsoft

2012 HPC Advisory Council Workshops

- **Upcoming events in 2012**
  - Israel – February 7
  - Switzerland – March 13-15
  - Germany – June 17
  - China – October
  - USA (Stanford University, CA) – December
- **The conference will focus on:**
  - HPC usage models and benefits
  - Latest technology developments around the world
  - HPC best practices and advanced HPC topics
- **The conference is free to attendees**
  - Registration is required
- **For more information**
  - www.hpcadvisorycouncil.com, info@hpcadvisorycouncil.com

University Award Program

- **University award program**
  - One of the HPC Advisory Council's activities is community and education outreach, in particular to enhance students' computing knowledge-base as early as possible
  - Universities are encouraged to submit proposals for advanced research around high-performance computing
  - Twice a year, the HPC Advisory Council will select a few proposals

- **Selected proposal will be provided with:**
  - Exclusive computation time on the HPC Advisory Council's Compute Center
  - Invitation to present the research results in one of the HPC Advisory Council's worldwide workshops, including sponsorship of travel expenses (according the to the Council Award Program rules)
  - Publication of the research results on the HPC Advisory Council website and related publications
  - Publication of the research results and a demonstration if applicable within HPC Advisory Council world-wide technology demonstration activities

- **Proposals for the 2012 Spring HPC Advisory Council University Award Program can be submitted between November 1, 2011 through May 31, 2011. The selected proposal(s) will be determined by June 10th and the winner(s) will be notified.**

Joining the HPC Advisory Council

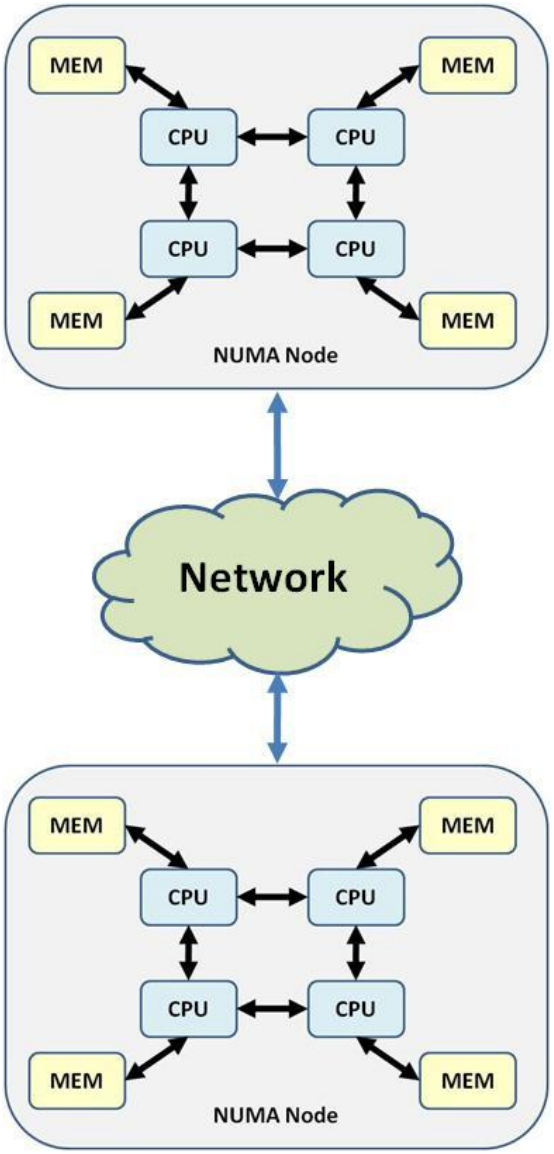
www.hpcadvisorycouncil.com

info@hpcadvisorycouncil.com

Note

- **The following research was performed under the HPC Advisory Council HPC|Scale working group:**
  - Analyzing the effect of Inter-node and Intra-node latency in the HPC environment
    - HPC System Architecture Overview
    - Illustration of Latency Components
    - Comparisons of Latency Components
    - Example: Simple Model
    - Example: Application Model Using WRF
    - MPI Profiling: Time usage in WRF

HPC System Architecture Overview



- **Typical HPC system architecture:**
  - Multiple compute nodes
  - Connected together via the cluster interconnect
- **Intra-node communications involves technologies such as:**
  - PCI-Express (PCIe)
  - Intel QuickPath Interconnect (QPI)
  - AMD HyperTransport (HT)
- **The intra-node latency is between**
  - CPU cores and memory within a NUMA node or between NUMA nodes
- **The inter-node latency is between**
  - Across different compute nodes over a network

Illustration of Inter-node Latency

- **The inter-node (network) latency can be expressed as:**

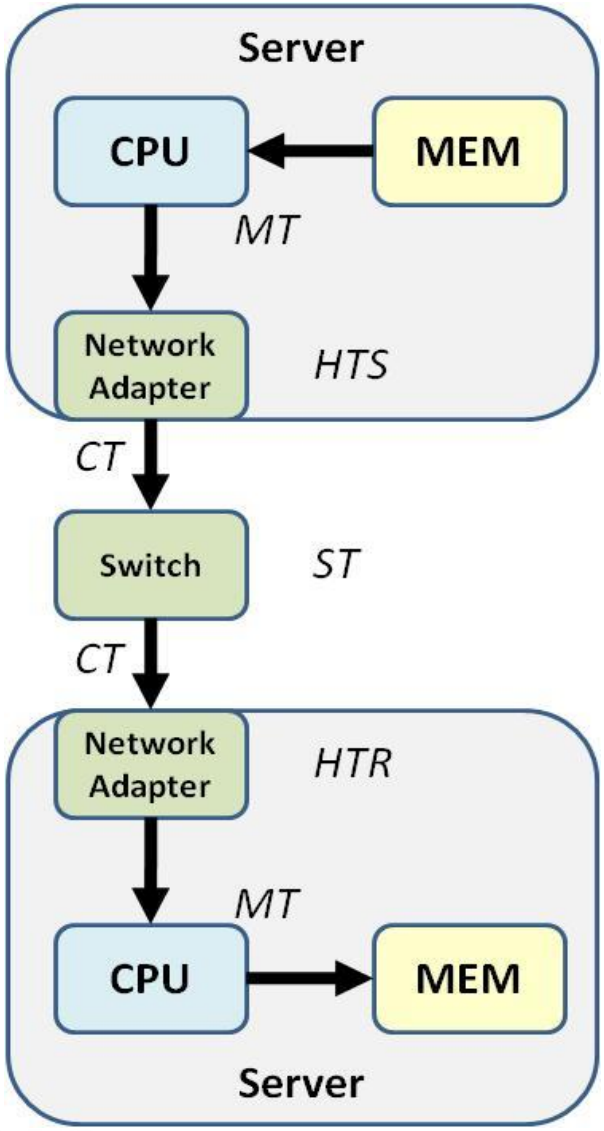
$$NT = \sum_1^{X-1} ST + \sum_1^{X} CT + HTS + HTR$$

- **When cable latency (CT) is much less than switch latency (ST), it can be expressed as:**

$$NT = \sum_1^{X-1} ST + HTS + HTR$$

- **Typically, *Network Adapter* latency for send and receive are in the same magnitude**
- **Overall network latency:**

$$Inter - Node\ T = NT + 2\ x\ MT$$



Comparisons of Latency Components

Component	Latency
Inter-Node Latency (X=0)	1050nsec
Inter-Node Latency (X=1)	1150nsec
Inter-Node Latency (X=5)	1550nsec
Intra-Node Latency	400nsec

InfiniBand QDR

- **The table shows the measured latency components**
  - The inter-node and intra-node latency are in the same order of magnitude
- **Network offloads is crucial to maintain low inter-node latency**
  - Transport offload (0-copy, RDMA, etc) saves the transport processing overhead at the CPU level, context switching, etc.
  - From the source to destination host memory through the IB network
- **Gap between inter-node & intra-node latency can be closed when**
  - Intra-node latency is increased: when the CPU is busy
  - Inter-node latency RTS: Transport offload networking solution is used

Application Synchronizations

- **An application runtime can be defined as:**

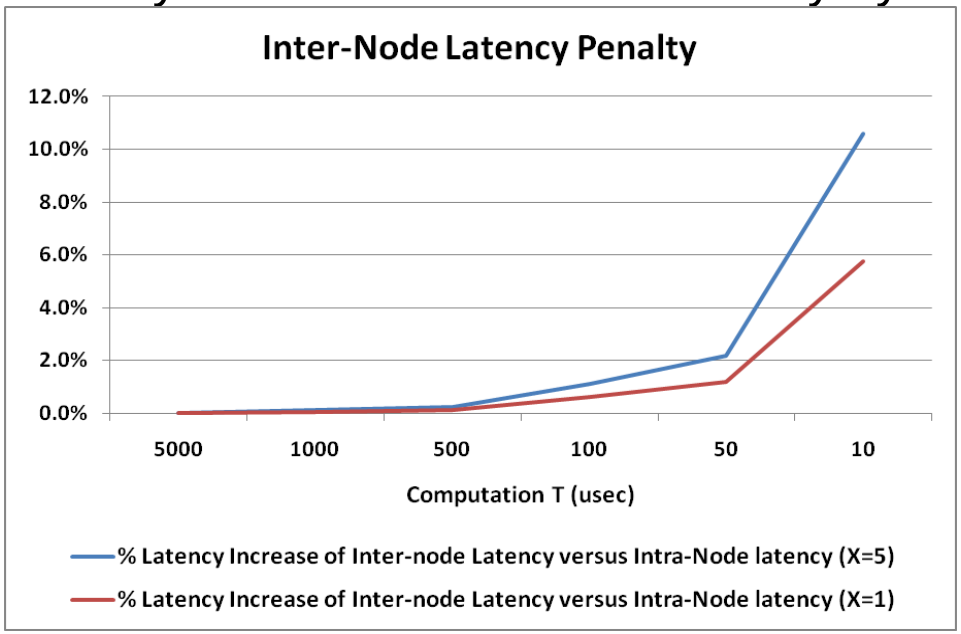
$$Application\ T = \sum_{i=1}^{n} (Computation\ T + Synchronization\ T)$$

- **Typical parallel HPC application consists of:**
  - Compute periods and global synchronization stages
- **Compute periods involves:**
  - Application compute cycles
  - With or without data exchange with other compute nodes
  - Any MPI collective operations (such as MPI_Bcast and MPI_Allreduce)
- **Global synchronization:**
  - Occurs at the end of compute cycles
  - Ends only after ALL of the MPI processes complete the compute period
- **Any delay in the compute cycle can affect the cluster performance**
  - The next compute cycle can start only after the global synchronization process has complete

Example: Network Switch Hops Impact

- **Inter-node latency penalty depends on the number of switch hops**
  - 1 hop (2 end points); 5 hops (11664 end points, in a non-blocking config)
- **In a highly parallel application:**
  - Application spends small time for synchronization (more time in compute)
  - The inter-node latency has effect is negligible (typical case)
- **If application spends shorter time in computation**
  - the inter-node latency increases the overall latency by 10% (worst case)

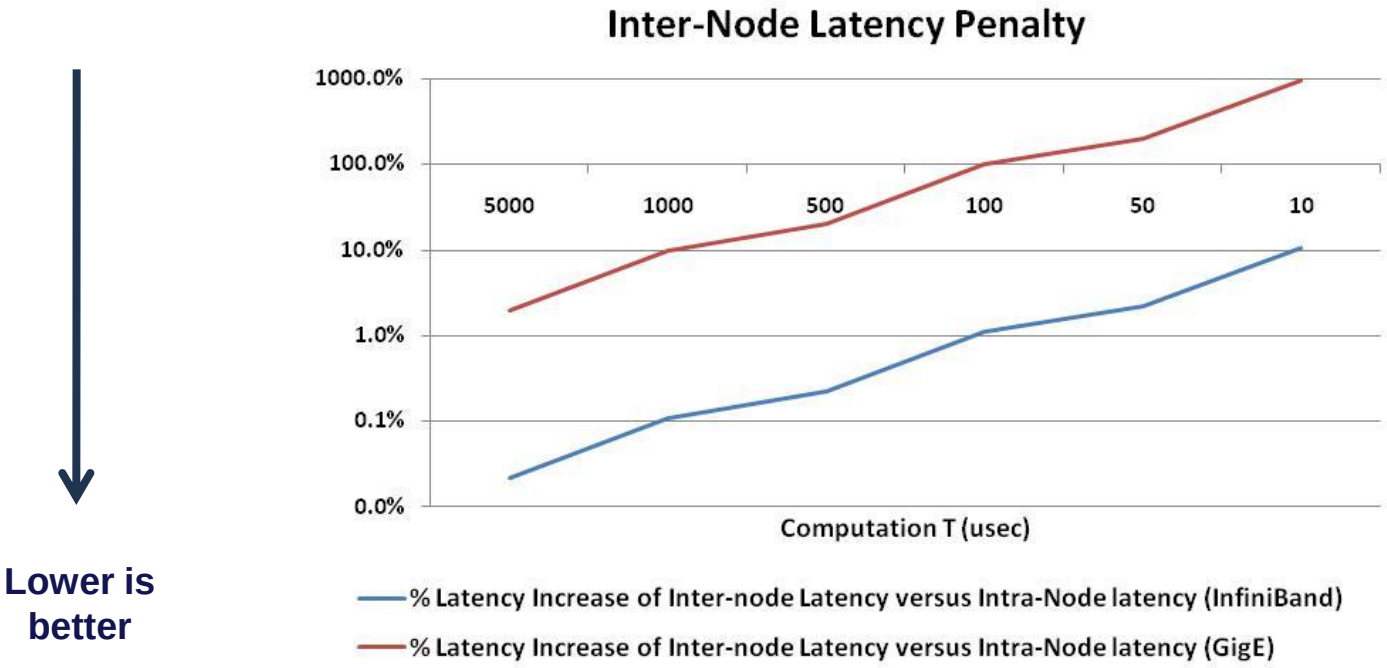
Lower is better



InfiniBand QDR

Example: Network Interconnect Impact

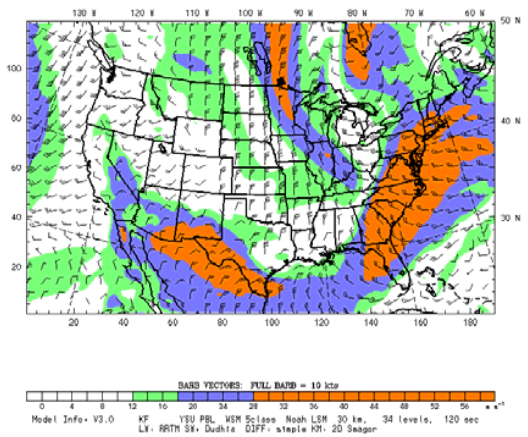
- **Comparison of low-latency to high-latency interconnects**
  - Low-latency (InfiniBand): shows lower ratio compared to intra-node
  - High-latency (1GbE): causes much higher penalty for outgoing access
- **High-latency network causes significant performance degradation**
  - About 2 orders of magnitude (100x) greater than low-latency networks

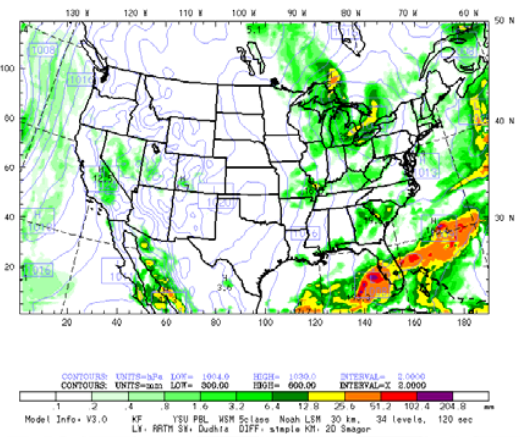


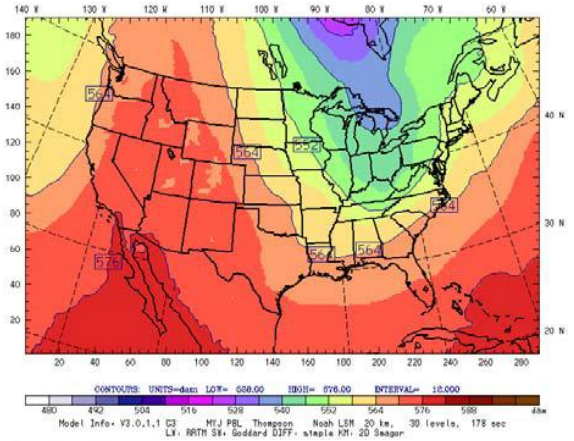
Lower is better

Weather Research and Forecasting (WRF)

- **The Weather Research and Forecasting (WRF) Model**
  - Numerical weather prediction system
  - Designed for operational forecasting and atmospheric research
- **The WRF model usage:**
  - Is designed to be an efficient massively parallel computing code
  - Can be configured for both research and operations
  - Offers full physics options
  - Real-data and idealized simulations
  - Applications ranging from meters to thousands of kilometers





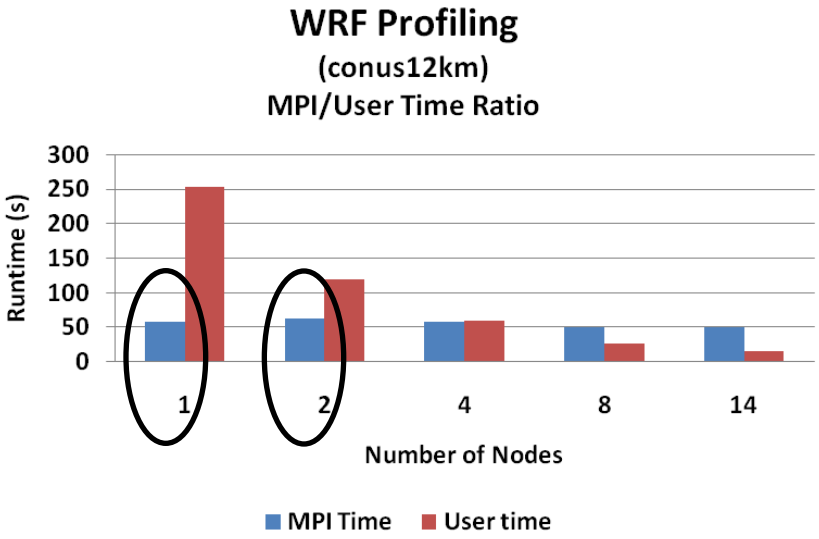


Test Cluster Configuration

- **Dell™ PowerEdge™ M610 14-node cluster**
  - Six-Core Intel X5670 @ 2.93 GHz CPUs
  - Memory: 24GB memory, DDR3 1333 MHz
  - OS: CentOS 5 Update 4, OFED 1.5.1 InfiniBand SW stack
- **Intel Cluster Ready certified cluster**
- **Mellanox ConnectX-2 InfiniBand adapters and switches**
- **MPI: Platform MPI 8.0.1**
- **Compilers: Intel Compilers 12.0.0**
- **Miscellaneous package: NetCDF 4.1.1**
- **Application: WRF 3.2.1**
- **Benchmarks: CONUS-12km - 48-hour, 12km resolution case over the Continental US from October 24, 2001**

WRF Profiling – MPI/Compute Time Ratio

- **WRF demonstrates the ability to scale as the node count increases**
  - As cluster scales, the runtime is reduced because compute time is reduced
- **The MPI time stays generally constant as the cluster scales up**
  - Time used to broadcast data to all cores is the same, regardless of cluster size
- **Intra-node does not provide any benefit over inter-node**
  - Shows **no differences** in MPI time between single node (over shared memory) and 2-node case (over InfiniBand)

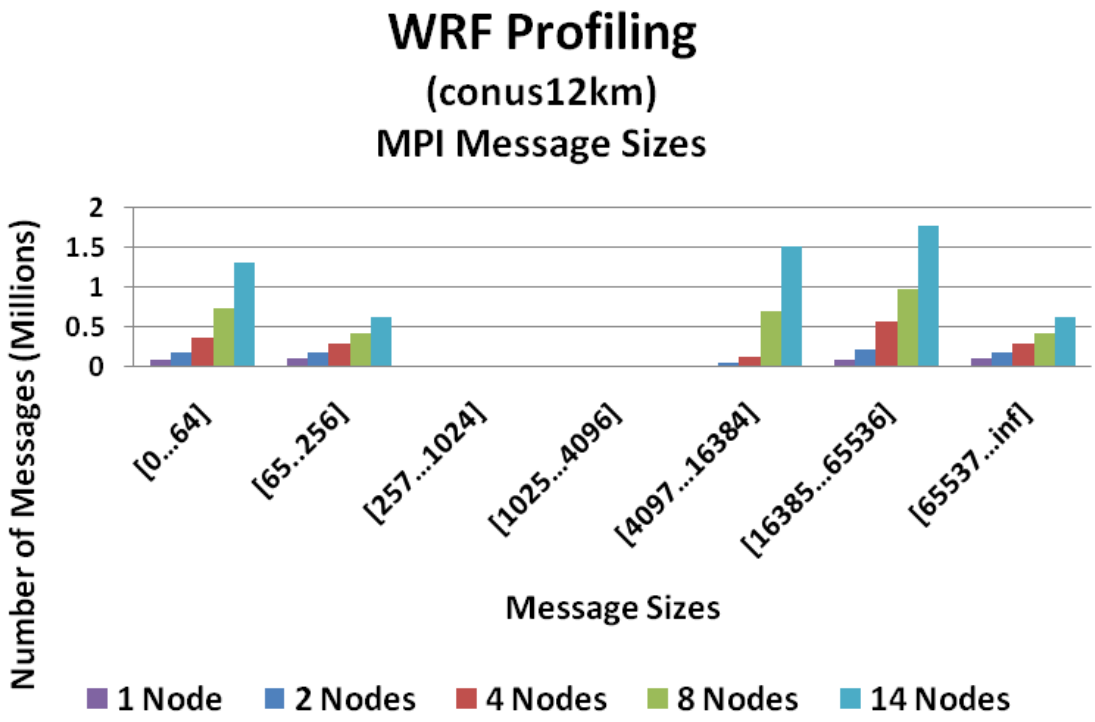


InfiniBand QDR

12 Cores/Node

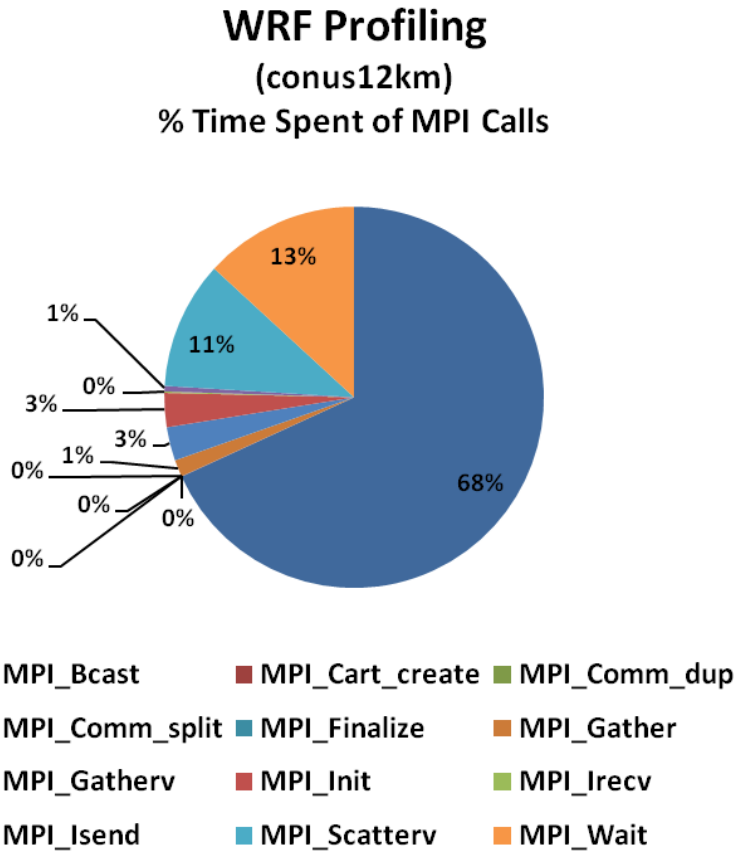
WRF Profiling – MPI Message Sizes

- **MPI message distributions from 1 to 14 nodes**
- **Messages increase proportionally with the cluster size**
  - WRF distributes the to-be-computed database among the cores
  - As more CPUs are used, the larger amount of messages are exchanged



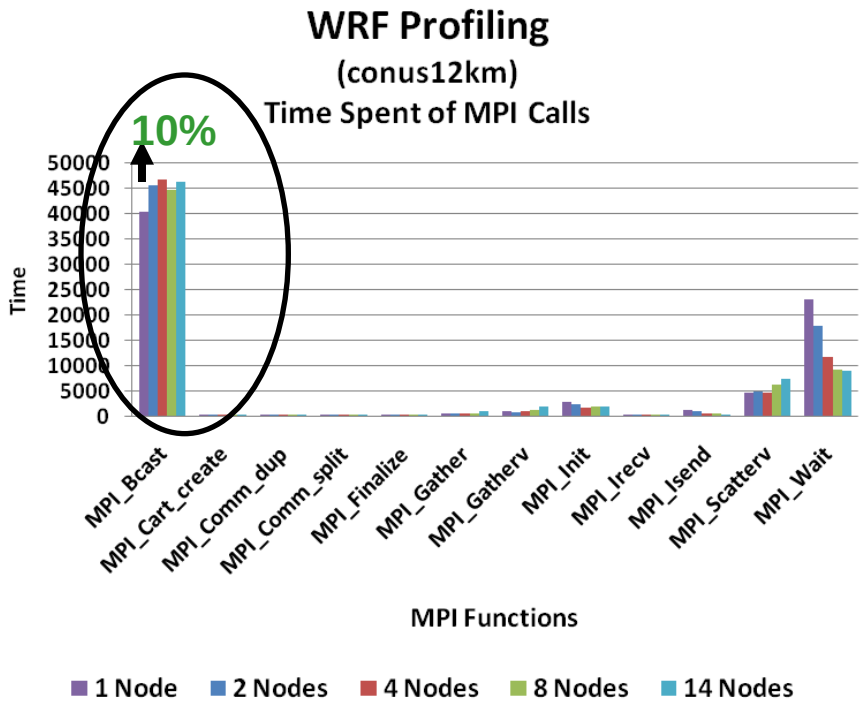
WRF Profiling – Time Spent by MPI Calls

- **MPI Profiling to illustrate the usage of various MPI collective ops**
- **Majority of communication time is spent on MPI_Bcast**
  - MPI_Bcast is accounted for 68% of time spent on a 14-node job



WRF Profiling – Time Spent by MPI Calls

- **Profiling shows timing of MPI collective operations per cluster size**
  - MPI_Wait reduces as more processes are used
- **The 10% time difference in MPI_Bcast between 1 and 2+ nodes**
  - Reflects the penalty of intra-node versus inter-node communications
  - Meets out expectations from the mathematical model of a few msec operations



Conclusions

- **Using low-latency interconnect to access remote CPUs or GPUs**
  - Has minimal penalty for applications compute durations
- **Network offloads is crucial to maintain low inter-node latency**
  - InfiniBand network with support for transport offload (0-copy, RDMA)
- **Inter-node latency introduced by switch hops is minimal**
  - ~10% for short duration tasks (worst case)
- **High-latency network (1GbE) has a large performance degradation**
  - The performance difference is in 2 orders of magnitude (about 100x)
- **Using WRF application:**
  - Overall MPI time:
    - Intra-node does not provide any benefit over inter-node (network)
  - On MPI_Bcast only:
    - Shows ~10% difference for broadcasting data between intra-node (in shared memory) and inter-node (InfiniBand) communications

Thank You

HPC Advisory Council

www.hpcadvisorycouncil.com