

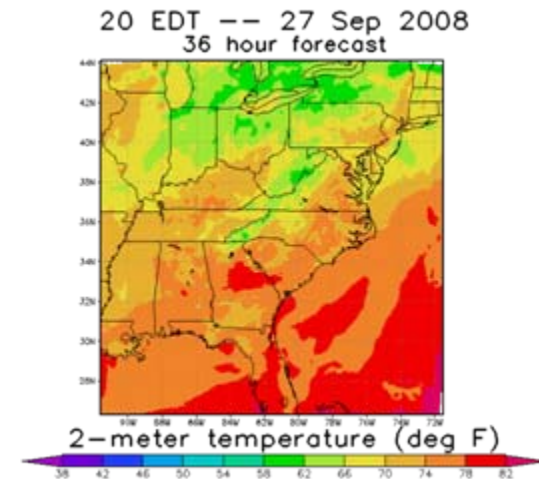
天气研究和预报模式 (WRF) 的性能分析与优化

2008年10月



- 该项研究由美国高性能计算咨询委员会完成(HPC Advisory Council)
 - AMD, 戴尔, Mellanox
 - 高性能计算咨询委员会计算机中心
- 感谢美国国家大气研究中心首席开发人员John Michalakes对这项研究的支持
- 更多信息请查询下列网站
 - www.mellanox.com, www.dell.com/hpc, www.amd.com

- **天气研究模式 (WRF) Model**
 - 数值天气预报系统
 - 为天气预报和大气研究而设计
- **WRF 由下列成员开发**
 - 美国国家大气研究中心 (NCAR)
 - 美国国家环境预报中心 (NOAA)
 - 美国天气预报研究院 (FSL)
 - 空军气象局 (AFWA)
 - 海军研究实验所
 - 俄克拉荷马大学
 - 联邦航空局 (FAA)



气候研究模式 (WRF) 用途

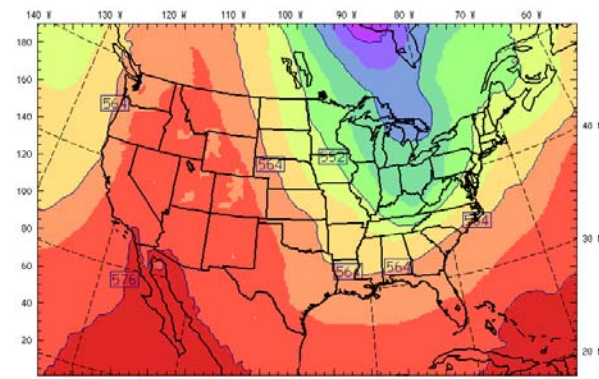
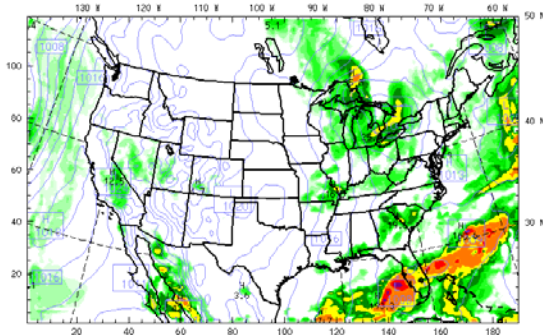
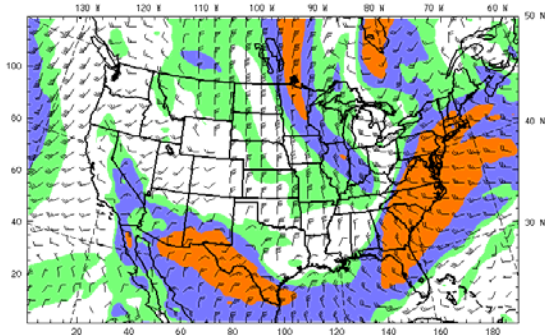
- 气候研究模式包括

- 实时与理想化数据的仿真模拟(Simulations)
- 多种侧边界处理方案
- 全物理方案
- 非静力与静力
- 程序支持的范围从几米到若干公里

ARW WRF - 30KM-NEST - NCAR/MMM
Fcst. 21 h Valid: 21 UTC Tue 30 Sep 08 (15 MDT Tue 30 Sep 08)
Supercell type (9-10 km rel. flow) sa= 5
Supercell motion vectors

ARW WRF - 30KM-NEST - NCAR/MMM
Fcst. 18 h Valid: 18 UTC Tue 30 Sep 08 (12 MDT Tue 30 Sep 08)
Total precip. since h 0
Total precip. since h 0
Sea-level pressure sw= 4

20km ARW WRF. GFS-init -- NCAR/MMM
Fcst. 18 h Valid: 18 UTC Wed 01 Oct 08 (12 MDT Wed 01 Oct 08)
1000 to 0500 hPa thickness sa= 2



Model Info: V3.0 KF YSU PBL WSM ScvLaso Noah LEM 30 km, 34 levels, 120 sec
LX: RRTH S8, Dudkita DIFF, stapsie KM: 20 Saagar

Model Info: V3.0 KF YSU PBL WSM ScvLaso Noah LEM 30 km, 34 levels, 120 sec
LX: RRTH S8, Dudkita DIFF, stapsie KM: 20 Saagar

Model Info: V3.0.1.1 C3 MJJ PBL Thompson Noah LEM 30 km, 30 levels, 178 sec
LX: RRTH S8, Goddard DIFF, stapsie KM: 20 Saagar

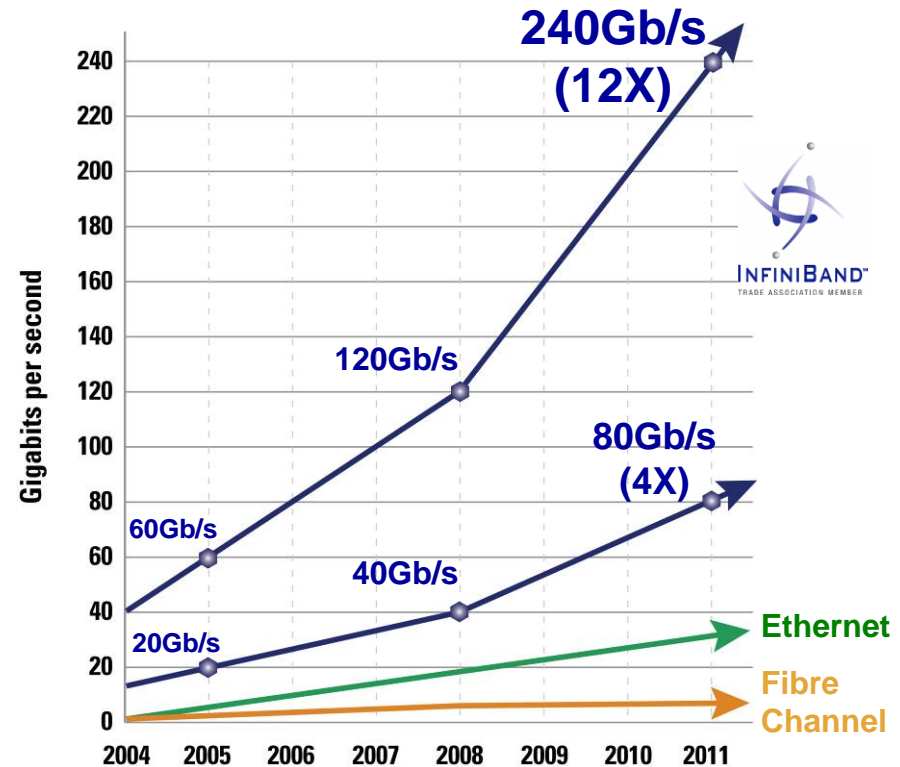
- 该项研究实现以下目的
 - WRF 模式性能基准分析/Benchmarking
 - 网络通讯设备的比较及其对WRF性能的影响
 - 提高WRF运算效率
 - WRF 网络通信特点的分析
 - MPI libraries 的比较

测试所用的计算机集群的配置

- 24台戴尔 PowerEdge SC 1435 服务器组成的集群
- 4核AMD Opteron™ 2358 SE CPUs
- Mellanox® InfiniBand ConnectX® DDR 网络适配卡 (HCA)
- 内存: 16GB, DDR2 677MHz
- 操作系统: RH 5.1, OFED 1.3 InfiniBand 软件包
- MPI库: Open MPI 1.3, MVAPICH 1.1, HP MPI 2.2.7
- 测试程序: WRF V3, 12km CONUS 基准分析数据源
- 编译器: Gfortran v4.2
 - 编译Flags: FCOPTIM= -O3 -ffast-math -ftree-vectorize -ftree-loop-linear -funroll-loops

- **基于工业标准**
 - 硬件，软件，网线，及适配器的管理
 - 为计算机集群以及存储设备的连接而设计
- **性价比**
 - 40Gb/s node-to-node
 - 120Gb/s switch-to-switch
 - 1us 延迟/latency
 - 业界最具进取性与前瞻性的产品设计蓝图
- **拥塞管理机制实现可靠性**
- **高效率**
 - RDMA and Transport Offload
 - Kernel bypass
 - CPU 被更加集中地投入到程序计算中
- **高可扩展性可满足千万亿次计算或更高的需求**
- **实现端到端服务质量的保证/Quality of service**
- **虚拟机的加速特性/Virtualization acceleration**
- **I/O整合(包括存储系统-storage)**

The InfiniBand Performance Gap is Increasing



InfiniBand Delivers the Lowest Latency

四核 AMD Opteron™ 处理器

- **性能/Performance**

- 四核/Quad-Core

- Enhanced CPU IPC
- 4x 512K L2 cache
- 2MB L3 Cache

- 直接联机架构/Direct Connect Architecture

- HyperTransport™ Technology 1.0
- Up to 8 GB/s

- 浮点运算/Floating Point

- 128-bit FPU per core
- 4 FLOPS/clock peak per core

- 内存/Memory

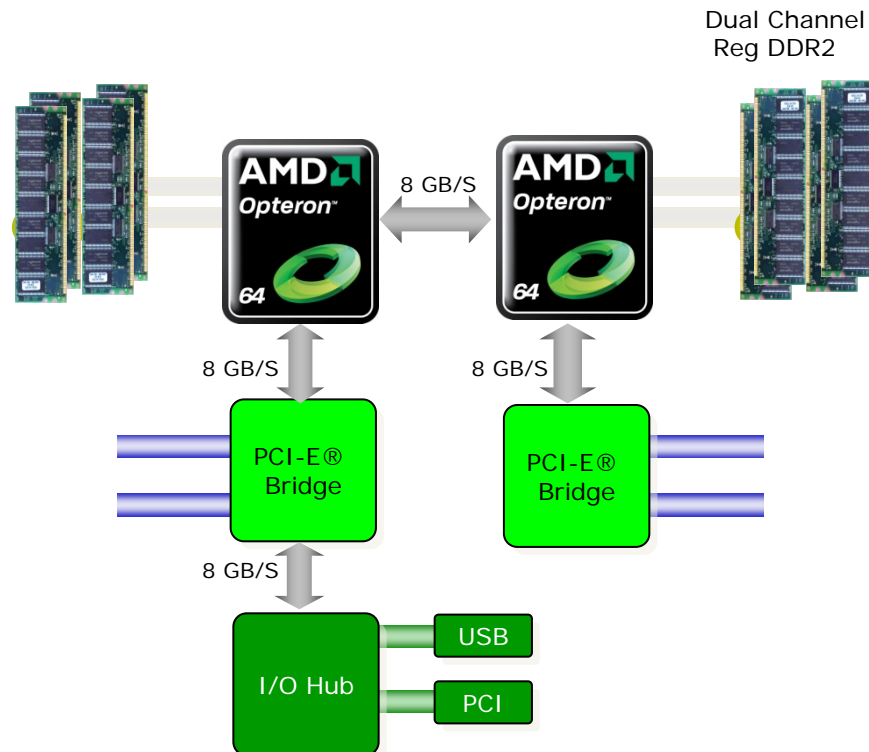
- 1GB Page Support
- DDR-2 667 MHz

- **扩展性/Scalability**

- 48-bit Physical Addressing

- **兼容性/Compatibility**

- Same power/thermal envelopes as Second-Generation AMD Opteron™ processor

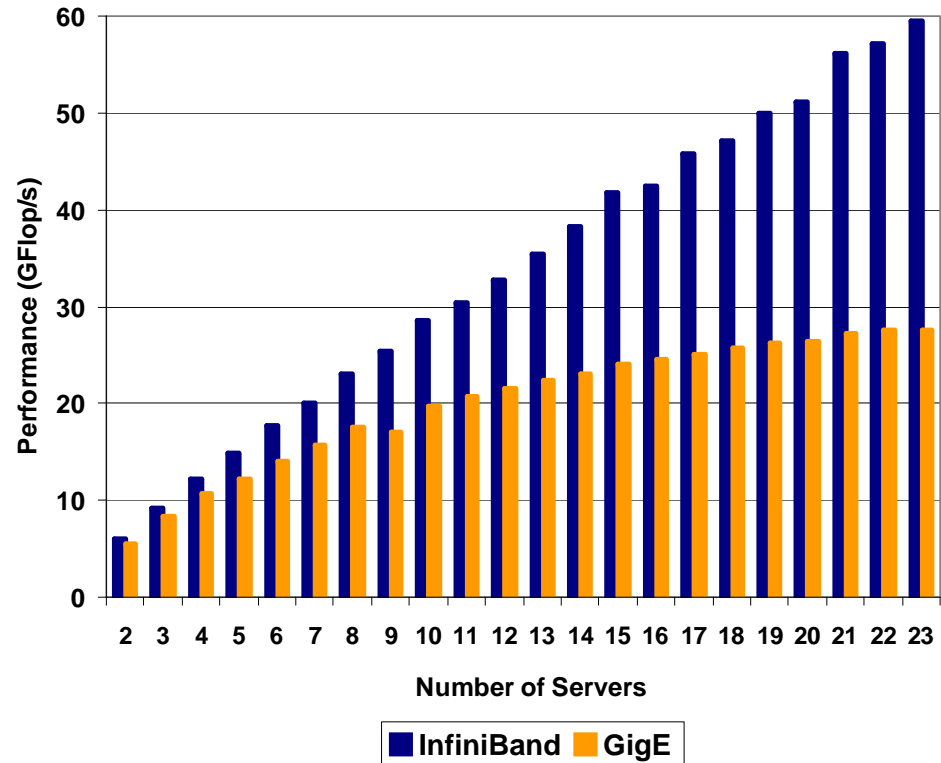


- **戴尔 HPC 解决方案**
 - 高性能与可扩展性的架构
 - 专业的实施与技术支持服务
 - 最大限度提高系统性能并减少成本与能耗
- **系统规模的设计指南/System Sizing Guidelines**
 - 系统实施布局的搭建
 - 提供经过整合与测试的系统架构
- **基于工作负荷建立解决方案/Workload Modeling**
 - 设计优化的系统规模及相关配置
 - 性能基准分析/Benchmarks
 - 应用程序特点分析/ ISV Applications Characterization
 - 最佳方案分析/ Best Practices & Usage Analysis

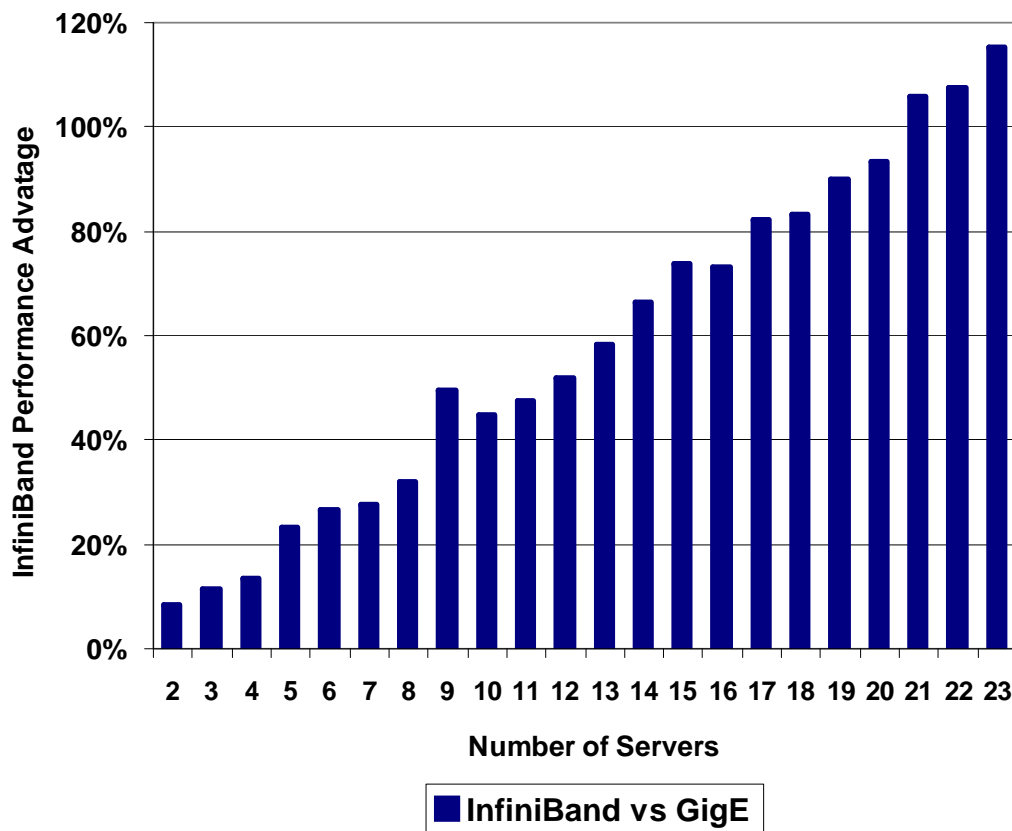


- 计算机集群网络连接设备的比较 - InfiniBand and GigE
- **InfiniBand 高速网络适配卡实现了线形扩展**
 - 最大限度的提高系统性能来实现更快速的气象模拟
- **Gigabit Ethernet网络适配器限制了WRF的运行速度**

WRF Benchmark Results - Conus 12Km



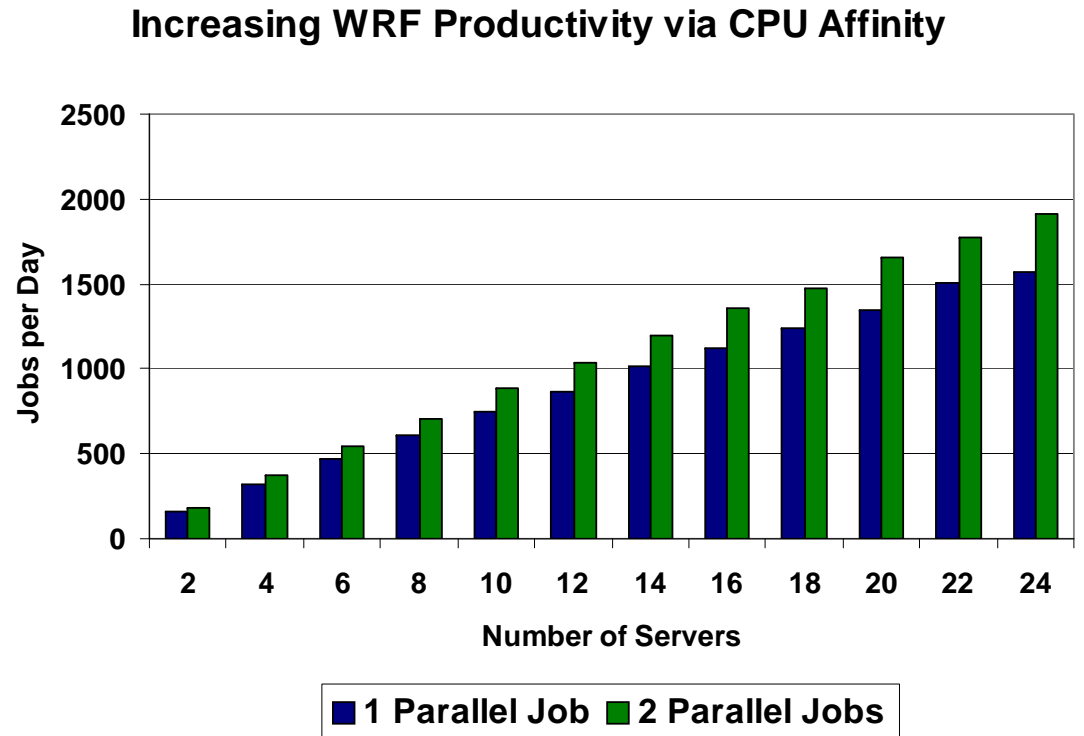
WRF Benchmark Results - InfiniBand vs GigE



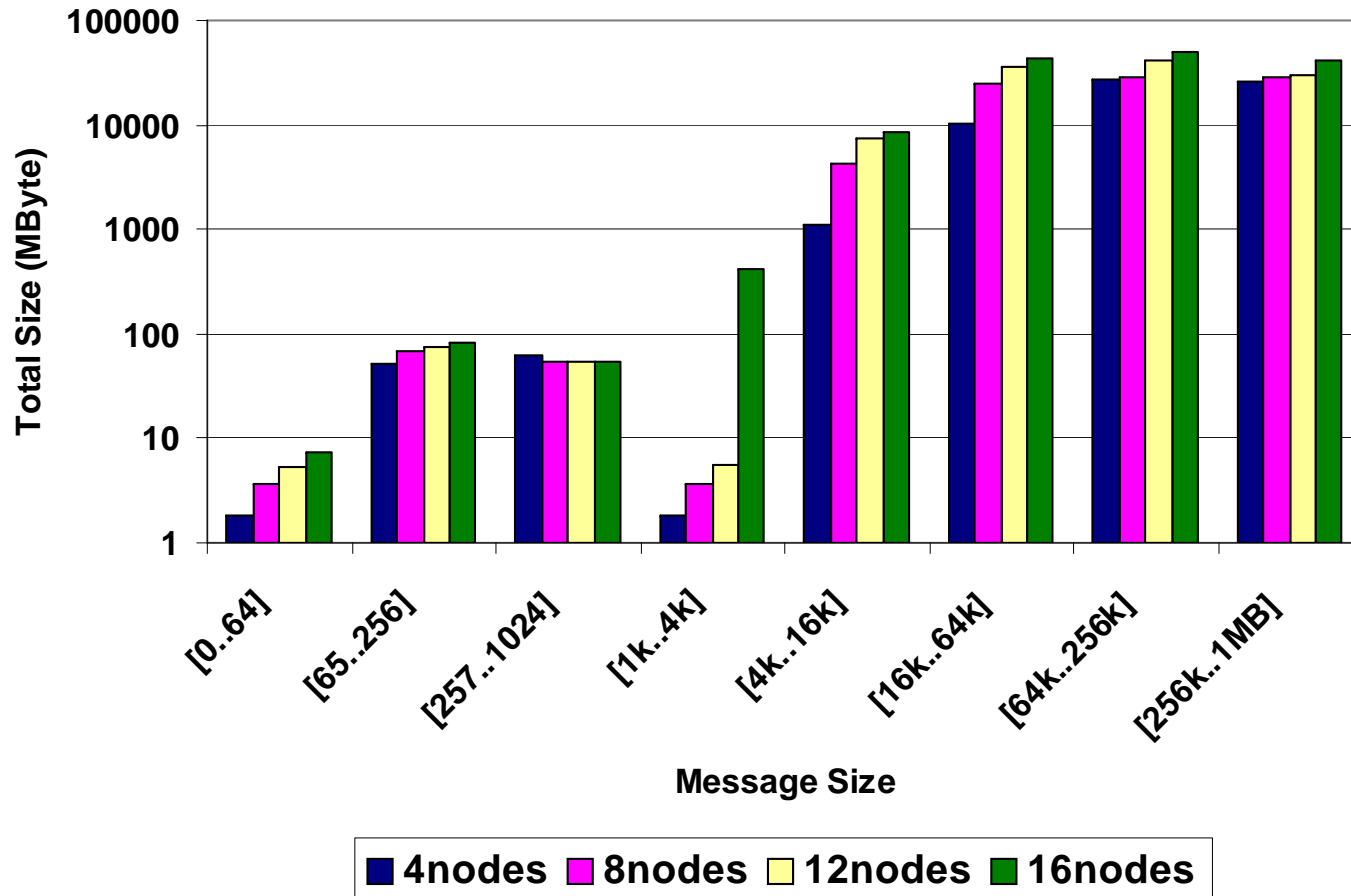
InfiniBand 提高WRF性能最高达到115%

该测试集群只有24个节点, 随着节点数增加性能差距会进一步加大

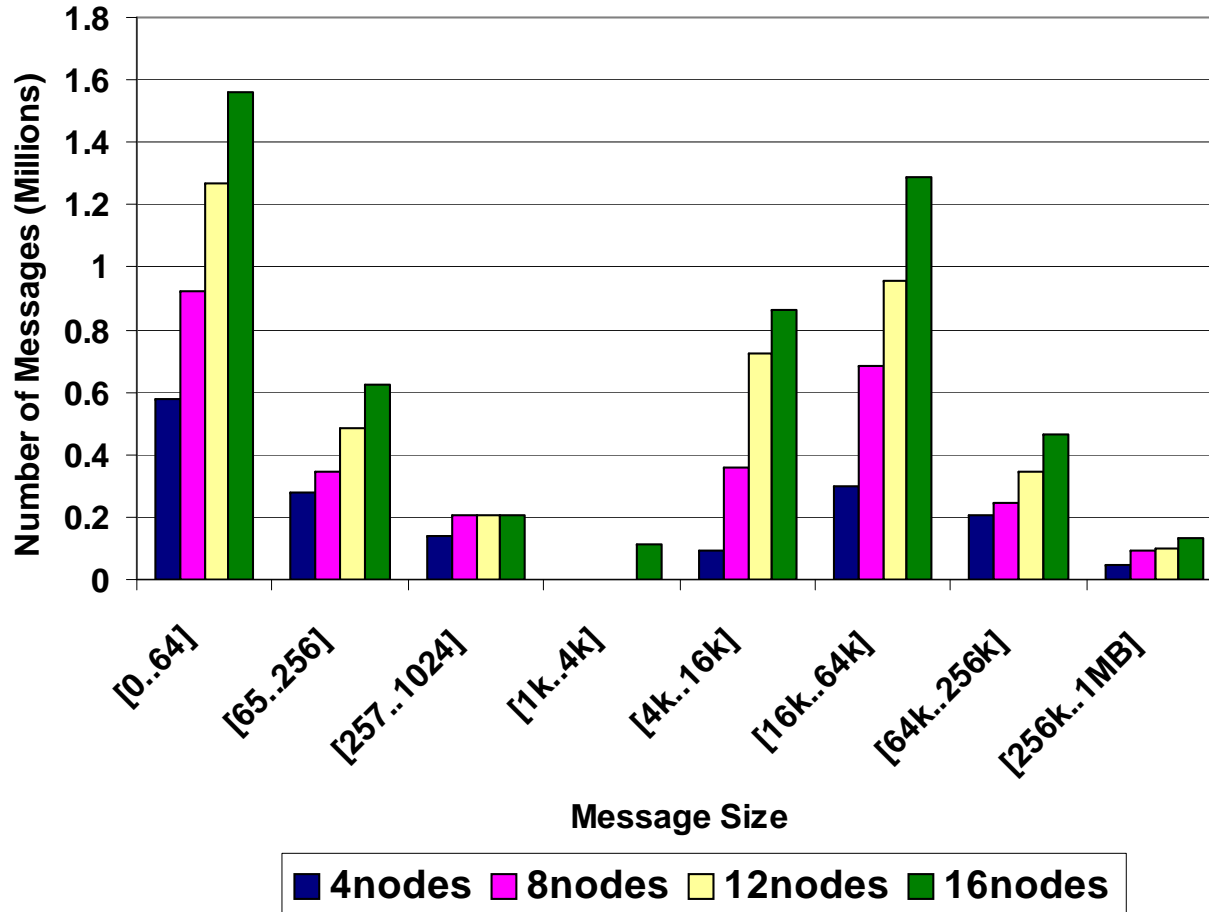
- 应用CPU affinity来实现更高的工作总量
- **Two cases**
 - 整个系统只运行一个job
 - 运行两个job, 每个job只占用每台计算机的一个CPU(同时使用CPU affinity)
- **CPU affinity 能提高工作总量20%**



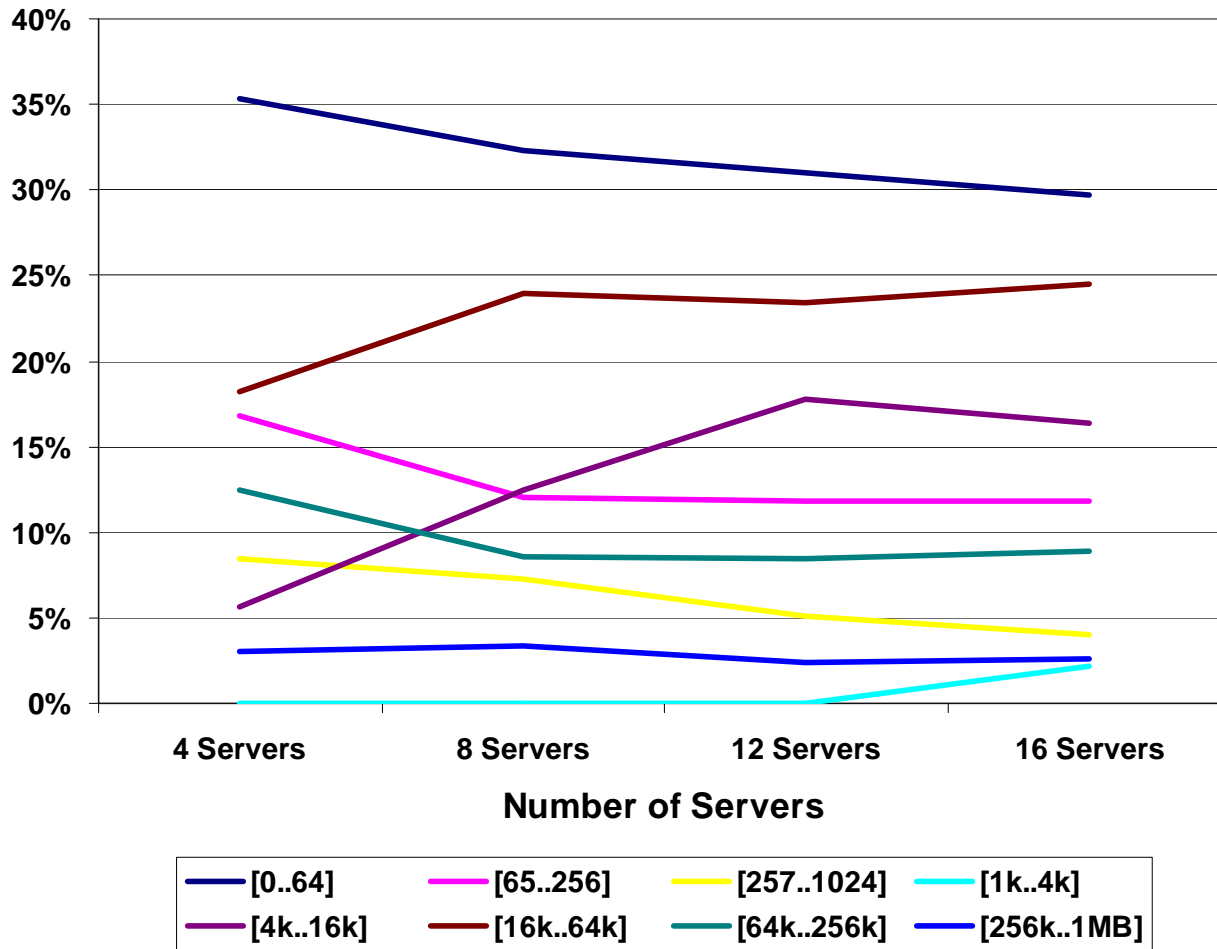
WRF MPI Profiling Total Data Send per Message Size per Cluster Size



WRF MPI Profiling Total Number of Messages per Cluster Size



WRF MPI Profiling Message Distribution

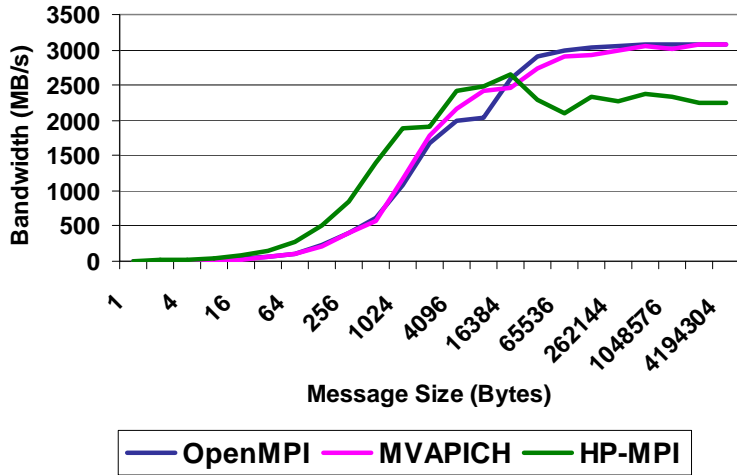


WRF程序特性分析总结

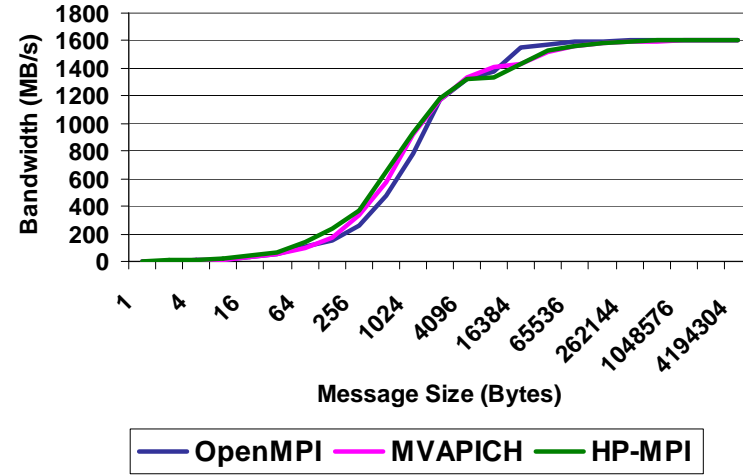
- 分析WRF 模式对网络通信的依赖
- 大部分传输于计算机之间的数据是
 - 大小位于16KB至1MB之间的message
 - 随着集群规模的加大数据传输量也会相应增加
- 大部分被应用的message尺寸
 - <64B messages – 主要用于信息同步/synchronizations
 - 16KB-64KB – 与程序计算相关
- **Message的分布**
 - 随着集群规模增大，大小message的数量都有增加
 - 在信息的总量当中
 - 大message的比率增加更快
- **WRF 取决于集群的网络连接延迟和带宽**
 - 延迟/Latency – 用于同步/synchronizations
 - 带宽/Throughput (interconnect bandwidth) – 用于计算/compute

MPI 低层性能比较 - 相对不同 MPI

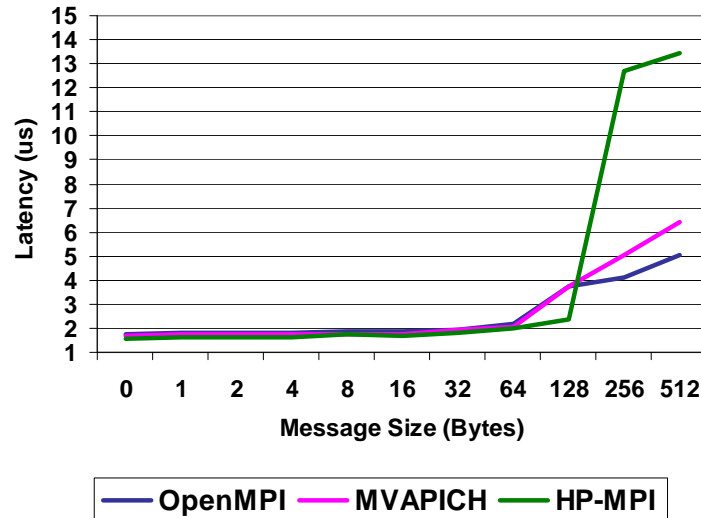
MPI Bi-Dir Bandwidth



MPI Uni-Dir Bandwidth

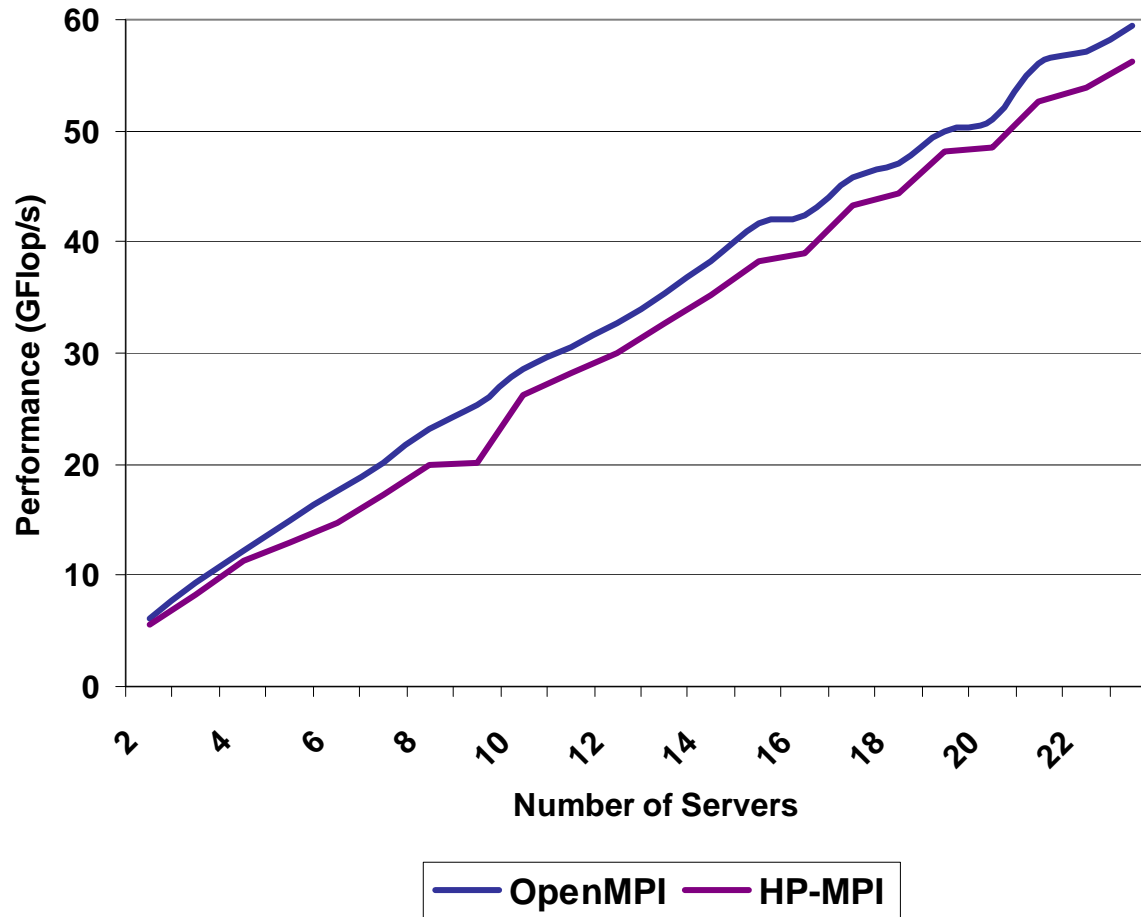


MPI Latency



WRF性能比较 - 相对不同MPI

WRF Benchmark Results - Conus 12Km



WRF性能比较总结 – 相对不同MPI

- **MPI的比较**

- 测试使用的MPI libraries– Open MPI, MVAPICH, HP-MPI
- 在Message大小达到128B以前, 所用的MPI都有相同延迟/latency
 - 之后, MVAPICH 和 Open MPI 显现出更低的延迟
- MVAPICH 和 Open MPI 有更高的双向带宽

- **WRF测试结果**

- 用MVAPICH 和 Open MPI, WRF具有相同的运行速度
- HP-MPI 则相对会产生略低的WRF性能(大约10%)
 - 由于相对略低的带宽/bandwidth和较高的延迟/latency

结论和未来工作

- **WRF是新一代天气预报的模式**
 - 实现强风暴预报与警报的关键工具
 - 2006年起被使用，是现在被最为广泛使用的天气预报模式
- **高效的WRF模式应用需要高性能计算集群系统(HPC)**
 - 实时，准确，与大范围的天气分析
- **WRF 程序特性分析证明它需要**
 - 高带宽与低延迟的网络连接设备
 - NUMA aware 程序来快速存取内存
 - 正确选择MPI Libraries
- **未来工作**
 - 能耗测试以及大的memory pages对WRF性能的影响
 - 优化的MPI collective 功能对WRF性能的影响

Thank You

HPC Advisory Council
HPC@mellanox.com



All trademarks are property of their respective owners. All information is provided "As-Is" without any kind of warranty. The HPC Advisory Council makes no representation to the accuracy and completeness of the information contained herein. HPC Advisory Council Mellanox undertakes no duty and assumes no obligation to update or correct any information presented herein