



CAS2K5

Jim Tuccillo

jtuccillo@Inxi.com

912.576.5215



Agenda

- Corporate Overview
- System Architecture
- Node Design
- Processor Options
- Interconnect Options
- High Performance File Systems – Lustre
- System Management
- System Software
- Services and Training
- WRF Scalability

Linux Networx Corporate Overview

- Mission
 - Focus exclusively on Linux clustered solutions, tools, and services
 - High Performance production computing systems
 - Government, Commercial, Academic
- Growth/Funding
 - Revenues have grown an average of over 100% per year for the past four years
 - Similar growth rates projected for 2005 and 2006
 - \$40 million in funding during 2004
 - Significant additions of key management and technical personnel from the HPC industry
 - 40% increase in integration capacity in 2005
- Future
 - Continued growth – Expertise in system engineering, software, and networking
 - Balanced introduction of dependable, leading edge technologies
 - Development of new technologies – focus on production HPC

Company Background

- Incorporated in 1989
- Headquarters: Salt Lake City, Utah
- Privately held
- Operations: Americas, Europe, Asia

What We Do

- Hardware Integration:
 - Commodity components – processors, memory, motherboard
 - Purpose built chassis, cooling system and power distribution units
- Software Integration:
 - Standard Linux OS distributions from Red Hat or SuSE
 - Clusterworx Cluster Management Software
 - Linux BIOS
 - Open Source Tools, MPI libraries, and applications software
- Installation & Overall System Integration
 - Test, validate, & integrate hardware and software components prior to shipment
 - Verify cooling and power requirements for your specific configuration and layout

Position in the Industry

- 1997 - First company to deliver a Linux cluster
- 1997 – First cluster management tools
- 2000 - First design & patent on vertical nodes
- 2002 – Built world's most powerful Linux cluster - first cluster to be ranked as a top 5 computer
- 2003 - Designed & delivered a 2,800 CPU cluster in under three months — an industry record
- Over 1000+ systems and 200+ sites

Linux Networx



Systems



Site: US Army Research Laboratory
Computer Name: John Von Neumann
Rmax: 8.770 TFlops
Best Top500 Ranking: 13
System Model: Evolocity II
Processors: 2048 Xeon 3.4 GHz



Site: Los Alamos National Laboratory
Computer Name: Lightning
Rmax: 8.051 TFlops
Best Top500 Ranking: 6
System Model: Evolocity
Processors: 2816 Opteron 2 GHz



Site: Lawrence Livermore National Lab.
Computer Name: MCR
Rmax: 7.634 TFlops
Best Top500 Ranking: 3
System Model: Evolocity II
Processors: 2304 Xeon 2.4 GHz



Site: Grid Technology Research Center
Computer Name: AIST Super Cluster
Rmax: 1.997 TFlops
Best Top500 Ranking: 97
System Model: Evolocity II
Processors: 512 Xeon 3.06 GHz

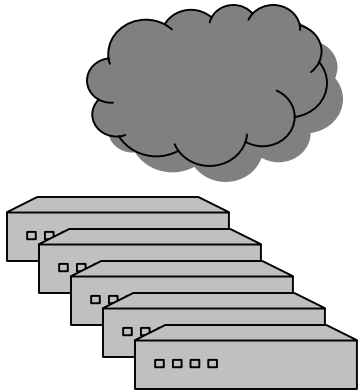


Site: Sandia National Laboratories
Computer Name: Catalyst
Rmax: 1.076 TFlops
Best Top500 Ranking: 111
System Model: Evolocity II
Processors: 256 P4 Xeon 3.06 GHz



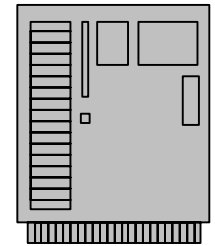
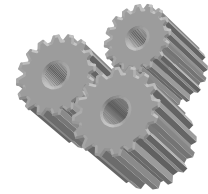
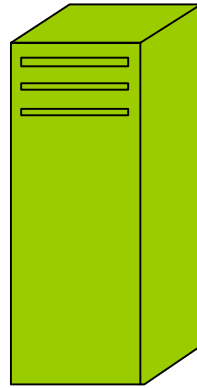
Site: US Army Research Laboratory
Computer Name: Powell
Rmax: 1.060 TFlops
Best Top500 Ranking: 113
System Model: Evolocity II
Processors: 256 Pentium4 Xeon 3.06 GHz

Linux Network



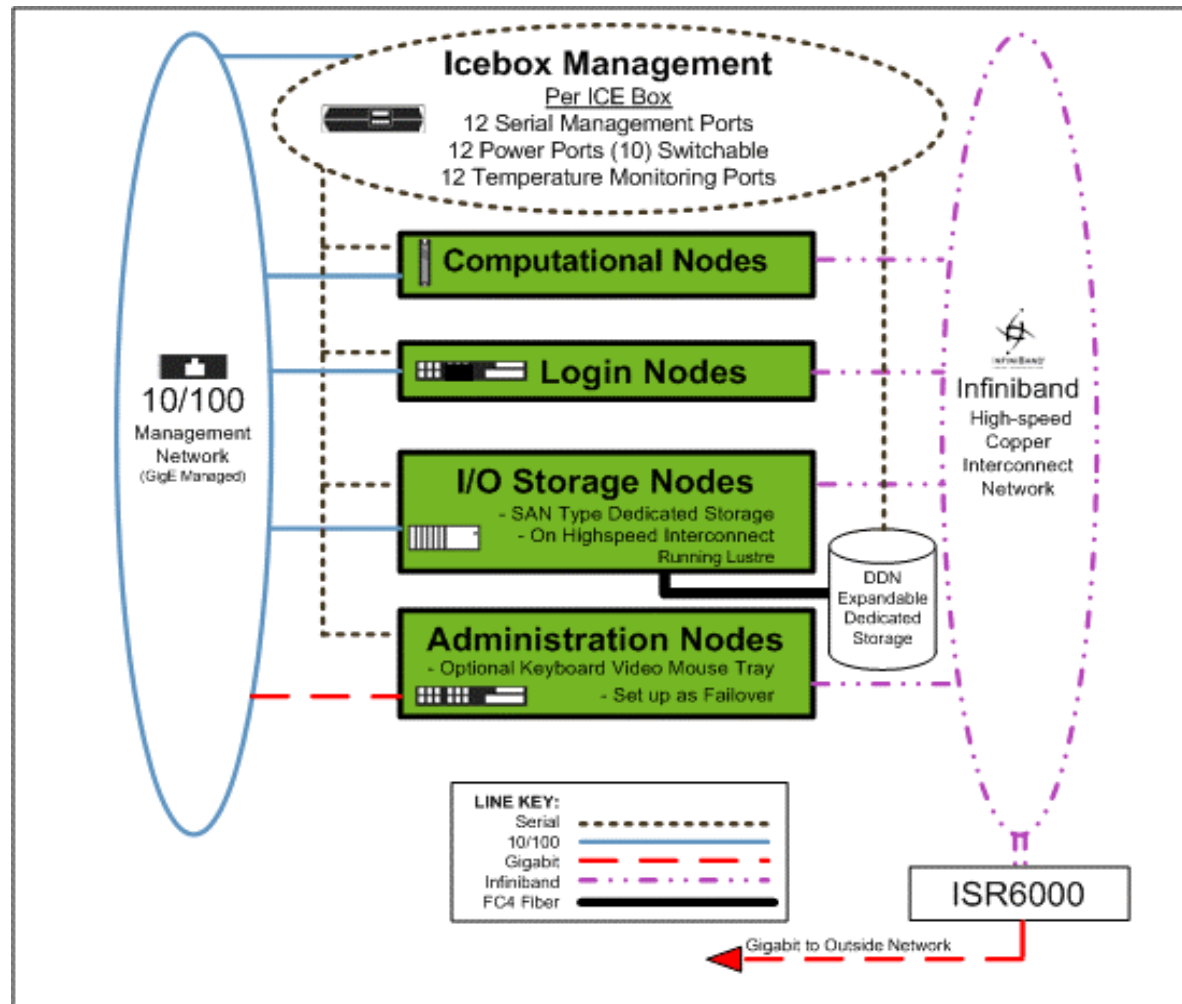
“White Box”
Vendors

Linux Network



Proprietary
Vendors

System Architecture



Node Overview

- **Compute Nodes**
 - ▶ Provide primary computational resource for applications
- **Administration Nodes**
 - ▶ Provide system management functions
 - ▶ Host the system management software
- **Login Nodes**
 - ▶ Provide user access point to cluster
 - ▶ Host authentication services, such as a DoD Kerberos ticketing system
 - ▶ Host development tools for users
- **I/O Storage Nodes**
 - ▶ Lustre MetaData Server (MDS) Nodes manage file names and locations. There are 1 or 2 MDS nodes per cluster.
 - ▶ Lustre Object Storage Target Server (OSS) nodes provide high-speed, scalable access to files

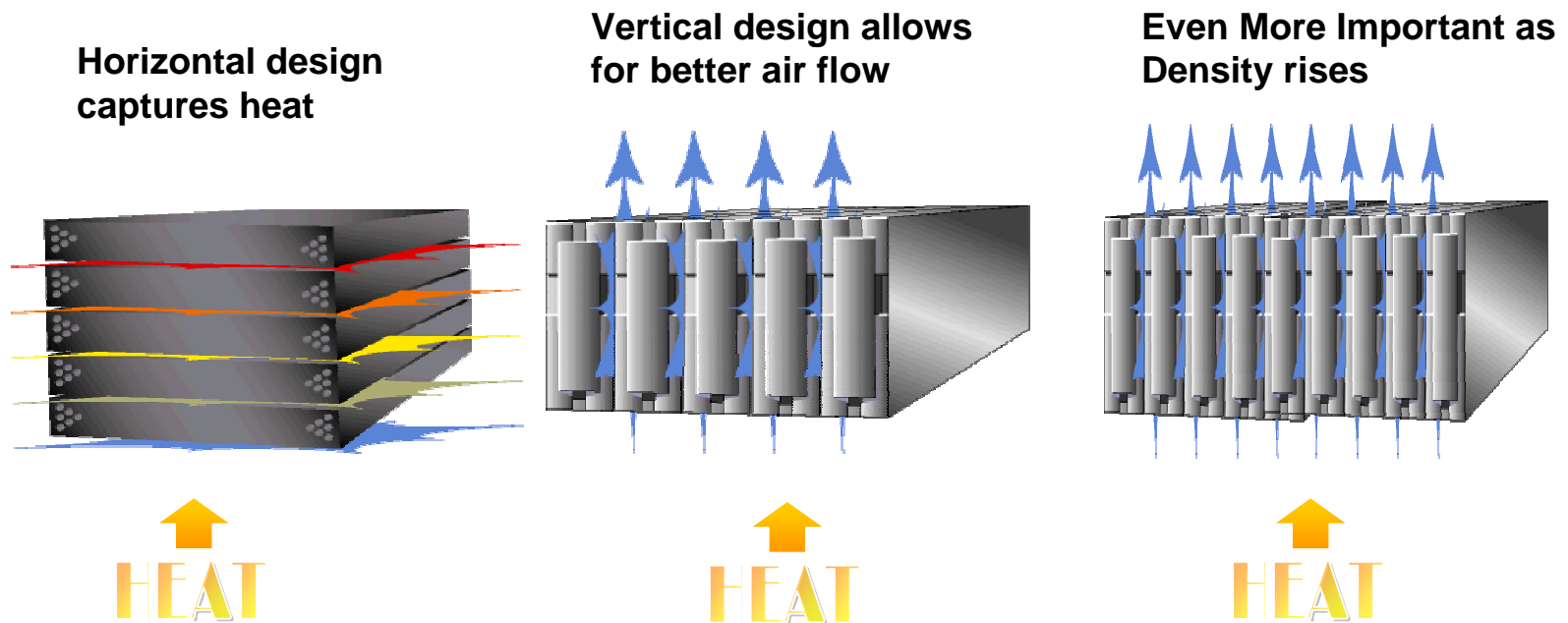
Node Design

- Evolocity II
 - Designed for HPC clustering
 - Compact form factor (0.8U)
 - Up to 50 Nodes per rack
 - Use standard motherboards
 - Optimized for maximum air flow and cooling
 - 2 processor sockets: 2-way or 4-way nodes



Node Design

Vertical rack-mount design benefits system cooling



Evolocity™ decreases temperature by 12° C compared to traditional solution

Processor Options

- Single-core AMD Opteron (64-bit)
- Single-core Intel (64-bit EMT-64)
- Dual-core AMD
- Dual-core Intel

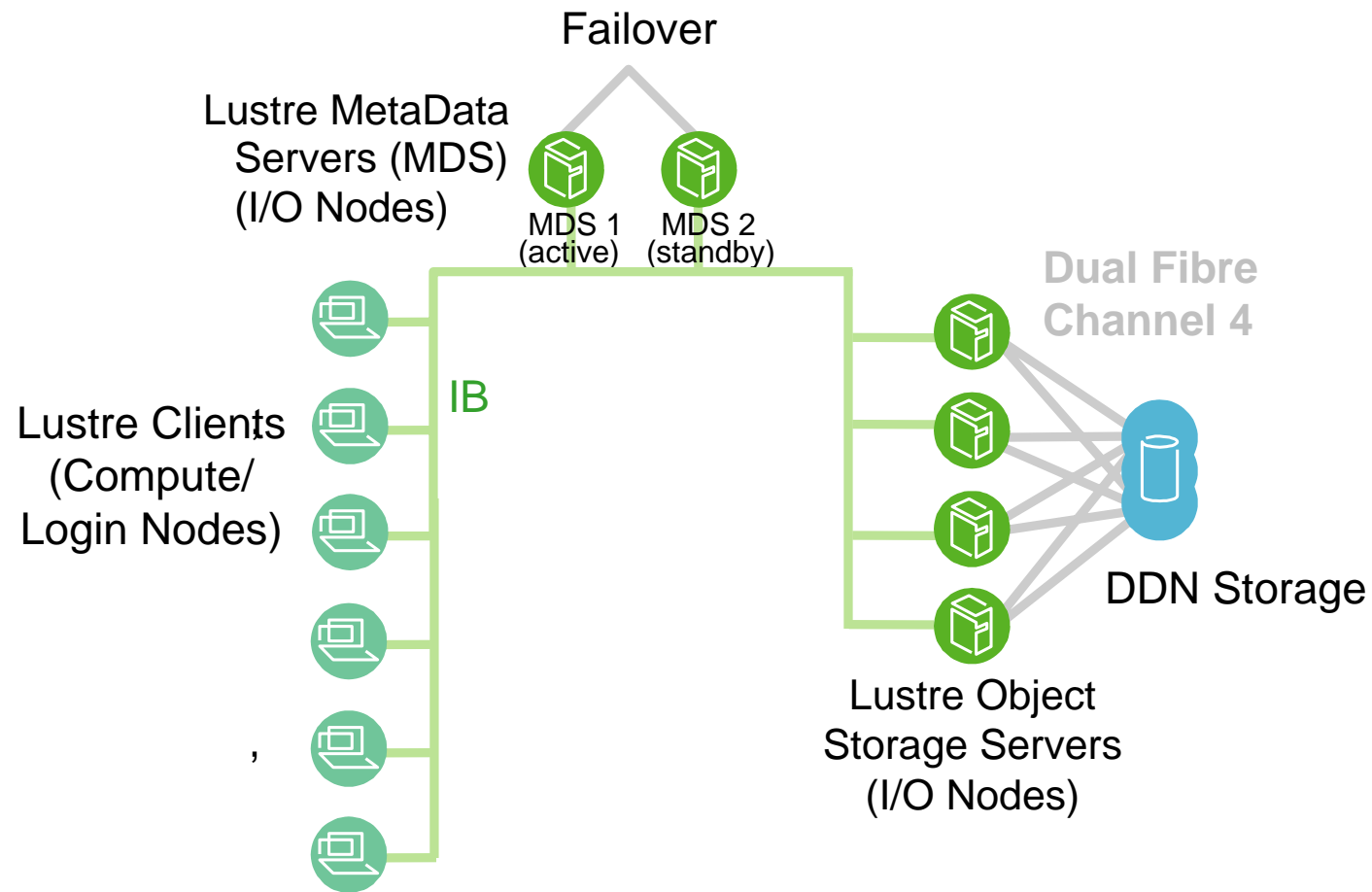
Processor Selection

- Typically we benchmark both AMD and Intel processors. A rough guideline would be:
 - AMD Opteron for memory bandwidth limited codes
 - Intel for codes that have high cache/register residency
- We can fine-tune price/performance through clock speed selection

Interconnect Fabric

- Infiniband
- Myrinet
- Gigabit Ethernet

Lustre Architecture



Lustre Glossary

- Metadata Server (MDS)
 - Exports one or more MetaData Targets (MDTs)
 - Stores file system metadata
- Object Storage Server (OSS)
 - Exports one or more Object Storage Targets (OSTs)
 - Stores file contents
- Lustre Client
 - Writes to and reads from files
 - Client installed on each Compute and Login Node for filesystem access
- Logical Object Volume (LOV)
 - Manages file striping across multiple OSTs
 - Size of stripe and number of OSTs selectable per directory or per file

Facilities, Power, and Cooling

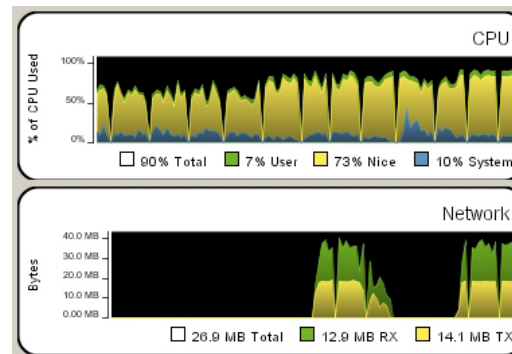
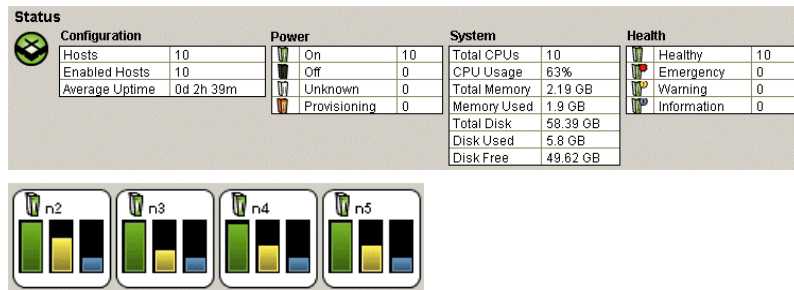
- System engineering approach treating facilities as part of system
- Include all major interfaces
 - Power
 - Thermal/Cooling
 - Physical requirements (space and loading)
 - Networking
 - Logistics
- Linux Networx has wide range of experiences deploying clustering in a variety of data centers
- We work closely with facilities engineers throughout design process
- Linux Networx provides Fluent airflow modeling to ensure proper system cooling at the customer facility
- 208V 3-phase power to racks
 - Most racks require two 50A circuits
- Low-power options are available, with reduced performance
- Temperature-controlled variable speed fans cool the system with reduced noise and airflow

System Management

- Consists of the ICE Box appliance, Clusterworx software, and LinuxBIOS
- Remotely controls each node's power cycling, booting, software image management
- Provides remote BIOS modifications
- Ongoing monitoring of software processes and physical characteristics including temperature, memory ECC, disk drive activity, and fan functionality

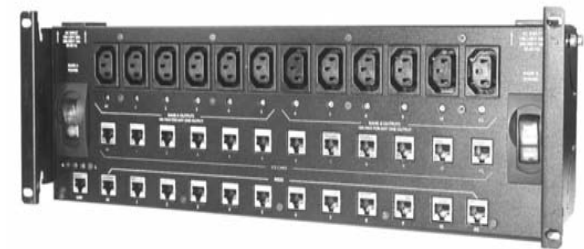
System Management: Clusterworx

- Integrated disk cloning installs the Linux OS and applications on a cluster system of any size in minutes
- Image Manager allows for easy creation of disk or nfsboot images
- Integrated version control system for cluster node payloads
- Historical graphing of system statistics creates dynamically generated charts to show cluster performance
- Monitoring of system properties including CPU usage, memory usage, disk I/O, and network bandwidth
- Automatic administration actions and notification
- Remotely accessible, easy-to-use GUI



System Management: ICE Box

- Designed specifically for HPC clusters
- Serial terminal server and concentrator
- Remote power/reset
- Each ICE Box supports up to 10 nodes and 2 auxiliary devices
- Unique zero U design that mounts to the back of a standard 19" rack.
- Each has a network connection and unique IP address
 - Allows multiple boxes to create a highly scalable IP-based communication
- Ability to monitor temperatures within nodes and adjust fan speeds
- Remotely reset motherboards through internally placed probes
- Emergency beacons
- Fully integrated with Clusterworx® via easy-to-use GUI



System Management: Linux BIOS

- An open source BIOS alternative
 - ▶ Boots quickly
 - ▶ Remotely accessible
 - ▶ Designed specifically for cluster systems
- LinuxBIOS performs the same basic functions as commercial (factory-installed) BIOS only 10-20 times faster.
 - ▶ Initializes the hardware
 - ▶ Checks for valid memory
 - ▶ Begins loading the operating system in about three seconds (Most commercial BIOS require about 30-60 seconds)
- LinuxBIOS can be configured and accessed from within the Linux operating system.
 - ▶ Changes and inspection of the BIOS can be made remotely to a single node or to all the nodes in a cluster system
 - ▶ Can be configured without re-booting
 - ▶ Can be upgraded/written without booting to dos

Software

- SuSe Linux Enterprise Server (SLES) 9 or Redhat
- Compilers
 - Intel
 - OpenMP and auto-parallelization
 - Portland Group (PGI) C/C++/Fortran
 - OpenMP and auto-parallelization
 - HPF
 - Pathscale (Opteron only) C/C++/Fortran compilers
 - OpenMP and auto-parallelization
 - GNU compilers C/C++/Fortran77
- Programming models
 - MPI
 - OpenMP, auto-parallelizing compilers
 - Sockets (IPoIB)
 - Hybrid MPI/OpenMP
 - SysV shared memory (on same node)
 - PVM

Software (cont.)

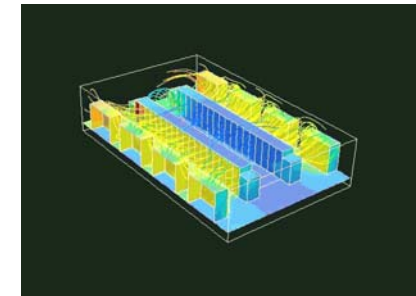
- Libraries
 - Intel MKL: optimized math libraries including BLAS, LAPACK, and FFT
 - AMD ACML
 - FFTW, ATLAS, etc are available
- Debuggers
 - GNU gdb
 - PGDBG Graphical debugger
 - Totalview
- Development Tools
 - Integrated development environments included with SLES
- Profiling tools
 - Intel VTUNE
 - Intel Trace Collector/Intel Trace Analyzer
 - mpiP, MPE, Papi are available
 - PGPROF parallel performance profiler

Batch Schedulers

- LSF
- PBSPro
- OpenPBS
- SLURM
- Checkpoint/Restart is currently being developed

Certified Cluster Services

- Linux cluster training
- Project management
- Comprehensive support
- Datacenter modeling and design
- Site planning
- Application consulting



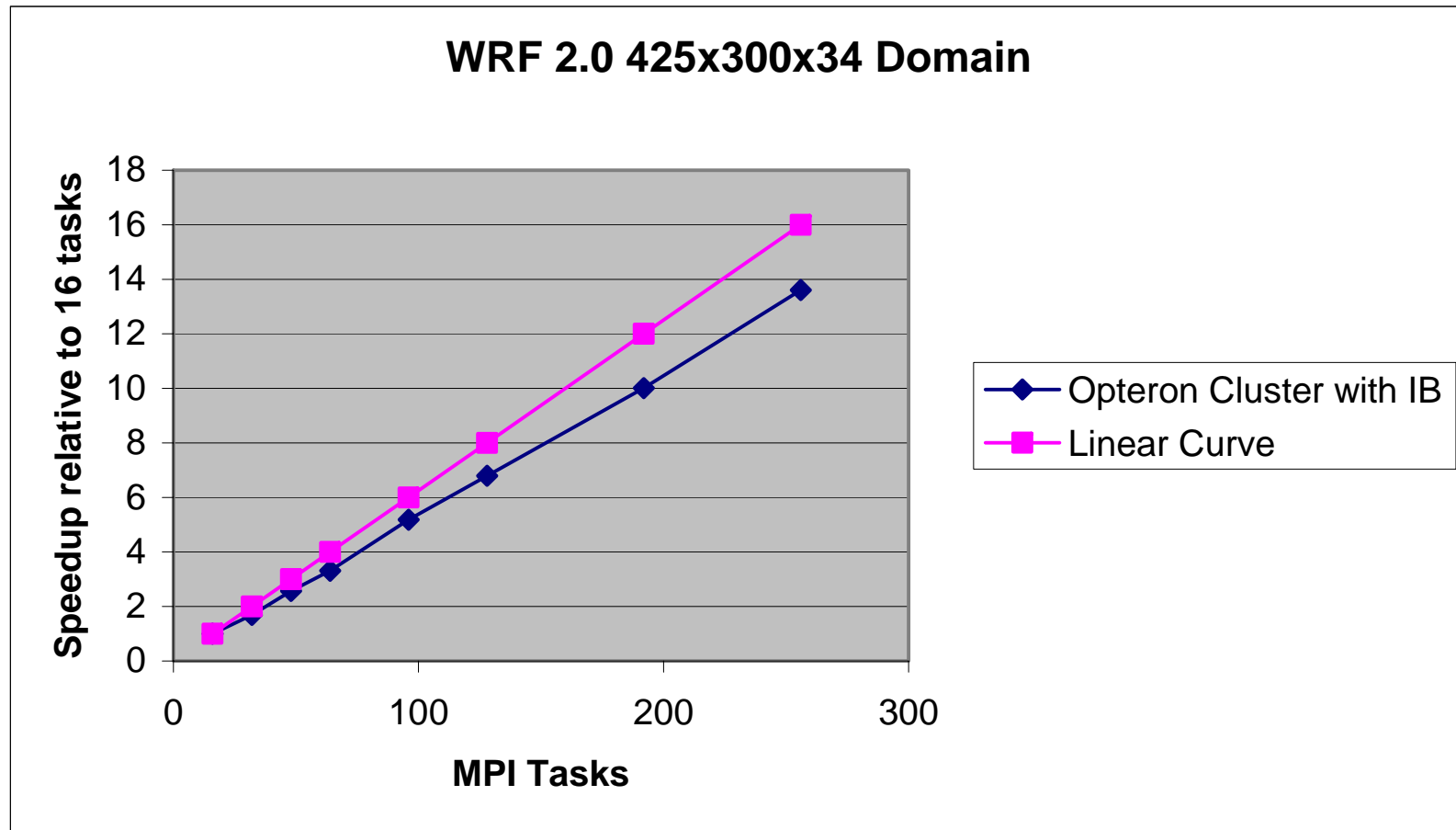
Training Offerings

- Cluster System Fundamentals
- Migrating HPC Applications to Clusters
- Linux Cluster Administration
- Storage and file systems
- Resource and Job Management
- Parallel Programming with MPI

Announcements

- Several new product announcements at SC05 in November
 - Mountaineer
 - Trailblazer

WRF Scalability





Thank You

