

Linux HPC Clusters

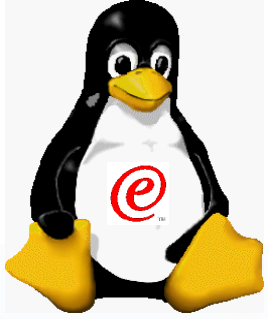
Dan Owsley

Pre-Sales Systems Engineer

Advanced Technical Support

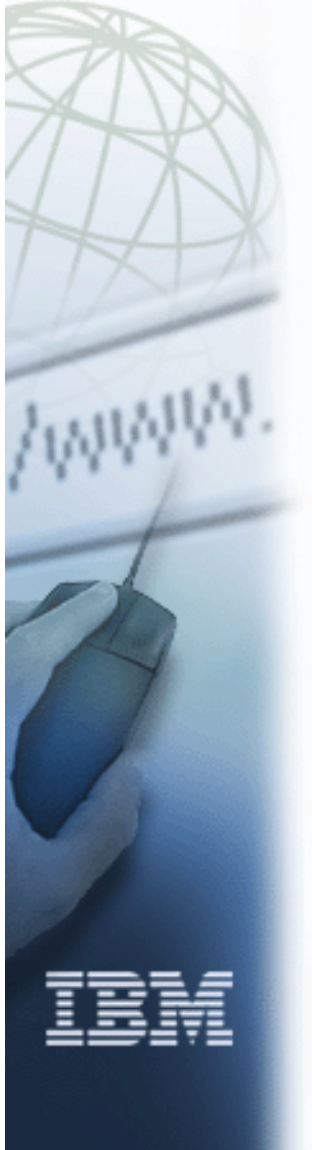
owsleyd@us.ibm.com

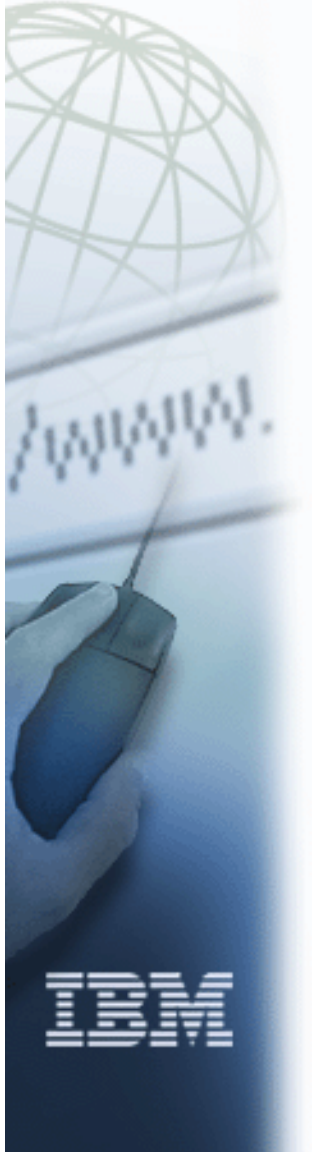
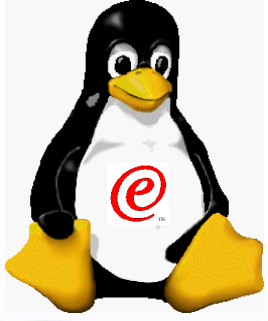
The IBM logo, consisting of the letters 'IBM' in a bold, sans-serif font, with horizontal lines through the letters.



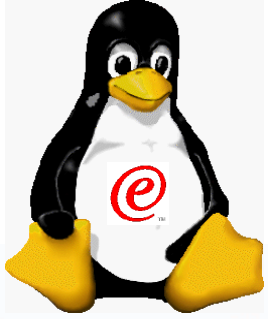
Overview

- Clusters
- Hardware
- Software
- Management
- Resources



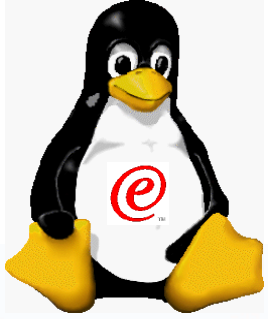


Clusters



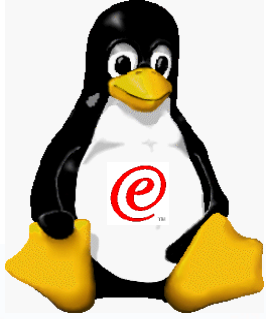
What is a Cluster?

- A cluster is a collection of interconnected computers used as a unified computing resource. (Gregory Pfister - In Search of Clusters)
- Egan's Linux Cluster Definition
 - ▶ It's a Linux Cluster if:
 - >1 interconnected identical nodes (HW and OS)
 - Central Management
 - User views as a single resource
 - Resource/Workload Manager
 - ▶ HPC (Parallel and Batch)
 - ▶ NetGet Farms (Web, DNS, Media, Application, etc...)
 - ▶ HA Clusters
- Clusters can mean different things to different people
 - ▶ High Availability
 - ▶ Scalability
 - ▶ High Performance



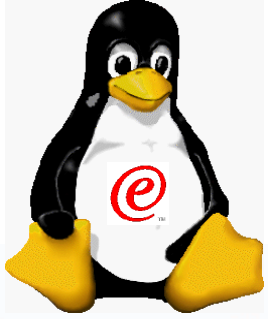
Types of Clusters

- High Availability
 - ▶ Tandem Himalaya
 - ▶ HACMP - RS/6000 clustering
 - ▶ Netware Cluster Services
 - ▶ Microsoft Cluster Server (Wolfpack)
 - ▶ SCO Nonstop Cluster (Linux perhaps?)
- Scalability
 - ▶ Web Clustering/Netgen Farms
 - Round Robin DNS, IP Sprayers, MS Load Balancing
 - ▶ Oracle Parallel Server
 - ▶ DB2/EEE
 - ▶ SAP
- High Performance Computing
 - ▶ SGI Origin 2000, RS/6000 SP2



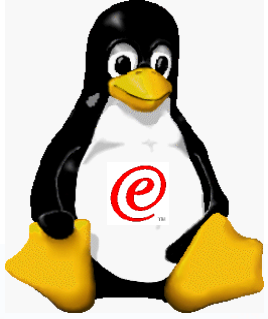
Types of Linux Clusters

- High Availability
 - ▶ Mission Critical Linux
 - ▶ Steeleye
 - ▶ GPFS
- Scalability
 - ▶ Web Clustering/Netgen Farms
 - Round Robin DNS, IP Sprayers
 - Pumpkin Networks
 - Websphere Net.Dispatch
 - ▶ Oracle Parallel Server
 - ▶ DB2/EEE
 - ▶ SAP
- High Performance Computing
 - ▶ Beowulf
 - ▶ Render Farms



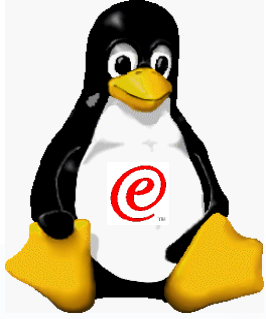
What is a "Beowulf" cluster?

- No strict definition...
- ...But Beowulf-class system generally means
 - ▶ "Commodity" hardware. (COTS, COTW)
 - Typically Intel, but also Alpha, SPARC, others.
 - Dedicated nodes, not a "Network of Workstations"
 - ▶ Must be networked.
 - Ethernet
 - Myrinet
 - ▶ Run parallel programs.
 - ▶ Open Source OS.
 - Typically Linux
- Everybody has one!



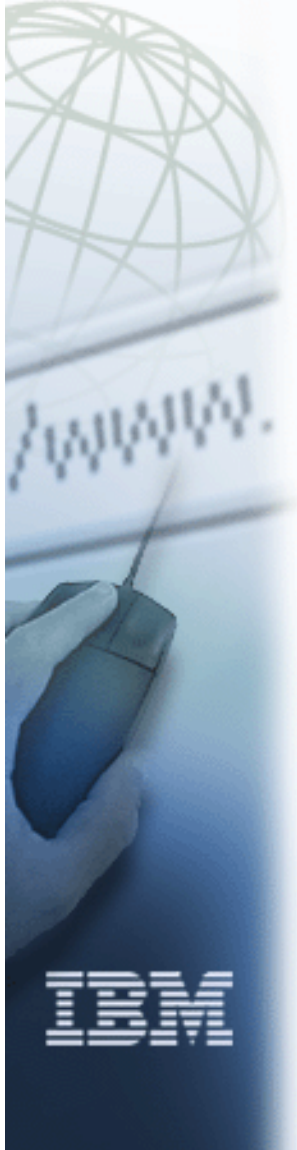
Why Linux?

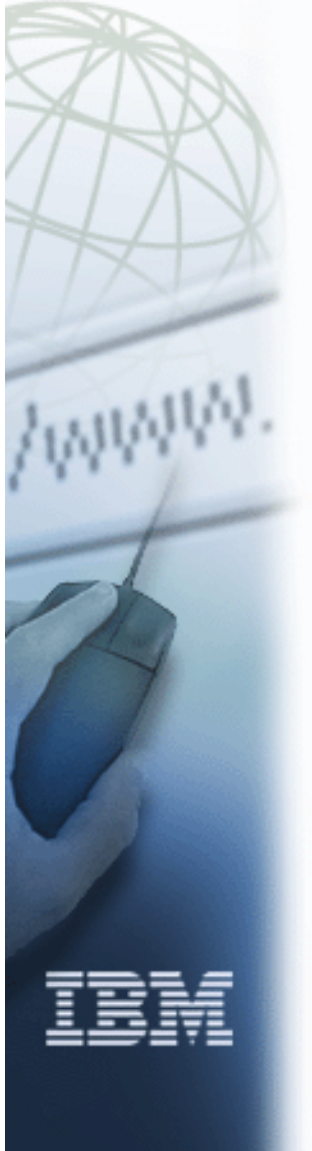
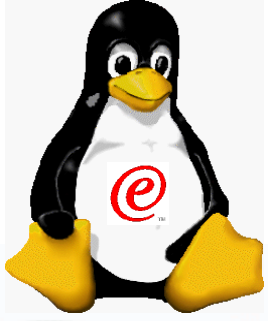
- The reasons are practical, not technical.
 - ▶ Open source, and free
 - ▶ Support for many processor families
 - Intel, Power, Alpha, etc...
 - ▶ Good environment for developing cluster infrastructure
 - ▶ Huge development effort means rapid improvement, support for new hardware.
 - ▶ Commercial applications
 - ▶ Talent pool
- The first four items are advantages over Windows/NT
- Enables the community to optimize the OS to support Super Computer type applications.
 - ▶ PAPI
 - ▶ Performance Counters
 - ▶ MPI TCP Patch



What are Beowulf type applications?

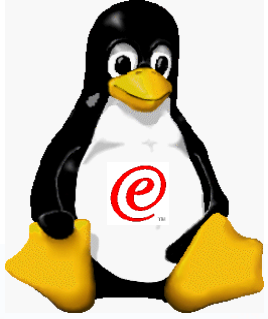
- Must be written to use parallel libraries*
 - ▶ MPI
 - ▶ PVM
- High-energy physics
- Weather Forecasting
- Protein Folding
- Geoscience
- etc...
- *Batch processing counts
 - ▶ Simulation (Intel)





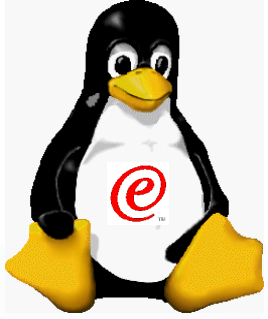
Hardware

Nodes, Network, Infrastructure



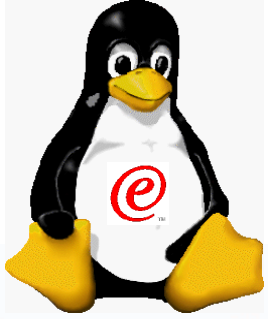
Networks - Heart of the Cluster

- Cluster has many Network Functions
 - ▶ IPC Communication (MPI, PVM)
 - ▶ Filesystems & I/O
 - NFS
 - Parallel filesystems
 - ▶ Management
 - YP/DNS
 - Process startup/management
 - Scheduler
 - Batch processing
- You will probably want to segregate functionality across multiple networks



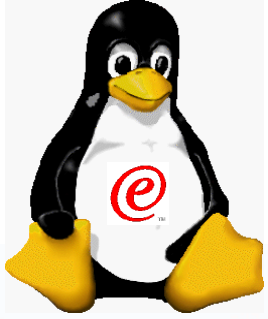
Network Choices

- High Speed IPC Network?
 - ▶ If you need it, you'll know
 - Get out your wallet
 - Prepare to spend up to 30% of budget on network
- I/O Network
 - ▶ Most clusters use either IPC Network or Management Network for I/O
 - ▶ Depend on application needs
- Management
 - ▶ 10/100/1000 Ethernet
 - ▶ Use intelligent switches - worth the cost!
 - SNMP



Networks - High Speed Fabrics

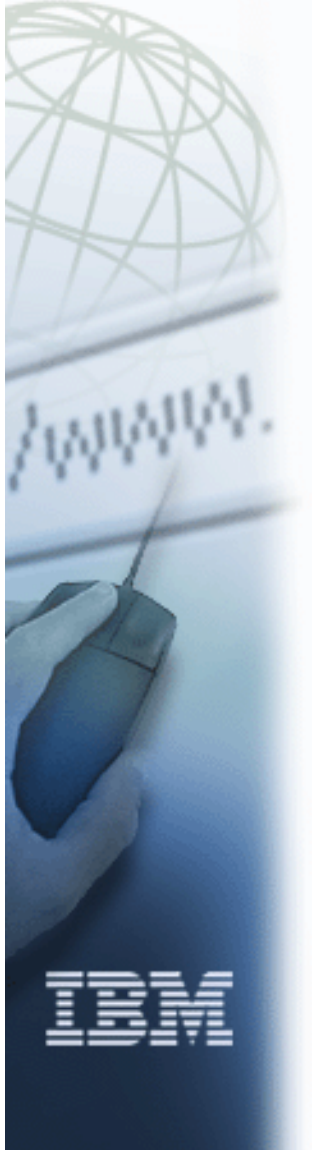
- Select a technology that scales to your needs
 - ▶ Choices - Myrinet, Gigaset, SCI, Quadrics, ServerNet
 - Future - InfiniBand
- Myrinet is by far the market leader - Myricom, Inc.
 - ▶ <http://www.myri.com>
 - ▶ > 200 MB/s
 - ▶ The only choice for building very large networks (large switches, proven scalability) 128-Port switches.
 - ▶ Supports user-space communication software is called “gm”
 - ▶ Programmable processor on NIC
 - ▶ Expect to pay ~\$1500/node
 - ▶ VIA support
 - ▶ IP support

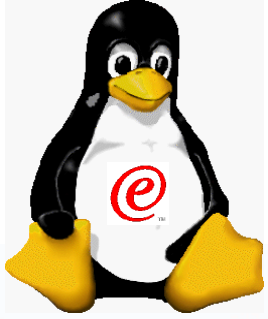


Network Performance

- "Bisection Bandwidth"

- ▶ Bisection Bandwidth is used to report the connectivity of the entire cluster
- ▶ Theoretic Bisection Bandwidth is calculated by dividing the machine in half, using the partition of worst connectivity, and summing the bandwidth of the links between the halves.
- ▶ Interconnection topologies have "Full Bisection Bandwidth" when the number of links between any two halves of the machine is $N/2$.
- ▶ Bisection Bandwidth is important for programs that do global communication





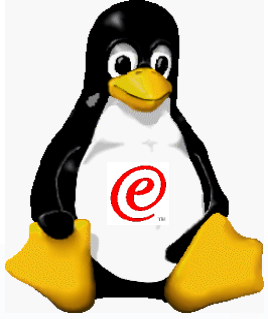
Network Performance

- User Space Communications

- ▶ TCP has many overheads, high latencies
 - Poor transport for MPI-style message passing
- ▶ For best performance, need protected user-space communication system (kernel not involved)
- ▶ Protected user-space communication requires special hardware support
- ▶ Until recently; Only Myrinet; Many API's

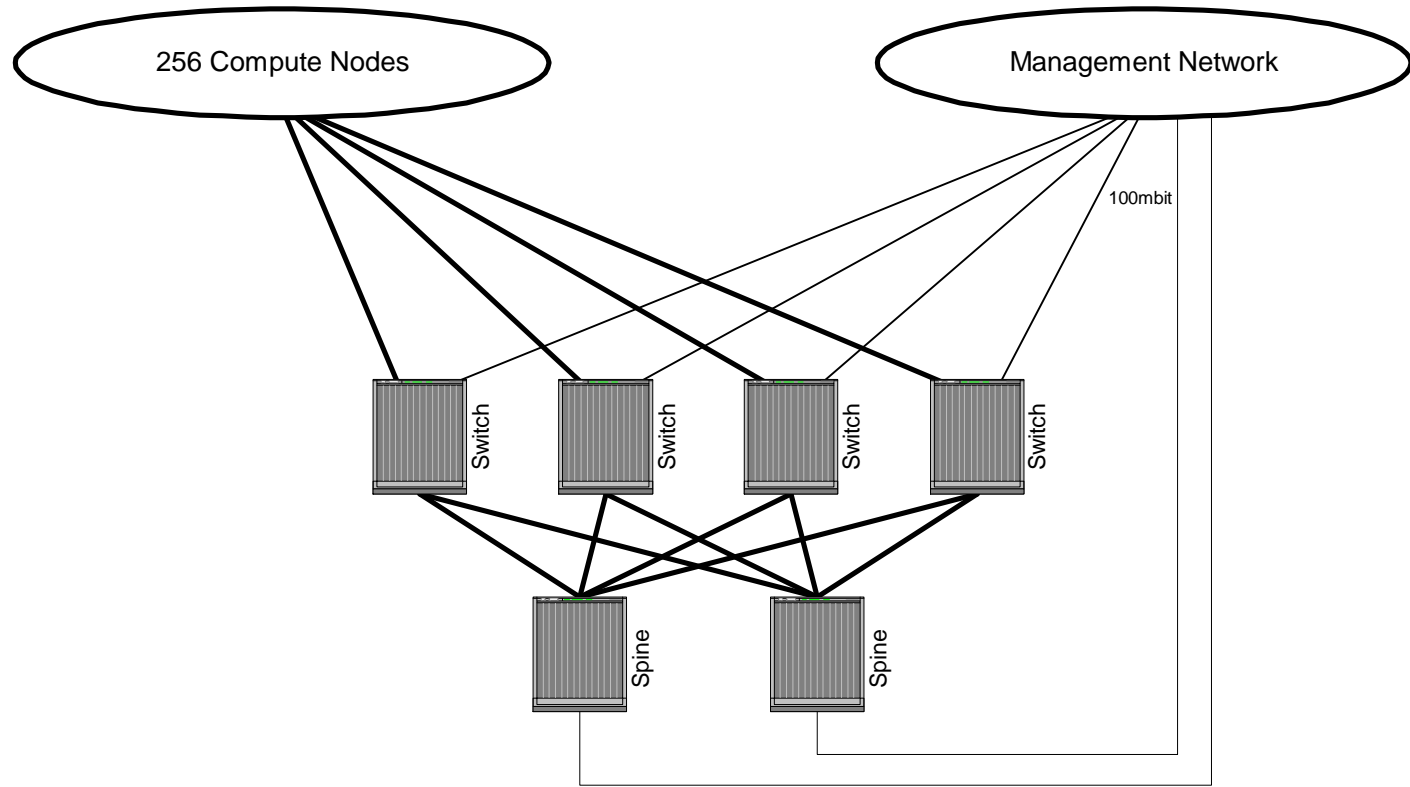
- VIA - Virtual Interface Architecture

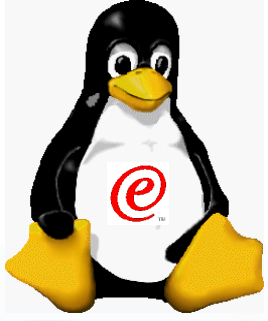
- ▶ New industry standard; used by InfiniBand
- ▶ Single API for any network



Building Big Switches

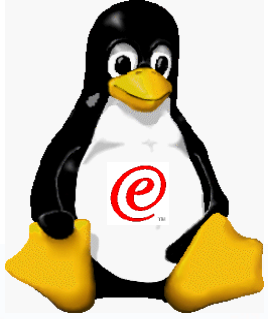
- Scalable networks require switches that can be connected arbitrarily
- Building a $2N$ -port full-bisection switch from N -port full-bisection switches requires $6N$ N -port switches and $2N$ cables!





Compute Nodes

- Form Factor - how many can I fit in a rack?
 - ▶ x330, x340
- System Architecture and implementation can be very important:
 - ▶ PCI busses - how many, how fast
 - ▶ SMP? Memory performance is an issue
 - More CPU's share memory bandwidth
 - ▶ Memory controller, chipset, FSB speed, ECC.
- Integrated Components
 - ▶ Ethernet - PXE?
 - ▶ Management processor?
 - Netfinity Service Processor, Intel EMP port
 - ▶ Floppy? CDROM? HDD?



Non-Compute Nodes

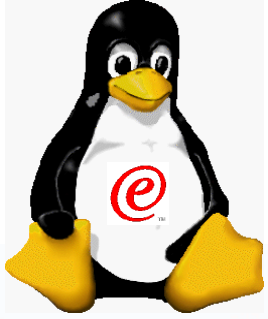
- Management Node

- ▶ High Availability
 - RAID Adapter (ServeRAID)
- ▶ Form factor not as critical, but still want rack optimized
 - x340
- ▶ May be distributed for large clusters

- Front-end/User Node

- ▶ High Availability
 - RAID Adapter (ServeRAID)
- ▶ Form factor not as critical, but still want rack optimized
 - x340
- ▶ Configured as compute node





I/O Nodes

- Priority is throughput, availability, rack density
 - x340
 - ▶ Throughput - multiple gigabit-speed Adapters likely
 - RAID Controllers (ServeRAID)
 - Fibre Channel
 - Gigabit Ethernet
 - Myrinet
 - ▶ Multiple Peer PCI busses, the more the merrier
 - 64-bit, 66Mhz if possible
 - ▶ Availability
 - Internal RAID Controller and mirrored OS drive
 - Redundant Hot-Swap power, fans



SMP or Uni-processor nodes

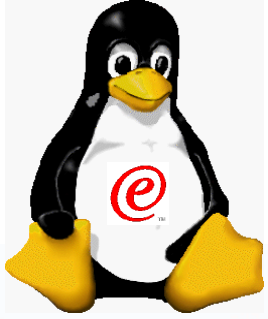
The obvious: Cost per CPU for a 2-way SMP cluster is less

■ Disadvantages to SMP

- ▶ Significantly complicates many layers of software
- ▶ Overall memory bandwidth per CPU is reduced
- ▶ Memory can be more expensive (higher density required)

■ Advantages to SMP

- ▶ Cost of fast interconnect split over 2 CPUs
- ▶ Compact form factor
- ▶ Price / Performance often better than Uni-processor nodes



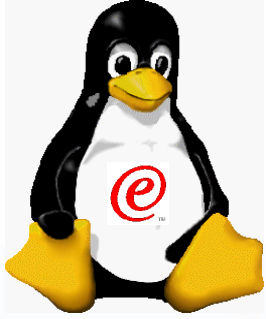
Rack-Optimized Systems

- **1U "Compute Node"**
 - Netfinity x330



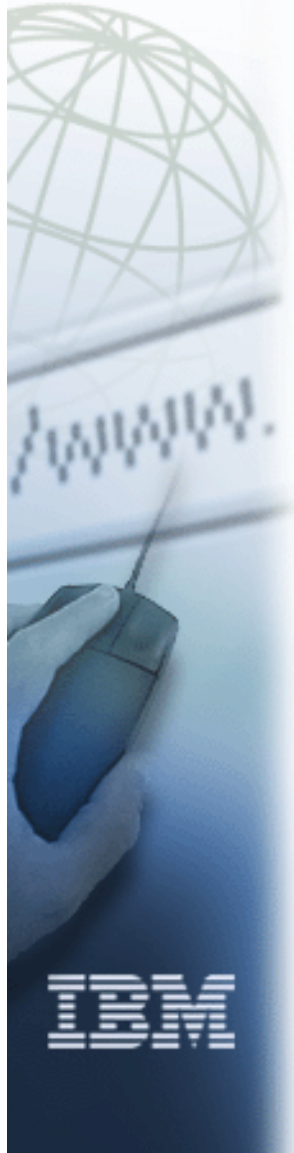
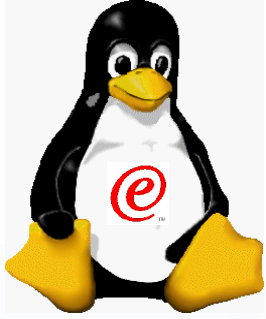
- **3U "Infrastructure Node"**
 - Netfinity x340





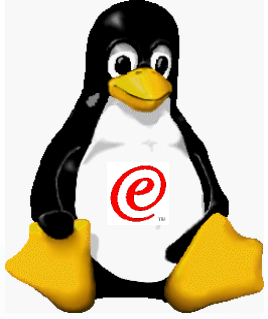
Concrete Example

- Scalable Unit (32 compute nodes, One Management Node)
 - ▶ 1 Management Node
 - Netfinity x340
 - ServeRAID 4L
 - 2 18GB SCSI drives, RAID 1
 - 1 Gigabit Ethernet Adapter
 - ▶ 32 Compute Nodes
 - x330, 1GB RAM
 - 1 9GB HD
 - Myrinet Card
 - ▶ 2 16-port Equinox terminal server
 - ▶ 1 48-port 10/100 Ethernet switch, 2 Gigabit ports
 - ▶ 1 Netfinity Rack
 - ▶ 1 64-port Myrinet Switch
 - 32 for nodes
 - 32 for interconnect

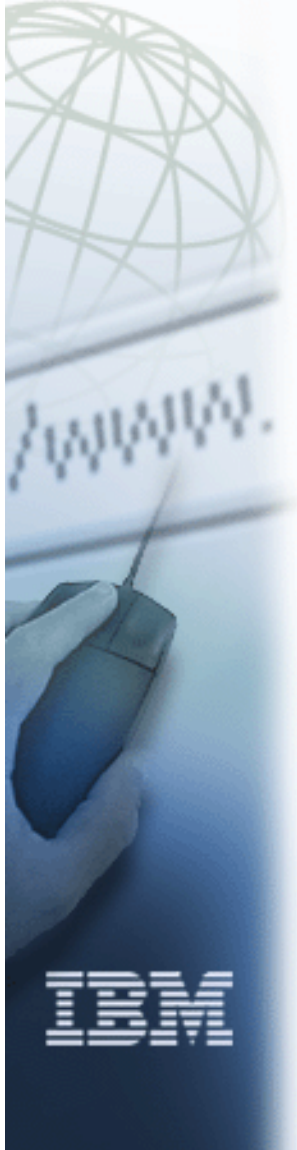


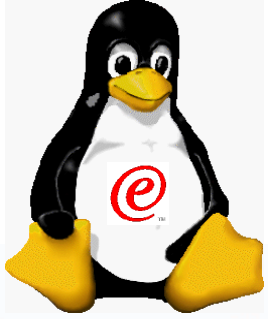
Supporting Infrastructure

- Space
 - ▶ Raised Floor?
- Power
 - ▶ UPS?
 - ▶ Got enough?
- Cooling
- Network Integration
 - ▶ LAN
 - ▶ WAN
- Installation and Maintenance
 - ▶ Have a hardware integrator or lots of free labor!
- Weight

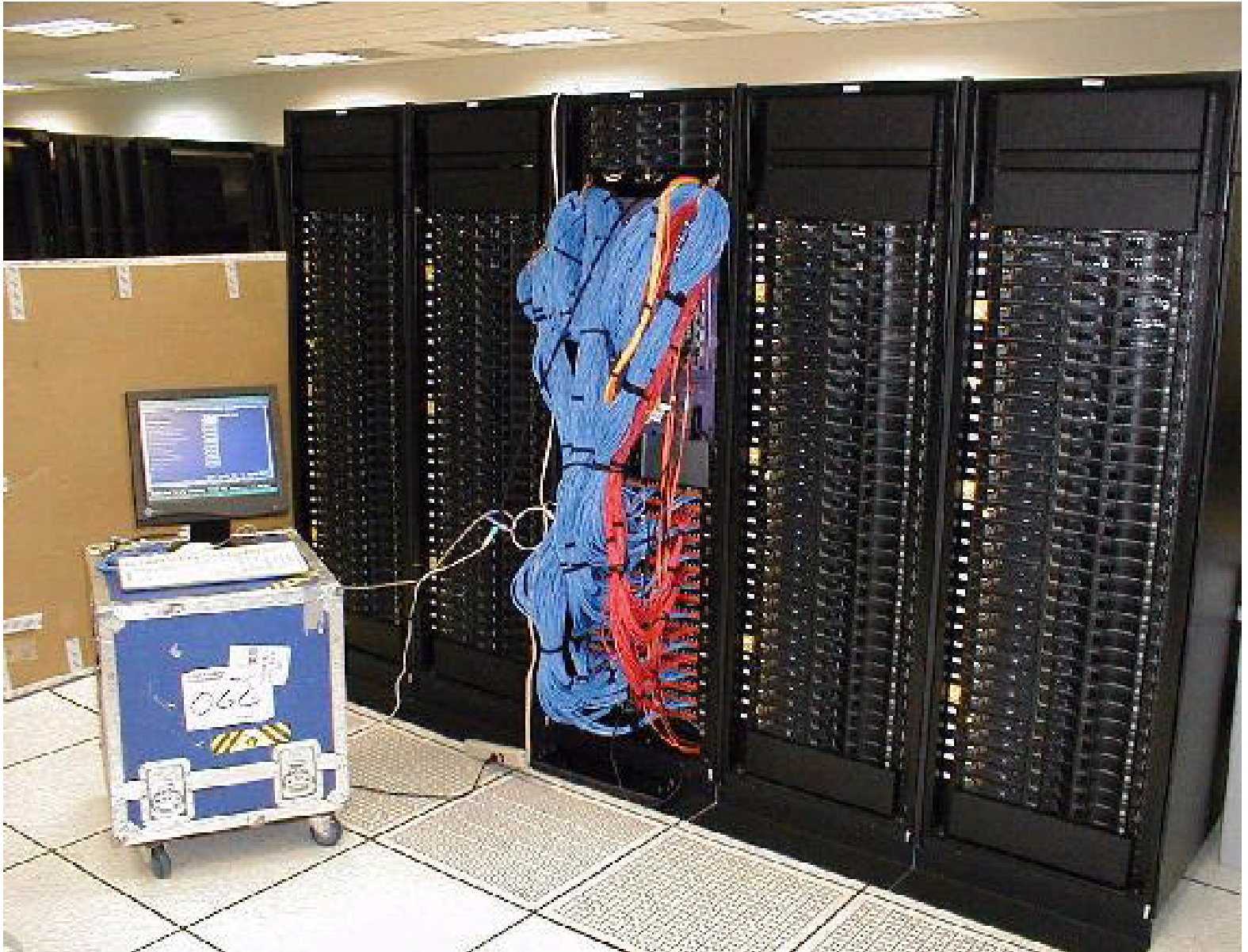


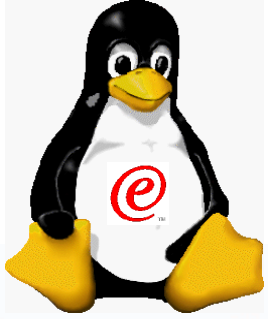
What does a cluster look like?





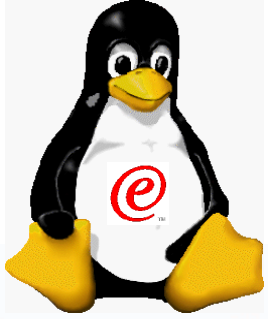
Cables are always a problem



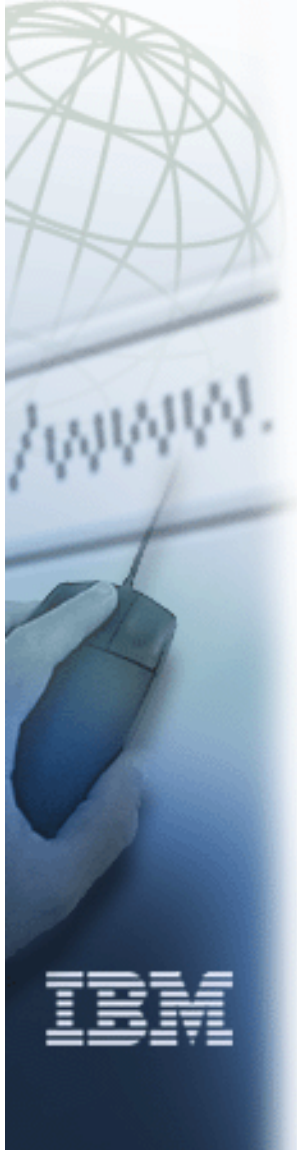
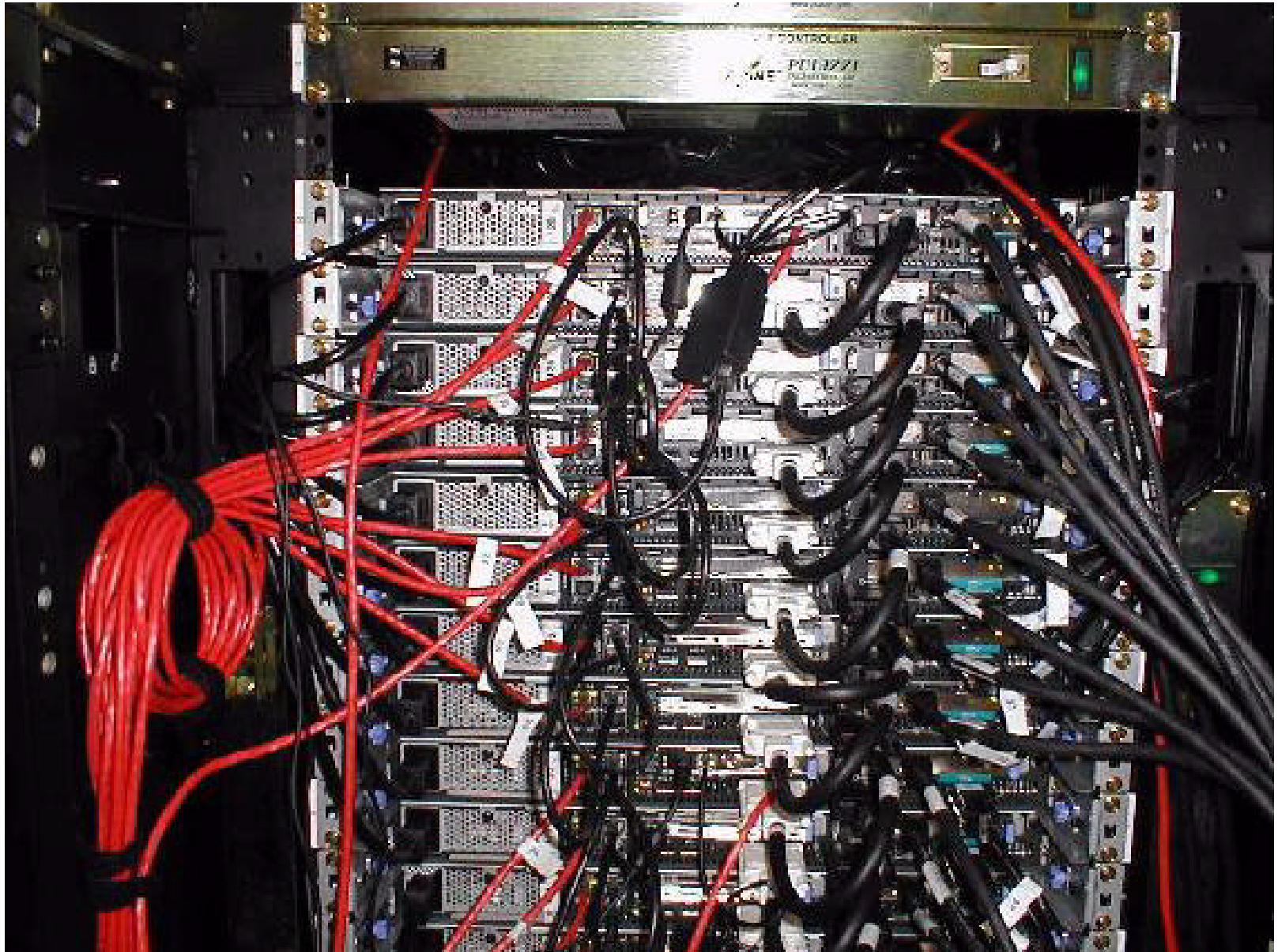


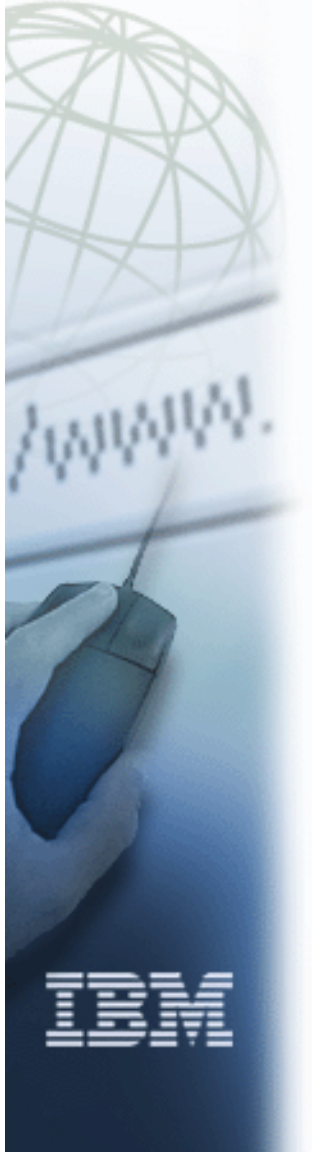
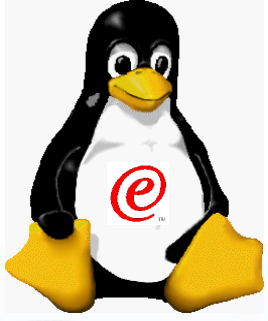
Individual node cables



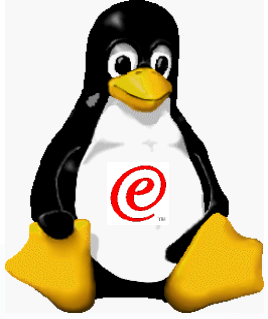


A closer look at cables



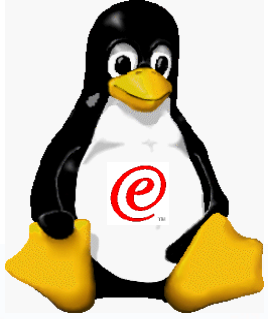


Software



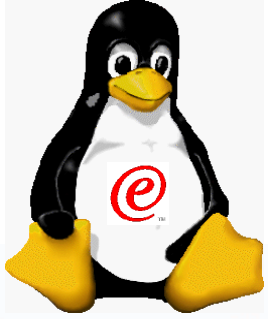
Linux Cluster Software

- Recommend Red Hat Linux (free)
 - ▶ Support (ISV, IHV, Community)
 - ▶ Network Installer
 - ▶ Complete
- Cluster Management
 - ▶ No clear standard exists
 - ▶ Hardware
 - CSM
 - xCAT
 - ▶ Installation
 - xCAT (Any method, Kickstart default)
 - LUI
 - System Imager
 - ▶ Software/OS/User/etc...
 - CSM
 - xCAT



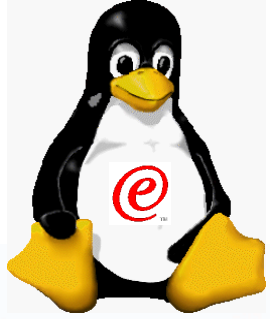
Linux HPC Software

- Messaging Passing
 - ▶ MPI Versions
 - MPICH (free)
 - LAMMPI (free)
 - MPIPro (commercial)
 - ▶ PVM (free)
- Compilers
 - ▶ Portland Group (commercial)
 - ▶ Linux Native (GNU)
- Math Libraries (free/commercial)
 - ▶ BLAS, LAPACK, ScaLAPACK, FFT, etc...
- Applications
 - ▶ BLAST (Free/Commercial)
- Benchmarks
 - ▶ Linpack/NPB



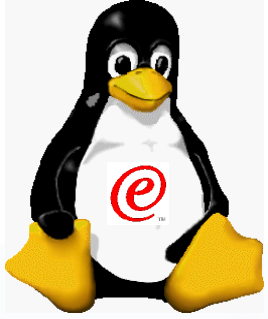
Linux HPC Software

- Resource Managers/Schedulers
 - ▶ PBS--Portable Batch System (free)
 - ▶ Maui Scheduler
 - ▶ LSF--Load Sharing Facility (commercial)
 - ▶ IBM Load Leveler (TBD)
- File Systems
 - ▶ GFS
 - Limited Hardware Support
 - Not parallel
 - Idea for Non-compute nodes
 - ▶ PVFS (free)
 - Research Project
 - Still under development
 - ▶ IBM GPFS (commercial)
 - Announced in June, 01



Management

Systems Management, Administration,
Control



Management Requirements

- Minimum

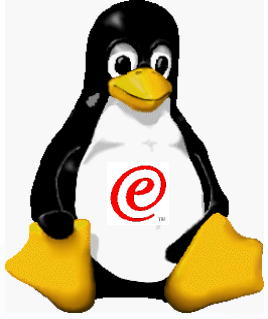
- ▶ Power Control
- ▶ Remote OS Console
- ▶ Automated Network Installation

- Desired

- ▶ Remote Reset (Soft and Hard)
- ▶ Remote Inventory (Serial number/Model/CPU/Memory/Disk/PCI)
- ▶ Remote BIOS/POST Console
- ▶ Remote Vitals (Fan speed/Temp/etc...)
- ▶ Remote Hardware Event Logs
- ▶ SNMP Hardware Alerts

- Other

- ▶ User/Security
- ▶ Queue/Accounting
- ▶ Software



SPNs

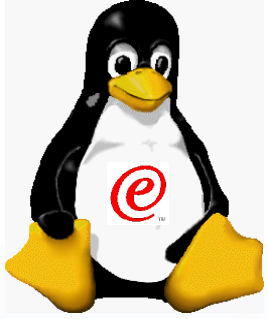
- Service Processor Networks

- ▶ OS Independent
- ▶ No software required
- ▶ Scriptable Interface (Telnet/HTTP)

- Functions

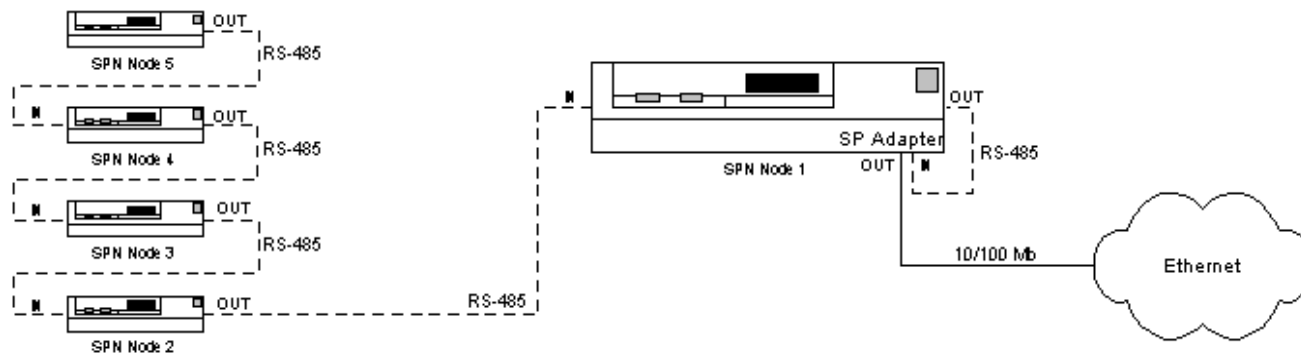
- ▶ Remote Power
- ▶ Remote Reset (Soft and Hard)
- ▶ Remote Inventory (Serial number/Model/CPU/Memory/Disk/PCI)
- ▶ Remote BIOS/POST Console
- ▶ Remote Vitals (Fan speed/Temp/etc...)
- ▶ Remote Hardware Event Logs
- ▶ SNMP Hardware Alerts

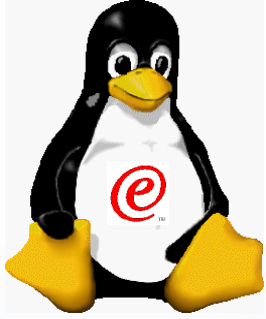
The IBM logo, consisting of the letters 'IBM' in a stylized, striped font, is located in the bottom left corner of the slide.



SPNs

- Design Requirements
 - ▶ 11 Nodes/SPN
 - ▶ One PCI ASMA Card required





Terminal Servers & Network Install

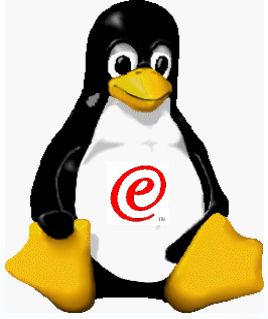
- Terminal Servers

- ▶ Provides Remote OS Console
- ▶ Often Required
- ▶ Recommended Solutions
 - Equinox ELS-16

- Network Install

- ▶ PXE
 - Provides Network Installation
 - Hardware and/or BIOS update may be required
 - Native on x330
 - IBM EtherJet II w/ flash
- ▶ Etherboot
 - For non-PXE systems

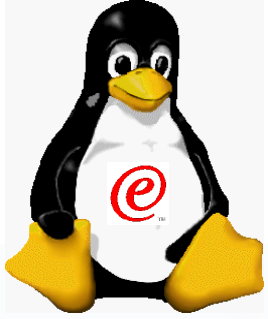




xCAT

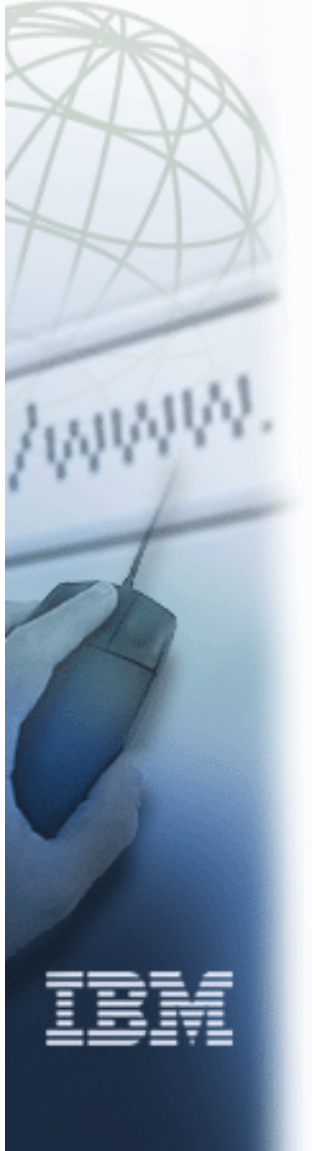
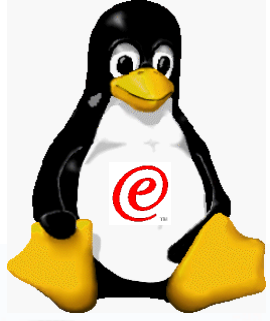
■ Features

- ▶ Remote power control (on/off/state)
- ▶ Remote hardware & software reset (Ctrl+Alt+Del)
- ▶ Remote OS/POST/BIOS console
- ▶ Remote vitals (fan speed/temp/etc...)
- ▶ Remote hardware event logs
- ▶ Remote hardware inventory
- ▶ Parallel remote shell/copy
- ▶ Command line interface (no GUI)
- ▶ Single operations can be applied in parallel to multiple nodes
- ▶ Network installation (PXE/Etherboot)
- ▶ Support for various user defined node types
- ▶ SNMP hardware alerts
- ▶ All scripts! No C code.



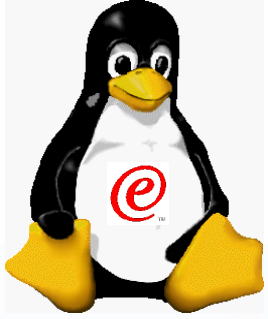
xCAT (continued)

- xCAT is Glue
 - ▶ OSS Requirements
 - Many
- HW Support
 - ▶ All Netfinity with Ranger/Wiseman
 - ▶ All Netfinity with Falcon (in 1.1)
 - ▶ APC Master Switch
 - ▶ Any Terminal Server
 - ▶ Intel/AMD NICs
- Documentation
 - ▶ Red Book - <http://www.redbooks.ibm.com/redbooks/SG246041.html>
 - ▶ HTML Documentation
 - ▶ Man Pages (maybe)



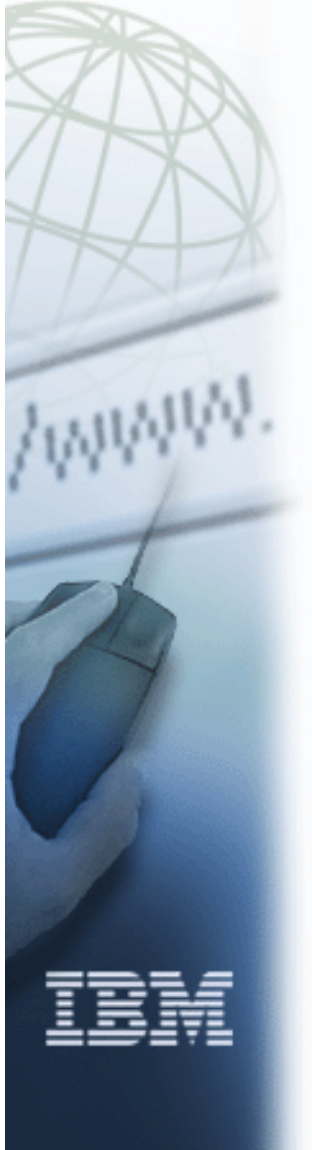
Resources

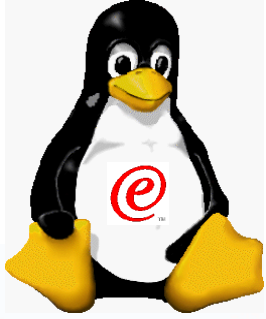
Where to find stuff



Information Concealed in Books

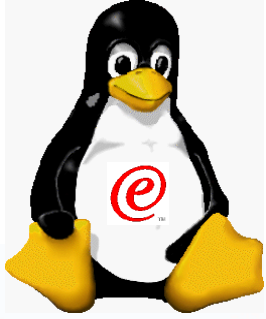
- In Search of Clusters, 2nd Ed (required reading)
- How to Build a Beowulf





Beowulf Software

- MPI - Message Passing Interface <http://www-unix.mcs.anl.gov/mpi/>
 - ▶ MPICH - defacto standard
 - <http://www-unix.mcs.anl.gov/mpi/mpich/>
 - ▶ LAM MPI - TCP/IP only
 - <http://www.mpi.nd.edu/lam/>
 - ▶ MPI/Pro - "Thread-friendly" MPI
 - <http://www.mpi-softtech.com/>
- PVM - Parallel Virtual Machine
 - ▶ http://www.epm.ornl.gov/pvm/pvm_home.html
- VIA - Virtual Interface Architecture
 - ▶ VIA over Myrinet available at:
 - ▶ <http://www.millennium.berkeley.edu/~philipb/via/>
- GM
 - ▶ <http://www.myri.com>



Beowulf Software

- Resource Managers/Schedulers

- ▶ PBS - Portable Batch System

- "Open Source" - but not GPL
 - <http://www.openpbs.org>

- ▶ Maui

- <http://www.supercluster.org>

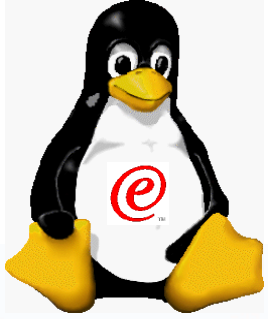
- ▶ LSF - Load Sharing Facility

- Commercial product, widely used
 - <http://www.platform.com/>

- ▶ LoadLeveler - IBM scheduler from RS/6000 SP

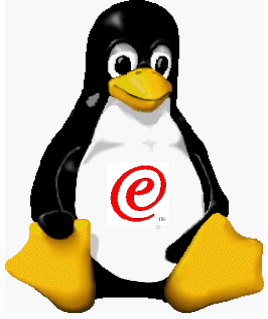
- Coming "Soon"





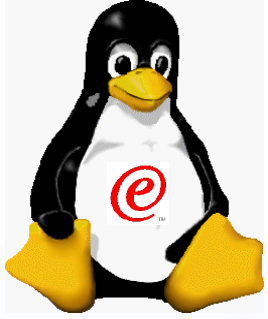
Beowulf Software

- PVFS - Parallel Virtual File System
 - ▶ Open Source parallel filesystem
 - ▶ Still considered "researchy" by some
 - ▶ Does not provide fault tolerance in case of node failure
 - ▶ <http://parlweb.parl.clemson.edu/pvfs/>
- GFS - Global File System
 - ▶ <http://www.globalfilesystem.org/>
- GPFS
 - ▶ <http://www-1.ibm.com/servers/eserver/clusters/software/gpfs.html>



Beowulf Information

- Beowulf Home page
 - ▶ <http://www.beowulf.org/>
- Beowulf-Underground
 - ▶ <http://www.beowulf-underground.org>
- Beowulf Mailing Lists
 - ▶ <http://www.beowulf.org/listarchives/>
 - ▶ Very high signal/noise ratio
- Extreme Linux
 - ▶ <http://www.extremelinux.org/>
- PARL - Parallel Architecture Research Lab
 - ▶ <http://www.parl.clemson.edu/>
- Top500
 - ▶ top500.org
 - ▶ clusters.top500.org



Questions?

- Dan Owsley
 - ▶ owsleyd@us.ibm.com
 - ▶ 714-438-5327

