Datacenter Fabric Workshop

Windows IB

Introduction to
Windows 2003 Compute Cluster Edition

Eric Lantz

Microsoft

elantz@microsoft.com

August 22, 2005

What this talk is **not** about…

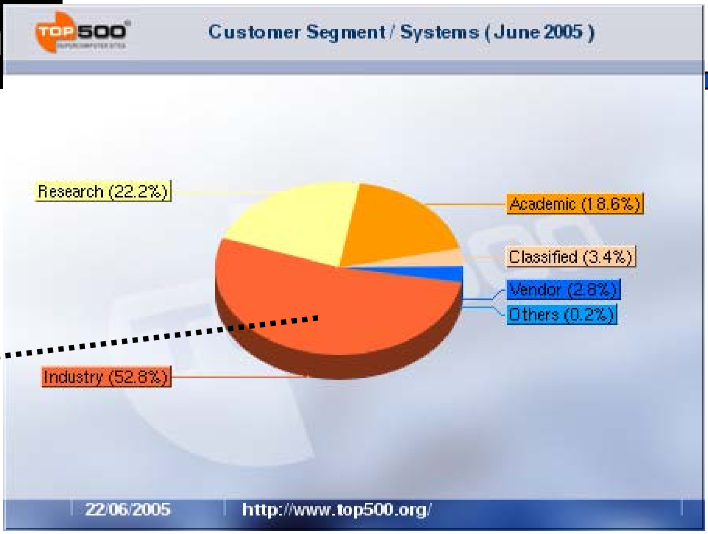
- High Availability, Fail-over clustering
- Scaling out general business applications (ie Exchange, SQL, SAP, etc)

Agenda

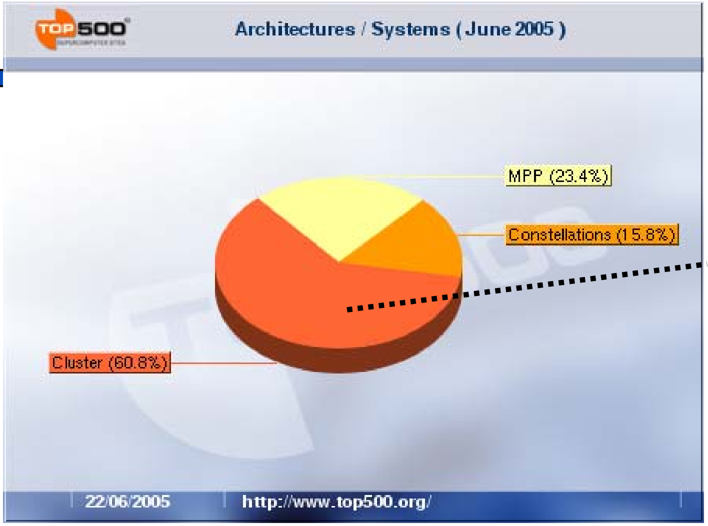
- HPC Market Definition & Trends
- Microsoft's Compute Cluster Solution
- CCE Key Features
  - Deployment
  - Job Scheduling
  - MPI & Networking
  - Development Tools
- Q&A

HPC Market Definition & Trends

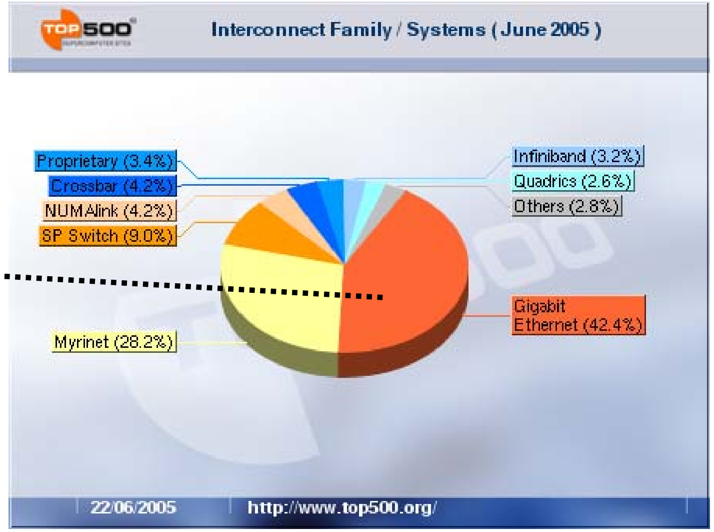
Top 500 Supercomputer Trends



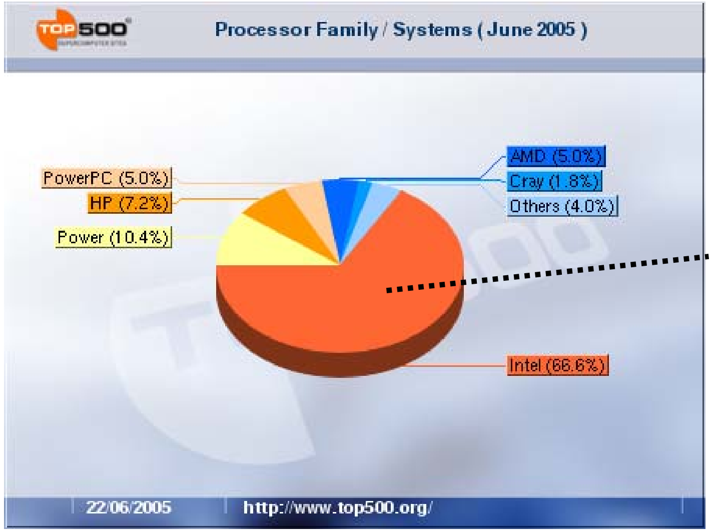
Industry usage rising



Clusters over 60%



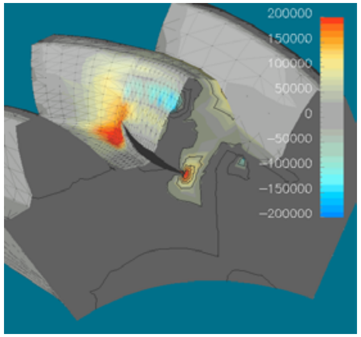
GigE is leading, IB is growing

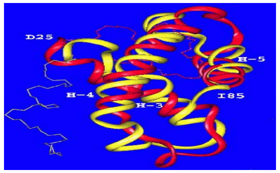


IA is winning

HPC Systems are Affecting Every Vertical…

- Leverage Volume Markets of **Industry Standard Hardware** and Software
- Rapid Procurement, Installation and Integration of systems
- Cluster Ready Applications Accelerating Market Growth
  - Engineering
  - Bioinformatics
  - Oil & Gas
  - Finance
  - Government



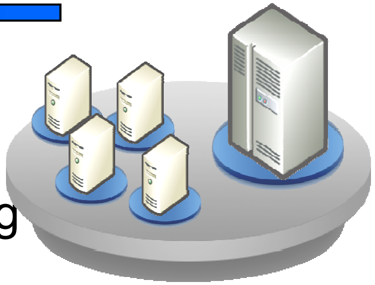


The convergence of affordable high performance hardware and commercial apps is making supercomputing personal

Supercomputing Goes Personal

System	Cray Y-MP C916 	Sun HPC10000 	Shuttle @ NewEgg.com
Architecture	16 x Vector 4GB, Bus	24 x 333MHz Ultra-SPARCII, 24GB, SBus	4 x 2.2GHz Athlon64 4GB, GigE
OS	UNICOS	Solaris 2.5.1	Windows Server 2003 SP1
GFlops	**~10**	**~10**	**~10**
Top500 #	1	500	N/A
Price	$40,000,000	$1,000,000 (40x drop)	< $4,000 (250x drop)
Customers	Government Labs	Large Enterprises	Every Engineer & Scientist
Applications	Classified, Climate, Physics Research	Manufacturing, Energy, Finance, Telecom	Bioinformatics, Materials Sciences, Digital Media

Solution Requirements



Customers require:

- An integrated supported solution stack
- Simplified job submission, status and progress monitoring
- Maximum compute performance and scalability
- Simplified environment from desktops to HPC clusters

Administrators require:

- Better cluster monitoring and management for maximum resource utilization
- Flexible, extensible, policy-driven job scheduling and resource allocation
- Maximum node uptime
- Secure process startup and complete cleanup

Developers Require:

- Programming environment that enables maximum productivity
- Availability and optimized compilers (Fortran) and math libraries
- Parallel debugger, profiler, and visualization tools
- Parallel programming models (MPI)

Microsoft Compute Cluster Solution

Key Scenarios

Departmental Cluster:

Conventional scenario

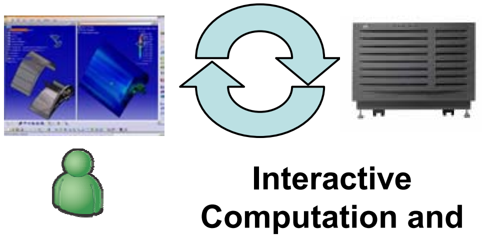
- IT owns large clusters due to complexity and allocates resources on per job basis
- Users submit batch jobs via scripts
- In-house and ISV apps, many based on MPI
- Very poor development tools

Manual, batch execution

Personal/Workgroup Cluster:

Emerging scenario

- Clusters are pre-packaged OEM appliances, purchased and managed by end-users
- Desktop HPC applications transparently and interactively make use of cluster resources
- Desktop development tools integration



Interactive Computation and Visualization

Design Goals

- Designing for (Corporate/Engineering IT)
- Appliance-like setup experience
  - Clear, prescriptive setup guidance
  - Simple deployment of head node and compute nodes
  - Minimized complexity for corporate IT integration
  - Most operations are scriptable
- Leverage existing infrastructure
  - Use Active Directory for user, resource and access management
  - Secure execution, resource access, management
  - Allow customers to use existing deployment tools

Compute Cluster Solution

Mission: Deliver the easiest to deploy and most cost effective solution for solving scaled-out business, engineering and scientific computational problems.

Windows Server 2003, Compute Cluster Edition + **Compute Cluster Pack** = **Microsoft Compute Cluster Solution**

Windows Server 2003, Compute Cluster Edition

- Support for high performance hardware (x64bit architecture)

Compute Cluster Pack

- Support for Industry Standards MPI2, RDMA on Ethernet & Infiniband
- Integrated Job Scheduler
- Cluster Resource Management Tools

Microsoft Compute Cluster Solution

- Integrated Solution out-of-the-box
- Leverages investment in Windows administration and tools
- Makes cluster operation easy and secure as a single system

CCS Key Features

Integration with existing Windows and management infrastructure

- Integrates with AD, Windows security and existing systems management and deployment tools

Node Deployment and Administration

- Compute nodes automatically imaged and added to cluster
- Node Management through UI and command line
- To Do List to configure head node

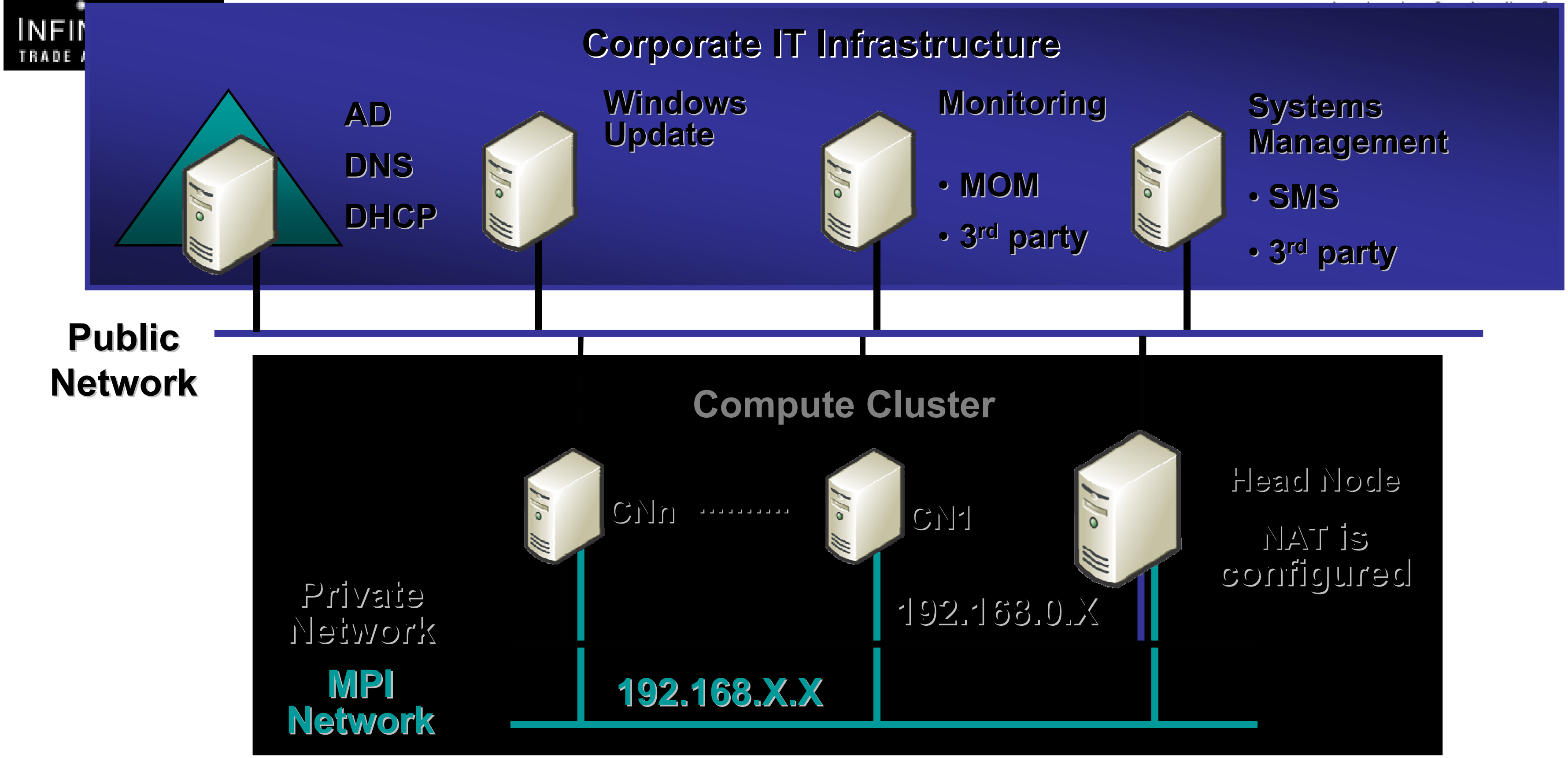
Extensible job scheduler

- 3rd party extensibility at job submission and/or job assignment
- Examples: admission policies and license verification
- Submit jobs from command line, UI, or directly from applications
- Simple management, similar to print queue management

Secure MPI

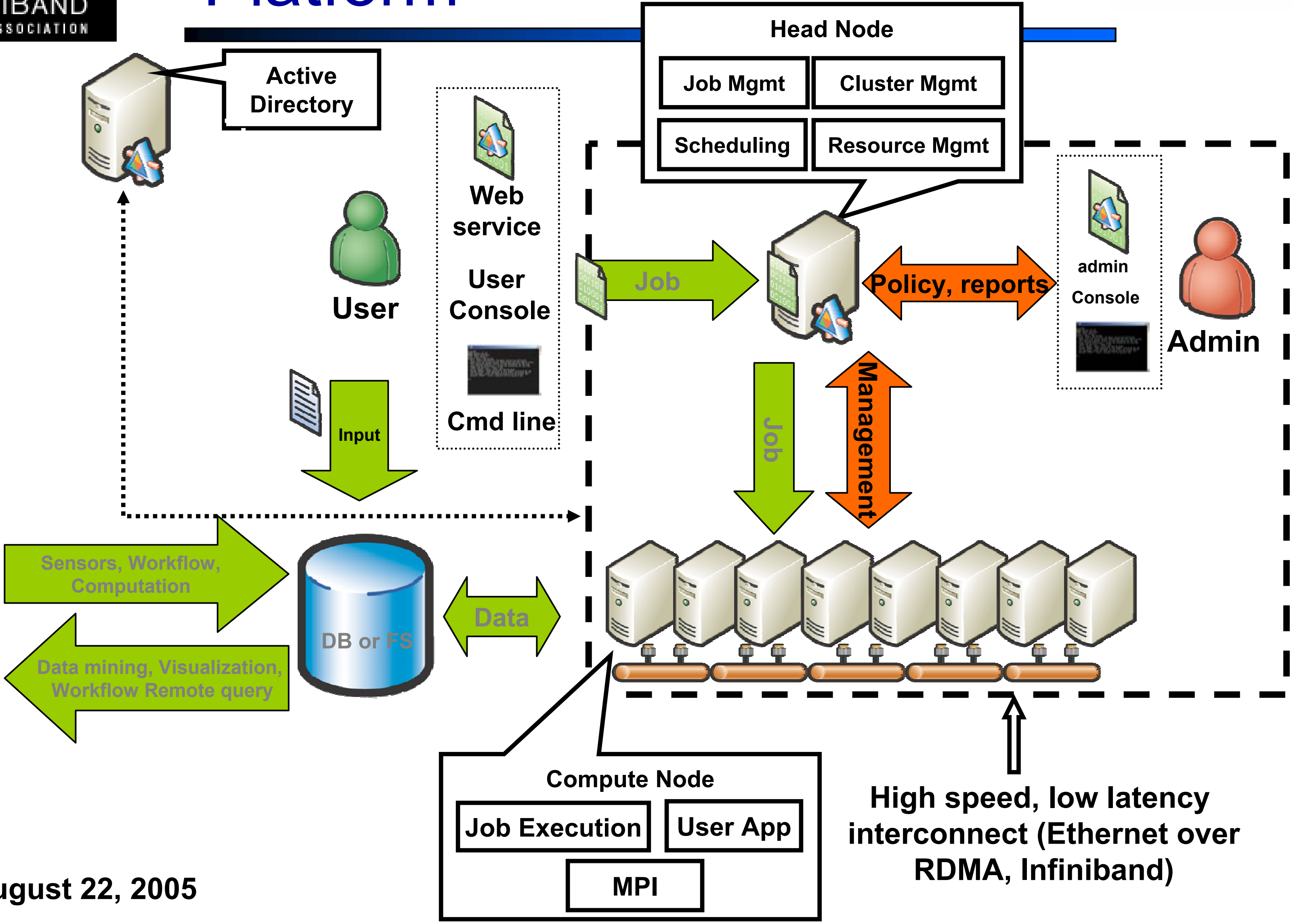
- User credentials secured in job scheduler and compute nodes
- Standardized MPI stack
- Microsoft provided stack reduces application/MPI incompatibility issues

Typical Cluster Topology



- Compute Nodes and Head Node are member servers in a domain in a corp Active Directory
- Public Network: Required for connectivity with existing corp network
- Private Network: Required to separate cluster management and deployment traffic
- MPI Network: Optional high-speed interconnect network (IB, Gig-Ethernet with RDMA) to separate the MPI traffic

Windows Cluster Computing Platform



Head Node

Job Mgmt

Cluster Mgmt

Scheduling

Resource Mgmt

Active Directory

User

Web service

User Console

Cmd line

Input

Job

Policy, reports

admin Console

Admin

Job

Management

Sensors, Workflow, Computation

DB or FS

Data

Data mining, Visualization, Workflow Remote query

Compute Node

Job Execution

User App

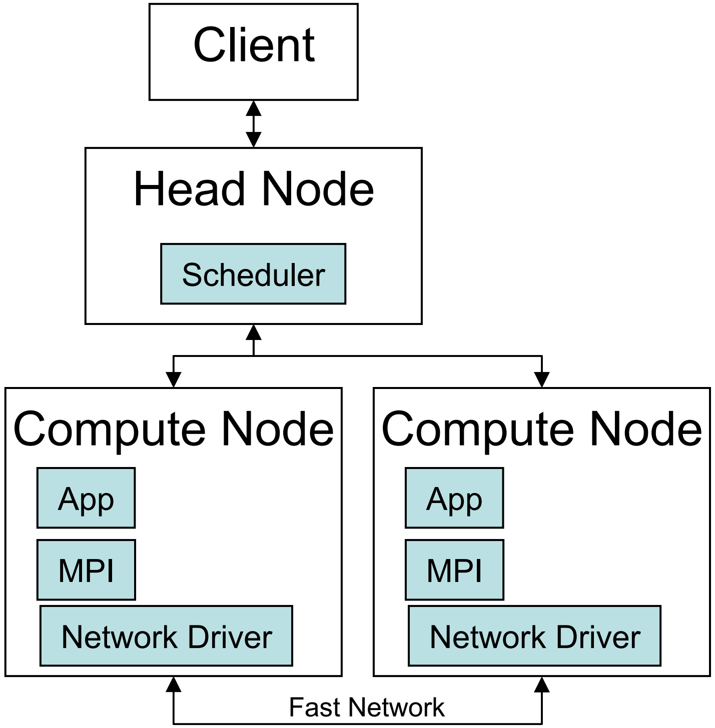
MPI

High speed, low latency interconnect (Ethernet over RDMA, Infiniband)

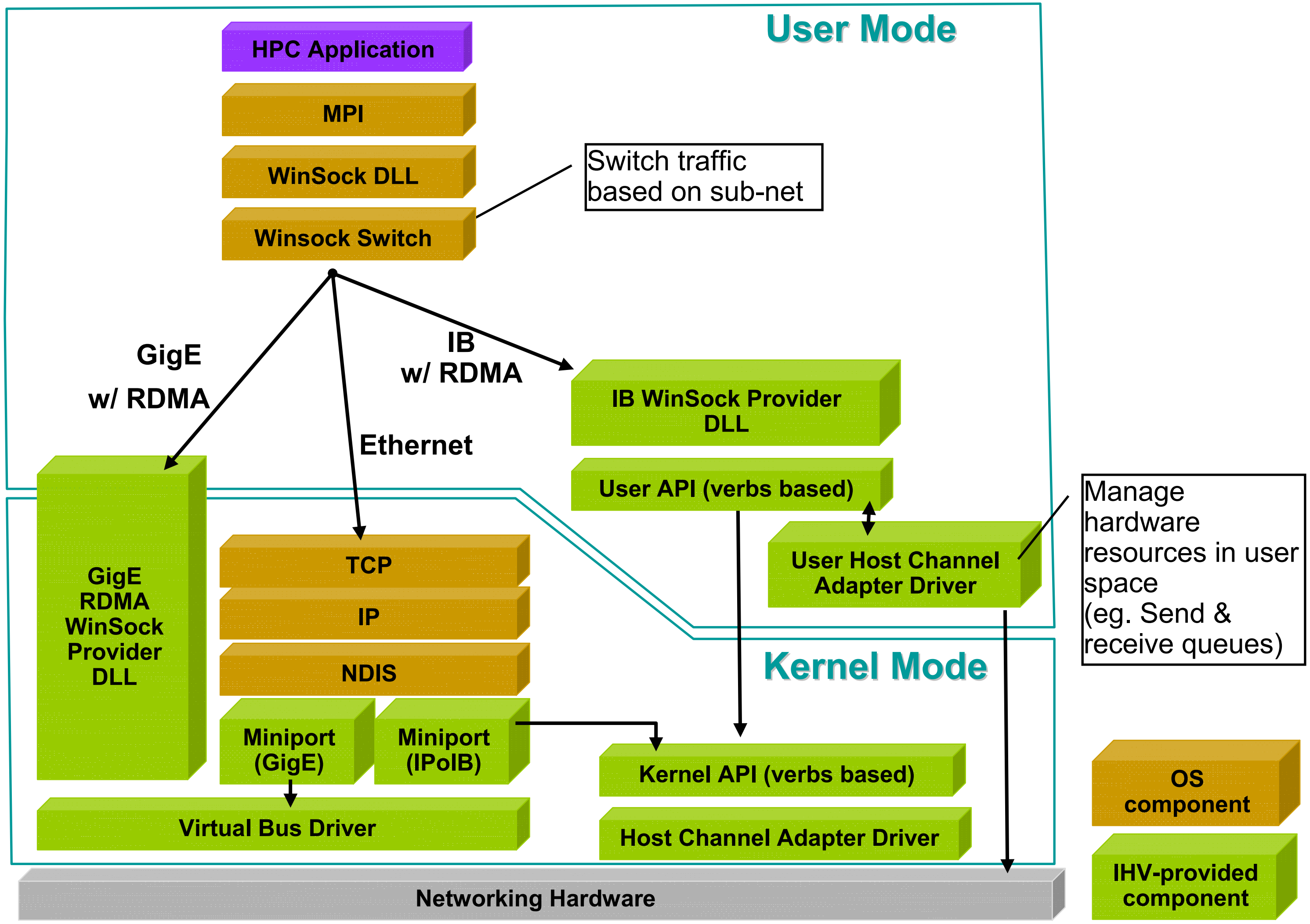
August 22, 2005

What Happens After I Press "Submit Job"?

- Task Execution
  - Scheduler orchestrates
    - Node allocation to the tasks
    - Timing, execution, and clean-up
    - Error Recovery
      - Re-try
      - Routing "around" un-responsive nodes
  - Within a security context
    - Compute nodes authenticate as the user
    - Secure client-scheduler-computeNode communication
- The "Other" Layers
  - Application: [your program here]
  - Message Passing Interface (MPI): API for messaging between compute nodes cooperating on a task
  - Networking: drivers that enable fast communication via WinSock Direct



Windows CCE Leverages Winsock Direct Architecture



Associated MS Products

- Visual Studio
  - Parallel Debugger
    - Automatic attach to MPI processes from IDE
    - Process level stepping
    - Process breakpoints
    - Process sensitive expression evaluation
  - OpenMP support
- Services For Unix (SFU)
  - Integrate Windows and UNIX/Linux environments
  - Migrate UNIX applications to Windows
  - Directory, File System and UNIX Subsystem
  - Tested and supported by Microsoft

To Learn More

To Learn More

- Microsoft HPC website:
  - http://www.microsoft.com/hpc/
- x64 Info:
  - http://www.microsoft.com/windowsserversystem/64bit/default.mspx

Q & A

Message Passing Interface

- What is it?
  - Minimalist Answer…
    - Software which is automatically installed on each compute node by Windows CCE's node management
    - It's plumbing…it has to be there for your HPC apps to run
  - MS MPI based on (and compatible with) an open-source, MPI reference implementation- **Argonne National Lab's MPICH2**
- Why did the MS team choose MPI?
  - MPI emerging as the dominant protocol for parallel compute messaging
- Do I have to use MS MPI if I use Windows CCP?
  - No, you can use any MPI stack you choose.
  - However, the security features MS HPC have added to MPICH may not be available in other MPI stacks.

MPI Description

- MPI is a standard specification, there are many implementations such as MPICH2, MS MPI, etc.
- MPI consists of 2 parts
  - For ISVs: Full-featured API of160+ functions (can do much work with ~10 functions!)
  - For Users: Command-line (mpiexec) or GUI tool to launch jobs
- Abstracts communication concepts so even I can create parallel programs! [But it's still not as simple as it must be for common usage]

MS MPI and MPICH2

- MS HPC goal is maximum compatibility with MPICH2 Reference Implementation
  - Full compatibility for ISV's using MPI API's.
- Exceptions made for with:
  - CCP Scheduler incompatibilities
  - CCP Security Goal incompatibilities
  - Windows-based Performance improvements that do not affect the API's
- Thus, differences concentrated in job launch/mgmt: MPIExec, MPI Daemon (SMPD)
  - pwdfile, delegate, impersonate, localroot, remove [uninstall smpd], sethosts, etc.

Parallel Debugger

- Basic features to debug MPI applications
  - Automatic attach to MPI processes from IDE
  - Process level stepping
  - Process breakpoints
  - Process sensitive expression evaluation

OpenMP

- A specification for multithreaded programs
  - Helps hyperthreading
- Conformance to the OpenMP 2.0 standard
- Support for .NET and OpenMP together
  - Compiler generates MSIL for OpenMP code
- It consists of a set of simple #pragmas and runtime routines
  - #pragma omp parallel
- A common technique:
  - Start with sequential code and parallelize by adding #pragmas
- Most value, where?
  - Parallelizing large loops without loop-dependencies
    - Can do more, but that's the big win

Windows – UNIX Interoperability

- The Challenge
  - Enable user productivity via Windows for UNIX administrators and developers
- Approaches
  - Use UNIX Interoperability Tools
  - Compile and Configure UNIX Tools from Source Code
  - Assemble a Collection of Third-party Tools

Microsoft's Solution - Services for UNIX:

- Services for UNIX v. 3.5 provides the tools and environment that IT professionals and developers need to:
  - Integrate Windows and UNIX/Linux environments
  - Migrate UNIX applications to Windows
- Services for UNIX is one of the most comprehensive interoperability solutions:
  - Directory, File System and UNIX Subsystem
  - Tested and supported by Microsoft
- Services for UNIX uniquely enables IT pros to easily extend the value of their knowledge and training

Services For UNIX 3.5

- Focused on two major customer “pain” areas
  - Seamless UNIX / Windows Interoperability
    - File & data sharing
      - NFS (Client, Server, Gateway)
    - UNIX / Windows cross-platform management
      - AD / NIS server directory services & interop
      - Bidirectional Password Sync, user name mapping
      - Remote exec tools, rlogin, xterm, telnet, UNIX scripting, Perl
  - UNIX to Windows Application Portability
    - UNIX Tools: C, C++, Fortran, scripts, build tools
    - Interix UNIX subsystem

- Leverage existing UNIX skills, methods and code

”Best System Integration” Award - LinuxWorld 2003