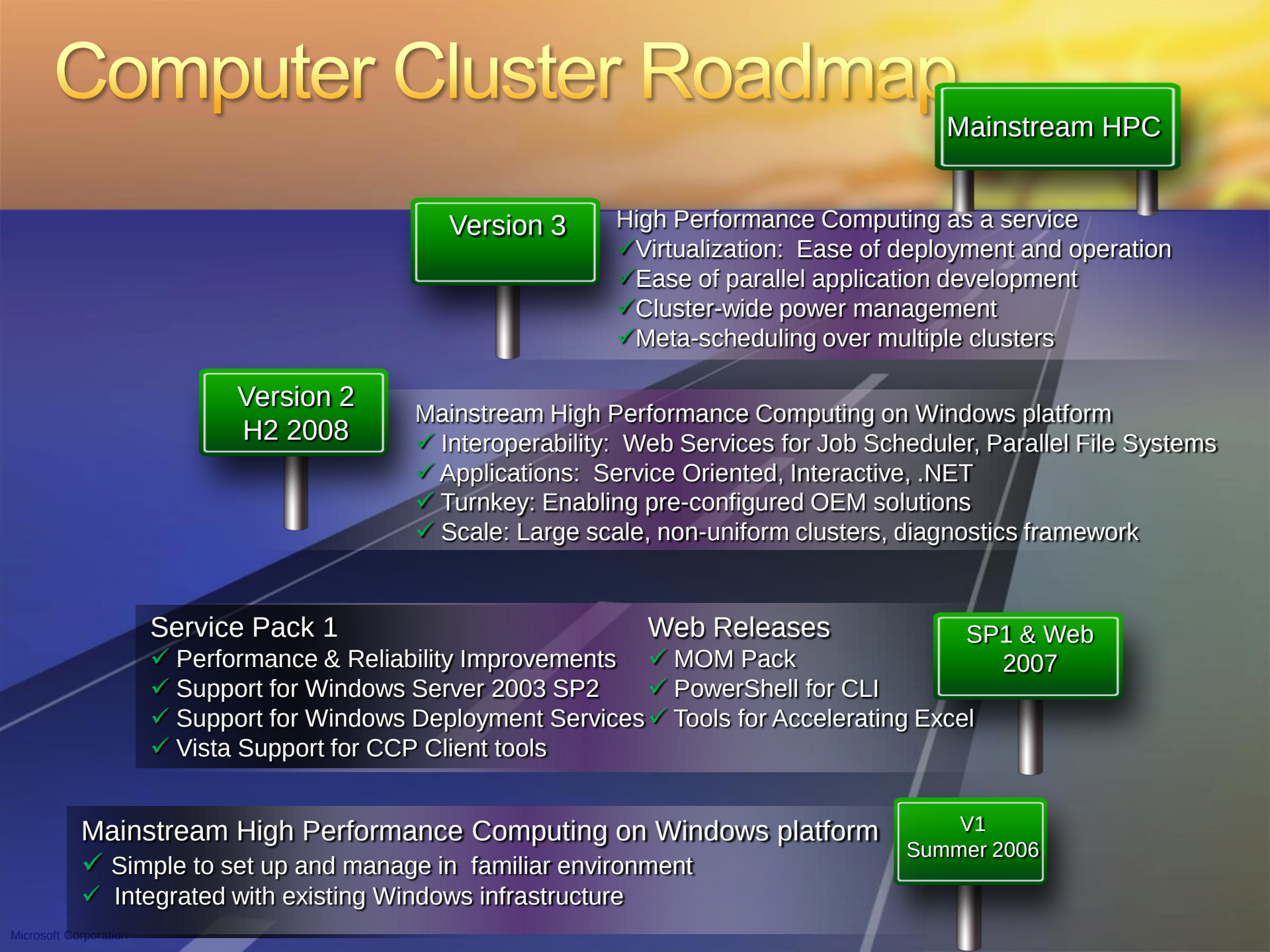
Microsoft

Windows Compute Cluster Server

Erez Haba
MPI & Networking Development
Microsoft Corporation

Computer Cluster Roadmap



Compute Cluster Server v1, Reminder

- Targeting ‘Personal Supercomputing’
- Windows Server 2003 OS
- Includes,
  - Deployment & Management
    - Using RIS, ICS
  - Job Scheduler
    - Fix job size, cpu allocation unit /numprocessors
    - Parametric sweep, MPI jobs
  - MPI
    - Derived from MPICH2; Integrated with CCS
- Primarily a batch system

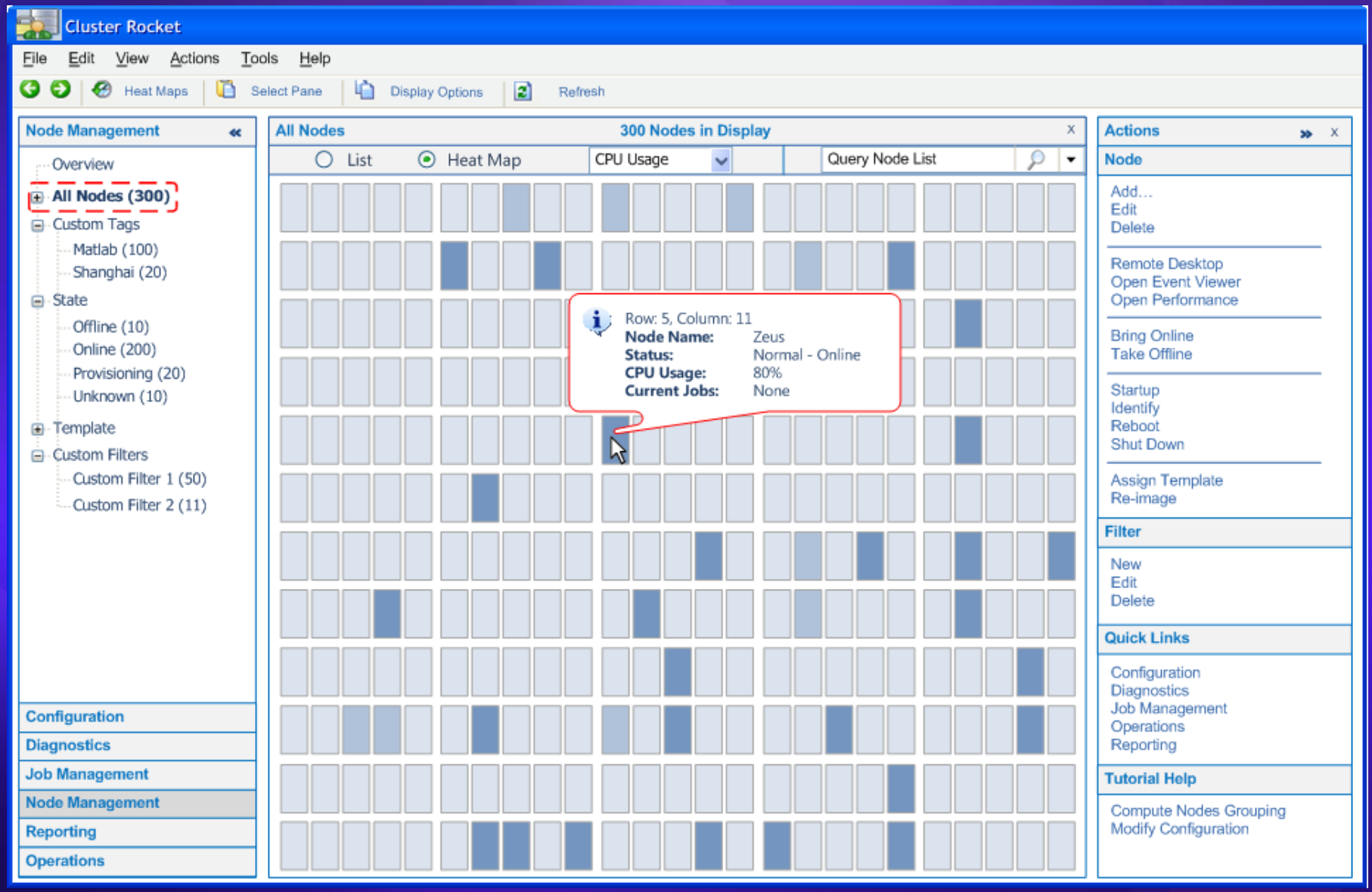
Compute Cluster Server v2

- Targeting ‘Divisional Supercomputing’
- Windows Server 2008 OS
- Extended market segments
  - Finance, CAE, Bioinformatics…
- Larger in-house test cluster
  - 256 nodes 8 cores Clovertown w/ InfiniBand

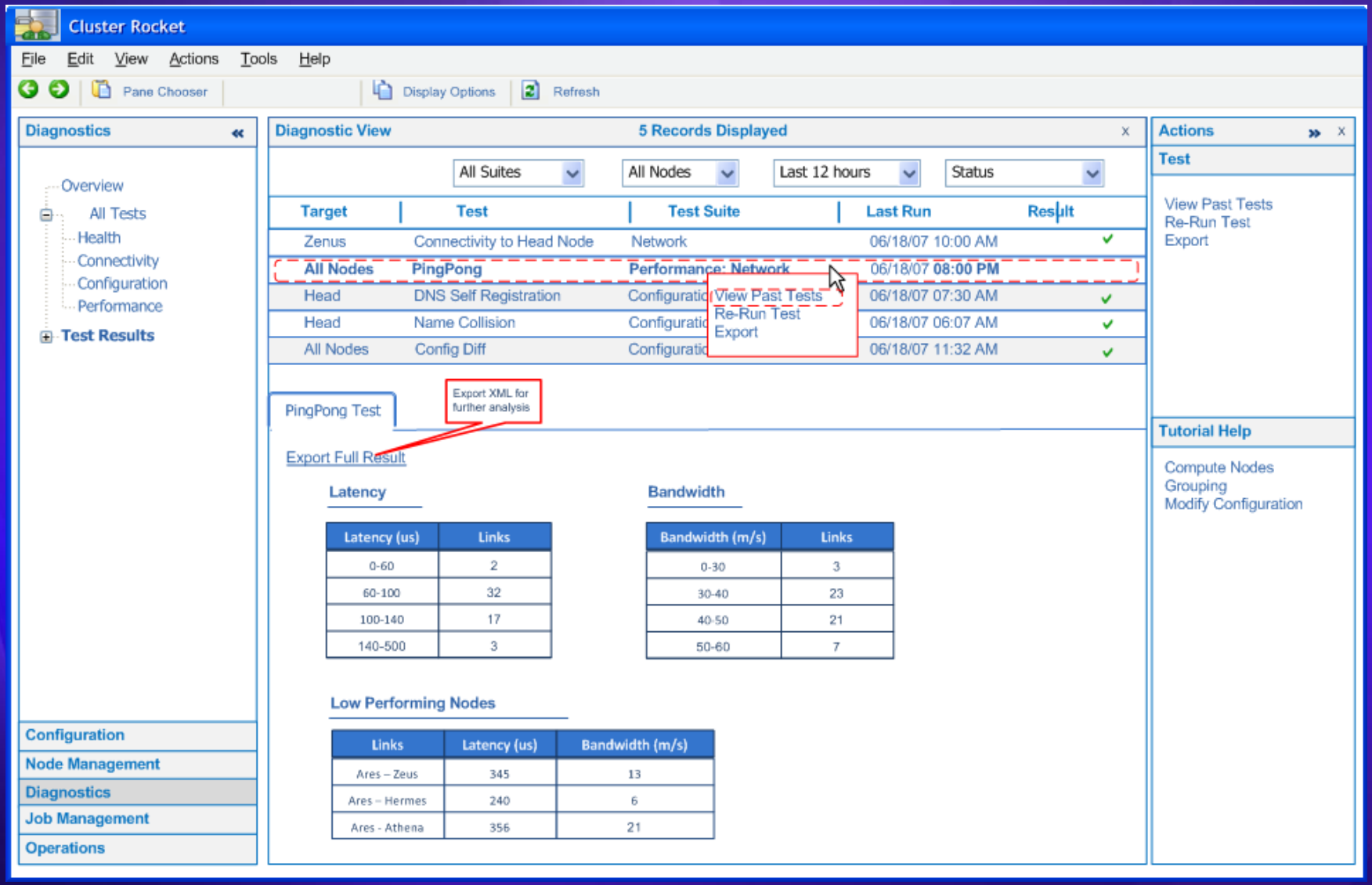
Deployment & Management

- Extended Deployment
  - WDS (multicast), Template based, incl. app deployment
  - RRAS, DHCP
- Extended Management & Diagnostics
  - Reporting
  - Diagnostics tools (pluggable)
  - Extended scripting using Power Shell
  - Microsoft Operations Manager (MOM)

Node Management - Monitoring



Diagnostics

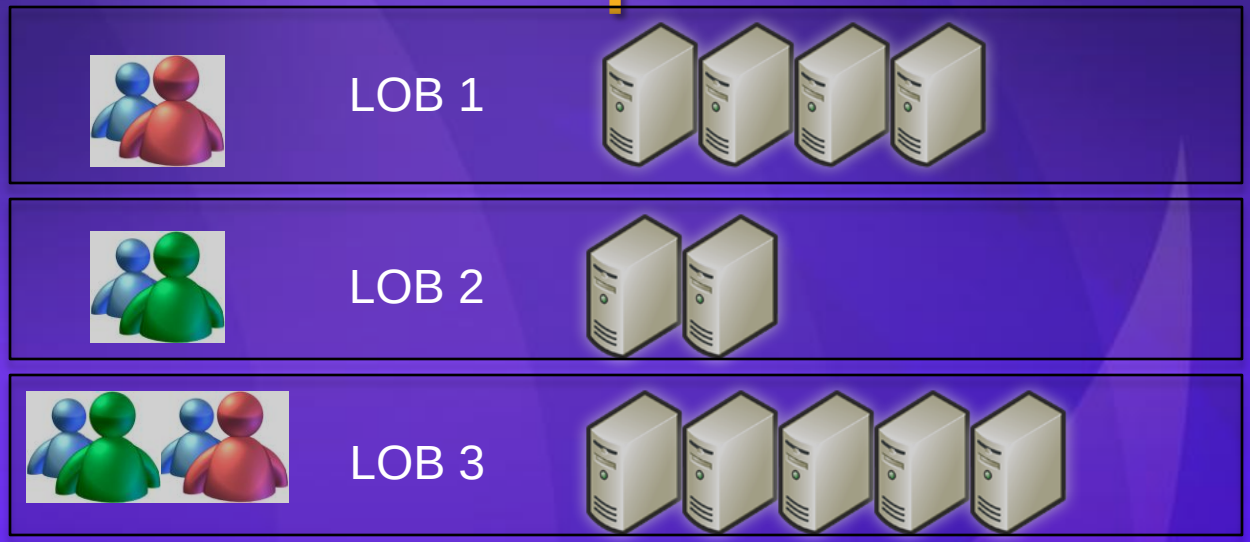


Compute Cluster Server v2

- Job Scheduler *(shared cluster)*
  - Cluster Administrator
    - Resource preemption
    - Job policies
  - Resource Utilization
    - Dynamic job resize (grow/shrink)
    - Resource units: new /numnodes /numsockets
  - Heterogeneous Clusters
    - Node tags; query string
  - Notifications

Policy Scenario: Multiple LOBs

Create Resource Partitions



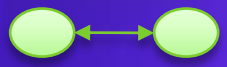
Configure LOB Level Admission Policies

Admission control	Descriptions	Definitions
Runtime to be mandatory	A supercomputing center wanting to enforce the runtime for all the jobs	Profile: default Runtime:required Default: none Users: All
Multiple Line of Businesses (LOBs) sharing a cluster	Admin would like to apportion resources to different nodes	Profile LOB1: Users: user1, user2 Priority: normal, Select:"sas && ib && processorspeed > 2000000" Uniform: switchId Profile LOB2: Users: user3, user4 Askednodes:host2 host3 host 4
Power user job priority	Power user userA can use all the nodes in the cluster	Profile PowerUser: Users: userA Askednodes: All Priority: Highest

Scenario: job right placement

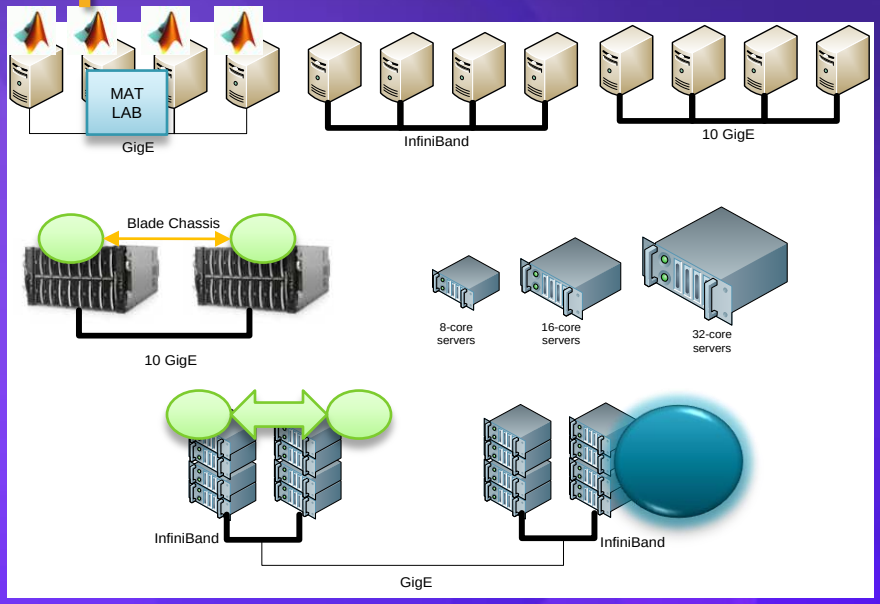
Matlab application (requires
Nodes where Matlab is installed)

MPI application - requires
Machines with same network
Switch

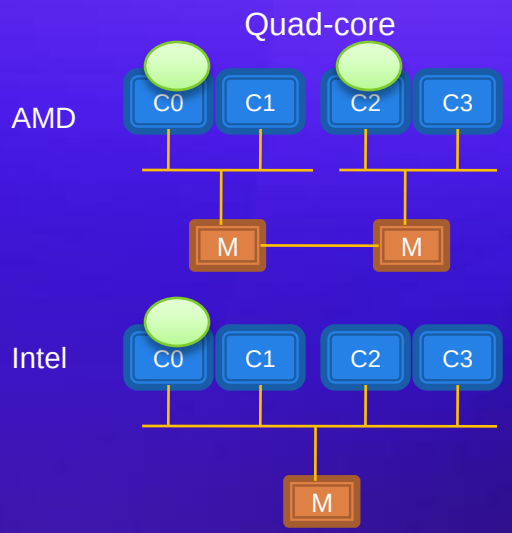


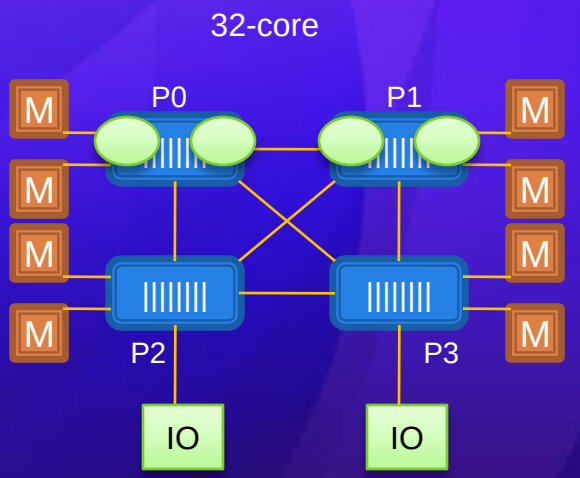
Large application (requires
Large memory machines)

MPI application - requires
High bandwidth and low latency



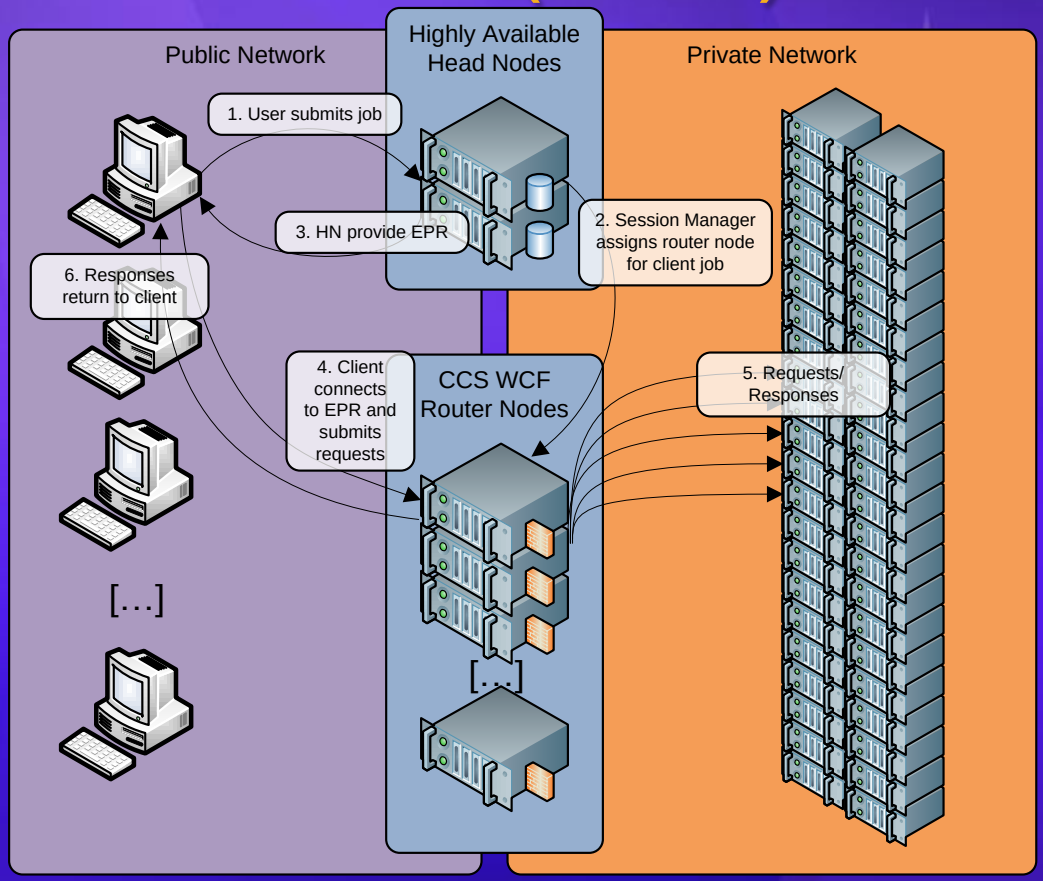
4-way Structural Analysis MPI Job





Interactive Applications
Service Oriented Architecture (SOA)

- Pre-deployed Web Service
  - Discovery
- Job Scheduler features
  - Most important jobs run first
  - Apply scheduling policies
- Clients submit to head node
  - Job is reservation of resources
- Head node assigns router
  - Assignment made when nodes available
  - Router starts WCF application on nodes
  - WAS and IIS hosting not supported in v2
- Client connects to router
  - HN provides EPR (router) to client
  - Client connects to EPR
  - Standard WCF request/response with stateless messages



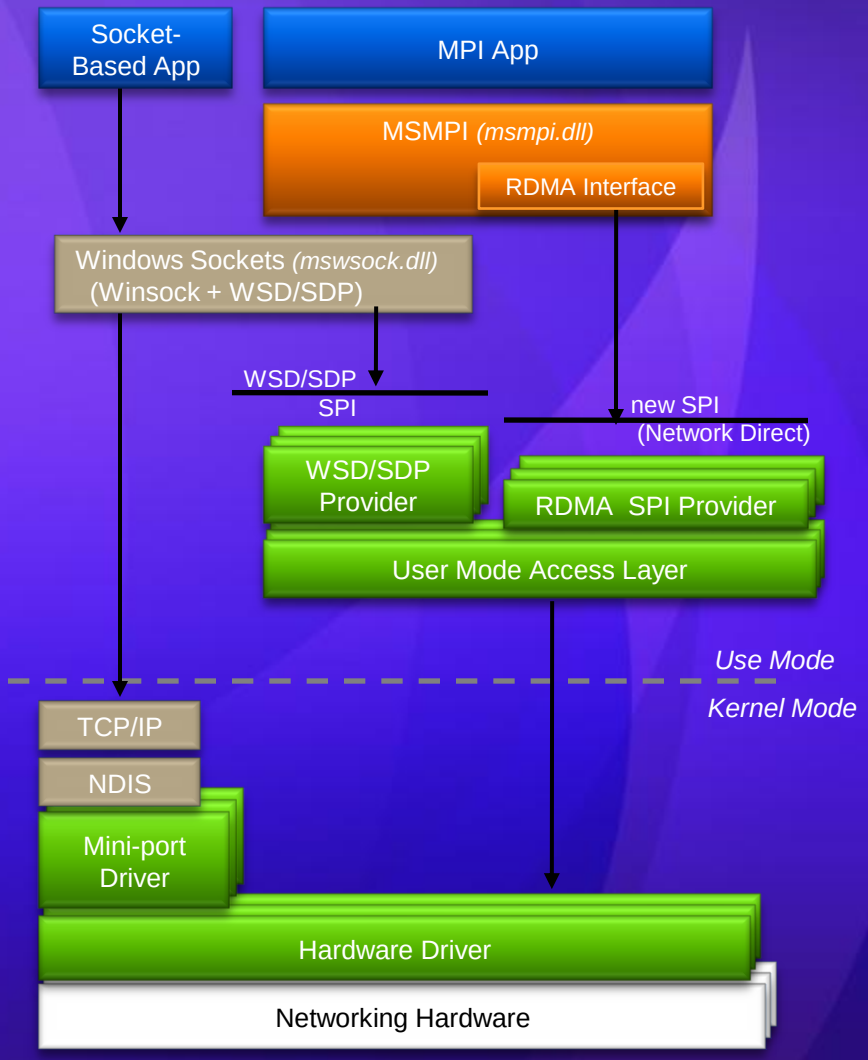
Microsoft MPI

- *"Gloves come off"* for MSMPI v2 Performance
  - Shiny new shared-memory interconnect plays nice with other interconnects. Pingpong latency < .6usec,throughput > 3.5GB/sec.
    - btw: checks are always on
  - MSMPI integrates Network Direct for bare-metal latencies
    - Network Direct, new industry standard SPI for RDMA on Windows
  - Benchmark and improve based on a set of commercial applications

- Devs really want to see how the apps execute on many nodes
  - Trace using high perf Event Tracing for Windows (ETW)
  - Provides OS, driver, MPI, and app events in one time-correlated log
  - CCS-specific feature…Ground-breaking trace log clock synchronization based solely on the MPI message exchange
  - Visualization as simple as high fidelity text or fully fledged graphic viewer
  - Convert ETW trace files to Vampir OTF or Jumpshot c2log/slog

Network Direct

- Designed for both IB & iWARP
  - Rely on IHV's Providers for CCSv2
  - iWARP, OFW, Myrinet
  - Coordinated w/ Win Networking team
- MSMPI
  - Retain MSMPI support for Winsock Direct
  - Uses bCopy and zCopy
  - Uses polling and notifications
  - Plays nice with other interconnects



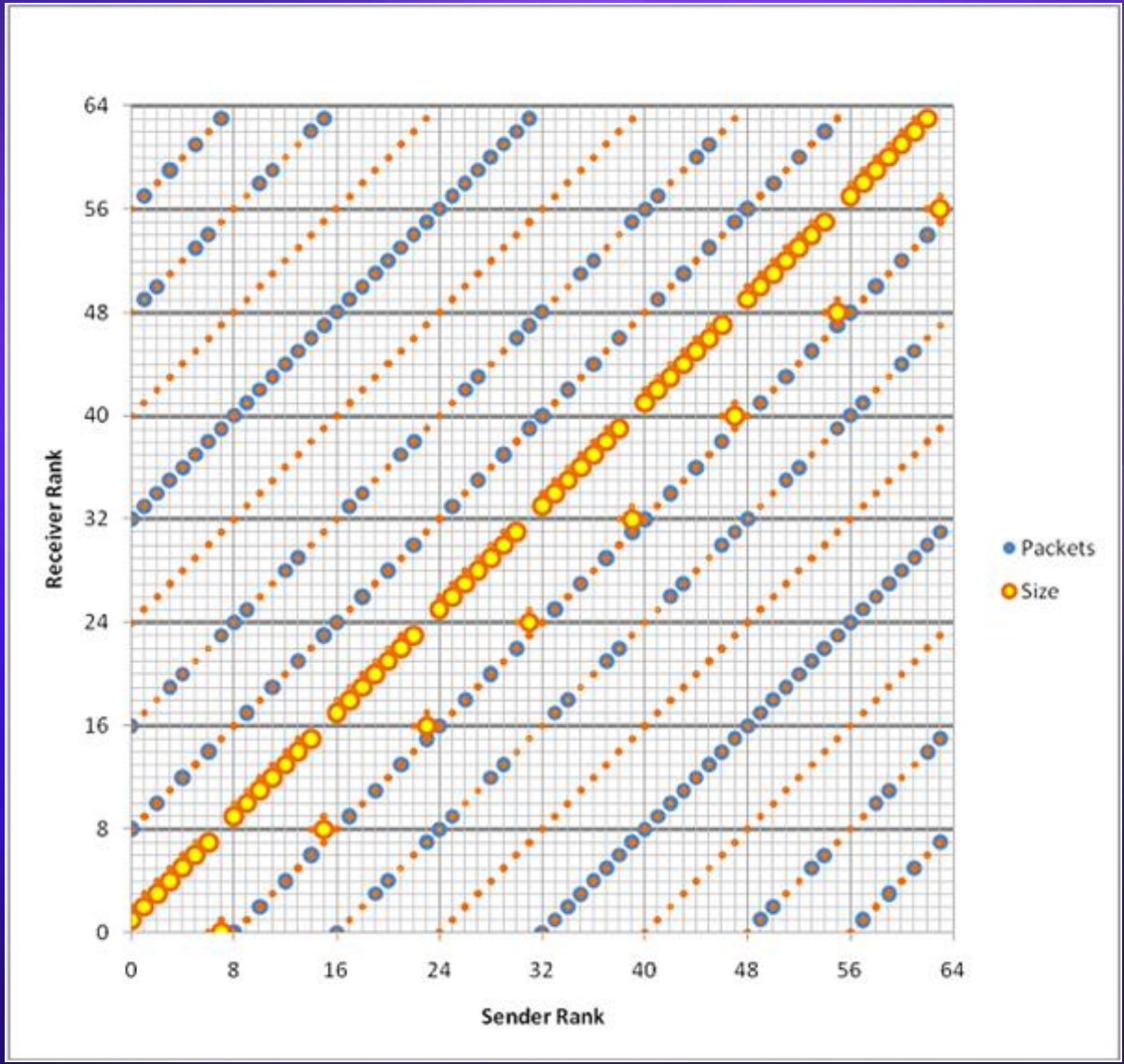
Tracing

- mpiexec –trace [filter] for the full run or,
- Turn on/off while the mpi app is running
- *Demo… stop @ms table*

```
c.pid.tid date       time     ms  ns     [component]    sig: free formatted text
0.954.900 06/01/2007-18:11:58.439.463000 [PMPI_Barrier] Enter:comm=44000000
0.954.900 06/01/2007-18:11:58.439.468400 [SOCK] Send:inln id={2.3.45} n_iov=1 size=36 type=0
0.954.900 06/01/2007-18:11:58.439.476100 [SOCK] Send:done id={2.3.45}
1.954.900 06/01/2007-18:11:58.556.206000 [SOCK] Recv:pkt id={1.2.40} type=0
1.954.900 06/01/2007-18:11:58.556.210000 [SOCK] Recv:done id={1.2.40}
1.954.900 06/01/2007-18:11:58.556.224900 [SHM] Send:inln id={2.0.85} n_iov=1 size=36 type=0
1.954.900 06/01/2007-18:11:58.556.231600 [SHM] Send:done id={2.0.85}
0.954.900 06/01/2007-18:11:58.556.276300 [SHM] Recv:pkt  id={0.2.45} type=0
0.954.900 06/01/2007-18:11:58.556.278800 [SHM] Recv:done id={0.2.45}
0.954.900 06/01/2007-18:11:58.556.281300 [PMPI_Barrier] Leave:rc=0

0.954.900 06/01/2007-18:11:58.556.284300 [PMPI_Gather] Enter:comm=44000000 sendtype=4c00080b sendcount=1……
0.954.900 06/01/2007-18:11:58.556.291400 [PMPI_Type_get_true_extent] Enter:datatype=4c00080b
0.954.900 06/01/2007-18:11:58.556.293400 [PMPI_Type_get_true_extent] Leave:rc=0 true_lb=0 true_extent=8
0.954.900 06/01/2007-18:11:58.556.294100 [PMPI_Type_get_true_extent] Enter:datatype=4c00010d
0.954.900 06/01/2007-18:11:58.556.294500 [PMPI_Type_get_true_extent] Leave:rc=0 true_lb=0 true_extent=1
0.954.900 06/01/2007-18:11:58.556.323400 [SOCK] Recv:pkt id={3.2.44} type=0
0.954.900 06/01/2007-18:11:58.556.325400 [SOCK] Recv:done id={3.2.44}
0.954.900 06/01/2007-18:11:58.556.327500 [PMPI_Get_count] Enter:status->count=8 datatype=4c00010d
0.954.900 06/01/2007-18:11:58.556.329000 [PMPI_Get_count] Leave:rc=0 count=8
0.954.900 06/01/2007-18:11:58.556.333300 [SHM] Send:inln id={2.0.86} n_iov=2 size=52 type=0
0.954.900 06/01/2007-18:11:58.556.336400 [SHM] Send:done id={2.0.86}
0.954.900 06/01/2007-18:11:58.556.338600 [PMPI_Gather] Leave:rc=0
```

Tracing - realtime



Tools

- Debuggers
  - VS, Allinea DDT
- Profilers
  - VS, Vampir
- Compilers
  - Fortran by PGI & Intel
- Libraries - boost.mpi & mpi.net
  - by Indiana University

Beyond v2

- Programming to MPI is easy! yes?
- Looking into languages and libraries to express parallelism
  - Use MPI as the transport
- Support distributed queries (Cluster LINQ)
- Extend Many cores to clusters
  - Microsoft researching many cores/cluster arch

Thanks,

- email
  - erezh@microsoft.com
- HPC web site
  - www.microsoft.com/hpc