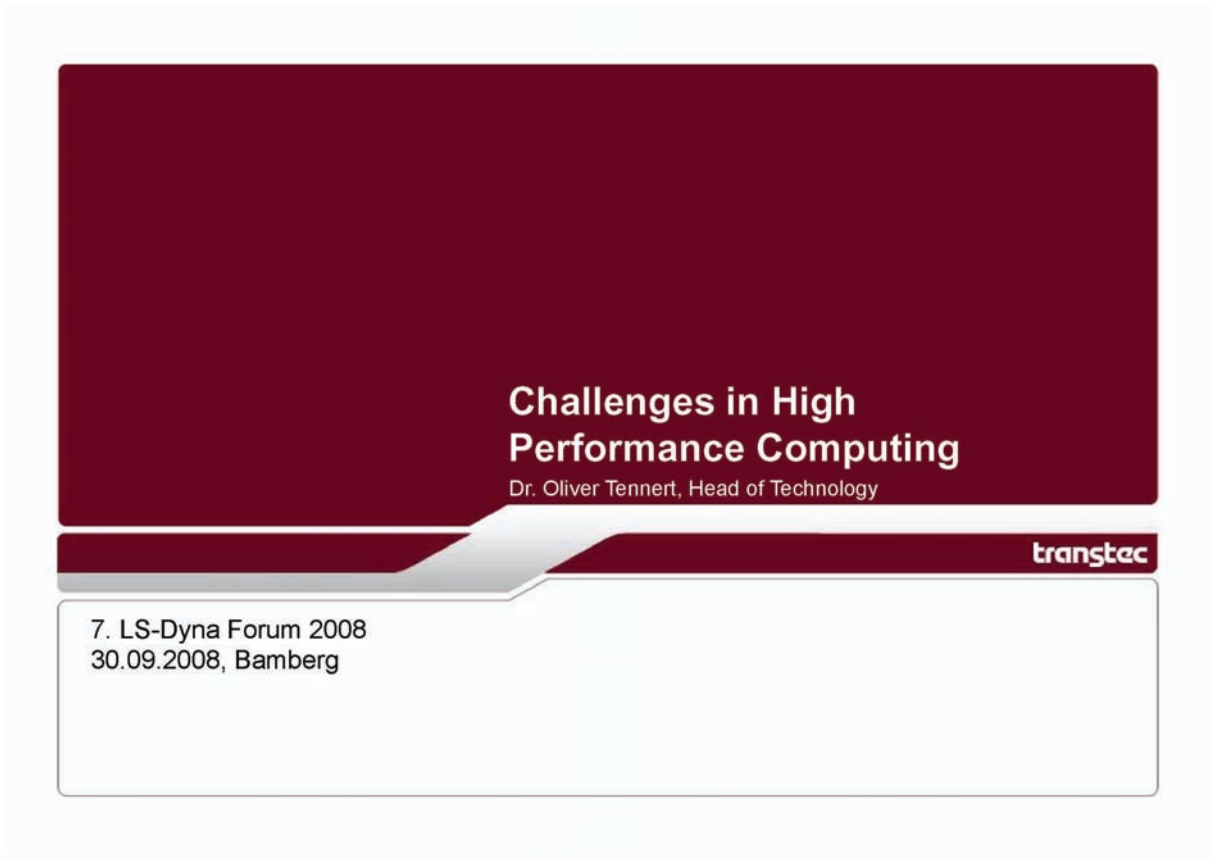
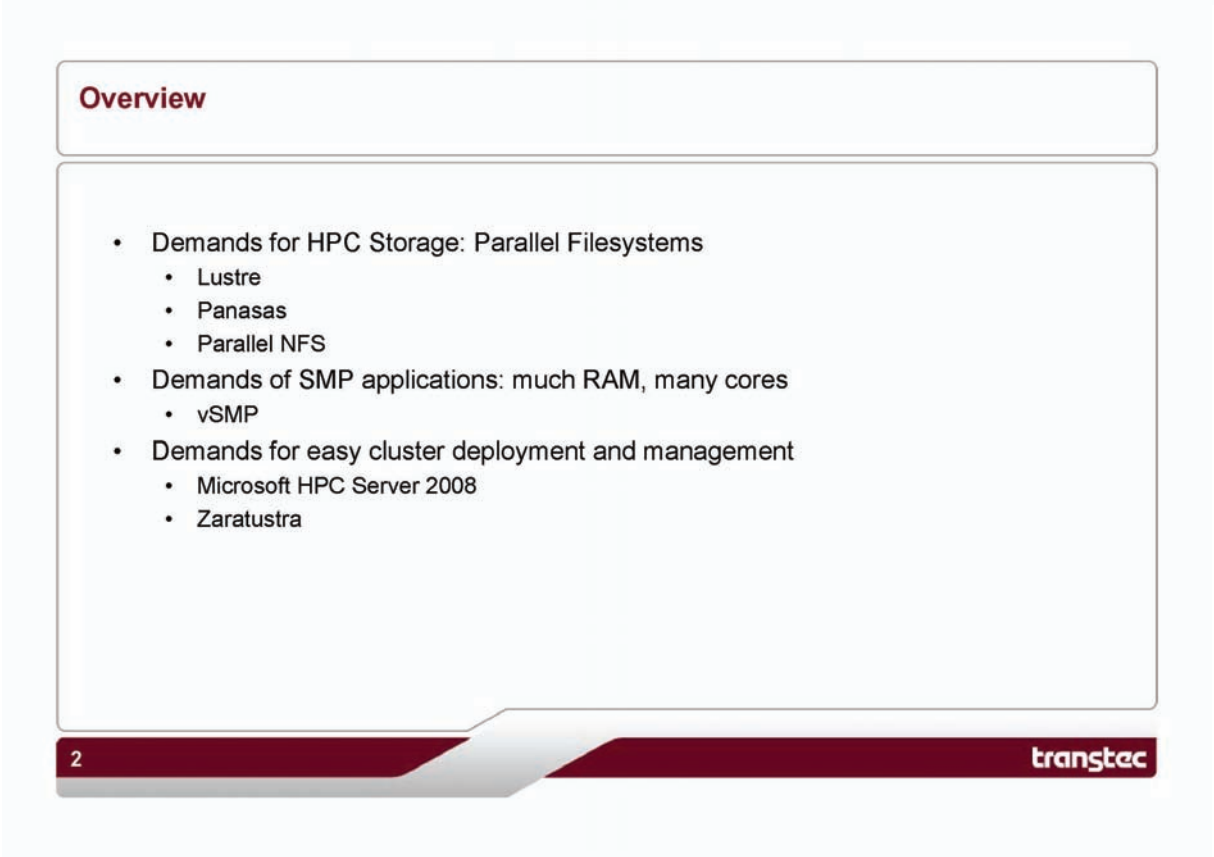
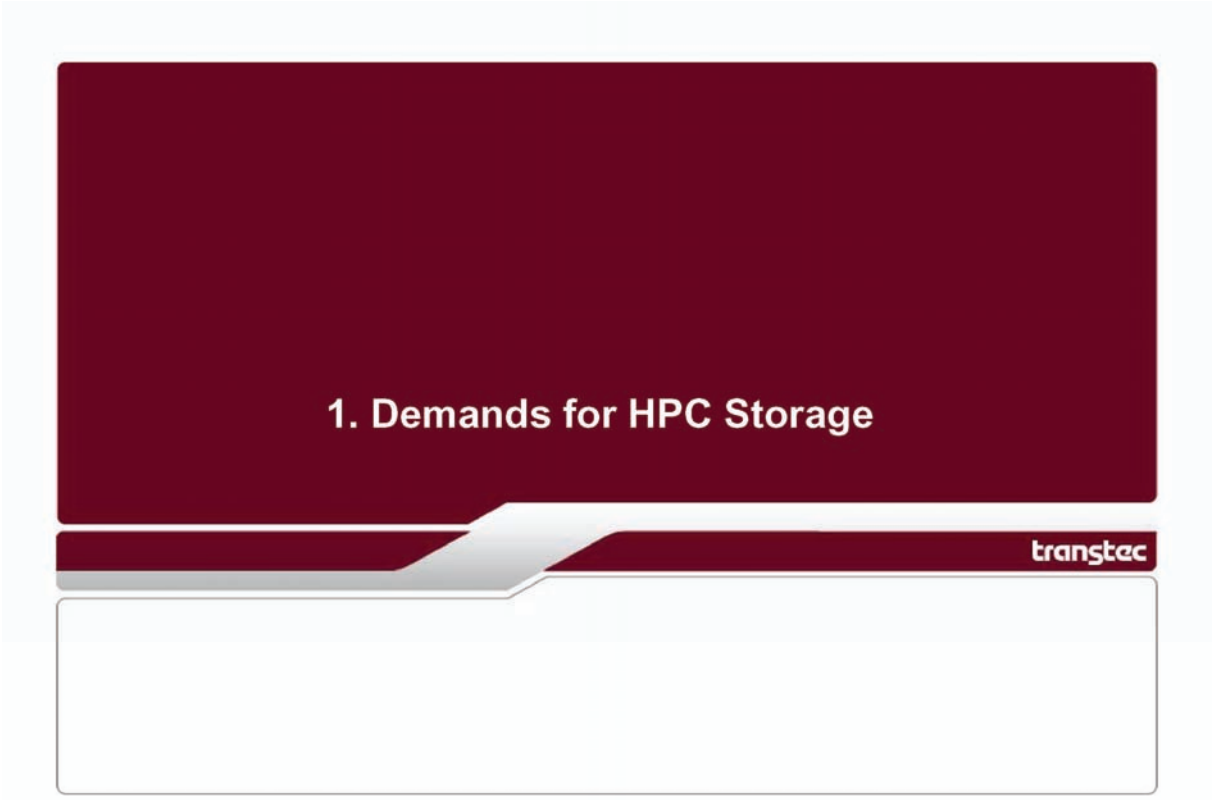
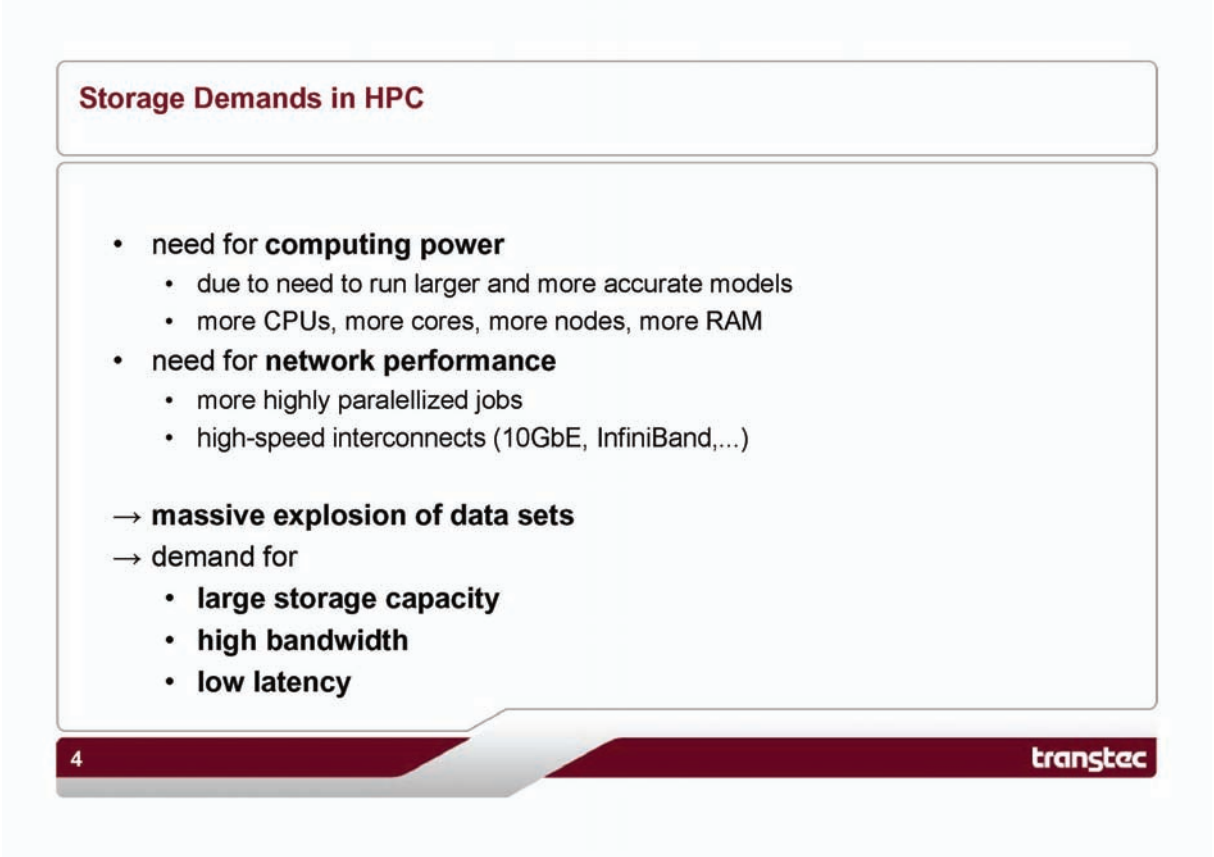
Challenges in High Performance Computing

O. Tennert (transtec ag)






1. Demands for HPC Storage
transtec


Storage Demands in HPC
• need for computing power
• due to need to run larger and more accurate models
• more CPUs, more cores, more nodes, more RAM
• need for network performance
• more highly paralellized jobs
• high-speed interconnects (10GbE, InfiniBand,...)
→ massive explosion of data sets
→ demand for
• large storage capacity
• high bandwidth
• low latency
4
transtec

Deficiencies of Today's Solutions

- most widespread solution: **single NFS server**
  - does not scale: NFS head is bottleneck
  - „high-speed" NFS server will be bottleneck by tomorrow
- **„clustered NFS":** problematic
  - either head-to-head synchronization limits scalability
  - or manual partitioning of global namespace is cumbersome
  - NFS is not suitable for dynamical load balancing (inherent state)

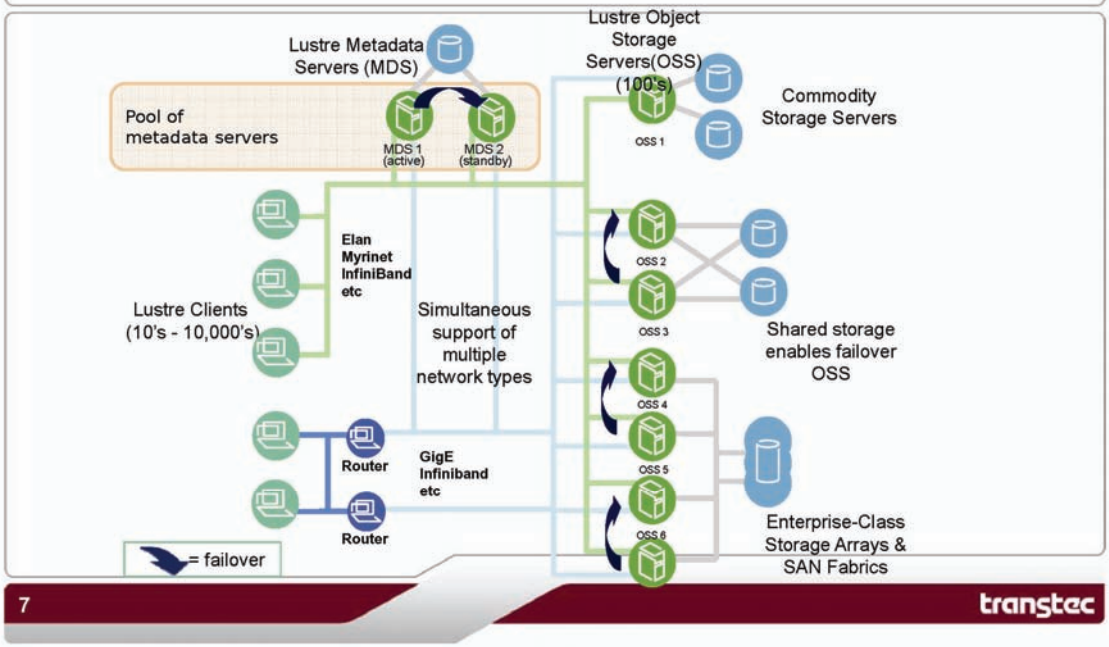
5 transtec

Parallel File Systems

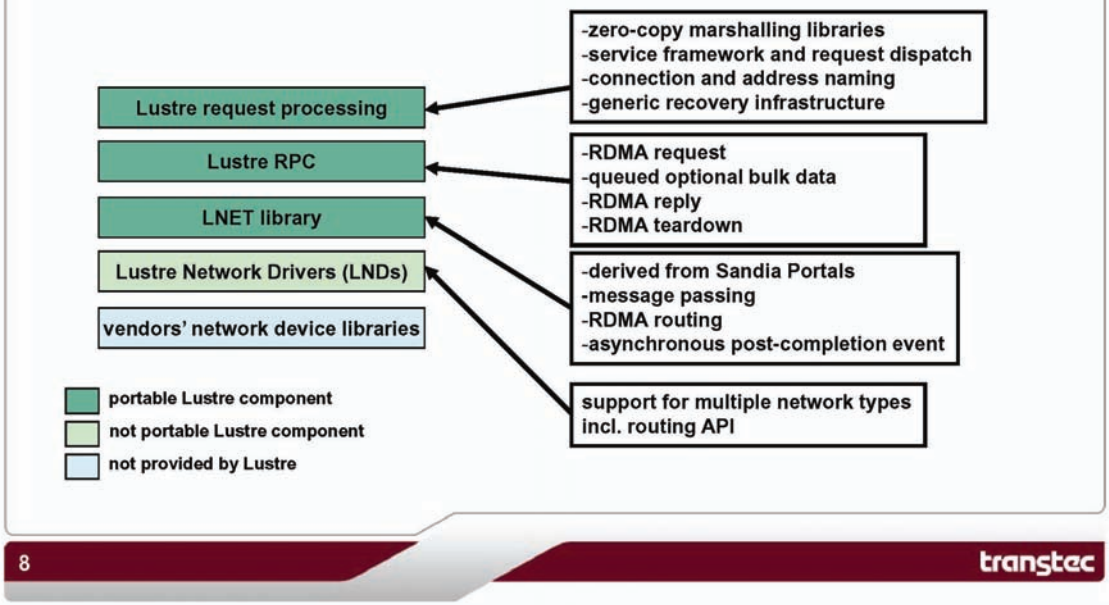
- major features:
  - **global namespace** eases filesystem management and job flow
  - **scalable** capacities and bandwidths
- **distributed** vs. **cluster** vs. **parallel filesystem:**
  - no shared storage → many-to-many access to data
- **proprietary solutions** already there:
  - IBM's GPFS
  - SGI's CXFS
  - Panasas' ActiveScale Filesystem (PanFS)
  - EMC's Celerra MPFS/MPFSi (pka High Road)
  - Lustre, PVFS2, ...

6 transtec

Lustre Overview



Lustre – Modular Network Implementation

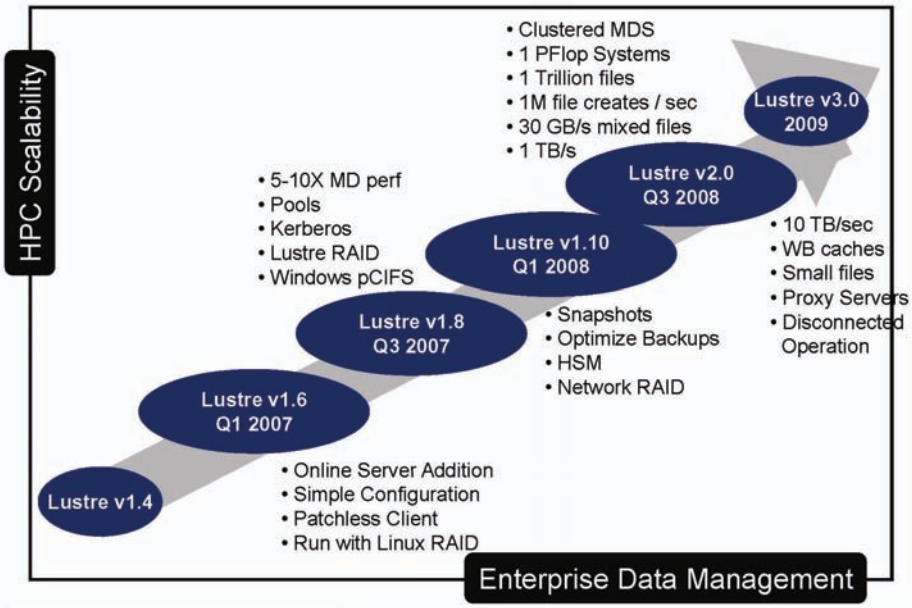


Lustre Limits

- maximum stripe count: 160
  maximum number of OSTs: 1020
  maximum number of clients: 32768 (tested: 22000)
  maximum size of filesystem: 8 TB * #OSTs

9 transtec

Intergalactic Strategy



transtec

NFS as a Standard

- need for **OS independent, interoperable, standardized** solution

→ **NFS is the ONLY standard!**

- **standards are good, because...**
  - they **protect** end user **investment** in technology
  - they **ensure** a base level of **interoperability**
  - while at the same time **provide choice** among products
  - commonality leads to **less training, simpler deployment, higher acceptance...**

11 transtec

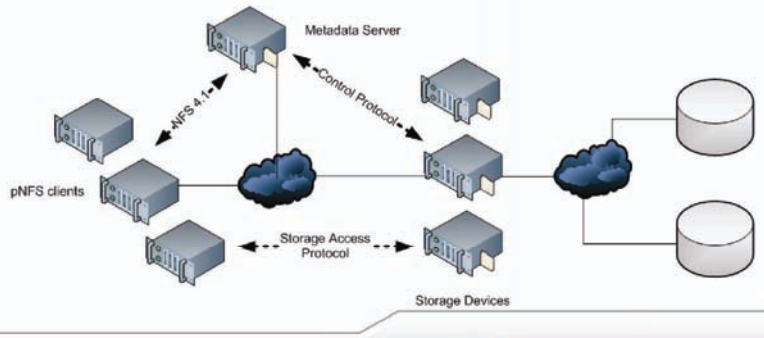
NFS 4.1 and Parallel NFS (pNFS)

- **NFS 4.1:** idea to use SAN FS architecture for NFS originally from Gary Grider (LANL) and Lee Ward (Sandia)
- development driven by Panasas, Netapp, Sun, EMC, IBM, UMich/CITI
- folded into NFSv4 minor version NFSv4.1 in 2006
- future **internet standard**
  (current draft 21: http://www.ietf.org/internet-drafts/draft-ietf-nfsv4-minorversion1-21.txt)
- major changes to NFS 4:
  - sessions
  - directory delegations
  - pNFS (optional feature)
- standardization expected some time in 2009

12 transtec

Parallel NFS (pNFS): Generic Architecture

- **separation** of **metadata** path and **data** path (**out-of-band** global namespace)
- built for **interoperability** and **backwards-compatibility**
- **flexible** design allows for different storage implementations (**layouts**)



13 transtec

What pNFS Does NOT Give You

- **improved cache consistency**
  - NFS has **open-to-close consistency**
- **perfect POSIX semantics** in a distributed file system
- **clustered metadata**
  - though a mechanism for this is not precluded

14 transtec

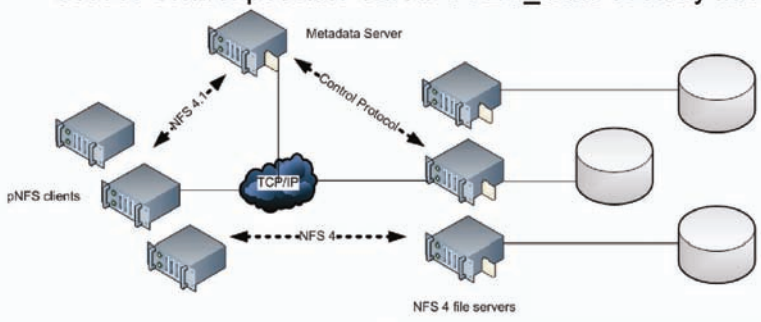
Parallel NFS (pNFS): Different Layout Formats

- a **layout** describes the location of file data, containing a list of device IDs and striping information
- possession of a layout grants access to storage devices, resp. files
- **file-based layout** (part of NFS 4.1/pNFS standard)
- **block-based layout:** http://www.ietf.org/internet-drafts/draft-ietf-nfsv4-pnfs-block-08.txt
- **object-based layout:** http://www.ietf.org/internet-drafts/draft-ietf-nfsv4-pnfs-obj-07.txt
- PVFS2 layout
- GPFS layout
- ...

15 transtec

pNFS: File Layout

- only storage access protocol directly specified in NFS 4.1 standard
- significantly co-designed by NetApp, Sun, IBM and others
- file layout simple, may be **heavily cached** by clients
- access control possible via RPCSEC_GSS security flavor

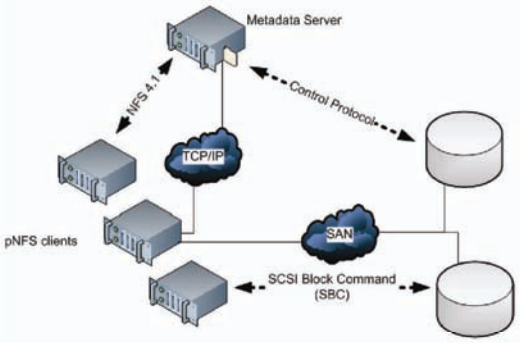


16 transtec

pNFS: Block Layout

- highly influenced by **EMC's** design of **Multi-Path File System MPFS(i) (**pka **High Road)**
- block layout uses **volume identifiers, block offsets and extents**
- secure authorization with **host granularity only**, file-level security cannot be enforced by storage devices

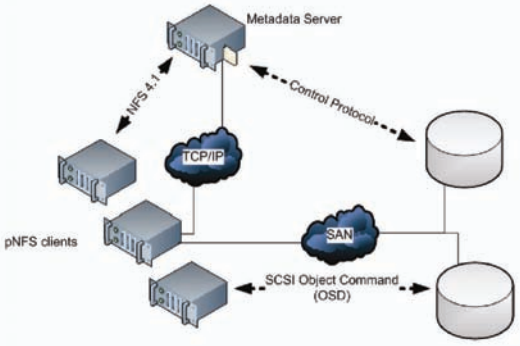
→ **clients must be trusted** (fundamental NFS problem ever since)



17 transtec

pNFS: Object Layout

- Panasas' contribution, based on **NASD** design (Network-Attached Secure Disk) developed at Carnegie Mellon University, later evolved into forthcoming **SCSI OSD** standard (**object-based storage device**)
- layout uses **SCSI object command** set
- **space management** built into devices
- designed for **secure access** and **high-performance data replication**
- cryptographically secured credentials ("**capabilities**") needed to access storage devices



18 transtec

pNFS: The Current State

- **Linux:** file layout, based on PVFS2 / based on NFS 4
- **OpenSolaris:** file (NFS 4) / object (OSD-1) layout driver will be completed soon, patches available: http://opensolaris.org/os/project/nfsv41/ http://opensolaris.org/os/project/osd/
- **Netapp:** file layout, based on NFS 4
- **IBM:** file layout, based on GPFS
- **EMC:** block layout, based on MPFS(i)
- **Panasas:** object layout, based on ActiveScale PanFS
- **Carnegie Mellon University:** performance and correctness testing

19 transtec

Weblinks Lustre

- Lustre Wiki: http://wiki.lustre.org
- Lustre Manual: http://manual.lustre.org
- Lustre Knowledge Base: https://bugzilla.lustre.org/showdependencytree.cgi?id=2374

20 transtec

Weblinks pNFS

- **NASD: Network Attached Secure Disks:** http://www.pdl.cmu.edu/NASD/
- **Panasas:** www.panasas.com
- **EMC Celerra Multi-Path File System:** http://www.emc.com/products/detail/software/celerra-multipath-file-system.htm
- **pNFS Information Portal:** http://www.pnfs.com
- **NFSv4 Status Pages:** http://tools.ietf.org/wg/nfsv4
- **Object-Based Storage Devices (now INCITS 400-2004):** http://www.t10.org/ftp/t10/drafts/osd/osd-r10.pdf
  Object-Based Storage Devices V2: http://www.t10.org/ftp/t10/drafts/osd2/osd2r03.pdf
- **Eisler's NFS Blog:** http://blogs.netapp.com/eislers_nfs_blog
- **NFSv4.1 Bakeathon at OpenSolaris.org:** http://opensolaris.org/os/project/nfsv41/nfsv41_bakeathon/

21 transtec

2. Demands of SMP Applications

transtec

Demands of SMP Applications

- **large memory** for applications
  - to enable workloads that cannot be run otherwise
- **shared memory** coupled with **many CPU cores**
  - to allow multi-threaded applications to benefit from shared-memory systems
- **BUT:** proprietary SMP machines (pka "mainframes") very expensive and not compatible!

→ **vSMP™ ("versatile SMP") technology:**

- aggregates multiple off-the-shelf x86 server boards into one virtual x86 system

ScaleMP™

23 transtec

vSMP Foundation™ Technology

Multiple off-the-shelf x86 boards, with processors and memory
Processors speed/amount or memory amount doesn't have to be same across all boards

+

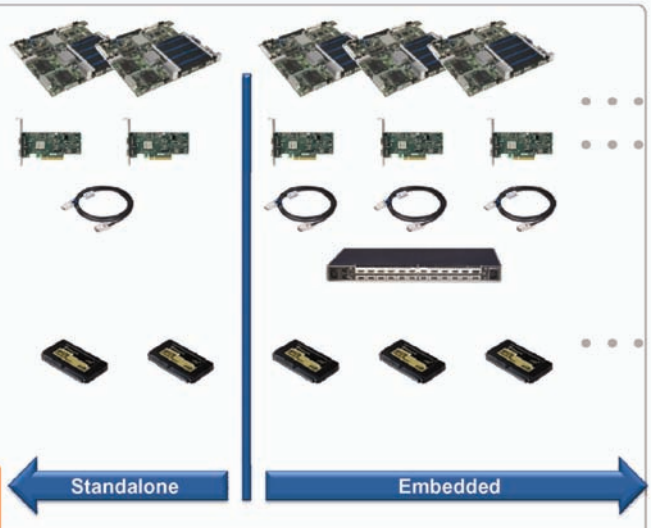
InfiniBand HCAs, cables and switch

+

vSMP Foundation™ Devices
The flash-devices plug into the boards and used as bootable device. vSMP Foundation is booted to present an aggregate coherent view to the OS

=

High-end x86 system, based on standard x86 components

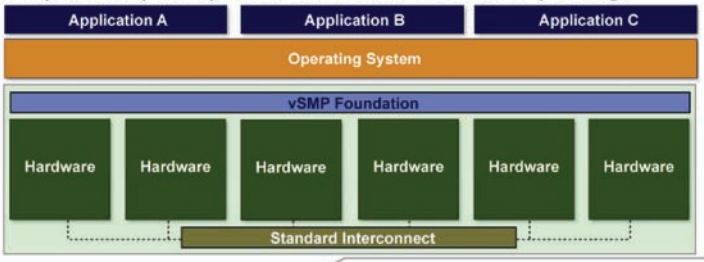


24 transtec

vSMP Foundation™ Technology

- one system to manage: fewer, larger nodes means less cluster management overhead
  - single operating system instance
  - avoid cluster file systems
  - complexities of InfiniBand hidden

→ compatible (=x86) SMP machine at cluster pricing



25 transtec

vSMP Foundation™ Technology: Behind the Scenes

One System

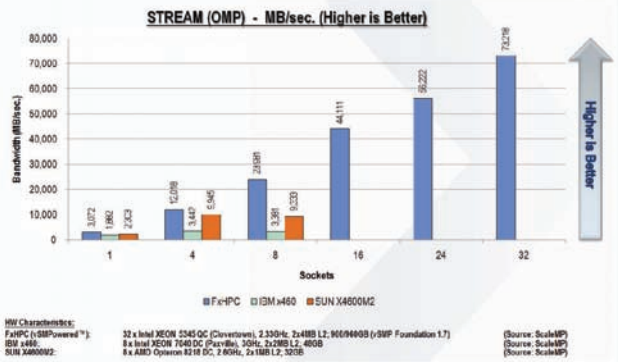
- Software interception engine creates a uniform execution environment
- vSMP Foundation creates the relevant BIOS environment to present the OS (and the SW stack above it) as single coherent system

Coherent Memory

- vSMP Foundation maintains cache coherency between boards
- Multiple concurrent memory coherency mechanisms, on a per-block basis, based on real-time memory activity access pattern
- Leverage board local-memory for caching

Shared I/O

- vSMP exposes all available I/O resources to the OS in a unified PCI hierarchy
- No need for cluster file systems



26 transtec

vSMP Compatible Systems (1)

- **system capabilities:**
  - max. boards: 16
  - max. memory per board: 4 GB / 64 GB
  - max. processors per board: 1 / 4
  - max. cores per board: 1 / 8
  - max. memory: 1 TB
  - max. processors: 64
  - max. cores: 128
- **fault-tolerant:** supported, minimum boards: 3

27 transtec

vSMP Compatible Systems (2)

- **supported hardware:**
  - **processors: Intel Xeon**
  - systems (excerpt):
    - discrete server systems:
      - Supermicro: 6015TW-INFB / 6015TW-INFV
    - blade-server systems:
      - Supermicro: SuperBlade: SBI-7125W-S6

- **supported operating systems:**
  - Linux OS, kernel level 2.6.11 or later
  - Red Hat Enterprise Linux 4/5 (RHEL4-5)
  - SUSE Linux Enterprise Server 10 (SLES10)
  - Fedora Core 4/5/6/7 (FC4-7)
  - OpenSUSE 10 or later

28 transtec

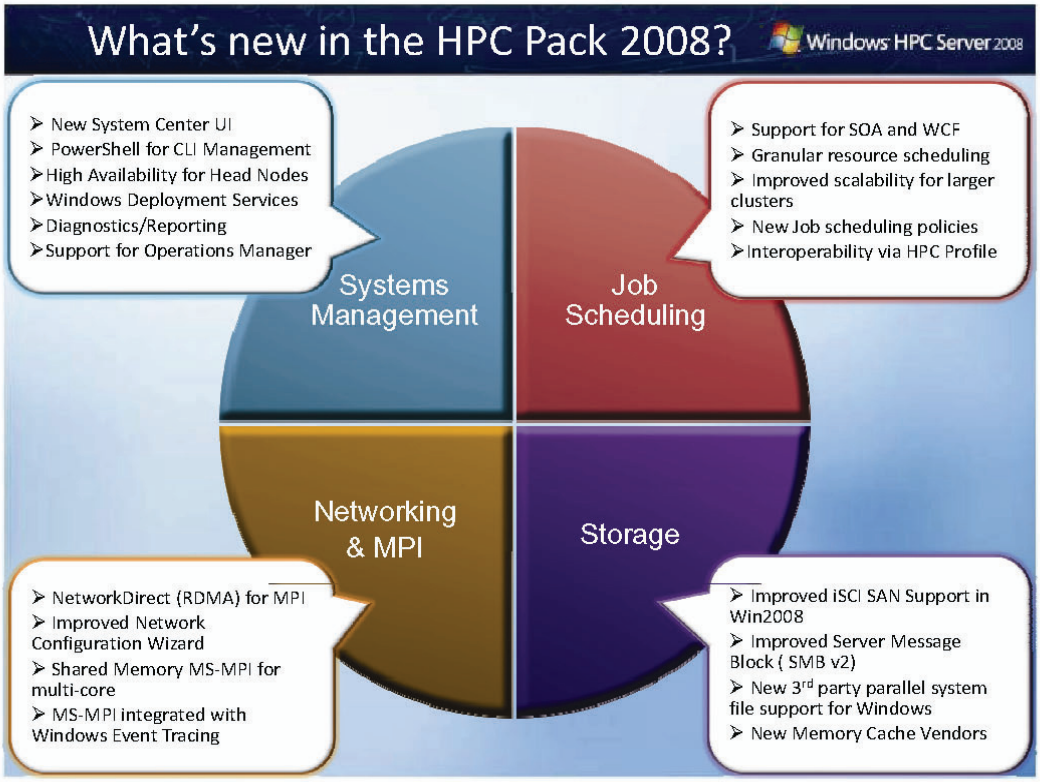
3. Demands for Easy Cluster Deployment and Management

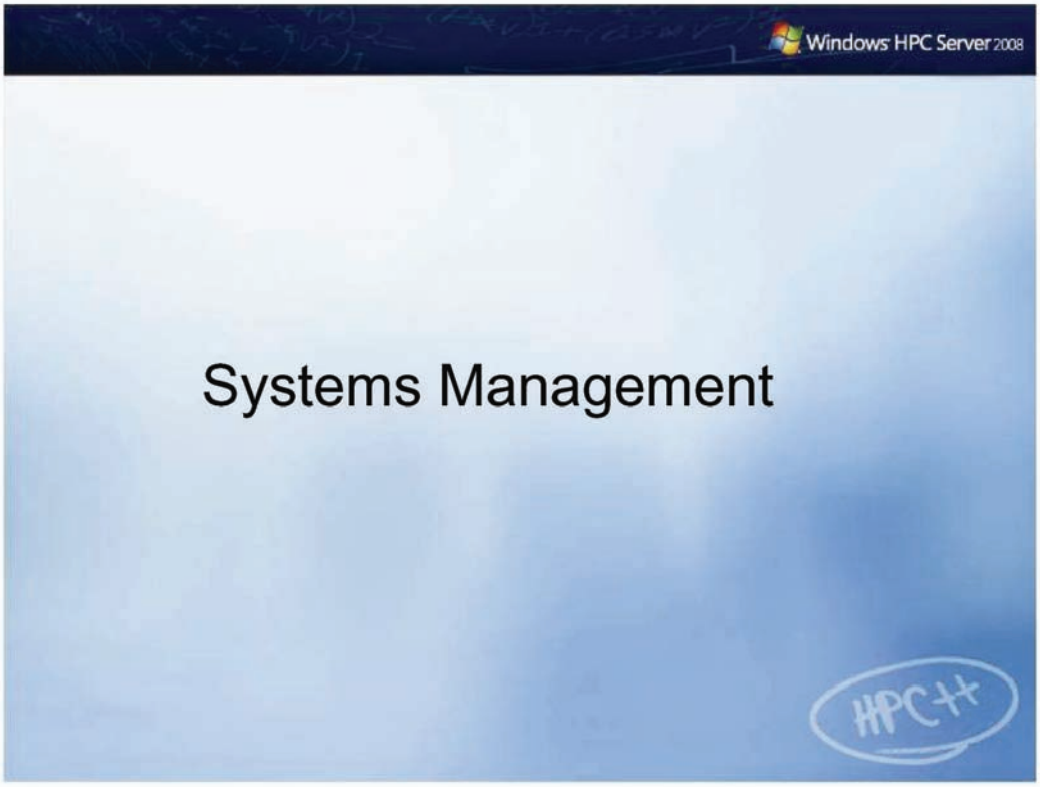
transtec

Demands for Easy Cluster Deployment

- **WANTED: versatility:**
  - clusters may be partitioned into different entities (different CPUs, different amount of RAM or disk space, various applications etc.)
  - different purchase generations and extensions during lifecycle lead to inhomogeneities in hardware
  - stateless vs. diskless vs. fully installed nodes
  - different OSses
- **BUT also simplicity:**
  - single point of administration
  - abstraction of node varieties from management software
- **AND: integratability:**
  - cluster should not be an „island“ where everything is different

30 transtec

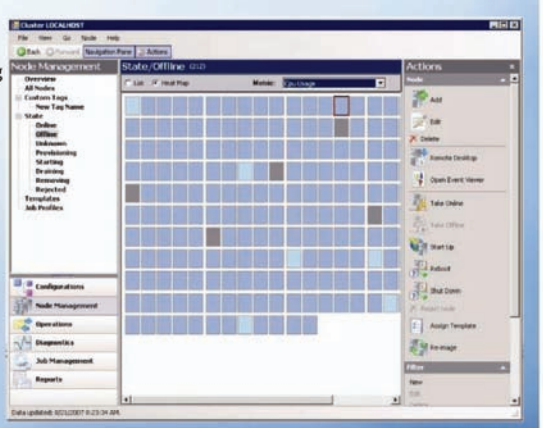

What's new in the HPC Pack 2008?
Windows HPC Server 2008
New System Center UI
PowerShell for CLI Management
High Availability for Head Nodes
Windows Deployment Services
Diagnostics/Reporting
Support for Operations Manager
Systems Management
Job Scheduling
Support for SOA and WCF
Granular resource scheduling
Improved scalability for larger clusters
New Job scheduling policies
Interoperability via HPC Profile
Networking & MPI
Storage
NetworkDirect (RDMA) for MPI
Improved Network Configuration Wizard
Shared Memory MS-MPI for multi-core
MS-MPI integrated with Windows Event Tracing
Improved iSCI SAN Support in Win2008
Improved Server Message Block (SMB v2)
New 3rd party parallel system file support for Windows
New Memory Cache Vendors


Windows HPC Server 2008
Systems Management
HPC++

Systems Management - Overview

Windows HPC Server 2008

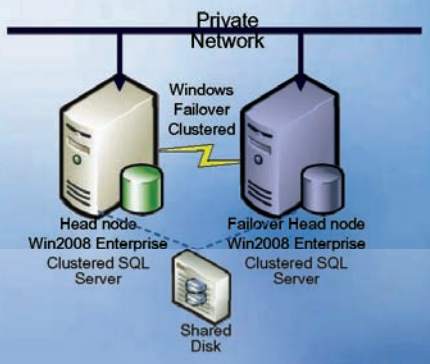
- Simple to setup and manage in a familiar environment
  - Turnkey cluster solutions through OEMs
  - Simplify system and application deployment
    - Base images, patches, drivers, applications
- Focus on ease of management
  - Comprehensive diagnostics , troubleshooting and monitoring
  - Familiar, flexible and "pivotal" management interface
  - Equivalent command line support for unattended management
- Scale up
  - Scale deployment, administration, infrastructure
  - Head node failover
  - Cluster usage reporting
  - Compute node filtering
- Better integration with enterprise management
  - Patch Management
  - System Center Operations Management
  - PowerShell
  - Windows 2008 high Availability Services

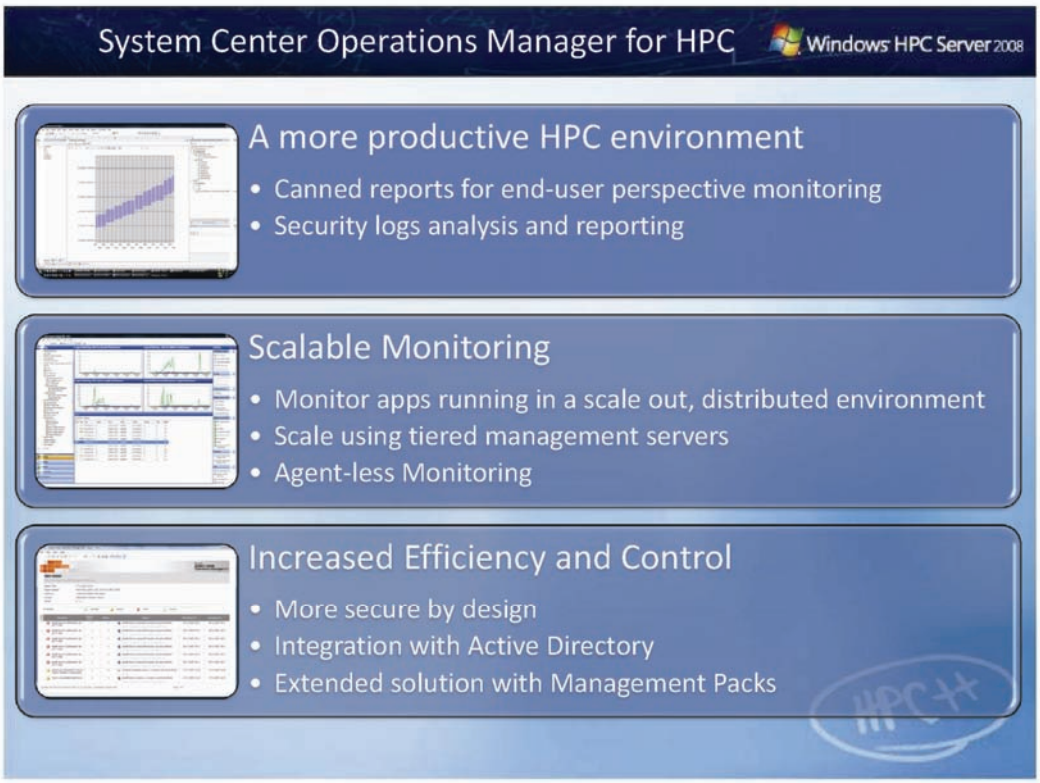


Head Node High Availability

Windows HPC Server 2008

- Eliminates single point of failure with support for high availability
- Requires Windows Server 2008 Enterprise *Failover Clustering Services*
  - Next generation of cluster services
  - Major improvement in configuration validation and management
- HPC Pack Includes
  - Setup integration with Failover Clustering Services
    - Head Node and Failover Node set up with SQL Failover Cluster
    - Job Scheduler services failover
  - Management console linked to Windows Server Failover Management console



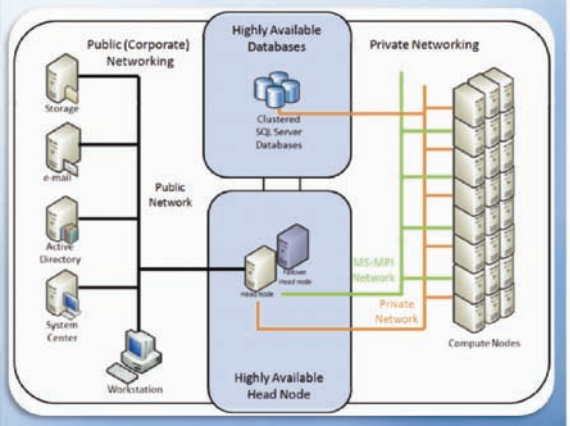

System Center Operations Manager for HPC
Windows HPC Server 2008
A more productive HPC environment
• Canned reports for end-user perspective monitoring
• Security logs analysis and reporting
Scalable Monitoring
• Monitor apps running in a scale out, distributed environment
• Scale using tiered management servers
• Agent-less Monitoring
Increased Efficiency and Control
• More secure by design
• Integration with Active Directory
• Extended solution with Management Packs
HPC++


Windows HPC Server 2008
Job Scheduling
HPC++

What's new in in Job Scheduling

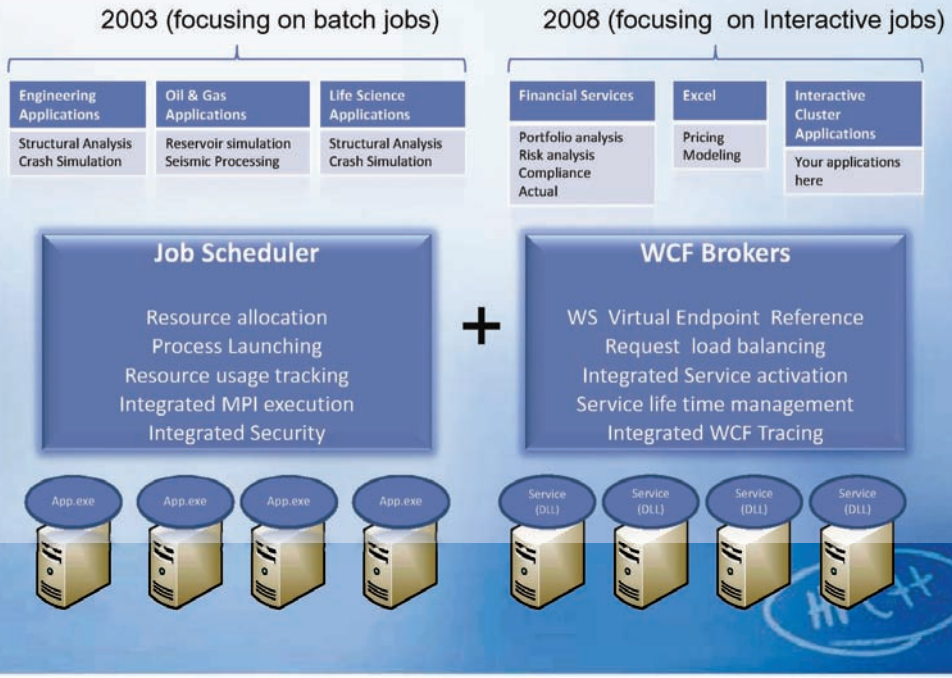
Windows HPC Server 2008

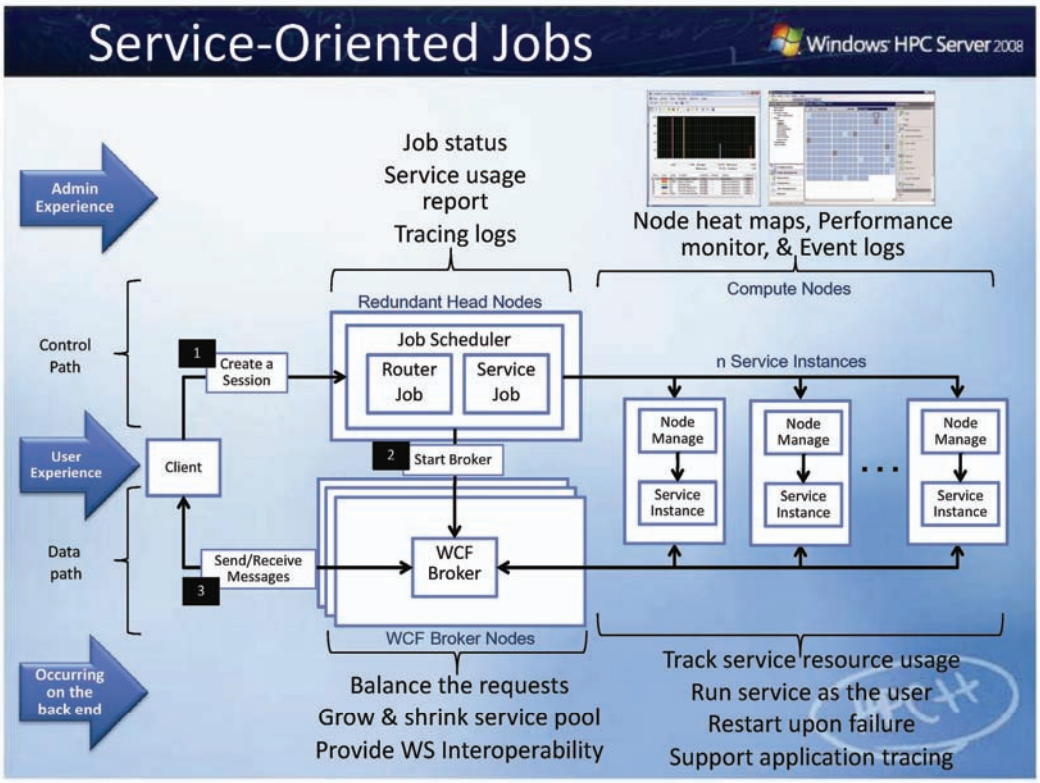
- Broader application support
- Expanded Job Policies
- Support for Job Templates
- Improve interoperability with mixed IT infrastructure

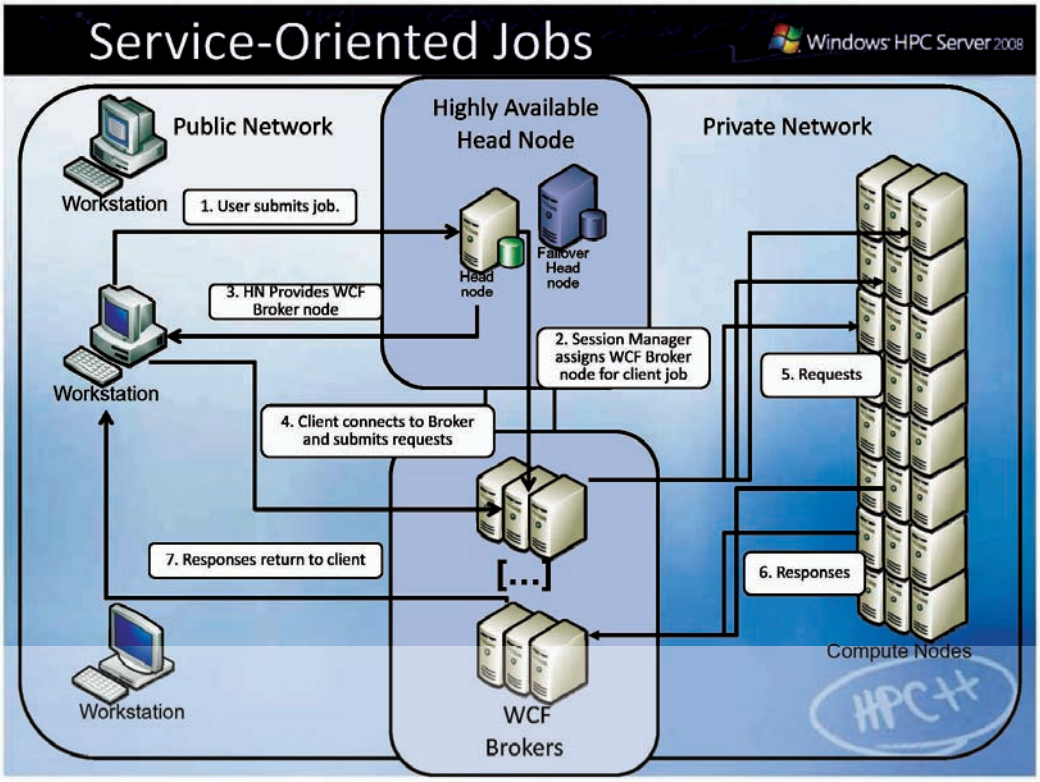


Broader Application Support

Windows HPC Server 2008




Service-Oriented Jobs
Windows HPC Server 2008
Admin Experience
Job status
Service usage report
Tracing logs
Node heat maps, Performance monitor, & Event logs
Redundant Head Nodes
Compute Nodes
Control Path
1
Create a Session
Job Scheduler
Router Job
Service Job
n Service Instances
User Experience
Client
2
Start Broker
Node Manage
Service Instance
Data path
Send/Receive Messages
3
WCF Broker
WCF Broker Nodes
Occurring on the back end
Balance the requests
Grow & shrink service pool
Provide WS Interoperability
Track service resource usage
Run service as the user
Restart upon failure
Support application tracing


Service-Oriented Jobs
Windows HPC Server 2008
Public Network
Highly Available Head Node
Private Network
Workstation
1. User submits job.
Head node
Failover Head node
3. HN Provides WCF Broker node
2. Session Manager assigns WCF Broker node for client job
5. Requests
4. Client connects to Broker and submits requests
7. Responses return to client
6. Responses
Compute Nodes
WCF Brokers

Job Templates

Windows HPC Server 2008

What are they?

- Encapsulates application resource requirements, job prioritization and resource allocation policies
- Created by the admin, ACL'ed to a set of users so that the resources can be delivered to the users in ways that meet the organization's business priorities

How to Define?

- Define the Value range: defines the range of values that the term can take. The value range can be empty. If the value provided at the job submission is not in the defined range, the job is rejected.
- Provide a default value: if a profile defines the default value, the user is not required to specify the job term

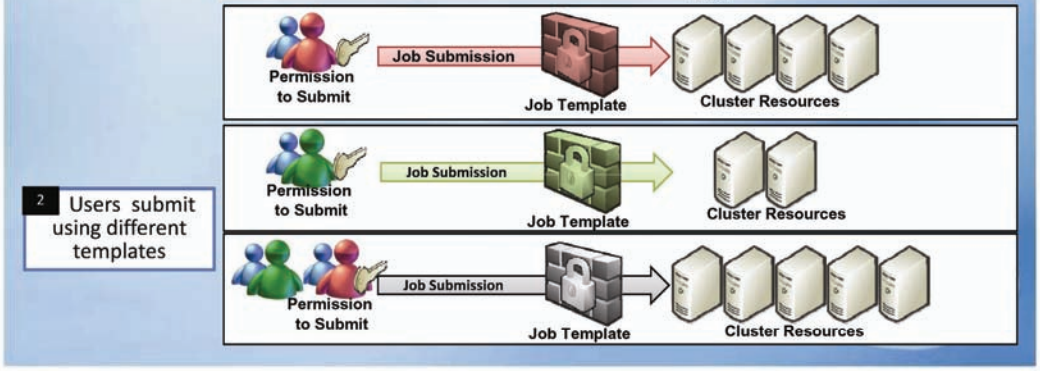
Job Templates

Windows HPC Server 2008

1 Admin creates resource partitions by creating job submission policies

Example 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 Range: N/A 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 Range:



2 Users submit using different templates

Five New Scheduling Policies

Windows HPC Server 2008

- Resource Matching
  - Job submit /nodegroup:appX myapp.exe
- Job Admission Policies via Templates
  - Job submit /template:groupY myapp.exe
- Multi-Level Processor Allocation
  - job submit /numsockets:4-8 myapp.exe
- Adaptive Allocation (Grow/Shrink)
  - job submit myapp.exe
- Preemption
  - Cluscfg PreemptionEnabled=true

Scenario: Placement via Job Context

Windows HPC Server 2008

node grouping, job templates, filters

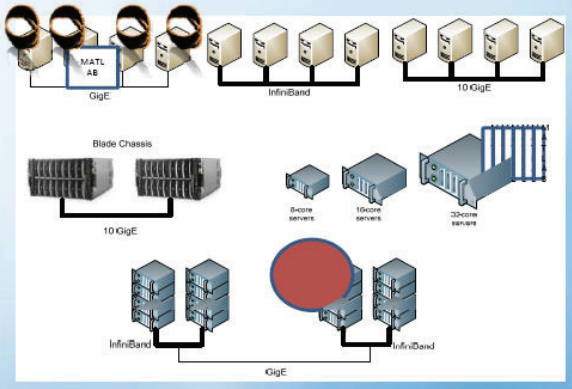
Application Aware

An ISV application (requires Nodes where the application is installed)

Capacity Aware

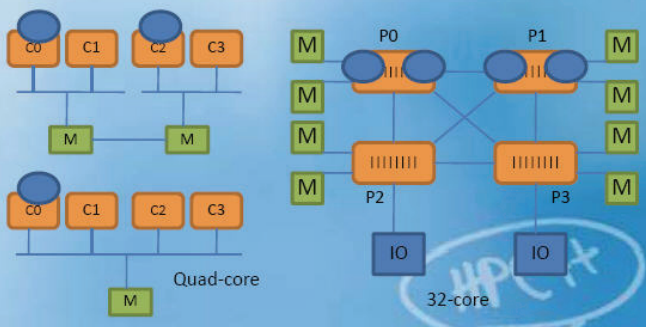
Multi-threaded application (requires machine with many Cores)

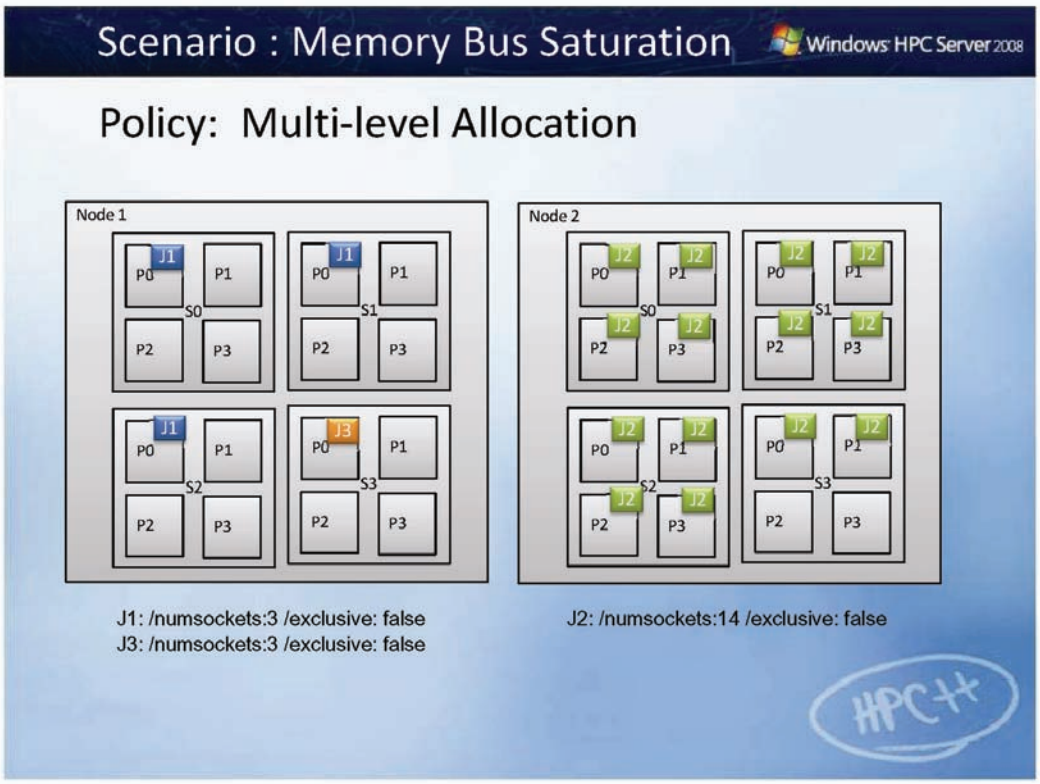
A big model (requires Large memory machines)

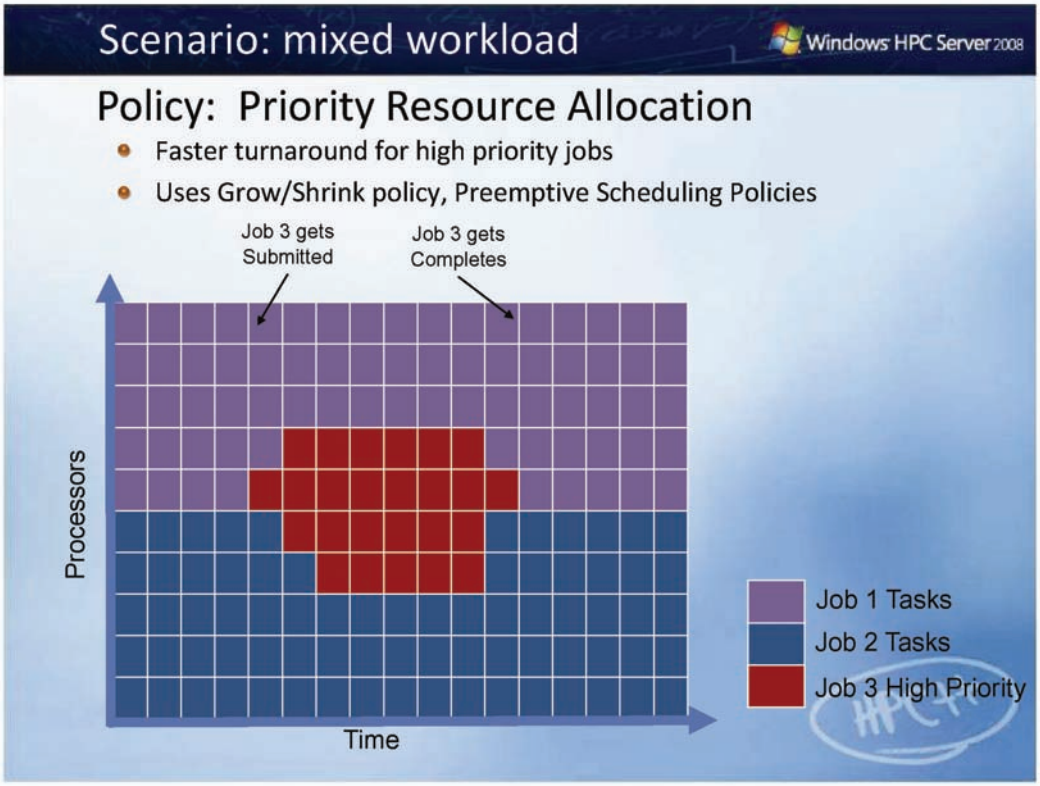


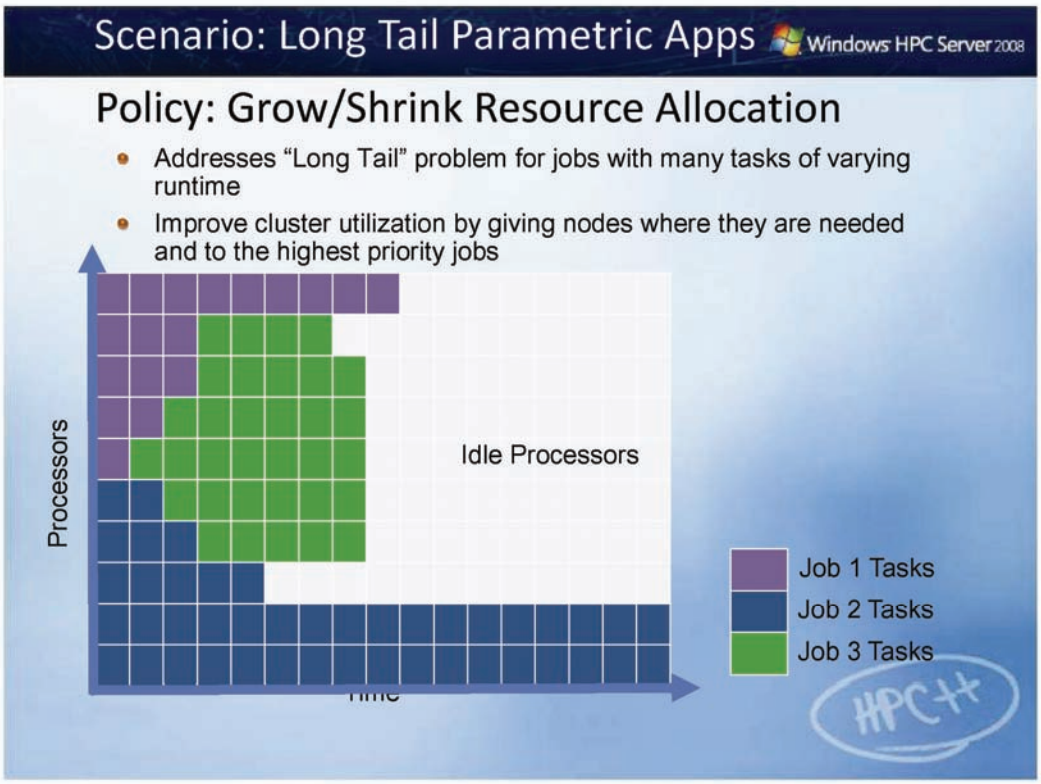
Numa Aware

4-way Structural Analysis MPI Job









Job Policies Summary

Windows HPC Server 2008

Features	Description	Version
Priority FCFS	Jobs are placed in the queue based on priority & submission Time. Jobs are in the queue on a First-Come First-Serve basis in the order they are submitted, except that ALL Highest Priority jobs are ahead of all Above Normal Priority jobs which are ahead of all Normal Priority jobs etc . . .	Window CCS 2003 Windows HPC 2008
Backfilling	Resources are reserved based on resource allocations. If the Job Scheduler identifies open windows for resources within the reserved timelines, it can backfill by selecting smaller jobs for execution within this window	Window CCS 2003 Windows HPC 2008
Exclusive Scheduling	When a job is "Exclusive" no other job can run on a node with that job. When a task is "Exclusive" no other task can run on a node with that task.	Window CCS 2003 Windows HPC 2008
Resource Match-making	Schedule against admin defined groups of nodes. We allow scheduling based on a limited amount of hardware properties and ordering by different hardware properties.	Windows HPC 2008
Multi-level Compute Resource Allocation	The Job Scheduler optimally places memory-intensive jobs to avoid contention of memory, delivering maximum and predictable application performance.	Windows HPC 2008
Preemption	Pre-emption occurs when high-priority jobs take resources away from lower-priority jobs which are already running.	Windows HPC 2008
Grow & Shrink Scheduling	A job may not have "uneven resource requirements" it may just need 100 resources to run 1000 tasks. With Grow Shrink, the job scheduler can give it 1 resource to get it started, and then additional resources as they become available. Once it has 100 resources and less than 100 tasks, job scheduler can begin taking away the un-needed resources.	Windows HPC 2008

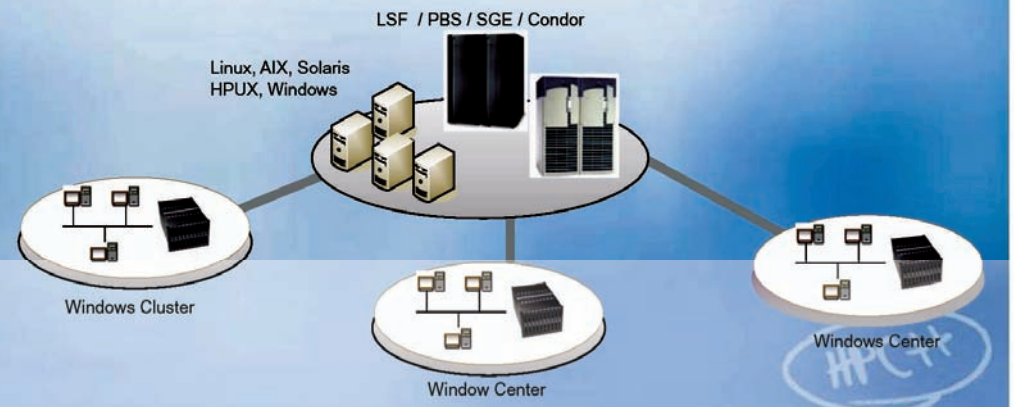
Interoperability & Open Grid Forum

Windows HPC Server 2008

What is it?
- A draft OGSA (Open Grid Services Architectures) interoperability standard for batch job scheduler task submission and management
- Based on web services standards (HTTP, XML, SOAP)

What is its value?
- Enables integration of HPC applications executing on different platforms and schedulers via web services standards

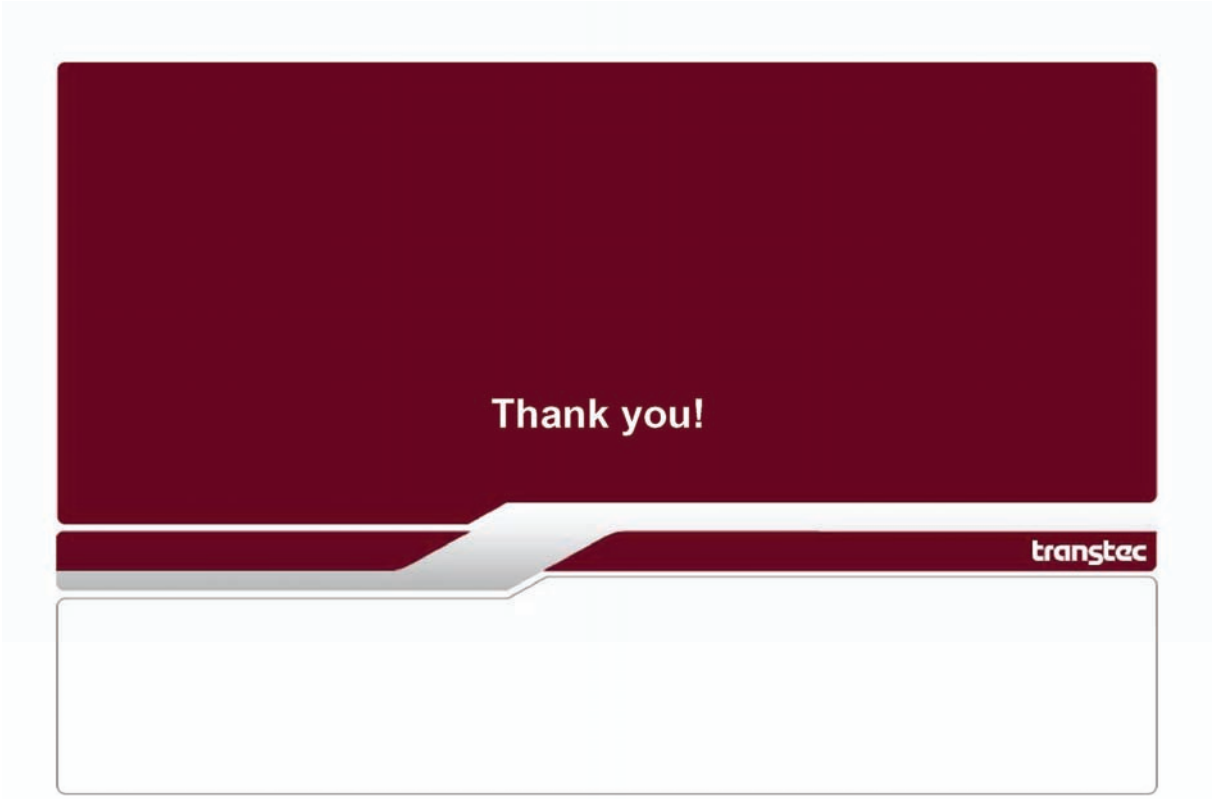
What's the Status?
- Passed the public comment period
- Working on new extensions

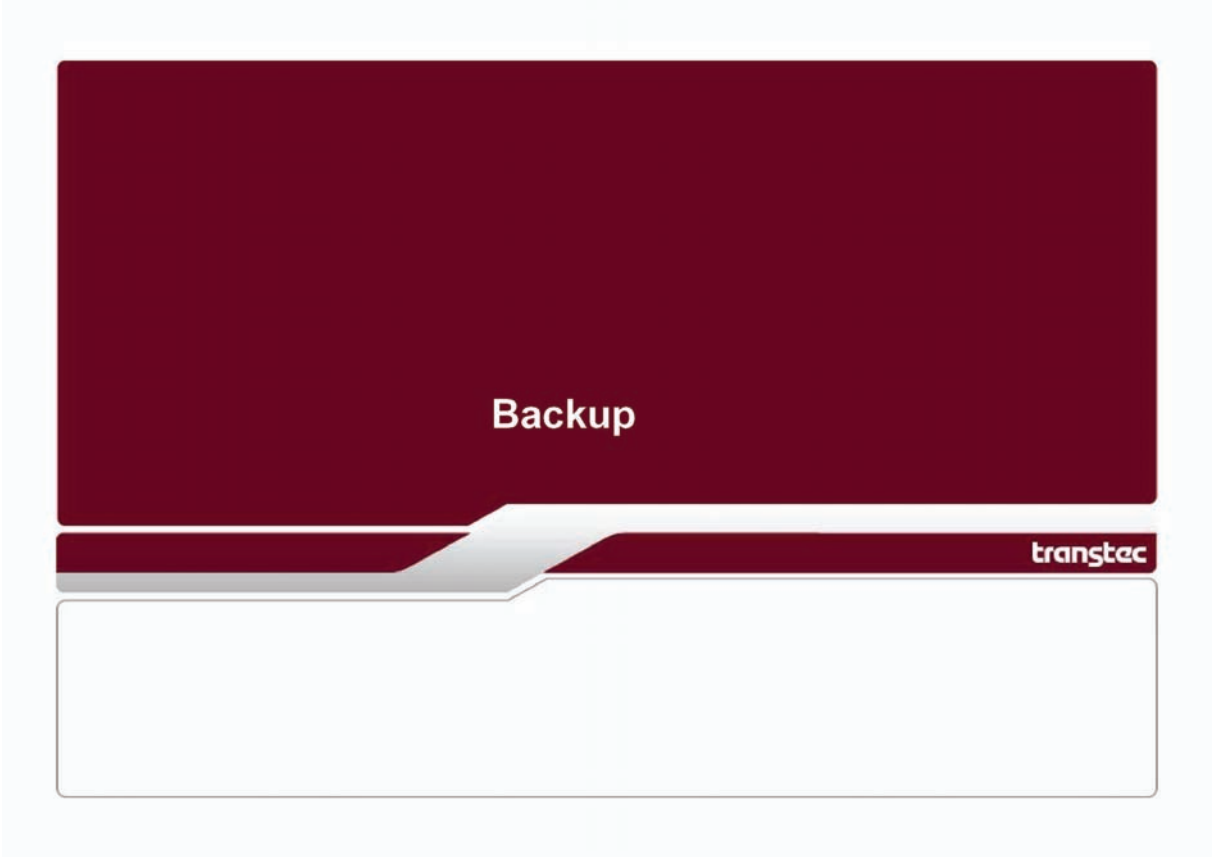


Job Scheduling Summary

Windows HPC Server 2008

Tools	Windows CCS 2003	Windows HPC 2008
End User Tools	1. CLI 2. Job Manager UI	1. Powershell CLI 2. New Job Manager UI with built-in parametric support & custom job filtering
Admin Tools	1. Cluscfg for configuration 2. clusrun	1. Configuration UI 2. Enhanced clusrun
Developer Tools	1. C# API 2. COM API	1. WCF Integration 2. Scalable API with Rowset and eventing support 3. Standard Job Submission Interface (HPC Profile)
Scheduling Policies	1. FCFS 2. Backfill 3. Exclusive Scheduling 4. License scheduling	1. Resource matchmaking 2. Job Template 3. Multi-level processor allocation 4. Adaptive allocation / grow & shrink 5. Preemption


Thank you!
transtec


Backup
transtec