LUSTRE ROADMAP and FUTURE PLANS

Sun HPC Consortium
June 15, 2008

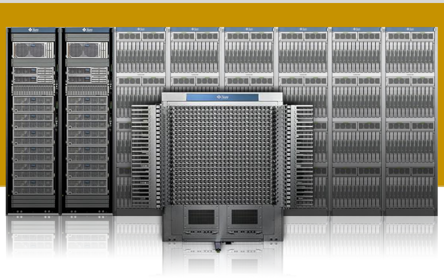
Peter Bojanic
Director Lustre Group
Sun Microsystems

Lustre HPC Filesystem

Open, Seamless and Comprehensive

Access	Developer	Management	OS	Interconnect	Storage/ Archive	Systems
Visualization Workstation, Thin Clients, Remote Access	Compilers, Debuggers, Optimization Tools, Libraries	Workload, Systems and Cluster Management	Linux, Solaris	InfiniBand or Ethernet	Cluster Storage, Backup, Archive, File Systems, HSM	Racks or Blades Variety of CPU Architectures

Sun Services



Sun Customer Ready

What is Lustre?

- Parallel, scalable shared POSIX file system
- Key benefits
  - > Petabytes of storage – one name space
  - > Tens of thousands of clients
  - > High-performance heterogenous networking and routing
  - > High availability
  - > Open source, multi-platform and multi-vendor
  - > Object-based architecture

Lustre Deployments Today

- Largest market share in HPC *(IDC's HPC User Forum Survey 2007)*
- Adopted by the largest systems in the world
  - 7 of top 10 run Lustre, including #1
  - 30% of top 100 (www.top500.org November 2007 list)
- Partners
  - Bull, Cray, DDN, Dell, HP, Hitachi, SGI, Terascala
- Growth in commercial deployments
  - Big wins – oil & gas, rich media, ISPs, chip design

Think

“Extreme HPC”

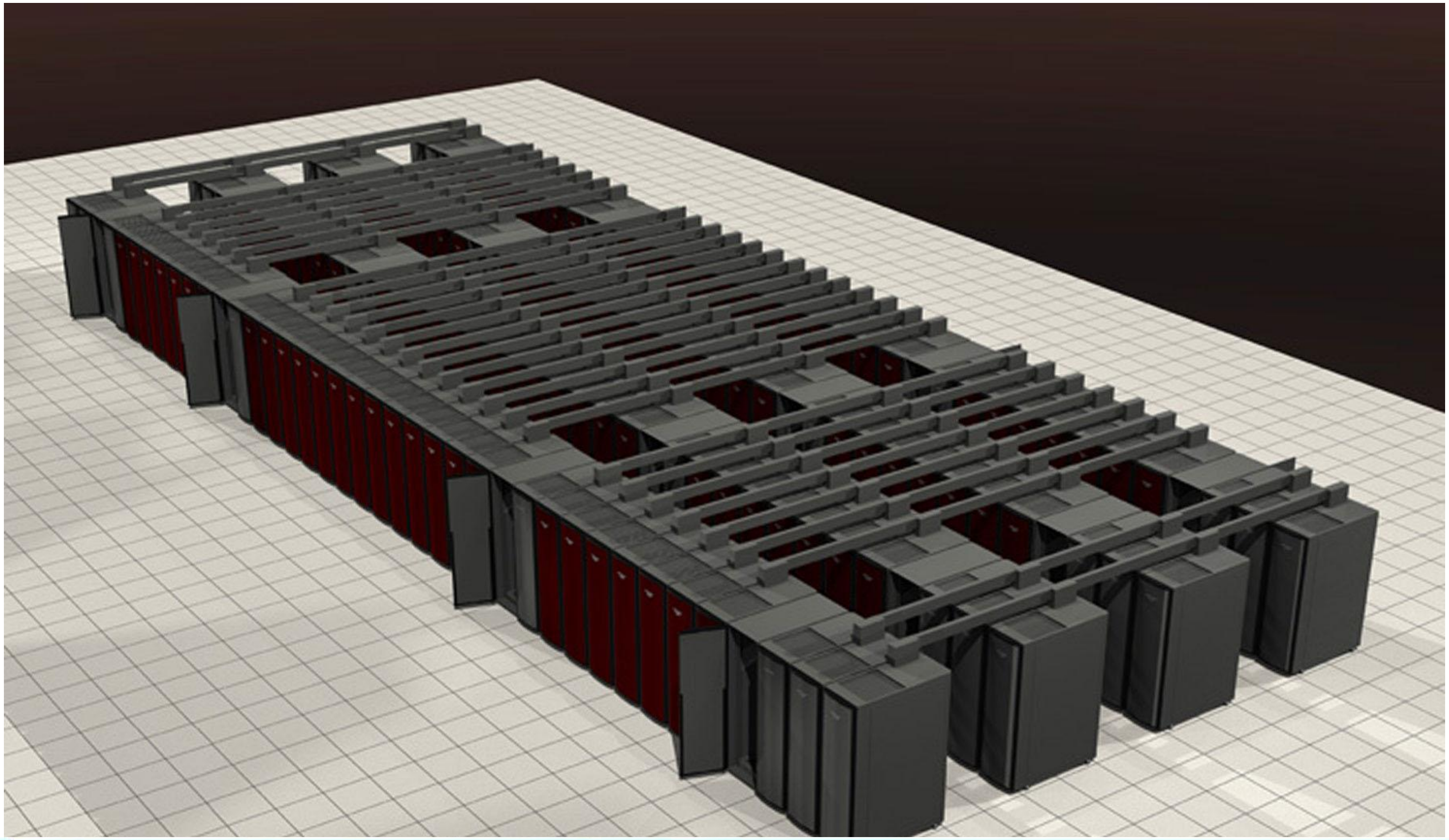
Livermore BlueGene/L (#1 in Top 500)

1.9 PB storage; 35.6 GB/s IO throughput;
212,992 client processes

TACC Ranger (Possible #2 in Top 500)

1.73 PB storage; 40 GB/s IO throughput;
3,936 quad-core clients

Sandia Red Storm (#6 in Top 500)



340 TB storage; 50 GB/s I/O throughput; 25,000 clients

CEA Tera-10 (#19 in Top 500)

1 PB storage; 100 GB/s I/O throughput;
4,352 dual-core clients

Lustre Today

WORLD RECORD	#Clients	Clients: 25,000 – Red Storm Processes: 212,992 – BlueGene/L Can have Lustre root file systems
WORLD RECORD	#Servers	Metadata Servers: 1 + failover OSS servers: up to 450, OST's up to 4000
	Capacity	Number of files: 2Billion File System Size: 32PB, Max File size: 1.2PB
WORLD RECORD	Performance	Single Client or Server: 2 GB/s + BlueGene/L – first week: 74M files, 175TB written Aggregate IO (One FS): ~130GB/s (PNNL) Pure MD Operations: ~15,000 ops/second
	Stability	Software reliability on par with hardware reliability Increased failover resiliency
	Networks	Native support for many different networks, with routing
	Features	Quota, Failover, POSIX, POSIX ACL, secure ports
	Varia	Training, Level 1,2 & Internals. Certification for Level 1

CFS Acquisition

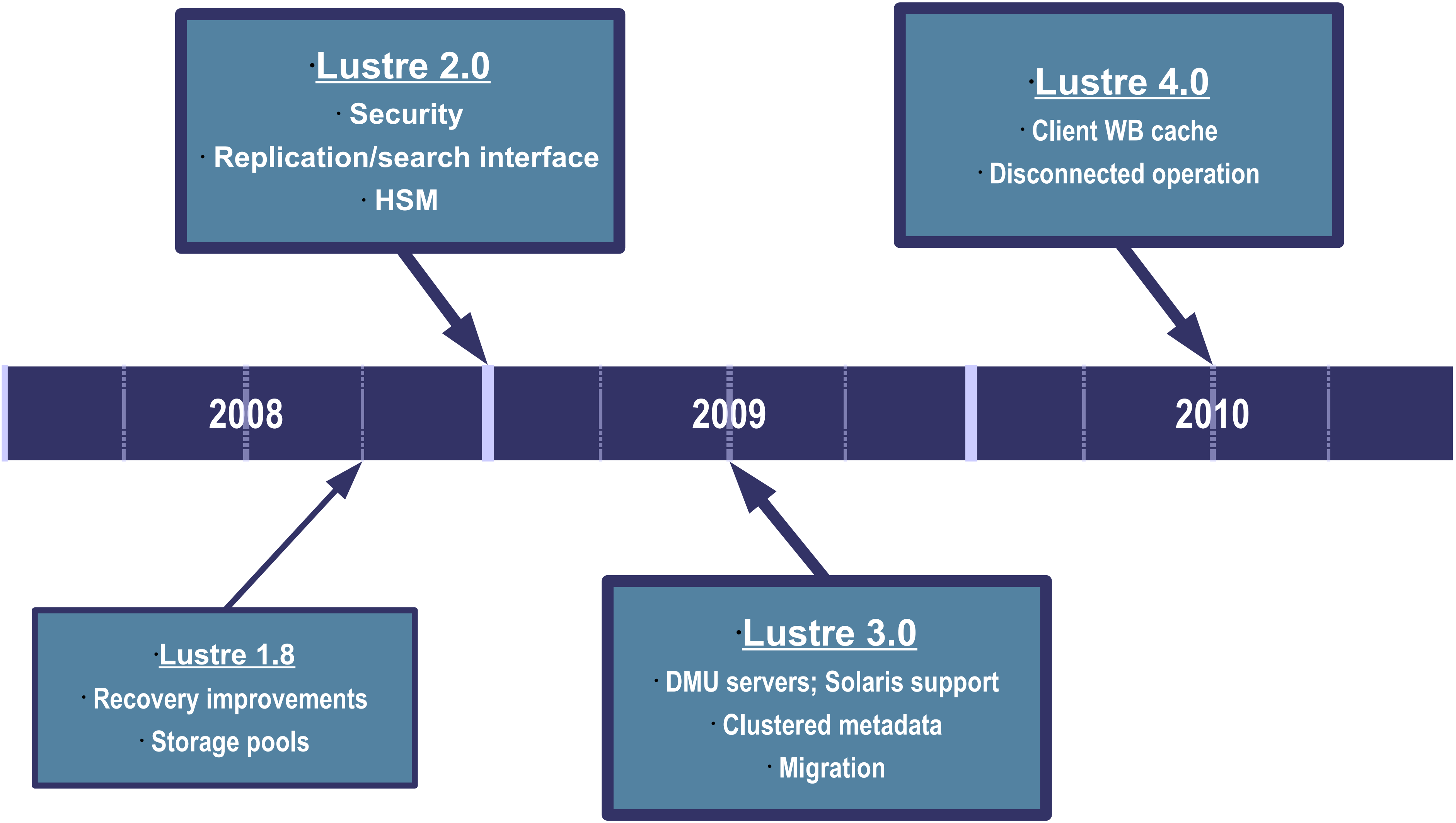
- Oct 1, 2007 the Sun acquisition of CFS closed
  - The theme is continuity
  - No employees, partners or customers were lost
- Lustre remains open source under GPL
  - All designs and the internals course are on lustre.org
  - CVS is open
  - Architecture discussions now on lustre-devel
- Sun continues to work with CFS partners
  - Expanding the network of partners
  - No special versions of Lustre for anyone

New Lustre Partnerships

System
Fabric
Works

STERLING
COMPUTERS

Lustre Roadmap – June, 2008



Lustre 2.0
- Security
- Replication/search interface
- HSM

Lustre 4.0
- Client WB cache
- Disconnected operation

2008

2009

2010

Lustre 1.8
- Recovery improvements
- Storage pools

Lustre 3.0
- DMU servers; Solaris support
- Clustered metadata
- Migration

Lustre Release Taxonomy

- Historically, Lustre release numbers did not accurately reflect major changes to the product
  - > Major architectural changes were previously targeted for releases like 1.6, 1.8
- Transition to a more conventional taxonomy (x.y.z)
  - > x: Major – architectural changes
  - > y: Minor – new features
  - > z: Maintenance – bug fixes

This will help both customers and partners better understand the risk and impact of new Lustre Releases

Lustre 1.8

- Introduces a modest set of new features
  - Recovery improvements
  - Protocol interoperability between b1_6 and HEAD
  - OST Pools
- ServiceTags support
  - Enable Lustre customers to *optionally* register their installations to receive service and training benefits
- Based on b1_6 instead of HEAD branch
  - Substantially reduced risk for customers that want to adopt these features
- Target release: September, 2008

Lustre 2.0

- Major new version of Lustre that introduces substantial architectural changes and features
  - Security (GSS/Kerberos)
  - Replication/search interface
  - Network Request Scheduler
  - HSM
- Based on HEAD branch
  - Provides a foundation for Clustered Metadata (CMD)
  - But no CMD support yet in this release
- Target release: December, 2008

Lustre 3.0

- Introduces Clustered Metadata
  - Code is already in HEAD but will be turned on by default
  - Complete full complement of recovery use cases
- DMU servers and related enhancements
  - Major dependency on ZFS quotas
- Migration and space management tools
  - To migrate from ldiskfs to ZFS storage volumes
  - For space balancing and storage volume evacuation
- Target release: June, 2009

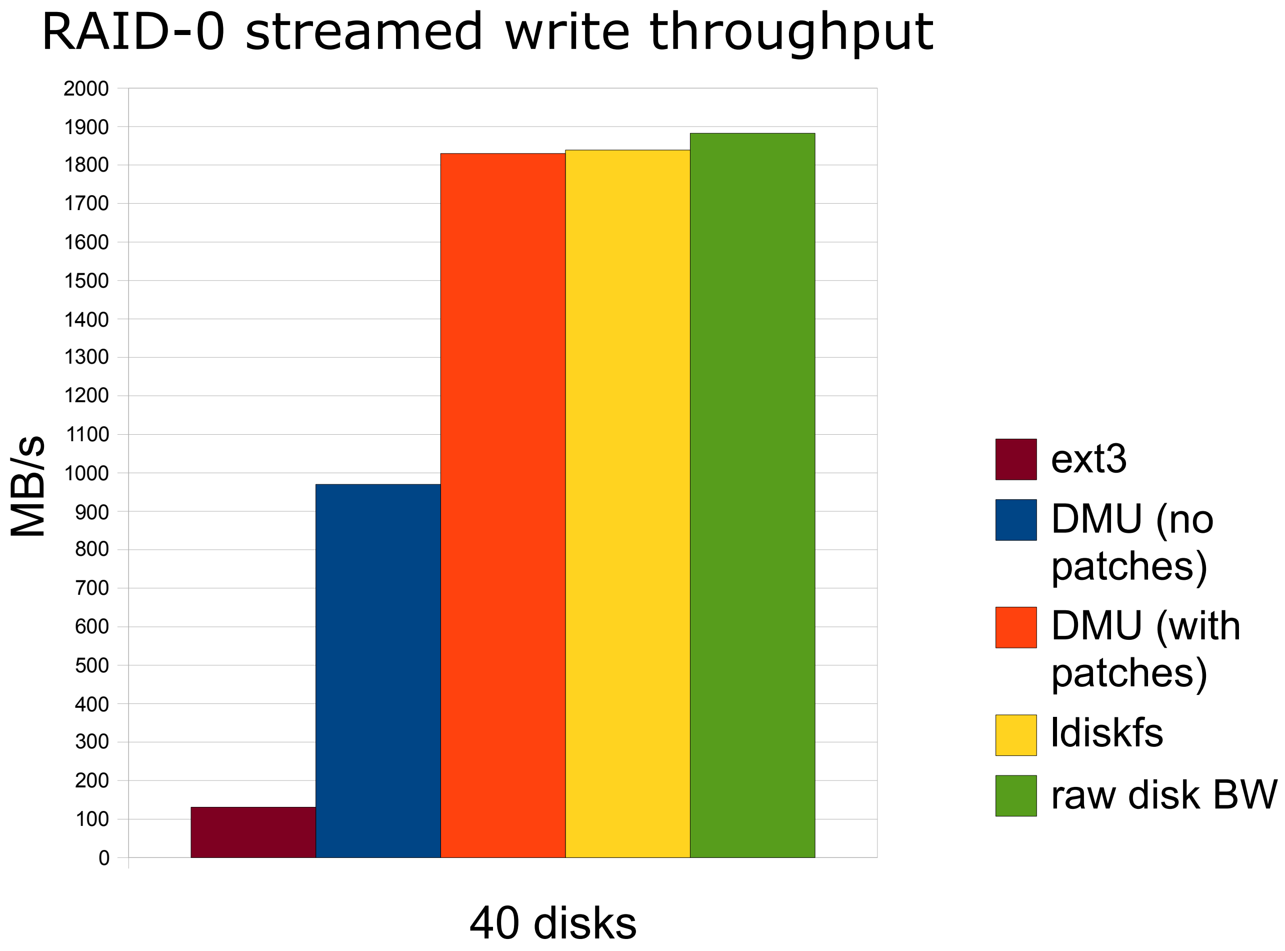
Data Integrity

- Lustre ensures data integrity over the network
  - Compare data before and after network DMA
  - When the feature was added it discovered a few network cards silently corrupting data!
- ZFS DMU has storage integrity
  - Copy-on-write, transactional design
  - Everything is checksummed
  - RAID-Z/Mirrored protection
  - Background disk scrubbing
  - Self-healing

Lustre + ZFS == End-to-end data integrity

Lustre/ZFS Performance

May, 2008



Data measured on Sun Fire X4500 (Thumper) RAID 0 with RHEL4Update6

Zero-copy and other patches buy us ~90% of raw disk!

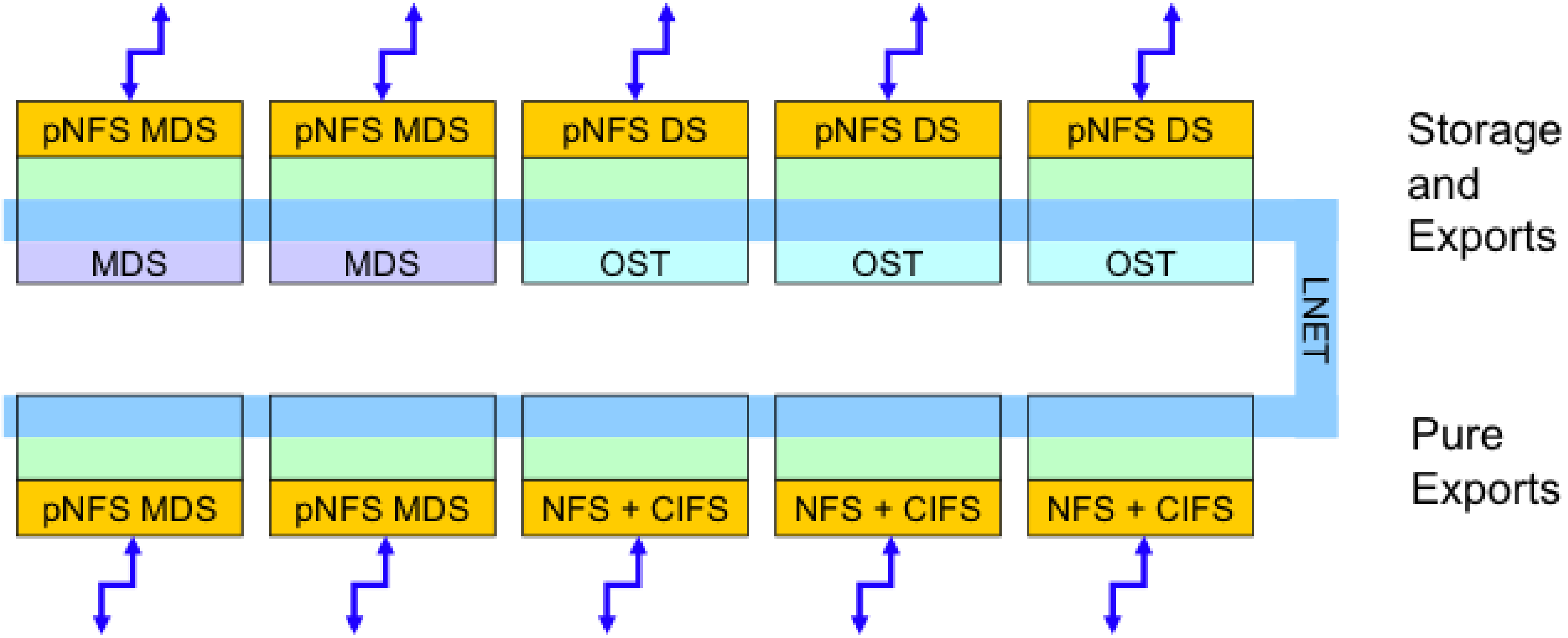
HSM

- Collaboration with CEA in France
  - Most ambitious Lustre community development effort ever undertaken
  - High Level Design (HLD) completed in January, 2008 and shared with Lustre community
  - Presentation by CEA at LUG in Sonoma, CA
- Modular architecture to support multiple HSM engines
  - First interface with HPSS
  - Early planning stages for SAM-QFS integration
  - Discussions underway with SGI for DMF integration

Lustre-pNFS

- pNFS essential for enterprise support
  - Higher performance than monolithic NFS servers
  - Support larger unified name space
- pNFS integration
  - > pNFS exports from Lustre client on Linux
  - > Solaris pNFS protocol servers layered on Lustre
- Make LNET an RDMA transport for NFS
  - > Offer proven Lustre features to NFS standards efforts

Example Clustered Server Config



Client Support

- pCIFS client for Windows
  - > Early customer evaluations in progress
- Clustered Samba (CTDB) Exports
  - > Good performance; purely Open Source solution
- Windows client port
  - > Technology preview expected by the end CY08; production version six months later
- Solaris client port in early planning stages
- Client portability library
  - > Facilitate porting Lustre to Windows and Solaris

Pushing the Limits

- Network Request Scheduler
  - > Achieve higher IO throughput by better coordinating IO across the cluster
- Flash Cache
  - > Read and write-optimized flash cache as “power assist” to DMU
  - > More advanced Flash Cache to accelerate client checkpoint operations
- Client Metadata Writeback Cache
  - > Achieve metadata performance comparable to a local file system

THANK YOU.

Peter Bojanic
pbojanic@sun.com