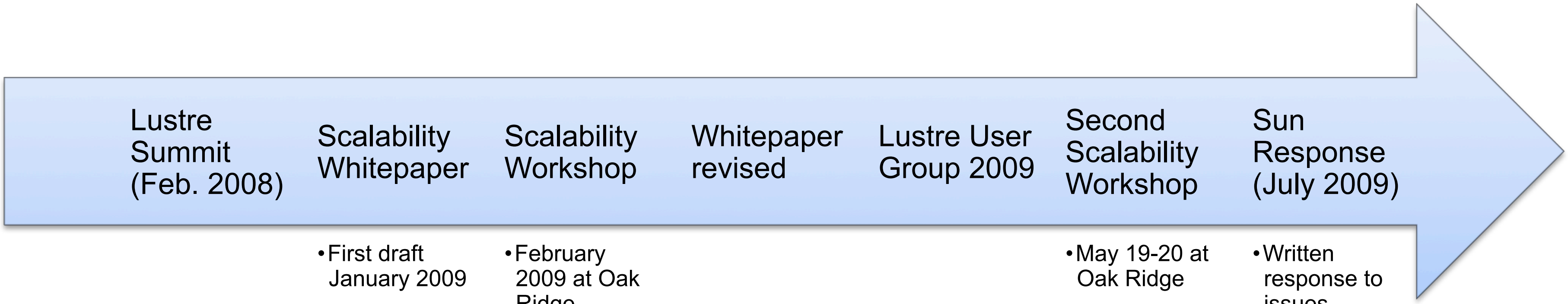
Lustre Scalability Workshop Initial Gap Response

John K. Dawson
Lustre Center of Excellence
Sun Microsystems

Lustre Scalability Planning



Lustre Summit (Feb. 2008)

Scalability Whitepaper

- First draft January 2009

Scalability Workshop

- February 2009 at Oak Ridge
- Focused on 2012 requirements
- Top 10 scalability issues identified

Whitepaper revised

Lustre User Group 2009

Second Scalability Workshop

- May 19-20 at Oak Ridge
- Focused on 2015 requirements

Sun Response (July 2009)

- Written response to issues
- Whitepaper revised
- Lustre roadmap updated

Lustre Scalability

Definition

- Performance / capacity grows nearly linearly with hardware
- Component failure does not have a disproportionate impact on availability

Requirements

- Scalable I/O & MD performance
- Expanded component size/count limits
- Increased robustness to component failure
- Overhead grows sub-linearly with system size
- Timely failure detection & recovery

Top 10 Scalability Issues

1. Asymmetric impact of failures
2. Metadata performance improvement
3. Lustre ZFS Licensing
4. Quality of Service support
5. Performance variability
6. Policy Engine
7. Manageable at scale
8. Failover duration
9. Small file performance
10. Wide stripe performance

Asymmetric impact of failures

- Disproportionate impact of failures
- Today
  - Depend on timeouts to detect failures
- Future
  - Imperative recovery and scalable health network to detect failures quickly and quickly evict client
  - Will be able to integrate with external RAS

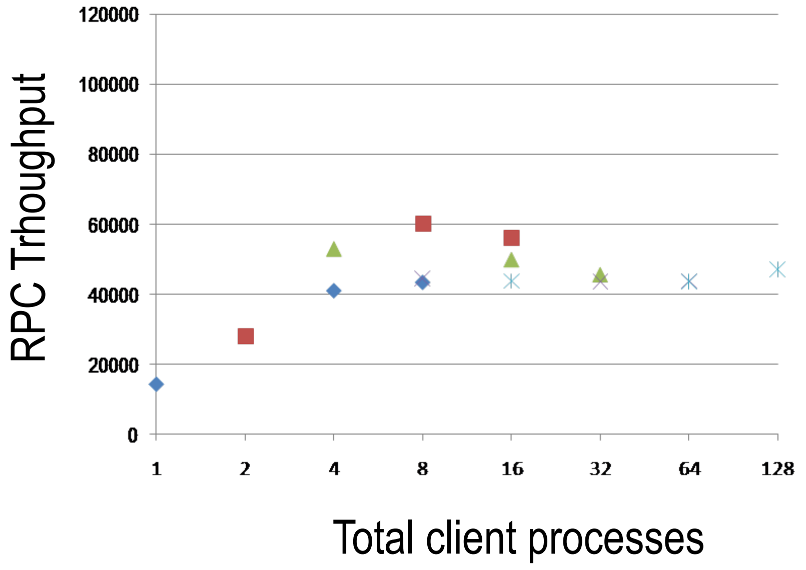
Metadata Performance Improvement

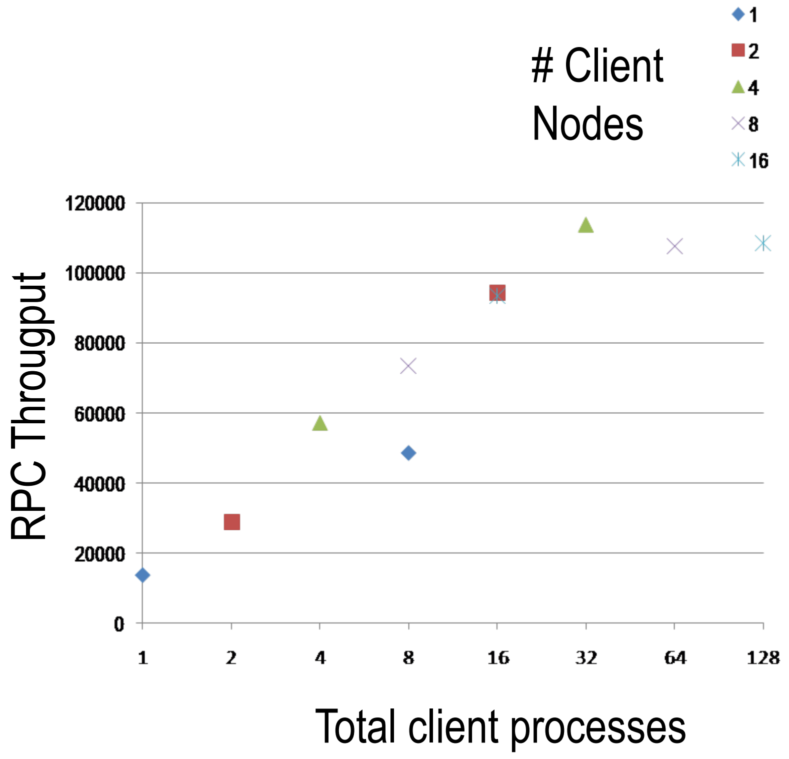
- Short term
  - SMP Scaling improvements
- Long term
  - Clustered Metadata

Performance Improvements

SMP Scaling

- Improve MDS performance / small message handling
- CPU affinity
- Finer granularity locking





Lustre ZFS Licensing

- We have a solution that we're confident will enable Lustre and ZFS to be linked, modified and redistributed by the entire Lustre community
- We have discussed it with OEM's and they support it
- Nest step is to discuss with strategic customers
- Plan to announce by the end of June

Quality of Service Support

- NRS will provide a basis for this
- Will provide quanta of service based on user, machine etc.
- Gang scheduling of quanta should be integrated with job schedulers

Performance Variability

- Mike Booth at LCE is investigating this
- One area to investigate is providing topology awareness for IO libraries so they can layout files in a kinder way
- NRS will help as well

Policy Engine

- There is a policy engine in HSM
- We will describe what hooks Lustre will provide

Manageable at Scale

- Still in progress

Failover Duration

- Health network and imperative recovery mentioned above address this

Small File Performance

- Key is aggregating multiple requests in single RPC
- WBC will allow MD ops to complete on client before they are flushed to the server and to be sent to the MDS in bulk
- Also considering keeping small files on MDS

Wide Stripe Performance

- Still under investigation
- Can potentially exploit collective IO libraries by extending them to make them layout aware

Q & A

Scalability Whitepaper online at:

http://ornl-lce.clusterfs.com/index.php?title=Image:LustreScalabilityWP_Updated.pdf

THANK YOU

John K. Dawson
Lustre Center of Excellence
Sun Microsystems

Lustre Scalability

Attribute	Today	Future
Number of Clients	Flat communication model (10,000’s)	Hierarchical communication Proxy servers IO forwarders 1,000,000’s
File system/LUN size	Ext3	ZFS
Metadata Performance	Single MDS	Clustered Metadata Servers
Recovery Time	Scales O(n)	Health Network to scale O(log n)