SNIA ESF | ETHERNET STORAGE

Clustered File Systems: No Limits

James Coomer, DDN
Jerry Lotto, Mellanox
John Kim, SNIA-ESF Chair, Mellanox

October 25, 2016

SNIA Legal Notice

SNIA ESF | ETHERNET STORAGE

- The material contained in this presentation is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

About SNIA

SNIA ESF | ETHERNET STORAGE

SNIA-At-A-Glance

160
unique member companies

3,500
active contributing members

50,000
IT end users & storage pros worldwide

Learn more: **snia.org/technical** **@SNIA**

Today's Presenters

James Coomer
Technical Director
DDN

Jerry Lotto
Director, HPC
Mellanox

John Kim
SNIA-ESF Chair
Mellanox

Alex McDonald
SNIA-ESF Vice Chair
NetApp

Agenda

- SNIA Background
- Overview
  - When should I consider a Clustered File System?
  - What are Clustered File Systems?
    - Scale-out NAS and Parallel file systems
  - What choices are there and how do I choose?
  - What sort of performance can I expect?

Who Needs Clustered File Systems?

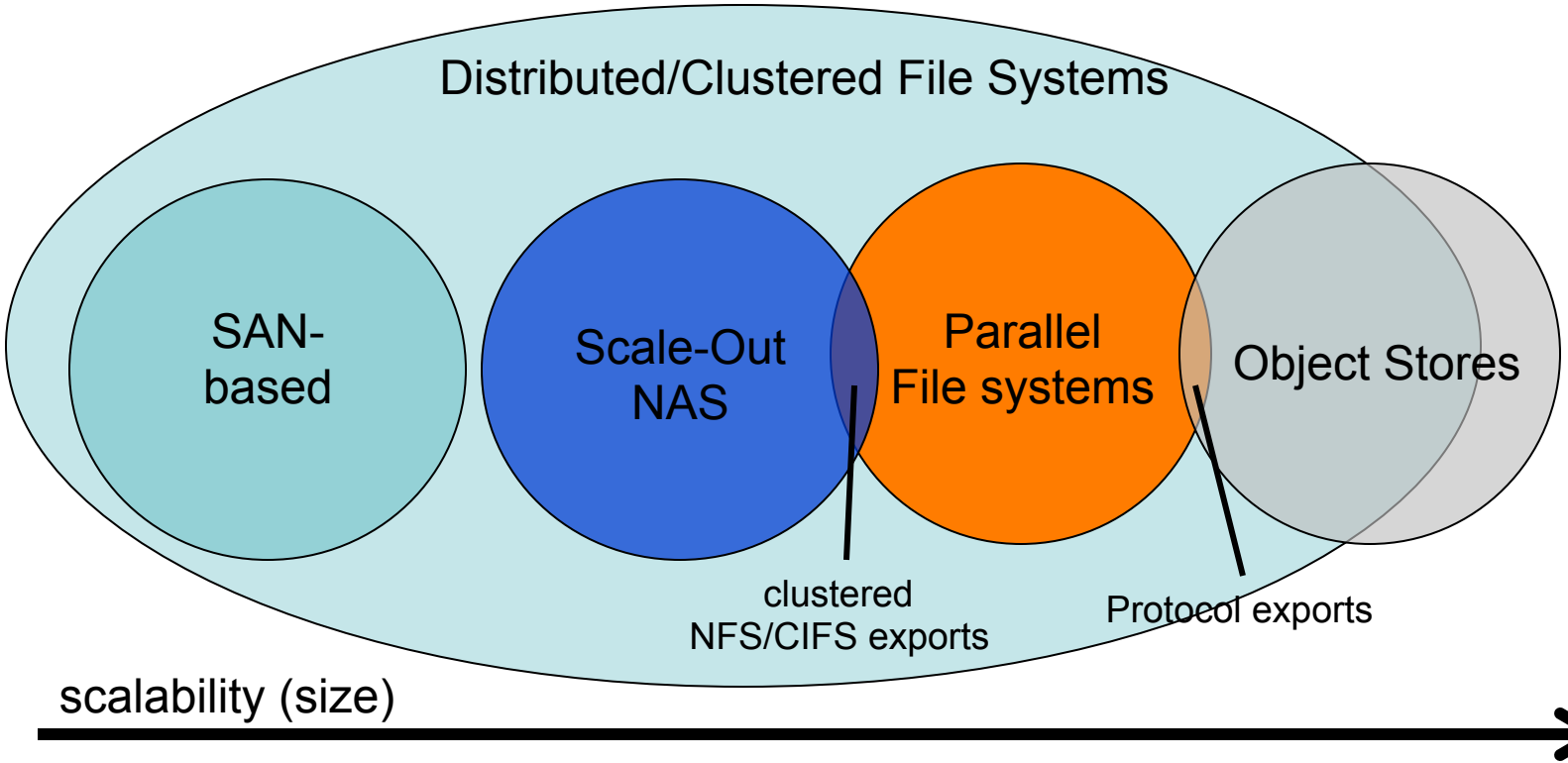
- Traditional storage is block (SAN), file (NAS), or object
  - SAN: High performance but no data sharing
  - NAS: Shared access but limited scalability
  - Object: Scalable but often slow, no file writes/edits
- What if you need it all?
  - More performance than one storage system can offer
  - Many hosts share read/write access to the same data
  - Expand performance and capacity simultaneously

What's the Difference?

- Clustered File Systems: multiple servers sharing IO load
- Distributed File Systems: no shared back-end storage
- Parallel File Systems: + native, intelligent client

- All can distribute data across multiple systems
  - Allow multiple clients to access data in parallel
  - Scale up to petabytes per cluster
  - Support high bandwidth

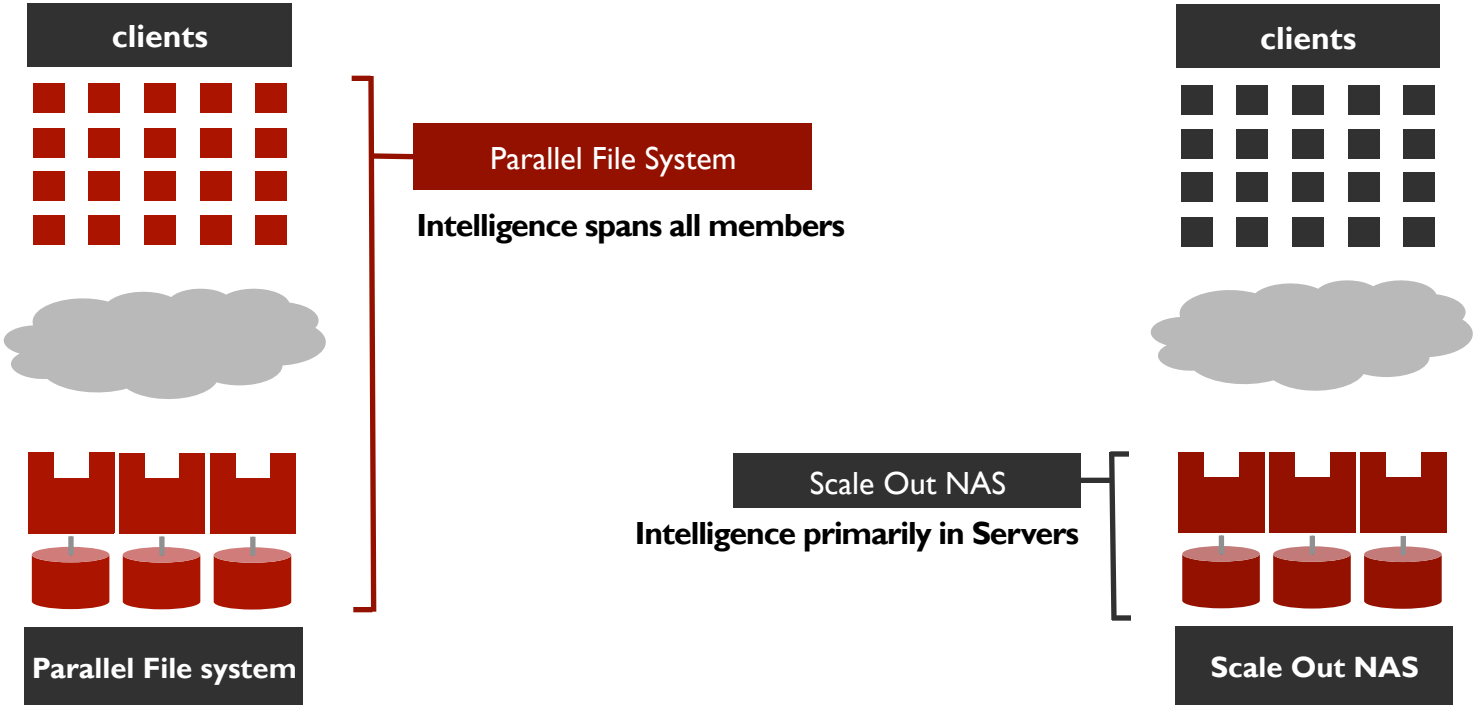
Clustered File System Types



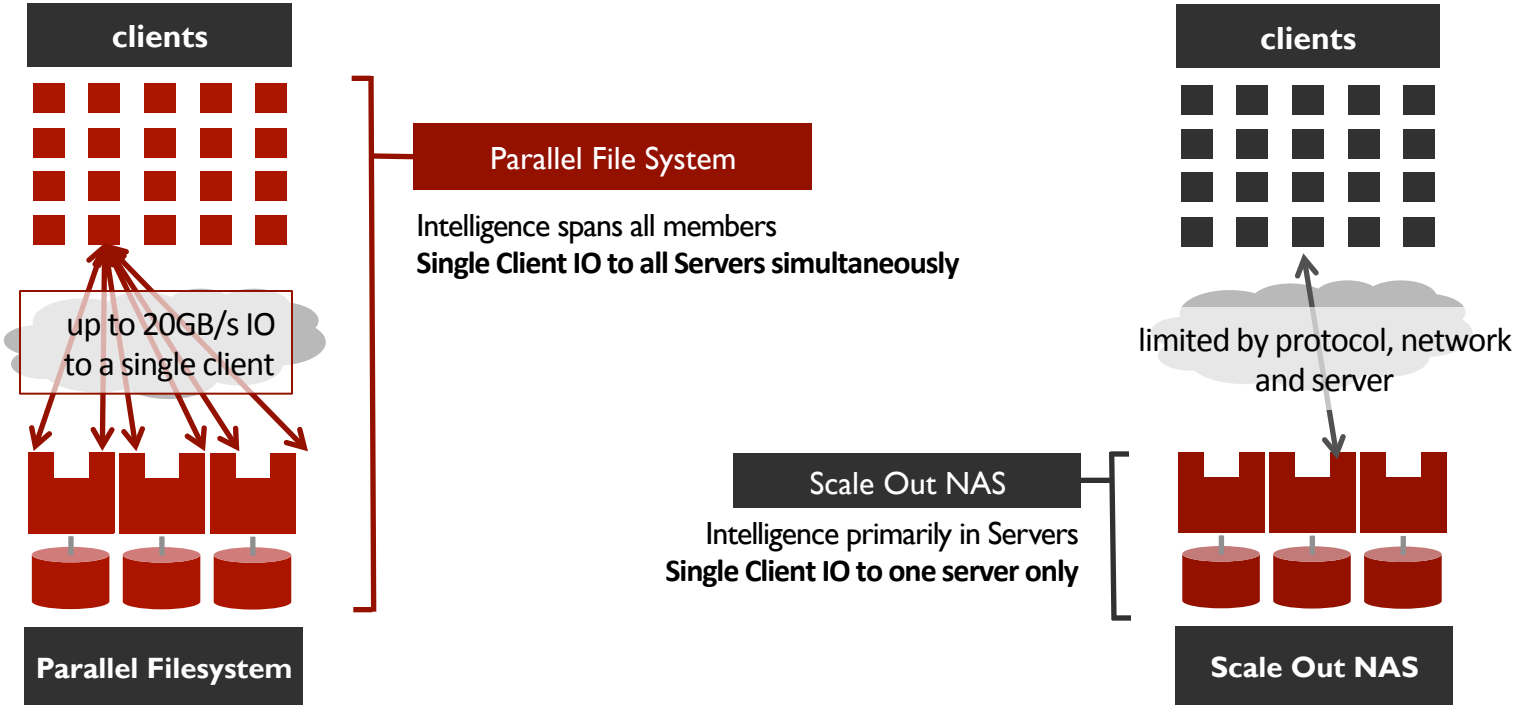
Scale-Out NAS vs. Parallel File Systems

- Both feature high availability and shared namespace
  - Access via special clients or NAS protocols
- Differ in how much file system responsibility shared with the clients
  - Scale-Out NAS: clients are relatively simple, NFS/CIFS + some optimization. Client-side setup is easy. All intelligence and scaling challenges must be handled in the servers
  - Parallel File Systems: clients software must be installed on all clients which need high performance. More complex, but very large performance and scalability gains. The intelligence and scaling challenges are shared with the clients.

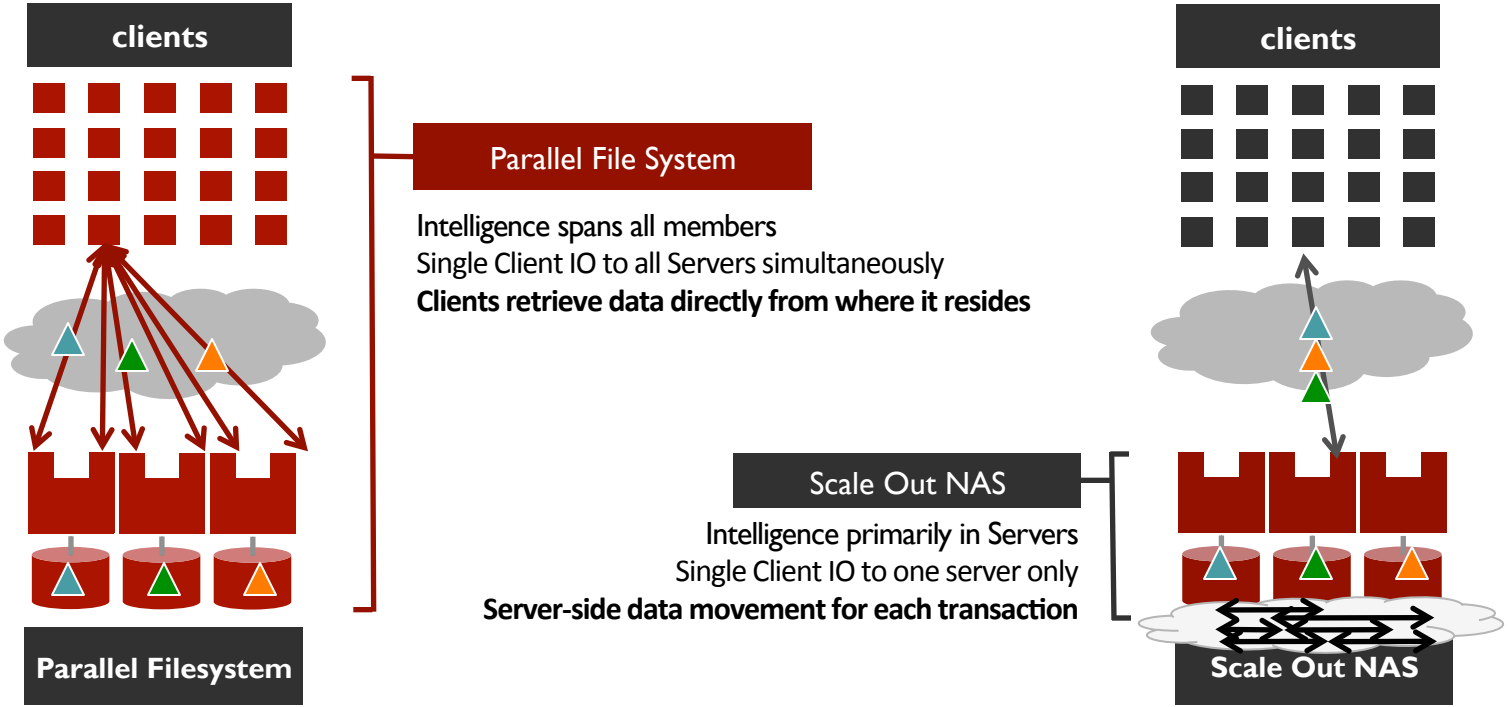
Parallel File System vs Scale-out NAS



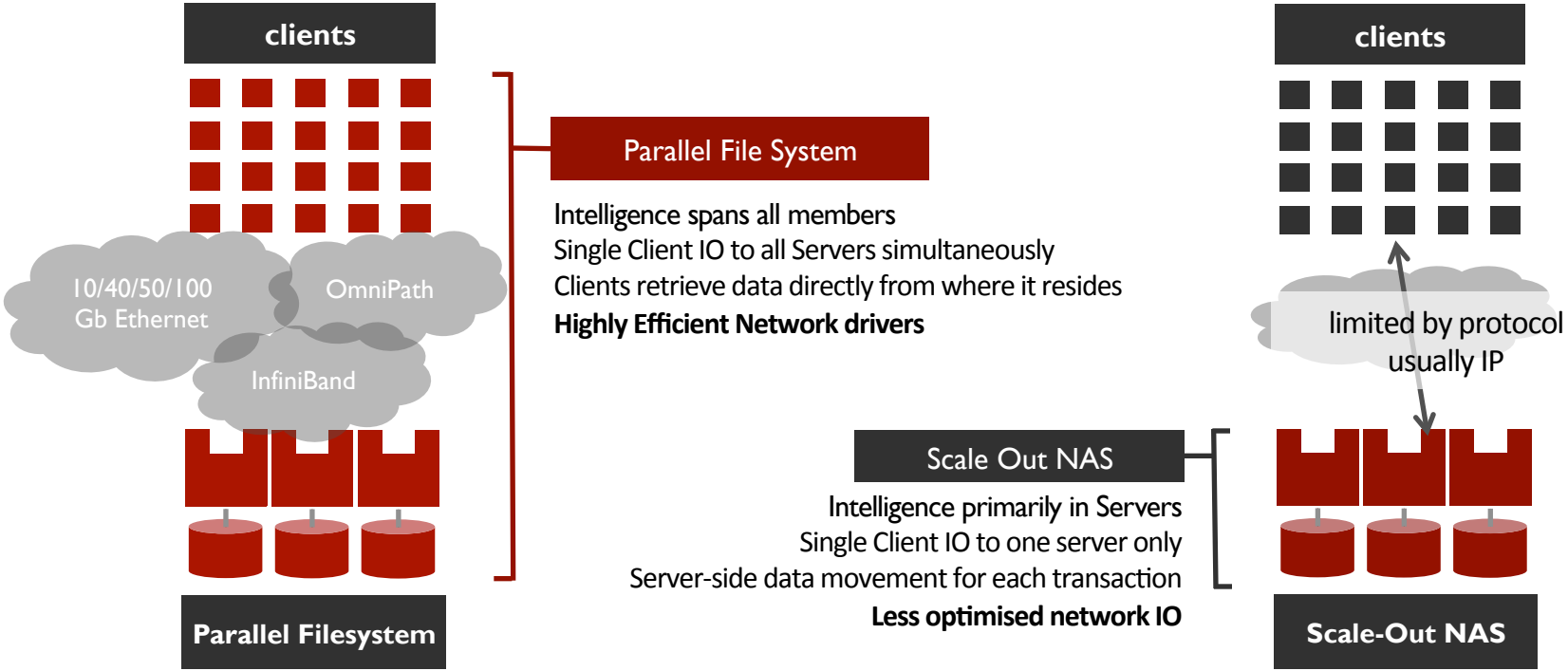
Parallel File System vs Scale-out NAS



Parallel File System vs Scale-out NAS



Parallel File System vs Scale-out NAS



So what are my choices?

- Parallel File Systems Include:
  - Lustre (Intel), Spectrum Scale/GPFS (IBM)
  - BeeGFS (Fraunhofer), OrangeFS/PVFS
  - Others: StorNext, HDFS, MooseFS, Gluster, Ceph, etc.
- Differences in data distribution, metadata, clients, licensing/cost, sharing/locks, data protection, etc.
- We shall concentrate on the most widely-deployed filesystems today: **Lustre** and **Spectrum Scale**

Lustre and Spectrum Scale

- Both are benefitting from strong recent development efforts:
  - **Spectrum Scale:** Active File Management, High Availability Write Cache, Local Read Only Cache, Encryption, GPFS Native RAID
  - **Lustre:** QoS, JobStats, Security
- Lustre development primarily at Intel, but significant features are developed by the wider community and other vendors

Super-High Level Comparison

Both Offer High Out-of-the-box Performance

- Lustre
  - Optimized for large-scale performance
  - flexible per-file/dir/fs striping policies
  - Strong QoS available
- Spectrum Scale
  - Optimized for small/ medium-scale performance
  - Mature Snapshot
  - Multi-protocol support
  - Data Policy Management Built-in

What about NFS/SMB?

- GPFS supports native Linux and Windows clients
- Lustre only supports native Linux clients
- Both support the ability to export the filesystem via clustered NFS/SMB
  - Combine extreme performance with native clients AND a range of other clients with NFS/SMB
  - GPFS can deliver an extremely strong scalable NAS

Multi-protocol Options

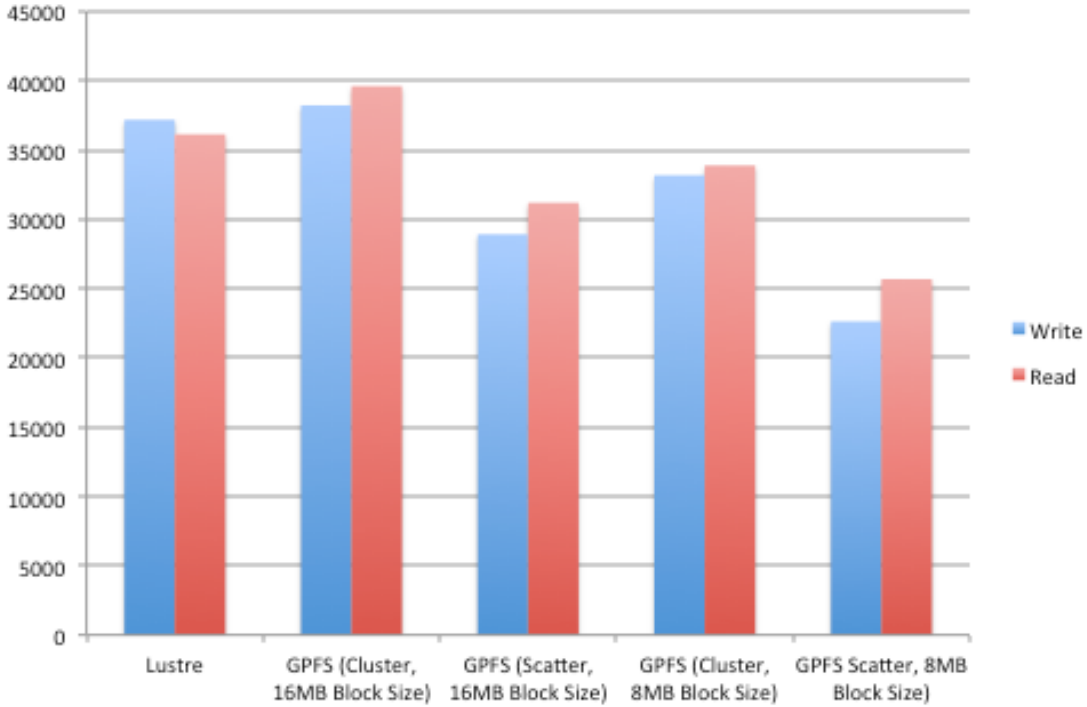
- In addition to native clients, both Lustre and GPFS support protocol gateways
- GPFS introduced a new protocol abstraction layer (CES) recently that supports any/all of object, SMB and NFS simultaneously. Runs on dedicated protocol nodes.
- Lustre supports re-exporting the filesystem with clients acting as Samba (SMB) and/or NFS/pNFS servers
- Performance typically less than native client access

Performance

- Both Lustre and GPFS can reach similar throughput with the same hardware...
  - Pushing the bandwidth limits of the underlying storage devices
- But the devil is in the details:
  - IOPs requirements
  - Metadata performance
  - File system block-size choices
  - Application IO characteristics (small files, mmap, directIO)

Parallel Filesystem Performance

Filesystem Throughput (GB/s): 400 NL-SAS drives



- Like-for-like comparison on DDN GS14K EDR
- Similar top-end throughput
- GPFS performance dependent on choice of data allocation method

Optimizing Clustered Filesystems

- "I have never met a file system or data-intensive workload that didn't respond well to tuning"
- Two key considerations
  - Architecture
    - Choice and design of the file system and interconnect – both are important!
  - Tuning
    - Parallel clustered filesystems have extensive tuning options
    - Spectrum Scale 4.2, for example, has more than 700 tuning parameters!
      - Only dozens are focused on end user tuning ☺

Architecture

- Hardware - storage and server choices
  - Technology and Capacity – data and metadata
    - Performance-limiting metric
    - Flash and archive tiers, from RAM to tape!
  - Disk aggregation
    - hardware RAID, ZFS, GNR (GPFS Native RAID)
    - Underlying filesystem(s)
  - Scaling capacity and performance – today and tomorrow?
    - Building block approach

Architecture - continued

- Interconnect choices
  - Disk to servers
    - Internal vs External
      - shared access (HA)
      - share-nothing (ie: FPO)
    - PCIe/NVMEoF, SAS, FC, InfiniBand, Ethernet
  - Servers to clients
    - InfiniBand, Ethernet, etc
    - RDMA supported by both Lustre and Spectrum Scale

Tuning

- Configuration files and parameters
  - Operating system kernel tuning
  - Spectrum Scale – mmlsconfig, mmchconfig
  - Lustre – /proc/fs/lustre and multiple configuration files
- Memory
  - cache (L2) – storage controllers, server AND client
  - other uses (daemons, tables, kernel, policy engines)
  - controlled (directly and indirectly) by many parameters

Tuning - continued

- Flash cache (L3) – write and/or read
  - Integral vs. policy-driven placement and migration
  - Sub-LUN, underlying file system level (L2ARC), AFM
- File system blocksize(s) – sub-blocks too
- Communications protocols and fabrics
  - TCP/IP required, RDMA optional but important for performance
  - Protocol overhead and gateway architectures

The Three “M”s

- Monitoring, Management, Maintenance
  - Monitoring and Management
    - Lustre – IEEL GUI
    - Spectrum Scale GUI – Originally GSS, now all
      - Excellent – health and performance monitoring
      - Good – code management, troubleshooting
      - “In development”– Deployment, configuration, tuning
    - Ganglia, Nagios and Splunk plugins
    - Daemons and command line

The Three “M”s

- Monitoring, Management, Maintenance
  - Maintenance
    - Lustre available as open source
    - Different cost models for support
    - Similar long-term costs for file system support
    - Client vs Server upgrades
      - Rolling vs. Downtime
      - Prerequisites and dependencies

More Information

- Spectrum Scale
  - http://www-03.ibm.com/systems/storage/spectrum/scale
  - https://www.ibm.com/support/knowledgecenter
    - Search for “Spectrum Scale”
- Lustre
  - http://lustre.org/
  - http://www.intel.com/content/www/us/en/lustre/intel-solutions-for-lustre-software.html
- Vendor partners – leveraging expertise and service
  - For example: DDN, IBM, Intel, Seagate, etc.

After This Webcast

SNIA ESF | ETHERNET STORAGE

- Please rate this Webcast and provide us with feedback
- This Webcast and a PDF of the slides will be posted to the SNIA Ethernet Storage Forum (ESF) website and available on-demand
- www.snia.org/forums/esf/knowledge/webcasts
- A full Q&A from this webcast, including answers to questions we couldn't get to today, will be posted to the SNIA-ESF blog: sniaesfblog.org
- Follow us on Twitter @SNIAESF
- Need help with all these terms? Download the 2016 SNIA Dictionary http://www.snia.org/education/dictionary

Thank You