High Performance NAS for Hadoop

HPC ADVISORY COUNCIL, STANFORD
FEB 8, 2013

DR. BRENT WELCH, CTO, PANASAS

PANASAS TECHNICAL DIFFERENTIATION

- **Scalable Performance**
  - Balanced object-storage building block [8TB SATA, 120GB SSD, 8GB RAM, 1 core, dual GE]
  - 40 TB to 8 PB single system supporting 100's to 1000's of active clients
- **Novel Data Integrity Protection**
  - File system and RAID are integrated
  - Highly reliable data w/ novel data protection systems
- **Maximum Availability**
  - Built-in distributed system platform manages 100's of blades
- **Simple to Deploy and Maintain**
  - Integrated storage system with appliance model
- **Application Acceleration**
  - Customer proven results
- **Standards Based**
  - pNFS, OSD

ActiveStor 14

ACTIVESTOR BLADE HARDWARE

Dual Power Supplies + Battery

4u

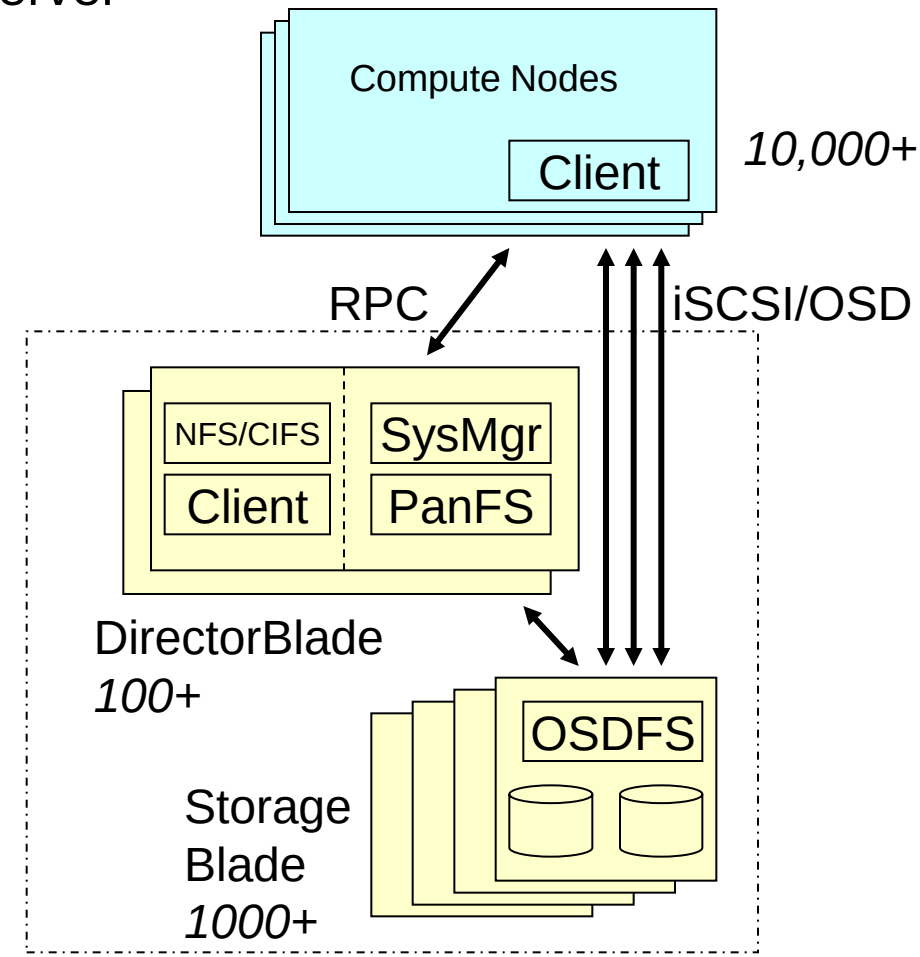
Dual 10GE uplinks

Scalable Metadata

Enterprise SATA + SSD => OSD

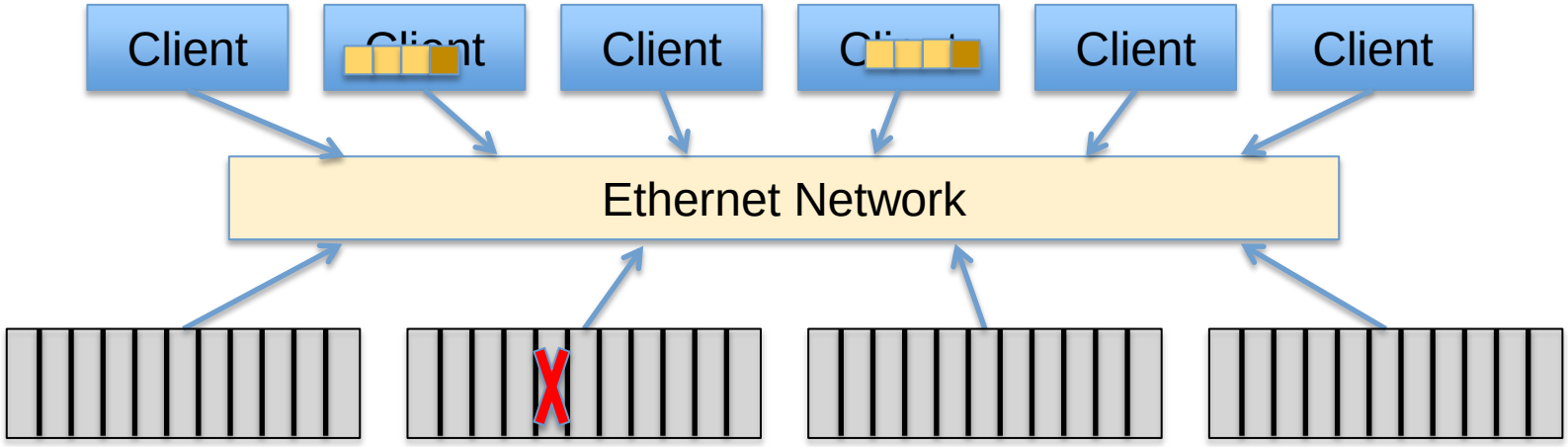
PANASAS SYSTEM VIEW

- **Complete "appliance" solution (HW + SW), blade form factor**
  - DirectorBlade = metadata server
  - StorageBlade = OSD
- **Clustered, fault tolerant metadata services**
- **Linux kernel module for parallel I/O**
  - DirectFlow, or pNFS
- **Object Storage**
  - Snapshots, Quota
- **Global namespace**
  - NFS & CIFS re-export



PANASAS PARALLEL DATA PATH

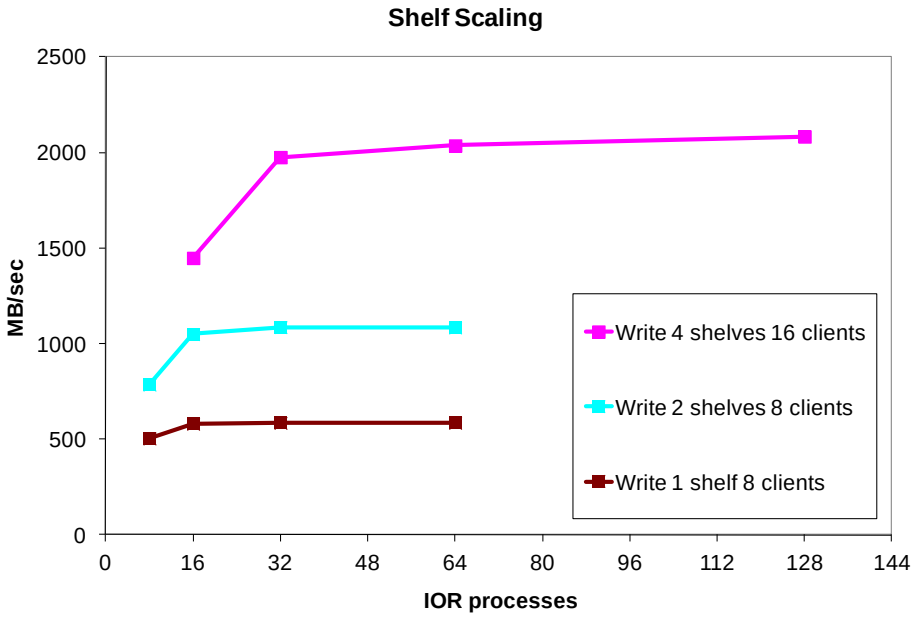
- **Data path by-passes RAID controllers and metadata servers**
  - Application writes data
  - DirectFlow/pNFS client layer generates redundant data for each stripe
  - Everything is written directly to storage
  - All blades work together on RAID rebuild



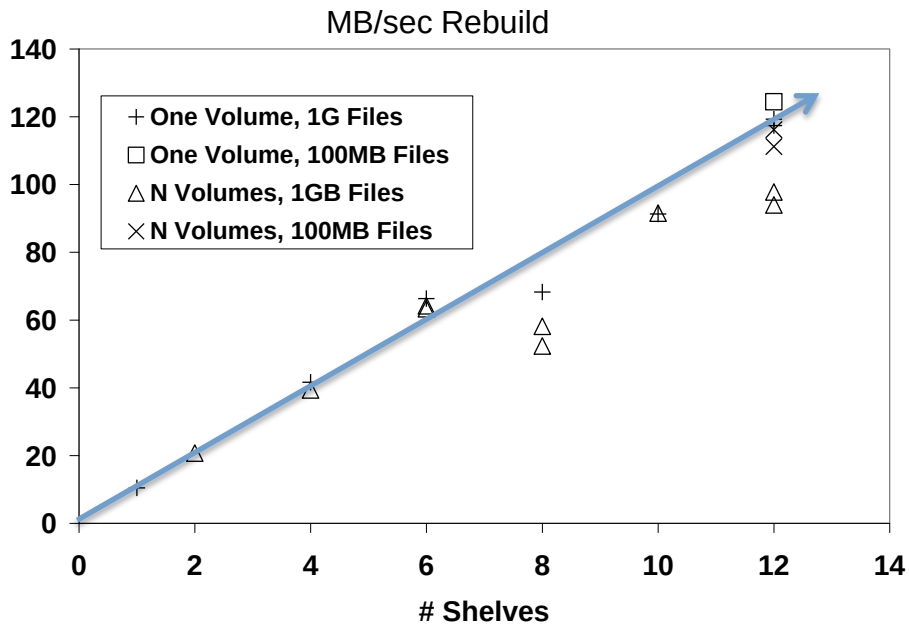
PANASAS PARALLEL ADVANTAGE

- **Scale-out storage system with true parallel architecture**
  - Scale performance and capacity at the same time
  - Rapid recovery from failure – shared RAID responsibility

4 Shelves are 4 times faster than 1

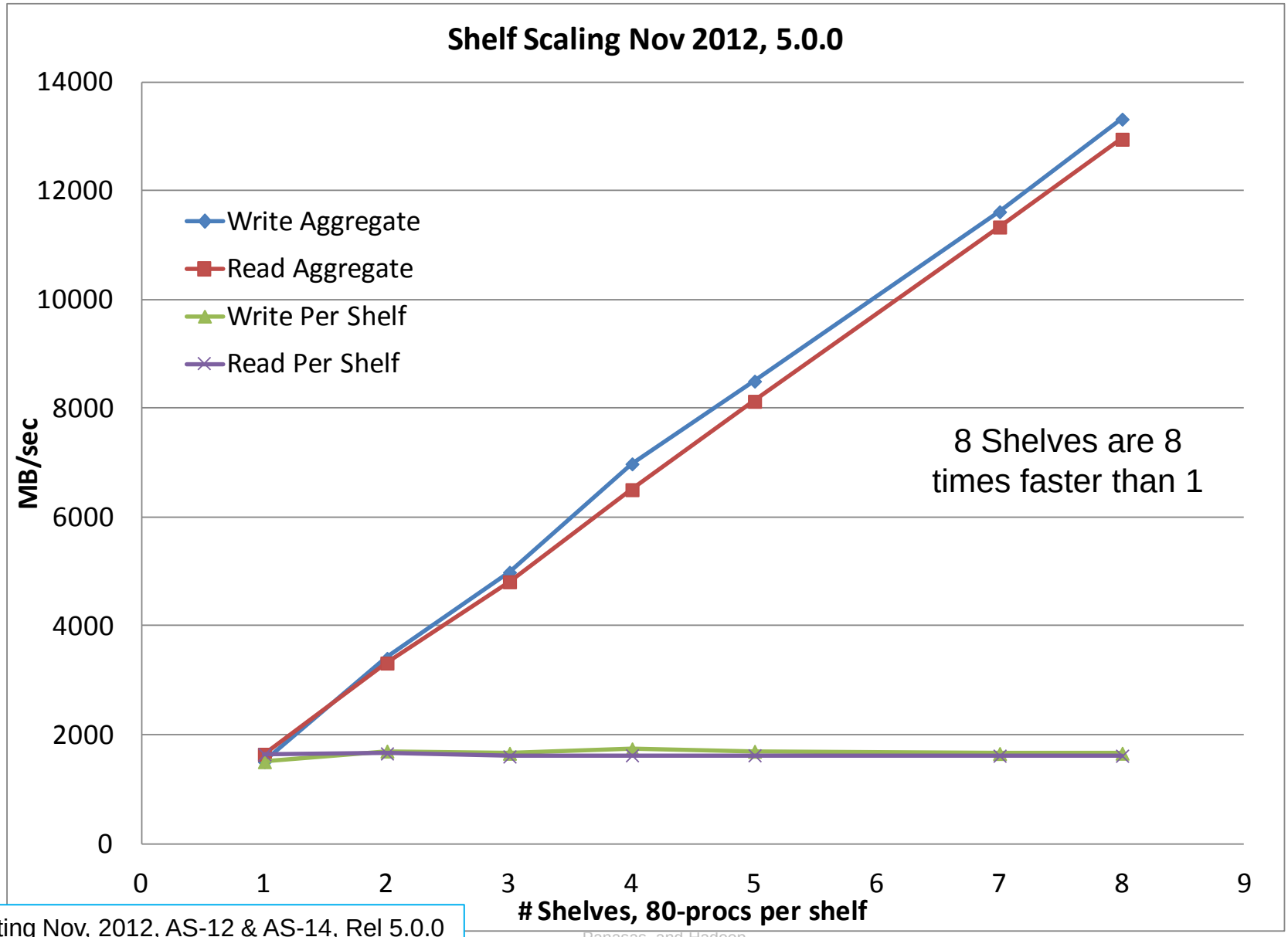


12 Shelves rebuild 12 times faster than 1



3.4 testing December 2008, PAS 8 10GE

SCALABLE BANDWIDTH



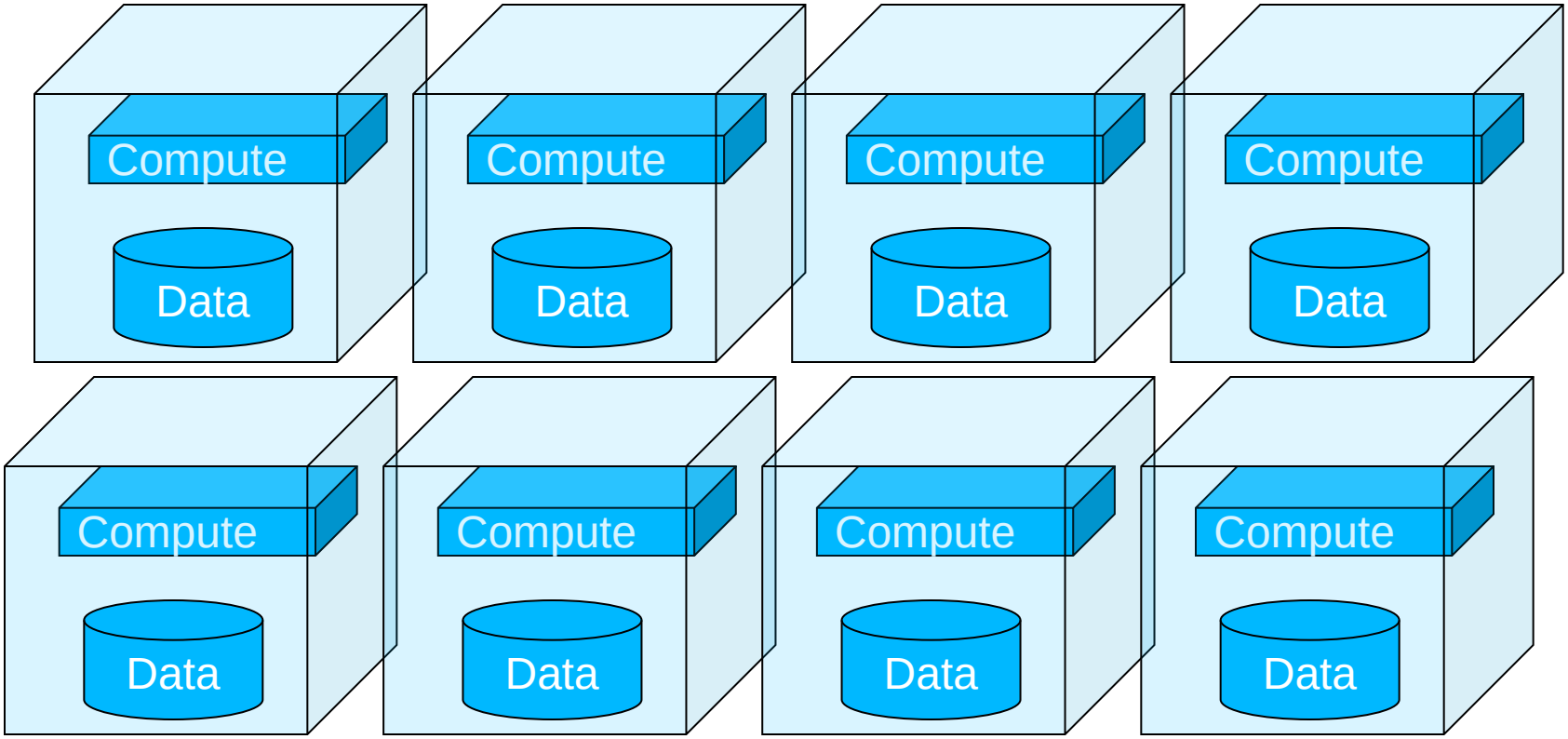
Testing Nov, 2012, AS-12 & AS-14, Rel 5.0.0

HIGH PERFORMANCE NAS FOR HADOOP

HADOOP HW ENVIRONMENT

Low cost hardware, run until failure, offline service

Network infrastructure often oversubscribed



HADOOP SW ENVIRONMENT

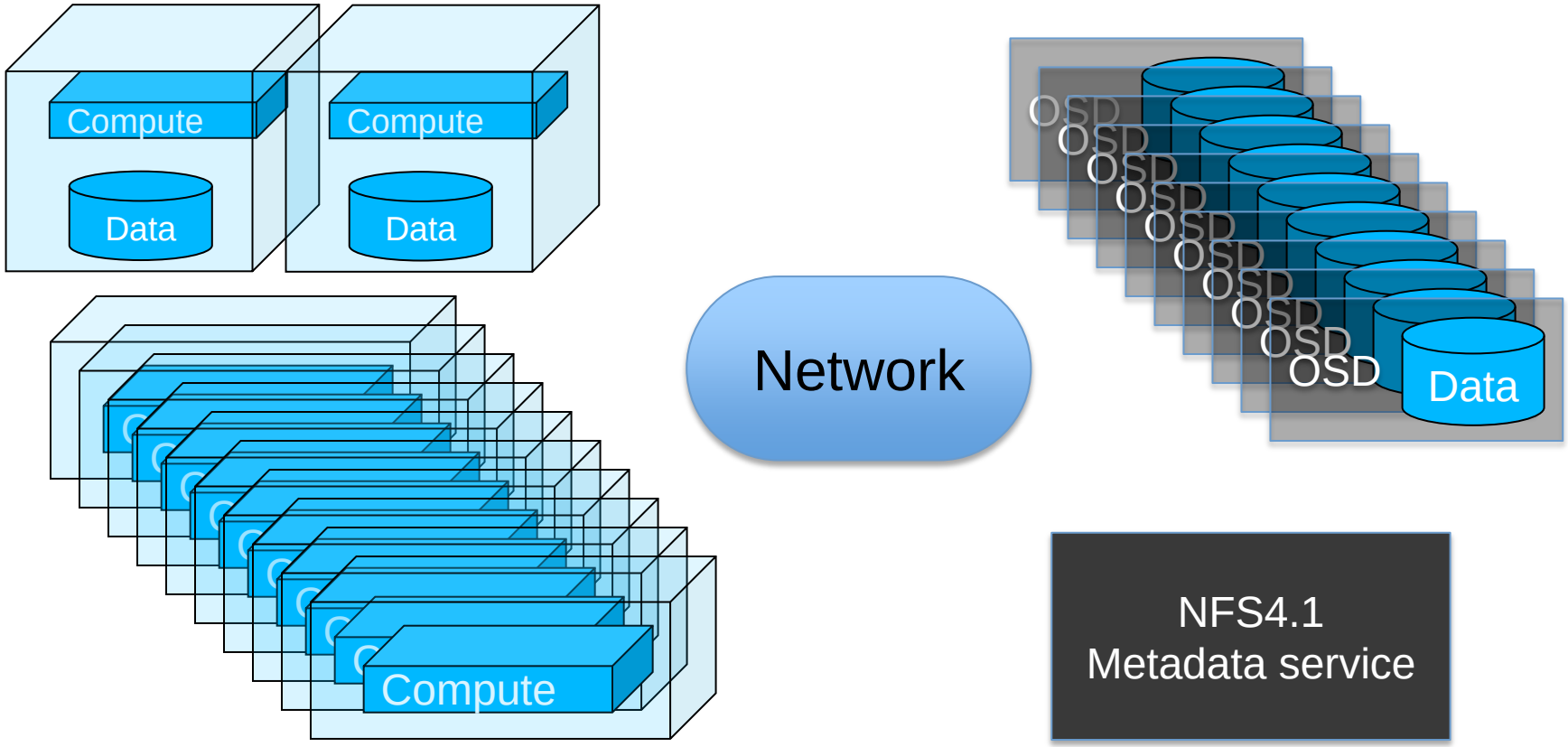
- **Hadoop environment is open Java implementation of a family of data and compute facilities**
  - Hadoop job scheduler for Map/Reduce applications
  - HDFS file system
  - Zookeeper configuration management
  - NoSQL key-value stores layered over HDFS
  - Query languages
  - Many more

LIMITATIONS OF THE ENVIRONMENT

- **Classic HW config mixes compute and data, with weak network**
  - Motivates function shipping instead of data shipping
  - Even so, local access to data is not always possible
  - Triplication is an expensive way to do data protection
  - Not easy to share HDFS data with “normal” applications
  - Classic model grew up in an environment skewed by Google requirements
  - Very different than classic HPC environment

DEDICATED COMPUTE AND STORAGE

Separating compute and storage demands a high quality network
Data is shared among different compute clusters
Hardware replacement cycles for compute and storage differ



HIGH PERFORMANCE NAS FOR HADOOP

panasas

- **A fast network and a good, scalable parallel file system**
  - Keep compute and data management separate
  - Mixed workflows with different kinds of application sharing data
- **Performance intuition**
  - A local disk goes at 50 to 100 MB/sec (large sequential workloads)
  - A good network file system can deliver 500-1000+ MB/sec to one client
  - A local SSD can deliver 250 to 2500 MB/sec
  - Tuning Map/Reduce is more about partitioning a problem so it fits into main memory of the nodes
- **Management intuition**
  - Data scattered among compute nodes makes them “heavy”
  - Hard to upgrade compute w/out affecting storage
  - Serviceability model of many hard drives or expensive PCIe card in every compute node is not very good

COMPARING PANFS AND HDFS

	Hadoop	Panasas	Comment
Data Availability	Triple Replication	Object RAID	Panasas at 15% overhead vs. 200%
File system support	Proprietary	POSIX	Panasas files can be shared with other big data workloads
Hardware	Compute and Storage scale together	Compute and Storage independent	Panasas allows independent scaling of compute and storage
Applications	Single task - Hadoop analytics	Multi-purpose workloads	Panasas designed for many big data workloads
Multi-client write to file	Not allowed - WORM	Supported – Write many	Panasas big data workloads require concurrent file access by multiple clients
Small File	No	Yes	Panasas well suited to mixed big data workloads

ENTERPRISE HADOOP ENVIRONMENT

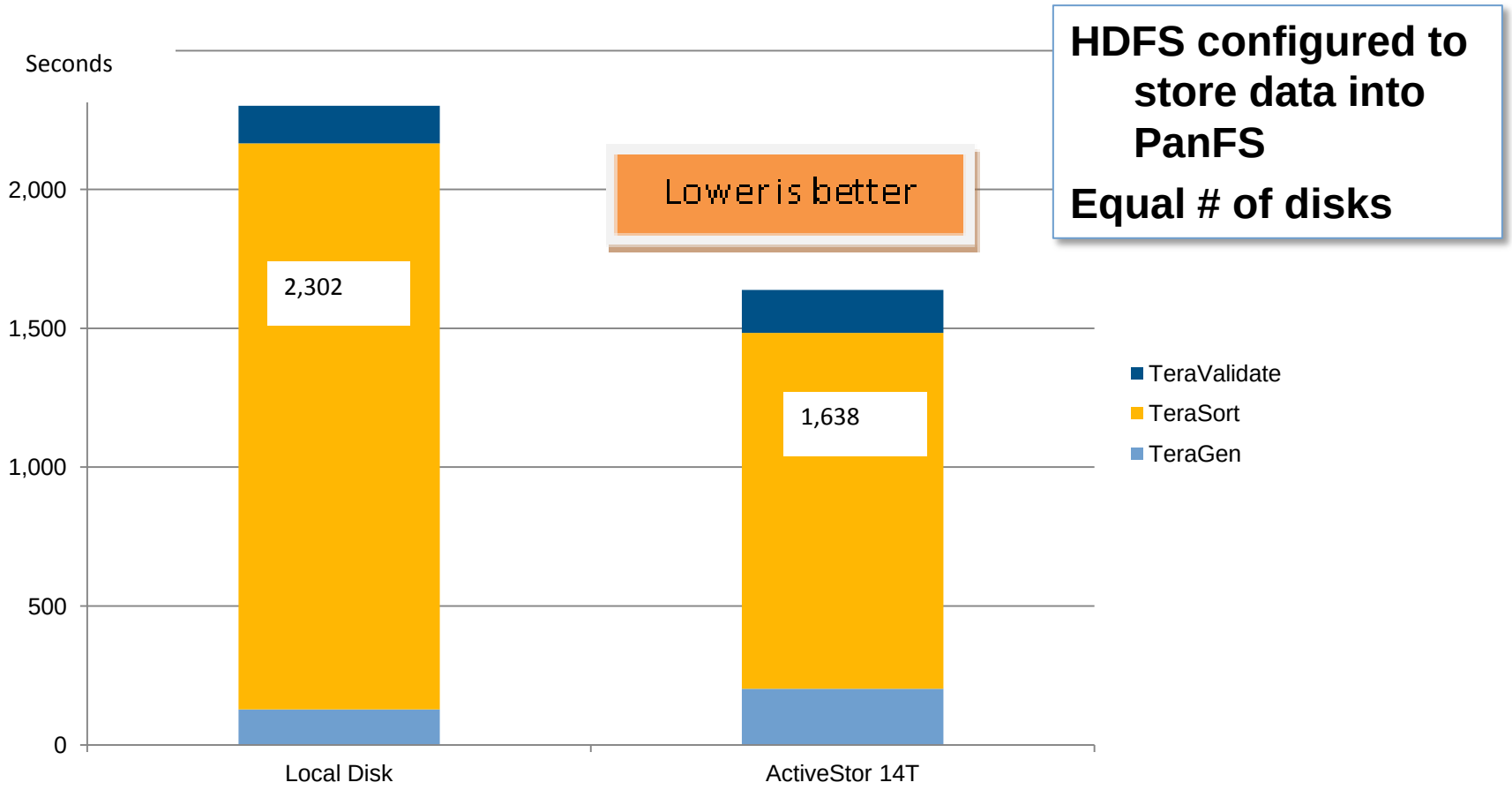
- **Reliable, trusted enterprise storage**
  - Panasas storage offers enterprise class features such as snapshots, user quotas, service and IT administration
- **Panasas allows users to scale computing and storage independently**
  - Features such as load balancing ensure all nodes are equally capable of participating in data transfers
  - Storage can be added to a live system and dynamically integrated into the available pool
- **Data management and data retention**
  - Supports data migration, old data can be moved to archives
  - It can integrate into with existing data management systems
    - Hadoop lacks any built-in data migration other than replication the entire data to another system
- **Scalable storage performance**
  - Tightly balanced system that scales performance linearly as more nodes are added to the system

USING NAS WITH HADOOP

- **Can run on any distribution and any version (Cloudera, Hortonworks, Apache)**
  - No updates required for newer versions of Hadoop
- **No need for proprietary software implementation**
  - Simple configuration setup
- **Can run on HDFS or run directly on PanFS**
  - Layer HDFS over PanFS
  - Configure HDFS pathnames to use /panfs
    - URL: hdfs://panfs/system/workspace
  - Bypass HDFS entirely
  - Configure file:// URLs to use /panfs
    - URL: file://panfs/system/workspace
- **Details captured in a white paper and configuration guide**
  - visit www.panasas.com to get a copy of the paper

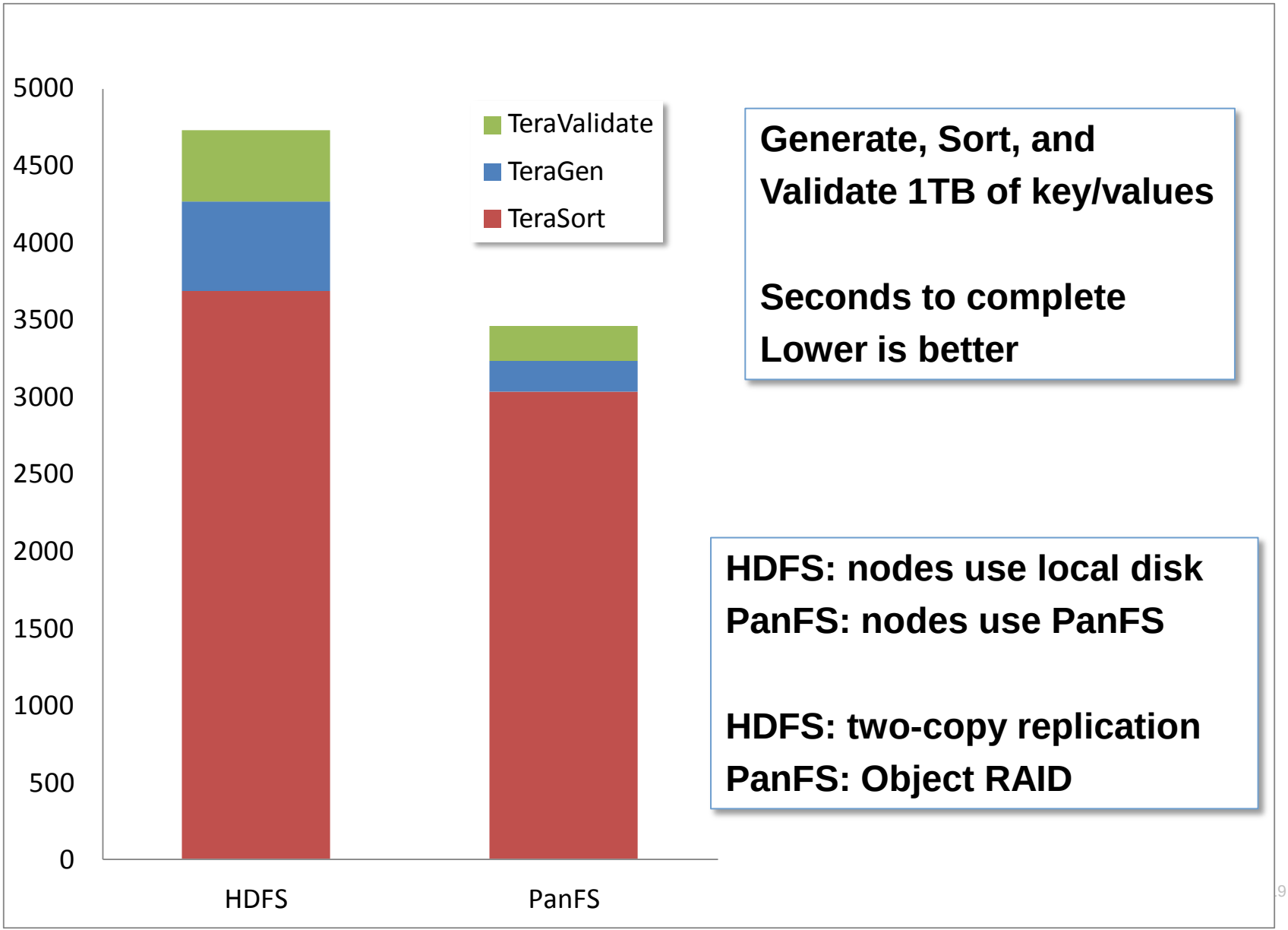
PERFORMANCE, HDFS OVER PANFS

- **41% faster than local disk on HDFS (1 copy)**
- **29% faster than local disk on HDFS (2 copy)**



Download Panasas whitepaper for detailed setup and results
http://www.panasas.com/sites/default/files/uploads/docs/hadoop_wp_lr_1096.pdf

PERFORMANCE, HDFS VS PANFS



Generate, Sort, and Validate 1TB of key/values

Seconds to complete
Lower is better

HDFS: nodes use local disk
PanFS: nodes use PanFS

HDFS: two-copy replication
PanFS: Object RAID

SUMMARY

- **The decisions around the original Hadoop hardware platform were driven by dedicated application specific requirements**
  - Direct attach dedicated server cluster works when the data set is small or when the entire business revolves around Hadoop
- **Mixed use environments, typical of the enterprise require a system that has flexibility, high-reliability, enterprise fault tolerance and supports typical Disaster recovery strategies**
- **Panasas Network attached storage is a viable option for many big data workloads including Hadoop analytics**
- **As networking continues to get faster and cheaper Networked storage will become an increasingly viable solution for Hadoop**
  - Large data sets are unwieldy on local disk
  - Management headache of the 1990's in the enterprise again?
- **Hadoop is first an application, the hardware choice depends on the business specific context. Panasas NAS is a viable, high performance solution for mixed-use workloads**

THANK YOU