DataDirect™
NETWORKS

Lustre Beyond HPC

Toward Novel Use Cases for Lustre?

Presented at LUG 2014

2014/03/31

Robert Triendl, DataDirect Networks, Inc.

this is a vendor presentation!

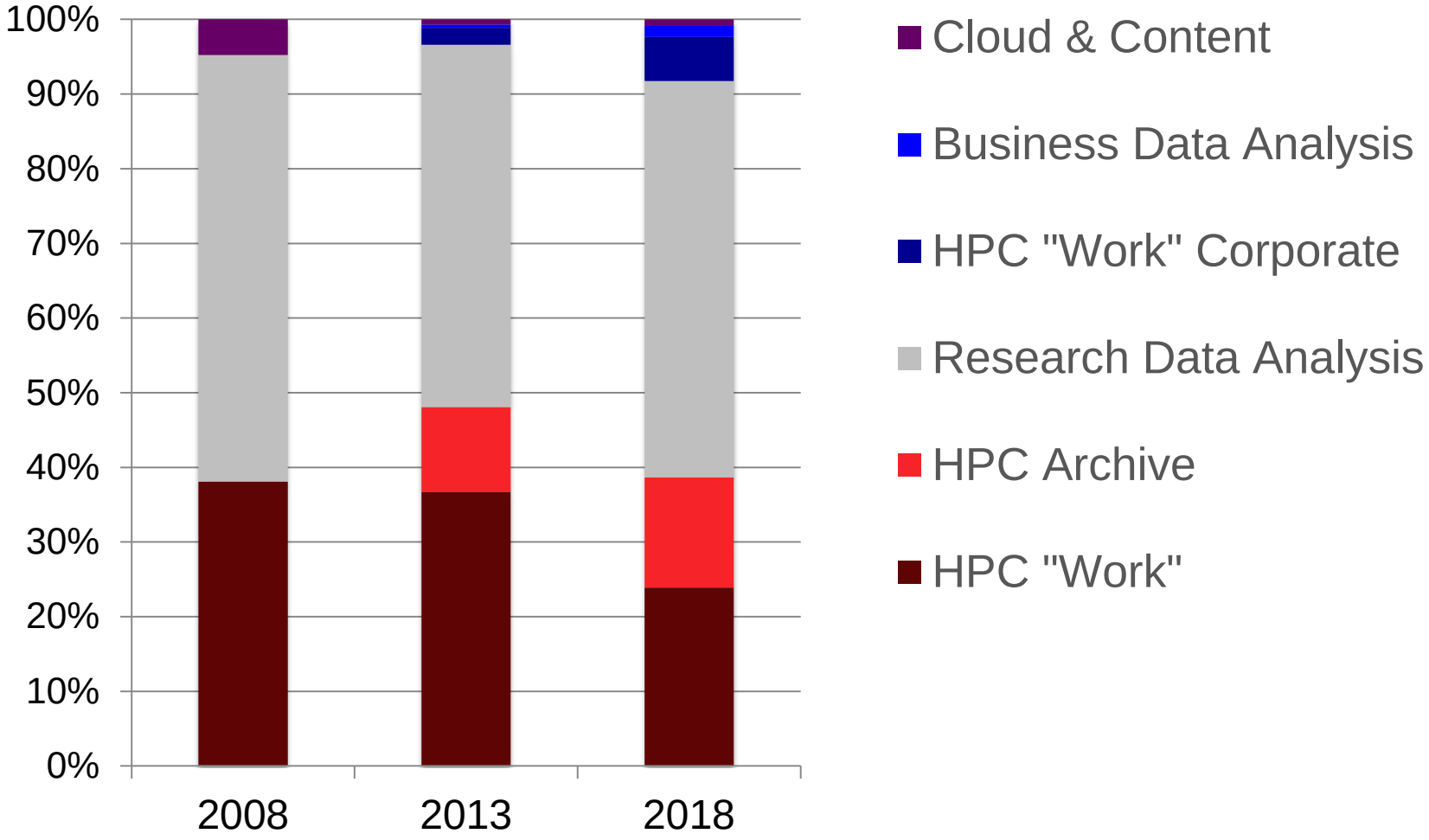
Magritte

Lustre Today

- The undisputed file-system-of-choice for HPC
  - large-scale parallel I/O for very large HPC clusters (several thousand nodes or larger)
  - applications that generate very large datasets but only a relatively limited amount of metadata traffic
- Alternative solutions
  - are typically more expensive (since proprietary)
  - and not nearly as scalable as Lustre

Lustre Market Overview

Data for the Japan Market



What Happened?
Example: Storage Market Japan

- Lustre market expansion
  - From a relatively small player to the dominant HPC file system
  - From tens of OSSs to hundreds of OSSs (and, when including the “K” system, literally **thousands** of OSSs)
  - But, also Lustre has become **synonymous** with large-scale HPC
  - Experiments to use Lustre in other markets and for other applications have **decreased**, rather than **increased**

Overall Storage Market Evolution

- "Software-defined storage" is replacing traditional storage architectures in large cloud deployments
- Many Lustre "alternatives" are now available
  - CEPH, OpenStack/Swift, Swift Stack, Gluster, etc.
  - Various Hadoop distributions
  - Commercial Object Storage (DDN WOS, Scality, Cleversafe, etc.)
- Lustre remains "exotic" and is rarely considered even an option

I/O Caching

Connectors

Client Performance

NFS/CIFS Access

Large I/O

Management

Cluster Integration

RAS Features

Data Management

Project Quota

Backup/Replicatio

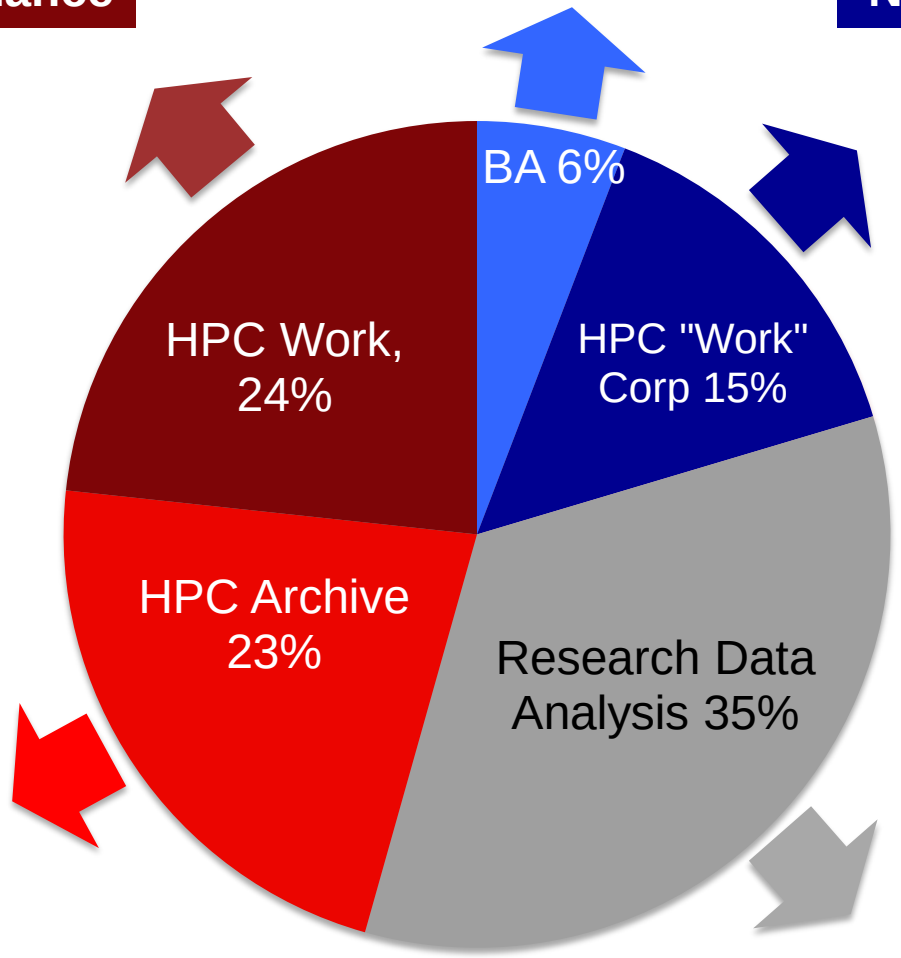
Small File I/O

Object/Cloud Links

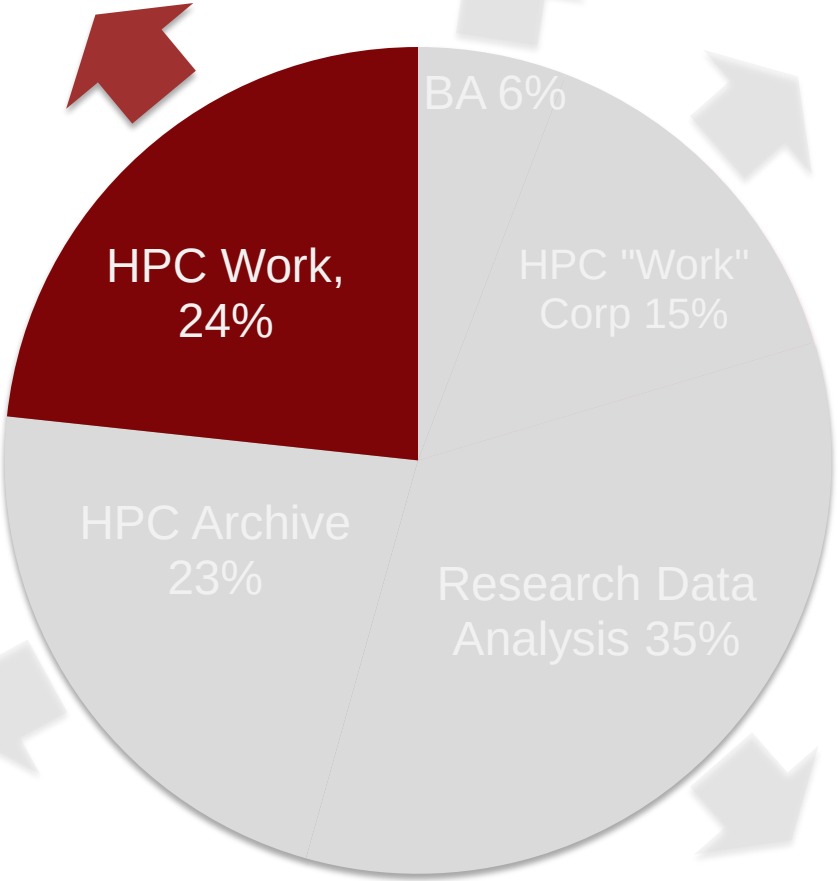
SSD Acceleration

HSM

Fine-Grained Monitoring



Large I/O

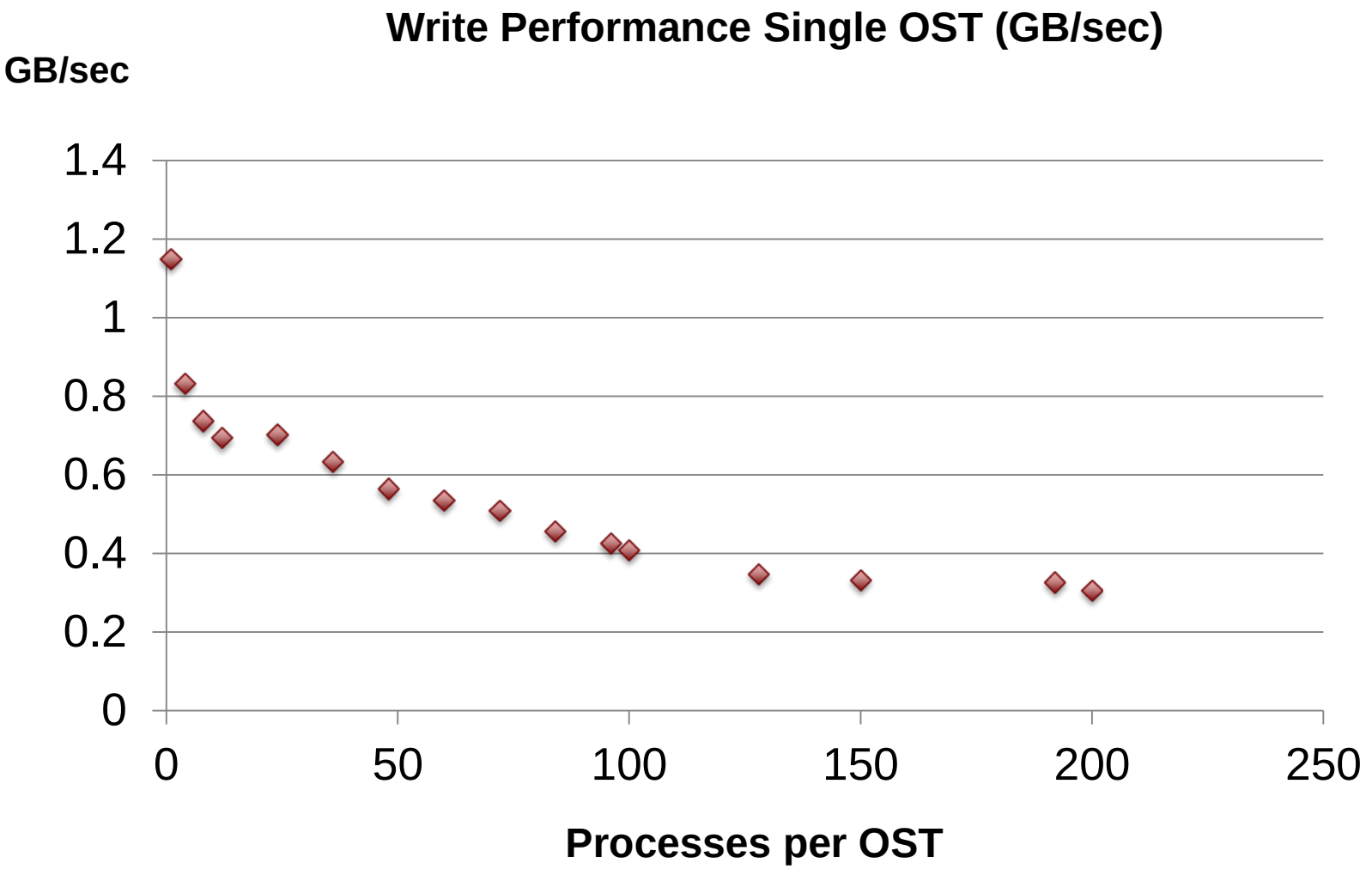

HPC Work, 24%
BA 6%
HPC "Work" Corp 15%
Research Data Analysis 35%
HPC Archive 23%

Nagoya University Acceptance BM

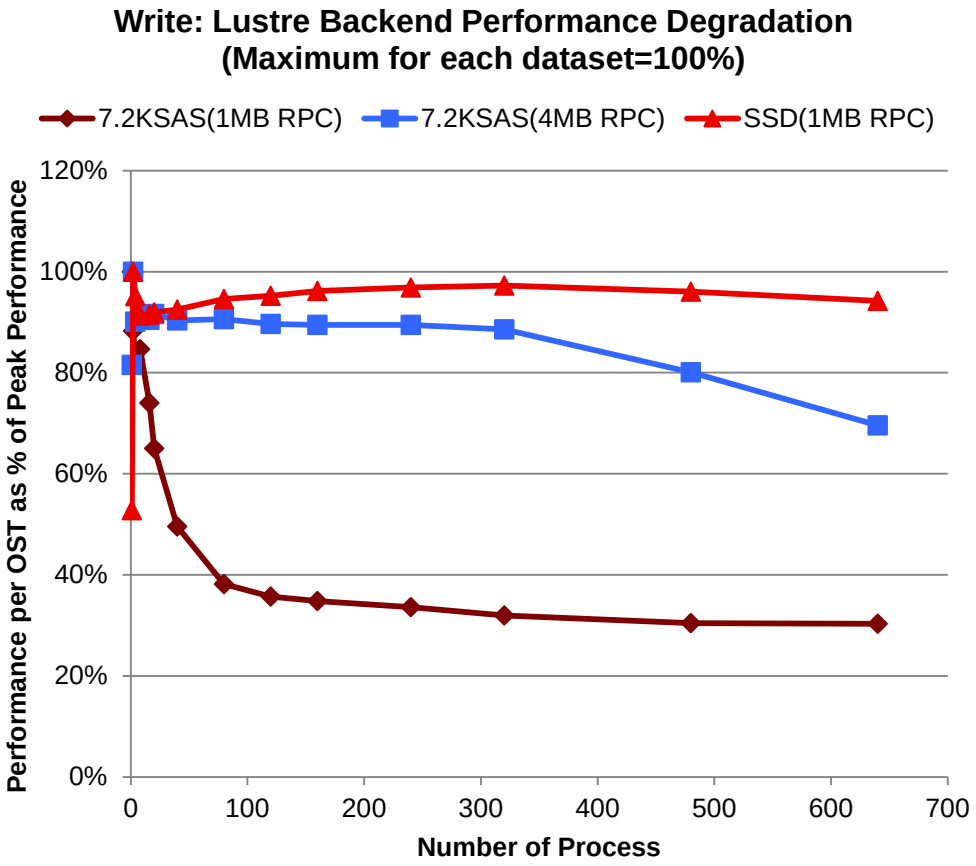
- Large Cluster
  - FFP from each core in the cluster
  - Most efficient configuration with 3 TB/4 TB drives
  - 350-400 threads per Lustre OST

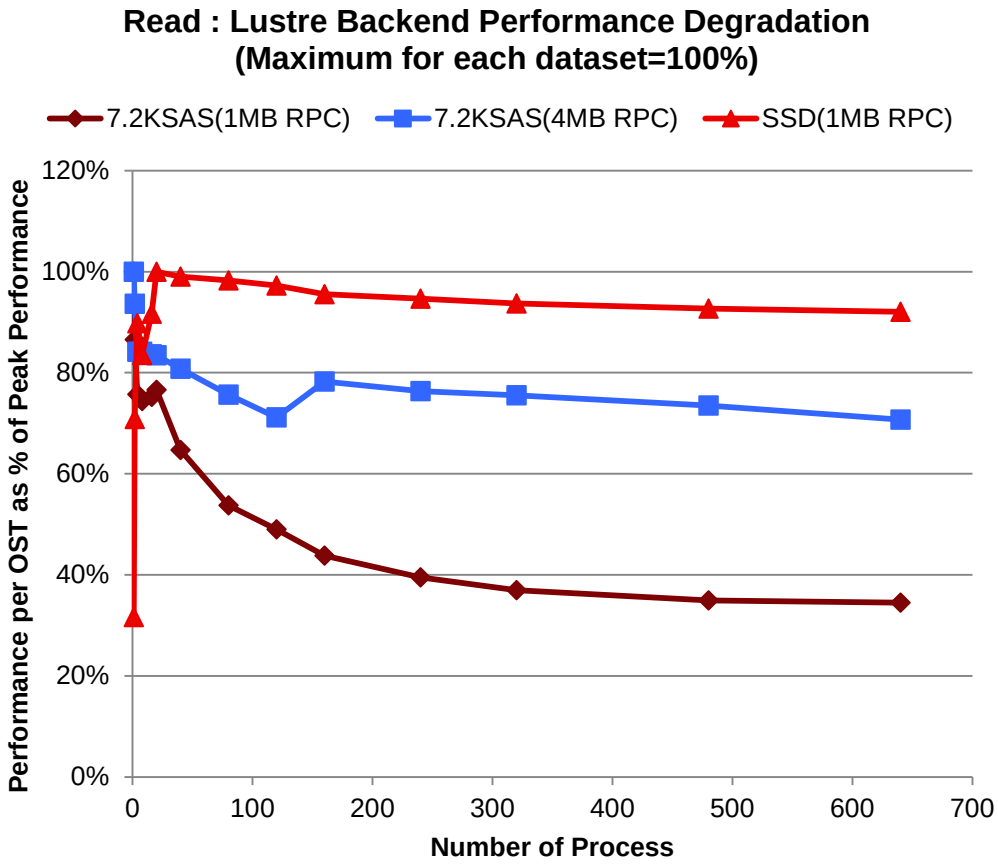
Nagoya University Initial Data

FPP with Large Number of Threads

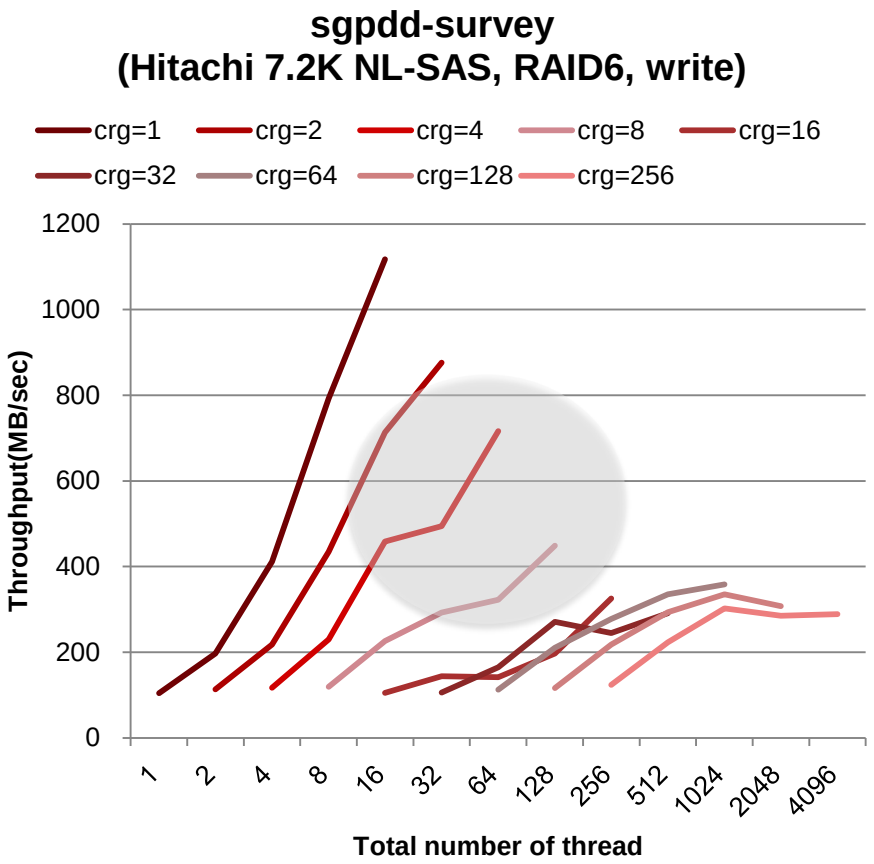


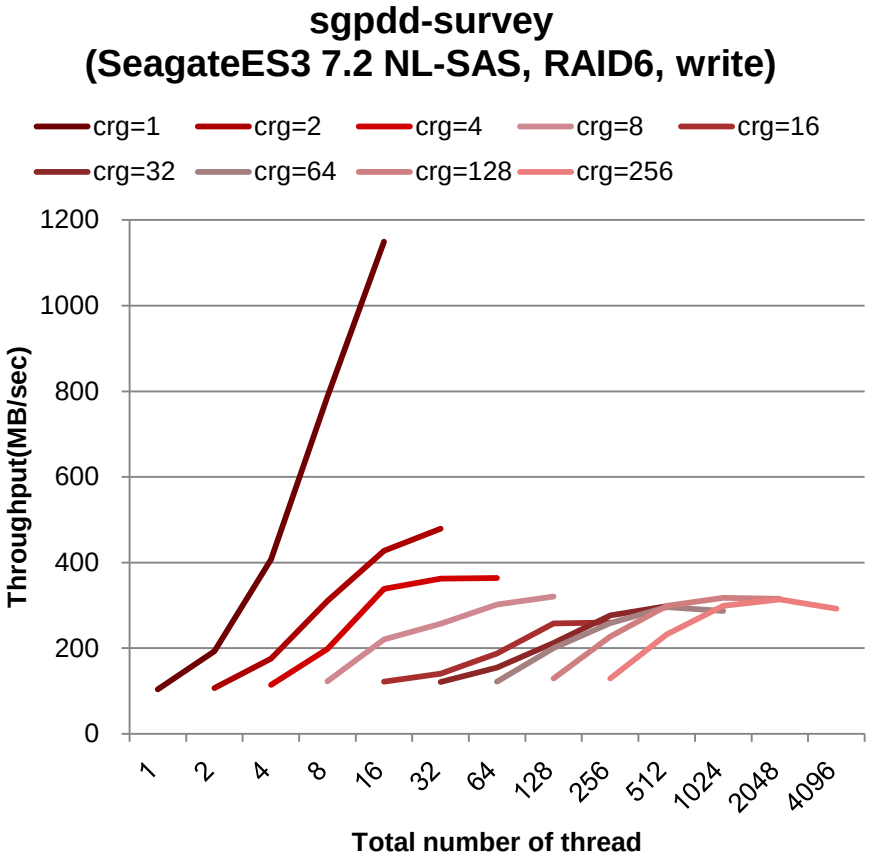
Large I/O Patches



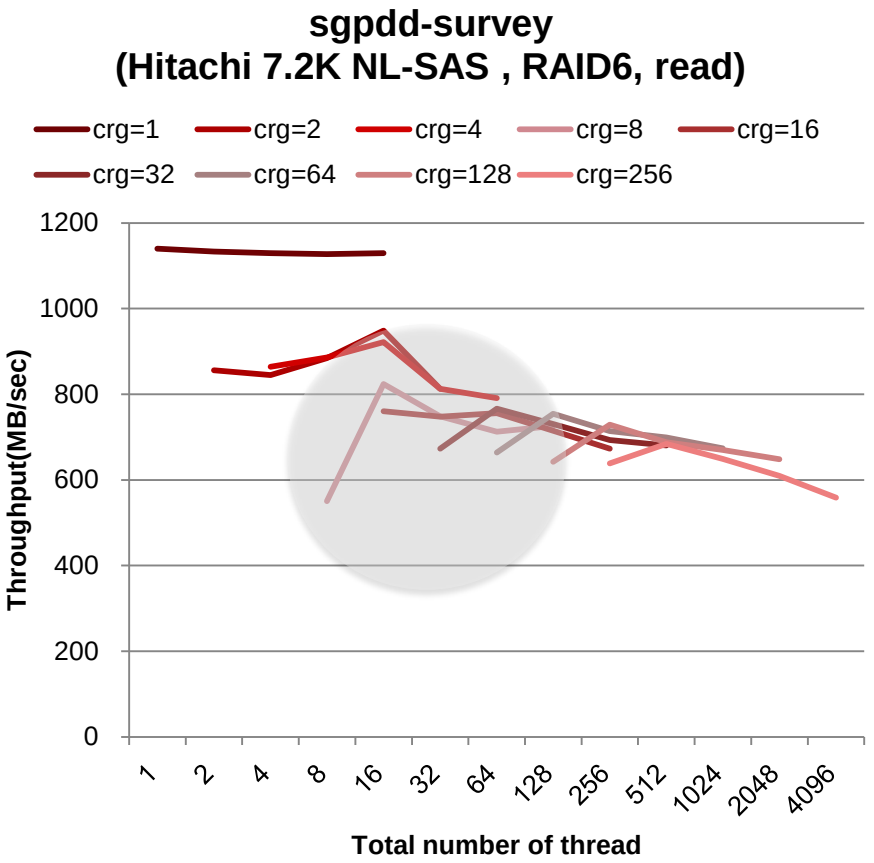


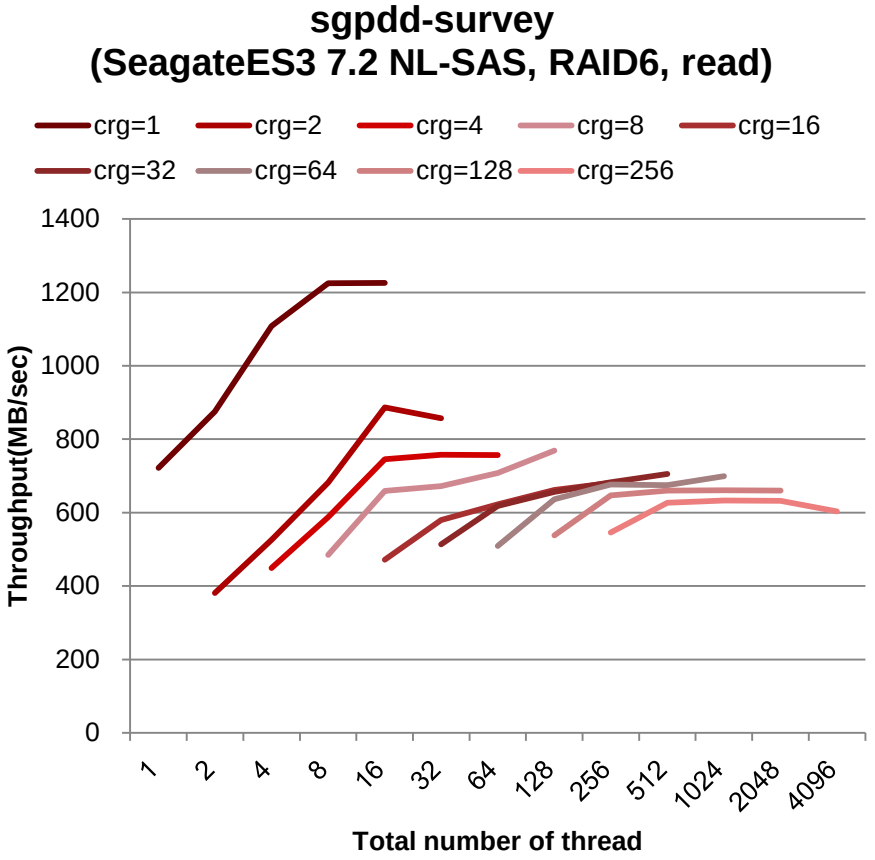
Raw Device Performance: Write

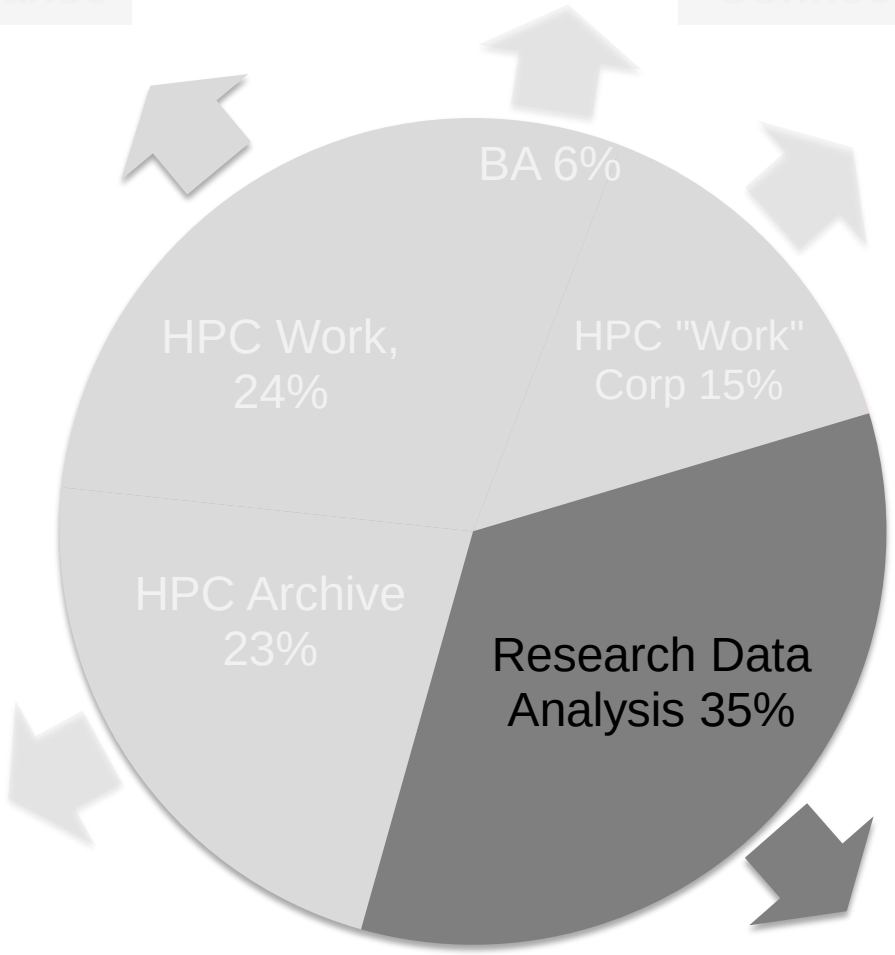




Raw Device Performance: Read





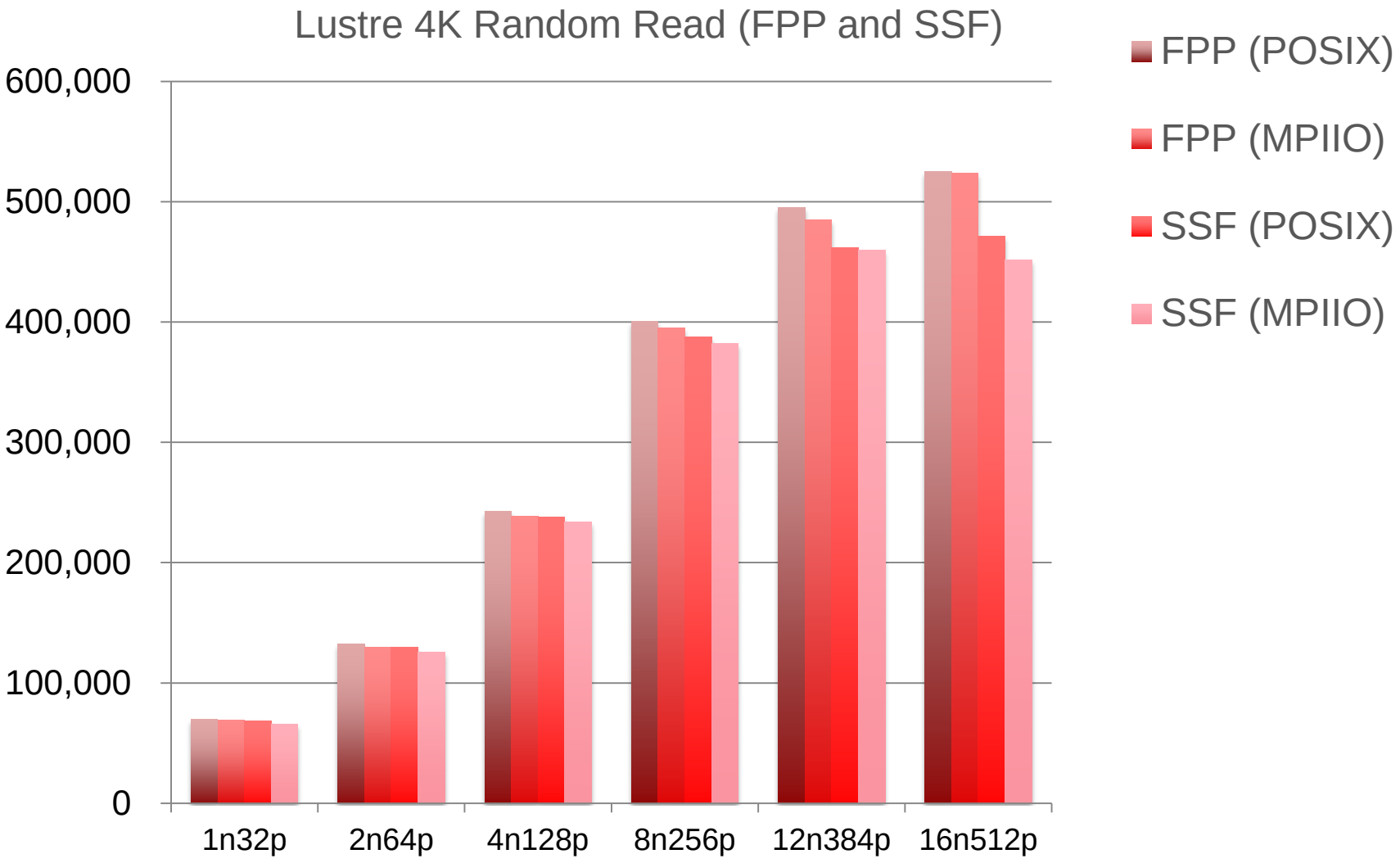


Output Data from a Simulation

- Requirements
  - Random reads against multiple large files
  - 2 million 4k random read IOPS
- Solution
  - Lustre file system with 16 OSS servers
  - Two SFA12K (or one SFA12KXi)
  - 40 SSDs as Object Storage Devices

Lustre 4K Random Read IOPS

Configuration: 4 OSSs, 10 SSDs, 16 clients



Genomics Workflos

- Mixed workflows
  - Ingest
  - Sequencing pipelines: large file I/O
  - Analytics workflows: mixed I/O
- Various I/O issues
  - Random reads for reference data
  - Small-file random reads

Random Reads with SSDs

Data Analysis “Workflows”

- Scientific data analysis
  - Genomics workflows
  - Seismic Data Analysis
  - Various types of accelerators
  - Large scientific instruments in astronomy
  - Remote sensing and environmental monitoring
  - Microscopy

Data Analysis “Workflows”

- Additional topics
  - Data ingest
  - Data management and data retention
  - Data distribution and data sharing

Hyperscale Storage
HPC, Cloud, Data Analysis

High Performance Computing	Cloud Computing
Mostly (very) large files (GBs)	Small and medium size files (MBs)
Mostly write I/O performance	Mostly read I/O performance
Mostly streaming performance	Mostly transactional performance
10s of Petabytes of Data	10s of Billions of files
Scratch data	WORM & WORN
100,000s cores	10s of millions of cores
Mostly Infiniband	Almost exclusively Ethernet
Single location	Highly distributed data
Very limited replication factor	High replication factor
High efficiency	Low efficiency

Data Analysis Workflows

I/O Caching

Client Performance

Large I/O

Cluster Integration

Connectors

NFS/CIFS Access

Management

Project Quota

Small File I/O

SSD Acceleration

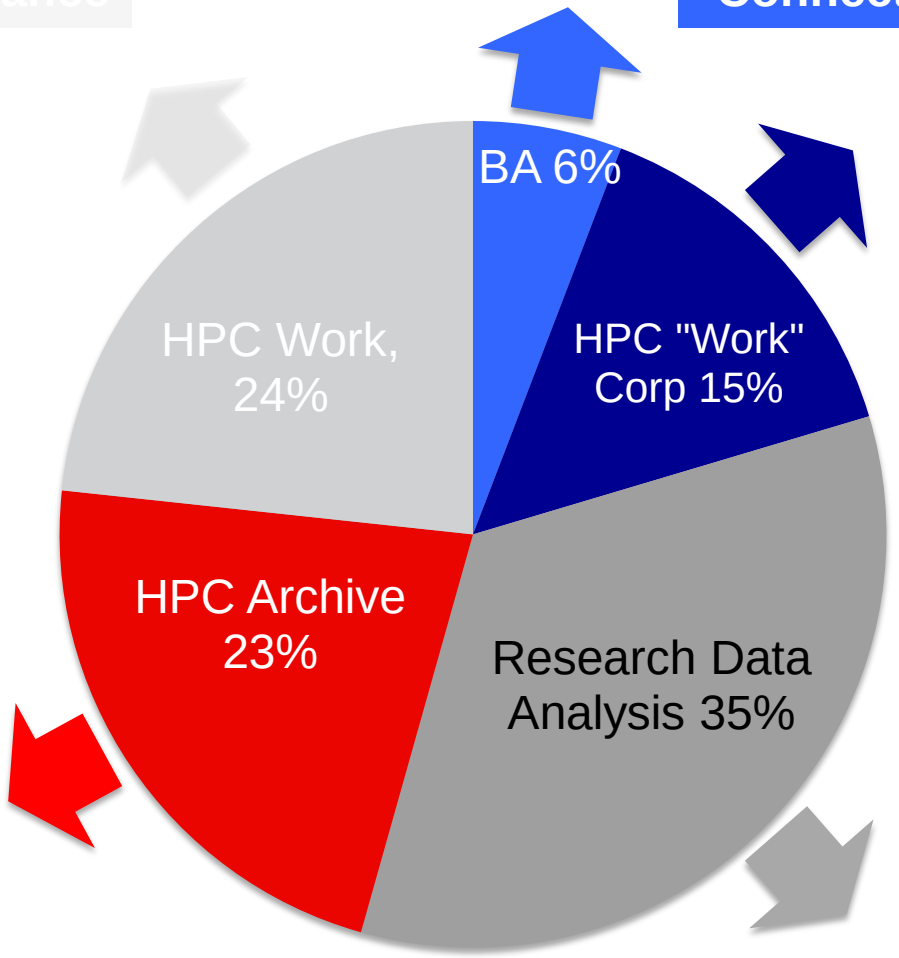
Fine-Grained Monitoring

HSM

Object/Cloud Links

Backup/Replicatio

Data Management



A (DDN) Vision for Lustre

- Maximum sequential and transactional performance per storage sub-system CPU
- Caching at various layers within the data path
- Increased single node streaming and small file performance
- Millions of metadata operations in a single FS
- Millions of random (read) IOPS within a single FS

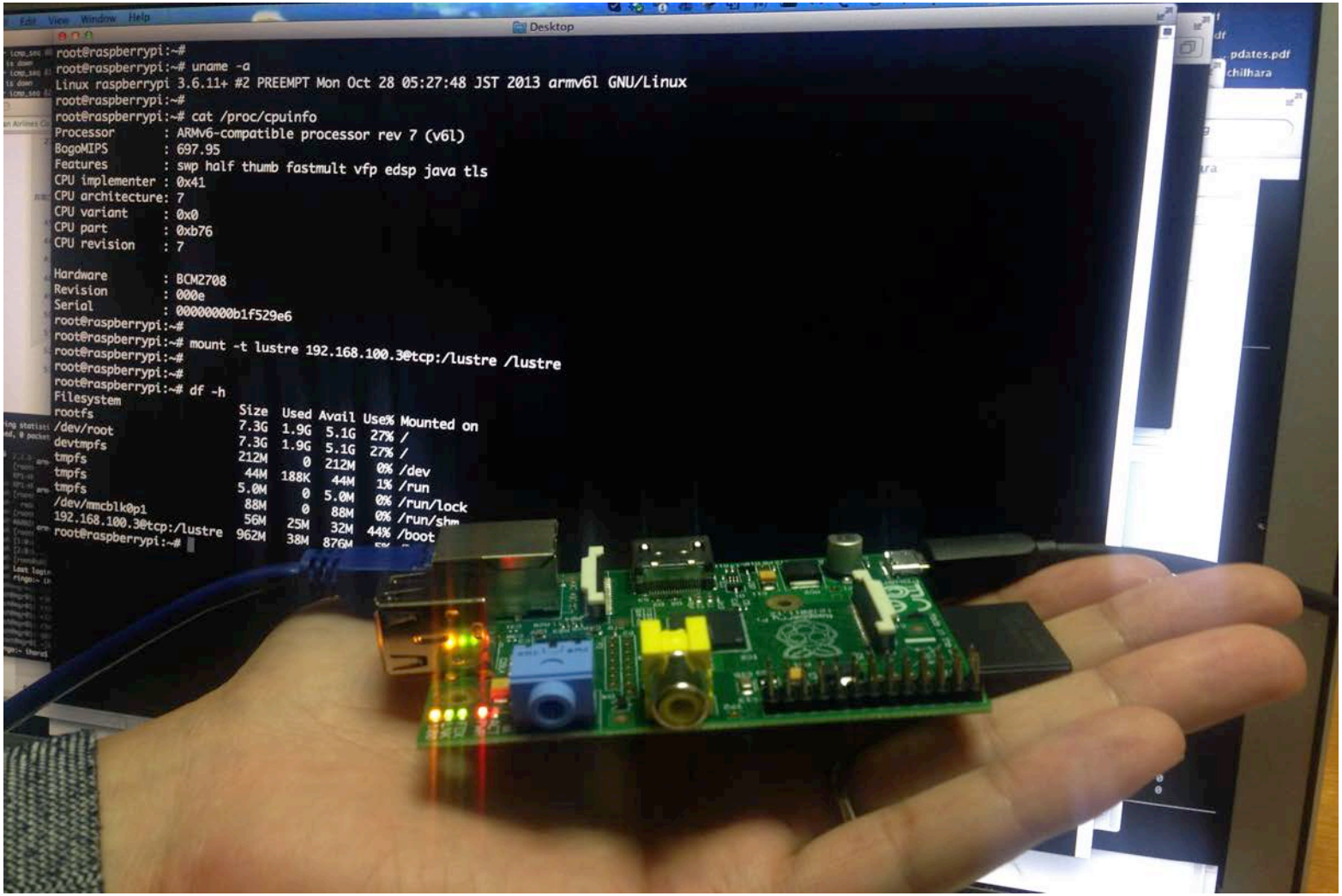
A (DDN) Vision for Lustre cont.

- Data management features, including (cloud) tiering, fast and efficient data back-up, and data lifecycle management
- Novel usability features such as cluster integration, QoS, directory-level quota, etc.
- Extremely high backend reliability for small and mid-sized systems

Futures for Lustre?

- Work Closely with Users
  - User problems are the best source for future direction
  - Translate user problems into roadmap priorities
- Work Closely with the Lustre Community
  - Work very closely with OpenSFS and Intel HPDD on Lustre roadmap priorities and various other topics

Futures for Lustre?



Ceci n'est pas une pipe.

Magritte