SSD Provisioning for Exascale Storage Systems

Devesh Tiwari

Oak Ridge National Laboratory

Sarp Oral

Feiyi Wang

Saurabh Gupta

Josh Judd

OAK RIDGE NATIONAL LABORATORY

MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

WARP MECHANICS

SSDs: The Good, the bad and the Ugly

High performance for random workloads

Low power consumption

Shock resistant

Write-endurance

High cost per Byte

Can we build an exascale storage system out of SSDs?

Write Endurance

Is SSD write-endurance a problem for HPC?

Projected SSD storage space: ~5-10 PB

If building blocks are typical 256 GB SSDs

Number of SSDs in the system = 20,000

5 year warranty for max. 40GB write per day*

Allowed write amount: 600TB write per day

*Samsung 840 Pro Data Sheet
http://www.samsung.com/us/pdf/memory-storage/840PRO_25_SATA_III_Spec.pdf

Is SSD write-endurance a problem for HPC?

Assuming write amplification factor = 1.3

Allowed user written data = ~460TB per day

Write-endurance becomes a roadblock if an application dumps even 10% of system memory as checkpointing data every hour

System-level checkpointing easy on the programmer, hard on SSD-based storage system

Is SSD write-endurance a problem for HPC?

Assuming write amplification factor = 1.3

Allowed user written data = ~460TB per day

OLCF: S3D 360 TB per day; GTC 240 TB per day

NERSC: <100 TB per day (Darshan instrumented)

ALCF: most jobs moving <100 TB

Carns et al., Understanding and improving computational science storage access through continuous characterization, ACM Transactions on Storage, 2011

Computational Requirements of Leadership Computing
http://www.olcf.ornl.gov/wp-content/uploads/2010/03/ORNL_TM-2007_44.pdf

Is SSD write-endurance a problem for HPC?

Even 1TB SSD as a building block will allow up to 150TB write per day

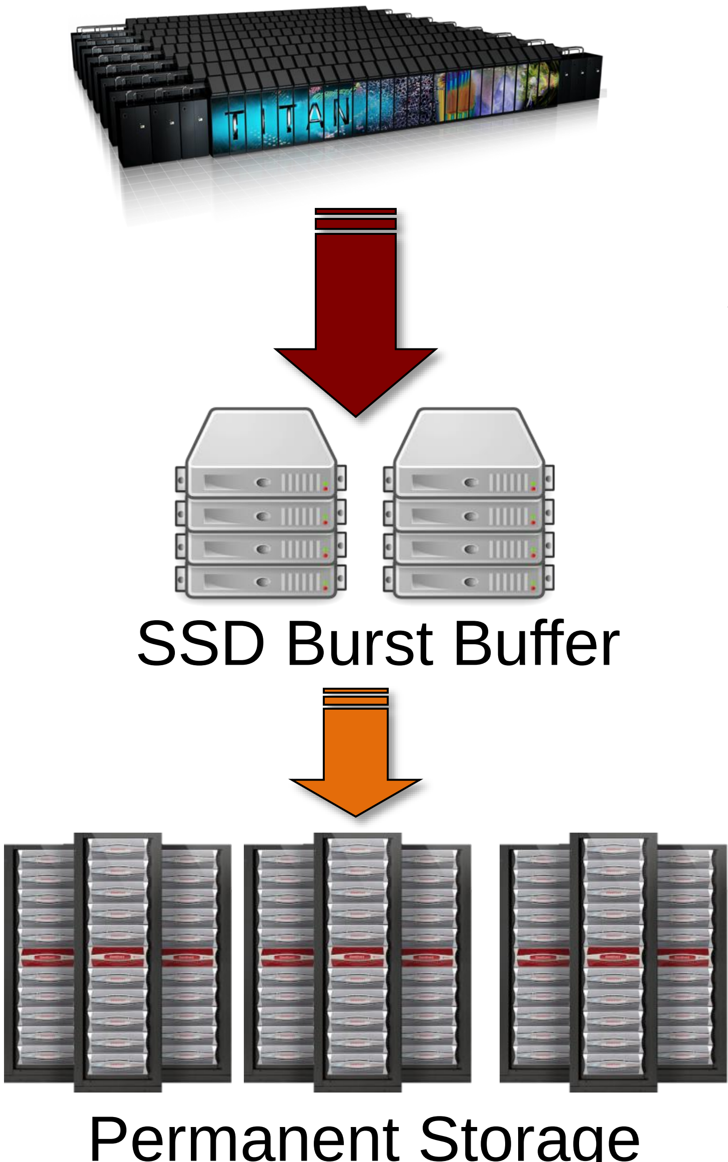
At higher price, up to 3 full writes per day

Write-endurance improving at a fast speed

Reducing the checkpointing size is the key to alleviating SSD write-endurance issue (application-level checkpointing strategies)

Where in the System?
And, how much?

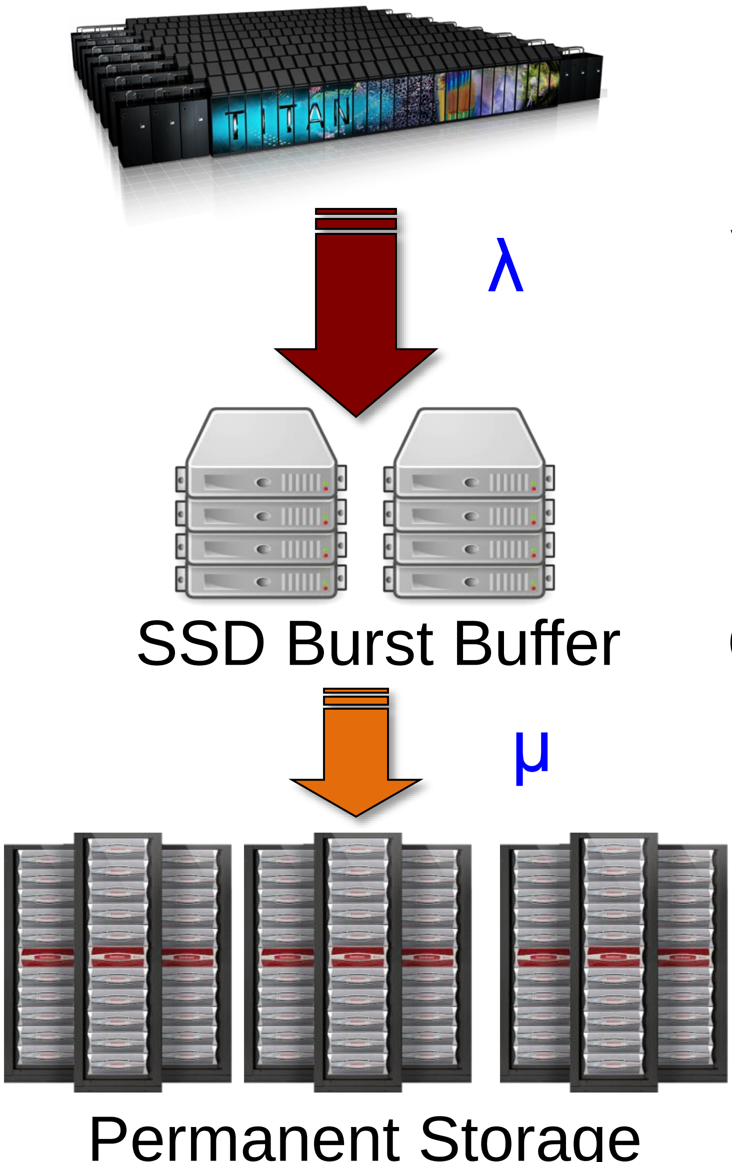
SSDs as a burst buffer in HPC



λ = Data production rate

μ = Bandwidth to Permanent Storage System

SSDs as a burst buffer in HPC

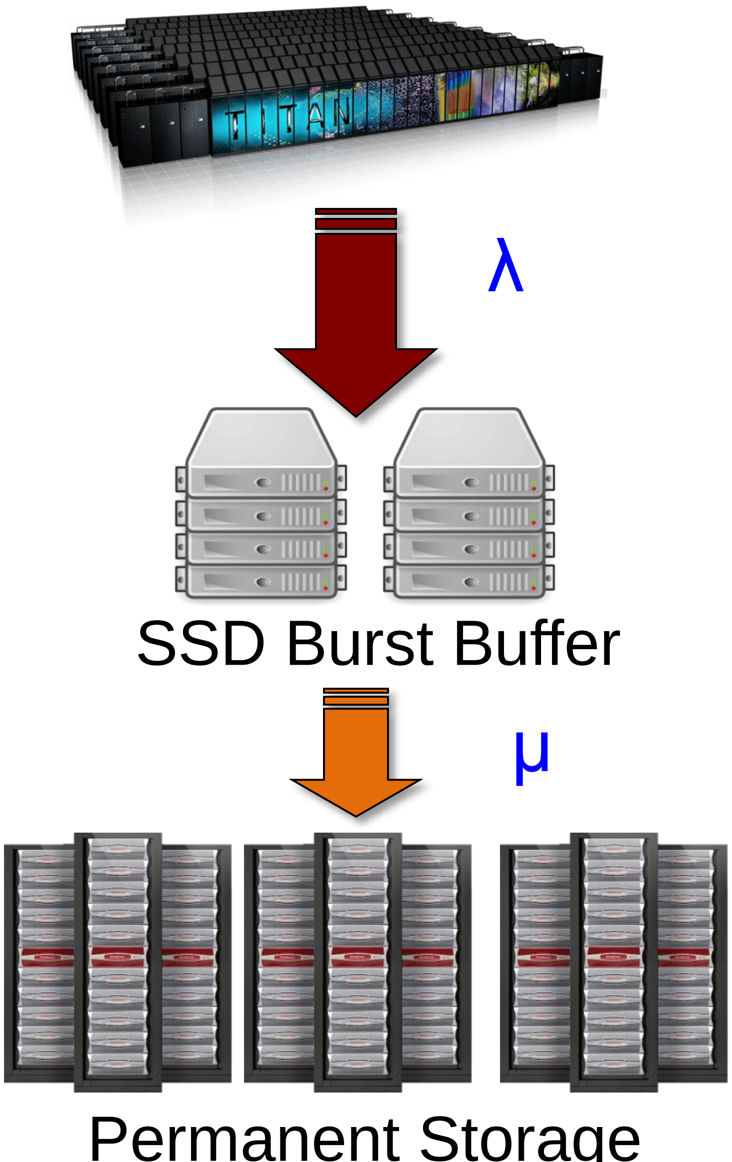


If $\lambda > \mu$:

SSD burst buffer capacity is unbounded

Intuition:
Higher incoming flux than outgoing flux will burst the pipe
Never enough time to drain out

SSDs as a burst buffer in HPC



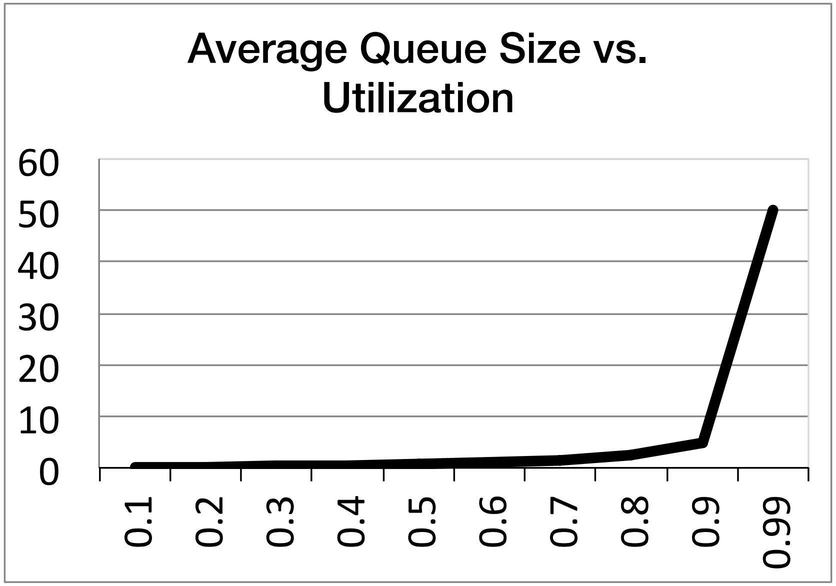
Implications:

Not all the data from burst buffer can be copied just in time

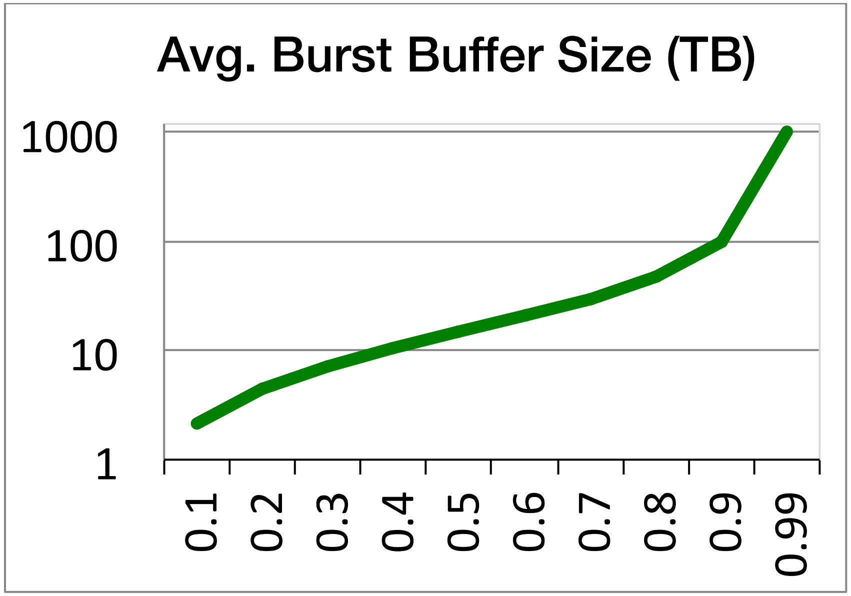
PFS Bandwidth still critical

Utilization = production rate/achievable PFS bandwidth = λ/μ

SSDs as a burst buffer in HPC



Utilization= λ/μ

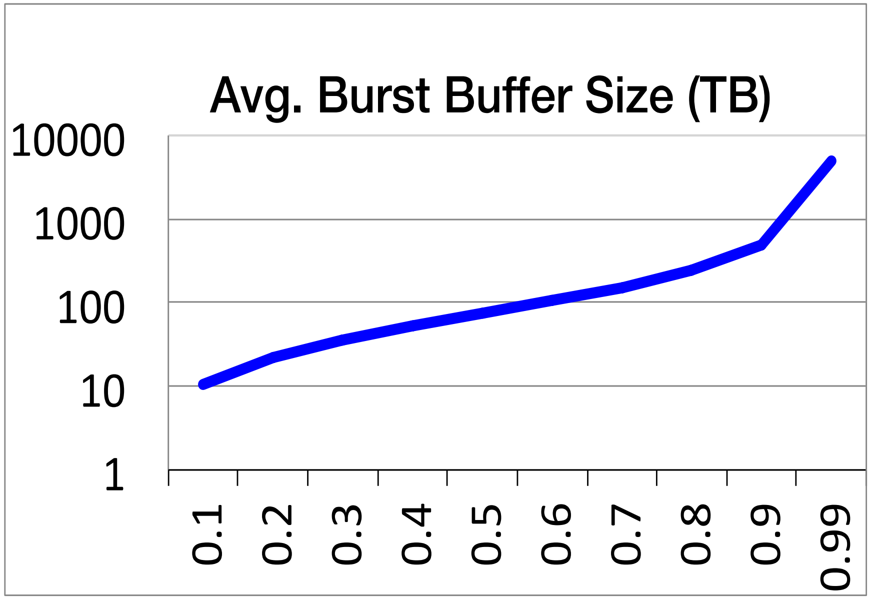


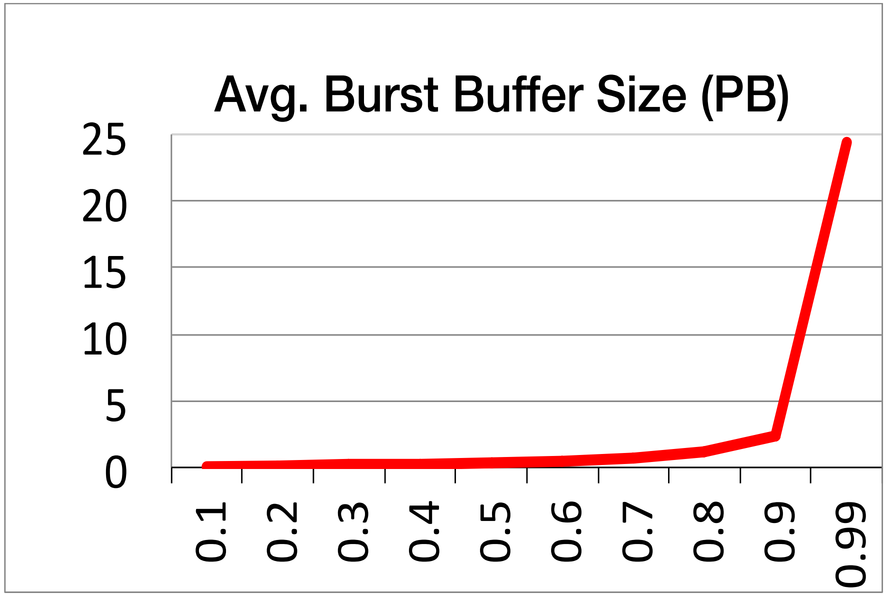
Utilization= λ/μ

Output data size at once = 20TB

Queuing models suggest that avg. queue size (burst buffer size) increases exponentially with the increase in utilization

SSDs as a burst buffer in HPC



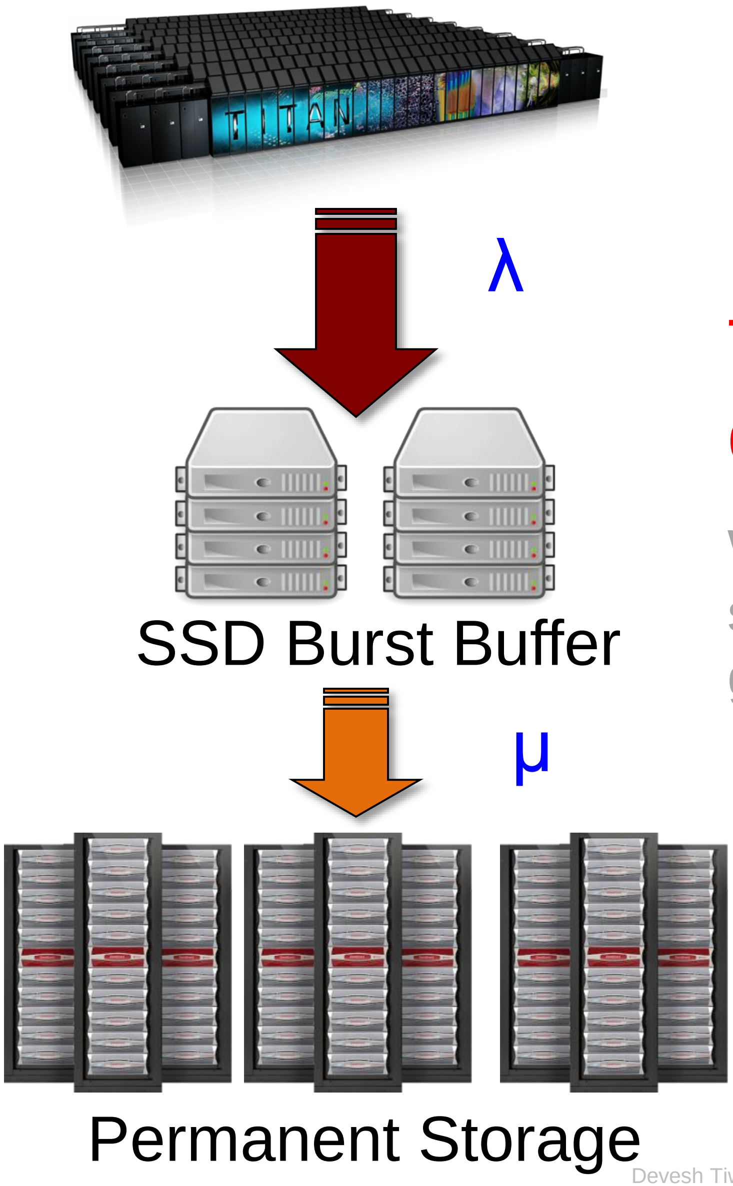


Utilization= λ/μ
Output data size at once = 100TB

Utilization= λ/μ
Output data size at once = 200TB

1-5 PB of SSD storage may suffice depending on the amount of data being produced at each step

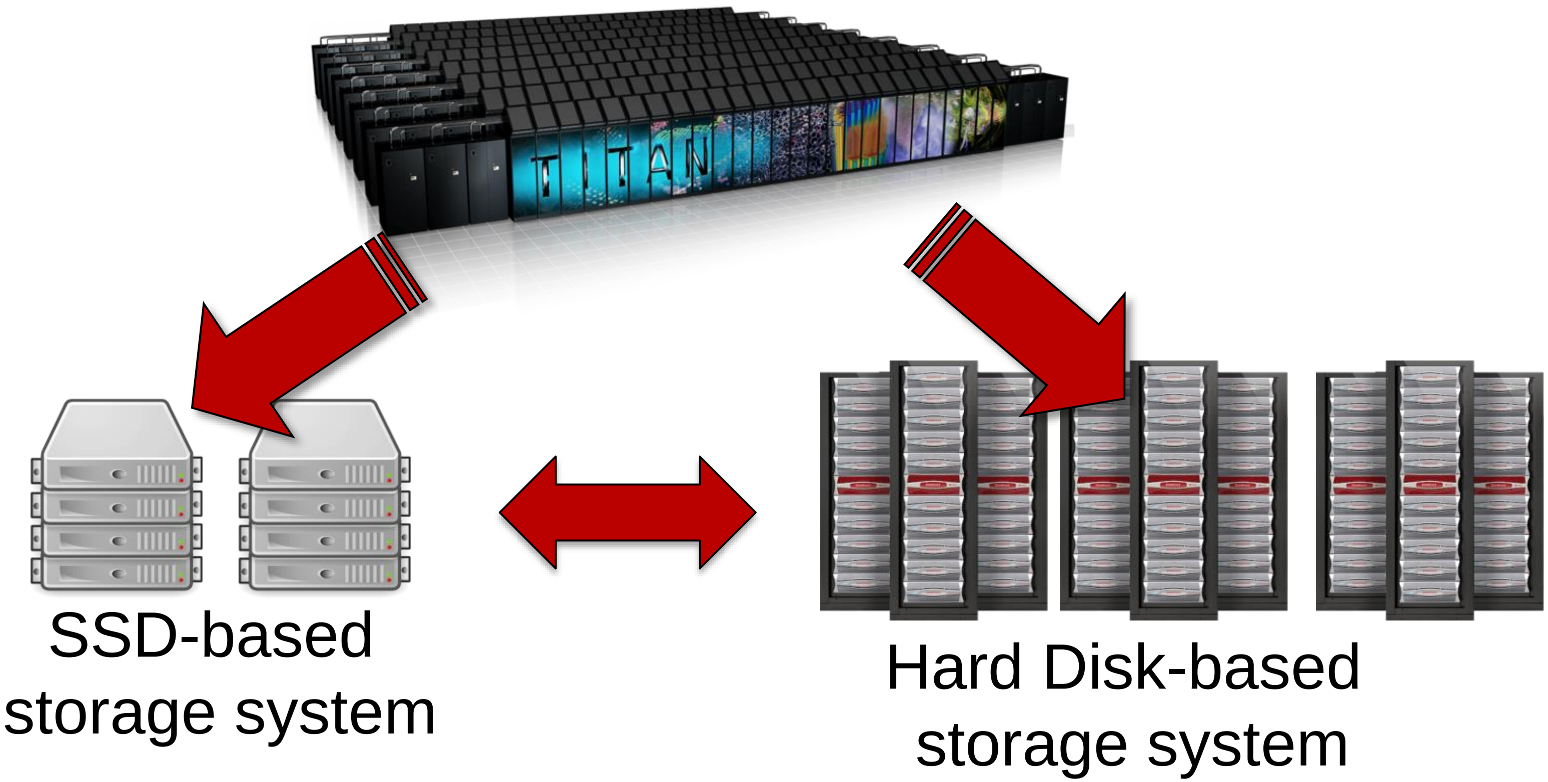
SSDs as a burst buffer in HPC



This model causes excessive data-movement

We are using SSDs for draining writes, something they are fundamentally not good at.

SSD in Exascale System Architecture



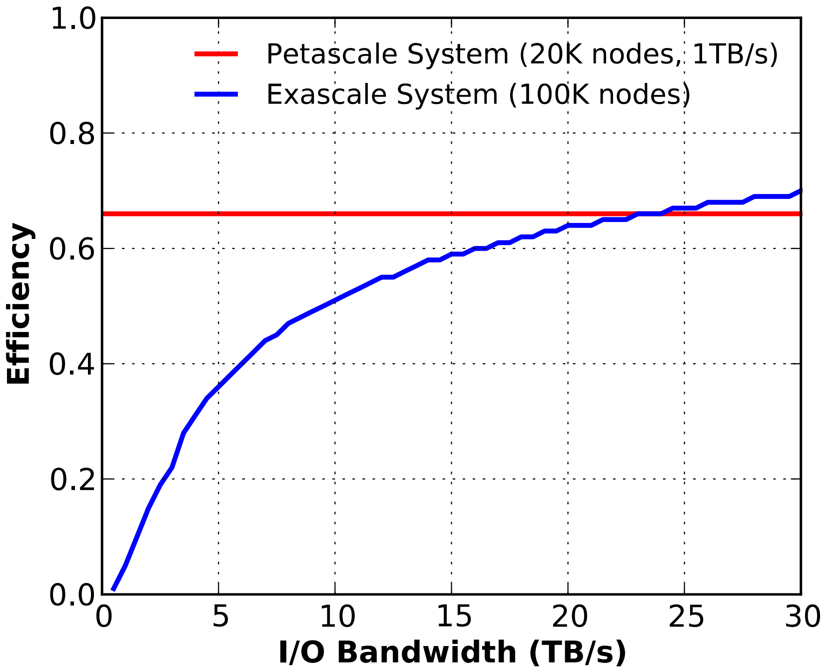
Users will be charged differently for different kind of storage system they use

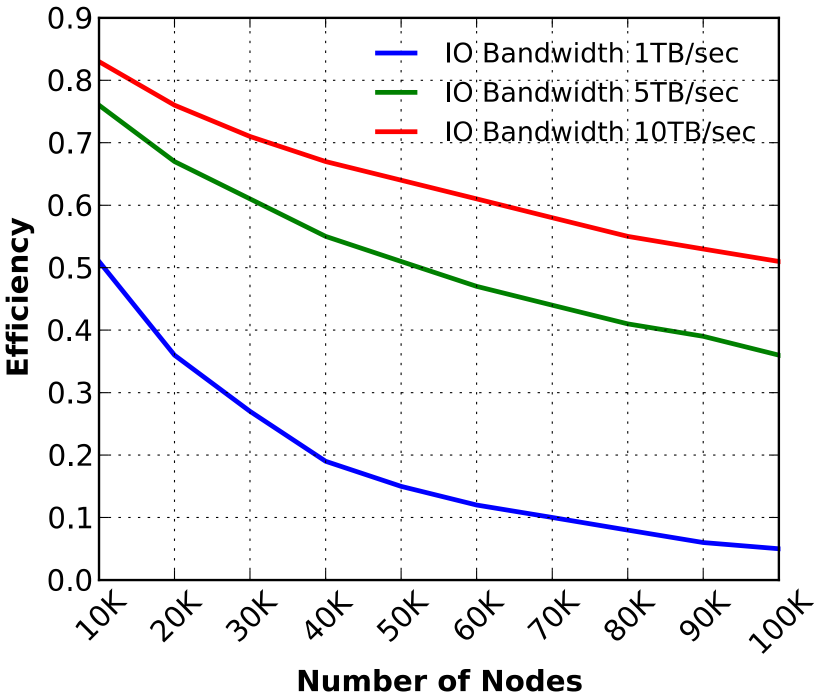
Disk system will become archival storage

Amount of waste work increases at larger scale system if we don’t have a fast enough storage system to quickly take a checkpoint

Capital investment in more expensive storage system promises higher pay-off during operation

Storage System Bandwidth

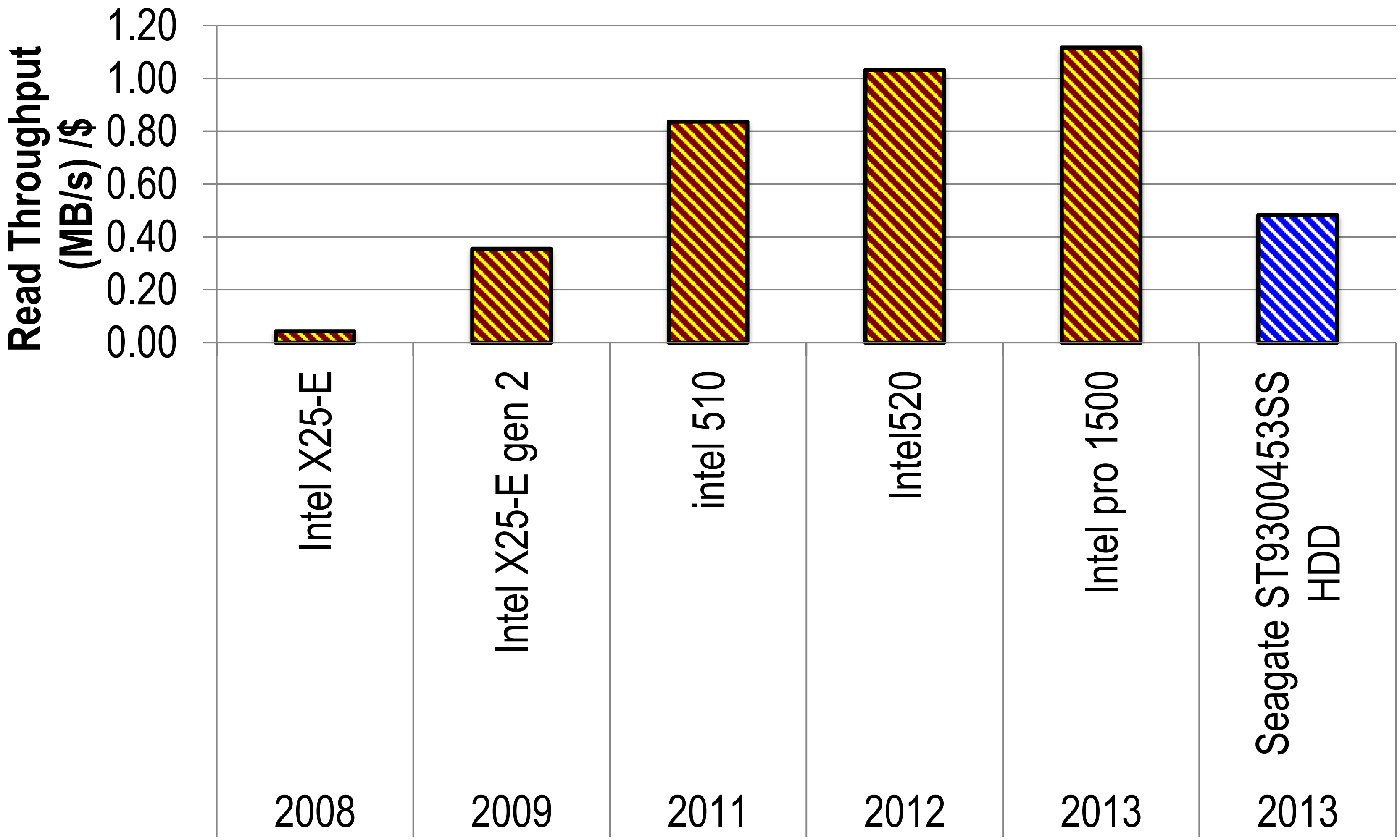




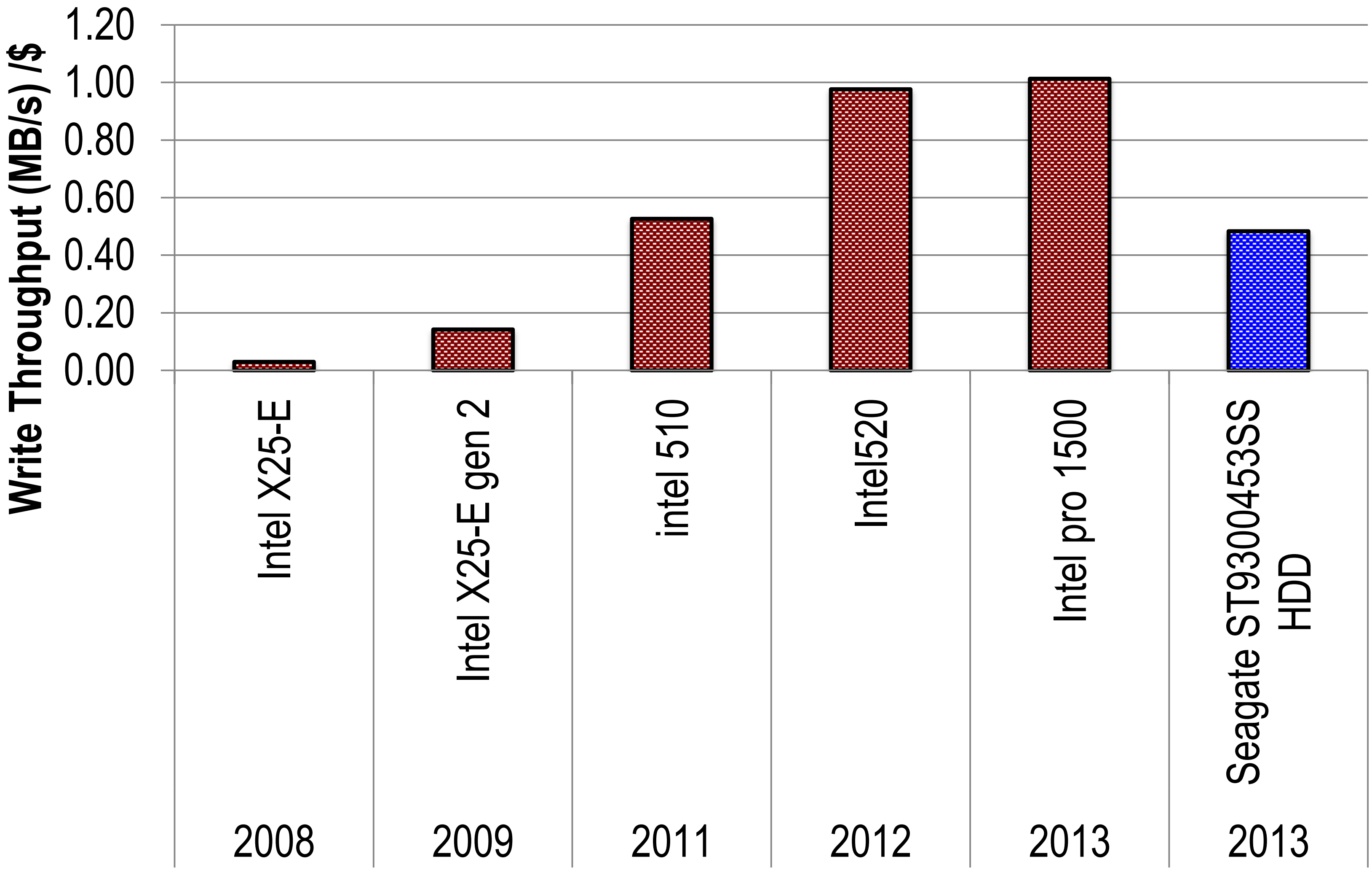
Storage system bandwidth determines the overall efficiency (i.e., amount of lost work due to failures)

Based on Daly's optimal checkpointing frequency (MTBF and time to checkpoint).

SSD vs HDD

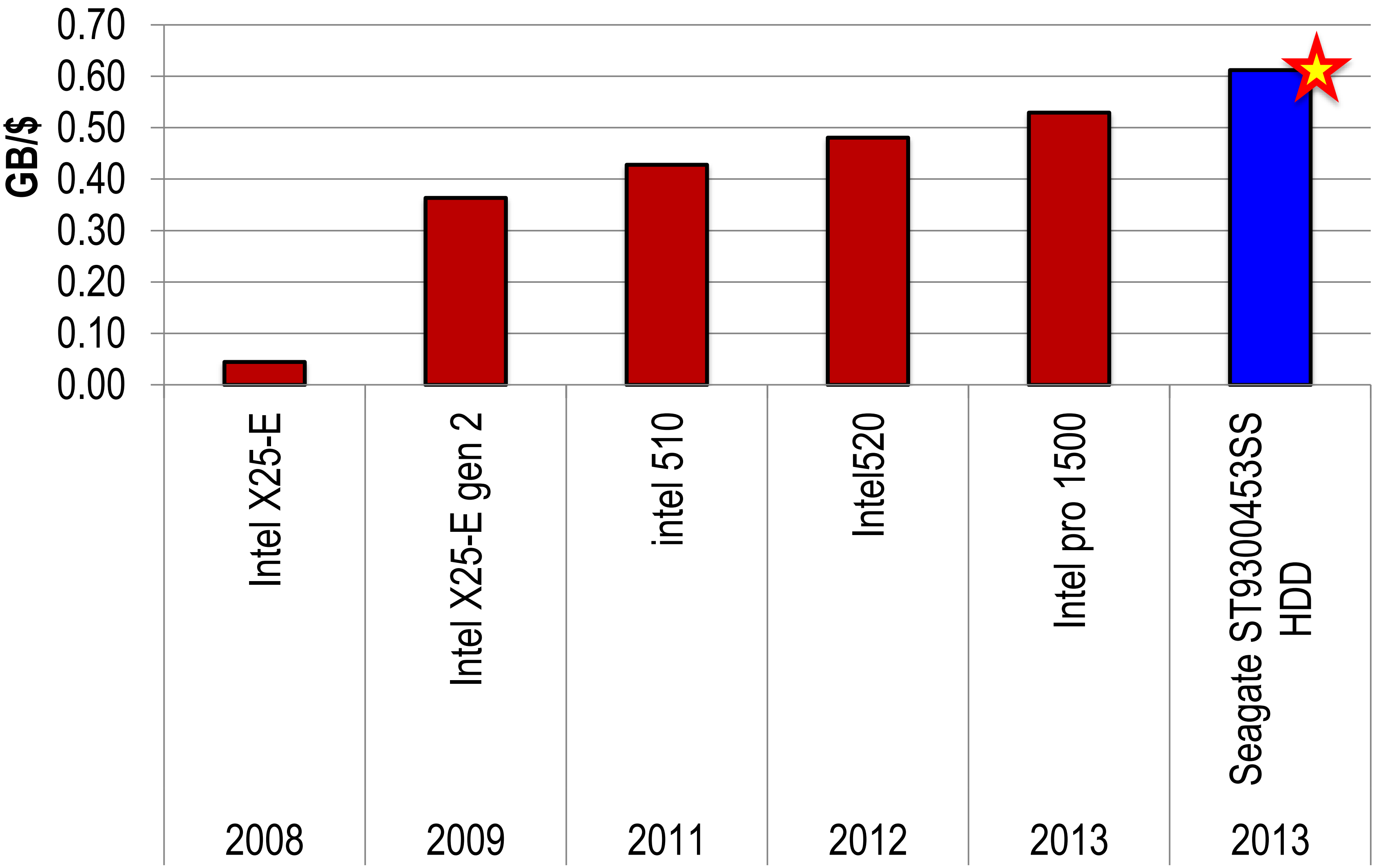

Read Throughput (MB/s) /$
1.20
1.00
0.80
0.60
0.40
0.20
0.00
Intel X25-E
Intel X25-E gen 2
intel 510
Intel520
Intel pro 1500
Seagate ST9300453SS HDD
2008
2009
2011
2012
2013
2013

SSD vs HDD



Write Throughput (MB/s) /$
1.20
1.00
0.80
0.60
0.40
0.20
0.00
Intel X25-E
Intel X25-E gen 2
intel 510
Intel520
Intel pro 1500
Seagate ST9300453SS HDD
2008
2009
2011
2012
2013
2013

SSD vs HDD


GB/$
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Intel X25-E
Intel X25-E gen 2
intel 510
Intel520
Intel pro 1500
Seagate ST9300453SS HDD
2008
2009
2011
2012
2013
2013

Building a HPC Storage System

Building blocks (cost factors):

- SSD/Hard Disk
- I/O Controller
- JBOD I/O module
- Enclosure
- Power and cooling
- RAID and firmware cost

Our preliminary study focuses on SSD/Hard disk and controllers costs.

Building a HPC Storage System

- HDD 1TB $200, 4TB $500, performance* 200MB/s
- SSD 256GB $250, 1TB $1000, performance* 500MB/s
- A pair of mid-scale controllers: throughput 8GB/s, $8K

- Each shelf can hold up to 60 drives
- 40 hard-disks saturate one pair of controllers
- 16 SSDs saturate one pair of controllers

*depends on the workload, lost on the route etc.

How to build a cost-effective high performance/high capacity HPC storage system?

Building a HPC Storage System

If planning to build a high capacity system:

GB/$ for SSDs is too low

Controller and other components do not matter

Hard-disks tend to become cheaper as density increases, but not true for SSDs

Building a HPC Storage System

Building 10TB/s system*

1280 pairs of controllers = ~$10million

Hard-disk based system 1280*40 (1TB) hard disks

disks cost = 1280*40*$200 = ~$10million

capacity = 50 PB, total cost = ~$20million

SSD based system 1280*40 (1TB) hard disks

disks cost = 1280*16*$250 = ~$5million

capacity = 5 PB, total cost = ~$15million

*numbers here only represent the trend, and are not (un)official quotes from WarpMechanics

Building a HPC Storage System

Controller cost very important factor in deciding the total capital cost of the system

This is is fundamentally different from enterprise computing that focuses on per drive IOPS

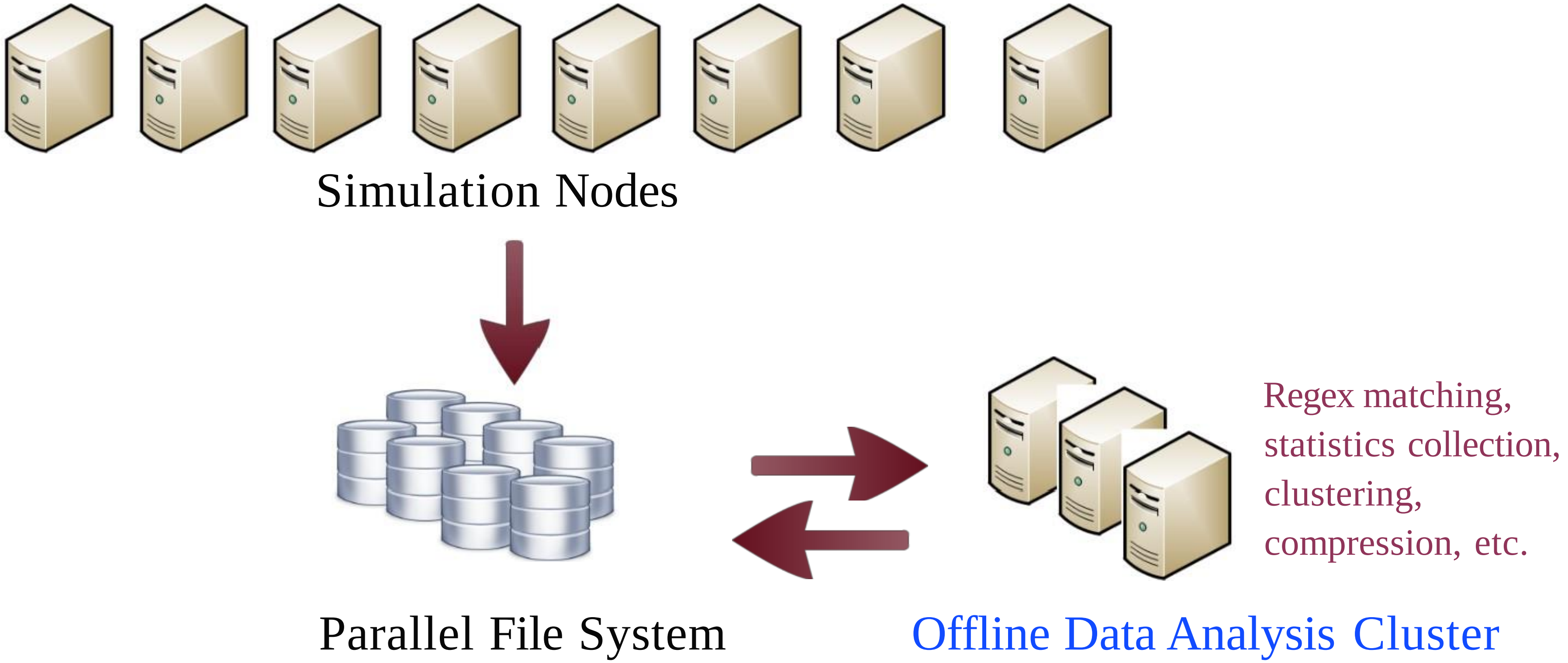
Need to account firmware, and other hardware cost

On-going work: workload characterization to explore a hybrid storage system, random vs. sequential performance, Lustre and RAID overhead etc.

What else can SSDs do for us?

Tiwari et al., Active Flash: Towards Energy-Efficient, In-Situ Data Analytics on Extreme-Scale Machines, USENIX FAST 2013.

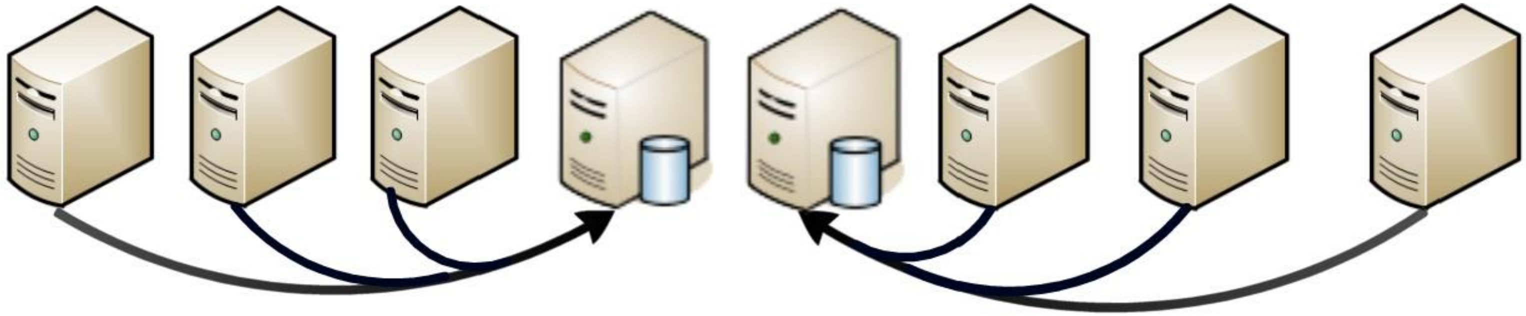
Traditional Scientific Data Analysis Approach



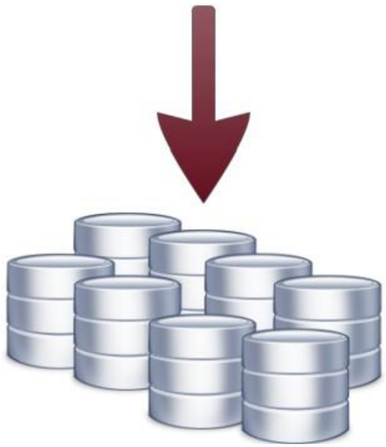
"Energy-cost for data movement at Exascale is likely to be of the same order of computation cost, if not more!"

-- Exascale Computing Study, 2008
Principle Investigator: Peter Kogge

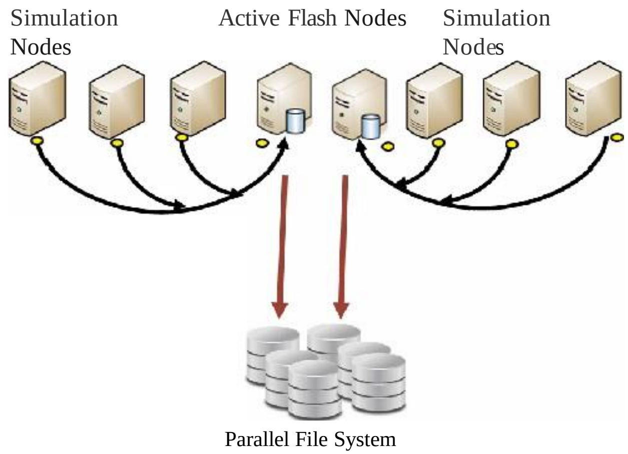
Active Computation on SSDs

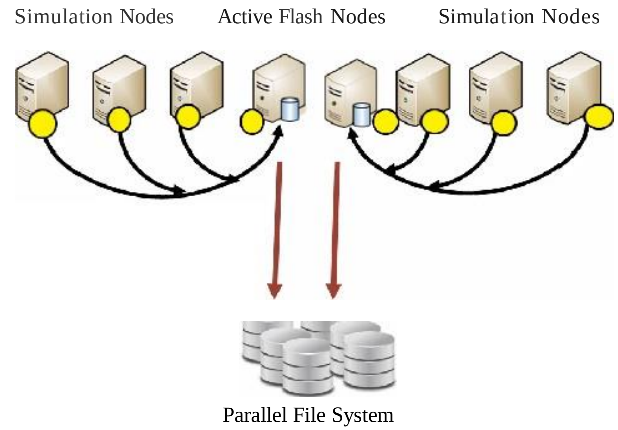


Scientific data analysis performed on SSD controllers
in-parallel with simulation without affecting it



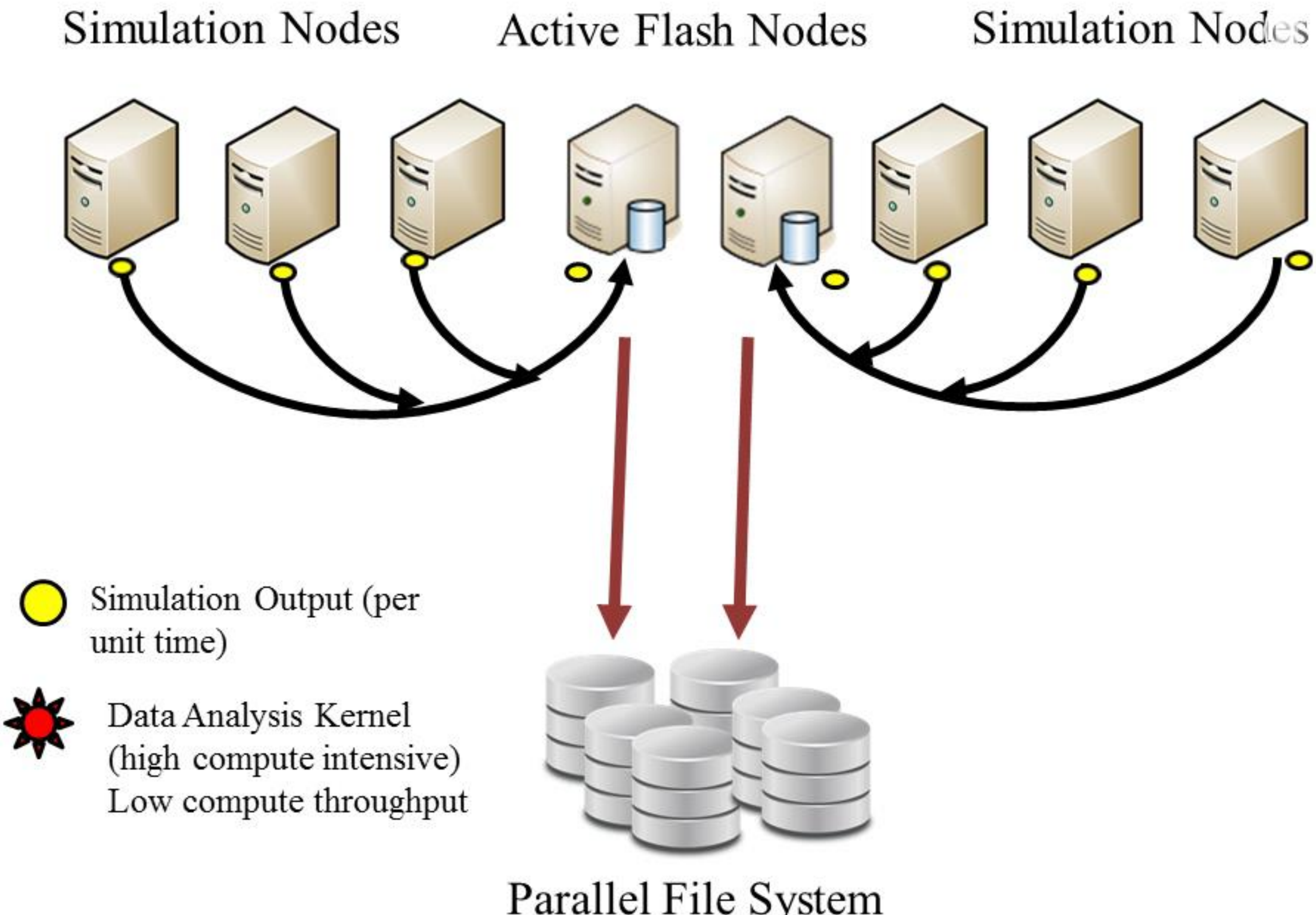
Parallel File System

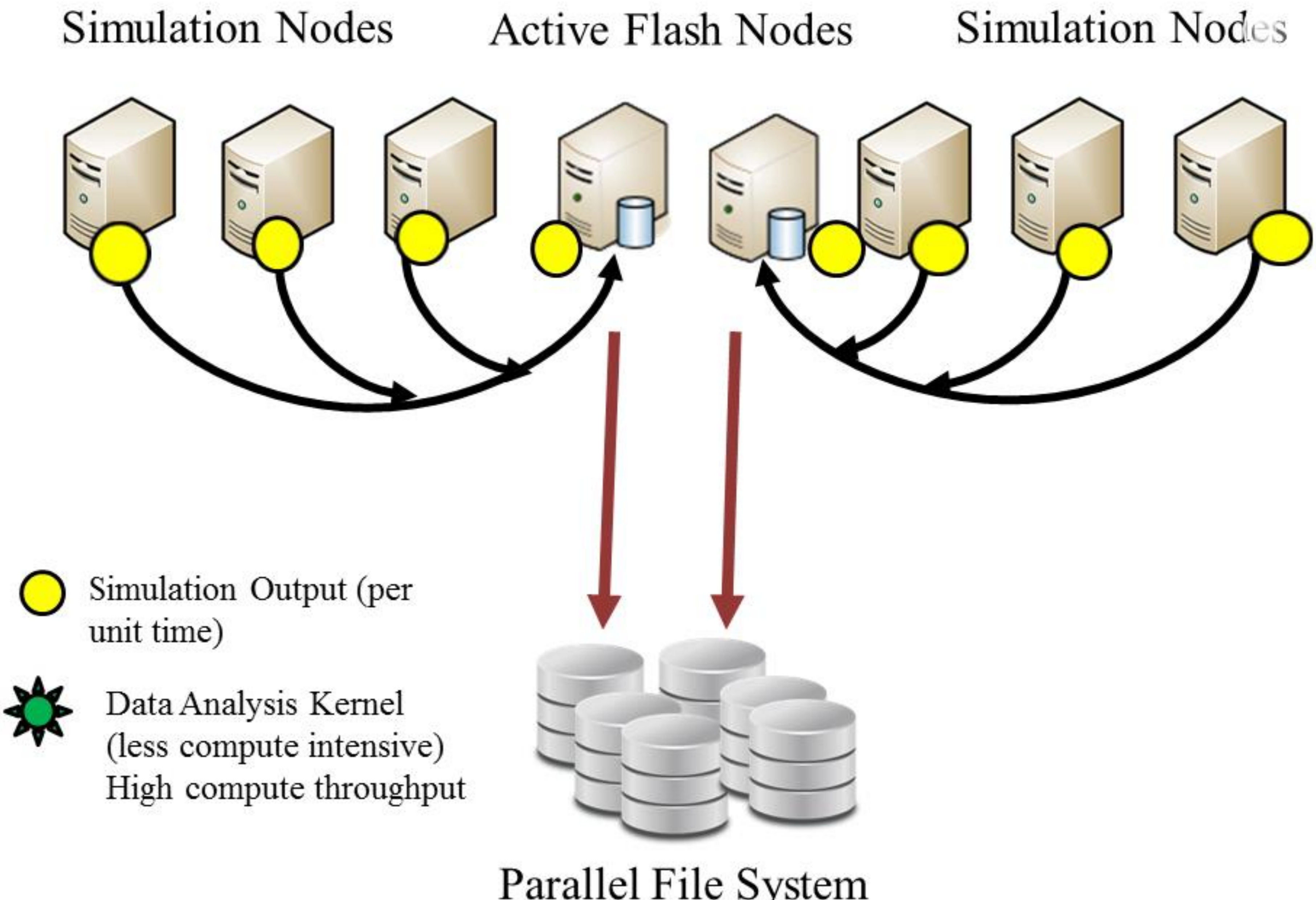


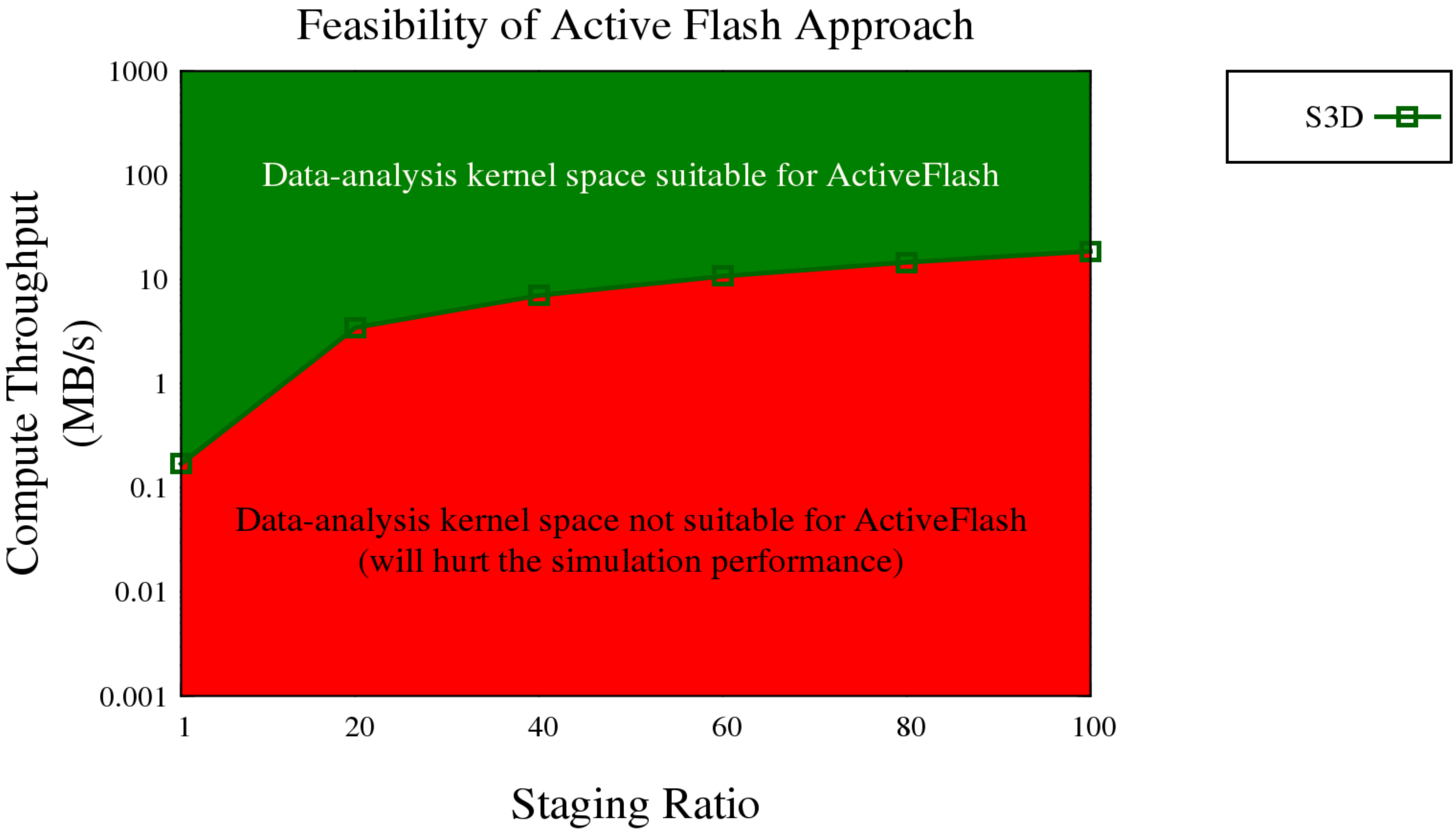


An analysis kernel needs to meet a "threshold compute throughput" to be placed on SSD controllers

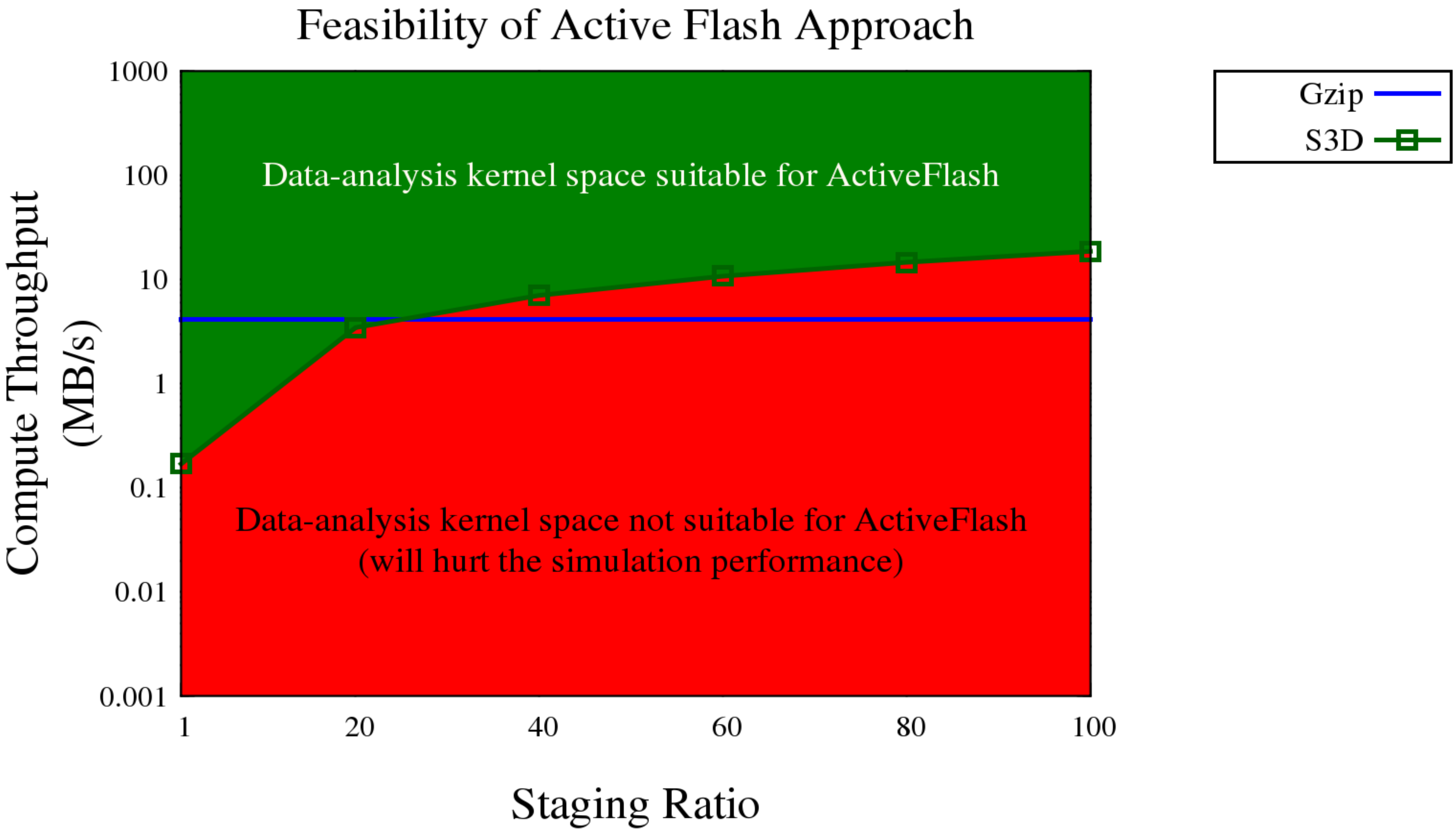
$$T_{SSD\_k} > \frac{\lambda_a \cdot R_{SSD}}{1 - \lambda_a \cdot R_{SSD} \cdot \left(\frac{1}{BW_{fm2c}} + \frac{1}{BW_{c2m}}\right) - \frac{N \cdot (\alpha \cdot \lambda_a + \lambda_c)}{BW_{PFS}} - \frac{t_i}{t_{iter}}}$$

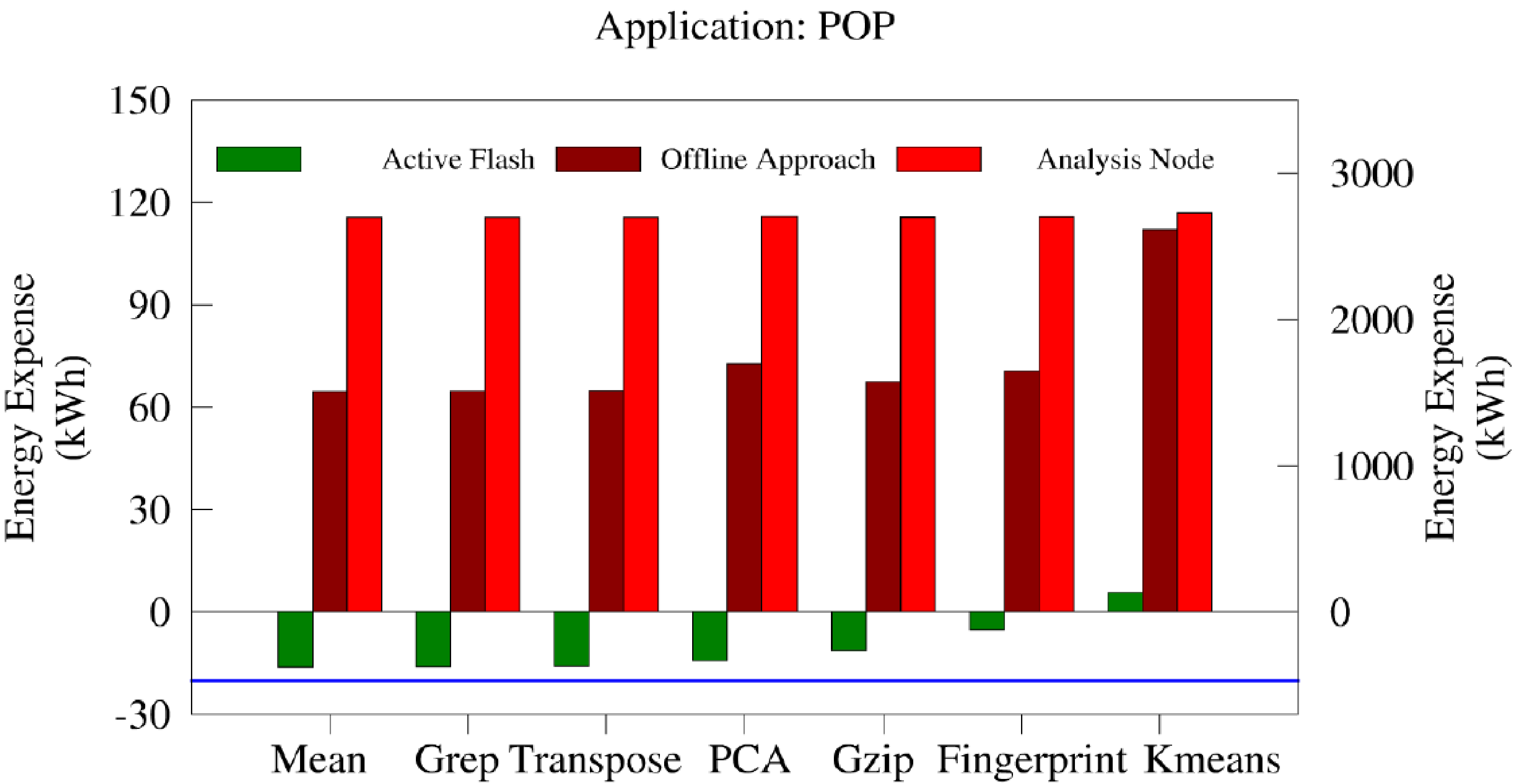

Simulation Nodes
Active Flash Nodes
Simulation Nodes
Simulation Output (per unit time)
Data Analysis Kernel (high compute intensive) Low compute throughput
Parallel File System


Simulation Nodes
Active Flash Nodes
Simulation Nodes
Simulation Output (per unit time)
Data Analysis Kernel (less compute intensive) High compute throughput
Parallel File System



Feasibility of Active Flash Approach
Data-analysis kernel space suitable for ActiveFlash
Data-analysis kernel space not suitable for ActiveFlash
(will hurt the simulation performance)
Compute Throughput (MB/s)
1000
100
10
1
0.1
0.01
0.001
1
20
40
60
80
100
Staging Ratio
S3D


Feasibility of Active Flash Approach
Data-analysis kernel space suitable for ActiveFlash
Data-analysis kernel space not suitable for ActiveFlash
(will hurt the simulation performance)
Compute Throughput (MB/s)
1000
100
10
1
0.1
0.01
0.001
1
20
40
60
80
100
Staging Ratio
Gzip
S3D


Application: POP
Active Flash
Offline Approach
Analysis Node
Energy Expense (kWh)
Energy Expense (kWh)
150
120
90
60
30
0
-30
3000
2000
1000
0
Mean
Grep
Transpose
PCA
Gzip
Fingerprint
Kmeans

Finding: Most data analysis kernels can be placed on SSD controllers without degrading simulation performance

Tiwari et al., Active Flash: Towards Energy-Efficient, In-Situ Data Analytics on Extreme-Scale Machines, USENIX FAST 2013.

Finding: Additional SSDs are not required for supporting in-situ data analysis on SSDs, beyond what is needed for sustaining the I/O requirements of scientific applications

Tiwari et al., Active Flash: Towards Energy-Efficient, In-Situ Data Analytics on Extreme-Scale Machines, USENIX FAST 2013.

ActiveFlash Prototype based on OpenSSD Platform

Prototype demonstrates the viability of our approach

Changes only in the FTL, no hardware changes

Preemption based scheduling

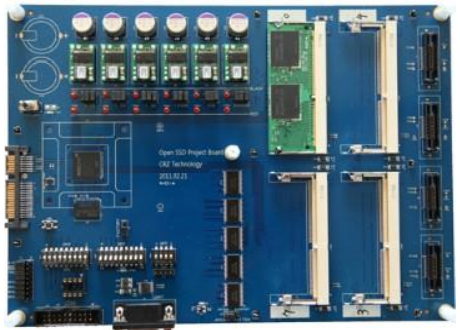
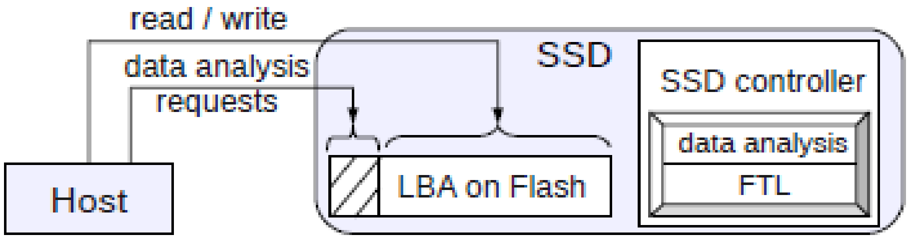


Figure courtesy: open-ssd project

See USENIX paper for the details and evaluation results



Contact

Devesh Tiwari
tiwari@ornl.gov