

Background



Energy and Cost Saving Analysis

"Active Flash" Energy Modeling

Modeled after Samsung PM1725 SSD

Total energy consists of multiple components

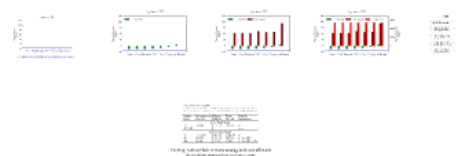
SSD energy during I/O, compute, and idle periods
Data movement energy cost in the interconnect

"Offline" and "Analysis Node" Approach Energy Modeling

Modeled after Intel Core i7 processors

Assumed idle when not doing data analysis

Optimistic modeling
cooling, assembly and installation costs ignored



Active Flash: Towards Energy-Efficient, In-Situ Data Analytics on Extreme-Scale Machines



Danyang Zhang



NC State University Oak Ridge National Lab Northeastern University

*Funding: NSFC is a joint funding agreement with ORNL



Active Computation Feasibility

Modeling SSD Deployment without Active Computation Support

Multiple constraints:
- Capacity: enough data to store original data
- Latency: high enough to allow for multiple read requests
- Power: low enough to allow for multiple read requests
- Cost: low enough to allow for multiple read requests

Modeling Active Computation Feasibility

Modeling active computation feasibility:
- Latency: low enough to allow for multiple read requests
- Power: low enough to allow for multiple read requests
- Cost: low enough to allow for multiple read requests

Staging Ratio

Staging ratio 1.0 seems to work well for all applications except CHIMERA

1.0



Staging Ratio

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Staging ratio 1.0 seems to work well for all applications except CHIMERA

Conclusion

Active computation on SSDs enables energy-efficient in-situ data analysis in Supercomputing

In most cases, Active Flash does not require extra SSDs

Active Flash may even help cut SSD deployment cost by reducing electricity bill

Active Flash for scientific data analysis viable with OpenSSD

Problems and Challenges

Offline approach to data analysis involves multiple rounds of I/O, causing

excessive data movement

excessive energy cost

"Throughput" for data movement is considered to be the same order of magnitude as "I/O cost"

"Throughput" for data movement is considered to be the same order of magnitude as "I/O cost"

"Throughput" for data movement is considered to be the same order of magnitude as "I/O cost"

Using simulation results for data analysis not accurate

High CPU utilization cost on a Supercomputer

Active Flash Approach for In-situ Scientific Data Analysis



ActiveFlash Prototype based on OpenSSD Platform

Prototype demonstrates the viability of our approach

Changes only in the FTL, no hardware changes

Preemption based scheduling

See paper for the details and evaluation results



Figure courtesy: open-ssd project



Thank You!

Active Flash: Towards Energy-Efficient, In-Situ Data Analytics on Extreme-Scale Machines



Devesh Tiwari



Simona
Boboila



Sudharshan
Vazhkudai



Youngjae
Kim



Xiaosong
Ma



Peter
Desnoyers



Yan
Solihin

NC State University

Oak Ridge National Lab

Northeastern University

*Xiaosong Ma holds a joint faculty appointment with ORNL.

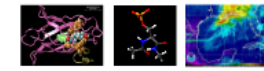
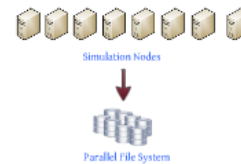
Background



Scientific Discovery: Two-Step Process



Traditional Scientific Simulation Setup

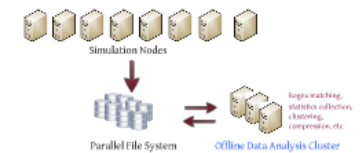


Astrophysics, climate modeling, combustion and fusion applications

Application	Analysis data generation rate (per node)	Visualization data generation rate (per node)
CHIMERA	440 GB/s	440 GB/s
VULCAN-3D	7.5 GB/s	9.5 GB/s
POP	16.7 GB/s	1.5 GB/s
CUF	100 GB/s	65 GB/s
OTW	100 GB/s	45 GB/s
OTW	11 GB/s	11 GB/s

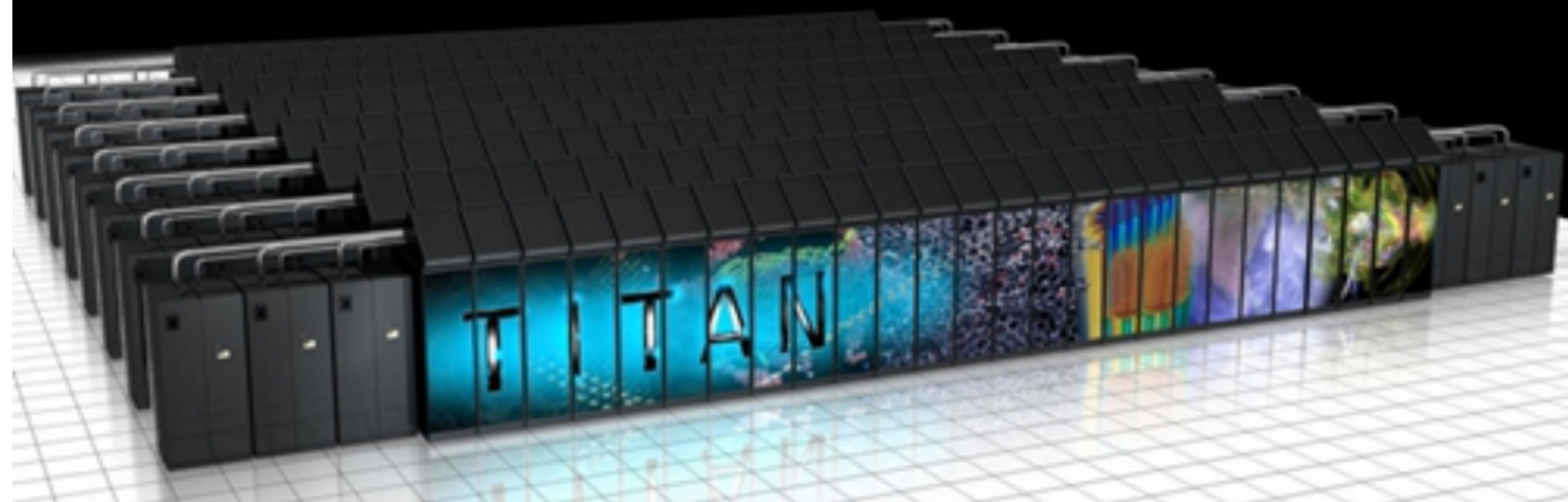
OTW produces ~50TB output data per hour at scale

Traditional Scientific Data Analysis Approach



World's #1 Open Science Supercomputer

Flagship accelerated computing system | 200-cabinet Cray XK7 supercomputer |
18,688 nodes (AMD 16-core Opteron + NVIDIA Tesla K20 GPU) |
CPUs/GPUs working together – GPU accelerates | 20+ Petaflops



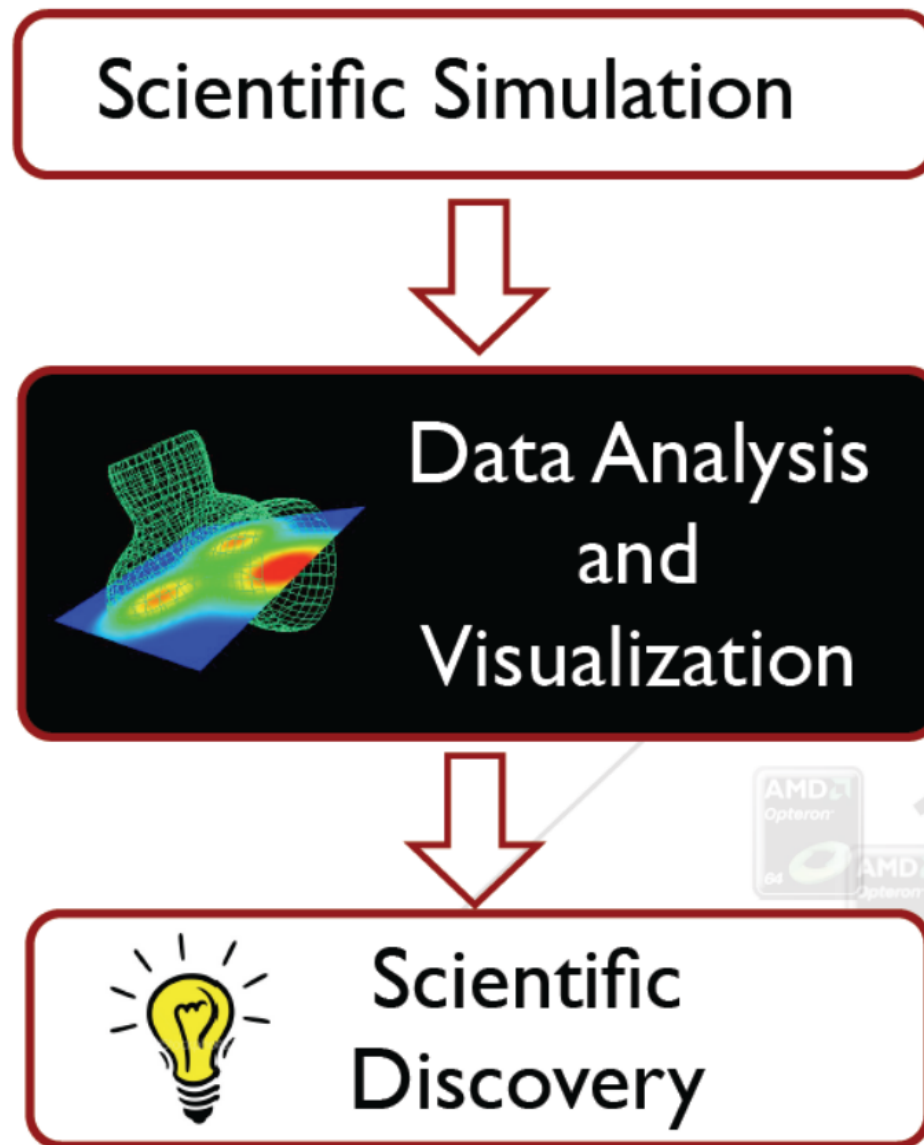


World's Most Powerful Computer For Science!

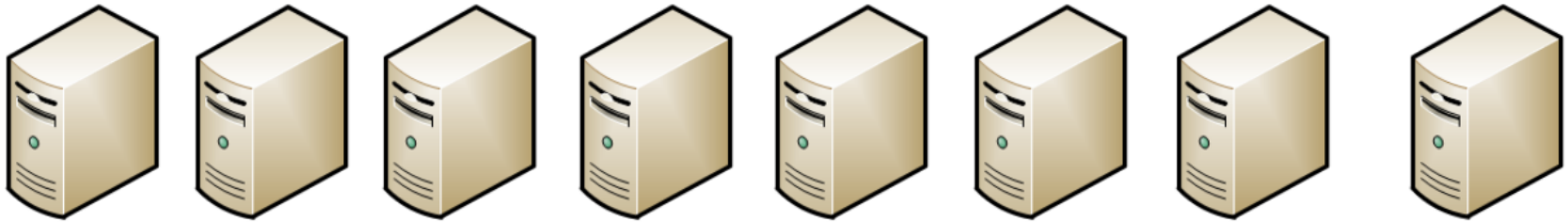
"The Jaguar system at ORNL provides immense computing power in a balanced, stable system that is allowing scientists and engineers to tackle some of the world's most challenging problems."

—2008, Kelvin Droegemeier, Meteorology Professor, University of Oklahoma.

Scientific Discovery: Two-Step Process



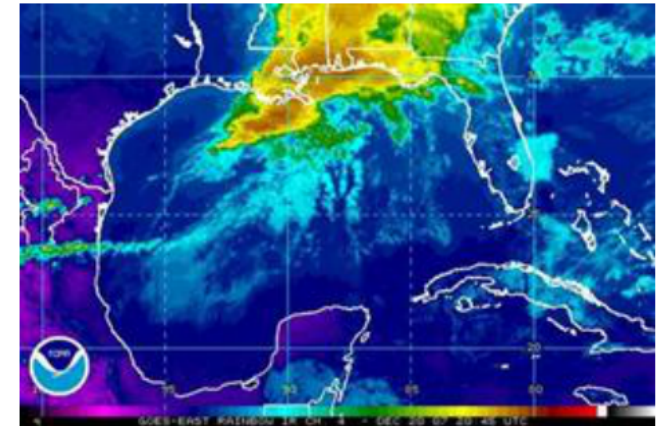
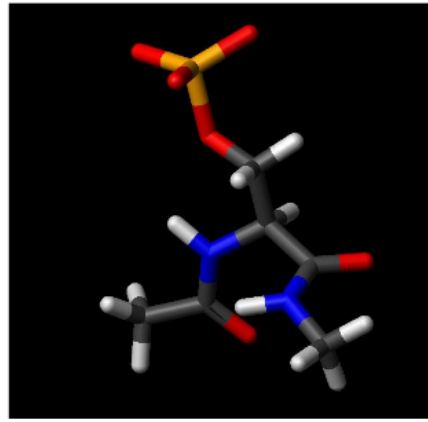
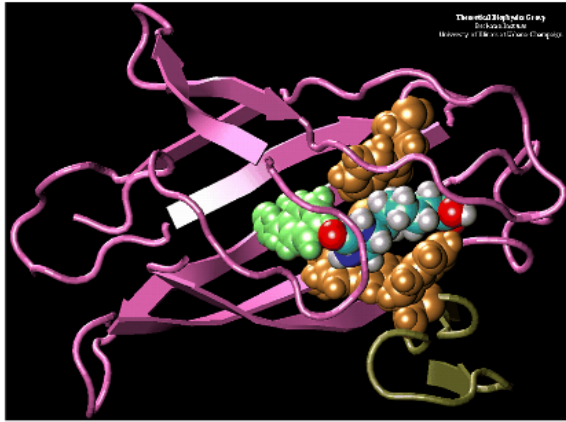
Traditional Scientific Simulation Setup



Simulation Nodes



Parallel File System



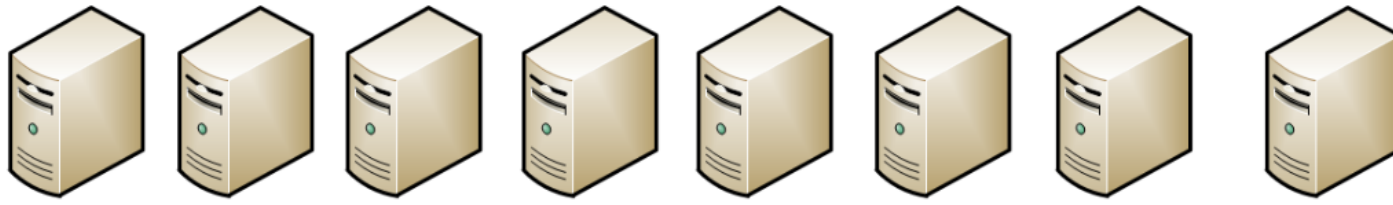
Large-scale leadership computing applications produce big-data

Astrophysics, climate modeling, combustion and fusion applications

Application	Analysis data generation rate (per node)	Checkpoint data generation rate (per node)
CHIMERA	4400 KB/s	4400 KB/s
VULCUN/2D	2.28 KB/s	0.02 KB/s
POP	16.3 KB/s	5.05 KB/s
S3D	170 KB/s	85 KB/s
GTC	14 KB/s	476 KB/s
GYRO	14 KB/s	11.6 KB/s

GTC produces ~30TB output data per hour at-scale.

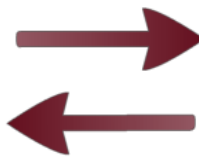
Traditional Scientific Data Analysis Approach



Simulation Nodes



Parallel File System



Regex matching,
statistics collection,
clustering,
compression, etc.

Offline Data Analysis Cluster

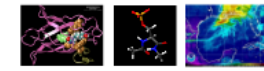
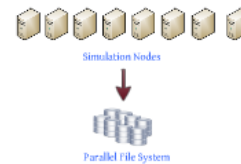
Background



Scientific Discovery: Two-Step Process



Traditional Scientific Simulation Setup

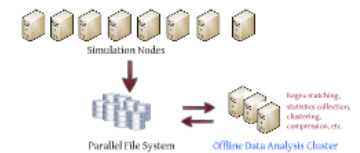


Astrophysics, climate modeling, combustion and fusion applications

Application	Analysis data generation rate (per node)	Visualization data generation rate (per node)
CHIMERA	440 GB/s	440 GB/s
VULCAN-3D	7.5 GB/s	9.5 GB/s
POP	16.7 GB/s	1.5 GB/s
CUF	100 GB/s	85 GB/s
QTM	100 GB/s	45 GB/s
QTM-2	11 GB/s	11 GB/s

QTM produces ~50TB output data per hour at scale

Traditional Scientific Data Analysis Approach



Problems and Challenges

Offline approach to data analysis involves multiple rounds of I/O, causing

- Excessive data movement
- Extra energy cost

"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

Using simulation nodes for data analysis not acceptable

- High CPU allocation cost on a Supercomputer

Offline approach to data analysis involves multiple rounds of I/O, causing

- Excessive data movement
- Extra energy cost

"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

Using simulation nodes for data analysis not acceptable

- High CPU allocation cost on a Supercomputer

Problems and Challenges

Offline approach to data analysis involves multiple rounds of I/O, causing

- Excessive data movement
- Extra energy cost

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

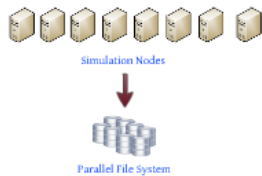
— Exascale Computing Study, 2008
Principle Investigator: Peter Kogge

Using simulation nodes for data analysis not acceptable

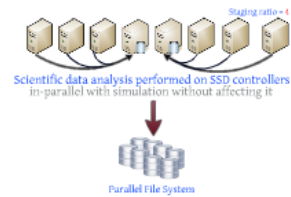
- High CPU allocation cost on a Supercomputer

Active Flash Approach for In-situ Scientific Data Analysis

Traditional Scientific Simulation Setup



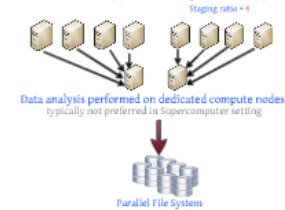
Active Computation on SSDs



Enabling Trends for Active Flash

- SSDs now being adopted in Supercomputers (e.g. Tsukuba, Gordon)
higher I/O throughput and storage capability
- SSD controllers becoming increasingly powerful
multi-core low-power processors
- Idle cycles at SSD controllers
I/O behavior of scientific workloads bursty in nature
- In-situ analysis inherently more energy efficient
reduction in data movement cost

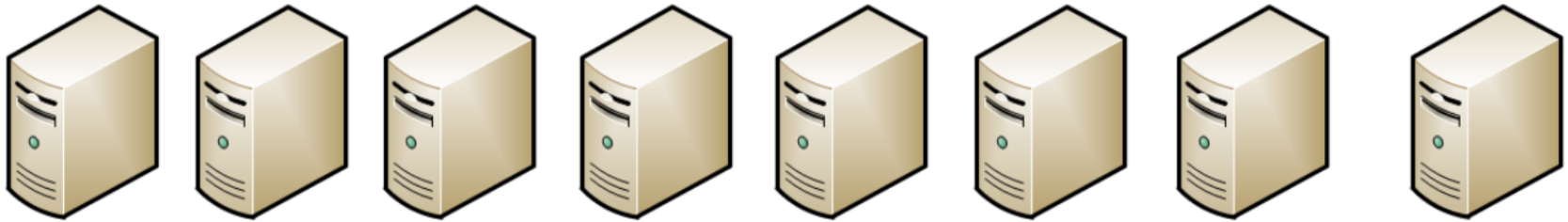
An Alternative Approach (Analysis Node Approach)



This work answers the following:

- If SSDs are deployed with only I/O performance in mind, then is active computation even feasible?
- Will additional SSD provisioning be required?
- Will active computation slowdown the main simulation nodes?
- How much energy and cost saving can Active Flash bring?

Traditional Scientific Simulation Setup

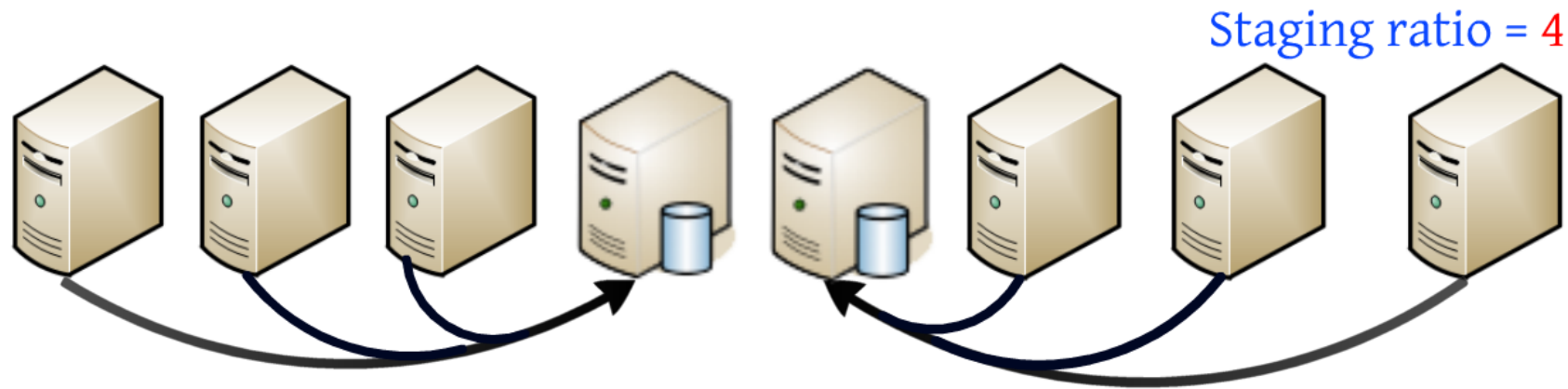


Simulation Nodes



Parallel File System

Active Computation on SSDs



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



Parallel File System

Enabling Trends for Active Flash



SSDs now being adopted in Supercomputers (e.g. Tsubame, Gordon)
higher I/O throughput and storage capability



SSD controllers becoming increasingly powerful
multi-core low-power processors



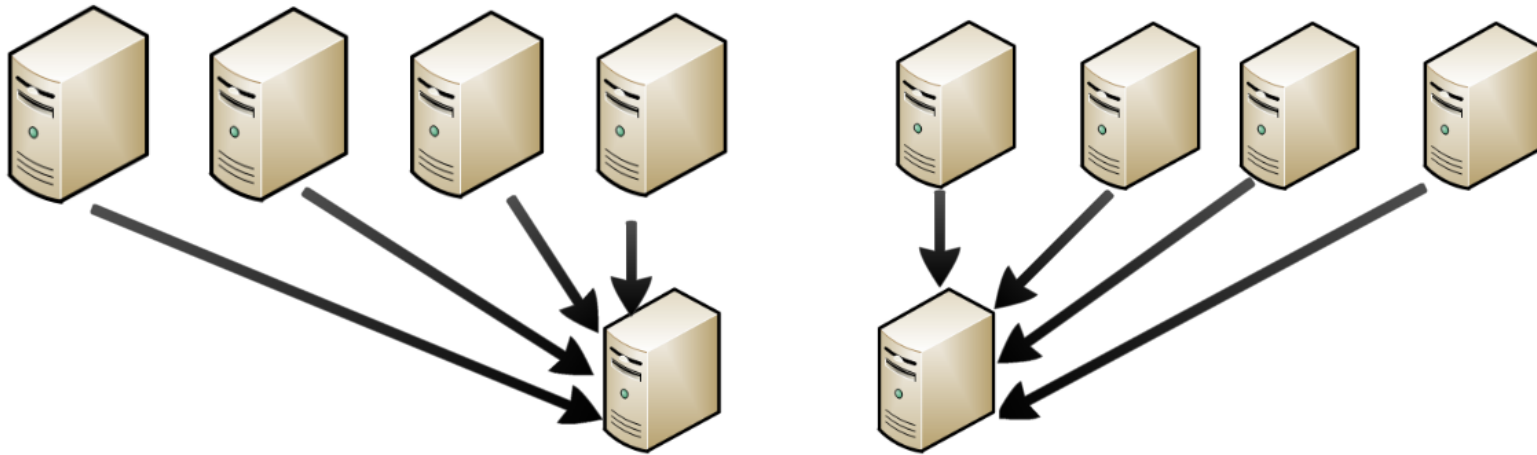
Idle cycles at SSD controllers
I/O behavior of scientific workloads bursty in nature



In-situ analysis inherently more energy efficient
reduction in data movement cost

An Alternative Approach (Analysis Node Approach)

Staging ratio = 4



Data analysis performed on dedicated compute nodes
typically not preferred in Supercomputer setting



Parallel File System

This work answers the following:



If SSDs are deployed with only I/O performance in mind, then is active computation even feasible?



Will additional SSD provisioning be required?



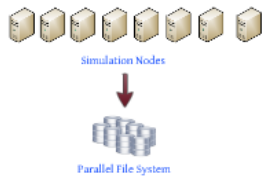
Will active computation slowdown the main simulation nodes?



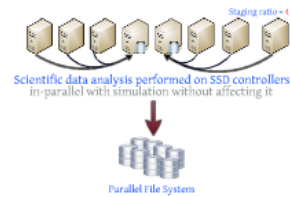
How much energy and cost saving can Active Flash bring?

Active Flash Approach for In-situ Scientific Data Analysis

Traditional Scientific Simulation Setup



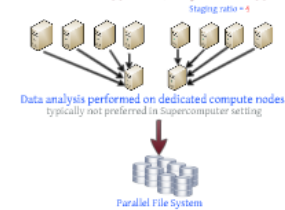
Active Computation on SSDs



Enabling Trends for Active Flash

- SSDs now being adopted in Supercomputers (e.g. Tsukuba, Gordon)
higher I/O throughput and storage capability
- SSD controllers becoming increasingly powerful
multi-core low-power processors
- Idle cycles at SSD controllers
I/O behavior of scientific workloads bursty in nature
- In-situ analysis inherently more energy efficient
reduction in data movement cost

An Alternative Approach (Analysis Node Approach)



This work answers the following:

- If SSDs are deployed with only I/O performance in mind, then is active computation even feasible?
- Will additional SSD provisioning be required?
- Will active computation slowdown the main simulation nodes?
- How much energy and cost saving can Active Flash bring?

Active Computation Feasibility

Modeling SSD Deployment **without** Active Computation Support

Multiple constraints:

Capacity

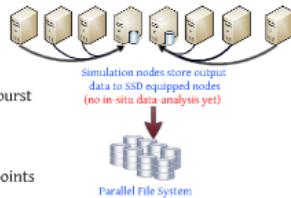
- Enough SSDs to sustain one output burst

Performance

- High I/O bandwidth to SSD space
- Fast restart from application checkpoints

Write durability

- SSD write endurance limits



Modeling Active Computation Feasibility

Simulation Applications

CHIMERA
VULCAN
POP
S3D
GTC
GYRO



Data Analysis Kernels

Statistics Collection
PCA
Grep
Gzip
Fingerprinting
Clustering



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

$$T_{SSD,k} > \frac{\lambda_k \cdot R_{SSD}}{1 - \lambda_k \cdot R_{SSD} \cdot \left(\frac{1}{BW_{SSD}} + \frac{1}{BW_{CPU}} \right) + \frac{N \cdot (k \cdot \lambda_k + k_k)}{BW_{CPU}}} = \frac{\lambda_k}{k_{thr}}$$

Relatively less compute intensive kernels better suited (e.g. regex matching) for active computation

Less computation intensive → high compute throughput

Dependent on multiple factors: simulation data production rate, staging ratio, I/O bandwidth, etc.

Staging Ratio

How many simulation nodes share one common SSD?



Staging ratio determined by the most restrictive constraint

$$R_{SSD} = \min(R_{capacity}, R_{bandwidth}, R_{restart}, R_{endurance})$$

$$R_{capacity} = \frac{C_{SSD}}{f_{app} \cdot (n + BW_{SSD} \cdot \tau) \cdot \lambda_{thr}}$$

$$R_{bandwidth} = \frac{BW_{SSD} - SSD}{BW_{CPU}}$$

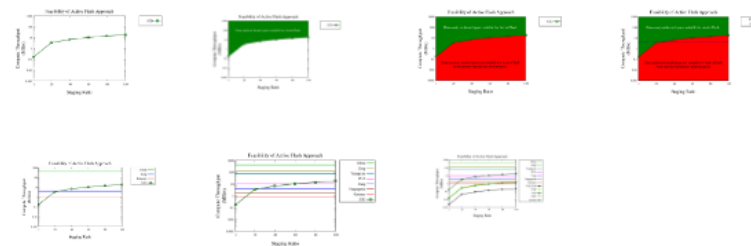
$$R_{restart} = \frac{f_{restart} \cdot BW_{SSD} - SSD}{\lambda_k}$$

$$R_{endurance} = \frac{W_{endurance}}{(\lambda_k + k_k) \cdot E_{thr}}$$

Modeled Jaguar Supercomputer consists of 18000 nodes
Staging ratio of 10 means 1800 SSDs

Active Flash Model						
	CHIMERA	VULCAN	POP	S3D	GTC	GYRO
$R_{capacity}(32GB)$	1	2571	233	18	6	166
$R_{capacity}(64GB)$	1	4500	461	36	12	333
$R_{bandwidth}$	29	29	29	29	29	29
$R_{endurance}$	1	2268	245	20	10	204
$R_{restart}$	4	896218	4054	240	42	1758

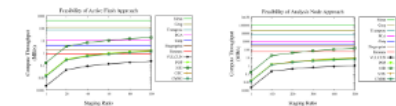
Staging ratio 10 seems to work well for all applications except CHIMERA



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

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

Feasibility of the Analysis Node Approach



Finding: Analysis node approach is feasible at higher staging ratios, but at additional infrastructure cost (see paper)

Modeling SSD Deployment **without** Active Computation Support

Multiple constraints:

Capacity

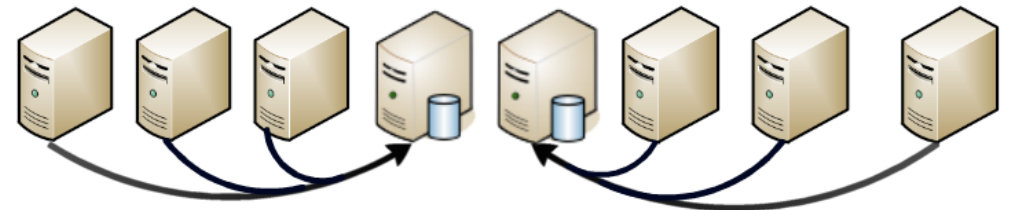
- Enough SSDs to sustain one output burst

Performance

- High I/O bandwidth to SSD space
- Fast restart from application checkpoints

Write durability

- SSD write endurance limits



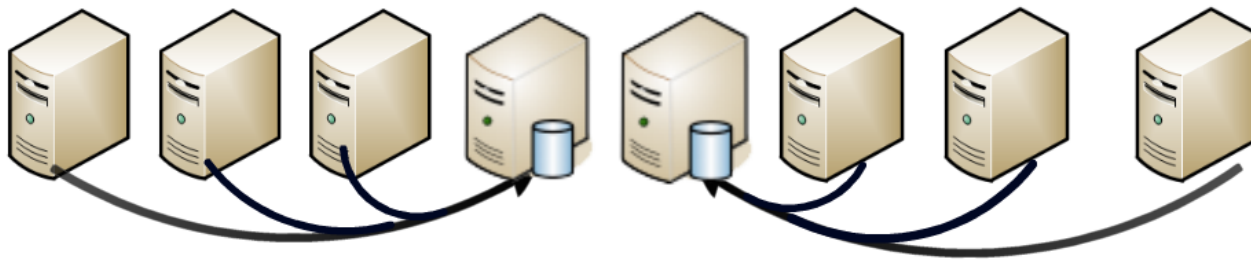
Simulation nodes store output data to SSD equipped nodes
(no in-situ data-analysis yet)



Parallel File System

Staging Ratio

How many simulation nodes share one common SSD?



Staging ratio = 4

Staging ratio determined by the most restrictive constraint

$$R_{SSD} = \min(R_{capacity}, R_{bandwidth}, R_{restart}, R_{endurance})$$



$$R_{capacity} = \frac{C_{SSD}}{f_{op} \cdot (a + num_{chkpts} \cdot c) \cdot t_{iter}} \quad R_{bandwidth} = \frac{BW_{sim-SSD}}{\frac{BW_{PFS}}{N}} \quad R_{restart} = \frac{f_{restart} \cdot BW_{sim-SSD}}{\lambda_c} \quad R_{endurance} = \frac{W_{endurance}}{(\lambda_a + \lambda_c) \cdot U_{time}}$$

Modeled Jaguar Supercomputer consists of 18000 nodes
Staging ratio of 10 means 1800 SSDs

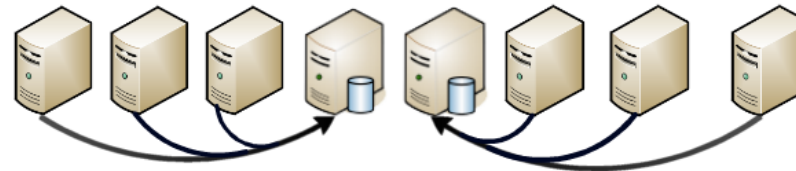
Active Flash Model						
	CHIMERA	VULCAN	POP	S3D	GTC	GYRO
$R_{capacity}(32\text{ GB})$	1	2571	233	18	6	166
$R_{capacity}(64\text{ GB})$	1	4500	461	36	12	333
$R_{bandwidth}$	29	29	29	29	29	29
$R_{endurance}$	1	2268	245	20	10	204
$R_{restart}$	4	896218	4054	240	42	1758

Staging ratio 10 seems to work well for all
applications except CHIMERA

Modeling Active Computation Feasibility

Simulation Applications

CHIMERA
VULCUN
POP
S3D
GTC
GYRO



Data analysis tasks need to finish before
next wave of data arrives at SSDs

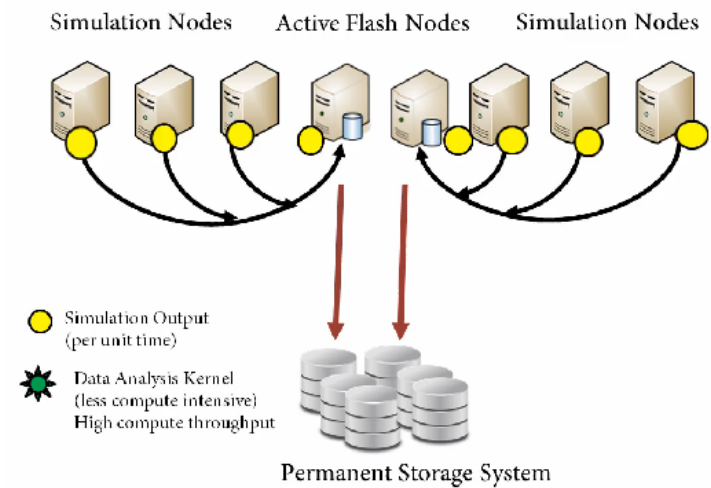
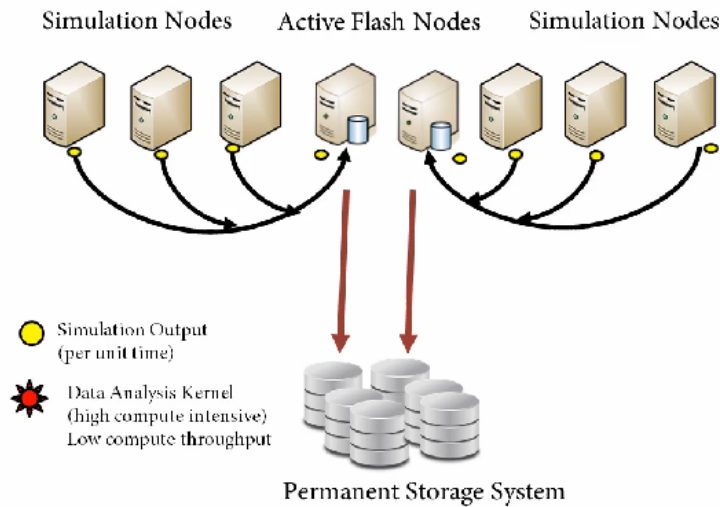
Data Analysis Kernels

Statistics Collection
PCA
Grep
Gzip
Fingerprinting
Clustering

Relatively less compute intensive kernels better suited (e.g. regex matching)
for active computation

Less computation intensive -> high compute throughput

Dependent on multiple factors: simulation data production rate, staging ratio,
I/O bandwidth, etc.



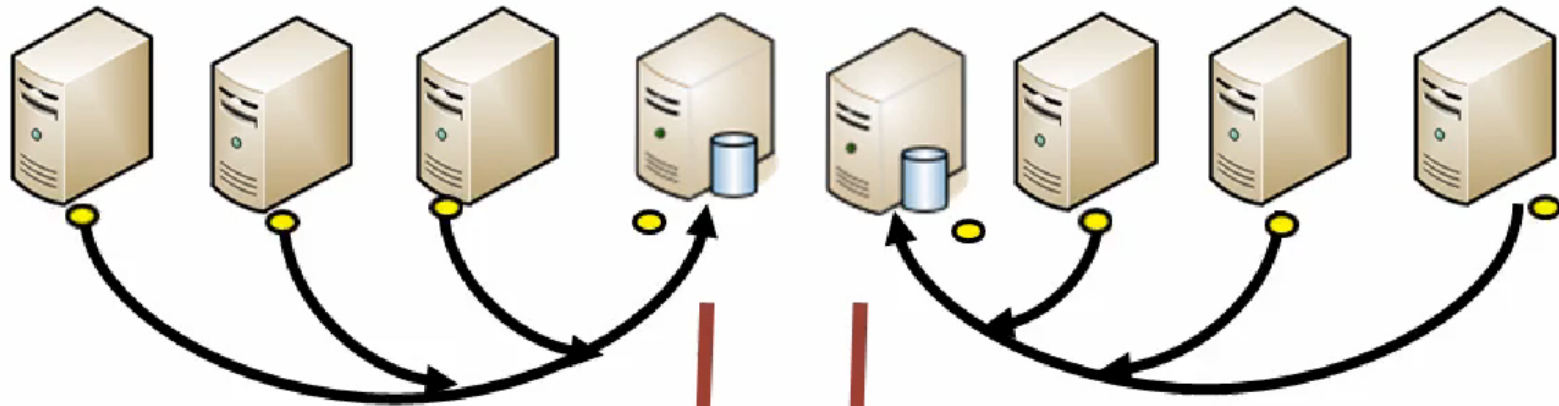
An analysis kernel needs to meet a "threshold compute throughput" to be placed on SSD controllers


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


Simulation Nodes

Active Flash Nodes

Simulation Nodes



 Simulation Output
(per unit time)

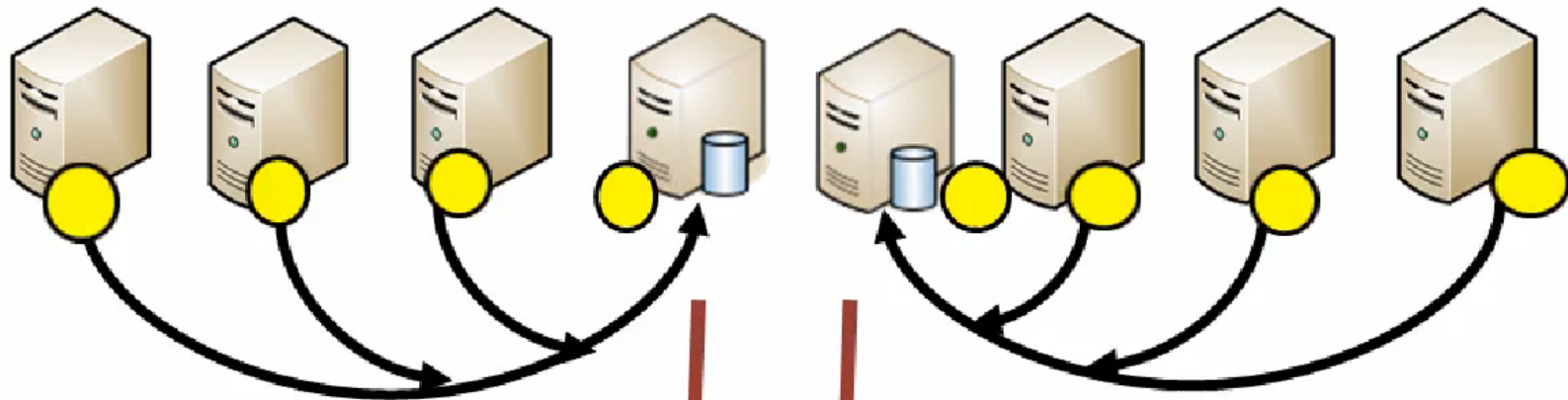
 Data Analysis Kernel
(high compute intensive)
Low compute throughput


Permanent Storage System


Simulation Nodes

Active Flash Nodes

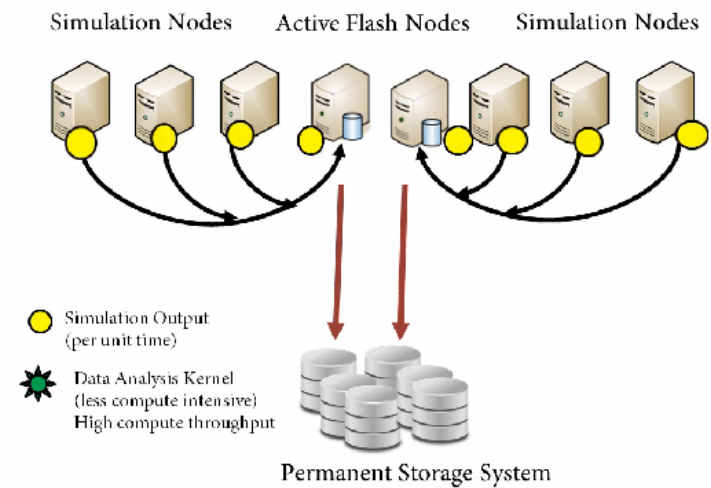
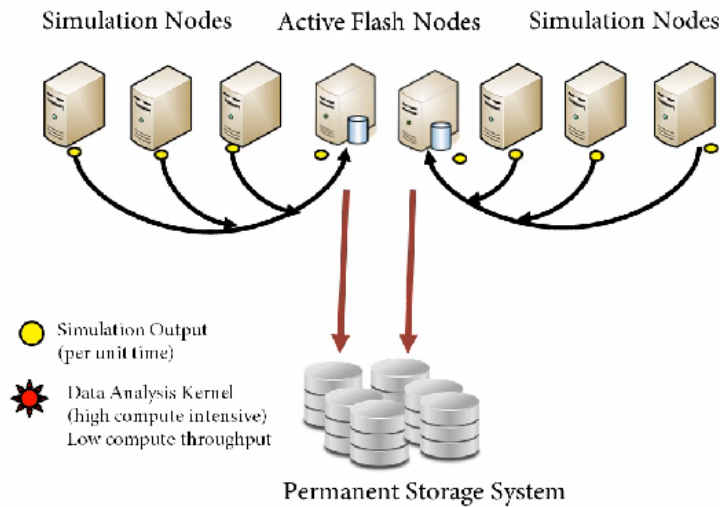
Simulation Nodes



 Simulation Output
(per unit time)

 Data Analysis Kernel
(less compute intensive)
High compute throughput

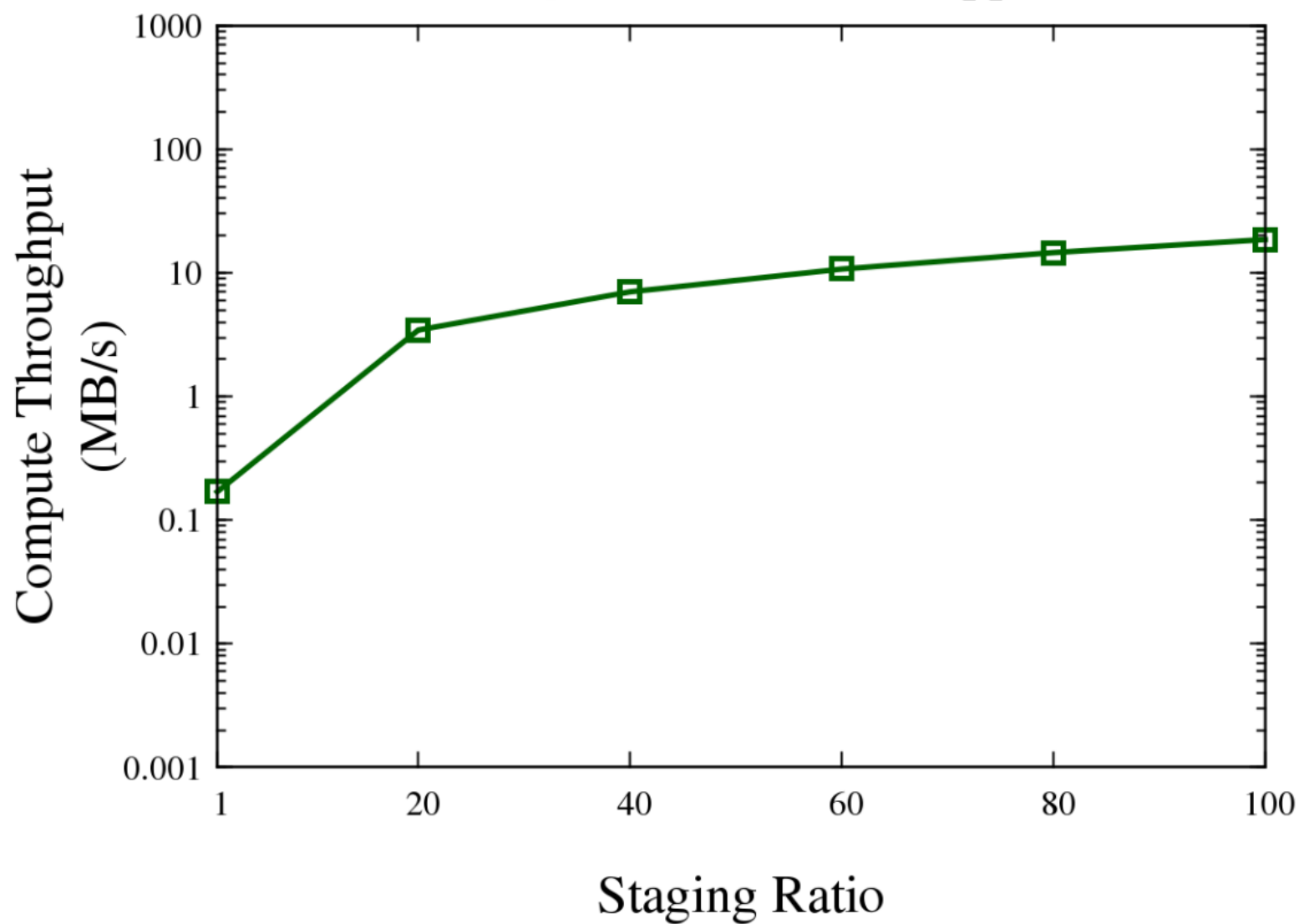
Permanent Storage System



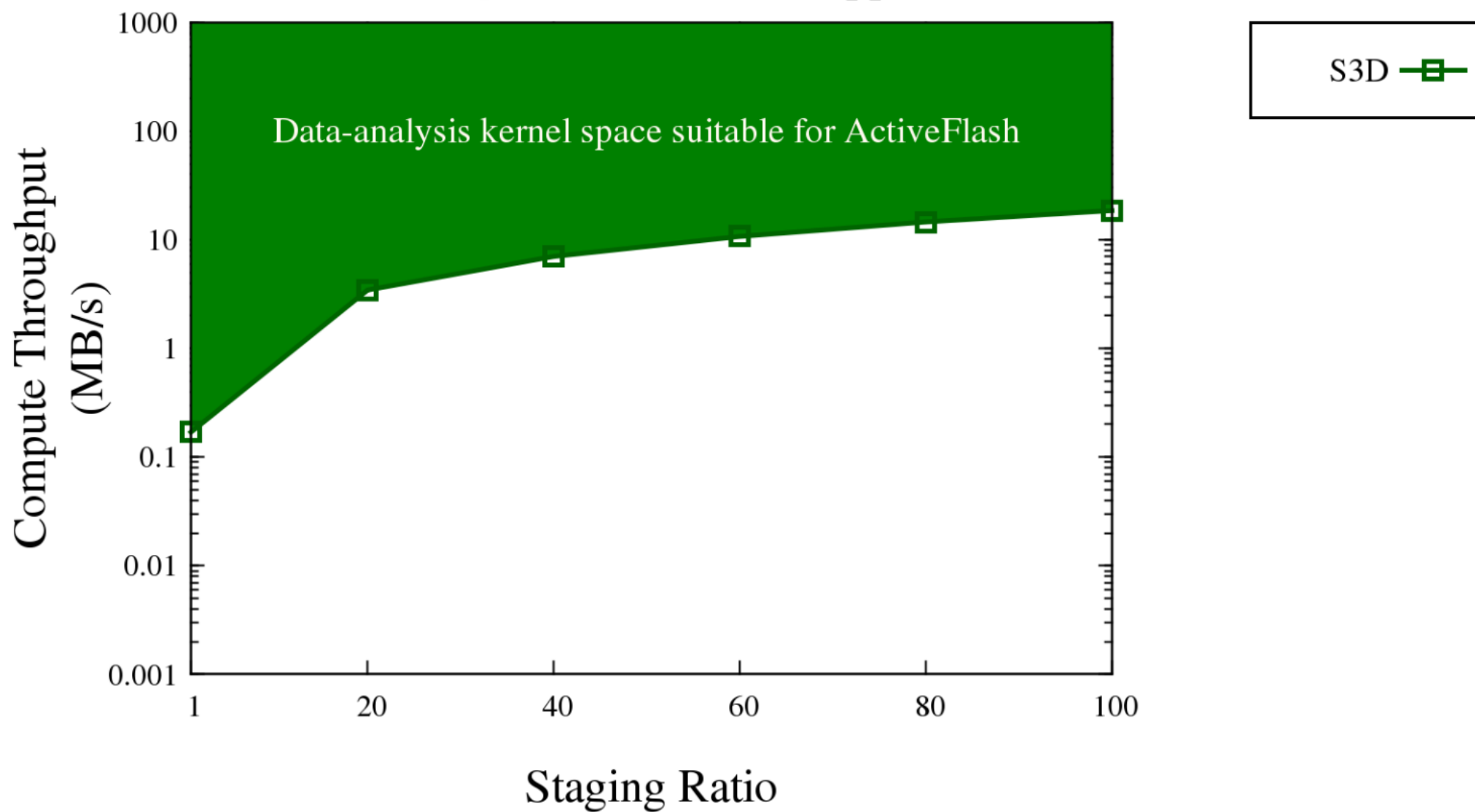
An analysis kernel needs to meet a "threshold compute throughput" to be placed on SSD controllers

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

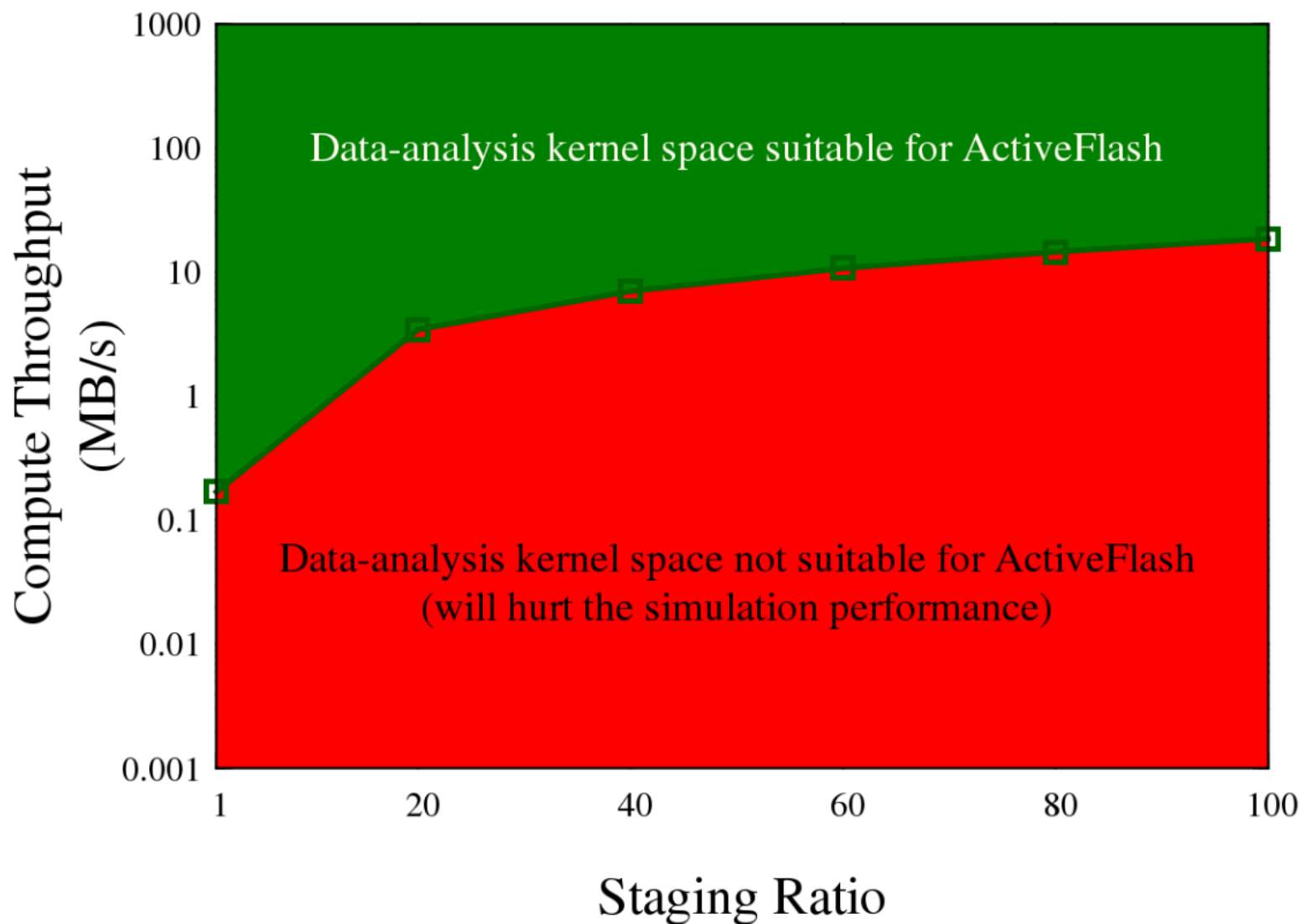
Feasibility of Active Flash Approach



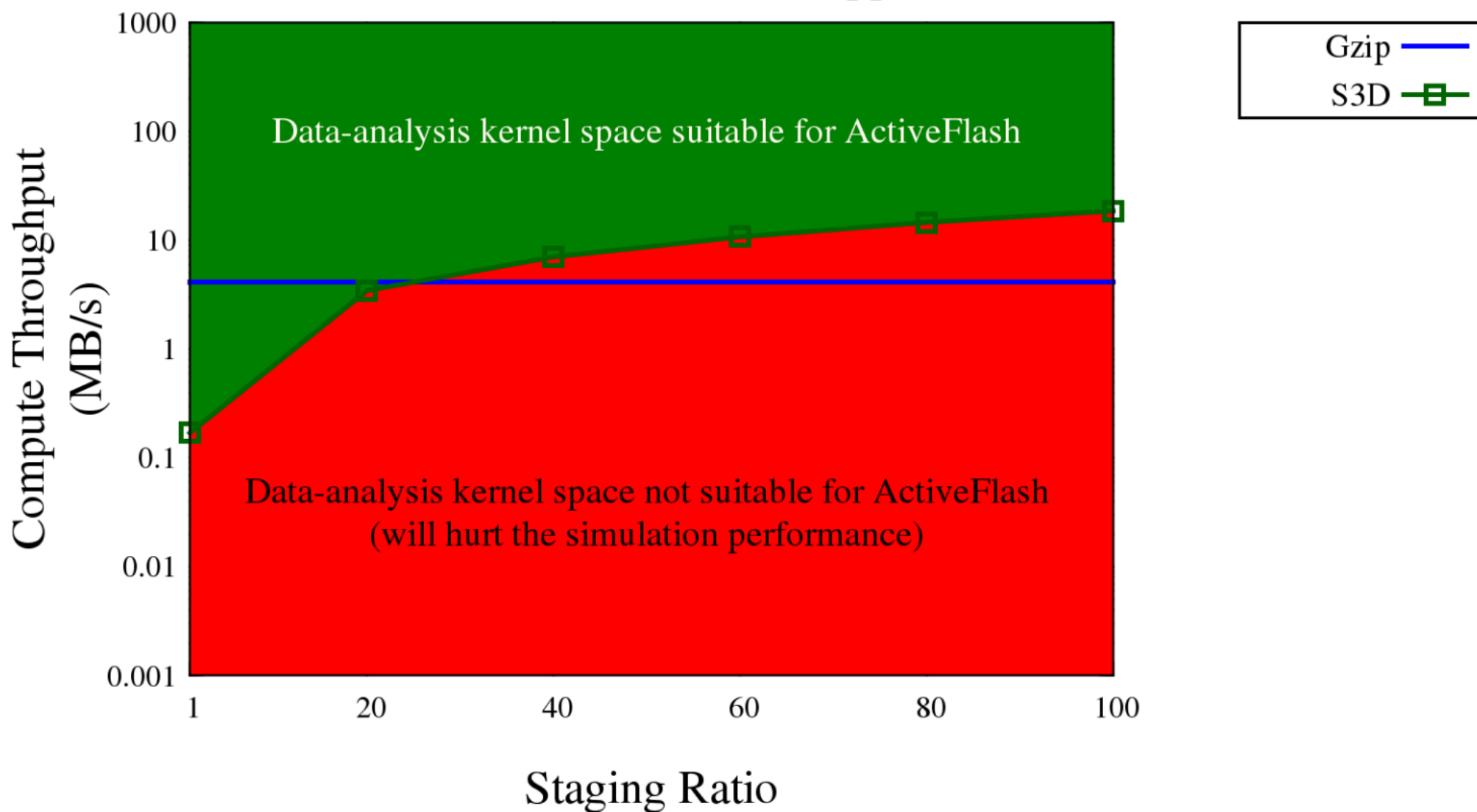
Feasibility of Active Flash Approach



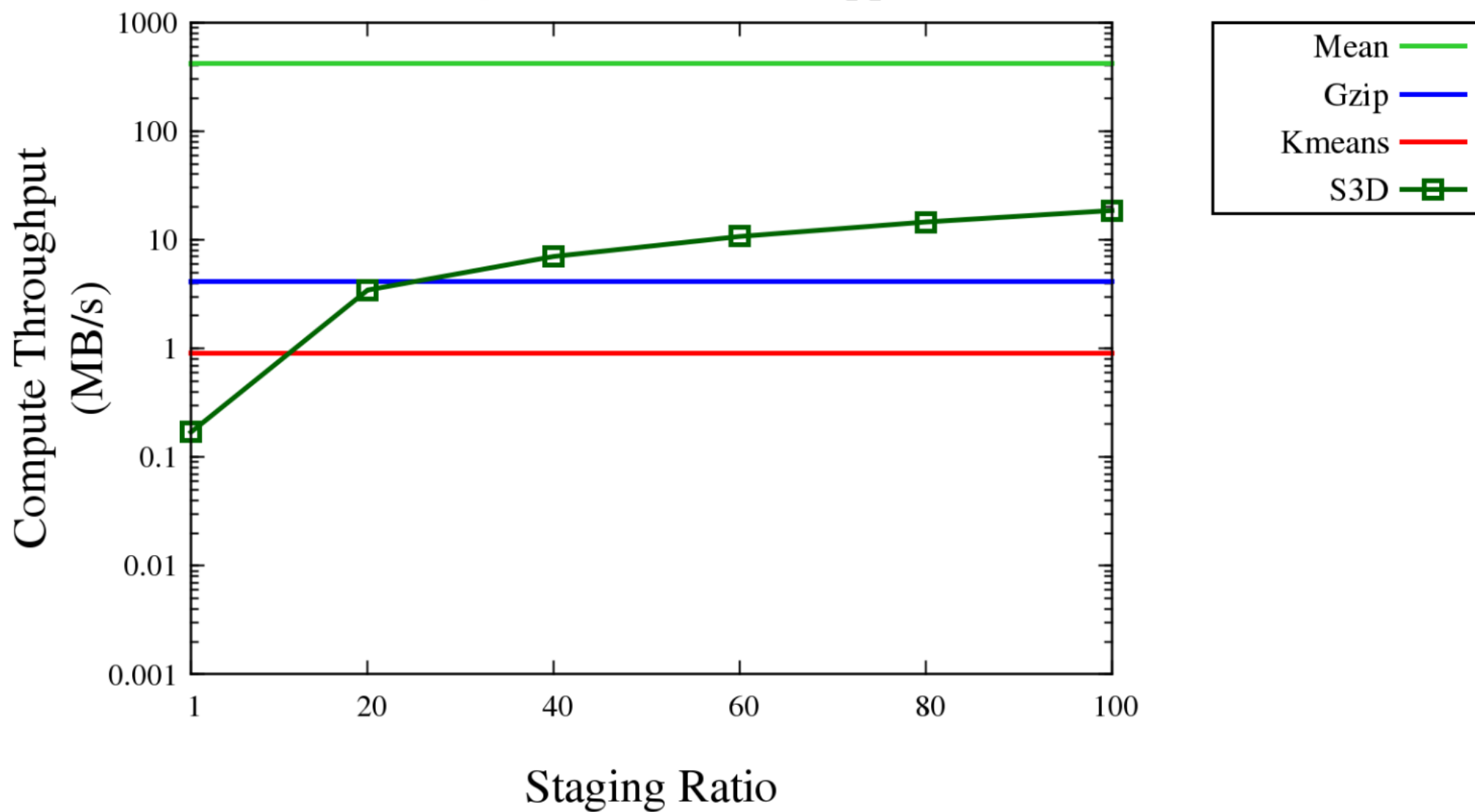
Feasibility of Active Flash Approach



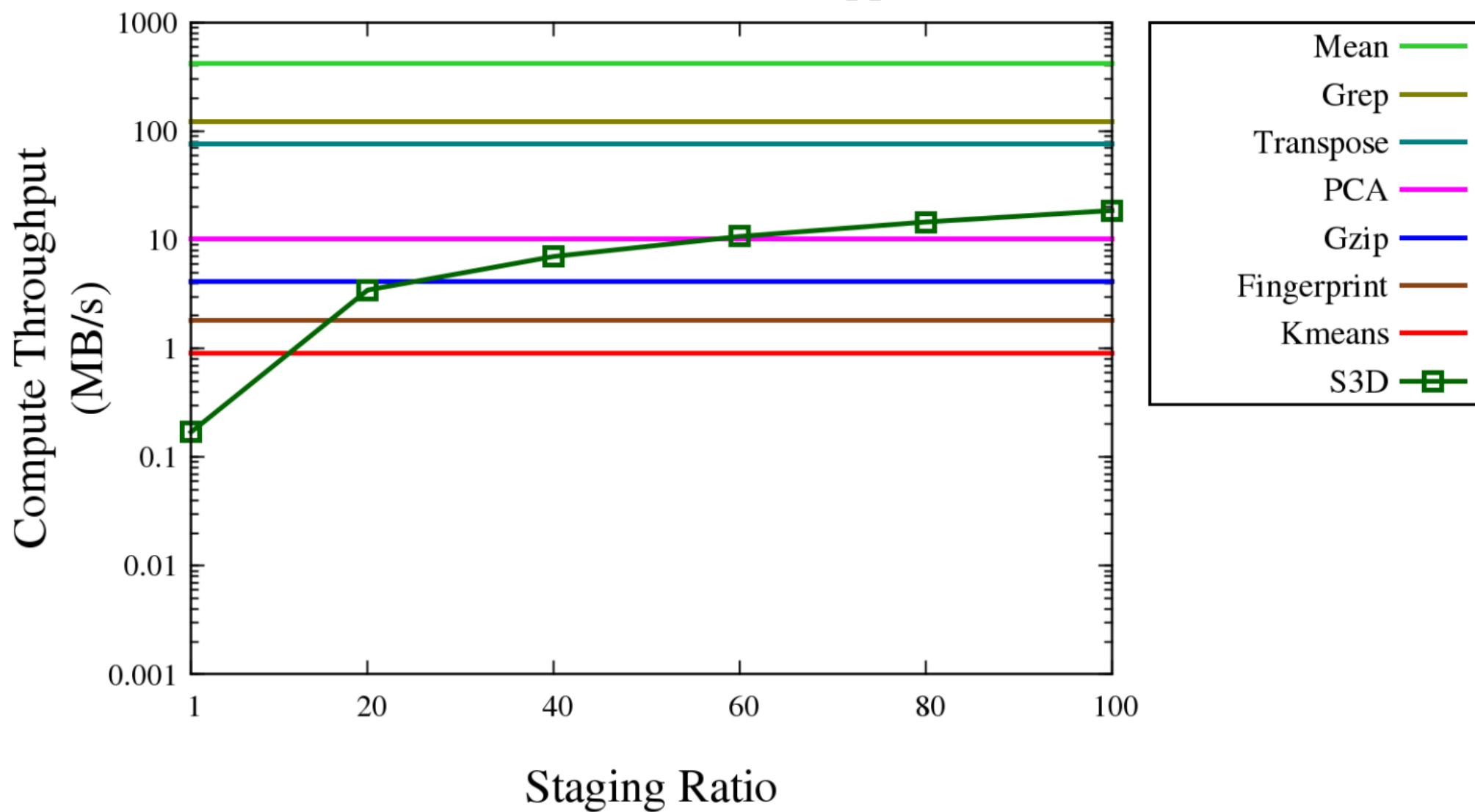
Feasibility of Active Flash Approach



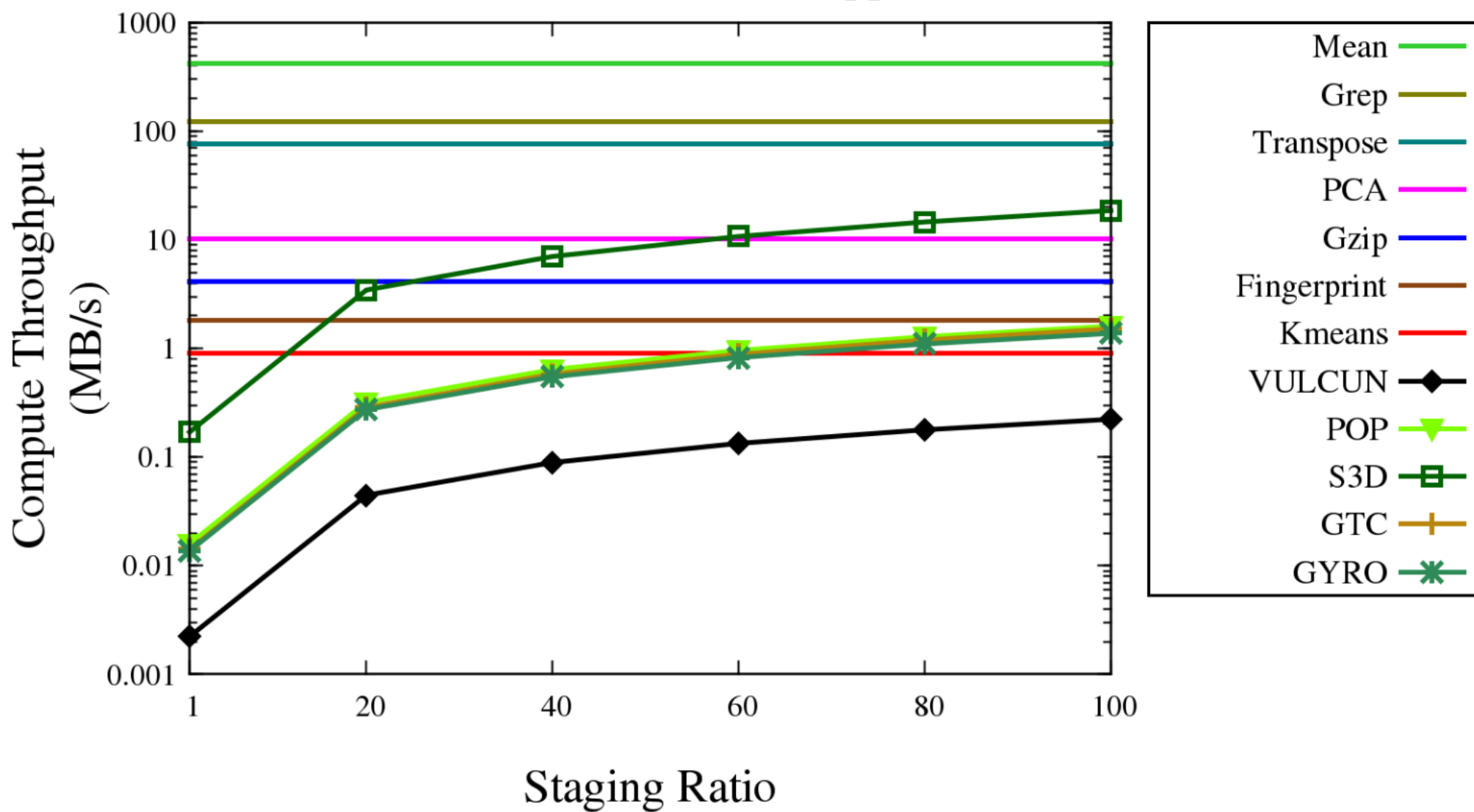
Feasibility of Active Flash Approach



Feasibility of Active Flash Approach



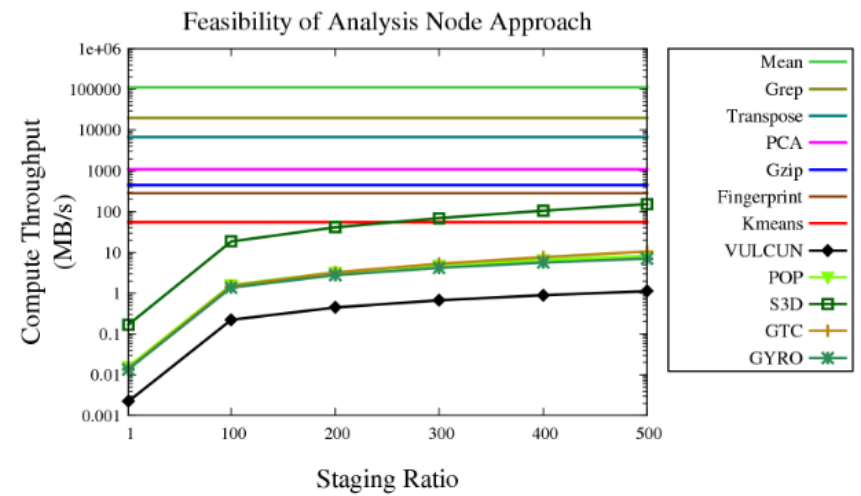
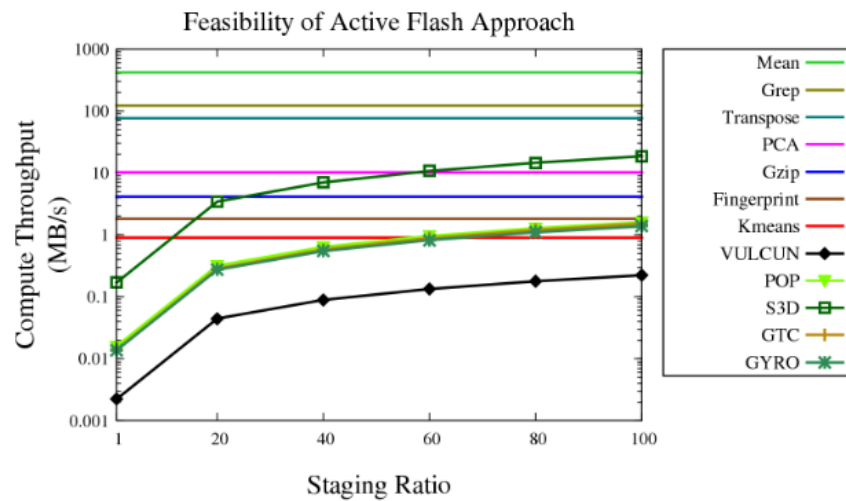
Feasibility of Active Flash Approach



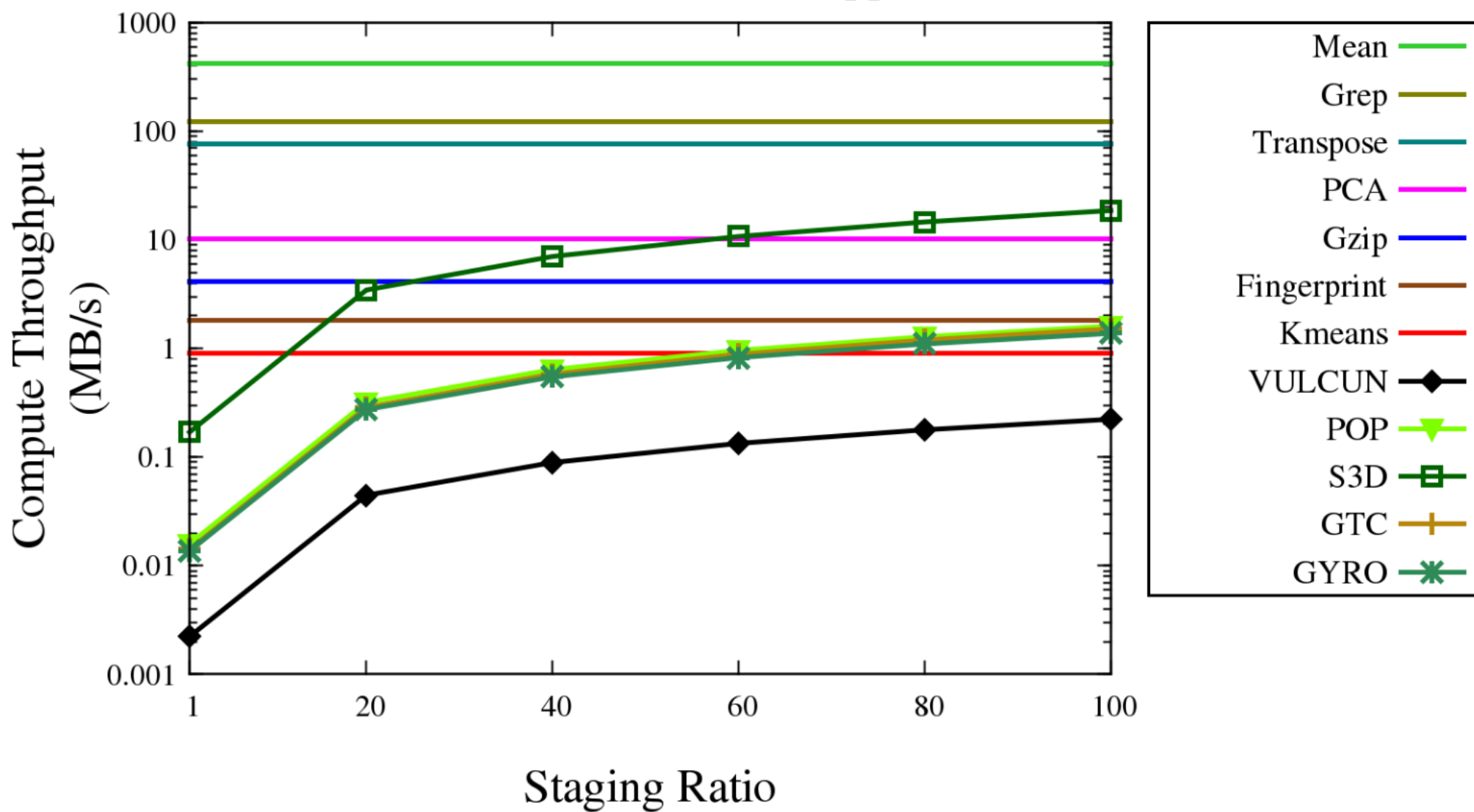
Finding: Most data analysis kernels can be placed on SSD controllers without degrading simulation performance

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

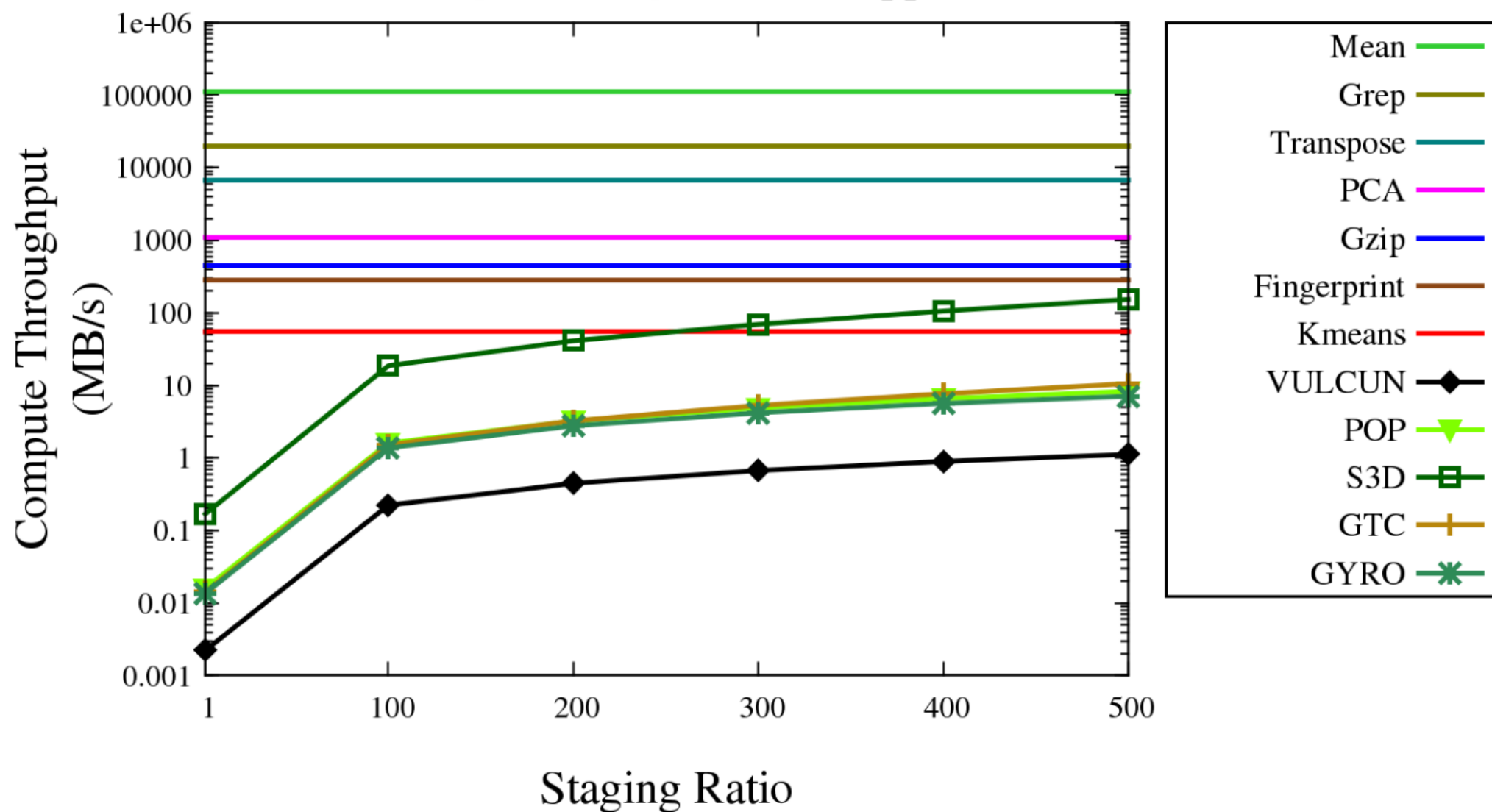
Feasibility of the Analysis Node Approach



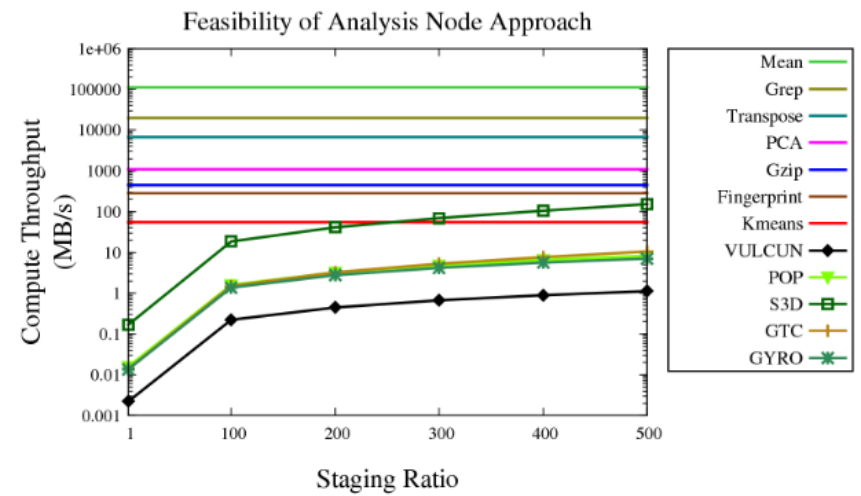
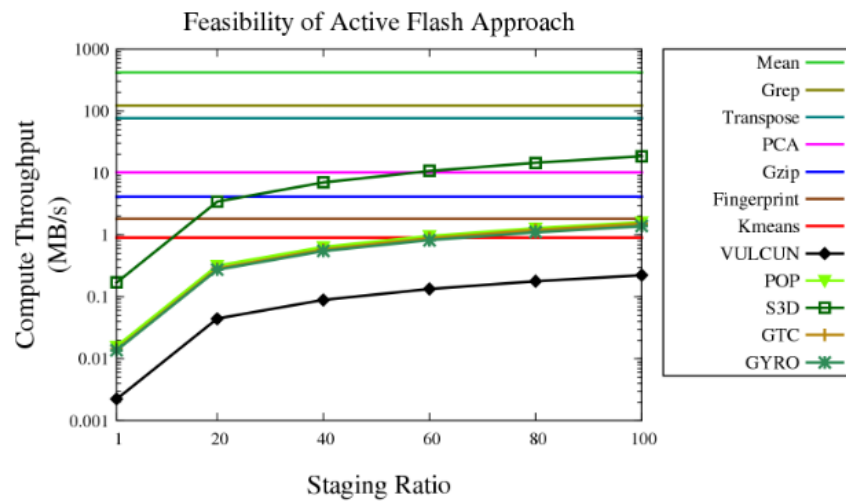
Feasibility of Active Flash Approach



Feasibility of Analysis Node Approach



Feasibility of the Analysis Node Approach



Finding: Analysis node approach is feasible at higher staging ratios, but at additional infrastructure cost (see paper)

Active Computation Feasibility

Modeling SSD Deployment **without** Active Computation Support

Multiple constraints:

Capacity

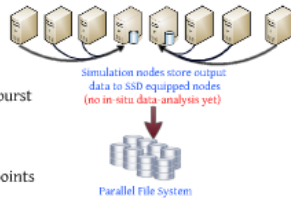
- Enough SSDs to sustain one output burst

Performance

- High I/O bandwidth to SSD space
- Fast restart from application checkpoints

Write durability

- SSD write endurance limits



Modeling Active Computation Feasibility

Simulation Applications

CHIMERA
VULCAN
POP
S3D
GTC
GYRO



Data Analysis Kernels

Statistics Collection
PCA
Grep
Gzip
Fingerprinting
Clustering



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

$$T_{SSD,k} > \frac{\lambda_k \cdot R_{SSD}}{1 - \lambda_k \cdot R_{SSD} \cdot \left(\frac{1}{BW_{SSD}} + \frac{1}{BW_{CPU}} \right) + \frac{N \cdot (k \cdot \lambda_k + k_c)}{BW_{CPU}}} = \frac{\lambda_k}{k_{thr}}$$

Relatively less compute intensive kernels better suited (e.g. regex matching) for active computation

Less computation intensive → high compute throughput

Dependent on multiple factors: simulation data production rate, staging ratio, I/O bandwidth, etc.

Staging Ratio

How many simulation nodes share one common SSD?



Staging ratio determined by the most restrictive constraint

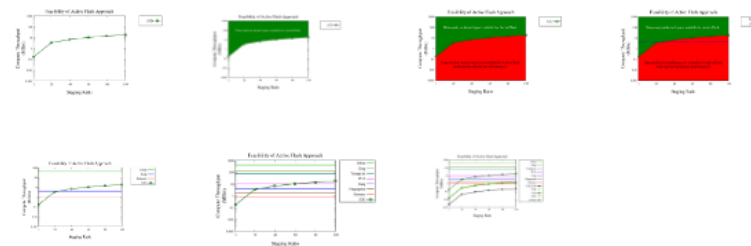
$$R_{SSD} = \min(R_{capacity}, R_{bandwidth}, R_{restart}, R_{endurance})$$

$$R_{capacity} = \frac{C_{SSD}}{f_{app} \cdot (n + BW_{SSD} \cdot \tau) \cdot \lambda_{thr}} \quad R_{bandwidth} = \frac{BW_{SSD} - SSD}{BW_{CPU}} \quad R_{restart} = \frac{f_{restart} \cdot BW_{SSD} - SSD}{\lambda_k} \quad R_{endurance} = \frac{W_{endurance}}{(\lambda_k + k_c) \cdot E_{thr}}$$

Modeled Jaguar Supercomputer consists of 18000 nodes
Staging ratio of 10 means 1800 SSDs

Active Flash Model						
	CHIMERA	VULCAN	POP	S3D	GTC	GYRO
$R_{capacity}(32GB)$	1	2571	233	18	6	166
$R_{capacity}(64GB)$	1	4500	461	36	12	333
$R_{bandwidth}$	29	29	29	29	29	29
$R_{endurance}$	1	2268	245	20	10	204
$R_{restart}$	4	896218	4054	240	42	1758

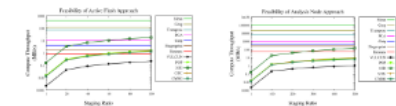
Staging ratio 10 seems to work well for all applications except CHIMERA



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

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

Feasibility of the Analysis Node Approach



Finding: Analysis node approach is feasible at higher staging ratios, but at additional infrastructure cost (see paper)

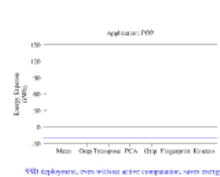
Energy and Cost Saving Analysis

"Active Flash" Energy Modeling

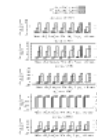
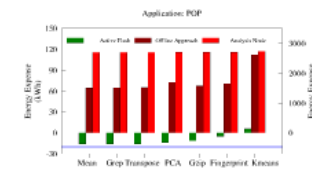
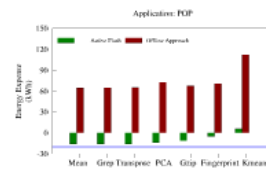
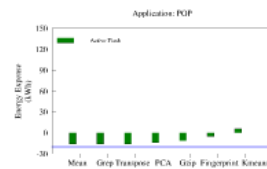
Modeled after Samsung PM830 SSD

Total energy consists of multiple components

SSD energy during I/O, compute, and idle periods
Data movement energy cost in the interconnect



Y00 employees, even without active computation, save energy



"Offline" and "Analysis Node" Approach Energy Modeling

Modeled after Inter Core i7 processors

Assumed idle when not doing data analysis

Optimistic modeling
cooling, assembling and installation costs ignored

Infrastructure and Energy cost
All five applications run simultaneously for 2 years (each application for 144 times, 24 hour long simulation) using ratio of 50:100:30 for core:GPU:SSD

Scaling Ratio	Infrastructure Cost (\$)	Energy Bill (\$)	Total Cost (\$)	Feasible Applications
10	180,000	15,131	195,131	all
25 & 300	—	—	—	none
Analysis Node Model				
10	1,014,000	500,375	1,514,375	all
30	806,000	138,193	944,193	all, w/o GTC
300	80,000	31,893	111,893	all, w/o GTC, SSD

Finding: Active Flash is more energy and cost efficient than other approaches in many cases

"Active Flash" Energy Modeling

Modeled after Samsung PM830 SSD

Total energy consists of multiple components

SSD energy during I/O, compute, and idle periods
Data movement energy cost in the interconnect

"Offline" and "Analysis Node" Approach Energy Modeling

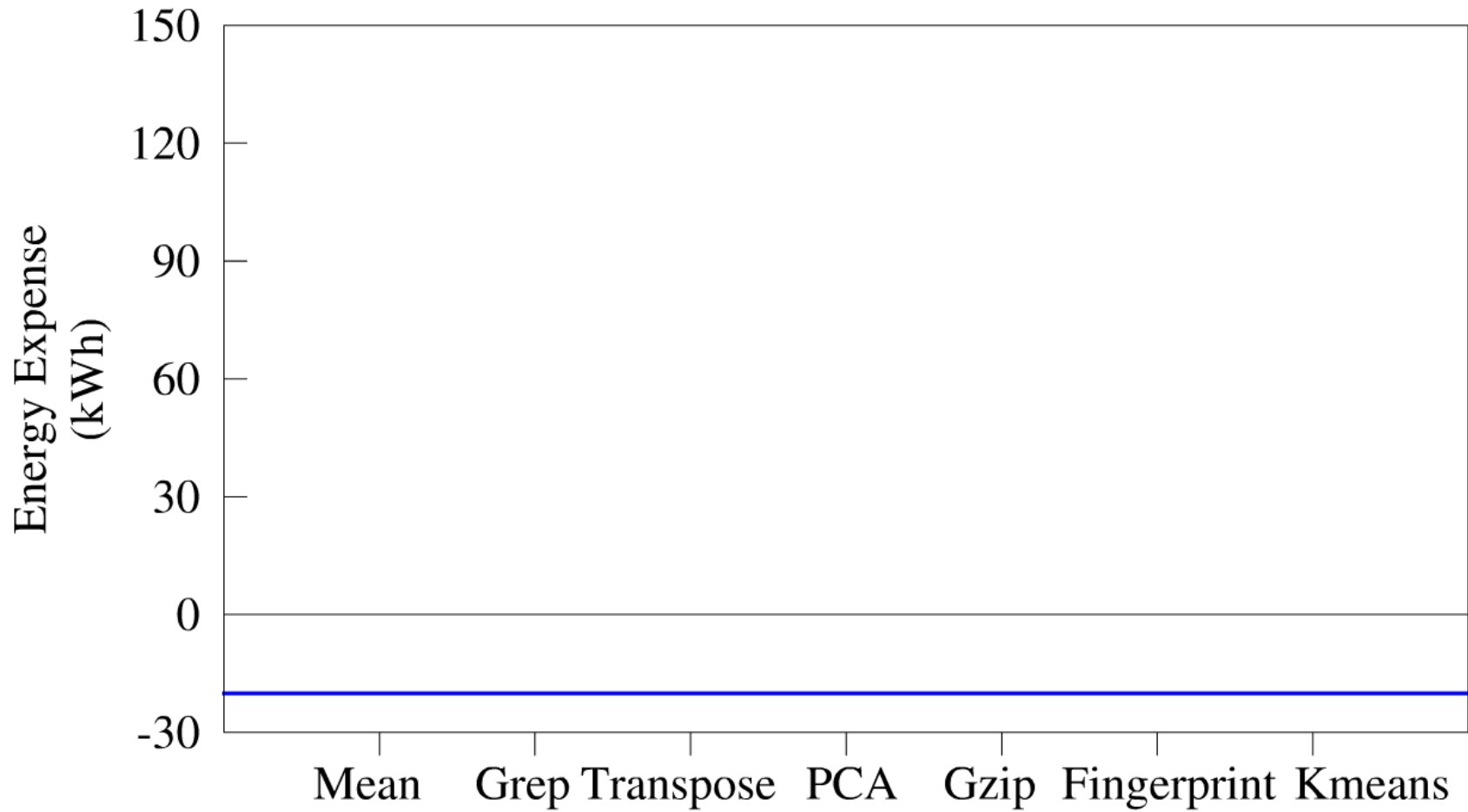
Modeled after Inter Core i7 processors

Assumed idle when not doing data analysis

Optimistic modeling

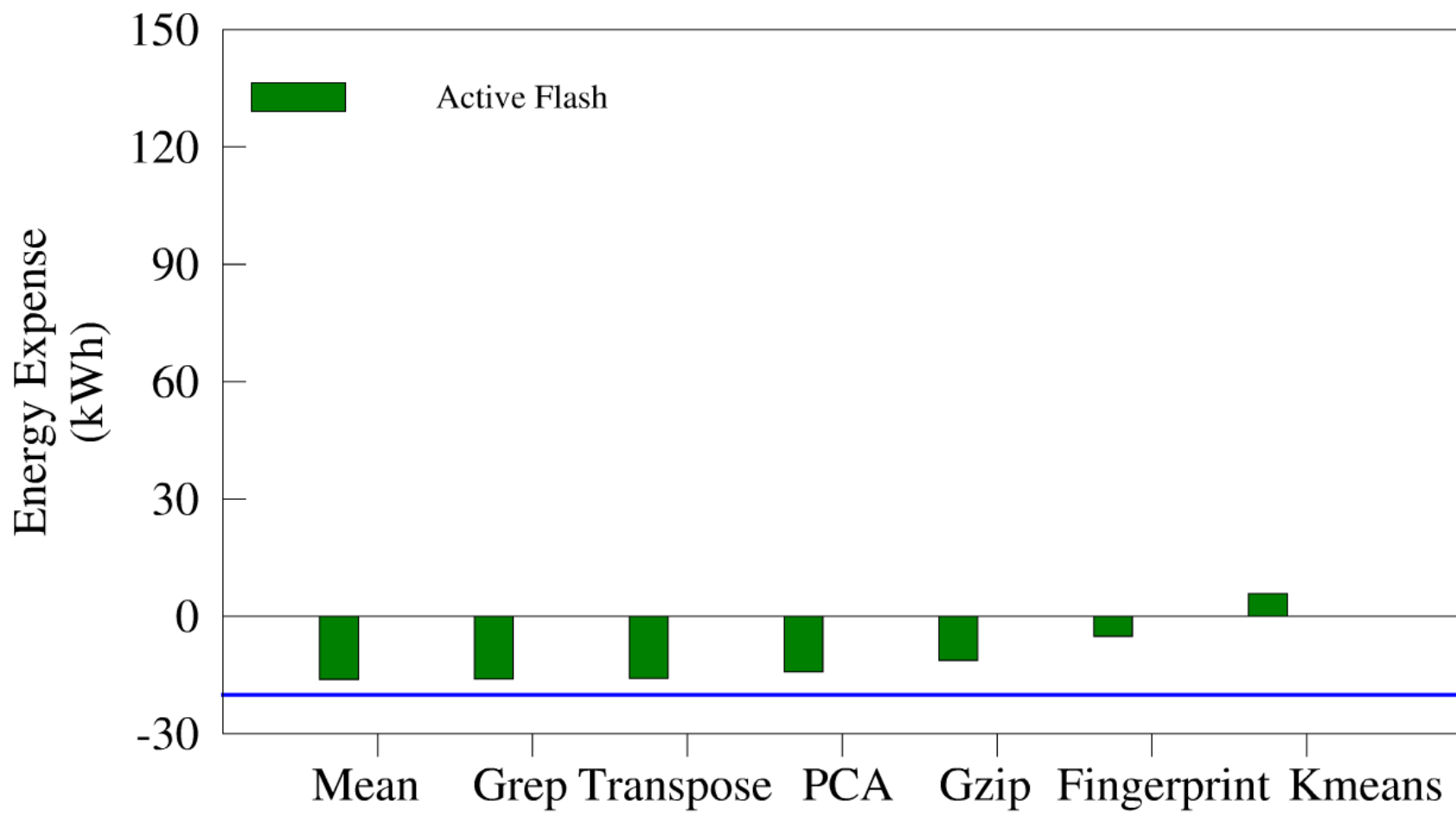
cooling, assembling and installation costs ignored

Application: POP

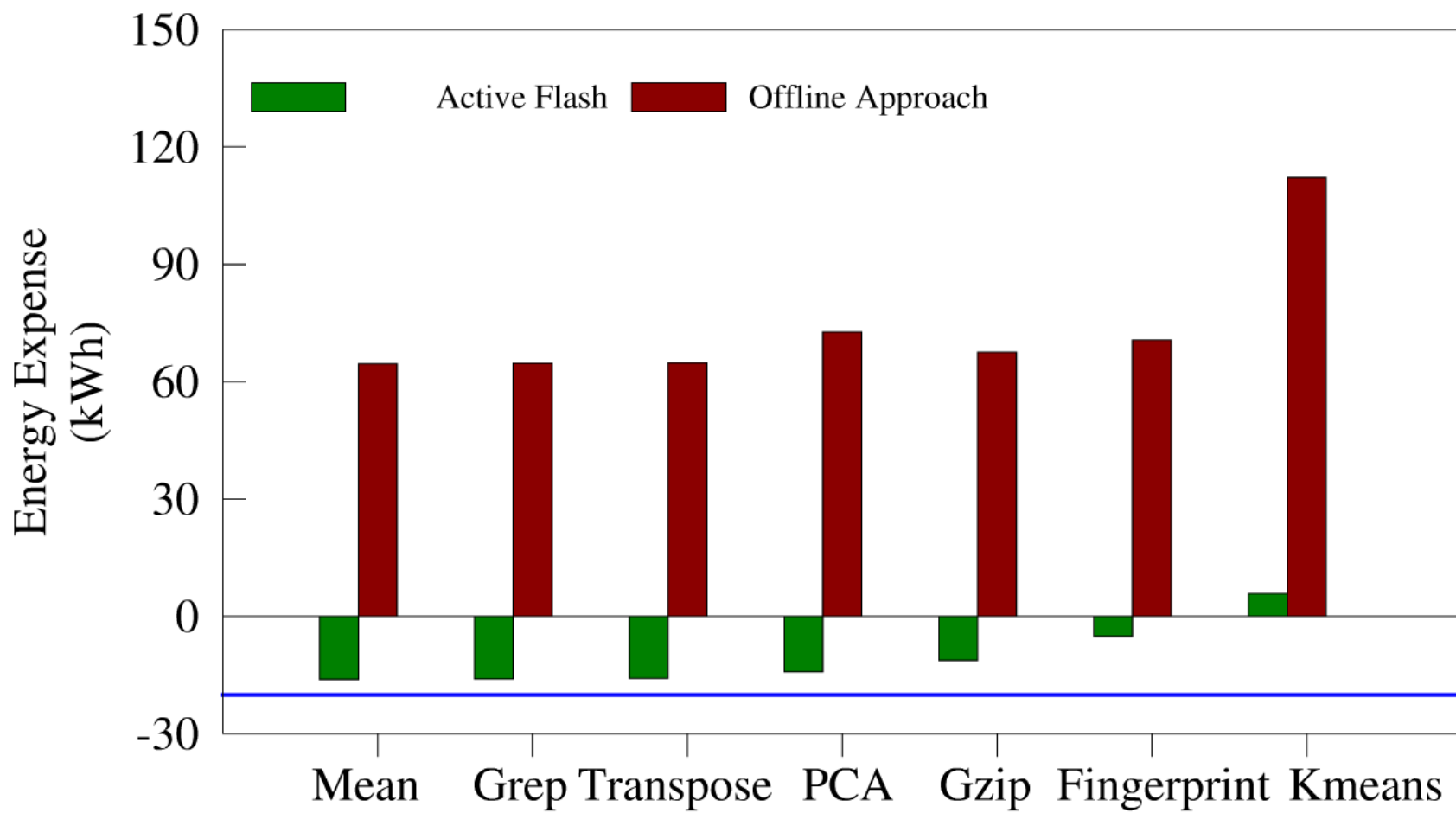


SSD deployment, even without active computation, saves energy

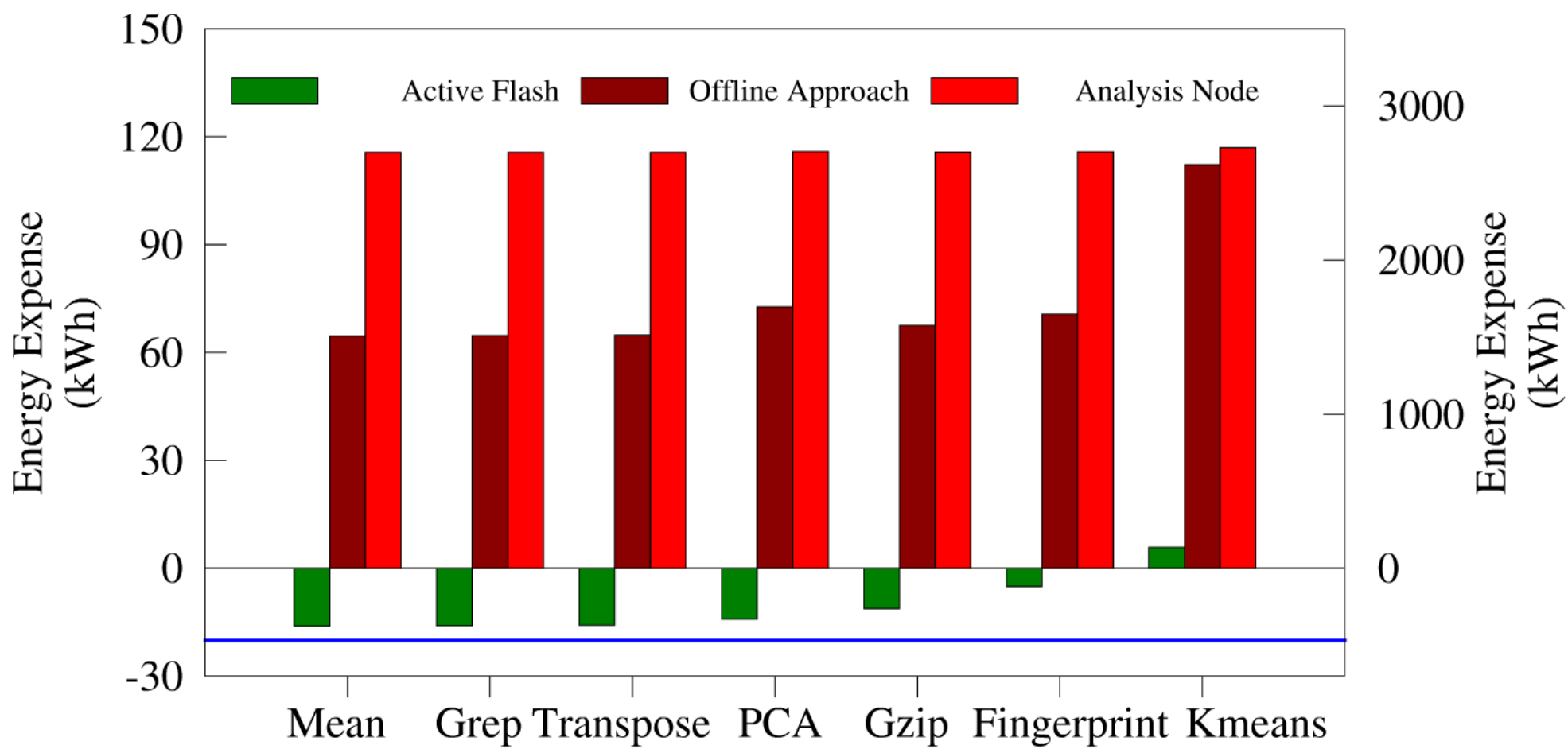
Application: POP






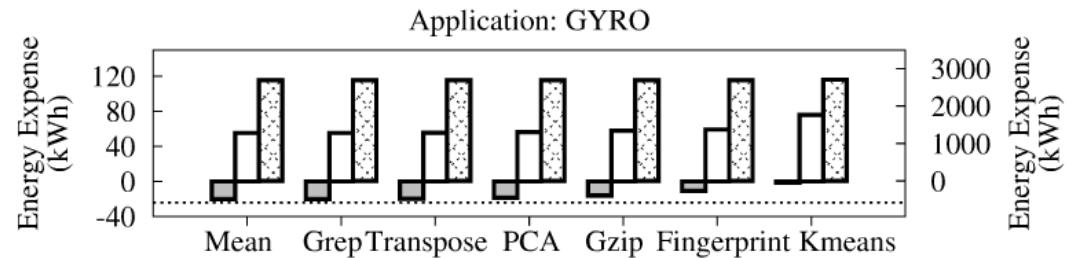
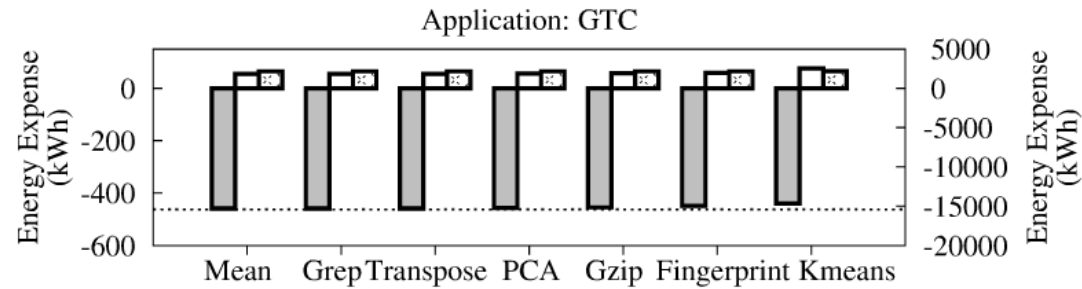
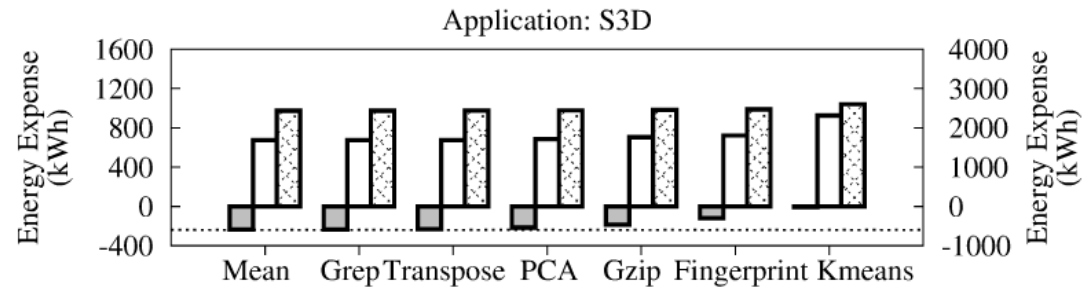
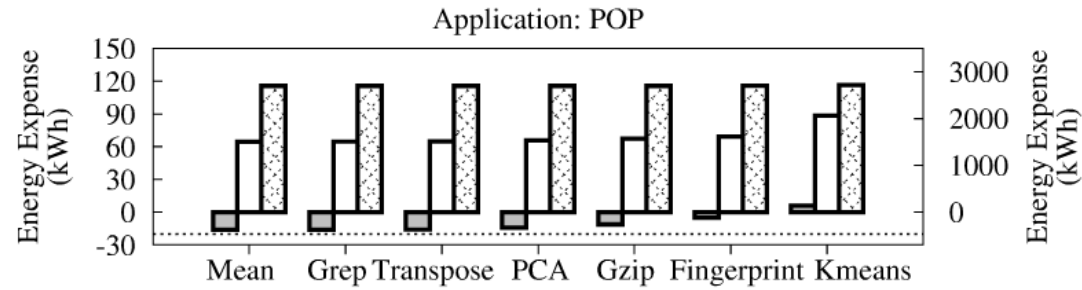
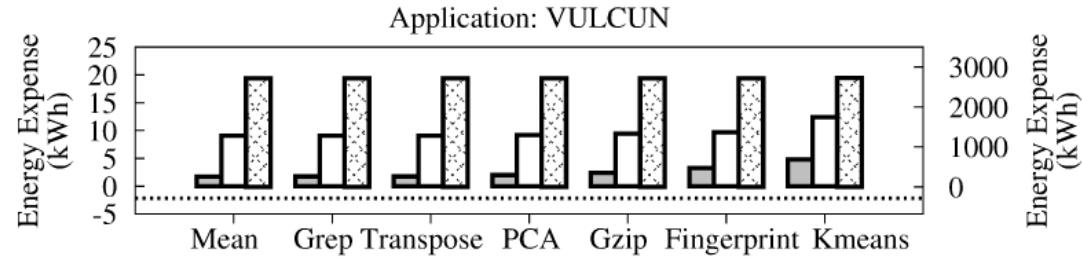
Application: POP



Application: POP



Active Flash (y1 axis) 
 Offline Approach (y1 axis) 
 Analysis Node (y2 axis) 



Infrastructure and Energy cost:

All five applications run continuously for 2 years (each application for 146 times, 24 hour long simulation time) Staging ratio of 10: 1800 SSDs in our 18000 node system

Staging Ratio	Infrastructure Cost (\$)	Energy Bill (\$)	Total Cost (\$)	Feasible Applications
Active Flash Model				
10	180,000	−19,131	160,866	all
30 & 300	—	—	—	none
Analysis Node Model				
10	1,818,000	566,375	2,384,375	all
30	606,000	158,193	642,993	all, w/o GTC
300	60,600	31,072	67,432	all, w/o GTC, S3D

Finding: Active Flash is more energy and cost efficient than other approaches in many cases

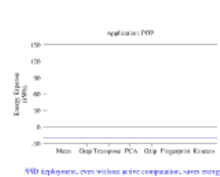
Energy and Cost Saving Analysis

"Active Flash" Energy Modeling

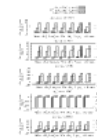
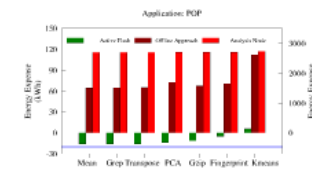
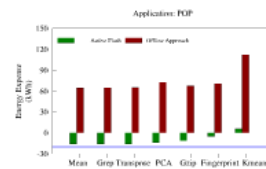
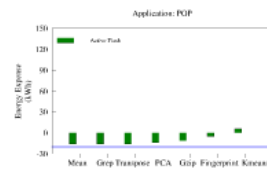
Modeled after Samsung PM830 SSD

Total energy consists of multiple components

SSD energy during I/O, compute, and idle periods
Data movement energy cost in the interconnect



Y00 employees, even without active computation, save energy



"Offline" and "Analysis Node" Approach Energy Modeling

Modeled after Inter Core i7 processors

Assumed idle when not doing data analysis

Optimistic modeling
cooling, assembling and installation costs ignored

Infrastructure and Energy cost
All five applications run simultaneously for 2 years (each application for 144 times, 24 hour long simulation) using ratio of 50:100:30 for core:storage:network

Scaling Ratio	Infrastructure Cost (\$)	Energy Bill (\$)	Total Cost (\$)	Feasible Applications
10	180,000	15,131	195,131	all
25 & 300	—	—	—	none
Analysis Node Model				
10	1,014,000	500,375	1,514,375	all
30	806,000	138,193	944,193	all, w/o GTC
300	80,000	31,893	111,893	all, w/o GTC, \$10

Finding: Active Flash is more energy and cost efficient than other approaches in many cases

ActiveFlash Prototype based on OpenSSD Platform

Prototype demonstrates the viability of our approach

Changes only in the FTL, no hardware changes

Preemption based scheduling

See paper for the details and evaluation results

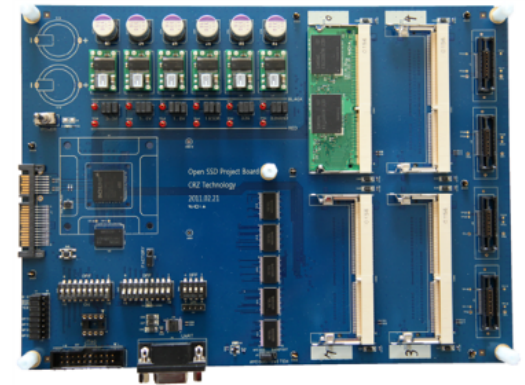
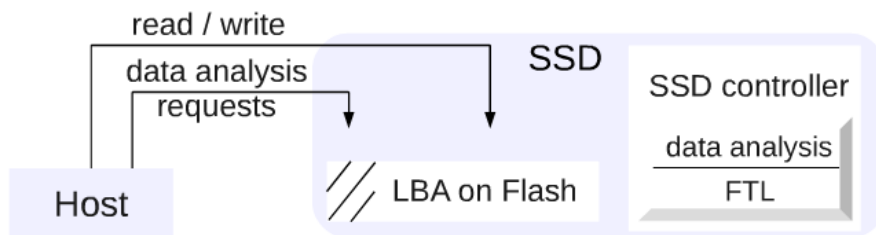


Figure courtesy: open-ssd project



Conclusion

Active computation on SSDs enables energy-efficient in-situ data-analysis in Supercomputing

In most cases, Active Flash does not require extra SSDs

Active Flash may even help cut SSD deployment cost by reducing electricity bill

Active Flash for scientific data analytics viable with OpenSSD

Thank You!



Background



Energy and Cost Saving Analysis

"Active Flash" Energy Modeling

Modeled after Samsung PM981 SSD

Total energy consists of multiple components

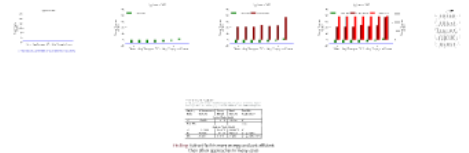
SSD energy during I/O, compute, and idle periods
Data movement energy cost (in the unaccessed)

"Offline" and "Analysis Node" Approach Energy Modeling

Modeled after Intel Core i7 processors

Assumed idle when not doing data analysis

Cryptic modeling
cooling, assembly and installation costs ignored



Active Flash: Towards Energy-Efficient, In-Situ Data Analytics on Extreme-Scale Machines



Problems and Challenges

Offline approach to data analysis involves multiple rounds of I/O, causing

- Excessive data movement
- Excessive energy cost

"Unpractical for data movement of thousands of bytes to be of the same order of computation cost, if not more"

Using simulation nodes for data analysis not acceptable

- High CPU allocation cost on a Supercomputer

Active Flash Approach for In-situ Scientific Data Analysis



Active Computation Feasibility

Modeling SSD Deployment without Active Computation Support

Multiple constraints

- Capacity
- Throughput
- High bandwidth to the host
- Low latency from application to host

Staging ratio

How many parallel write channels are common SSD?

Staging ratio determined by the host interface constraints

Staging ratio 10 seems to work well for all applications except CHIMERA

Modeling active computation feasibility

Modeling active computation feasibility

Modeling active computation feasibility

Modeling Active Computation Feasibility

Modeling active computation feasibility

Modeling active computation feasibility

Modeling active computation feasibility

Modeling active computation feasibility

Modeling active computation feasibility

Modeling active computation feasibility

Modeling active computation feasibility

Conclusion

Active computation on SSDs enables energy-efficient in-situ data analysis in Supercomputing

In most cases, Active Flash does not require extra SSDs

Active Flash may even help cut SSD deployment cost by reducing electricity bill

Active Flash for scientific data analytics viable with OpenSSD

Thank You!

ActiveFlash Prototype based on OpenSSD Platform

Prototype demonstrates the viability of our approach

Changes only in the FTL, no hardware changes

Preemption based scheduling

See paper for the details and evaluation results



Figure courtesy: open-ssd project

