ILLINOIS INSTITUTE OF TECHNOLOGY

HPIS3: Towards a High-Performance Simulator for Hybrid Parallel I/O and Storage Systems

Bo Feng, Ning Liu, Shuibing He, Xian-He Sun

Department of Computer Science

Illinois Institute of Technology, Chicago, IL

Email: {bfeng5, nliu8}@hawk.iit.edu, {she11, sun}@iit.edu

ILLINOIS INSTITUTE OF TECHNOLOGY

Outline

- Introduction
- Related Work
- Design and Implementation
- Experiments
- Conclusions and Future Work

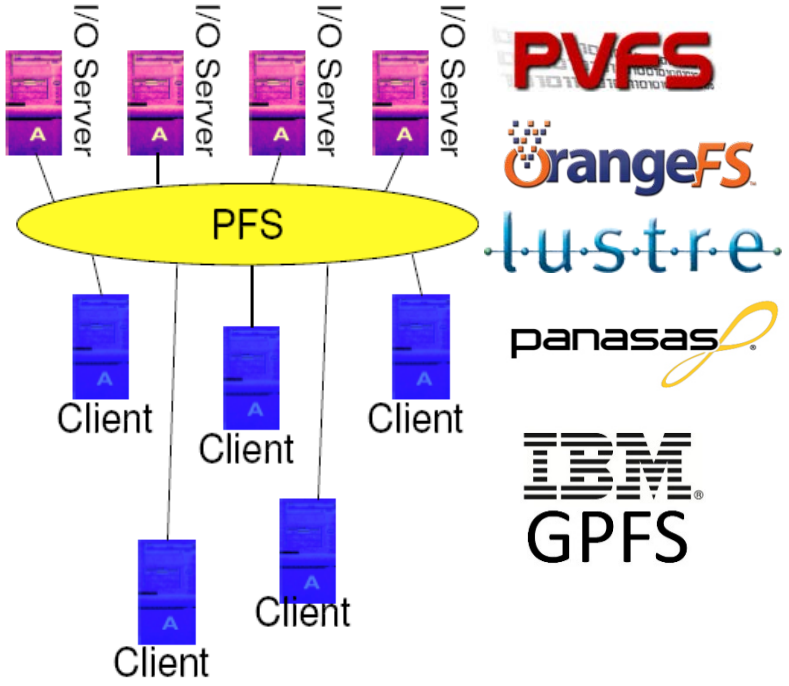
ILLINOIS INSTITUTE OF TECHNOLOGY

Outline

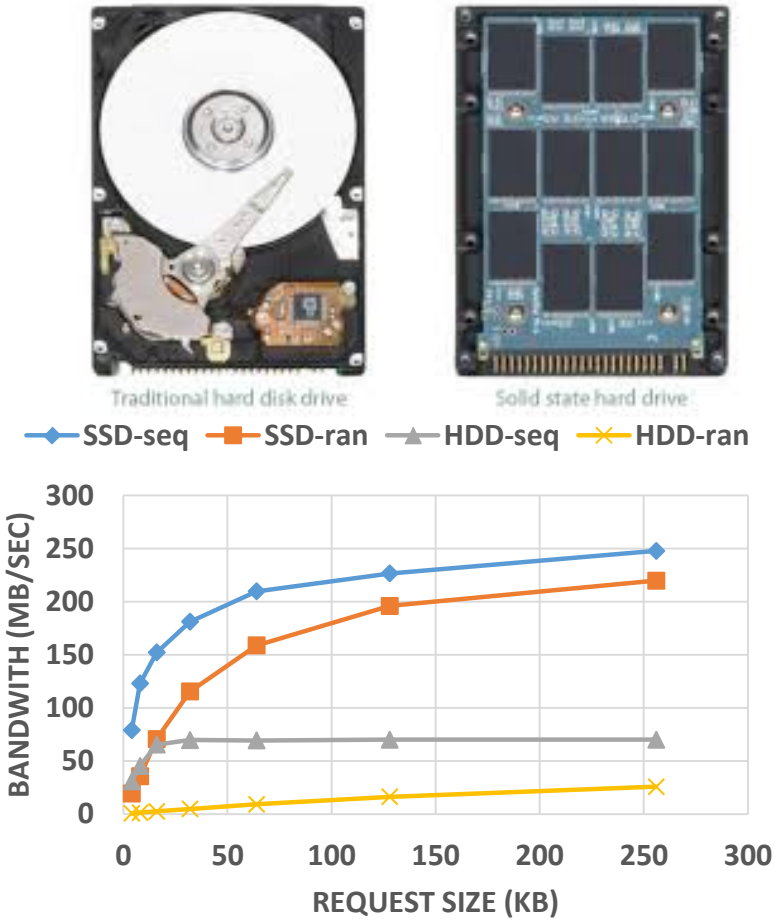
- Introduction
- Related Work
- Design and Implementation
- Experiments
- Conclusions and Future Work

To Meet the High I/O Demands

1. PFS



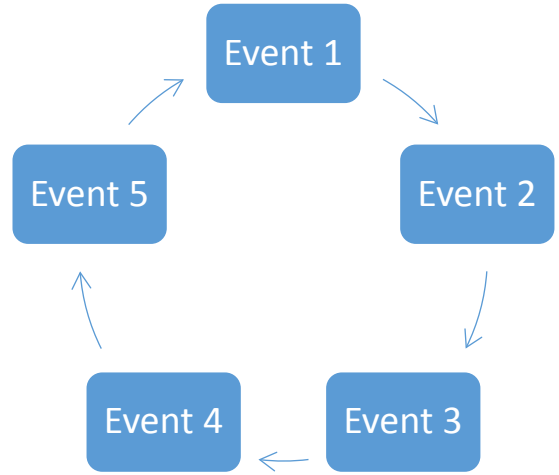
2. SSD



HPIS3: Hybrid Parallel I/O and Storage System Simulator

ILLINOIS INSTITUTE OF TECHNOLOGY

- Parallel discrete event simulator
- A variety of hardware and software configurations
- Hybrid settings
  - Buffered-SSD
  - Tiered-SSD
  - …
- HDD and SSD latency and bandwidth under parallel file systems
- Efficient and high-performance



ILLINOIS INSTITUTE OF TECHNOLOGY

Outline

- Introduction
- Related Work
- Design and Implementation
- Experiments
- Conclusions and Future Work

ILLINOIS INSTITUTE OF TECHNOLOGY

Related Work

Co-design tool for hybrid parallel I/O and storage systems

- S4D-Cache: Smart Selective SSD Cache for Parallel I/O Systems [1]
- A Cost-Aware Region-Level Data Placement Scheme For Hybrid Parallel I/O Systems [2]
- On the Role of Burst Buffers in Leadership-Class Storage Systems [3]
- iBridge: Improving Unaligned Parallel File Access with Solid-State Drives [4]
- More…

[1] S. He, X.-H. Sun, and B. Feng, "S4D-Cache: Smart Selective SSD Cache for Parallel I/O Systems," in *Proceedings of International Conference on Distributed Computing Systems (ICDCS)*, 2014.
[2] S. He, X.-H. Sun, B. Feng, X. Huang, and K. Feng, "A Cost-Aware Region-Level Data Placement Scheme for Hybrid Parallel I/O Systems," in *Proceedings of 2013 IEEE International Conference on Cluster Computing (CLUSTER)*, 2013.
[3] N. Liu, J. Cope, P. Carns, C. Carothers, R. Ross, G. Grider, A. Crume, and C. Maltzahn, "On the Role of Burst Buffers in Leadership-Class Storage Systems," in *Proceedings of 2012 IEEE 28th Symposium on Mass Storage Systems and Technologies (MSST)*, 2012.
[4] X. Zhang, K. Liu, K. Davis, and S. Jiang, "iBridge: Improving unaligned parallel file access with solid-state drives," in *Proceedings of the 2013 IEEE 27th International Parallel and Distributed Processing Symposium (IPDPS)*, 2013.

ILLINOIS INSTITUTE OF TECHNOLOGY

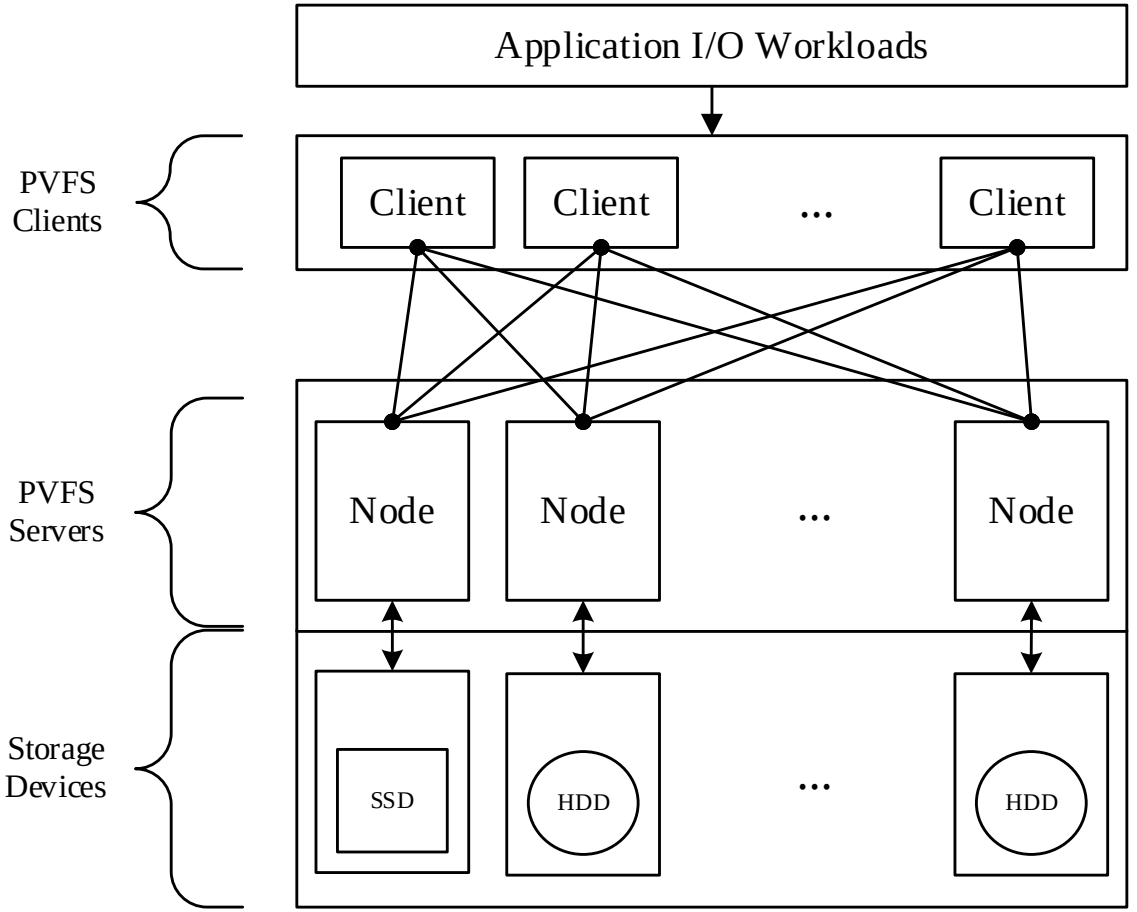
Outline

- Introduction
- Related Work
- Design and Implementation
- Experiments
- Conclusions and Future Work

ILLINOIS INSTITUTE OF TECHNOLOGY

Design Overview

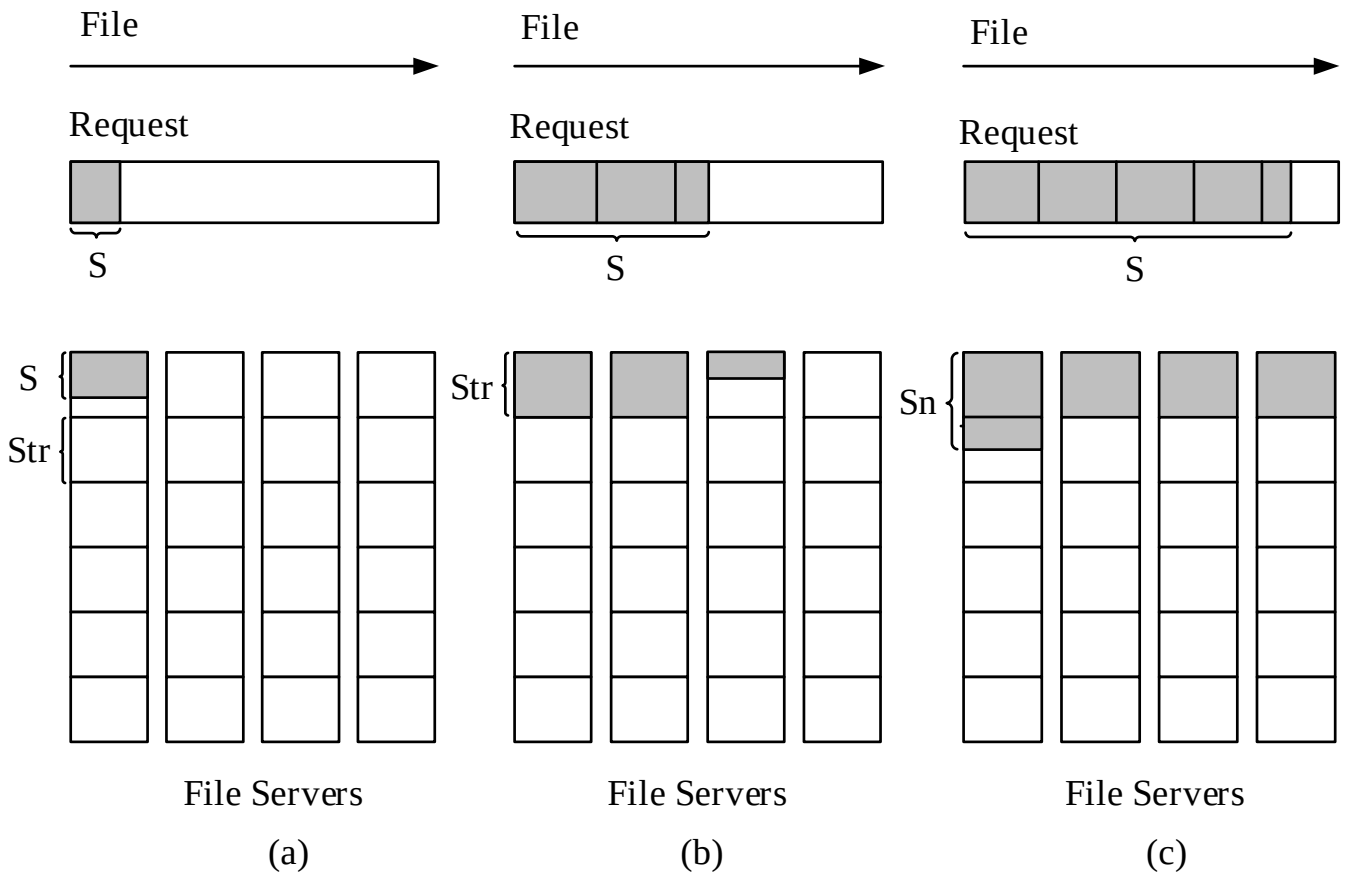
- Platform: ROSS
- Target: PVFS
- Architecture Overview
  - Client LPs
  - Server LPs
  - Drive LPs
- Note: LP is short of logical process. They act like real processes in the system and are synchronized by Time Warp protocol.



ILLINOIS INSTITUTE OF TECHNOLOGY

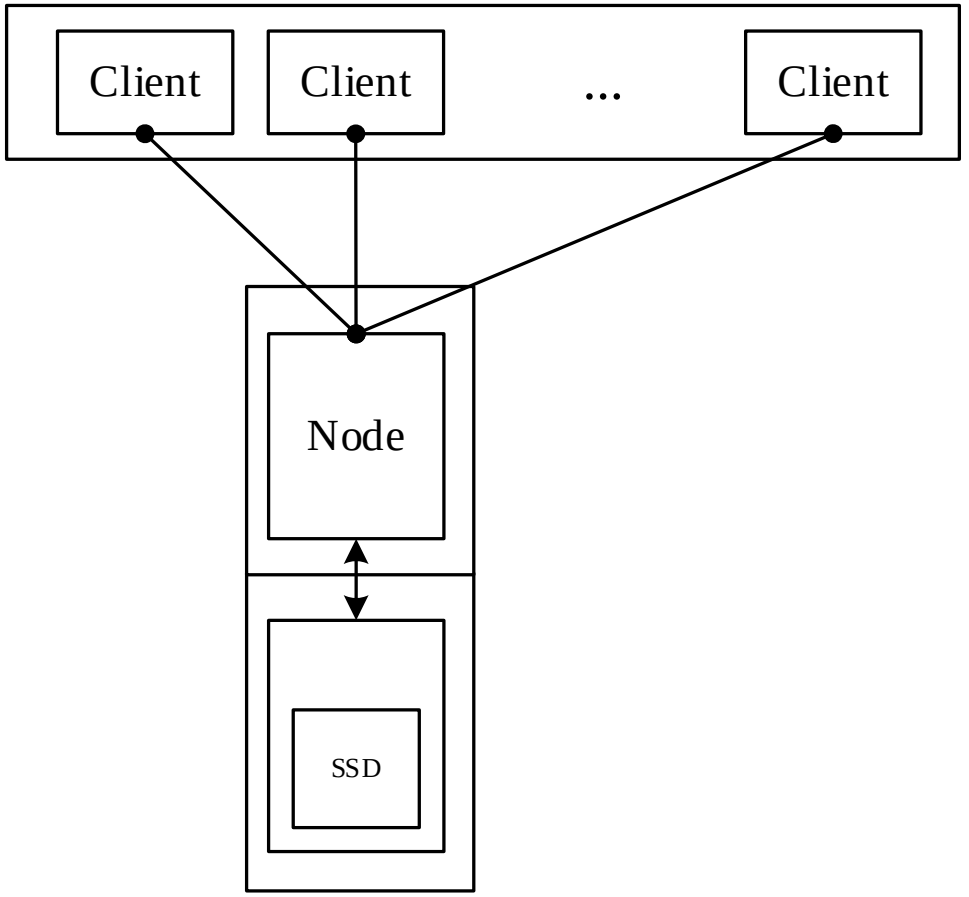
File, Queue and PVFS Client Modeling

- File requests and request queue modeling
  - <file_id, length, file_offset>
  - State variables define queues
- PVFS client modeling
  - Stripping mechanism



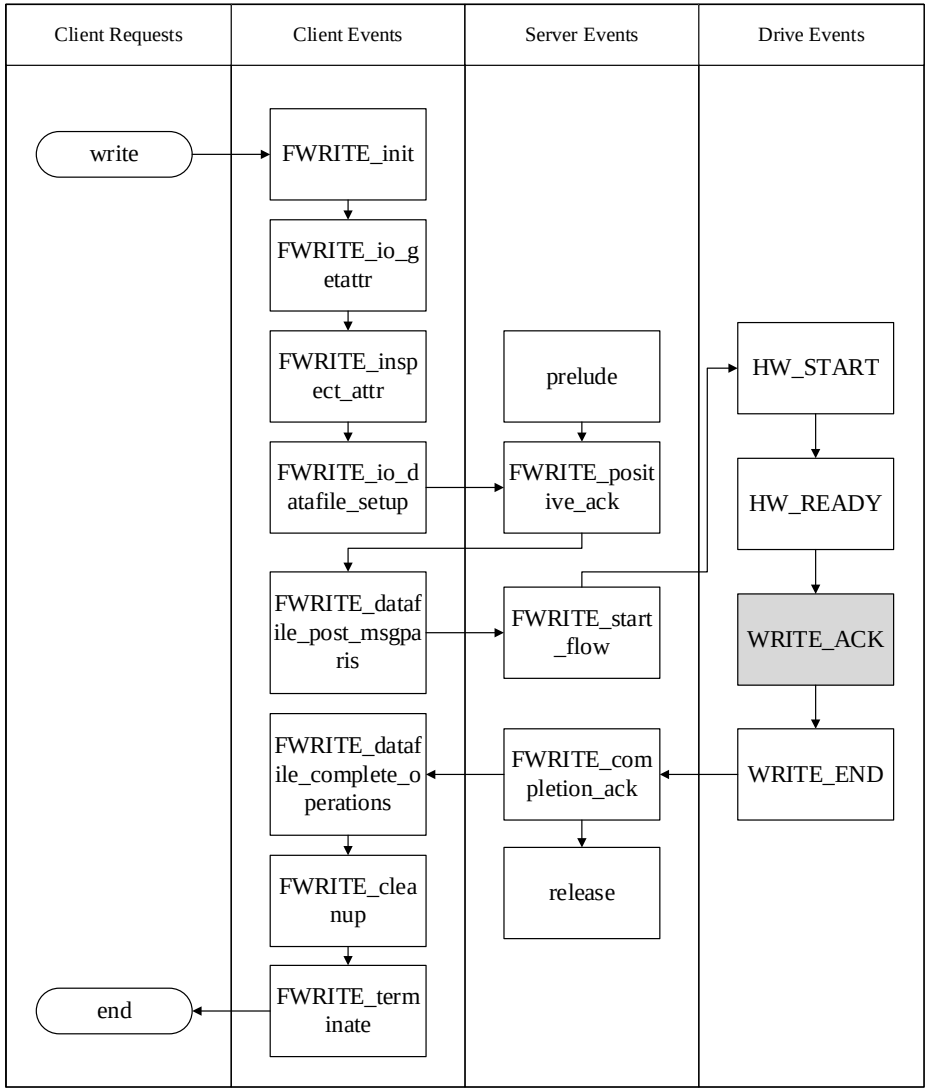
PVFS Sever Modeling

- Connected with clients and drives



Event flow in HPIS3: a write example

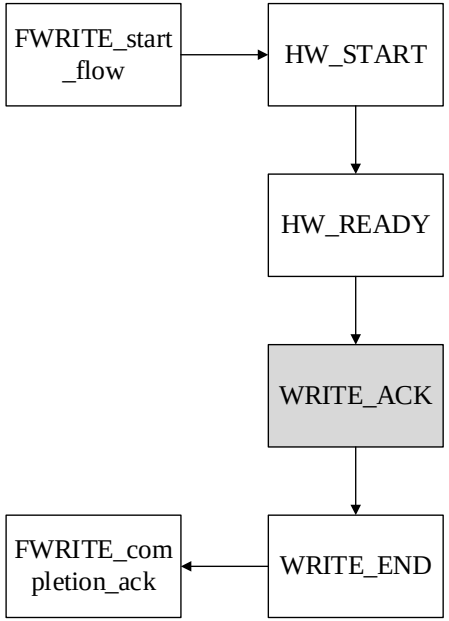
- Write event flow for HDD
  - Single-queue effect
- Write event flow for SSD
  - Multi-queue effect



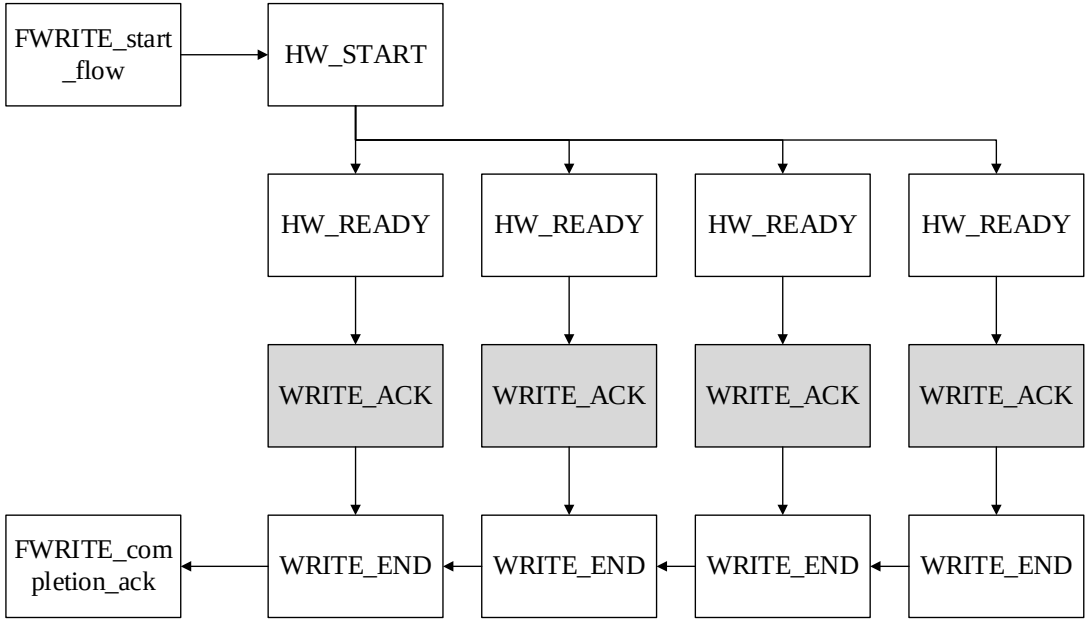
Storage Device Modeling: HDD vs. SSD (1)

ILLINOIS INSTITUTE OF TECHNOLOGY

HDD



SSD



Storage Device Modeling: HDD vs. SSD (2)

ILLINOIS INSTITUTE OF TECHNOLOGY

HDD

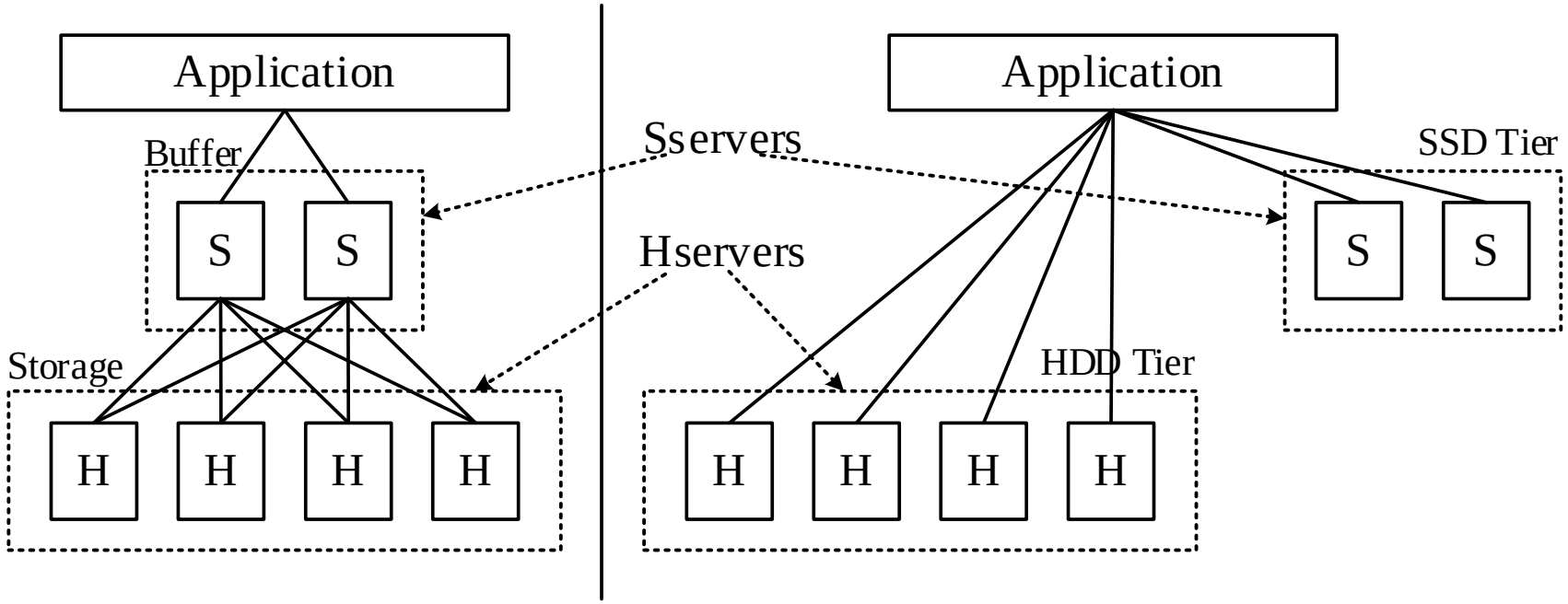
- Start up time
- Seek time
- Data transfer time

SSD

- Start up time
- FTL mapping time
- GC time

	Read	Write
Sequential	SR	SW
Random	RR	RW

Hybrid PVFS I/O and Storage Modeling



Buffered-SSD setting

Tiered-SSD setting

- S is short for SSD-Server, which is a server node with SSD.
- H is short for HDD-Server, which is a server node with HDD.

ILLINOIS INSTITUTE OF TECHNOLOGY

Outline

- Introduction
- Related Work
- Design and Implementation
- Experiments
- Conclusions and Future Work

ILLINOIS INSTITUTE OF TECHNOLOGY

Experimental setup

	65-nodes SUN Fire Linux Cluster
CPU	Quad-Core AMD Opteron(tm) Processor 2376 * 2
Memory	4 * 2GB, DDR2 333MHz
Network	1 Gbps Ethernet
Storage	HDD: Seagate SATA II 250GB, 7200RPM SSD: OCZ PCI-E X4 100GB
OS	Linux kernel 2.6.28.10
File system	OrangeFS 2.8.6

- 32 nodes used throughout our experiments in this study

ILLINOIS INSTITUTE OF TECHNOLOGY

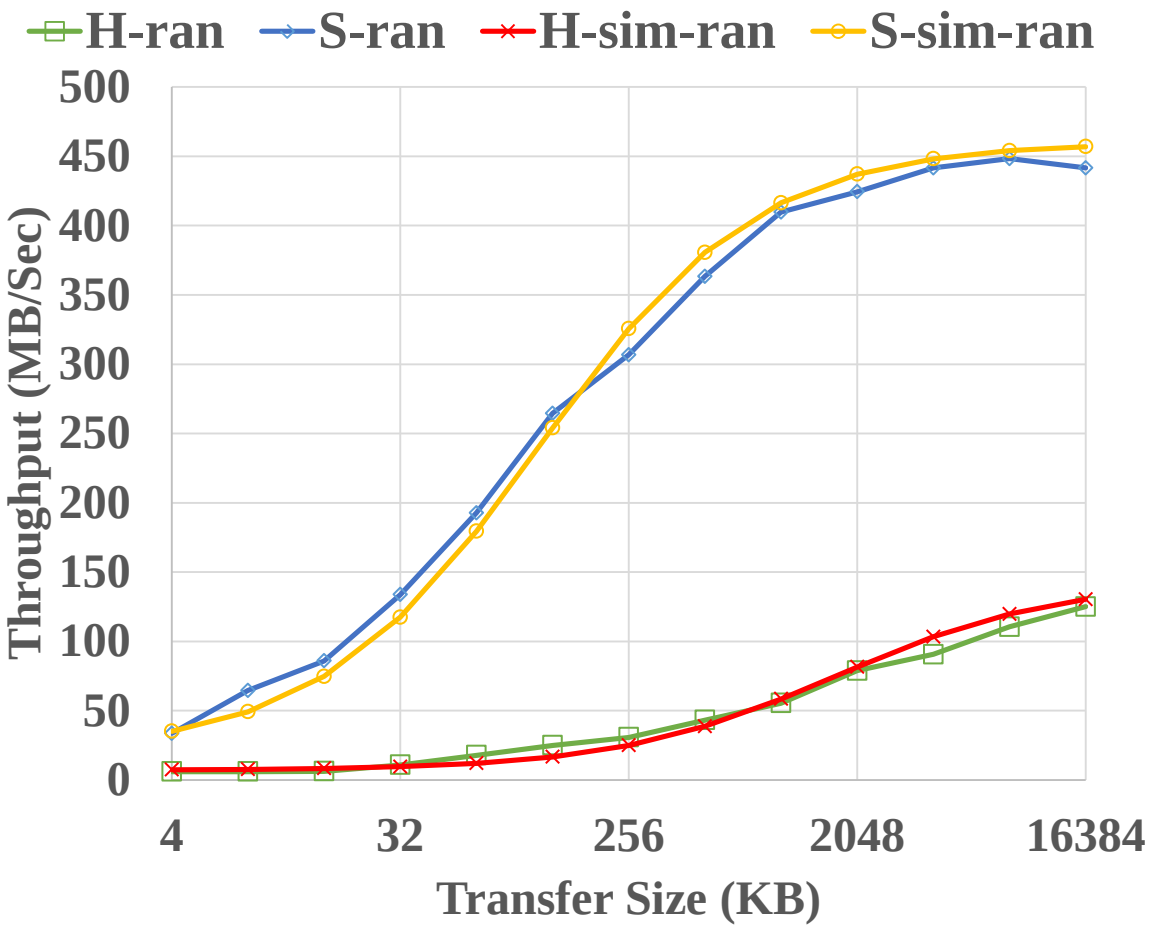
Benchmark and Trace tool

- IOR
  - Sequential read/write
  - Random read/write

- IOSIG
  - Conducted from trace replay to trigger events

ILLINOIS INSTITUTE OF TECHNOLOGY

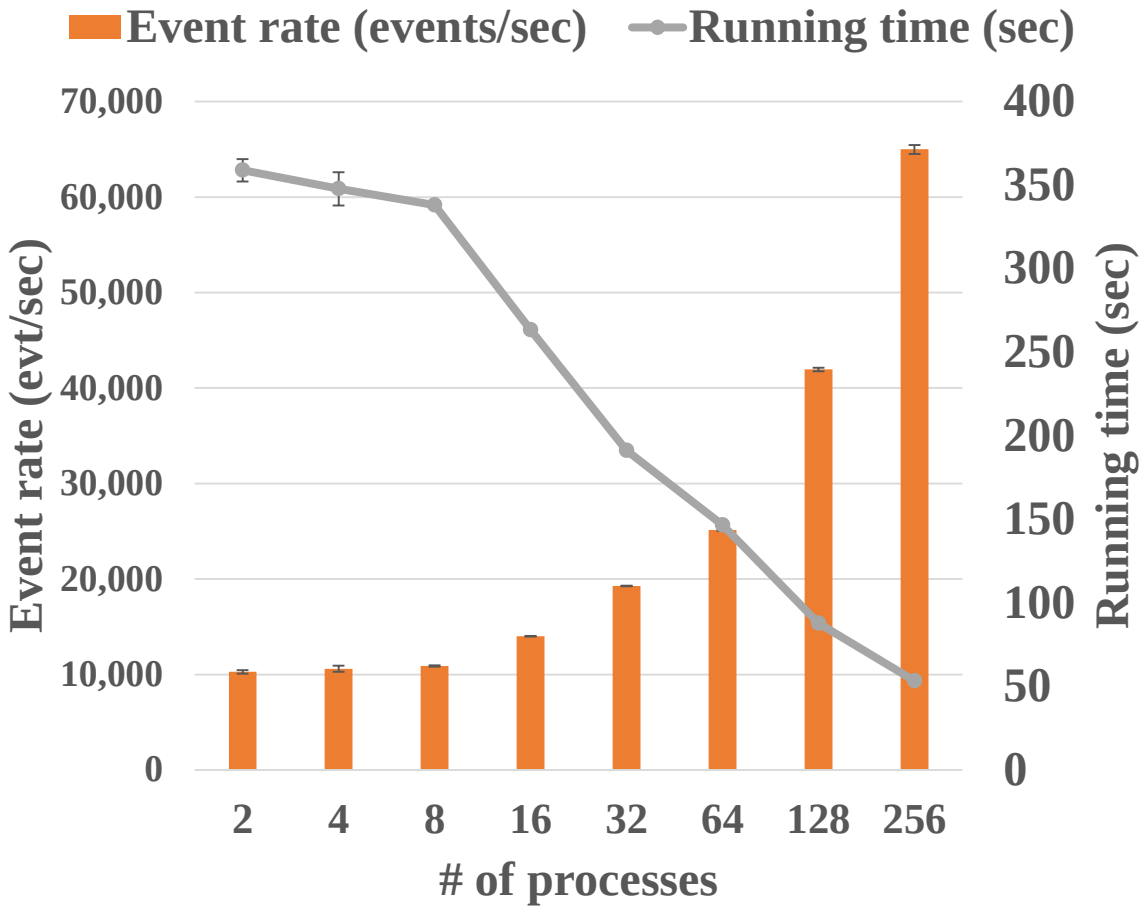
Simulation Validity



- 8 clients
- 4 HDD-servers
- 4 SSD-servers
- Lowest error rate is 2%
- Average error rate is 11.98%

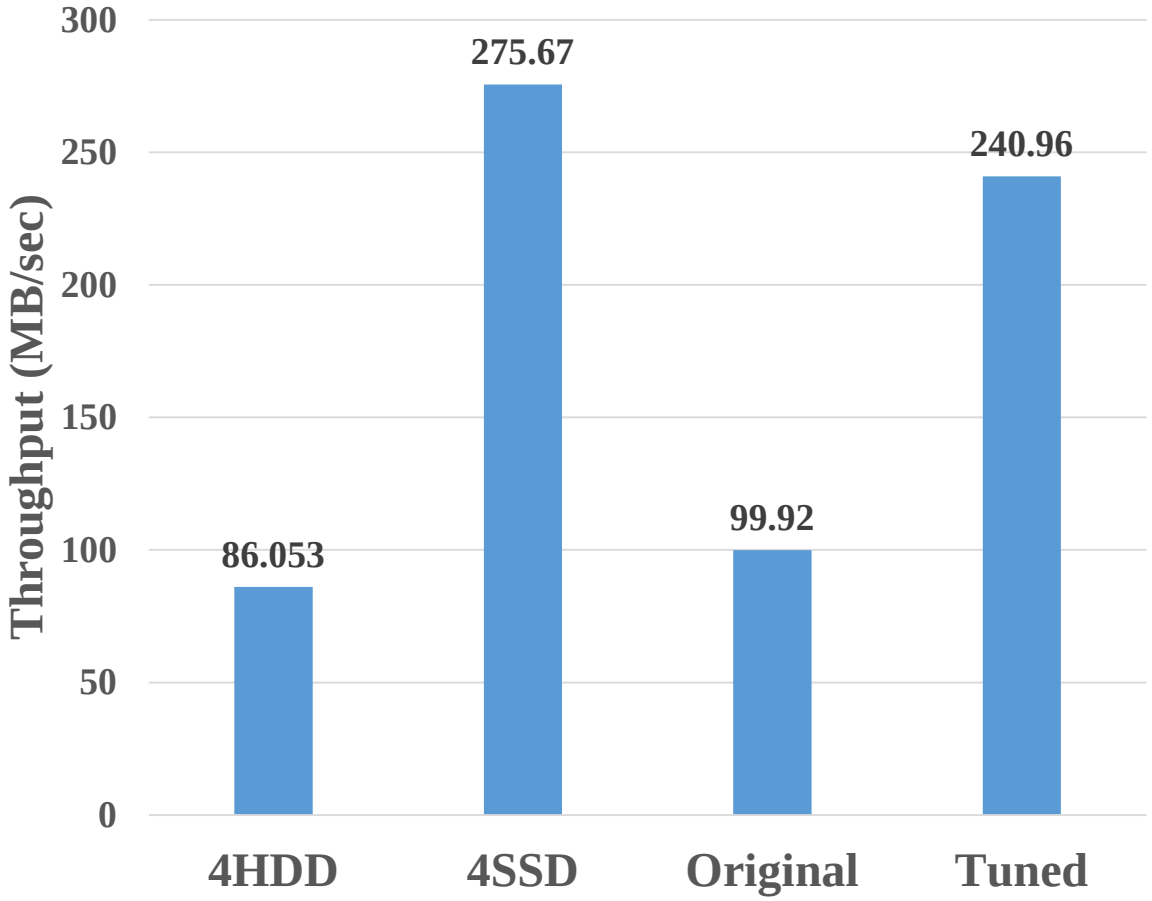
ILLINOIS INSTITUTE OF TECHNOLOGY

Simulation Performance Study



- 32 physical nodes
- 2048 clients
- 1024 servers
- # of processes from 2 to 256

Case study: Tiered-SSD Performance Tuning



- 16 clients
- 64K random requests
- 4 HDD-servers + 4 SSD-servers
- Performance boosts about 15% for original setting
- Performance boosts about 140% for tuned setting

ILLINOIS INSTITUTE OF TECHNOLOGY

Outline

- Introduction
- Related Work
- Design and Implementation
- Experiments
- Conclusions and Future Work

ILLINOIS INSTITUTE OF TECHNOLOGY

Conclusions and Future Work

- HPIS3 simulator: a hybrid parallel I/O and storage simulation system
  - Models of PVFS clients, servers, HDDs and SSDs
  - Validate against benchmarks
  - Minimum error rate is 2% and average is about 11.98% in IOR tests.
  - Scalable: # of processes from 2 to 256
  - Showcase of tiered-SSD settings under PVFS
  - Useful to find optimal settings
  - Useful to self-tuning at runtime
- Future work
  - More evaluation for tiered-SSD vs. buffered-SSD
  - Improve accuracy by detailed models
  - Client-side settings and more

ILLINOIS INSTITUTE OF TECHNOLOGY

Thank you

Questions?

Bo Feng

bfeng5@hawk.iit.edu