SCALABLE COMPUTING SOFTWARE LABORATORY

ILLINOIS INSTITUTE OF TECHNOLOGY

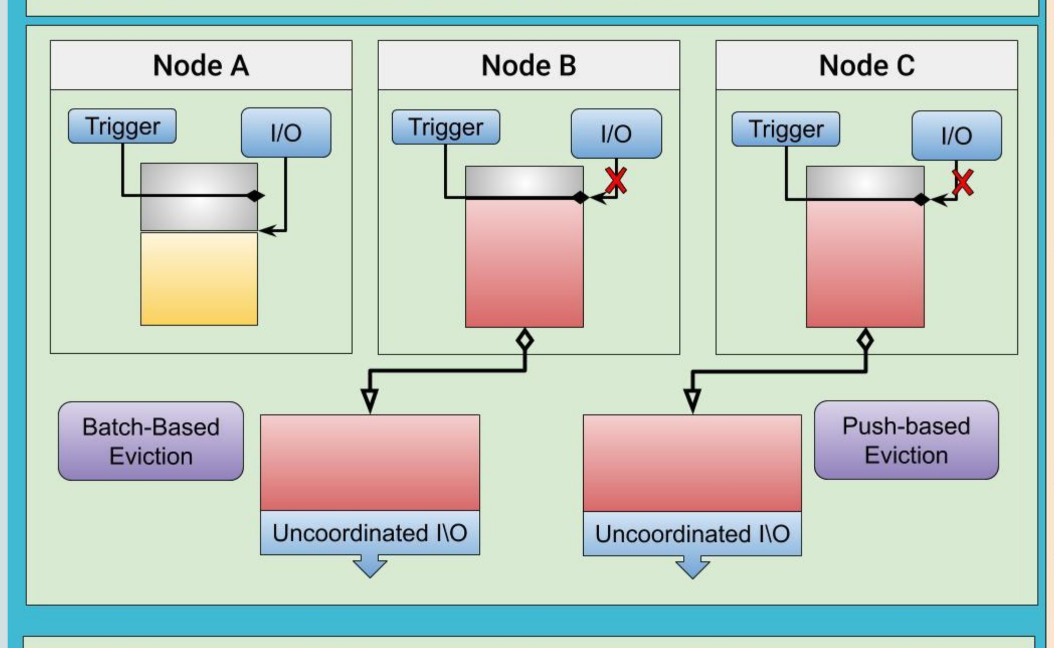
HFlush

Realtime Flushing for Modern Storage Environments

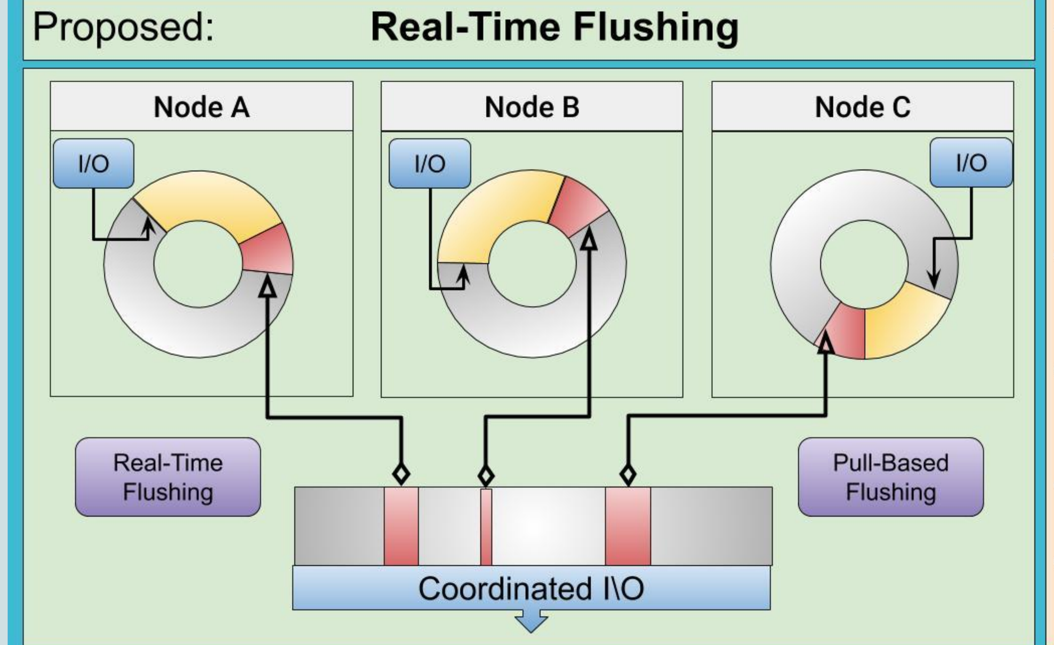
Jaime Cernuda, Hugo Trivino, Hariharan Devarajan, Anthony Kougkas, Xian-He Sun

OVERVIEW

Current: **Event-Based Evictions**



Proposed: **Real-Time Flushing**



History

→A disparity in speedup between CPU and Storage access time has created what is known as the I/O bottleneck.

→To solve this issue, traditional solutions have involved data buffering and aggregations on fast storage mediums.

→However, faster tiers of data storage, such as RAM, have lower storage capacity which eventually require eviction of data to a lower tier, traditionally a Parallel File System (PFS).

Observations

→Current eviction solutions are event-based and stall I/O when performing evictions.

→Evictions are initiated by individual nodes, without scalability, and provide writing patterns not favorable to the PFS.

→Enhanced capabilities of the new storage devices (e.g., NVMe SSDs) such as increased hardware concurrency are not taken into account by existing system software.

Approach

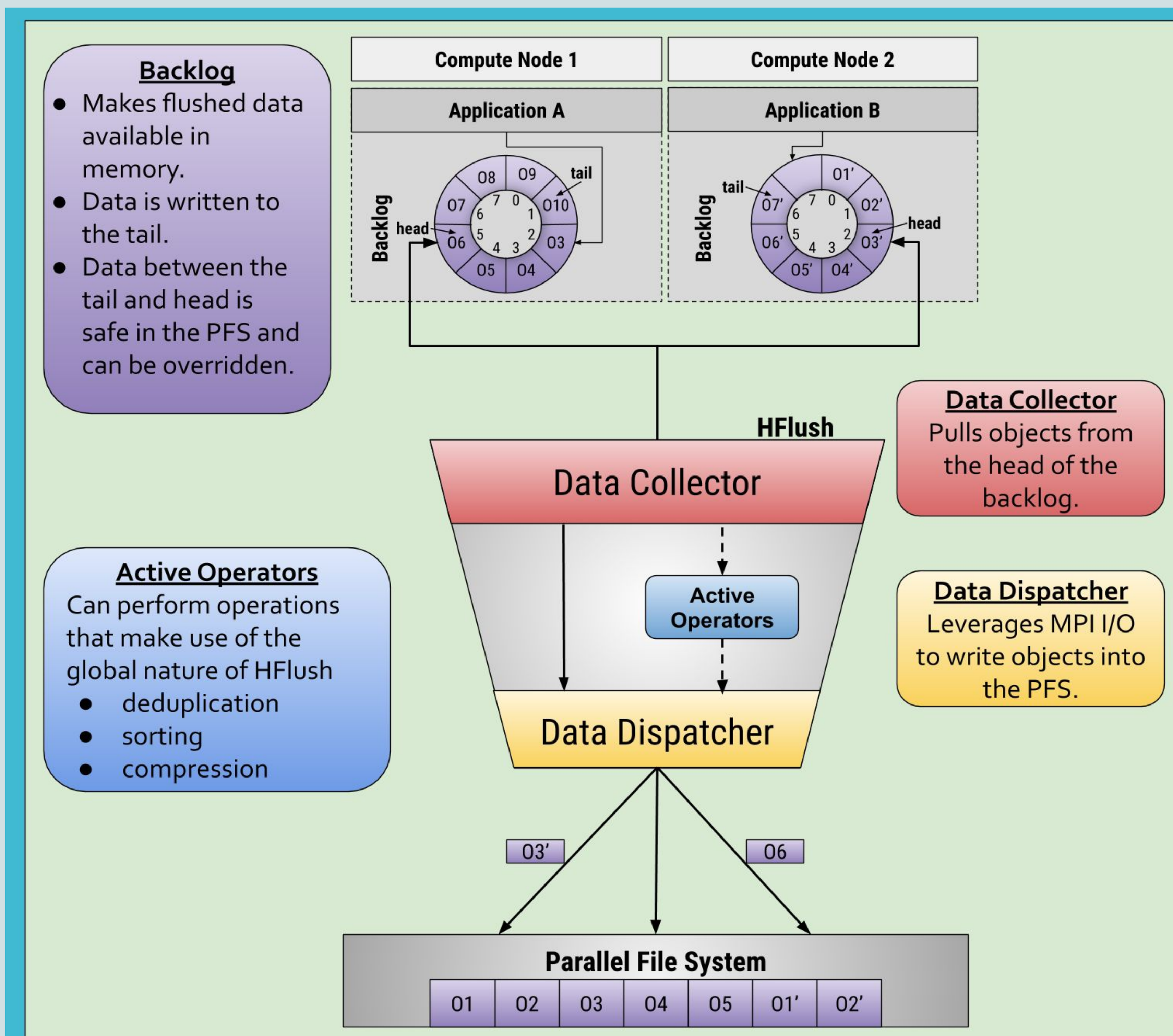
→Amortize the cost of evictions into small continuous flushing operations instead of irregularly stalling I/O operations.

→Globally coordinate all evictions to provide better writing patterns to the PFS and match the demand by offering elastic resources.

→Leverage the hardware concurrency to perform device-specific eviction optimizations.

→Move to a server-pull eviction model to provide a continuous stream of evictions.

HFLUSH DESIGN



DESIGN IMPLICATIONS

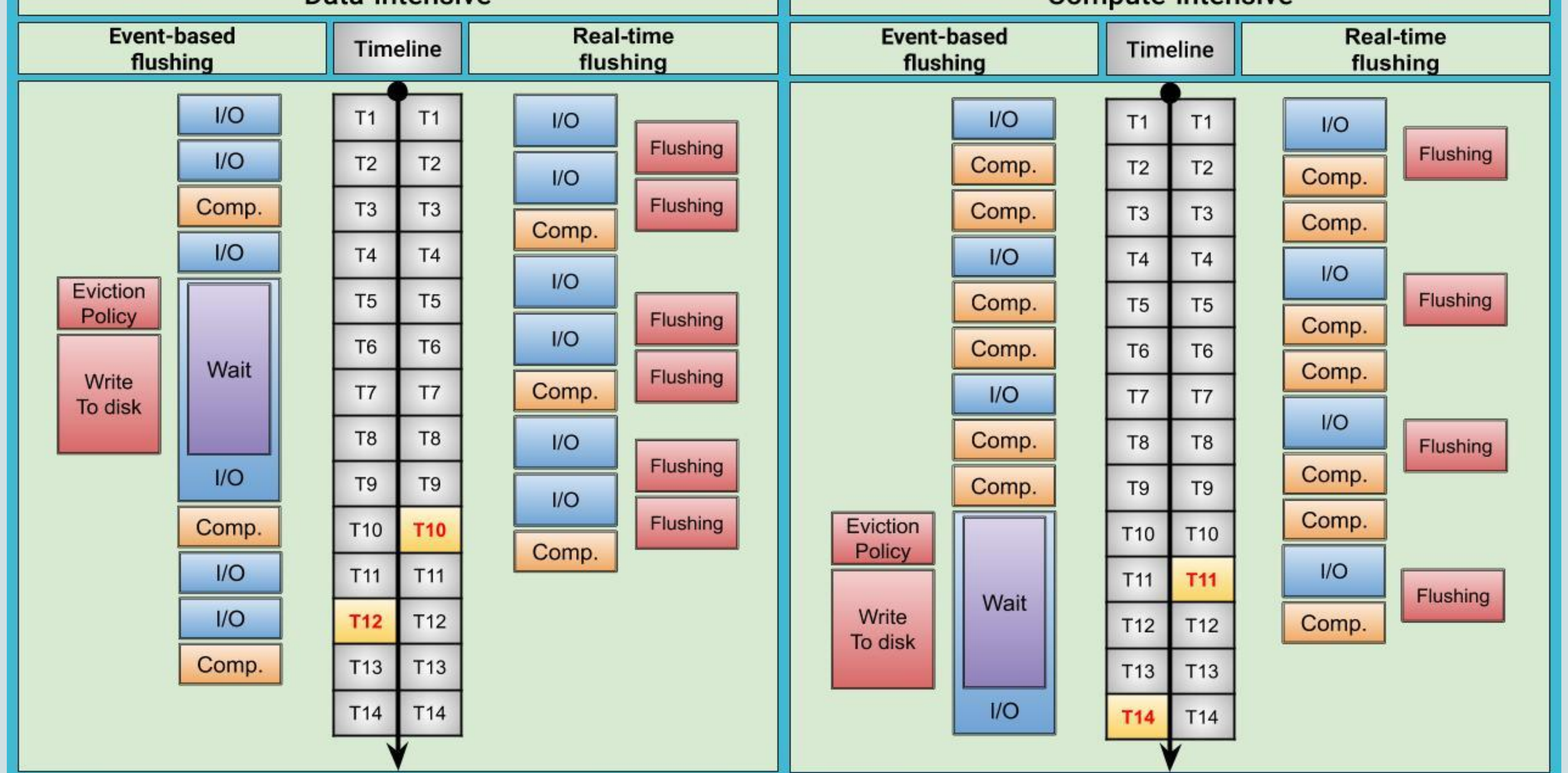
Leveraging a data streaming paradigm enables:

- Autoscaling
- Data durability
- Pipelined flushing in multi-tiered environments
- Matching hardware properties from source and destination

RELATED WORK

- Anthony Kougkas, Hariharan Devarajan, and Xian-He Sun. 2018. Hermes: A Heterogeneous-Aware Multi-Tiered Distributed I/O Buffering System. In Proceedings of the 27th International Symposium on High-Performance Parallel and Distributed Computing (HPDC '18). ACM, New York, NY, USA.
- A. Kougkas, H. Devarajan, X. Sun and J. Lofstead, "Harmonia: An Interference-Aware Dynamic I/O Scheduler for Shared Non-volatile Burst Buffers," 2018 IEEE International Conference on Cluster Computing (CLUSTER), Belfast, 2018, pp. 290-301.
- D. Zhao, K. Qiao and I. Raicu, "HyCache+: Towards Scalable High-Performance Caching Middleware for Parallel File Systems," 2014 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, Chicago, IL, 2014, pp. 267-276.

WORKLOADS

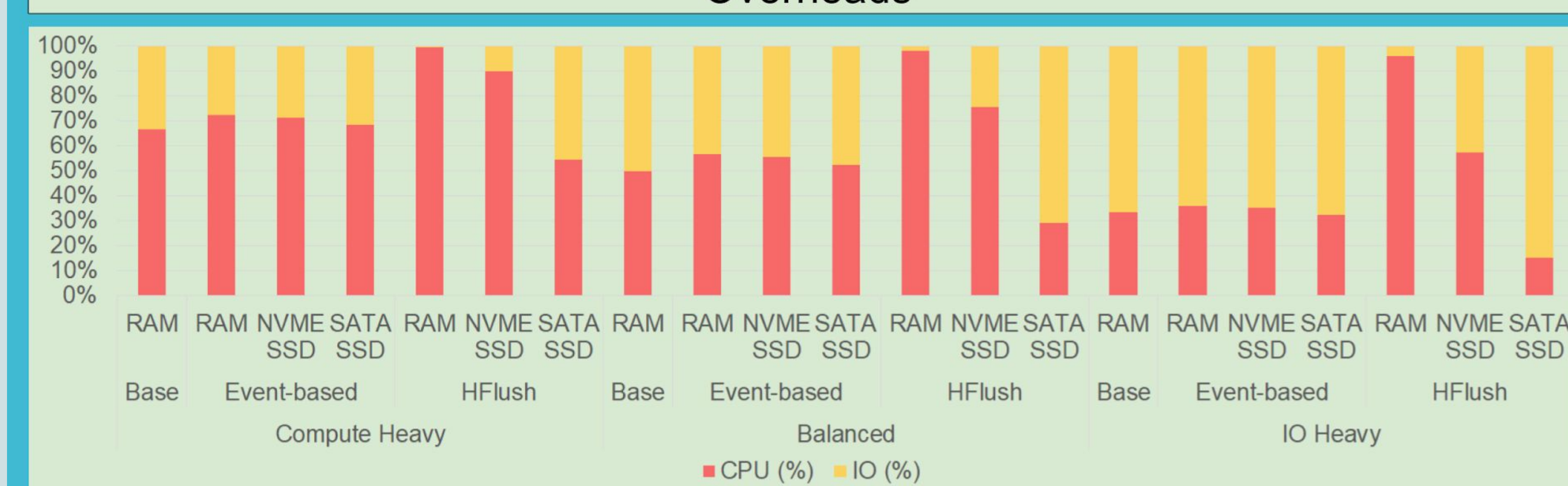


* Yellow boxes indicate I/O end time

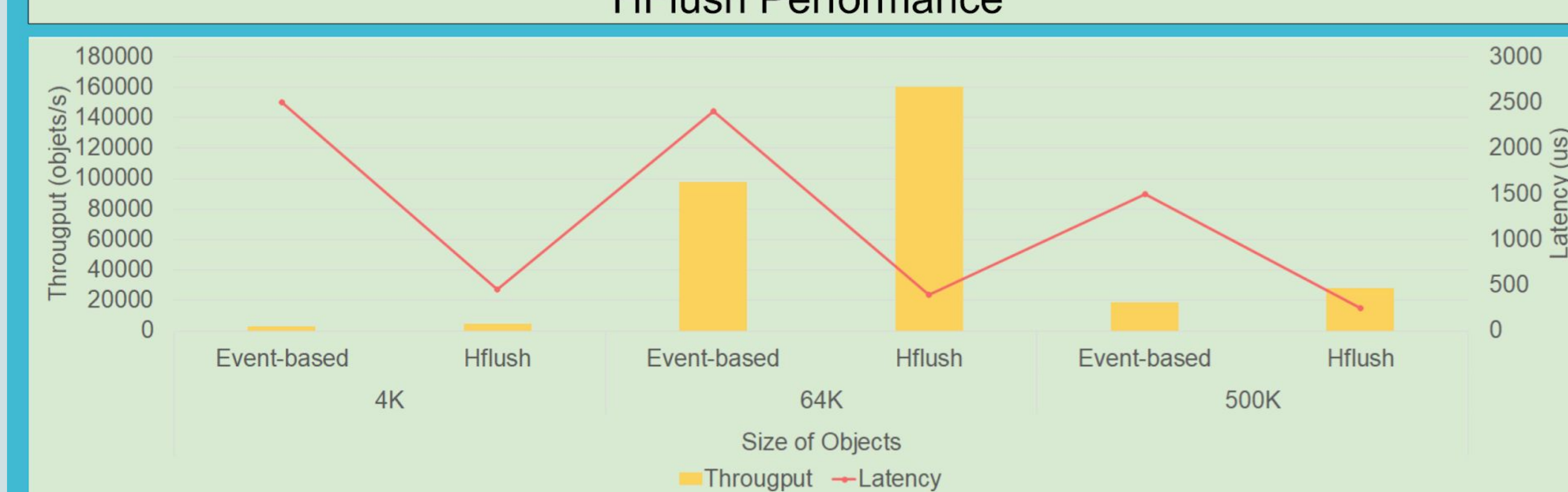
** I/O in Real-Time Flushing are drawn larger because of higher hardware interference

INITIAL EVALUATIONS

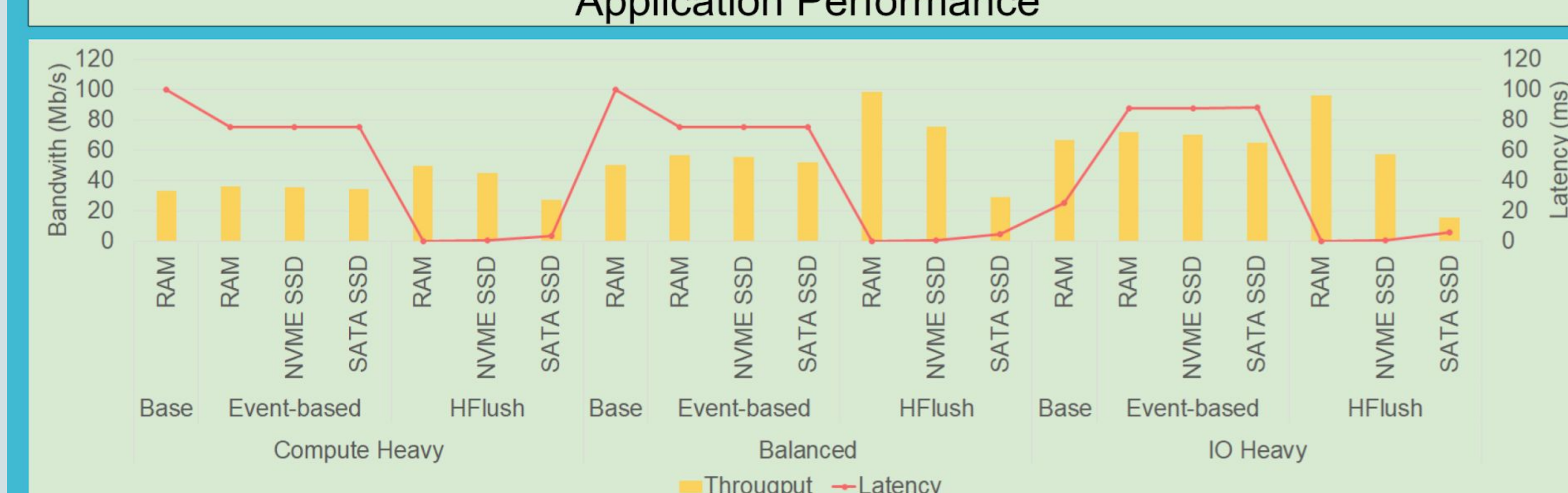
Overheads



HFlush Performance



Application Performance



Testbed: Ares supercomputer at the Illinois Institute of Technology

Compute Nodes:
- CPU: Dual Intel(R) Xeon Scalable Silver 4114 @ 2.20GHz (40 nodes)
- RAM: 96 GB RAM
- Network: 10Gbit Ethernet with RoCE
- Storage: local 512GB NVMe SSD.

Storage Nodes:
- CPU: Two quad-core Opteron 2376 @ 2.3GHz (40 nodes)
- RAM: 32GB DDR2-667
- Storage: 1 250GB Samsung 860 Evo SATA SSD, 1TB Seagate 7200K SATA hard drive

CONCLUSIONS

- HFlush is a pull-based data flusher that implements a continuous data eviction mechanism.
- Initial results have shown HFlush to be a promising solution to the growing challenge of extreme scale data generation, especially in case of workloads with periodic I/O or in systems that make use of modern hardware with high concurrency.
- The ability of HFlush to amortize the I/O stall time allows applications using it to significantly increase the CPU usage by over **50%**.
- The near real-time nature of the eviction provides an improved overall latency on the data flushing with **7x** latency reduction and a **2X** bandwidth increase over batch-based flushing solutions, which reflects in a lower I/O stall time for the application.

Jaime Cernuda
Illinois Institute of Technology
jcernudagarcia@hawk.iit.edu

Hugo Trivino
Illinois Institute of Technology
hhernandeztrivino@hawk.iit.edu

SCAN ME

tinyurl.com/hflush