IOCost: Block IO Control for Containers in Datacenters

Tejun Heo, **Dan Schatzberg**, Andrew Newell, Song Liu, Saravanan Dhakshinamurthy, Iyswarya Narayanan, Josef Bacik, Chris Mason, Chunqiang Tang, ‡Dimitrios Skarlatos

Session 5B: Data Center and Cloud Services
Thursday, March 3 @ 2pm

ASPLOS 2022

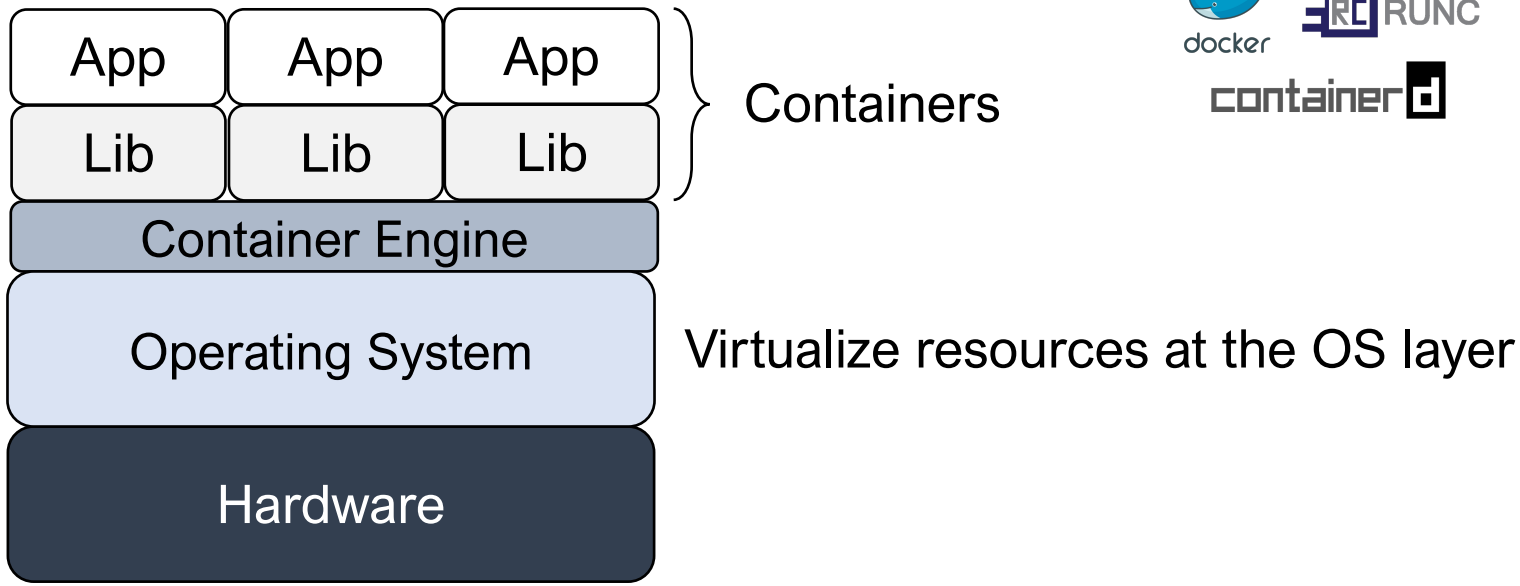
Meta

‡ Carnegie Mellon University
Computer Science Department

New Era in Datacenter Computing

Meta fleet runs on containers

- Lightweight
- Fast bring up
- Higher consolidation



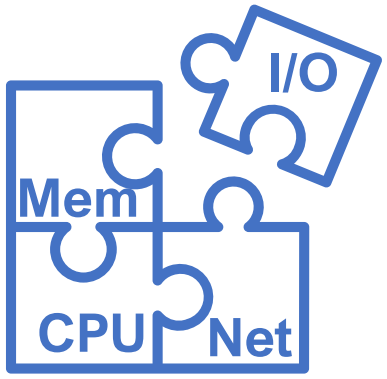
Twine: A Unified Cluster Management System for Shared Infrastructure – OSDI'20
RAS: Continuously Optimized Region-Wide Datacenter Resource Allocation – SOSP'21

Problem: Resource Isolation

Ensure sufficient resources despite collocated workloads

Well understood problem for CPU, Memory, Network

Missing piece: Block I/O for storage devices (SSDs, HDDs)



Resource Isolation

Why do we need Block I/O?

Storage Systems

APACHE ZooKeeper™

MySQL®

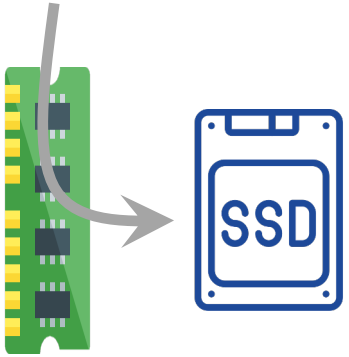
System Updating
Package Fetching

puppet

CHEF™

ANSIBLE

Memory Offloading



TMO: Transparent Memory Offloading @ASPLOS 2022

Challenge A → Hardware Heterogeneity

Multiple generations of SSDs, HDDs, Local/Remote

Different performance characteristics within each type

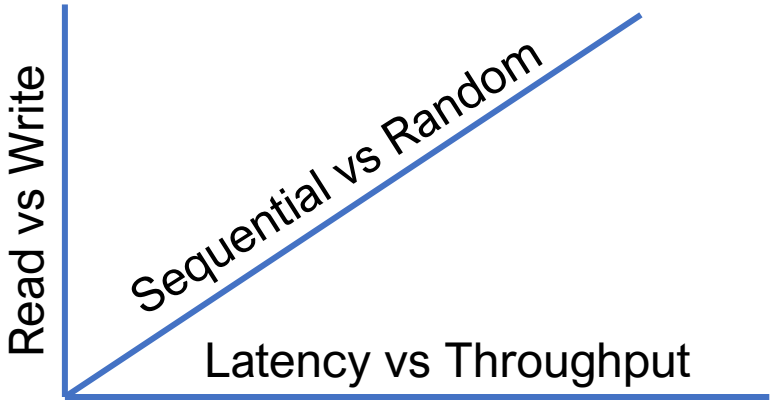
Idiosyncrasies of each device and hardware features

Challenge B → Workload Heterogeneity

IO control needs to cater to a wide variety of workloads

Varying characteristics:

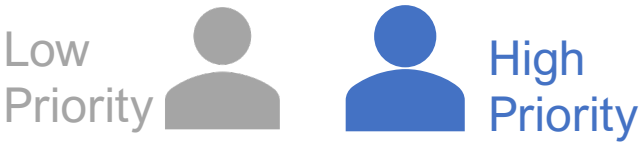
- Latency vs Throughput
- Read vs Write
- Sequential vs Random



Challenge C → Datacenter Requirements

Datacenter isolation of I/O needs to provide:

Ease of Use

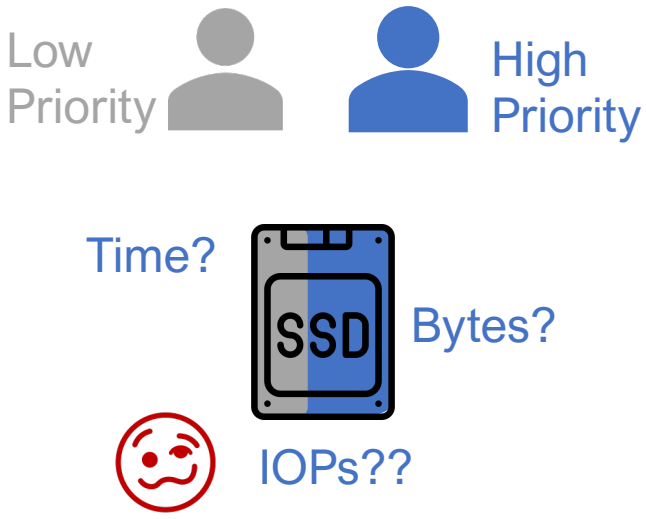


Difficult to configure

Challenge C → Datacenter Requirements

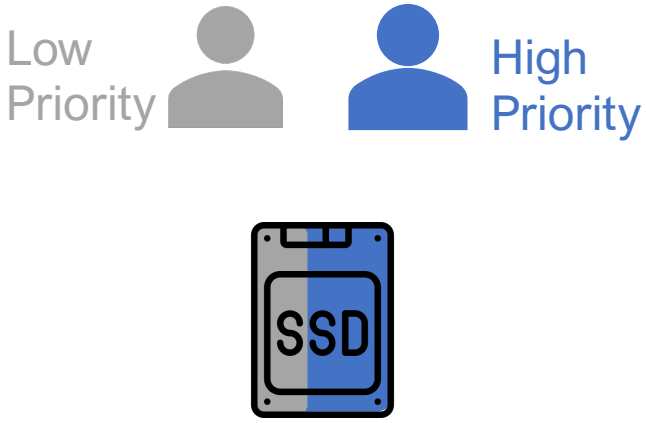
Datacenter isolation of I/O needs to provide:

Ease of Use



Difficult to configure

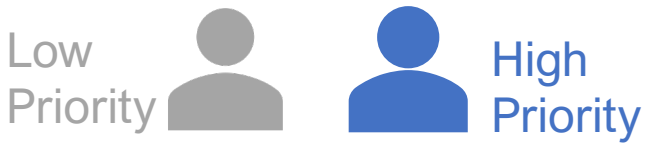
Work Conservation



Challenge C → Datacenter Requirements

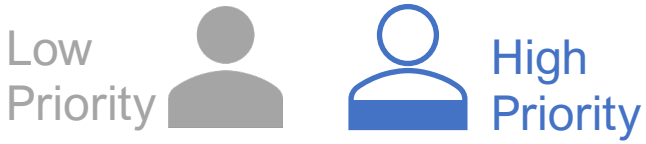
Datacenter isolation of I/O needs to provide:

Ease of Use



Difficult to configure

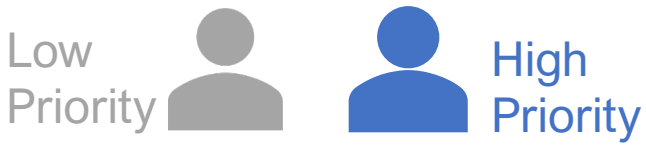
Work Conservation



Challenge C → Datacenter Requirements

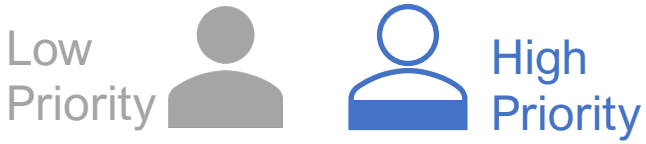
Datacenter isolation of I/O needs to provide:

Ease of Use



Difficult to configure

Work Conservation

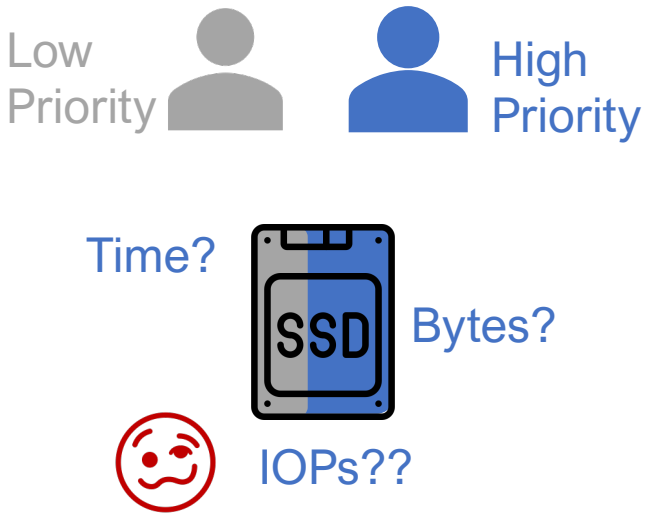


Underutilization

Challenge C → Datacenter Requirements

Datacenter isolation of I/O needs to provide:

Ease of Use

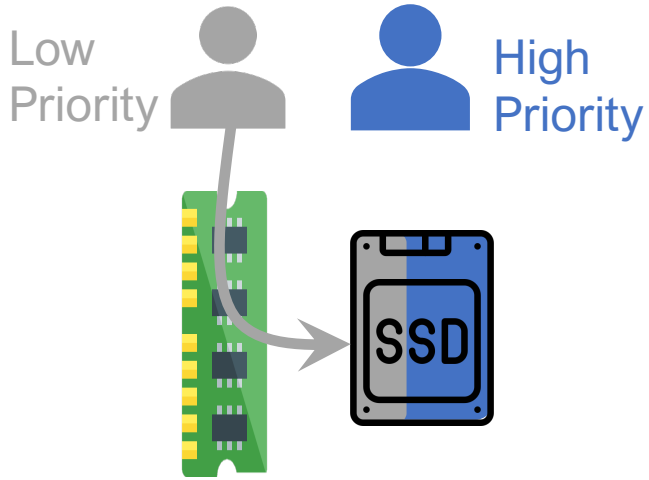


Difficult to configure

Work Conservation

Underutilization

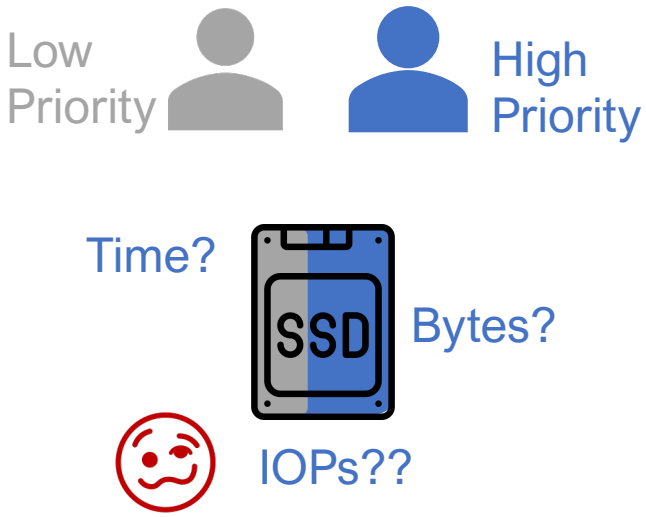
Memory-awareness



Challenge C → Datacenter Requirements

Datacenter isolation of I/O needs to provide:

Ease of Use

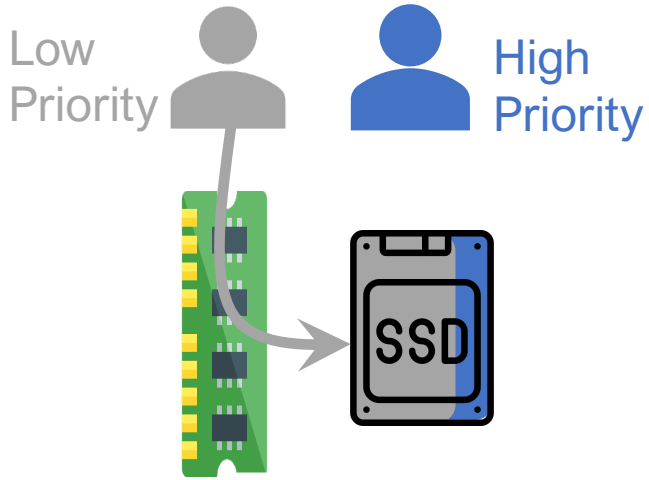


Difficult to configure

Work Conservation

Underutilization

Memory-awareness

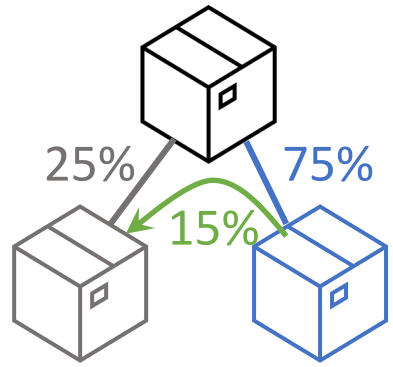


Priority Inversion

Contribution → IOCost
IO Control for Containers in Datacenters

Device occupancy with offline cost and QoS models

Proportional IO control through weights across containers

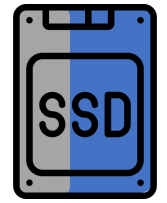


Lightweight work-conserving budget donation algorithm

IOCost manages IO across the entire Meta fleet

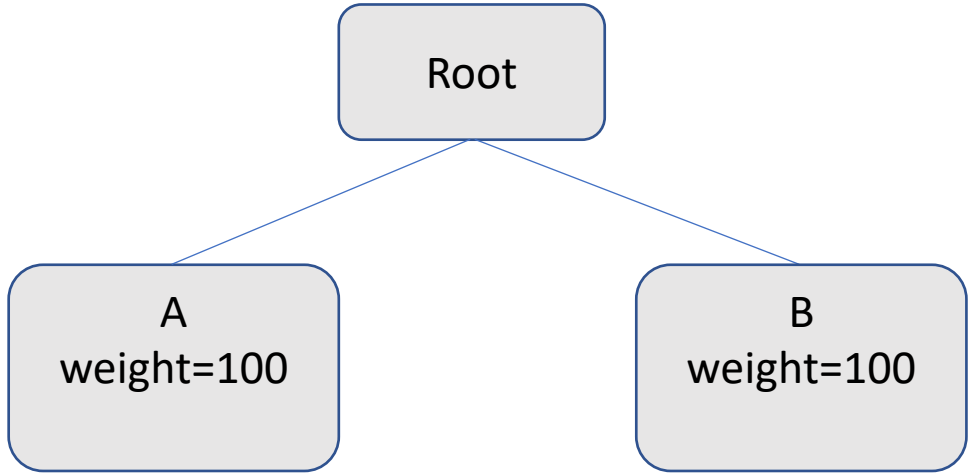
Open-source device profiling and benchmarking tools

Upstreamed in Linux

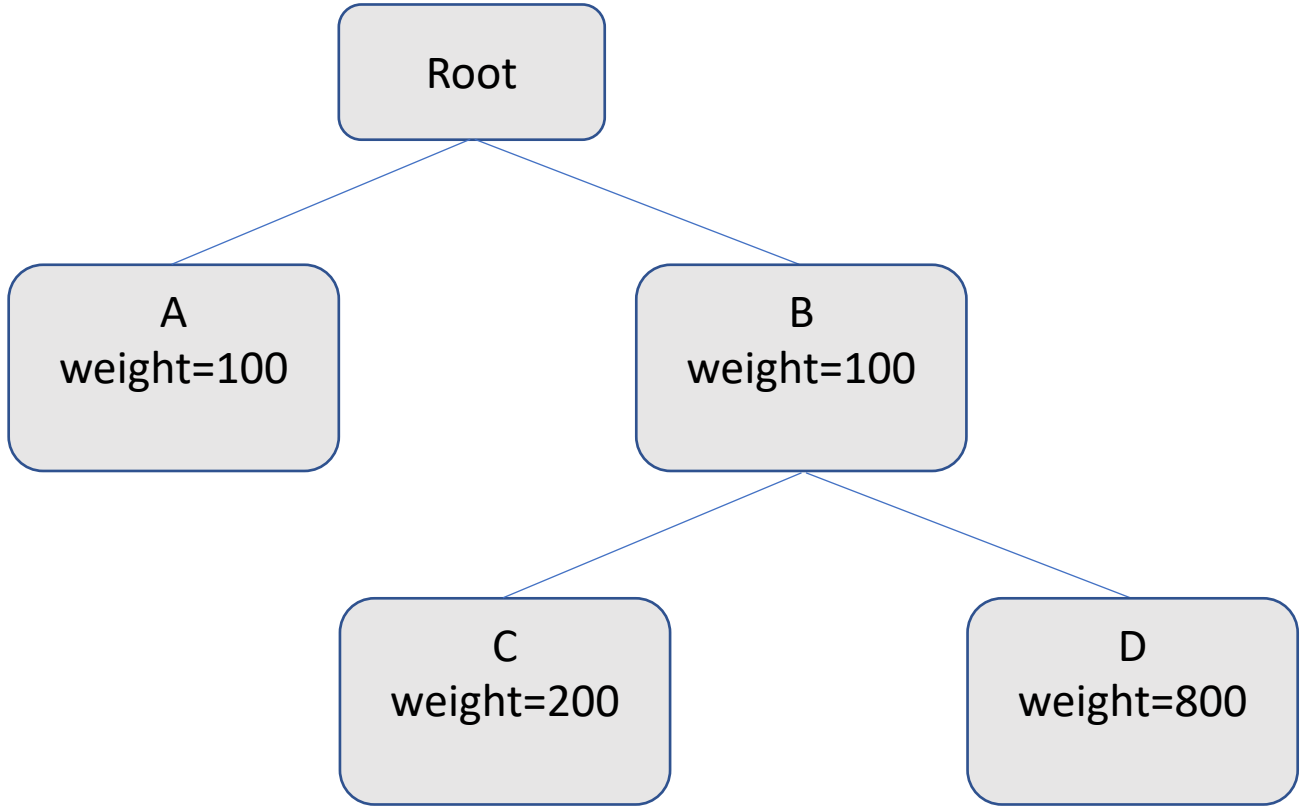


IOCost Design

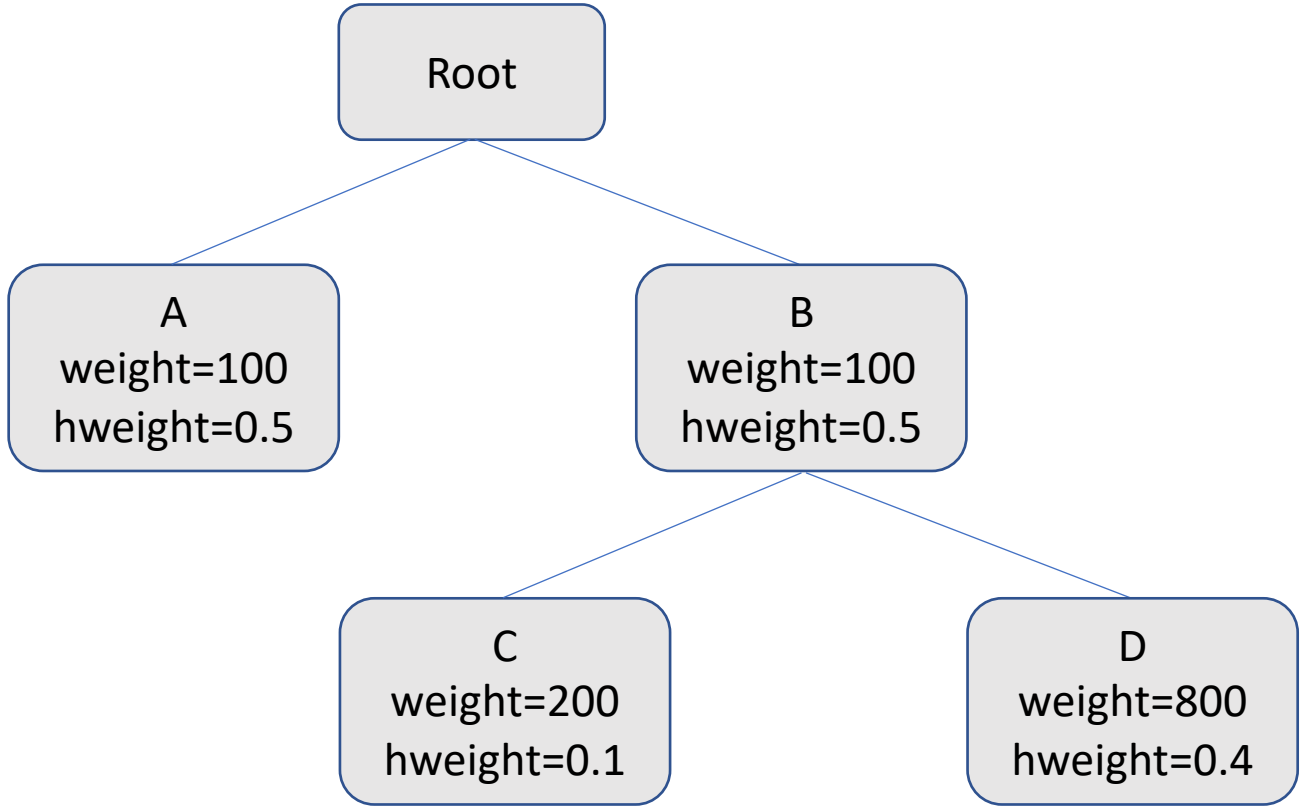
IOCost Is Configured With Weights



IOCost Is Configured With Weights

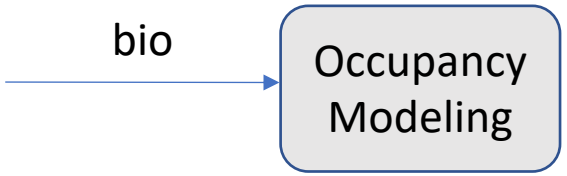


IOCost Is Configured With Weights

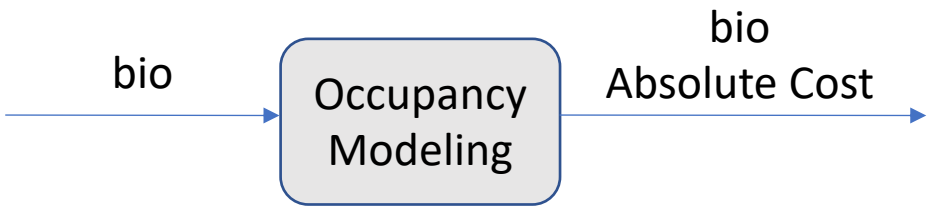


IOCost Issue Path Estimates Occupancy

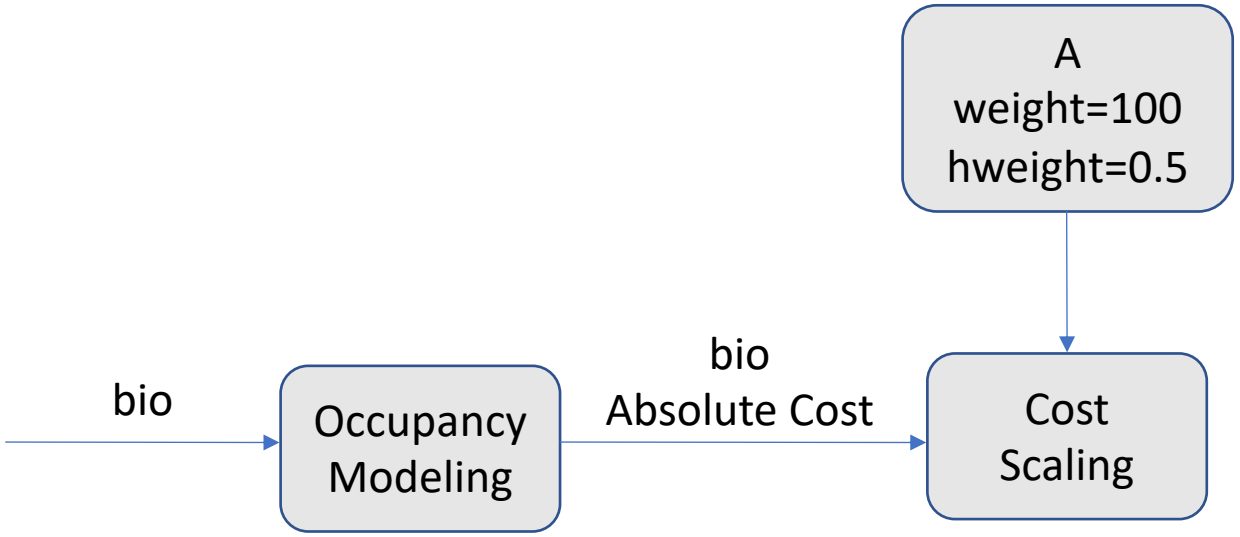
IOCost Issue Path Estimates Occupancy



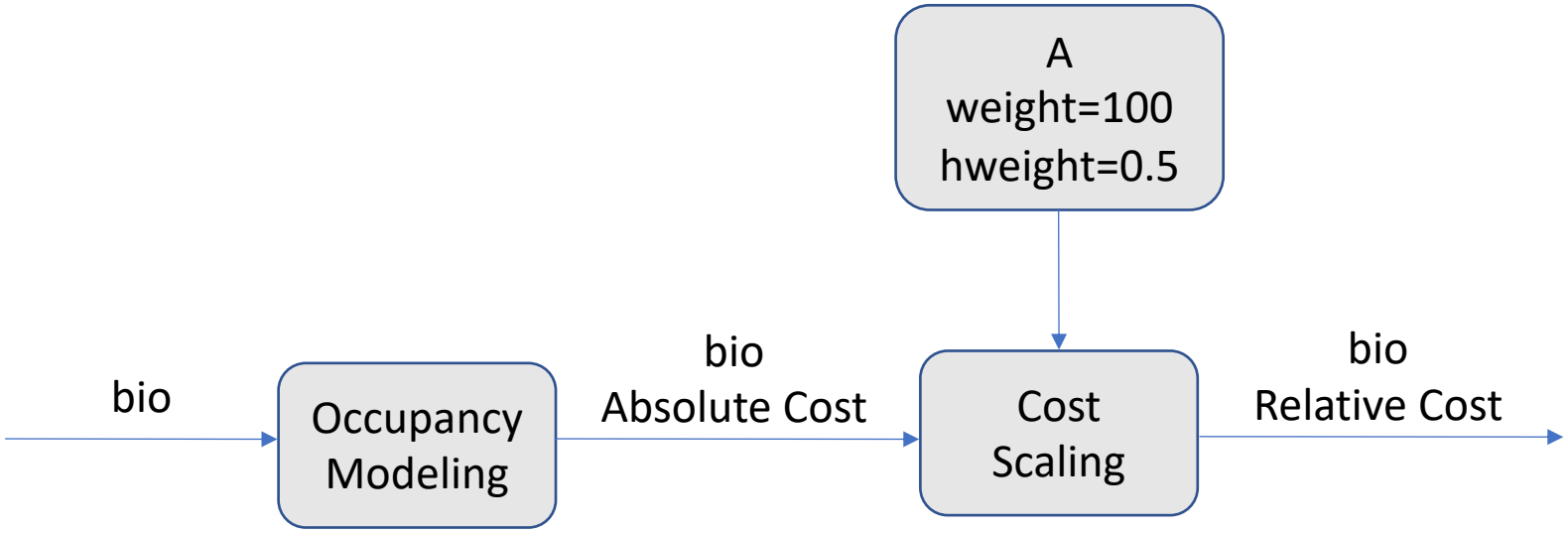
IOCost Issue Path Estimates Occupancy



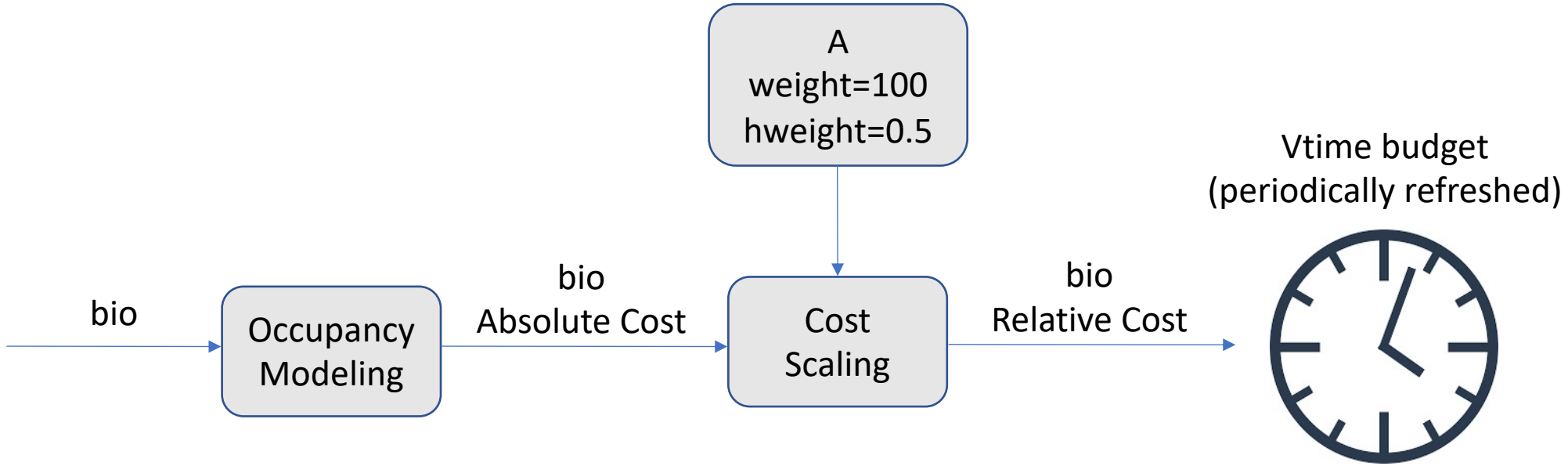
IOCost Issue Path Estimates Occupancy



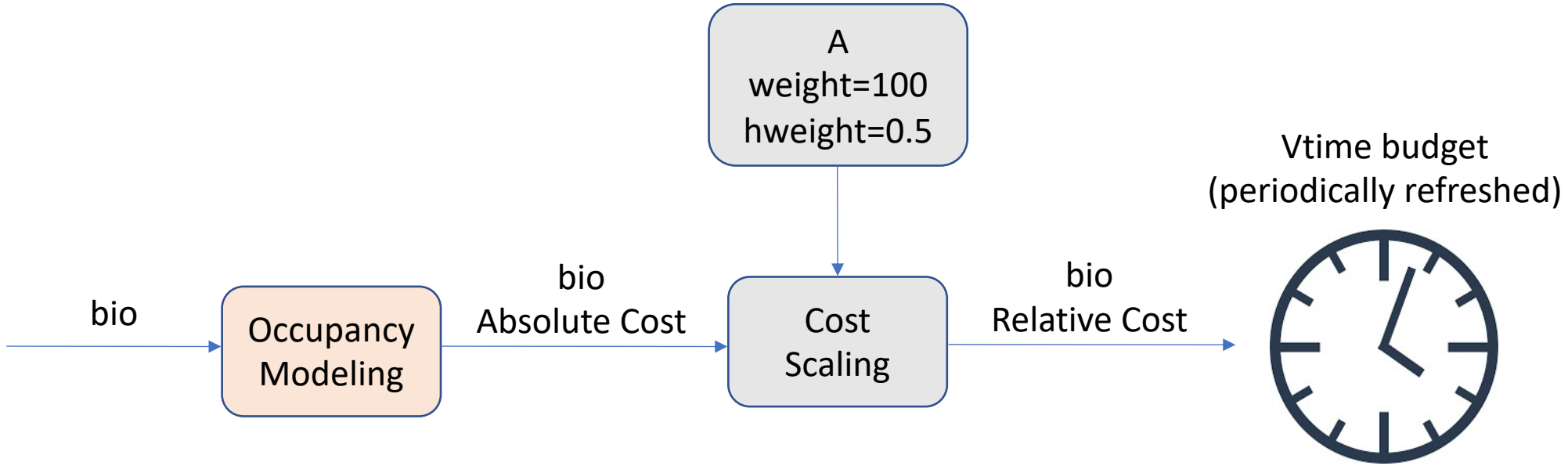
IOCost Issue Path Estimates Occupancy



IOCost Issue Path Estimates Occupancy



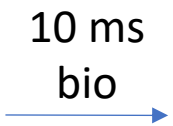
IOCost Issue Path Estimates Occupancy



Occupancy Modeling Captures Device Differences

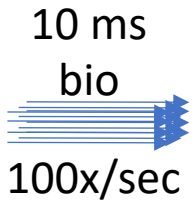
Occupancy Modeling Captures Device Differences

IO Cost is in units of time but not a measure of latency



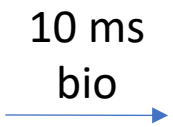
Occupancy Modeling Captures Device Differences

IO Cost is in units of time but not a measure of latency



Occupancy Modeling Captures Device Differences

IO Cost is in units of time but not a measure of latency

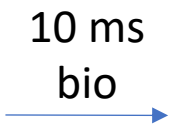


Linear Model estimated offline via synthetic saturating workloads

io_cost = base_cost + size_cost_rate * bio_size

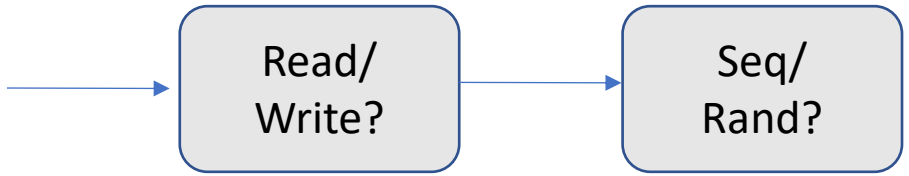
Occupancy Modeling Captures Device Differences

IO Cost is in units of time but not a measure of latency



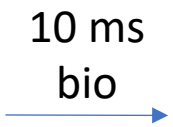
Linear Model estimated offline via synthetic saturating workloads

io_cost = base_cost + size_cost_rate * bio_size



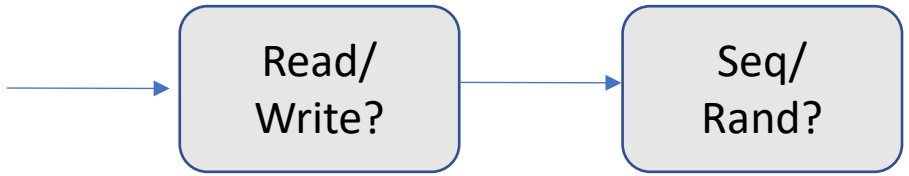
Occupancy Modeling Captures Device Differences

IO Cost is in units of time but not a measure of latency



Linear Model estimated offline via synthetic saturating workloads

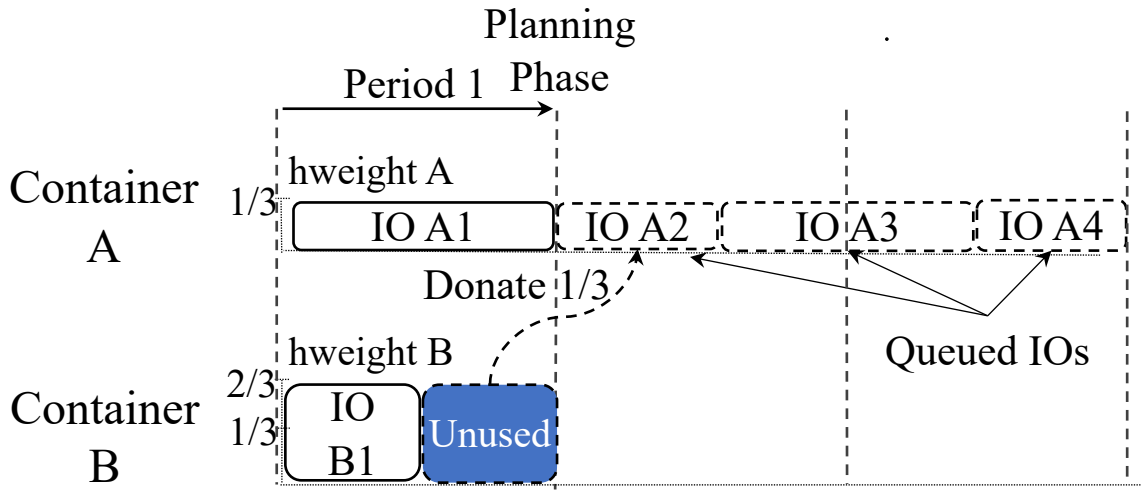
io_cost = base_cost + size_cost_rate * bio_size



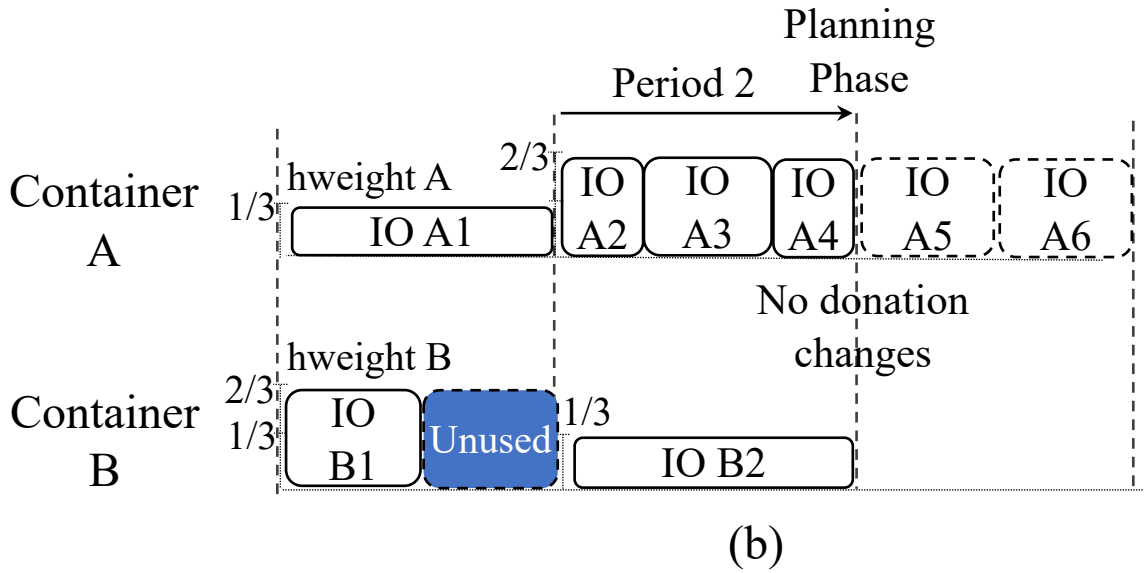
How to adjust for inaccurate models? VRate adjustment (see paper for details)

Budget Donation Ensures Work Conservation

Budget Donation Ensures Work Conservation

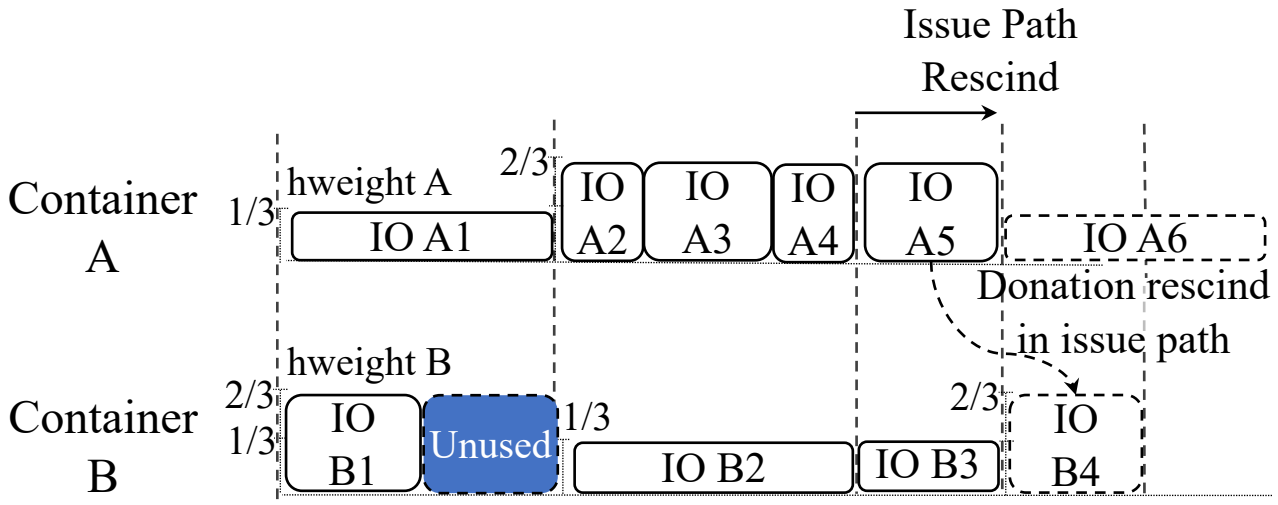


Budget Donation Ensures Work Conservation



(b)

Budget Donation Ensures Work Conservation

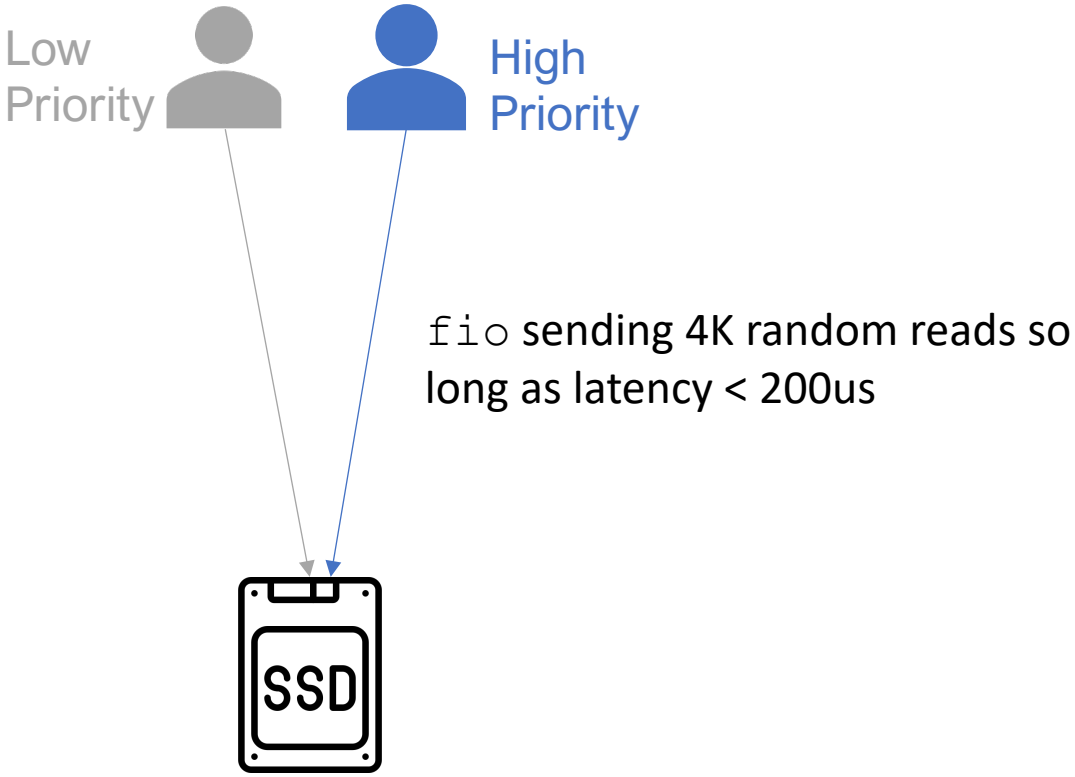


(c)

Evaluation

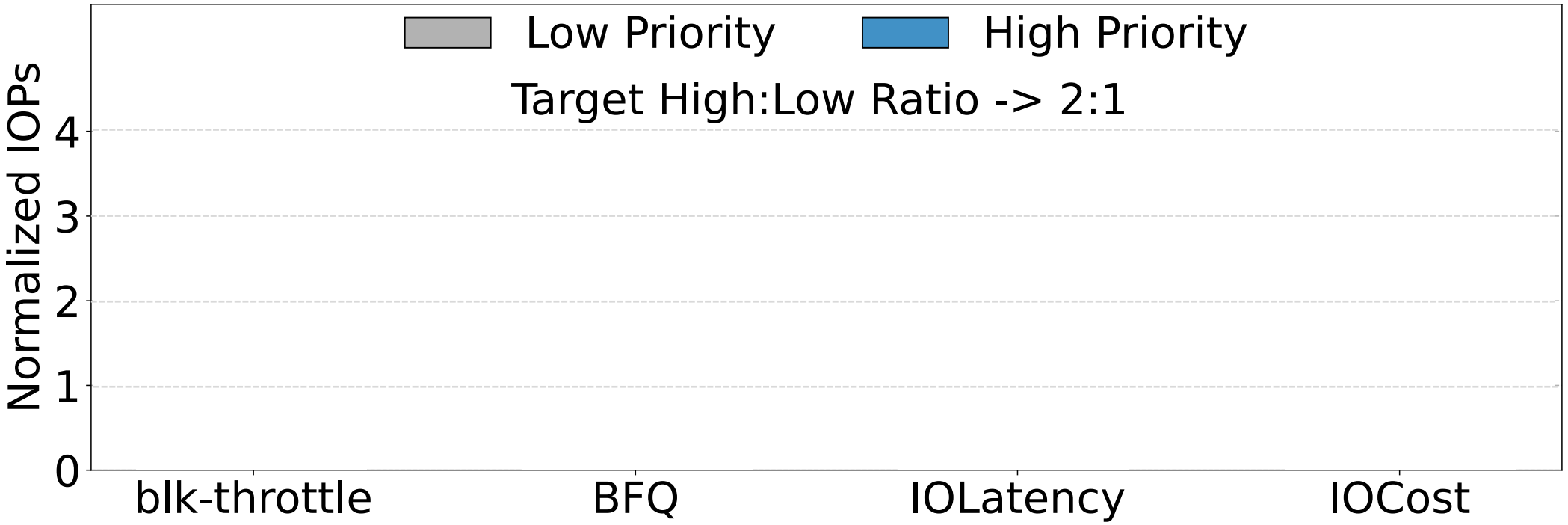
Evaluation – Proportional Control

Evaluation – Proportional Control

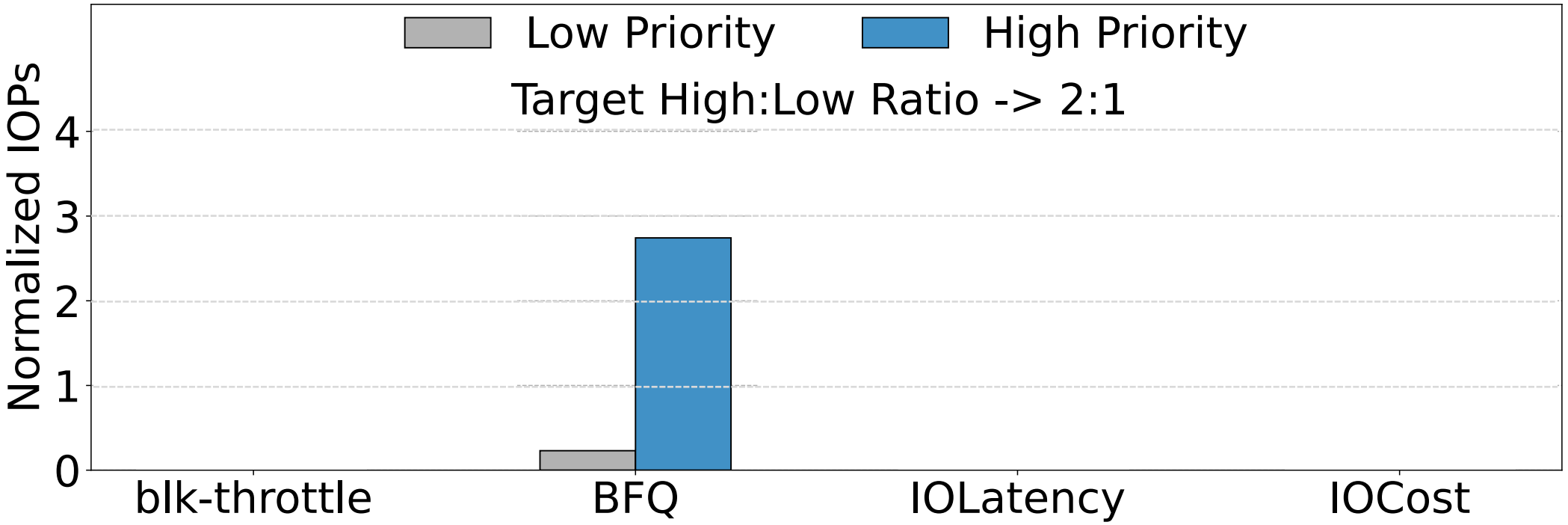


`fio` sending 4K random reads so long as latency < 200us

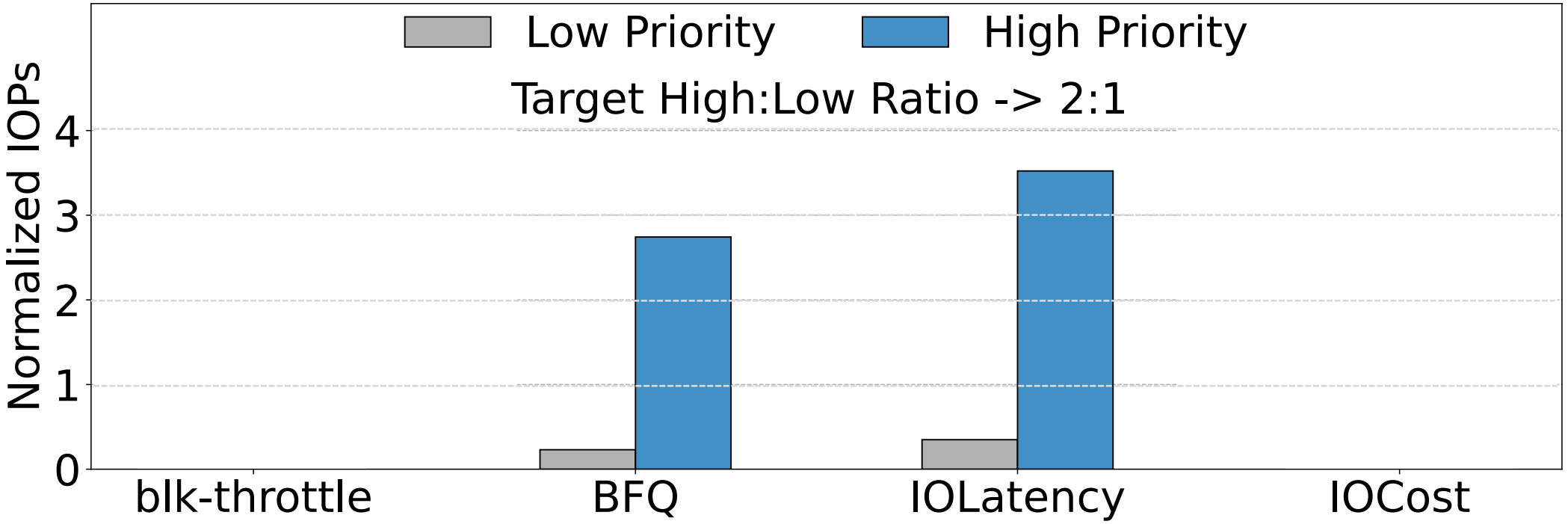
Evaluation – Proportional Control



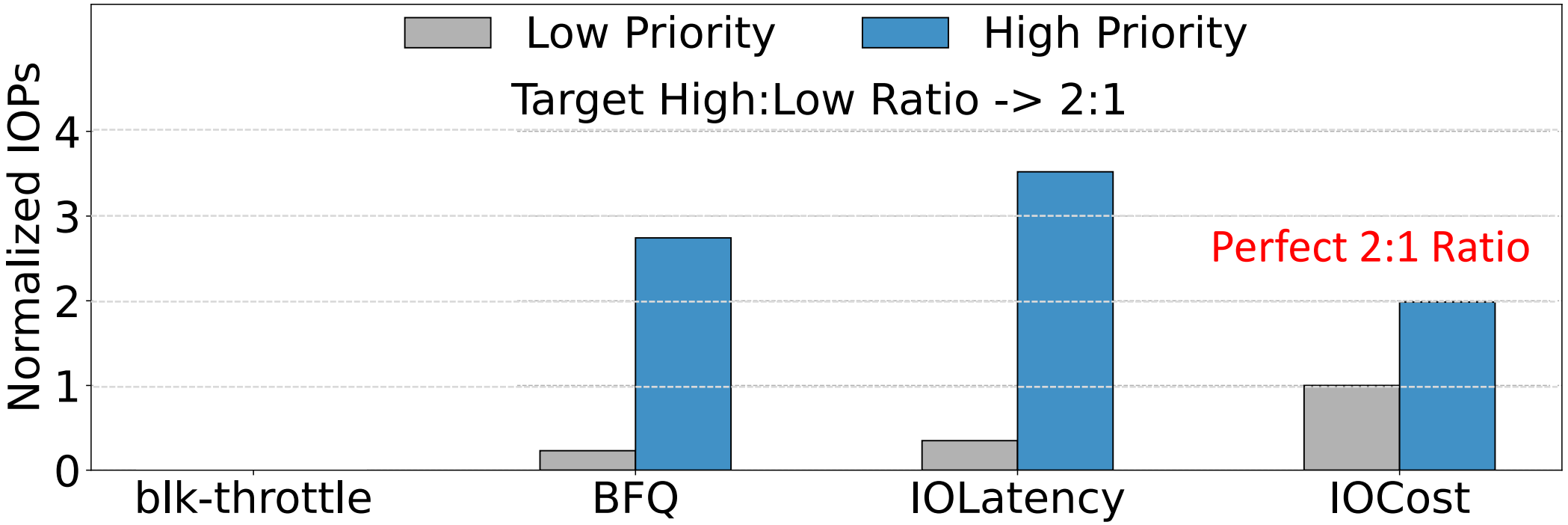
Evaluation – Proportional Control



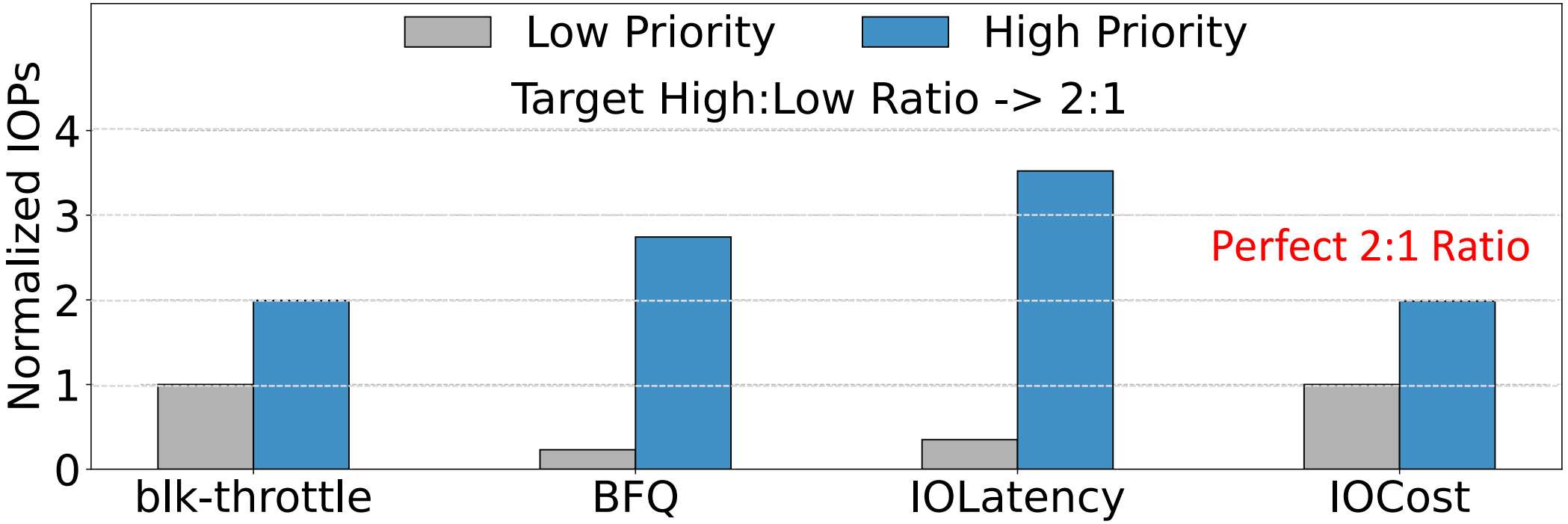
Evaluation – Proportional Control



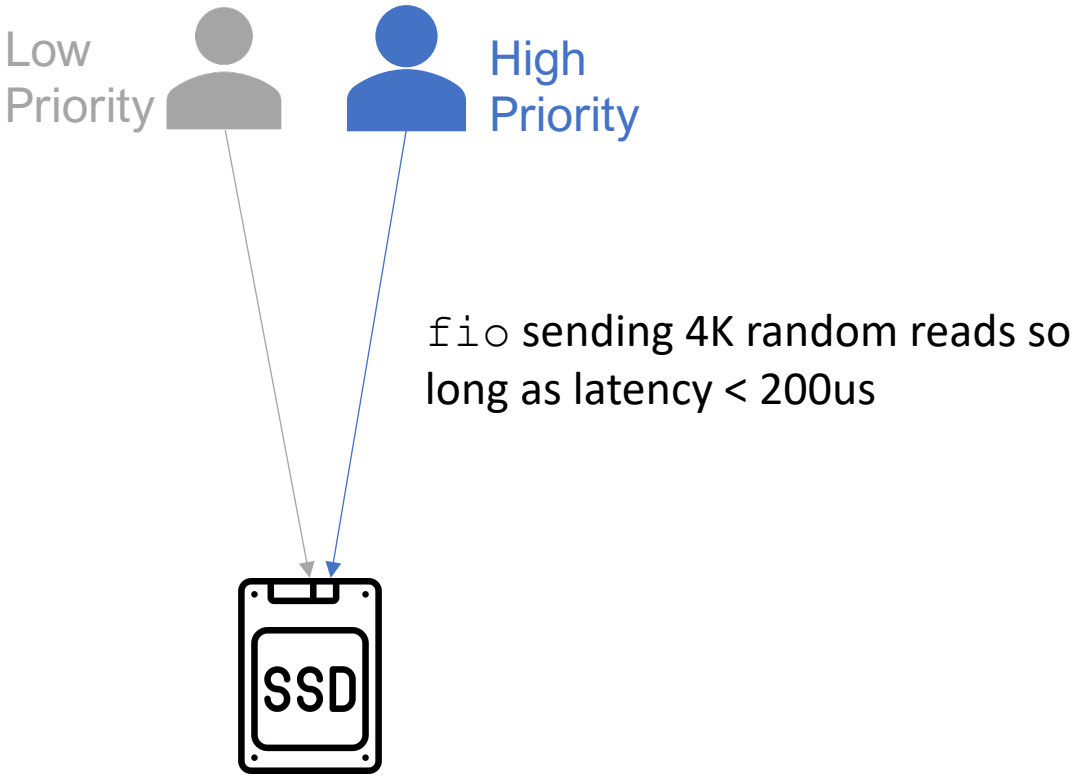
Evaluation – Proportional Control



Evaluation – Proportional Control

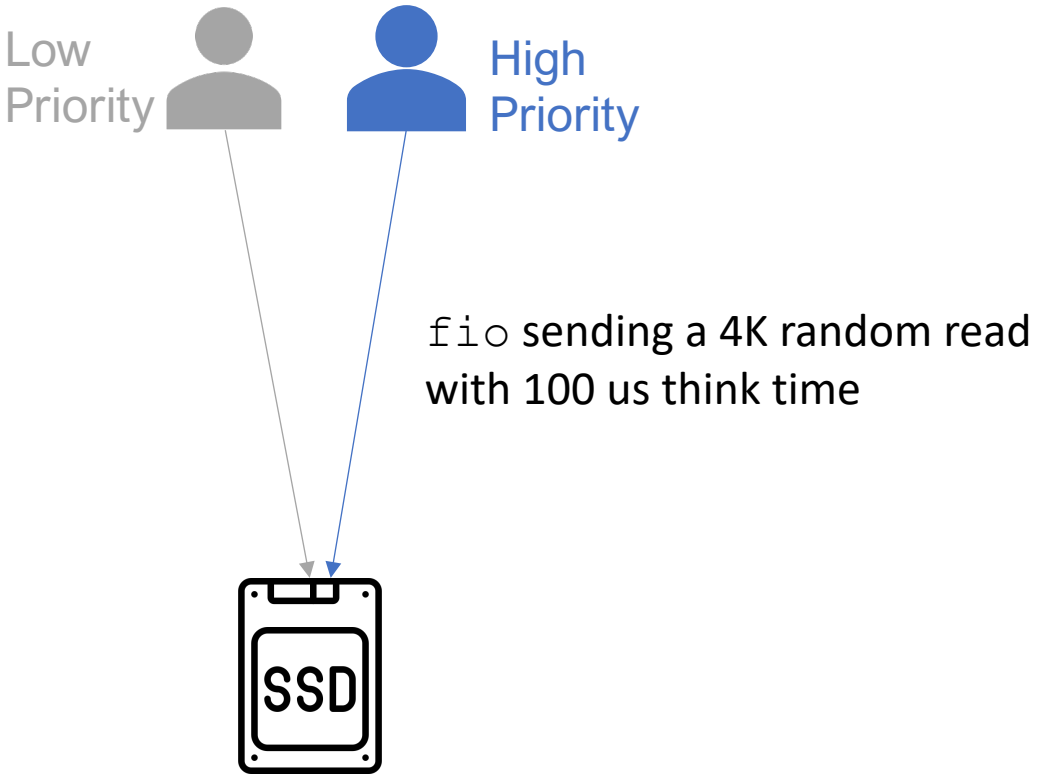


Evaluation – Work Conservation



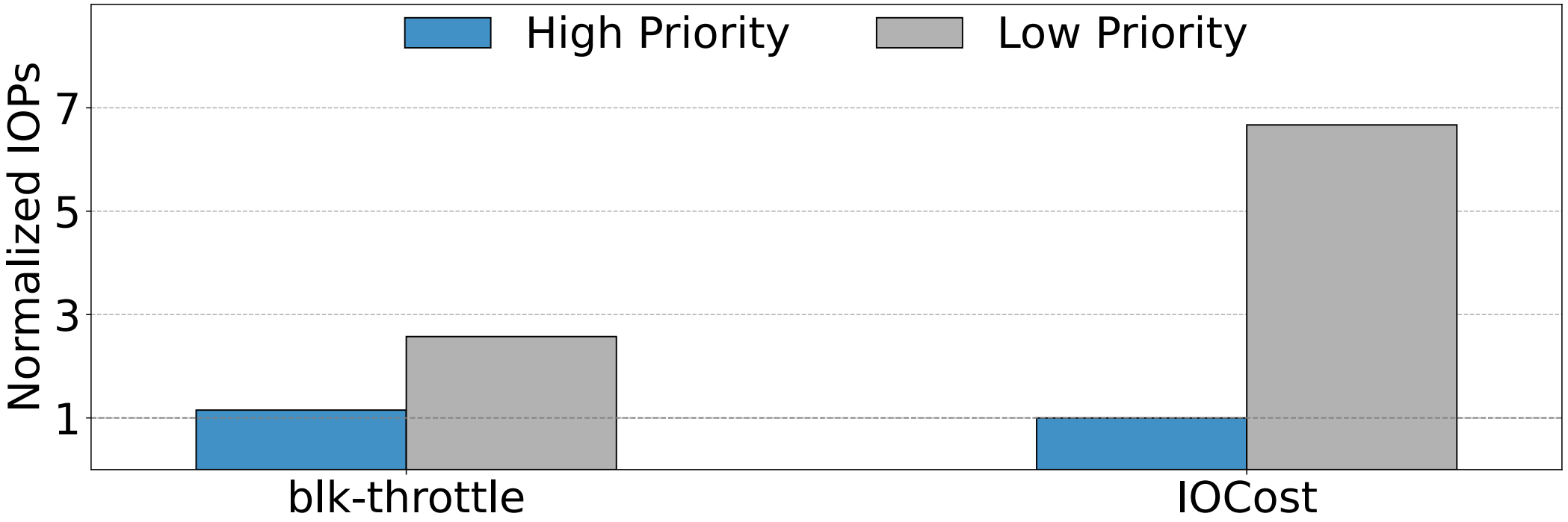
`fio` sending 4K random reads so long as latency < 200us

Evaluation – Work Conservation



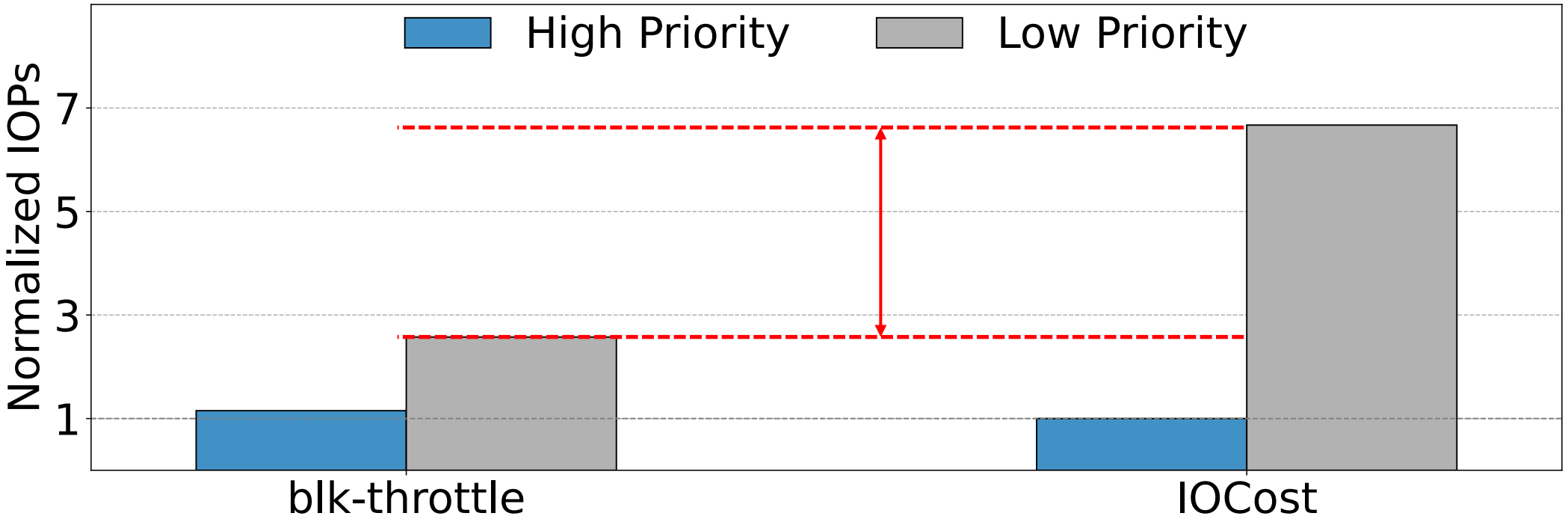
Evaluation – Work Conservation

The low priority workload should use up all available capacity



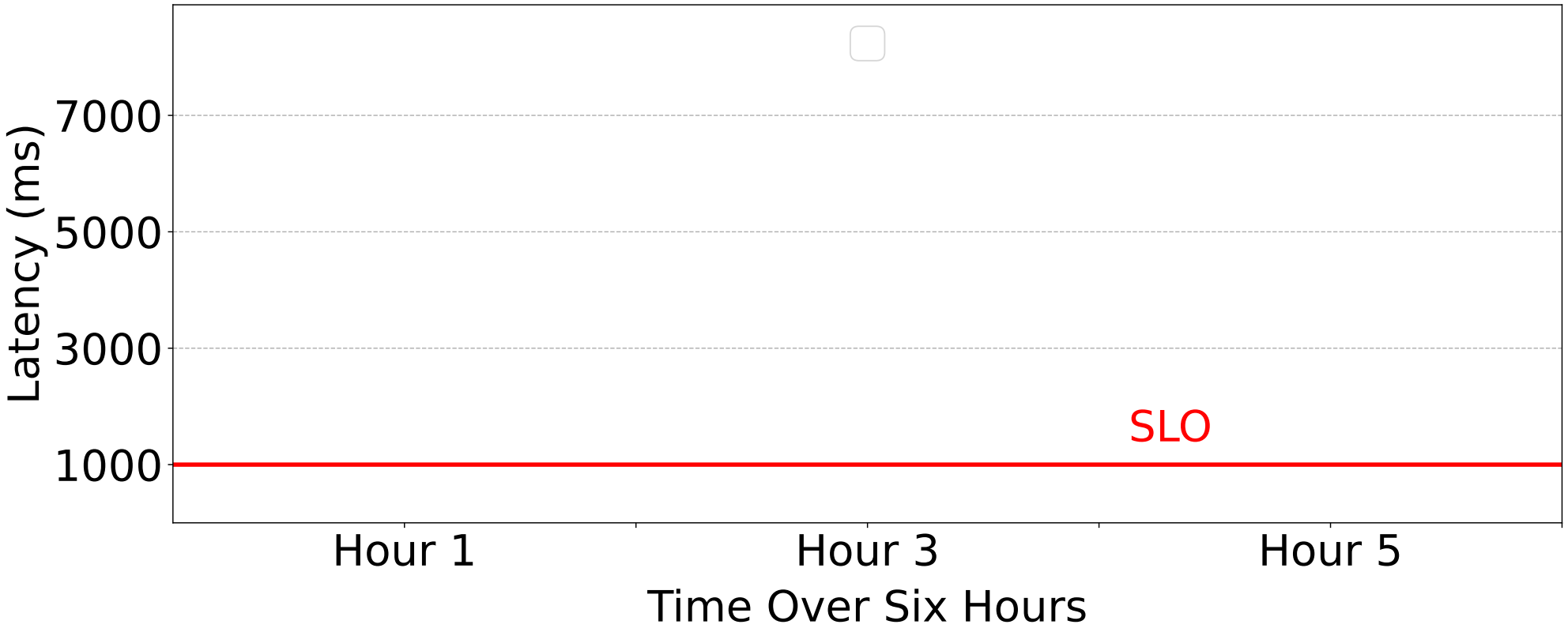
Evaluation – Work Conservation

The low priority workload should use up all available capacity

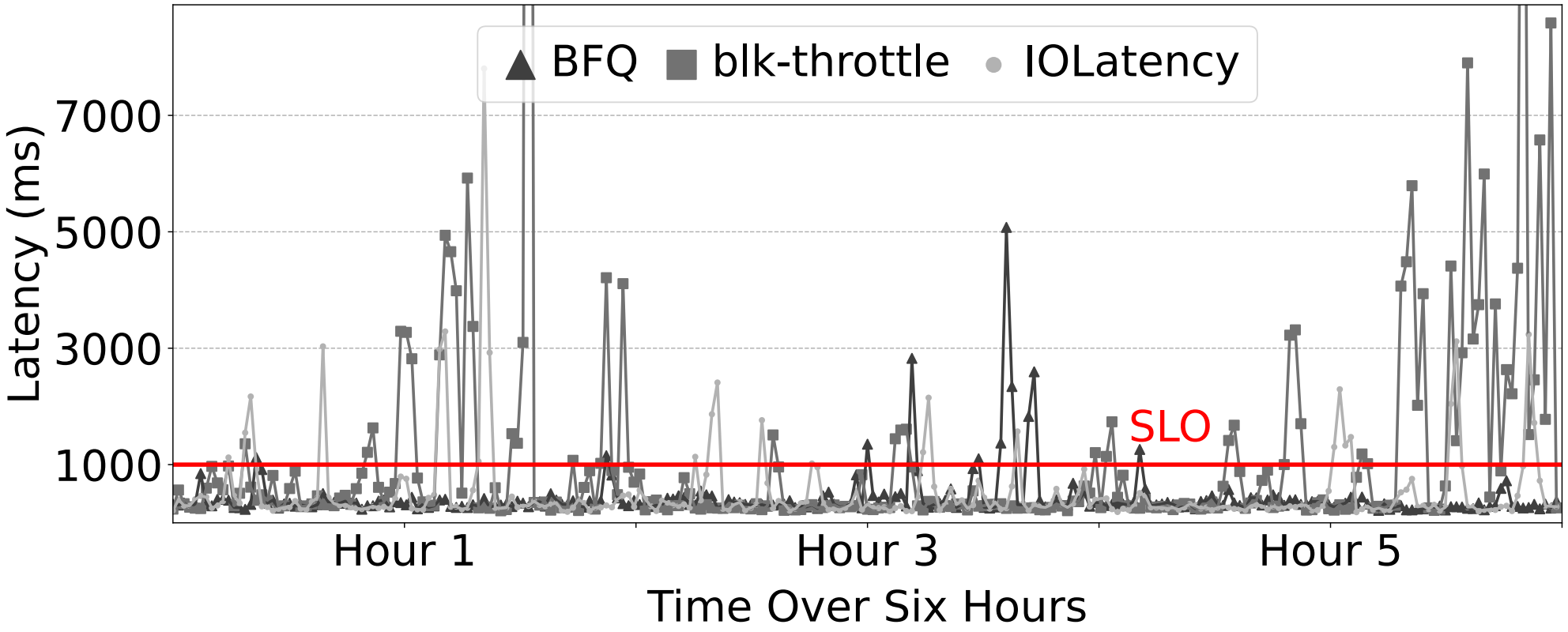


IOCost allows much more aggregate IO

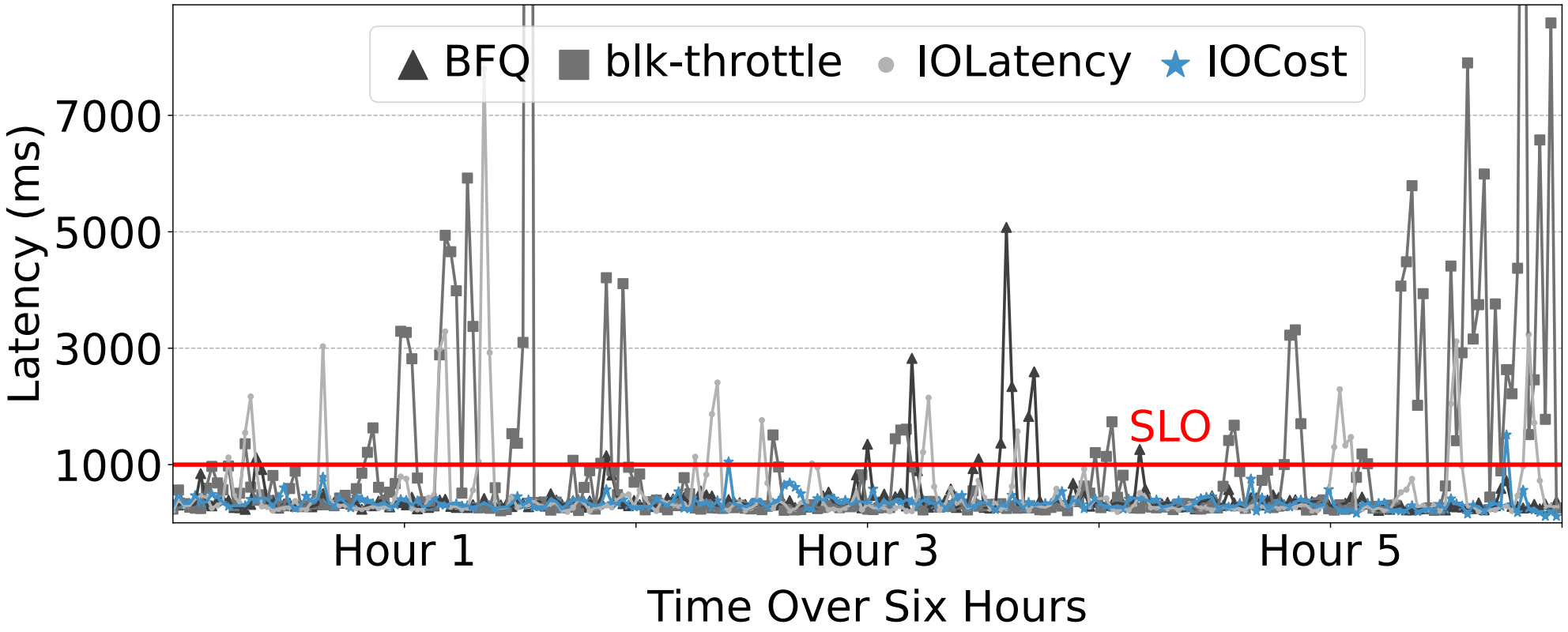
Evaluation – Zookeeper



Evaluation – Zookeeper



Evaluation – Zookeeper



More in the Paper

Memory awareness

Budget donation algorithm

SSD and workload heterogeneity across the Fleet

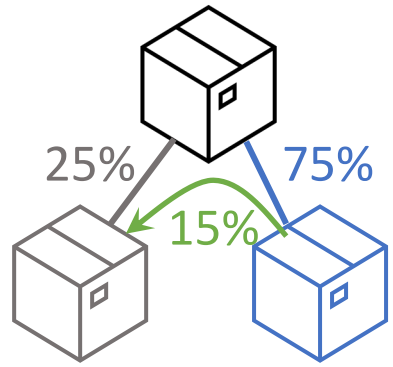
Additional evaluation:

- IO control overhead
- Disk modeling
- QoS and Vrate adjustment
- Overcommitted environments
- Remote storage and VMs
- Package fetching and Container Cleanups

Takeaway→ IOCost
IO Control for Containers in Datacenters

Device occupancy with offline cost and QoS models

Proportional IO control through weights across containers

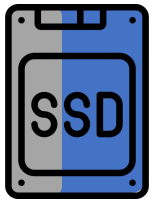


Lightweight work-conserving budget donation algorithm

IOCost manages IO across the entire Meta fleet

Open-source device profiling and benchmarking tools

Upstreamed in Linux



IOCost: Block IO Control for Containers in Datacenters

Tejun Heo, **Dan Schatzberg**, Andrew Newell, Song Liu, Saravanan Dhakshinamurthy, Iyswarya Narayanan, Josef Bacik, Chris Mason, Chunqiang Tang, ‡Dimitrios Skarlatos

Session 5B: Data Center and Cloud Services
Thursday, March 3 @ 2pm

ASPLOS 2022

Meta

‡ Carnegie Mellon University
Computer Science Department