IBM Research

OASIS: Self-tuning Storage for Applications

Kostas Magoutis, Prasenjit Sarkar, Gauri Shah

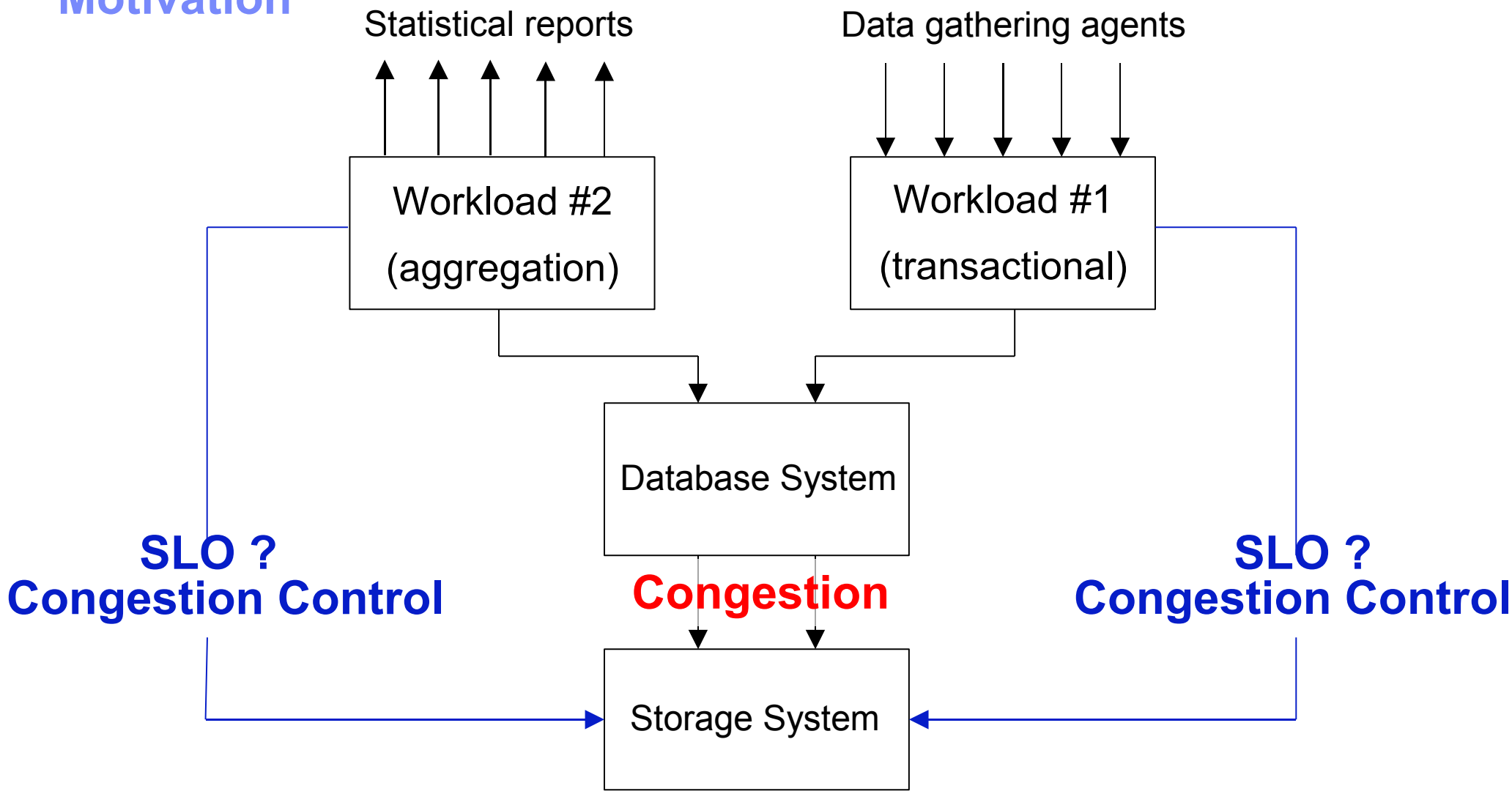
IBM Research

14th NASA Goddard- 23rd IEEE Mass Storage Systems Technologies, College Park, MD, May 17, 2006

Outline

- Motivation and background
- OASIS Architecture
  - Environment
  - Protocol
  - Algorithms
- Evaluation
- Conclusions

Motivation


Statistical reports
Data gathering agents
Workload #2
(aggregation)
Workload #1
(transactional)
Database System
SLO ?
Congestion Control
Congestion
SLO ?
Congestion Control
Storage System

Background

- Allocating storage resources to match application needs is hard
  - Best-effort allocations are often insufficient
- Systems offering Service Level Objectives (SLOs) have drawbacks
  - Administrators cannot give detailed description of storage requirements
  - Instead, they rely on rough estimates based on intuition or past experience
  - Descriptions of storage resource requirements can only be statistical
  - SLOs specified in the largest data centers are often punitive rather than descriptive
- Complexity of applications and storage infrastructure growing
  - Need simple interface between applications and storage infrastructure

Overview of OASIS

- Manages the allocation of storage resources to applications
  - Monitors application requirements and conveys to a storage manager
  - Applies *proportional fairness* mechanisms in resource allocation
  - Improvement over best-effort systems
- Self-tuning: The allocations are set automatically and dynamically
  - Does not require operator involvement
  - OASIS does not conflict (and can co-exist) with SLO-based systems
- Introduces simple interface between applications and storage

Proportional Fairness

- Derives from economic framework
  - Participants are associated with utility functions
  - Suitable choice for utility functions is logarithmic (“diminishing returns”)
  - Optimal allocation maximizes sum of all utilities
  - Applied to network bandwidth sharing by Kelly [1998]
- Resembles fairness achieved by TCP protocol
  - Proportional revocation of resources
- OASIS applies proportional fairness mechanisms to storage systems
- SLO-based storage systems offer performance isolation/differentiation

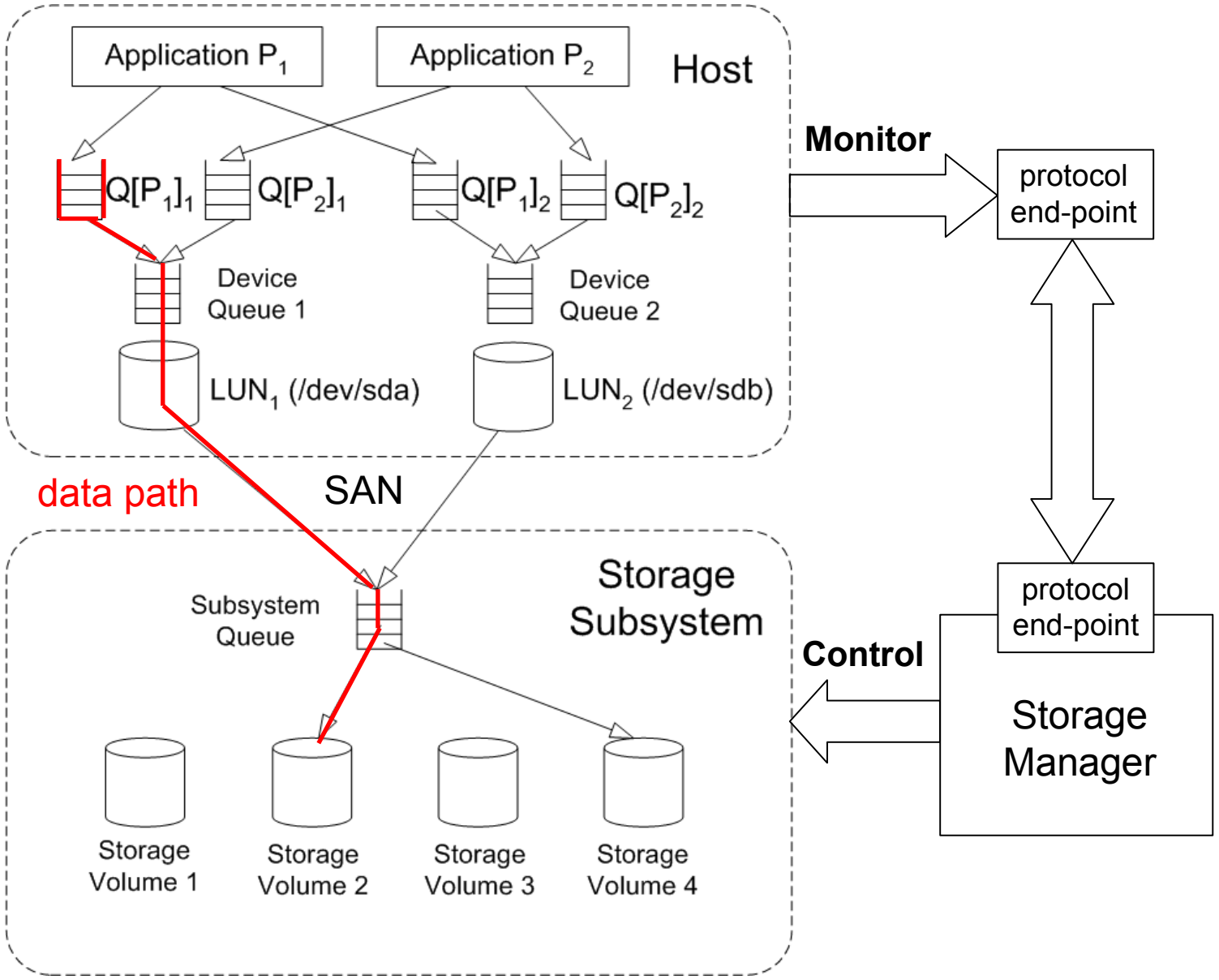
OASIS Architecture

- Environment
  - Multiple host servers running multiple applications
  - Storage access through storage-area network (SAN)
  - Applications, storage are complex domains in different areas of expertise

- OASIS Protocol
  - Insulates the application domain from the storage domain
  - Conveys application requirements to Storage Manager

- Resource Management Algorithms
  - Map application requirements to storage resources
  - Govern allocation of resources subject to proportional fairness

Environment



OASIS Protocol

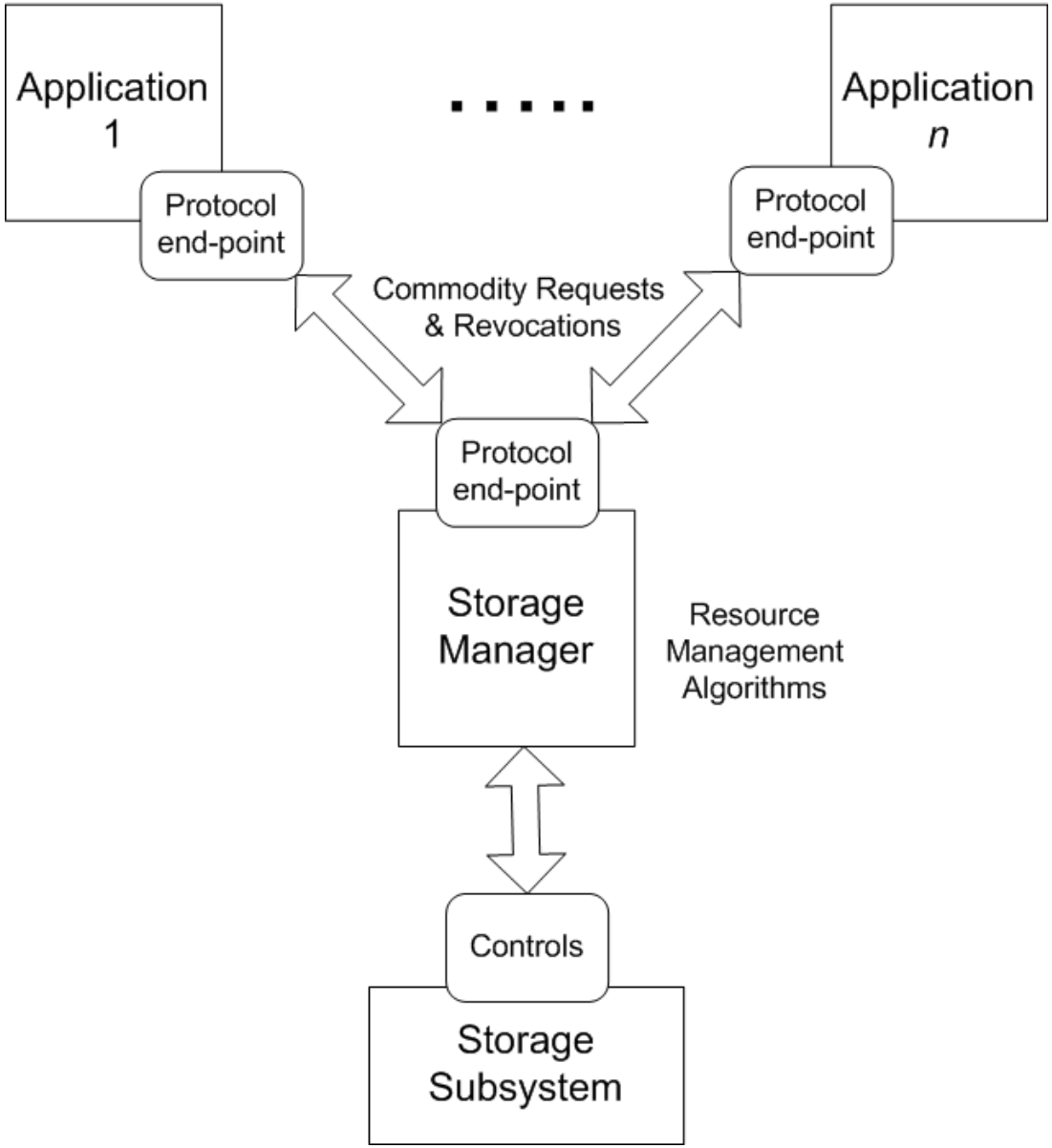
Commodities

- Throughput
- Latency

Mapping

Resources

- Disk bandwidth
- Cache



Application Protocol End-point

- **Functionally distinct from the application**
  - Does not require application modifications or human involvement
  - Operates using standard interfaces (e.g., UNIX iostat, database logs, etc.)
- **Determines application requirements of commodities**
  - For data path sourced at $Q[P_i]_j$ over $\Delta t = (t, t + \delta t)$, app end-point measures
    - $D_{i,j}(\Delta t)$: Average I/O arrival rate
    - $X_{i,j}(\Delta t)$: Average I/O completion rate
    - $R_{i,j}(\Delta t)$: Average I/O latency
    - $S_{i,j}(\Delta t)$: Average I/O size
  - Bandwidth or latency request:
    - $c_{i,j}(\Delta t) = D_{i,j}(\Delta t) * S_{i,j}(\Delta t)$.
    - $l_{i,j}(\Delta t) = R_{i,j}(\Delta t) - R_{i,j}(t)$, if $R_{i,j}(\Delta t) > R$
- **Communicates requirements to Storage Manager protocol end-point**

Storage Manager Protocol Endpoint

- Maps requested commodities to resources
  - E.g., maps throughput, latency to disk bandwidth, cache, etc.
- Identifies data paths involved in commodity requests
  - Data paths connect host I/O queues and storage volumes through SAN
- Manages (allocates, revokes) resources
  - Determines quantity of resources required
    - Resource type #1: Measurable, fixed quantity at time *t*; e.g., cache
    - Resource type #2: Estimable, varies over time; e.g., disk bandwidth
  - Determines whether system resource-constrained
  - Current implementation revokes resources proportionately

Storage Manager Protocol Endpoint: Bandwidth

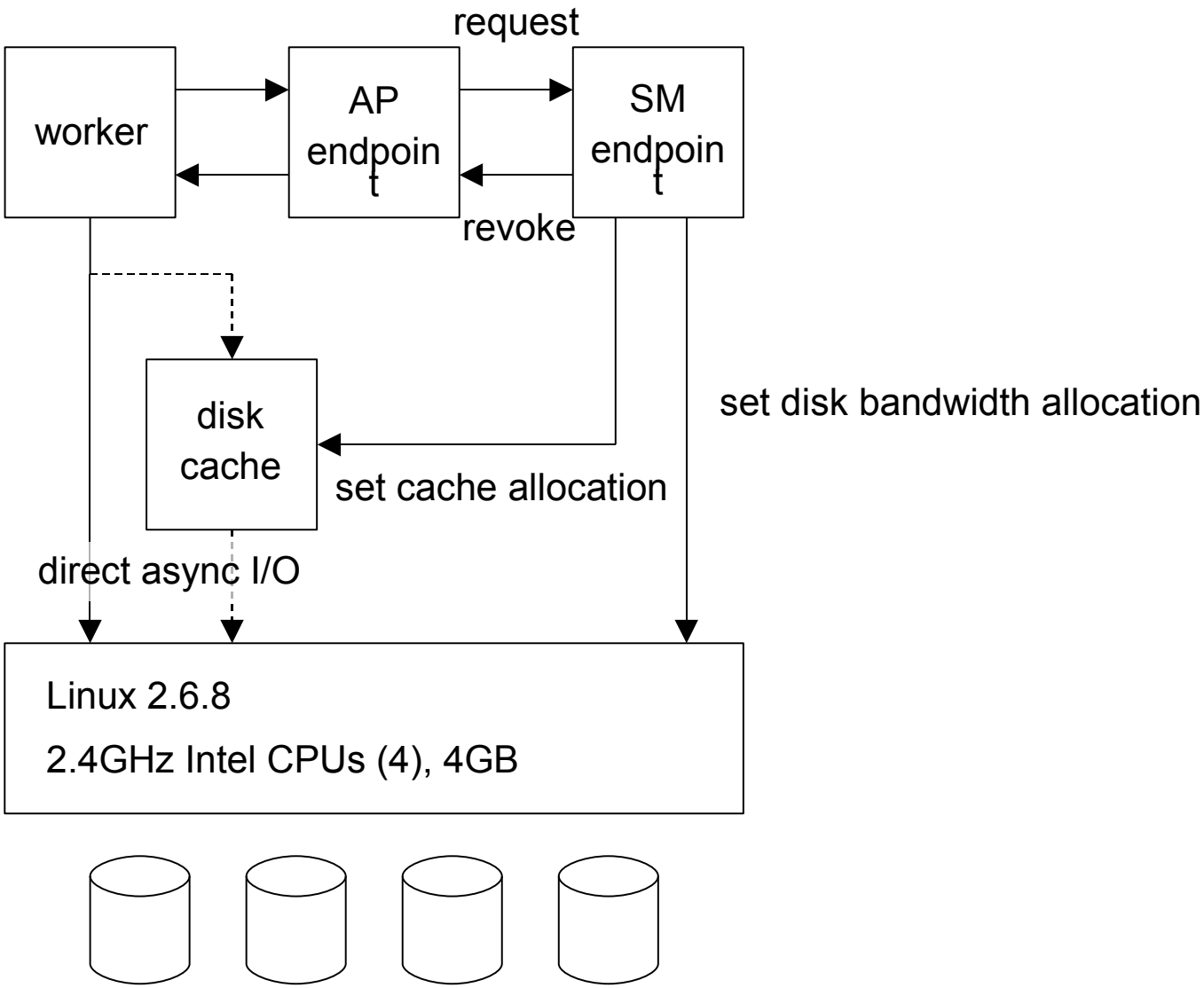
- Current state:
  - Bandwidth allocated to application *i* is A_i
  - Bandwidth used by application *i* is U_i

- Upon receiving bandwidth request *B* from application *p*
  - if $A_p = U_p$ then
    - Allocate bandwidth *B* to *p*, i.e., $A_p = A_p + B$
  - else ($A_p > U_p$)
    - $A = \sum_{i=1..n} A_i$
    - $U = \sum_{i=1..n} U_i$
    - Revoke $A_i - (A_i / A) * U$ from all applications except *p*

Storage Manager Protocol Endpoint: Latency

- Current state:
  - Cache allocated to application *i* is A_i
- Upon receiving request to reduce latency for application *p*
  - Convert request to hit ratio requirement
  - Determine additional cache *C* required to meet new hit ratio
- If cache space *C* not available
  - $A = \sum_{i=1..n} A_i$
  - Revoke $[A_i / (A - A_p)] * C$ from all applications except *p*
- Allocate cache space *C* to *p*

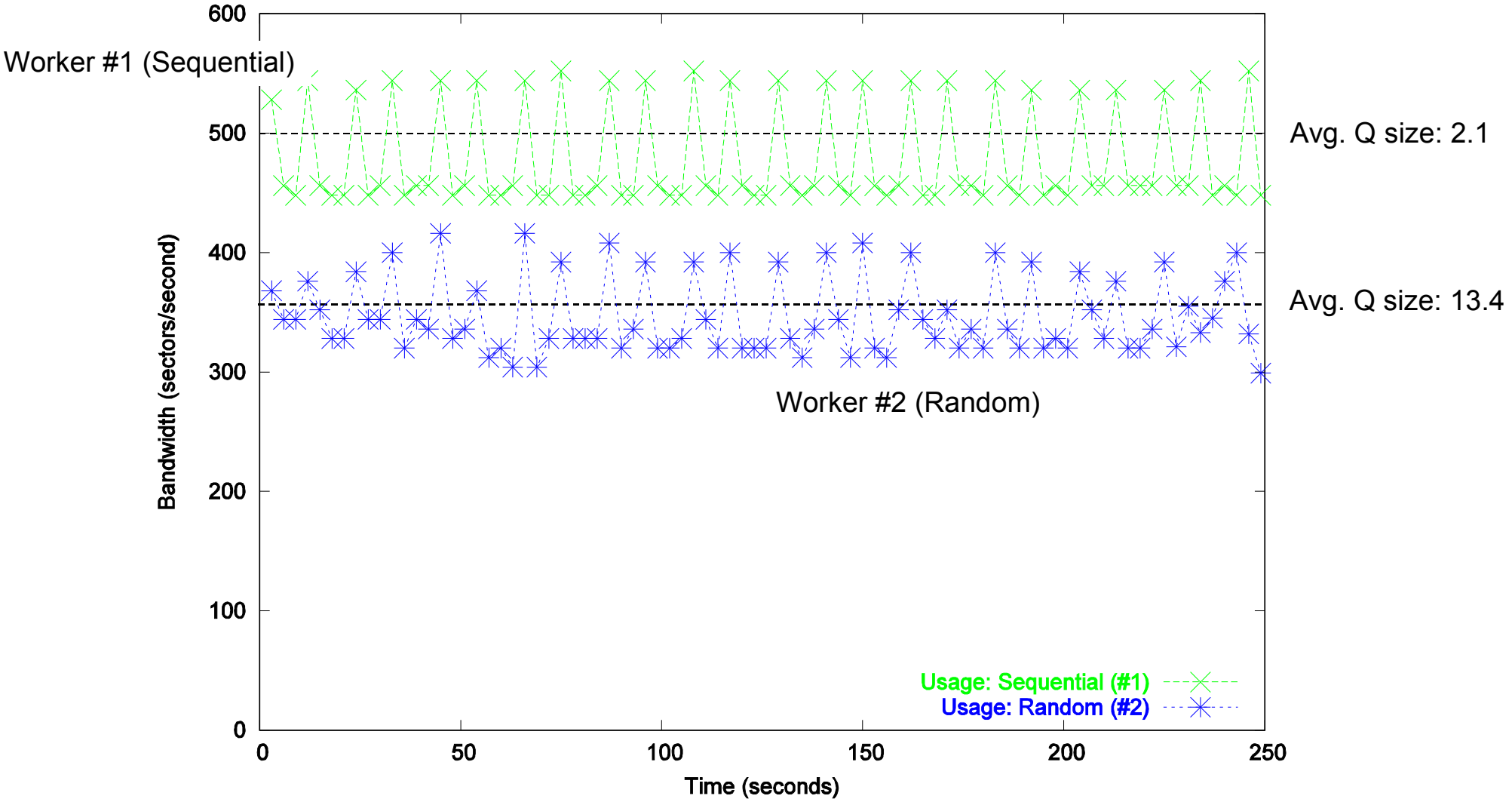
Experimental Setup



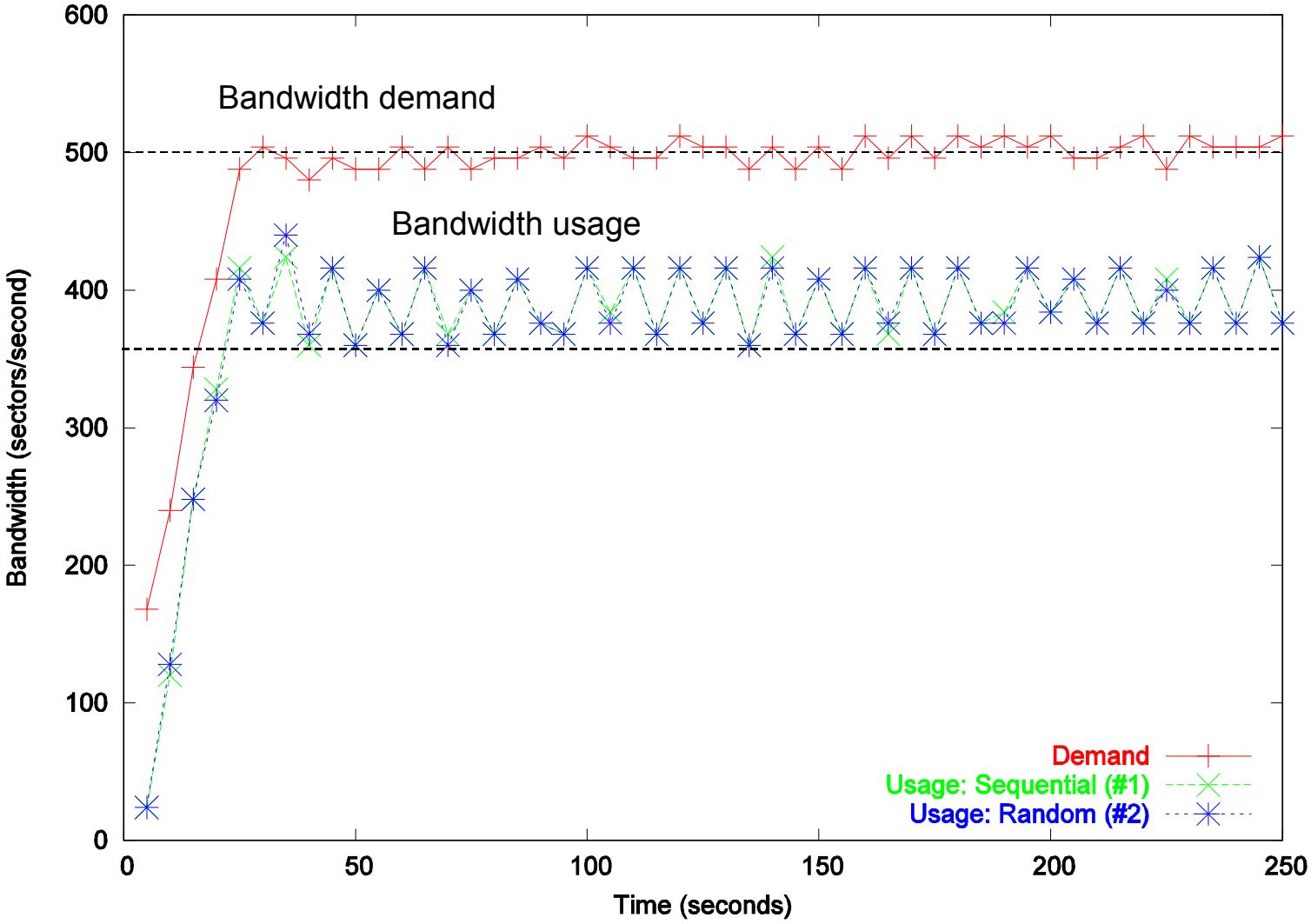
Evaluation: Bandwidth

- Two competing workers
  - worker #1: sequential access
  - worker #2: random access
- Workers identically configured
  - raw (unbuffered) 4KB reads to logical disk device
  - requests produced at constant rate 60 IOPS
  - logical disk saturates at about 100 IOPS for random 4KB I/Os
- Worker #1 overwhelms #2 in a best-effort system

Evaluation: Bandwidth, Best Effort

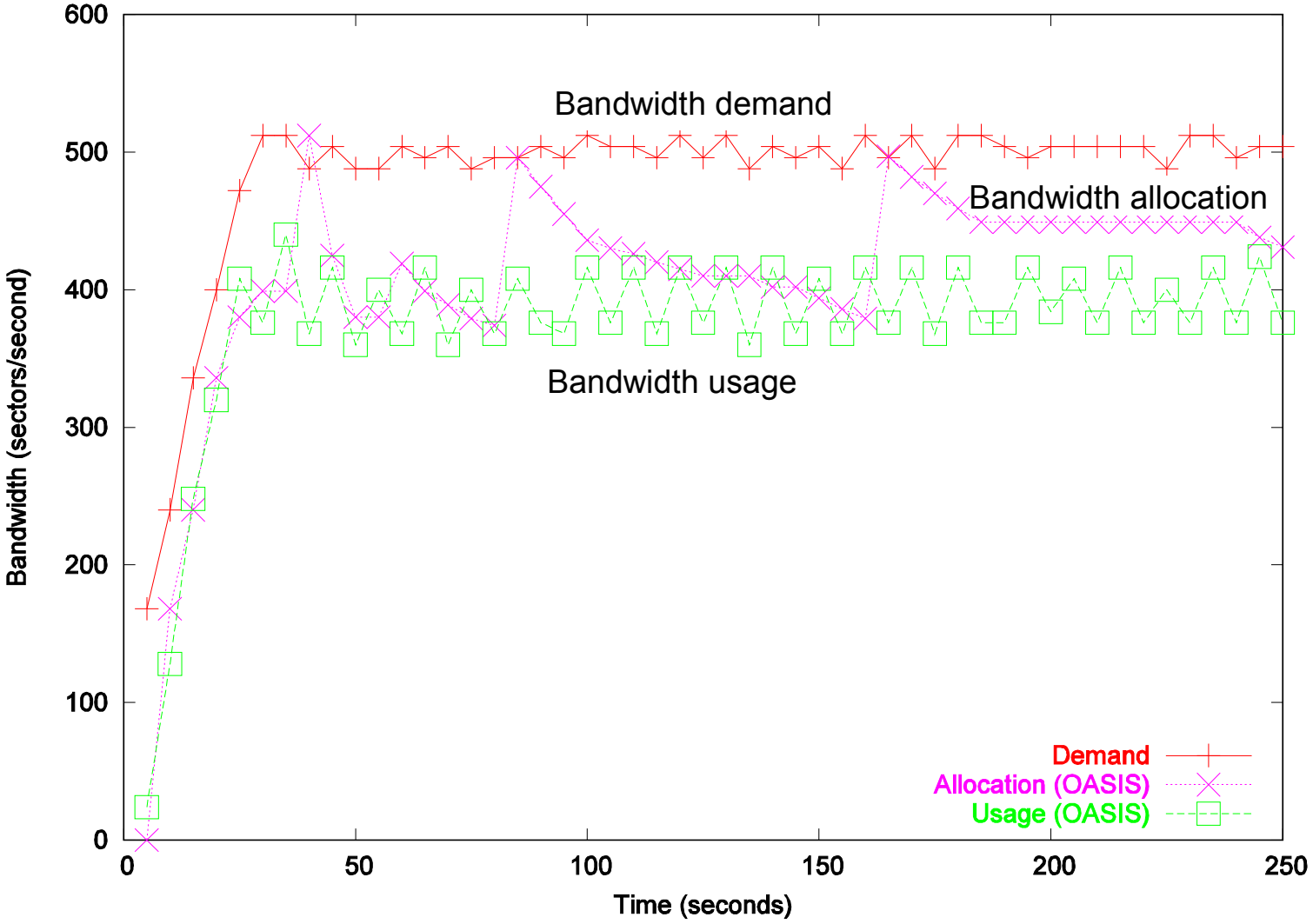


Evaluation: Bandwidth, OASIS



Equal avg. Q size

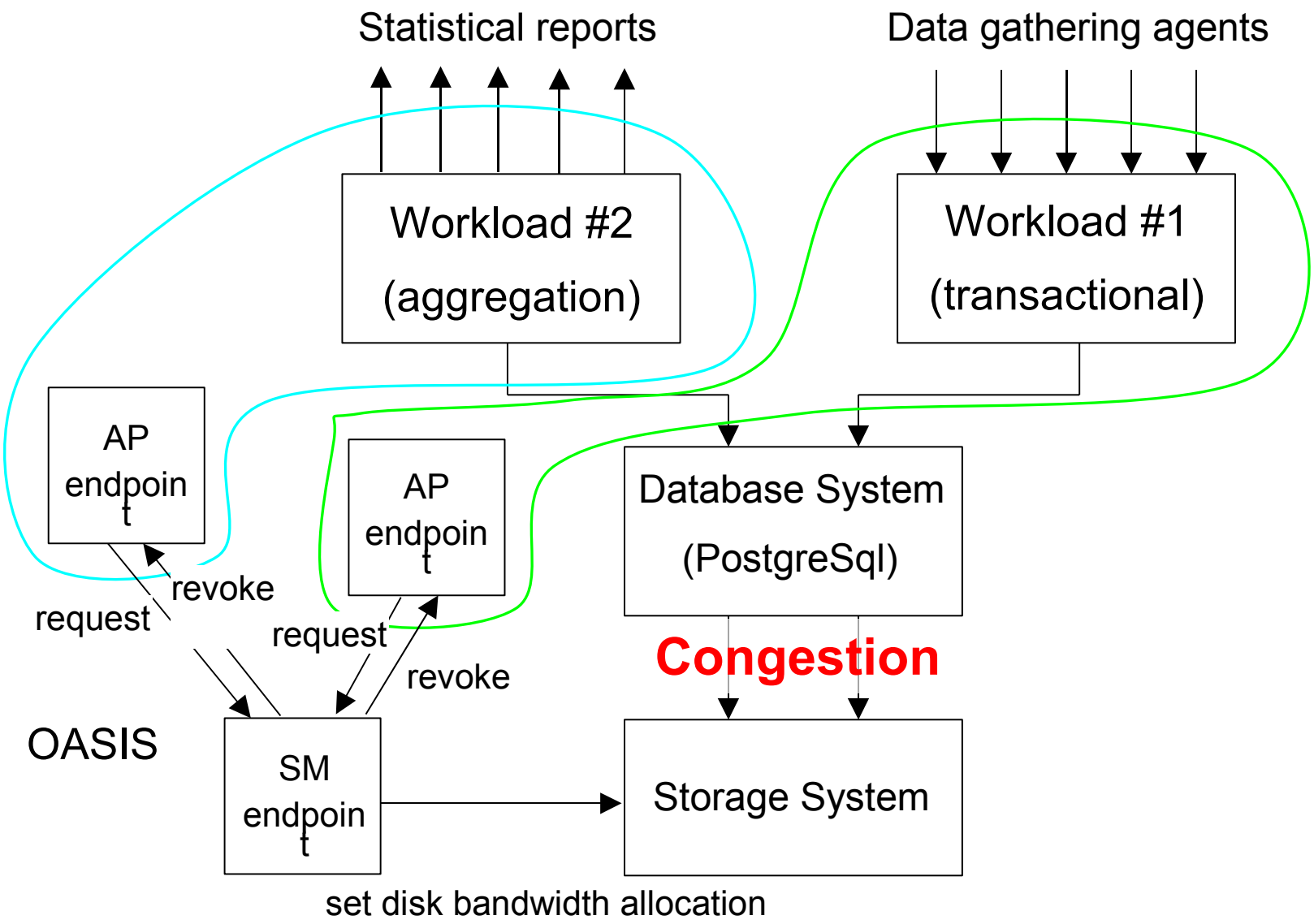
Evaluation: Bandwidth, OASIS Allocation vs. Usage



Evaluation: Latency (Summary)

- For simplicity, the OASIS prototype maps latency to cache via hit ratio
- Cache allocation is triggered by explicit requests for latency reduction
- OASIS achieves faster convergence to a fair allocation of cache
- However, adaptation is slower than in the case of bandwidth due to statistical nature of latency shift estimates

Evaluation: Database Workload



Database Workload

- Transaction workload (#1):
  - Mostly random access
  - Reads: 6%
  - Average I/O size: 0.8 KB
  - Average inter-arrival time: 0.121 ms

- Aggregation workload (#2):
  - Mostly sequential access
  - Reads: 99.9%
  - Average I/O size: 3.9 KB
  - Average inter-arrival time: 0.049 ms

Database Workload (contd.)

Mode	Best Effort				OASIS			
Trans Workload Scaling	Trans Perf (TPS)	Trans Queue Len	Aggr Perf (TPS)	Aggr Queue Len	Trans Perf (TPS)	Trans Queue Len	Aggr Perf (TPS)	Aggr Queue Len
1	49	7.2	26780	0	163	0	20472	0
2	82	15.8	25178	0	360	0	17290	0
4	142	19.5	22933	0	803	0	15400	0
8	235	31.8	20787	0	1029	0	13584	0

- Improvement of up to 5x for transaction workload
- Aggregation workload impacted by about 30-40%

Future work

- Fairness issues: Take efficiency of I/O streams into account?
- Complex mappings of commodities to resources
- Coupling between inter-related commodities
- Re-organization of data layout as a response to overload

Conclusions

- Self-tuning protocol to fairly share storage resources
  - Does not require operator involvement
  - Improves over best-effort systems
- OASIS automatically
  - Detects congestion
  - Ensures underperforming workloads receive fair share of resources
- Simple interface between applications and storage

Backup

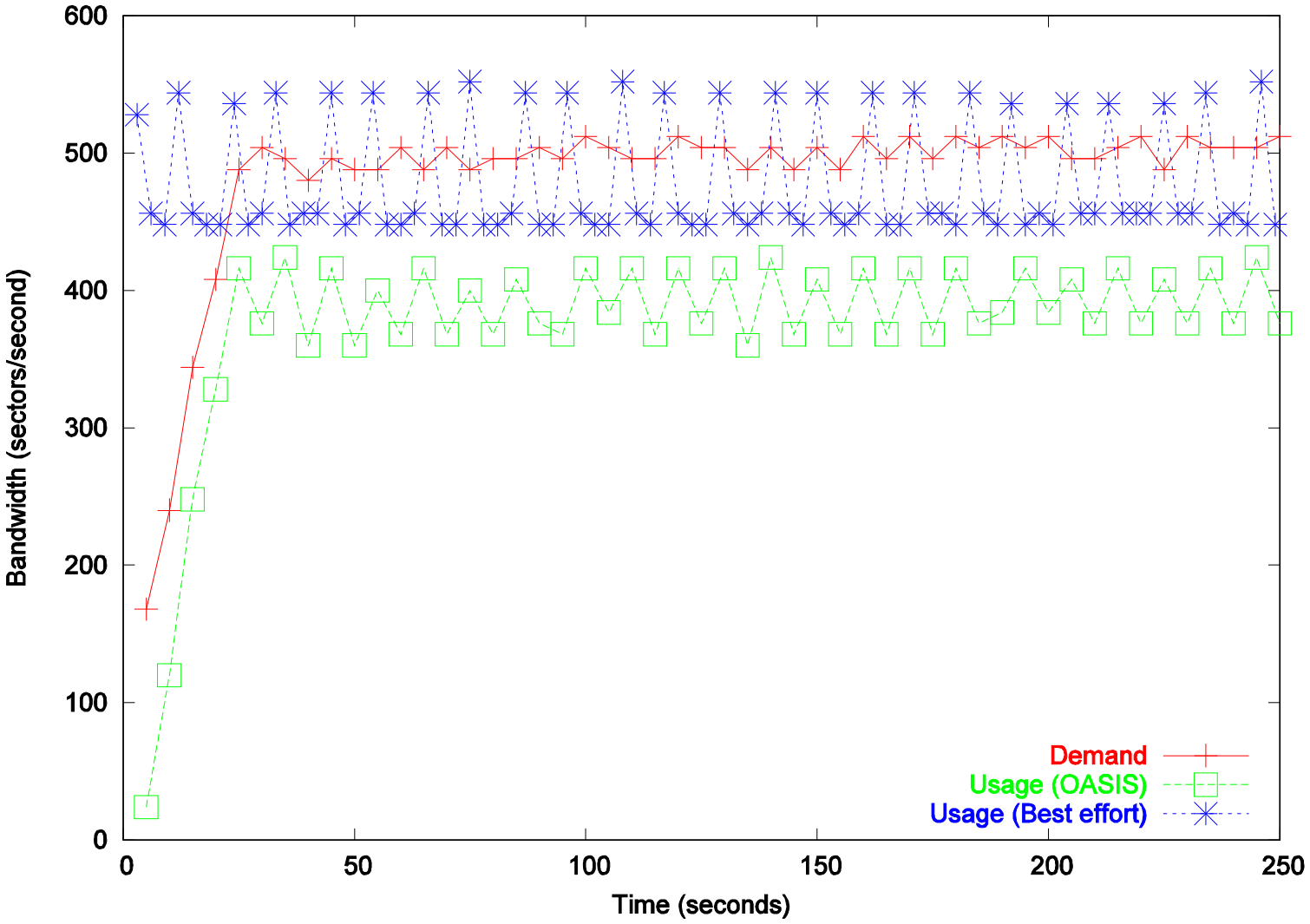
Related work

- “Black-box” systems
  - OASIS maps commodities to storage resources and requires control of the storage subsystem
- SLO-based systems
  - OASIS can co-exist with SLO-based systems
  - OASIS is fundamentally about the protection of overwhelmed I/O flows

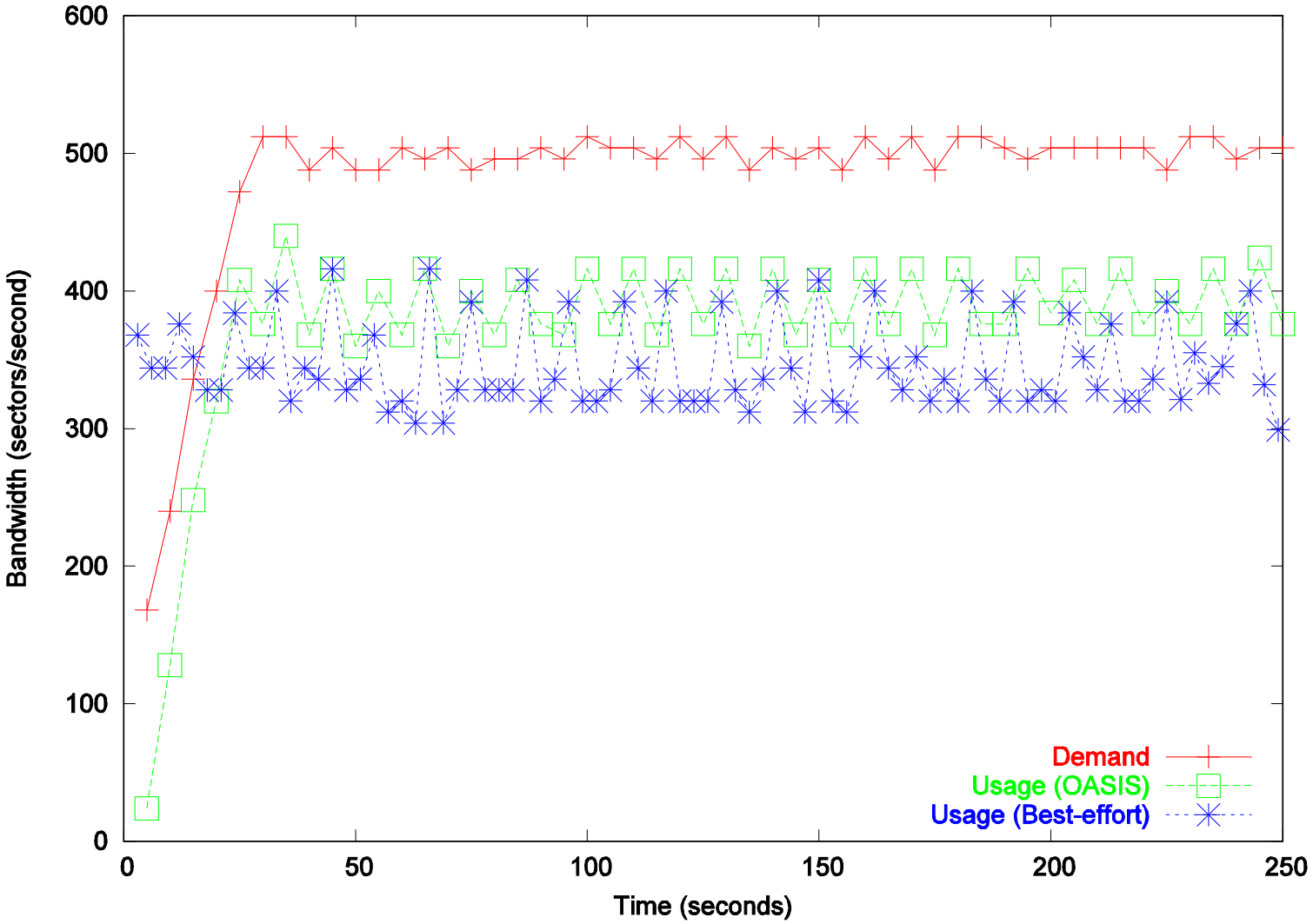
Fairness

- Weighted Fair Queuing (WFQ)
  - $R_i/R_j = p/q$
  - Performance isolation & differentiation
- Economic frameworks
  - Based on utility functions
    - Proportional Fairness [Kelly98]
- TCP
  - Additive increase/multiplicative decrease

Evaluation: Bandwidth, Worker #1 (Sequential)



Evaluation: Bandwidth, Worker #2 (Random)



Evaluation: Bandwidth, Worker #1 OASIS Allocation vs. Usage