Storage QoS Guarantee

Tzi-cker Chiueh

Experimental Computer Systems Laboratory
Computer Science Department
Stony Brook University

The Team

- PI: Tzi-cker Chiueh
- Ph.D. students: Maohua Lu and Shibiao Lin
- Collaborator: Dharmesh Satish (Symantec Research Labs)
- Project page: *http://www.ecsl.cs.sunysb.edu/stonehenge/index.html*

QoS Scheduling Theory

- Given a workload specification (e.g. input rate and maximum input burst size) and a performance requirement (e.g. delay, bandwidth, jitter), a given real-time request scheduling algorithm (i.e. weighted fair queuing or WFQ) fully determines
  - Correlation between bandwidth reservation and worst-case service delay
  - Criterion on when to admit a new reservation (admission control)

Applying This Theory to Storage

1. How to integrate traditional efficiency-driven disk scheduler with QoS-driven disk scheduler
2. How to accurately and fairly account for non-data-transfer disk service overhead in real-time request scheduling algorithm
3. How to maximize disk resource utilization while guaranteeing each virtual disk's QoS requirement
   - How to exploit statistical multiplexing to increase the number of virtual disks admitted without violating bandwidth and delay guarantee
   - How to accommodate the fact that input workloads cannot be fully characterized a priori

Disk Resource Scheduler

- High Disk Bandwidth Utilization
  - Candidates: SATF, CSCAN, etc.
- QoS/SLA Guarantee: more than just prioritization
  - Satisfy requests' deadlines or delay bounds: mainly focus on queuing delay
  - Fair bandwidth allocation among VDs
  - Candidates: Delay-EDD, Weighted Fair Queuing (WFQ), Virtual Clock (VC), etc.
- Our choice: Integration of VC (based on physical time rather than virtual time) and CSCAN

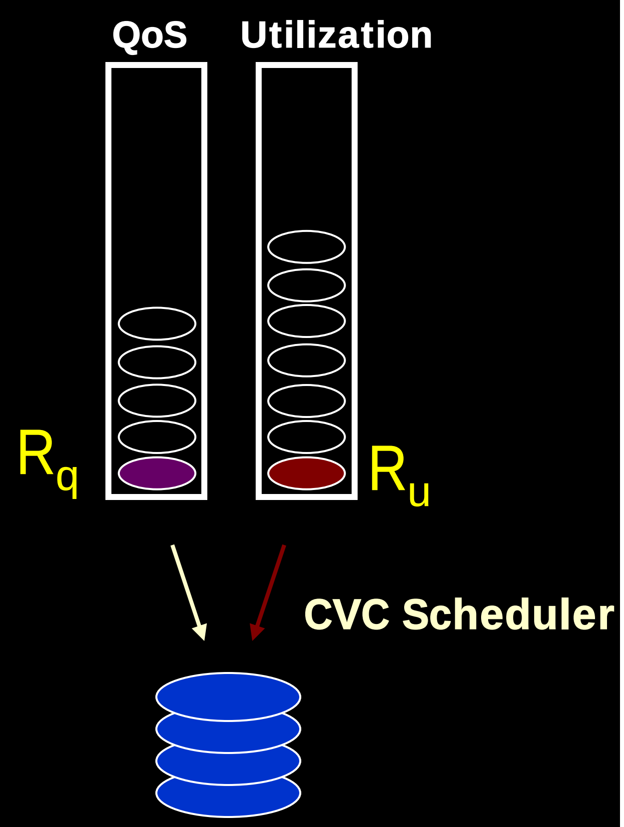
CSCAN-based Virtual Clock (CVC) Scheduler

- Two queues
  - QoS: ordered by latest start time (LST)

    FT(i) = max(FT(i-1), arrival_time) + normalized_service_time

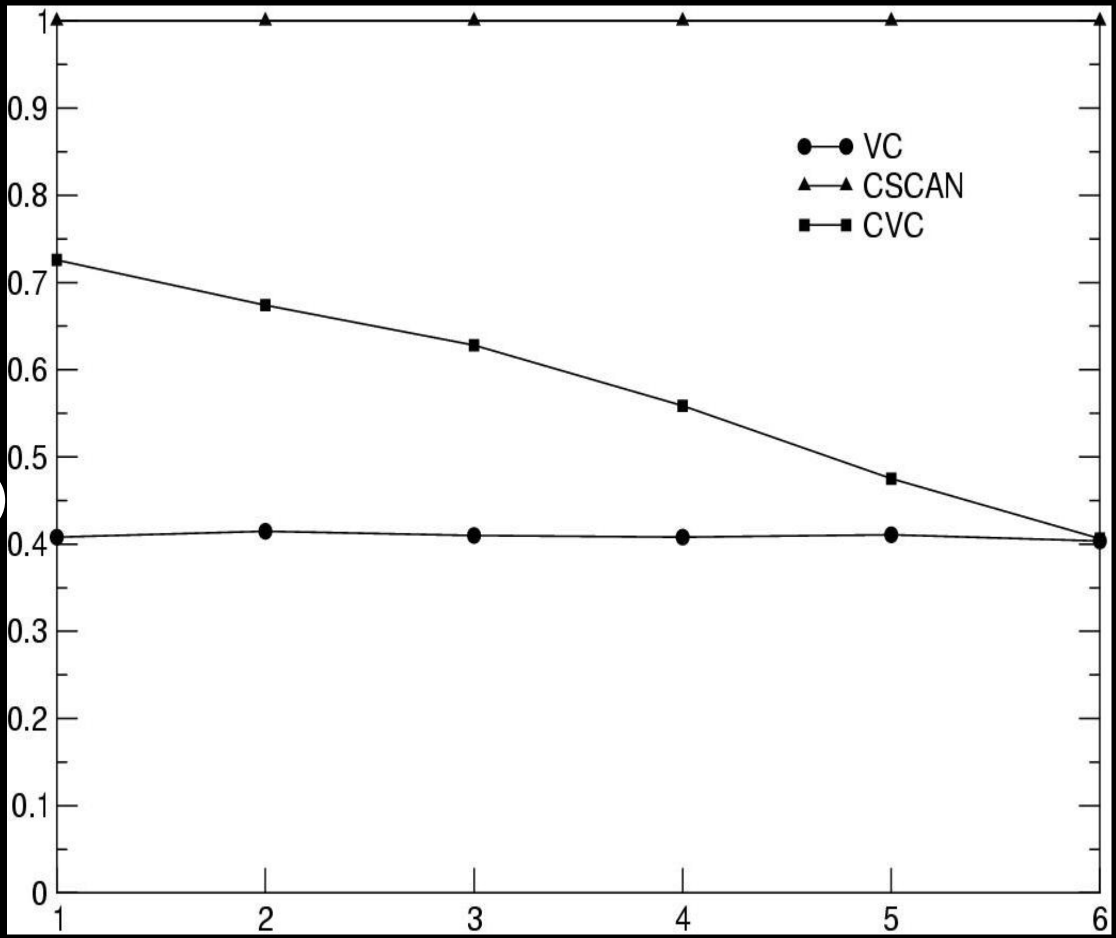
    LST(i) = FT(I) – physical_service_time
  - Utilization: ordered by disk request's target position
- Request from utilization queue is dispatched only if:

  Current_time + service_time(R_u) < Latest_start_time(R_q)



CVC's Utilization Efficiency

Normalized Disk Bandwidth Utilization (Video Stream Trace)



Load Index

Normalized Service Time

Finish_Time(i) = max(Finish_Time(i-1), arrival_time) + normalized_service_time

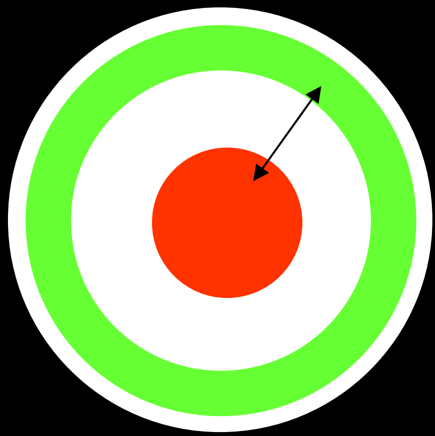
normalized_service_time = request_size / reserved_bandwidth

reserved_bandwidth = reserved_transfer_bandwidth/ $(1+\alpha)$

$\alpha =$ percentage of non-transfer-delay overhead

Virtual Disk Switching Overhead (VDSO)

- Multiplexing multiple VDs on the same physical disk(s) incurs additional overhead, which, like tax, should be distributed fairly among the sharing VDs
- Without fair attribution of VDSO, VDs with better locality suffer more when multiplexed with other VDs



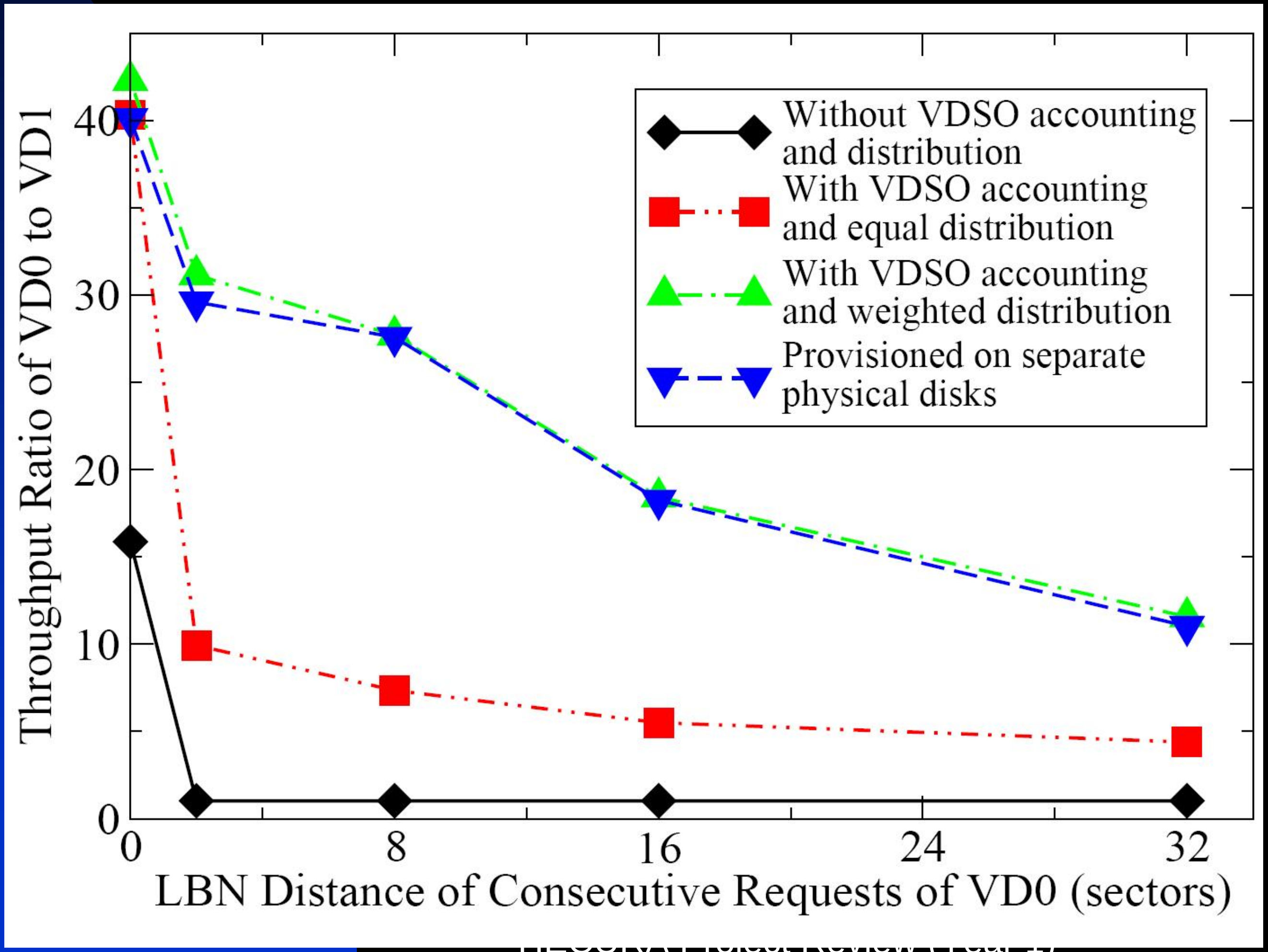
Ideal Fair Attribution of VDSO

- Goal: Throughput ratio between virtual disks multiplexed on the same physical disk should be the same as if they are serviced by separate physical disks

Distribution of VDSO

- Distributing VDSO proportional to total IOH of each individual virtual disks
  - $AVDSO_i = VDSO * IOH_i / \Sigma(IOH_j)$
  - $\alpha_i = IOH_i + AVDSO_i$
- Correctness:
  - $(IOH_i+AVDSO_i) / (IOH_j+AVDSO_j) = IOH_i/IOH_j$

Evaluation of Fair Attribution of VDSO

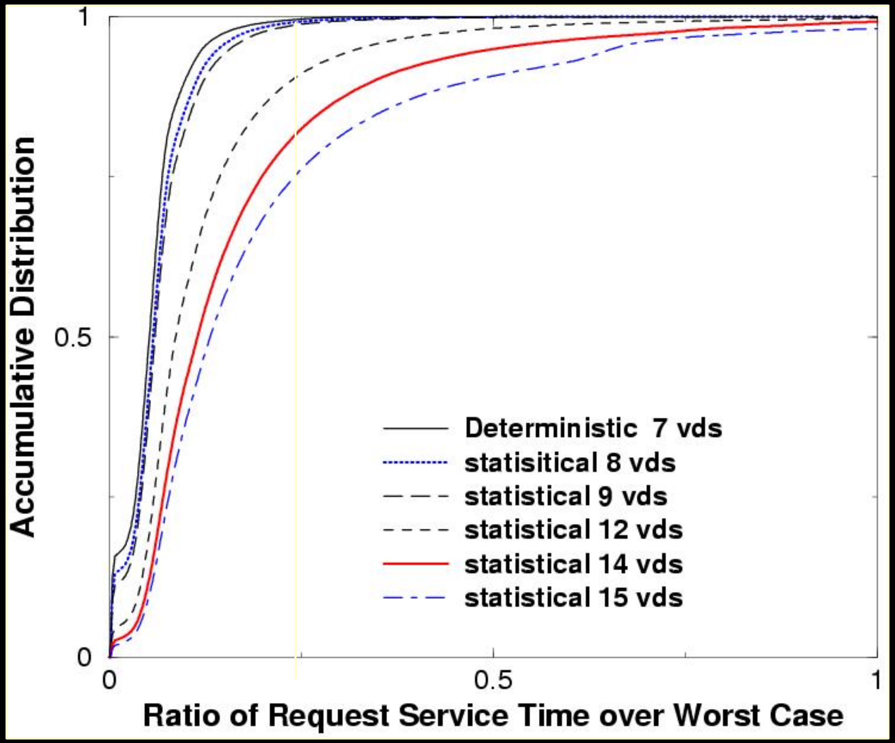


Delay Measurement-Based Admission Control

- Big deal: How to exploit statistical multiplexing while supporting (probabilistic) delay guarantees?
- Key idea: Ratio of measured delay and delay bound
- Deterministic delay guarantee vs. statistical delay guarantee with probability 1

Service Delay Measurement: $P_{service}$

- With a probability E_i, the actual delay bound of the i-th VD is $P^{-1}_{service}(E_i)$ of its original delay bound



Key Idea

- Fact: Given a bandwidth reservation B, empirically 90% of the requests experience a delay that is less than 25% of worst_case_delay(B)
- Deduction: To guarantee that at least 90% of requests experience a delay less than worst_case_delay(B), the bandwidth reservation required is the one whose corresponding worst_case_delay is 4 (=1/0.25) times of worst_case_delay(B)

MBAC Performance – Latency Bound

Run	VD Type	Probability	Deterministic	MBAC	Oracle
1	Financial	95%	7	20	22
2	Mixed	95%	7	14	14
3	Mixed	85%	7	17	17

Resource Reservation

Next Steps

- Virtual clock algorithm is long-term fair, but its short-term unfairness can be unbounded ➔ Need a disk scheduling algorithm that can trade off short-term fairness, long-term fairness and disk resource utilization efficiency
- Distributed disk resource scheduling across a fault-tolerant and load-balancing storage server cluster
- Integrate multi-dimensional storage virtualization technology with CPU/memory virtualization technology to build a complete virtual machine resource management system

Questions?

Thank You!

chiueh@cs.sunysb.edu

Publications

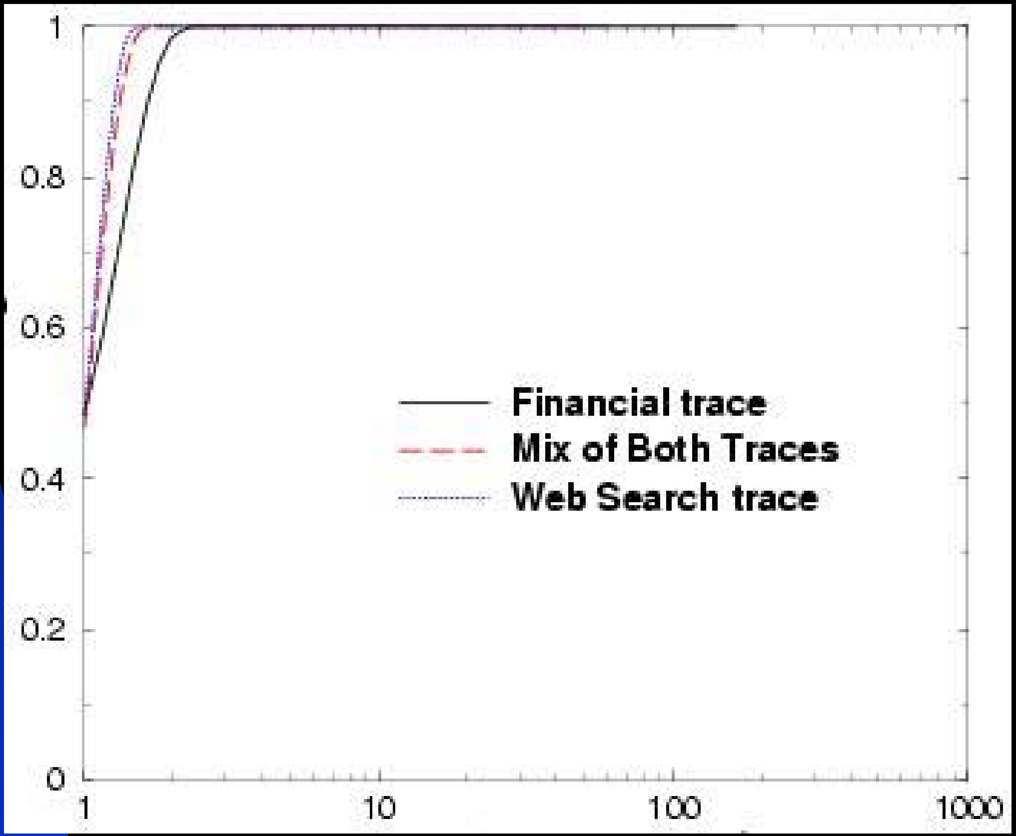
- Gang Peng, "Availability, Fairness, and Performance Optimization in Storage Virtualization Systems", Ph.D. Dissertation, Computer Science Department, Stony Brook University, October 2006.
- Ningning Zhu, Tzi-cker Chiueh, ``Portable and Efficient Continuous Data Protection for Network File Servers,'' in the 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, July 2007.
- Shibiao Lin, Maohua Lu, Tzi-cker Chiueh, ``Transparent Reliable Multicast for Ethernet-Based Storage Area Networks,'' in the 6th IEEE International Symposium on Network Computing and Applications, July 2007.
- Maohua Lu, Shibiao Lin, Tzi-cker Chiueh, ``Efficient Logging for Comprehensive Data Protection,'' in the 2007 IEEE Mass Storage and Systems Technology Symposium, September 2007.

Extraction of VDSO

- Inherent Overhead (IOH) of a VD tracks the VD's workload locality
- Only disk head movement counts
  - Need to detect disk cache miss
- Req N is Request X in VD_i, Req N+1 is Request Y in VD_j
- $VD_i \neq VD_j$
  - Req Y close to Req Y-1 – overhead attributed to VDSO
  - Otherwise – overhead attributed to VDSO and VD_j
- $VD_i = VD_j$
  - Attributed to IOH of VD_j

Spare Bandwidth Distribution: P_{spare}

Probability



No. of Virtual Disks

Measurement-based Admission Control (MBAC)

- The j^{th} VD: (B_j, C_j, D_j, E_j)
- Calculate $B_{i,latency}$ for $0 < i <= j$

  $D_i <= P^{-1}_{service}(E_i) * [(N+1) / IOPS_i + 1/IOPS_{full}]$
- Check if

  $\Sigma\ MAX(B_i, B_{i,\ latency}) <= IOPS_{full}$
- If the above inequality holds, accept the j_{th} VD; otherwise, reject it

Exploiting Statistical Multiplexing

- Delay bound of virtual clock scheduling

  $DB_i = (N+1)/IOPS_i + 1/IOPS_{full}$

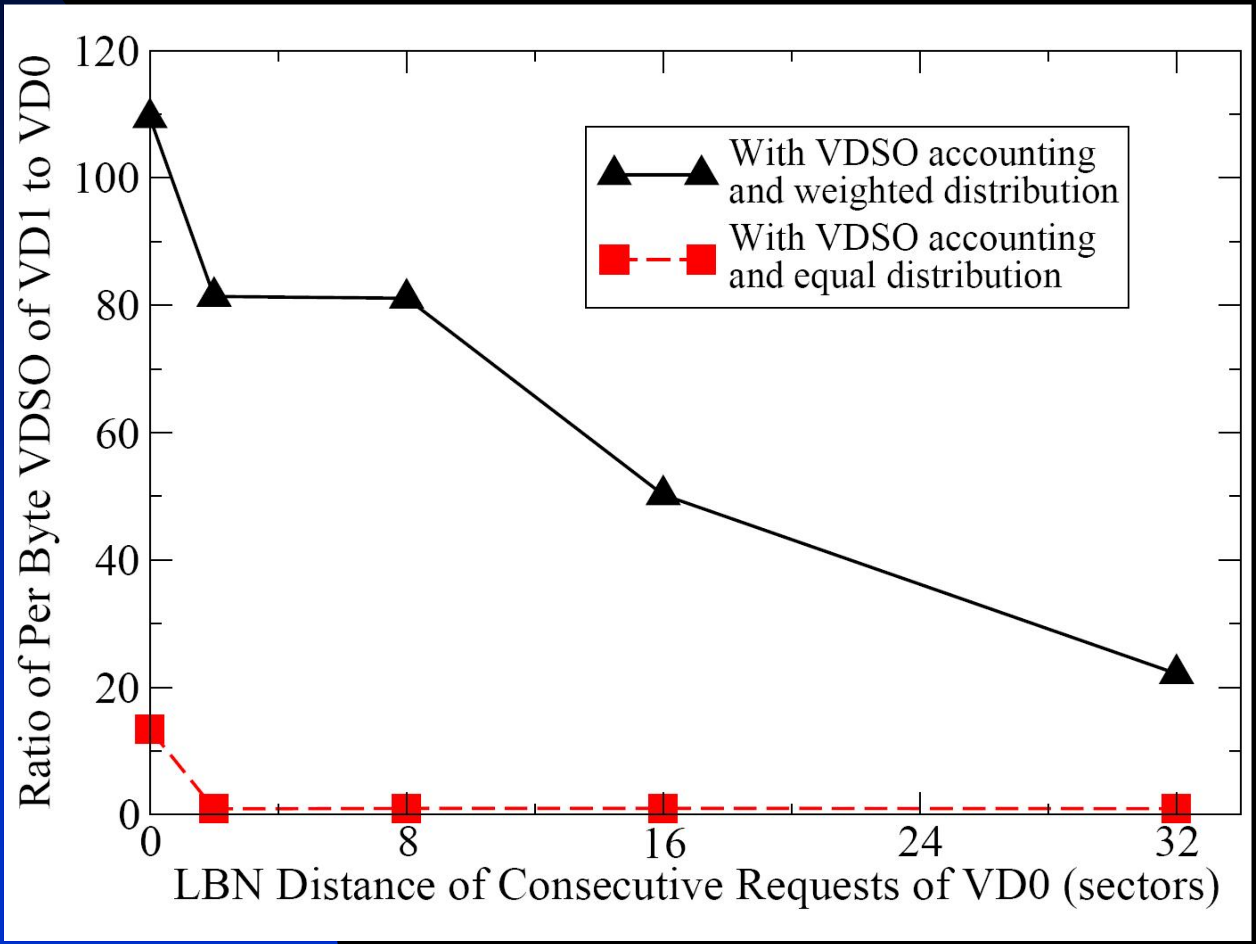
  DB_i : i-th VD's delay bound   N: burst length

  $IOPS_i$: i-th VD's bandwidth reservation

  $IOPS_{full}$: Measured physical disk array's raw bandwidth in I/Os/sec

- Observation: Worst-case delay rarely happens, so bandwidth reservation to achieve a certain delay bound can be reduced
- Why?
  - Not all resources are reserved
  - Not all resources reserved are used

Evaluation of Fair Attribution of VDSO



Dealing with Unknown Workload Features

- Request size (N) and read/write ratio (*fw*) affect resource reservation but are unknown at admission control time
- To use measurement to correct resource over-provisioning
  - Worst-case reservation first
  - Use MBAC to adjust reservation later on based on actual usage measurements at run time