BASIL: Automated IO Load Balancing across Storage Devices

Ajay Gulati, VMware, Inc.

Irfan Ahmad, VMware, Inc.

Chethan Kumar, VMware, Inc.

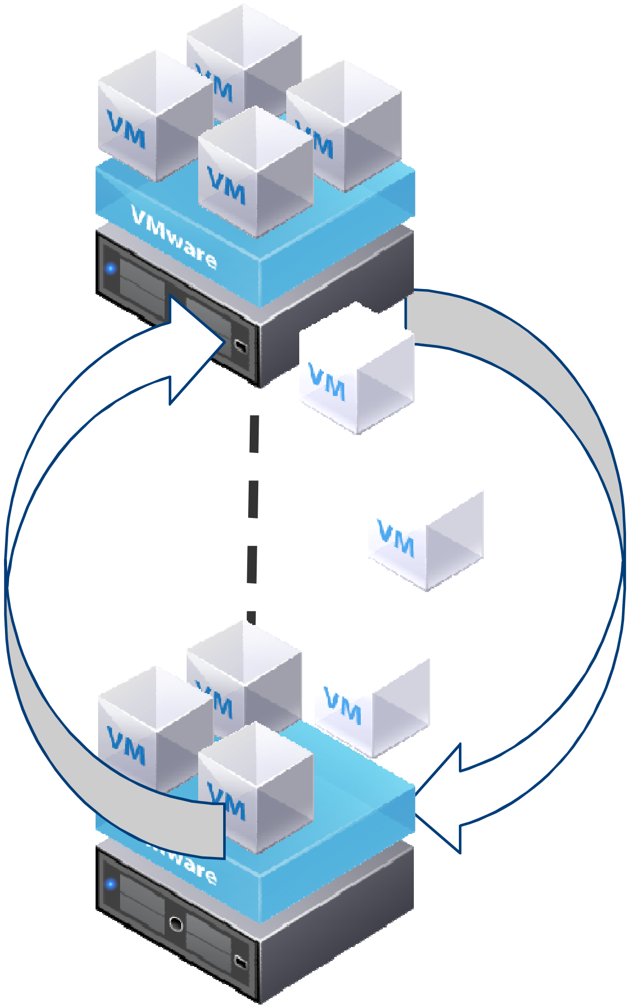
Karan Kumar, CMU

USENIX FAST – February 25, 2010

Outline

- **Problem Description & Motivation**
- BASIL – Modeling & Load Balancing
- Experimental Framework & Results
- Conclusions & Future Work

Datacenter Automation—State of the Art

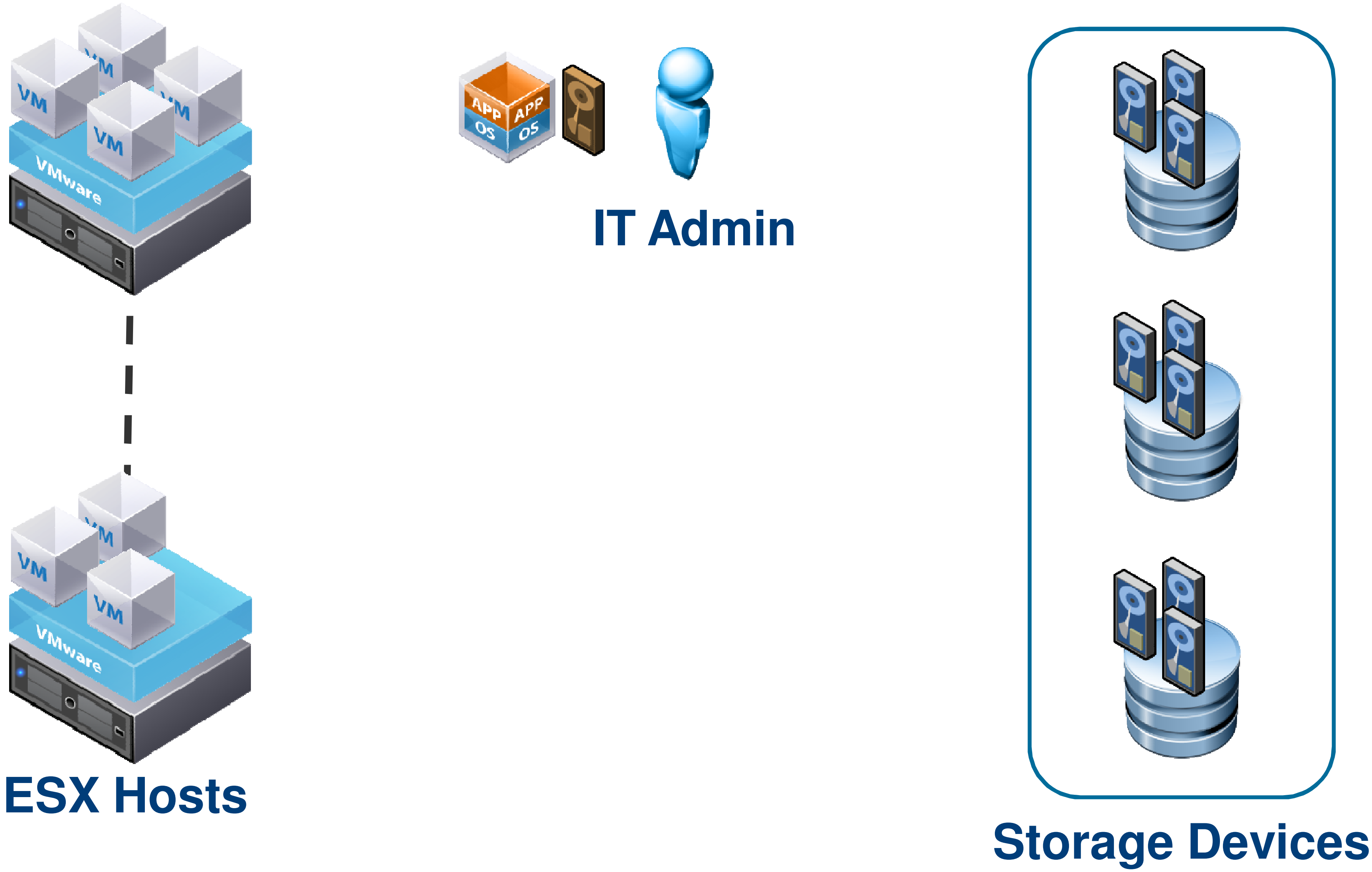


ESX Hosts

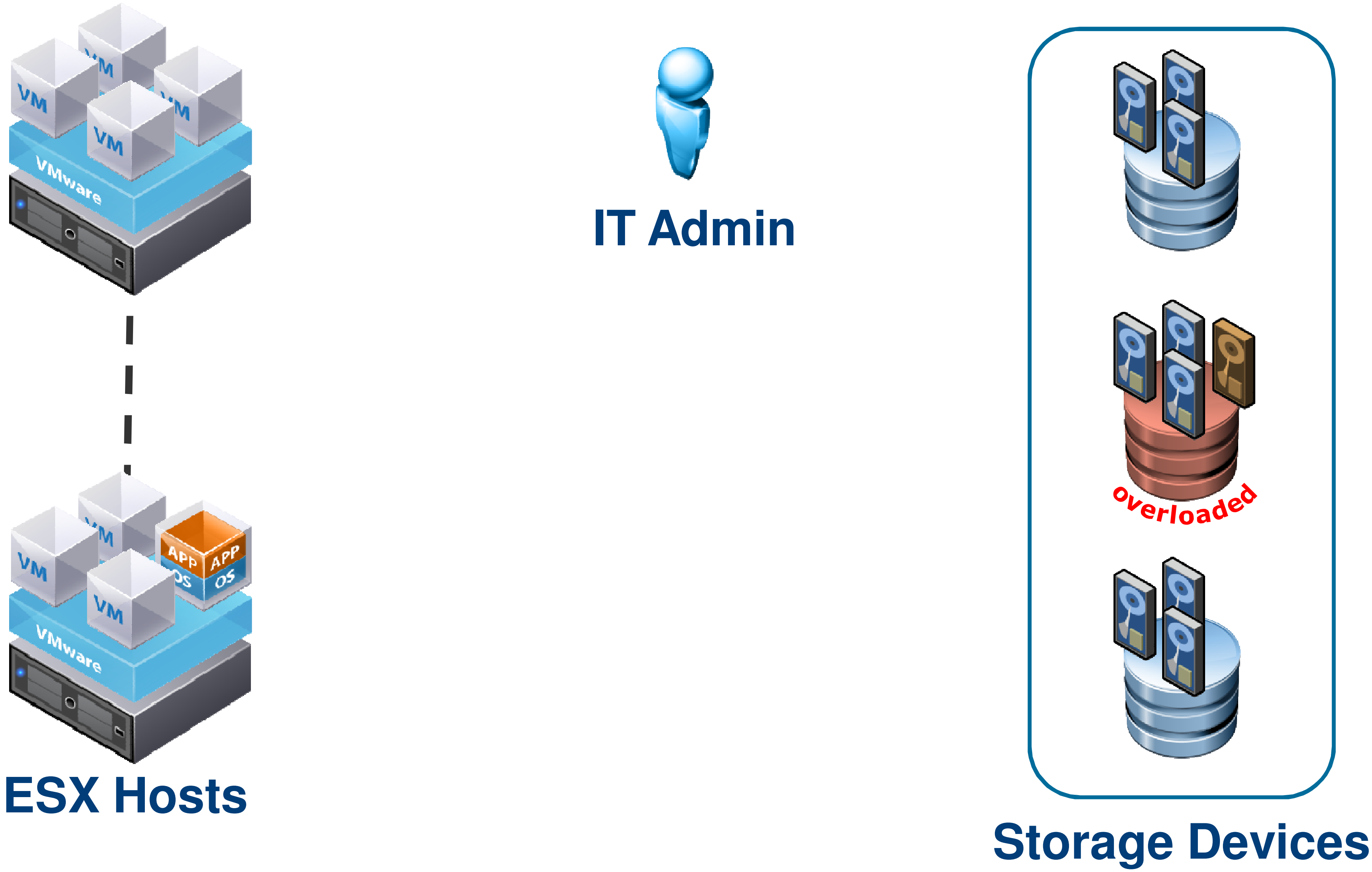
Automated Load Balancing of CPU and Memory resources across a cluster of servers using **live migration.**

e.g., **VMware DRS** (Distributed Resource Scheduler)

The Problem—Storage Management Not Automated

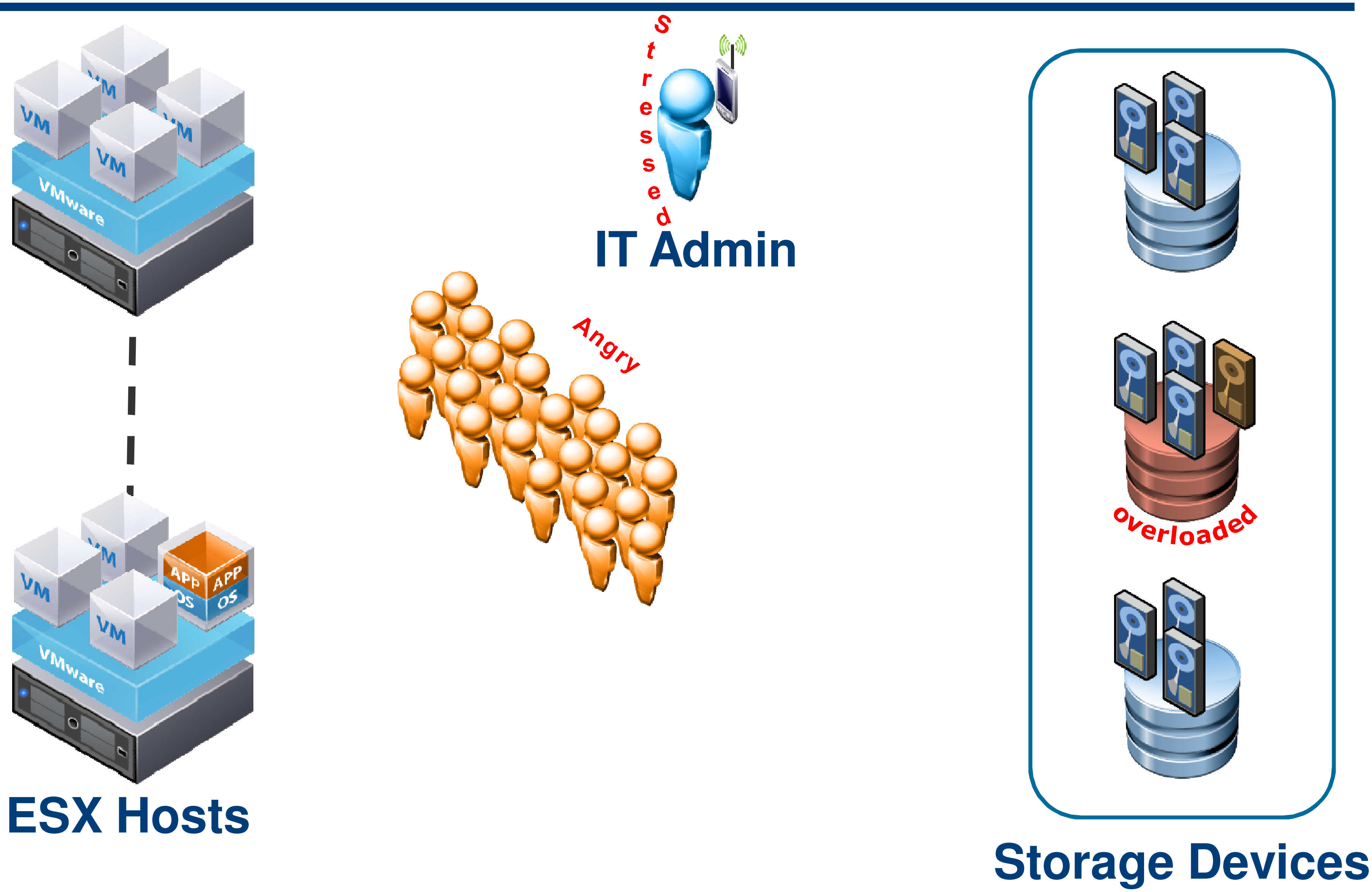


The Problem—Storage Management Not Automated

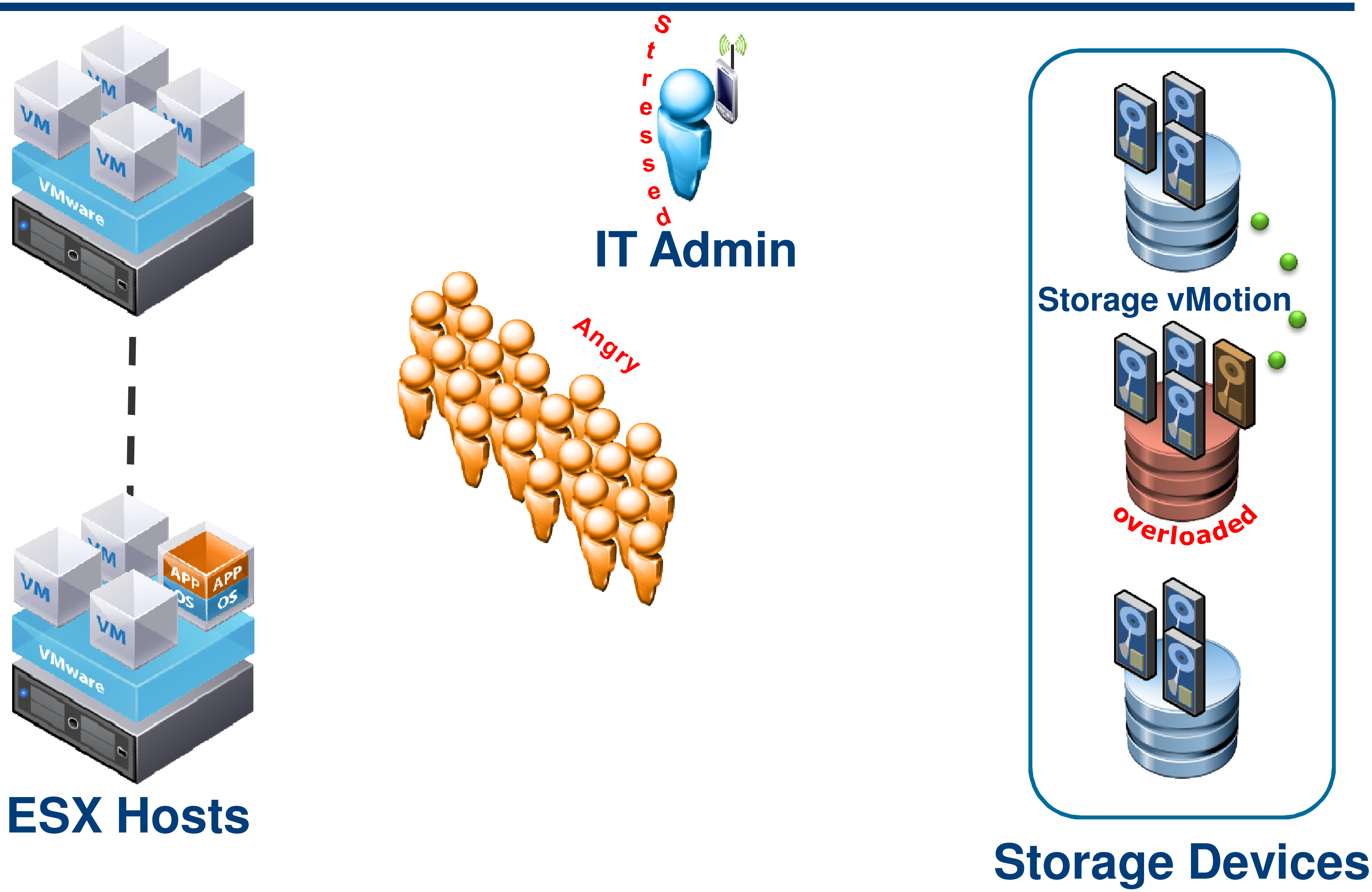


vmware

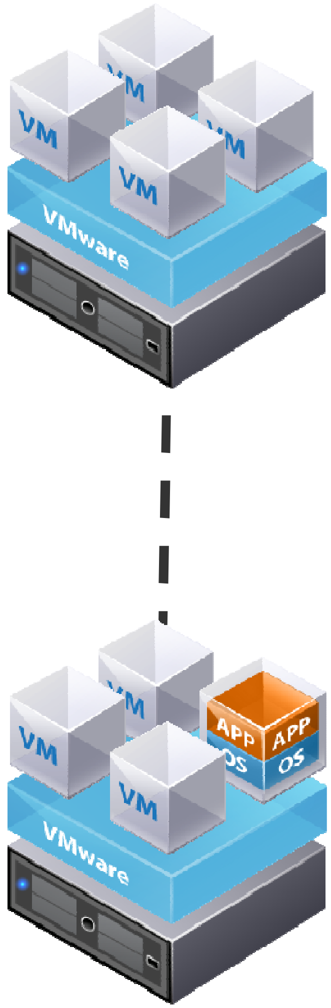
The Problem—Storage Management Not Automated



The Problem—Storage Management Not Automated



The Problem—Storage Management Not Automated



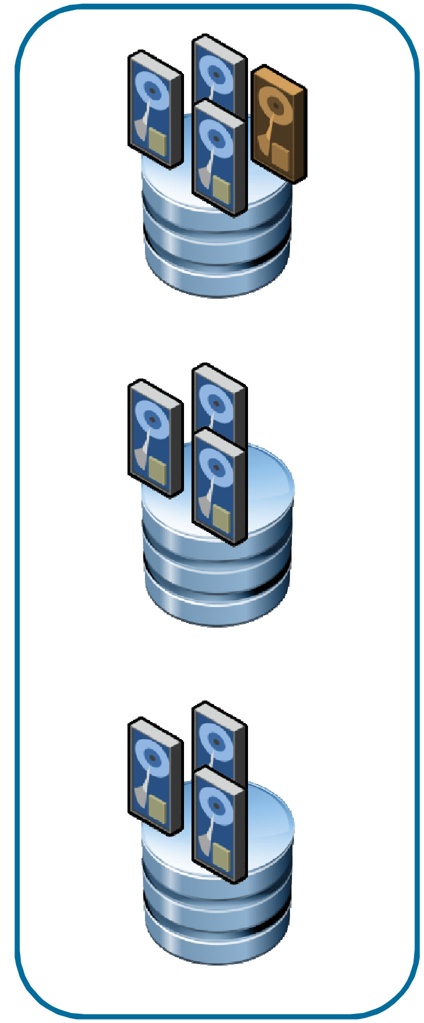
ESX Hosts

IT Admin

Management Nightmares

IO load balancing?

Virtual disk placement?



Storage Devices

Example Scenario

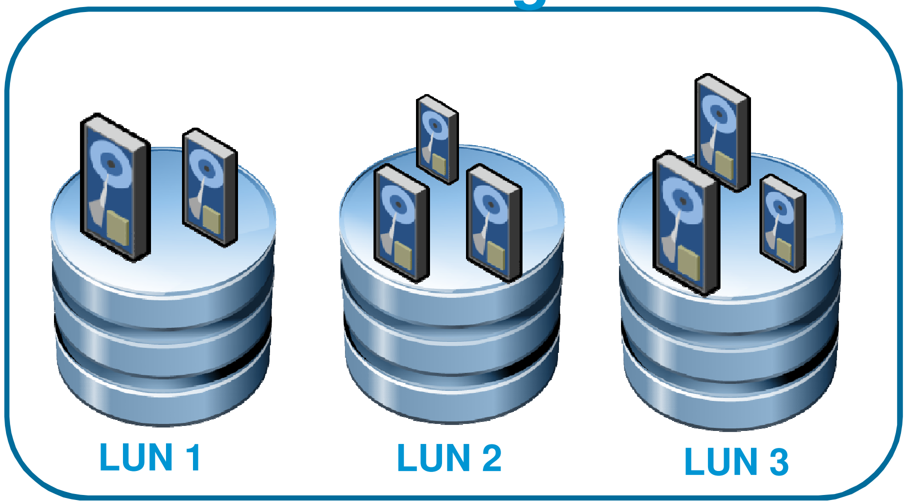
IOPS	Latency (in ms)
4172	16.7

% Change

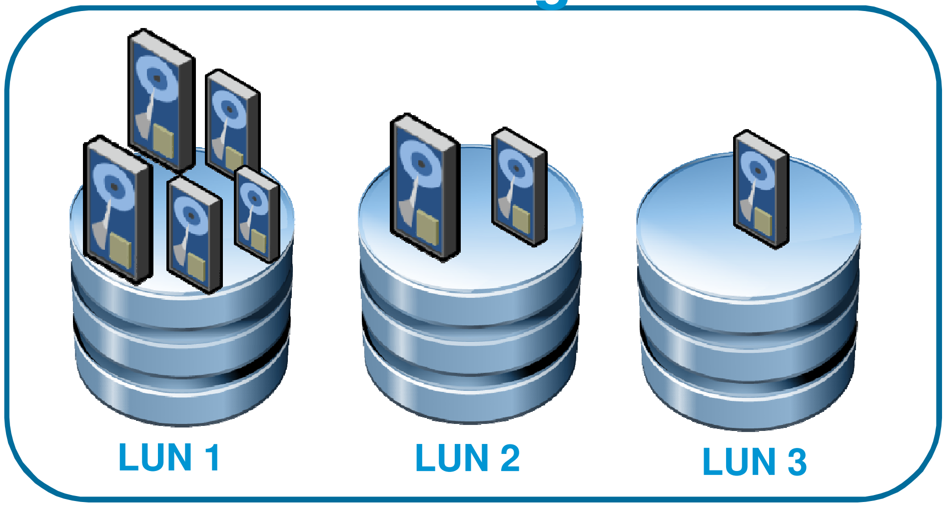
IOPS	Latency (in ms)
35%	-11%

IOPS	Latency (in ms)
5631	14.9

Initial Configuration



Final Configuration



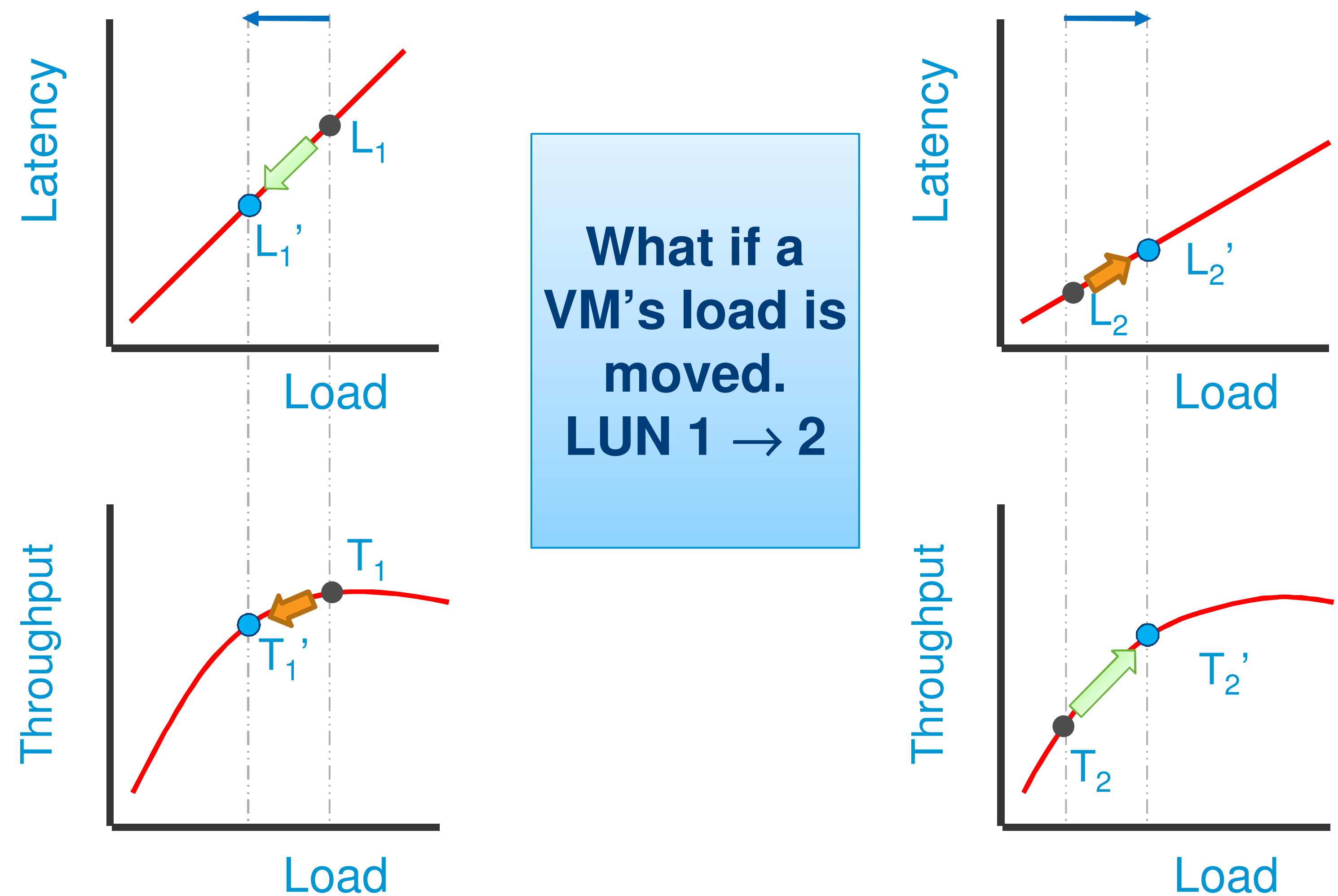
Shoulders of Giants

Much characterization & modeling work precedes us

- **Workload Characterization**
  Kavalanekar et al: IISWC ’08
  Gulati et al: VPACT ‘09
- **Minerva, Hippodrome, Table-based**
  Alvarez et al: ACM Trans. On Computing ’01
  Anderson et al: FAST ‘02
- **Analytical device models**
  Uysal et al: MASCOTS ’01
  Shriver et al: SIGMETRICS ‘98
  Merchant et al: IEEE Trans. Computing ’96
  Ruemmler & Wilkes: IEEE Computer ’94
- **Relative fitness modeling**
  Mesnier et al: SIGMETRICS ‘07
- **CART models**
  Wang et al: MASCOTS ’04

- **Novel features**
  - **Latency as primary metric**
  - **Online and lightweight**
  - **Different goal compared to existing literature**

Latency as Main Metric—Why?



LUN 1

Average Latency is lower.
Overall throughput is similar or higher.

LUN 2

Outline

- Problem Description & Motivation
- **BASIL – Modeling & Load Balancing**
- Experimental Framework & Results
- Conclusions & Future Work

BASIL Sketch

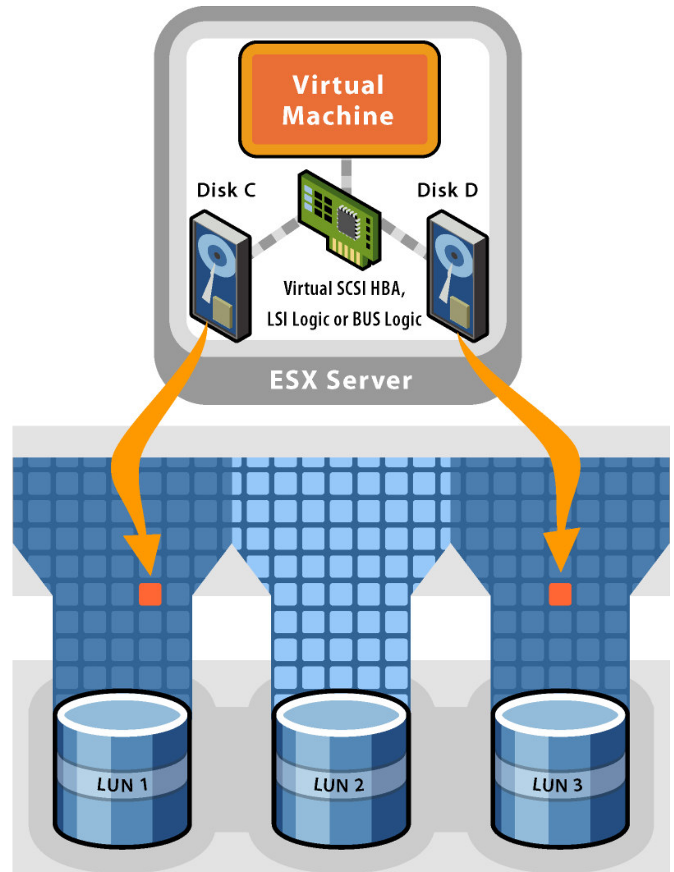
- **Online modeling**
  - Workload : capture dynamic behavior
  - Device : capture device performance
- **Load balancing based on**
  - Workload and device models
  - Assign workloads to device in proportion to their metrics

Workload Modeling

I/O Workload Modeling

- **Percentiles Data collected per-virtual disk**
  - Outstanding IOs
  - IO Size
  - Read/Write Ratio
  - Randomness

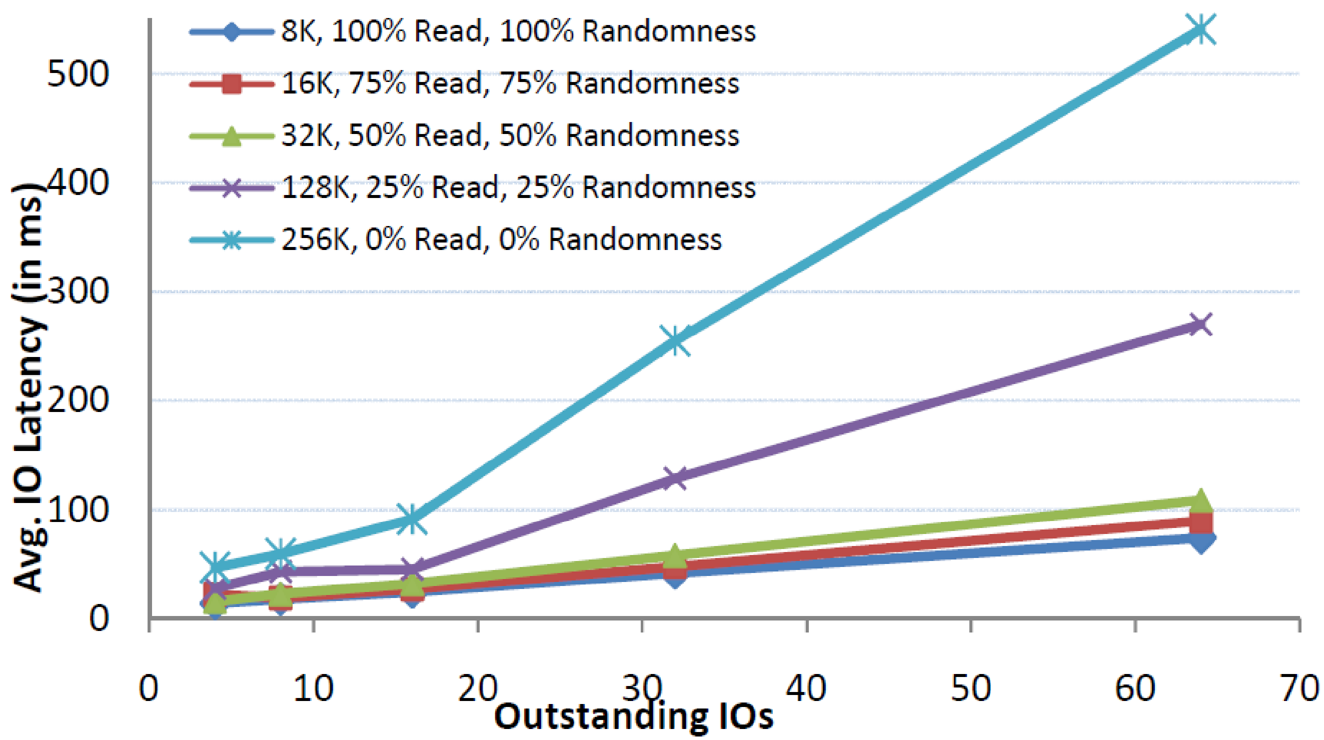
- **Methodology**
  - Analyze impact of each parameter on latency



I/O Workload Modeling

- **Percentiles Data collected per-virtual disk**
  - Outstanding IOs
  - Read/Write Ratio
  - IO Size
  - Randomness

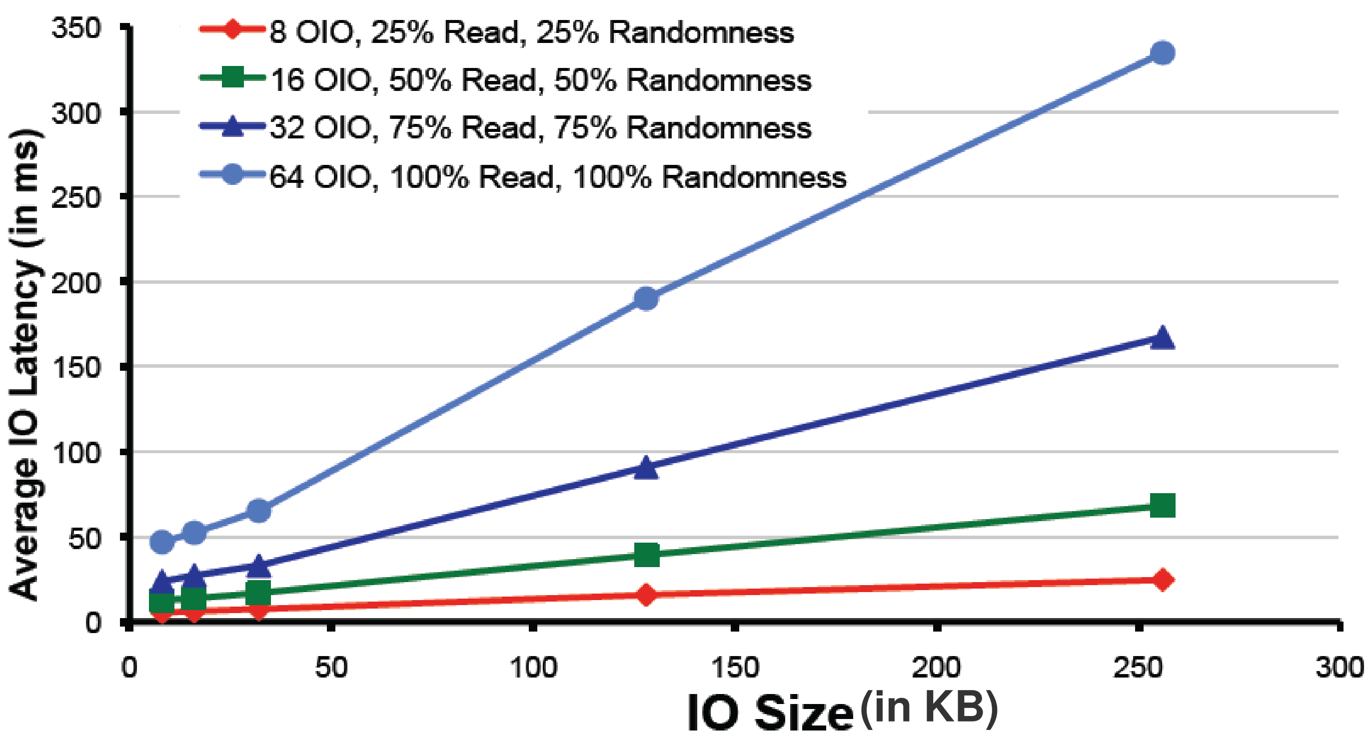
Latency varies linearly with #Outstanding IOs



I/O Workload Modeling

- **Percentiles Data collected per-virtual disk**
  - Outstanding IOs
  - Read/Write Ratio
  - **IO Size**
  - Randomness

Latency varies linearly with IO Size

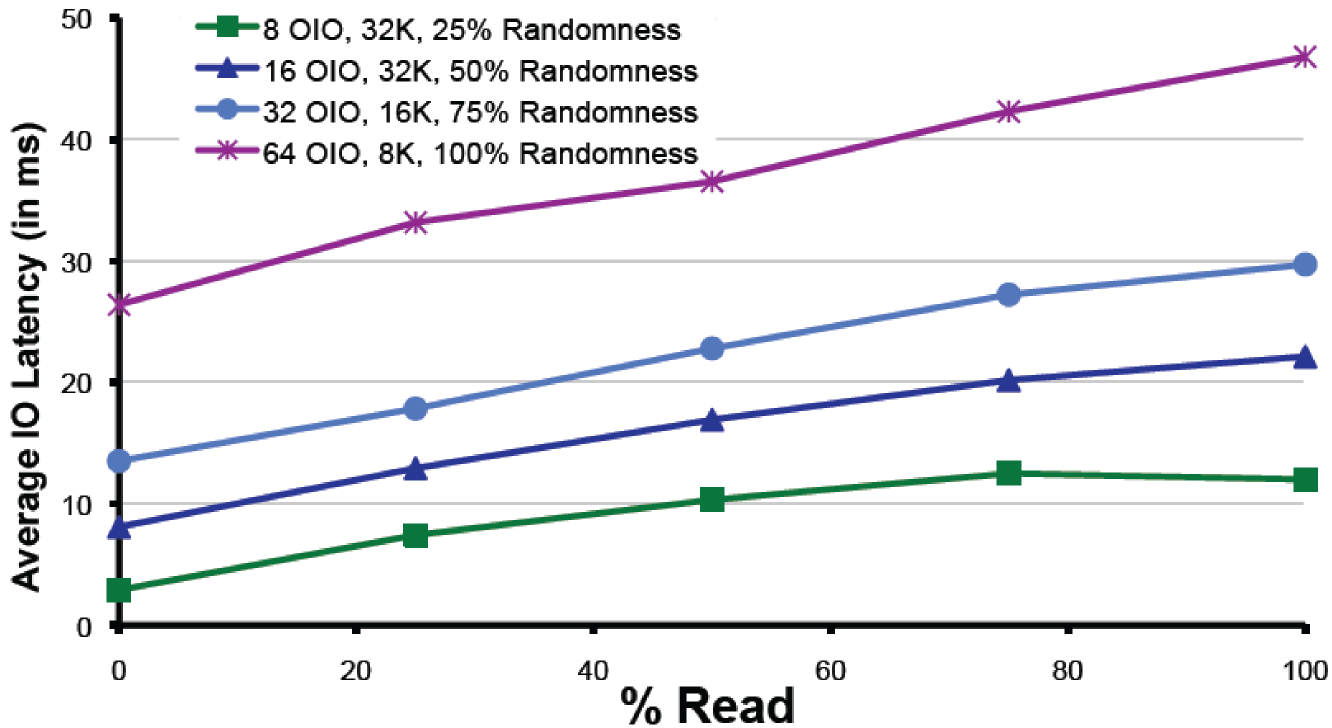


vmware

I/O Workload Modeling

- **Percentiles Data collected per-virtual disk**
  - Outstanding IOs
  - **Read/Write Ratio**
  - IO Size
  - Randomness

Latency varies linearly with %Reads

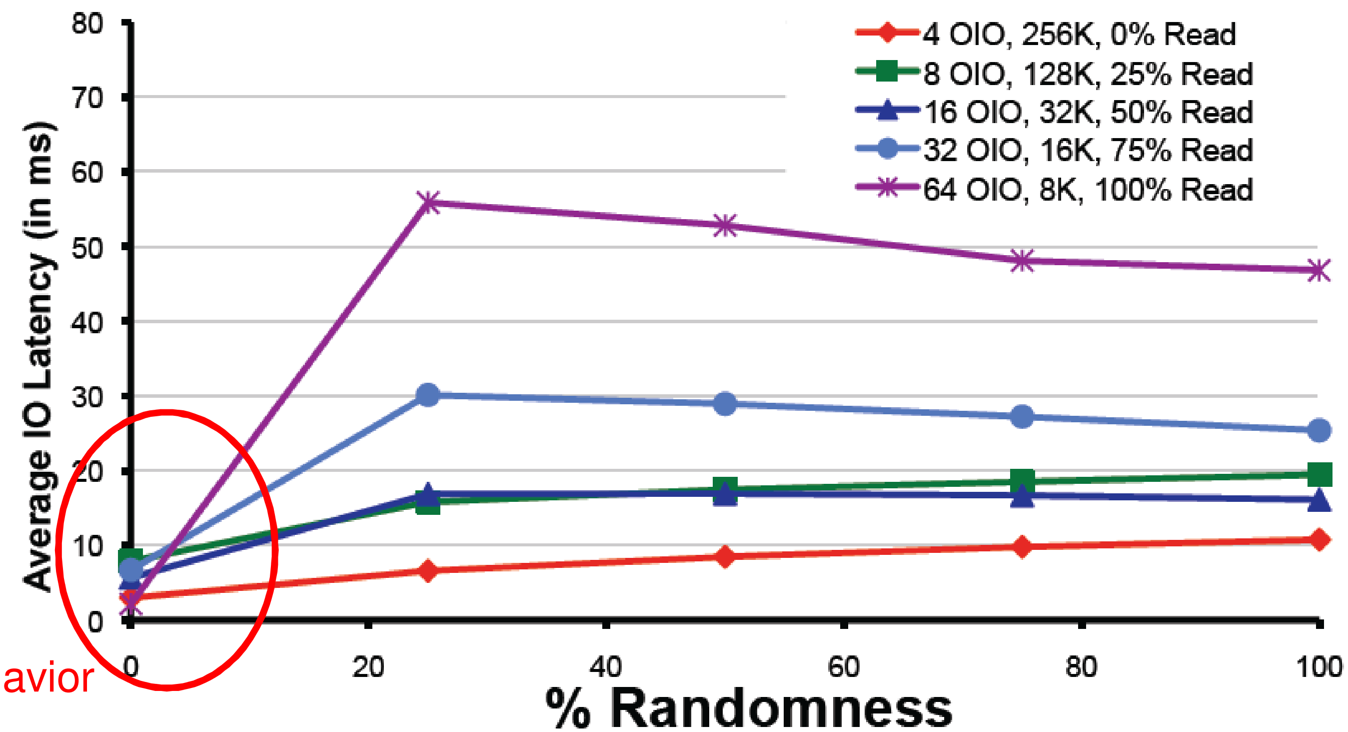


vmware

I/O Workload Modeling

- **Percentiles Data collected per-virtual disk**
  - Outstanding IOs
  - Read/Write Ratio
  - IO Size
  - **Randomness**

Latency varies linearly or Remains flat with %Randomness



Anomalous behavior for extreme cases

vmware

I/O Workload Modeling

- **Percentiles Data collected per-virtual disk**
  - Outstanding IOs
  - IO Size
  - Read/Write Ratio
  - Randomness
- **Workload Model denoted as *W***

$$W = (OIO + K_1) \cdot (IOsize + K_2) \cdot (Read\%/100 + K_3) \cdot (Random\%/100 + K_4)$$

- **K values fit from empirical data**
  - $K_1 = 1.3$
  - $K_2 = 51$
  - $K_3 = 0.4$
  - $K_4 = 0.6$

OIO is the main contributor for most cases

IO Size impacts only when change is large

Read% and Random% have less impact, except extreme scenarios

Device Modeling

Device Modeling

- **Device performance can vary widely**
  - Different number of disks: 4 vs.16 disk LUN
  - Different disk types: FC vs. SATA
  - RAID type
  - % Disk occupancy
- **BUT, device characteristics are hidden from hosts**

1 TB
20 disks
FC

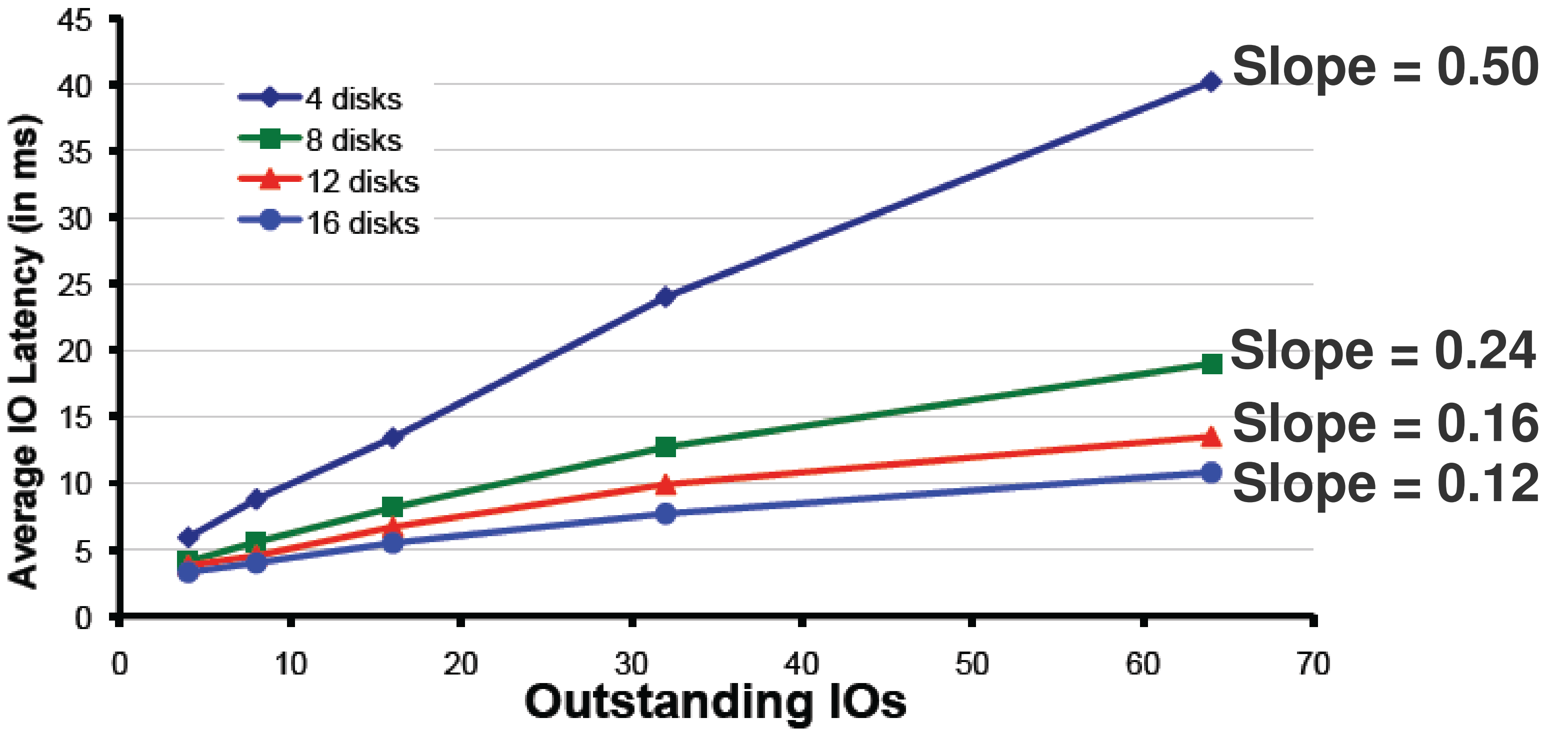
1 TB
10 disks
SATA

Device Modeling

- **Device Performance estimation**
  - *<**OIO, Latency**>* pairs collected using a reference workload
- **Linear fit approximation of the pairs**
- **Slope indicates relative performance of the device**

Device Modeling

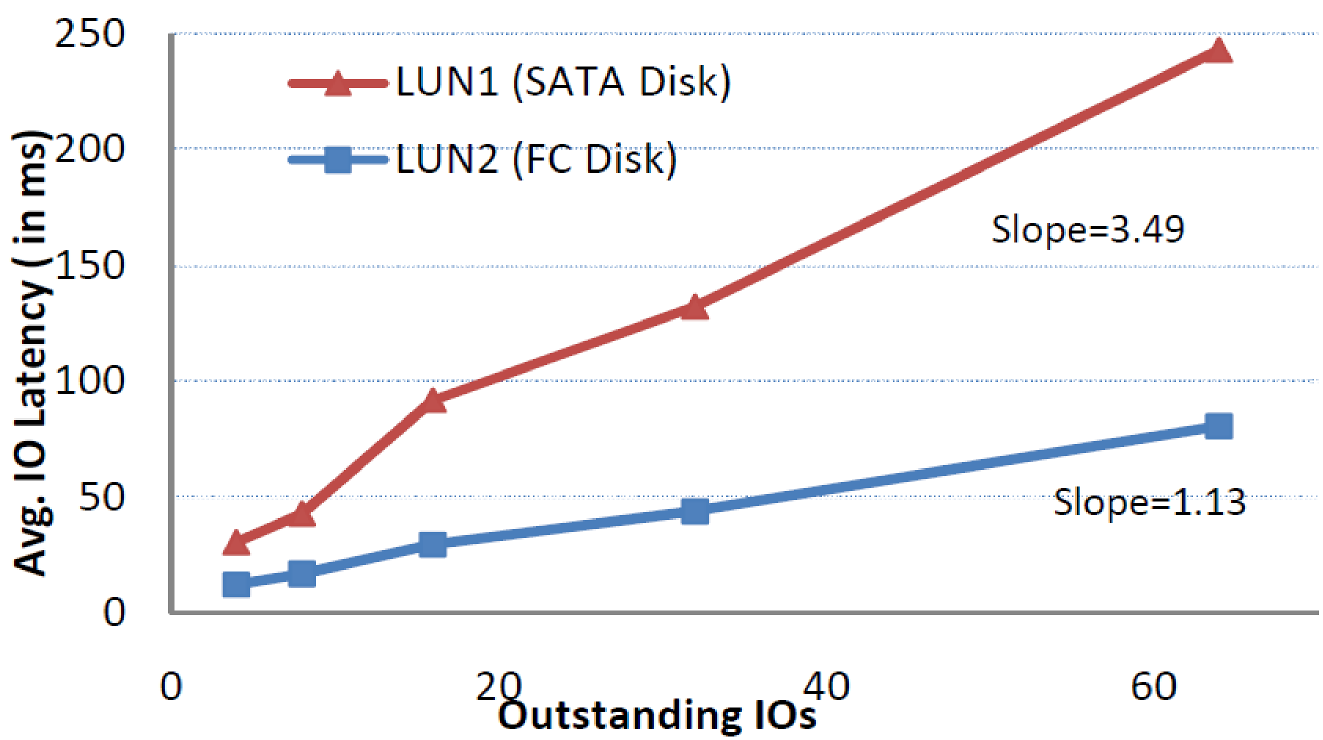
- **Device Performance estimation**
  - *<**OIO, Latency**>* pairs collected using a reference workload
- **Linear fit approximation of the pairs**
- **Slope indicates relative performance of the device**



vmware

Device Modeling

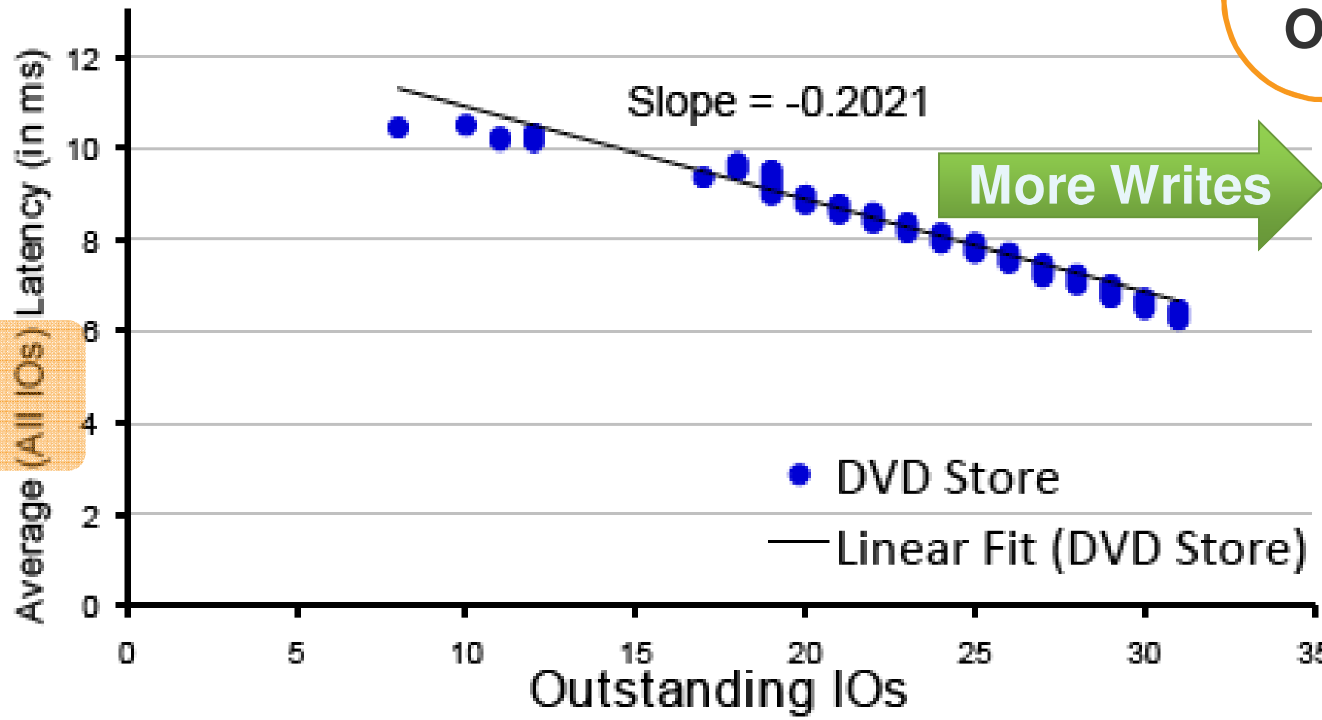
- **Device Performance estimation**
  - *<OIO, Latency>* pairs collected using a reference workload
- **Linear fit approximation of the pairs**
- **Slope indicates relative performance of the device**



Online Device Modeling—Issues

- **Generally expect positive slope values**
- **We observe negative slope values in some cases**
  - Large write IO bursts in real applications going to cache
  - IO size variation for different Outstanding IOs
  - Large sequential bursts

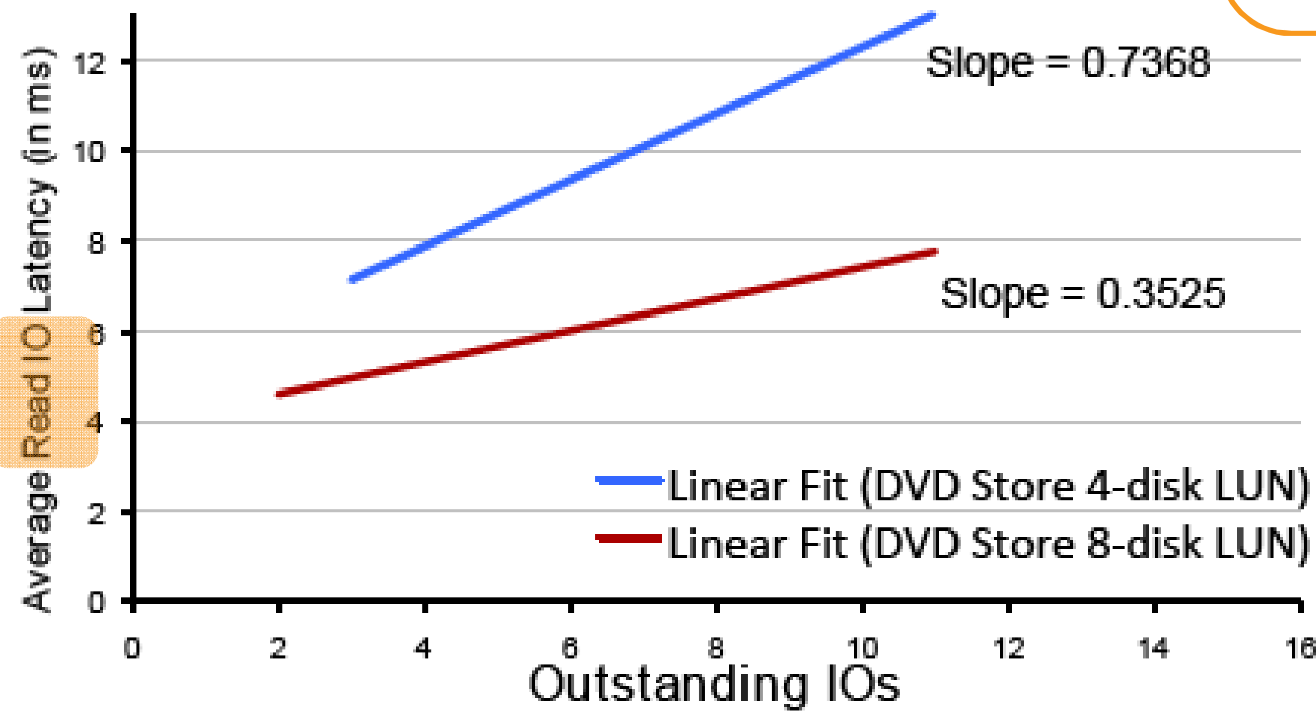
Write Bursts during high Outstanding IOs



Online Device Modeling—Solution

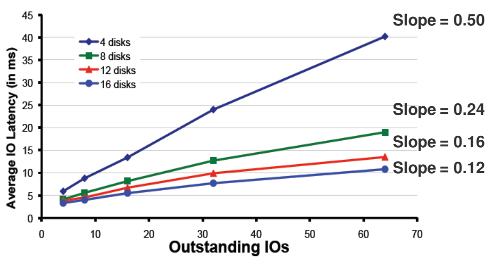
- **Filter out data from collected samples**
  - Writes: <***Read OIOs, Read latency***> pairs
  - Large IOs: filter out if IO size > 32 KB
  - Sequential IOs: filter out if sequentiality > 90 %

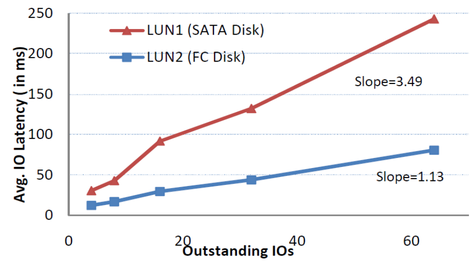
Considering only Read IOs

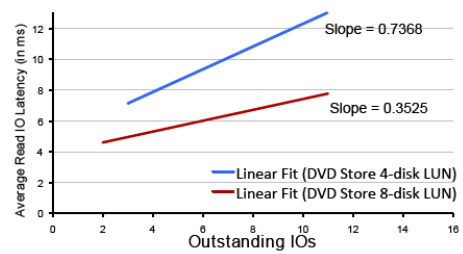


Key Takeaways

- **Slopes are indicative of relative performance**
  - 4 vs. 8 disks, other factors are constant
  - FC better than SATA, other factors kept constant
- **Incorporates cache effects**
  - Lower slope for arrays with smaller cache







- **Online modeling**
  - Online modeling is highly useful in practice
  - Filtering of online input needed to handle extreme workloads

vmware

Load Balancing

Load Balancing

- **Recall Workload metric: W_i**

$$W_i = (OIO + K_1) \cdot (IOsize + K_2) \cdot (Read\%/100 + K_3) \cdot (Random\%/100 + K_4)$$

- **Recall Device metric: P_j**
  - 1 / slope of linear fit between <Read OIO, Read latency>
- **Define Normalized Load on a device: NL**

$$NL = \frac{\sum \text{Workload metric } W_i \text{ on a device } j}{P_j}$$

- **Load balancing**
  - Assign workloads to devices in proportion to their performance
  - Heuristic: Equalize NL across data stores
- **Initial placement of virtual disks**
  - Pick device with minimum NL

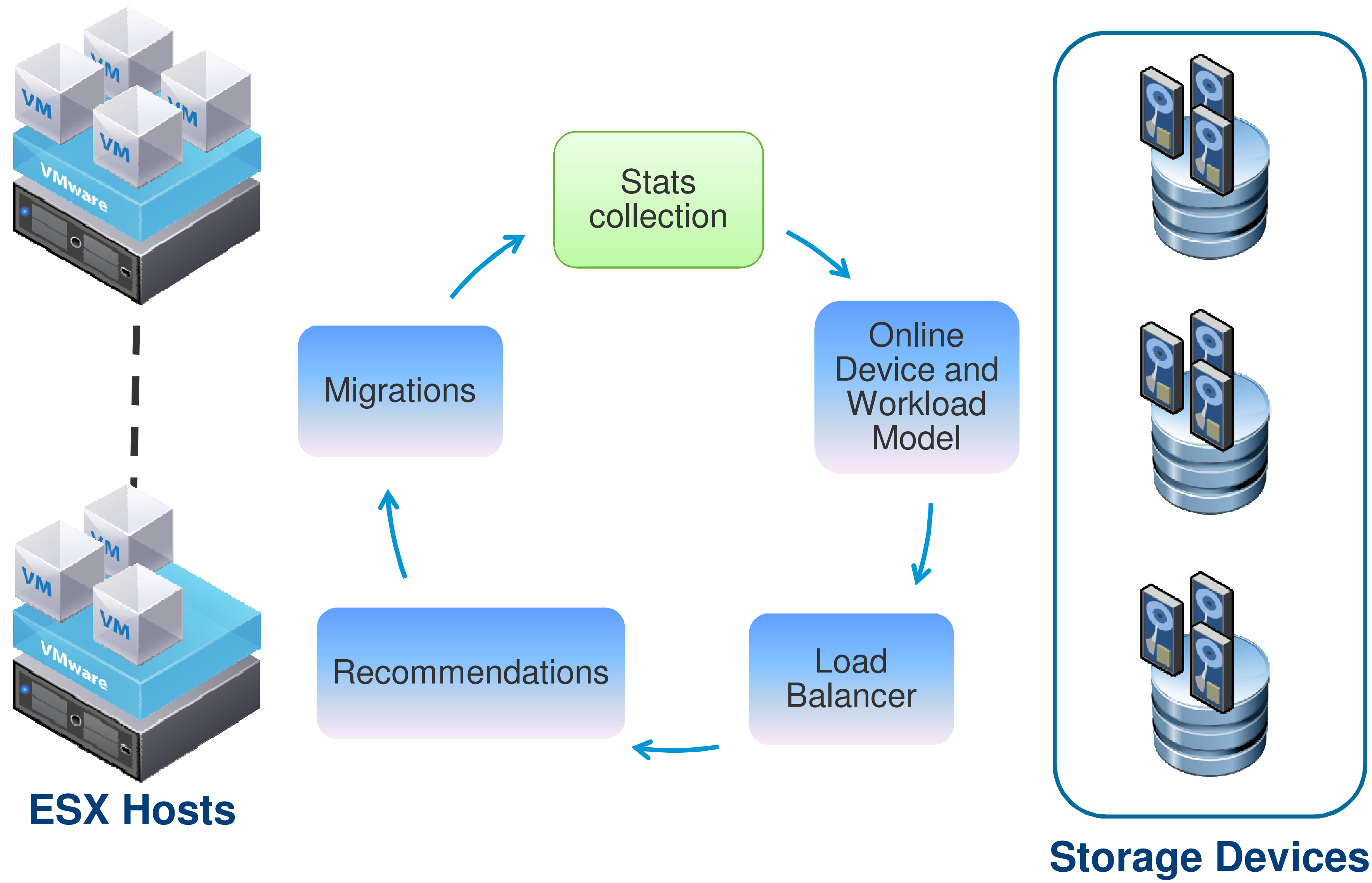
Outline

- Problem Description & Motivation
- BASIL – Modeling & Load Balancing
- **Experimental Framework & Results**
- Conclusions & Future Work

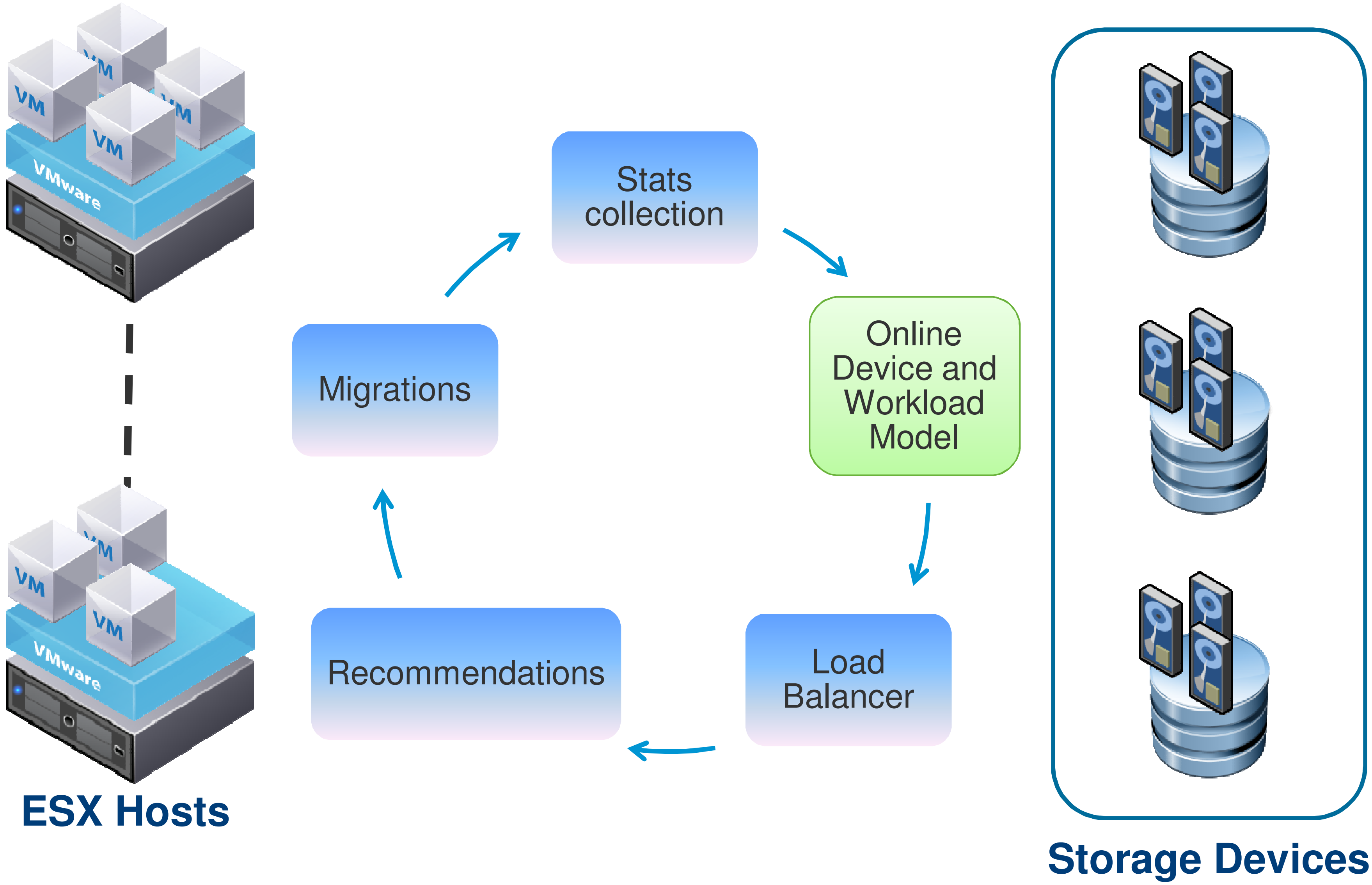
Experimental Setup

- **2 hosts running VMware ESX 4.0 hypervisor**
  - 8 to 13 virtual machines (VMs) – mix of Windows, Linux OSes
  - 6 Data stores
- **Devices (LUNs) spread across EMC CLARiiON & NetApp FAS-3140**
- **Workloads**
  - Real Apps: Swingbench (DBMS: Oracle), DVD Store (DBMS: SQL)
  - Filebench: varmail, OLTP, webserver
  - Iometer configurations: OLTP, Workstation, Exchange Server, Web Server
    - http://blogs.msdn.com/tvoellm/archive/2009/05/07/useful-io-profiles-for-simulating-various-workloads.aspx

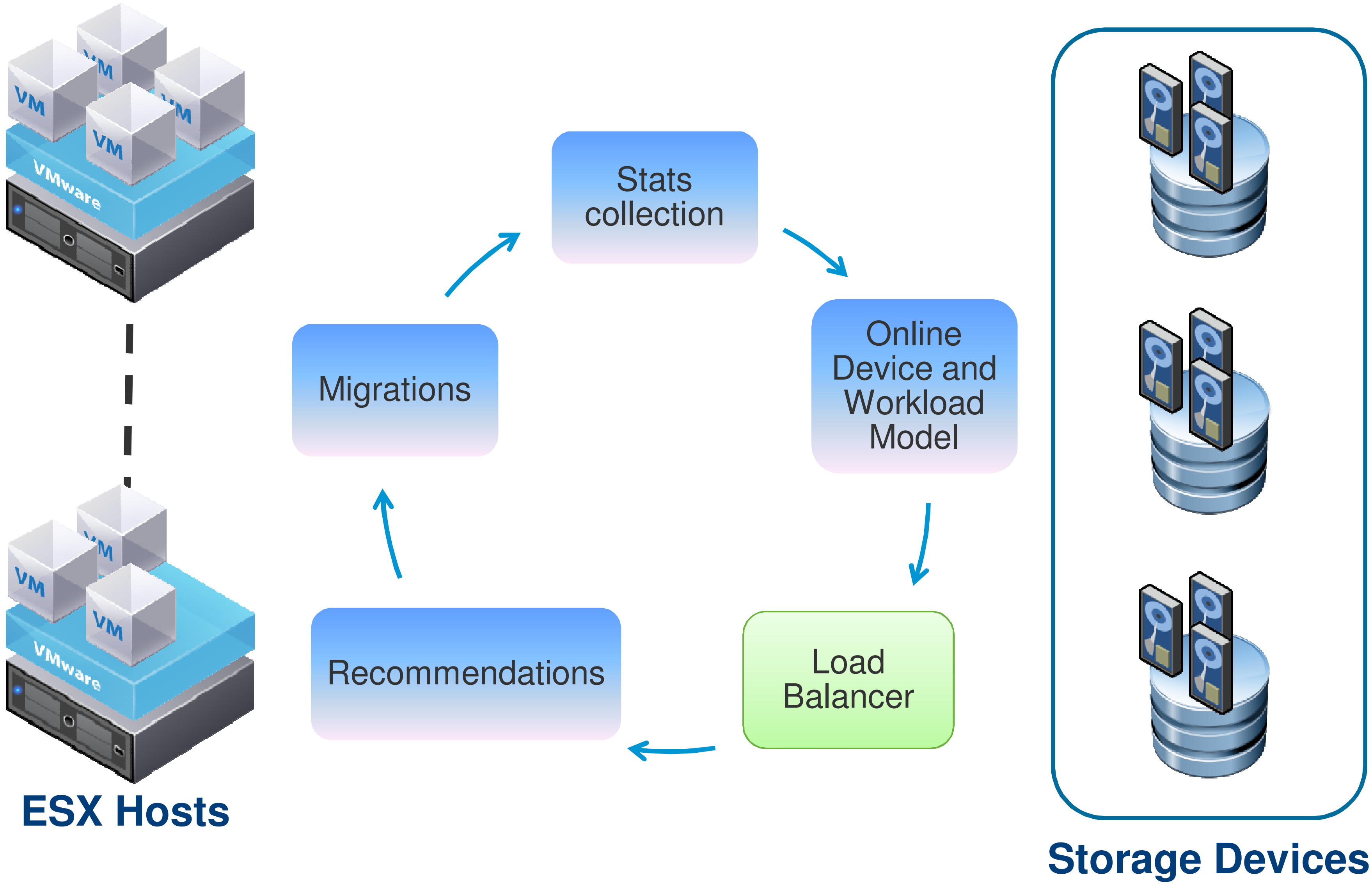
The Problem—Storage Management Not Automated



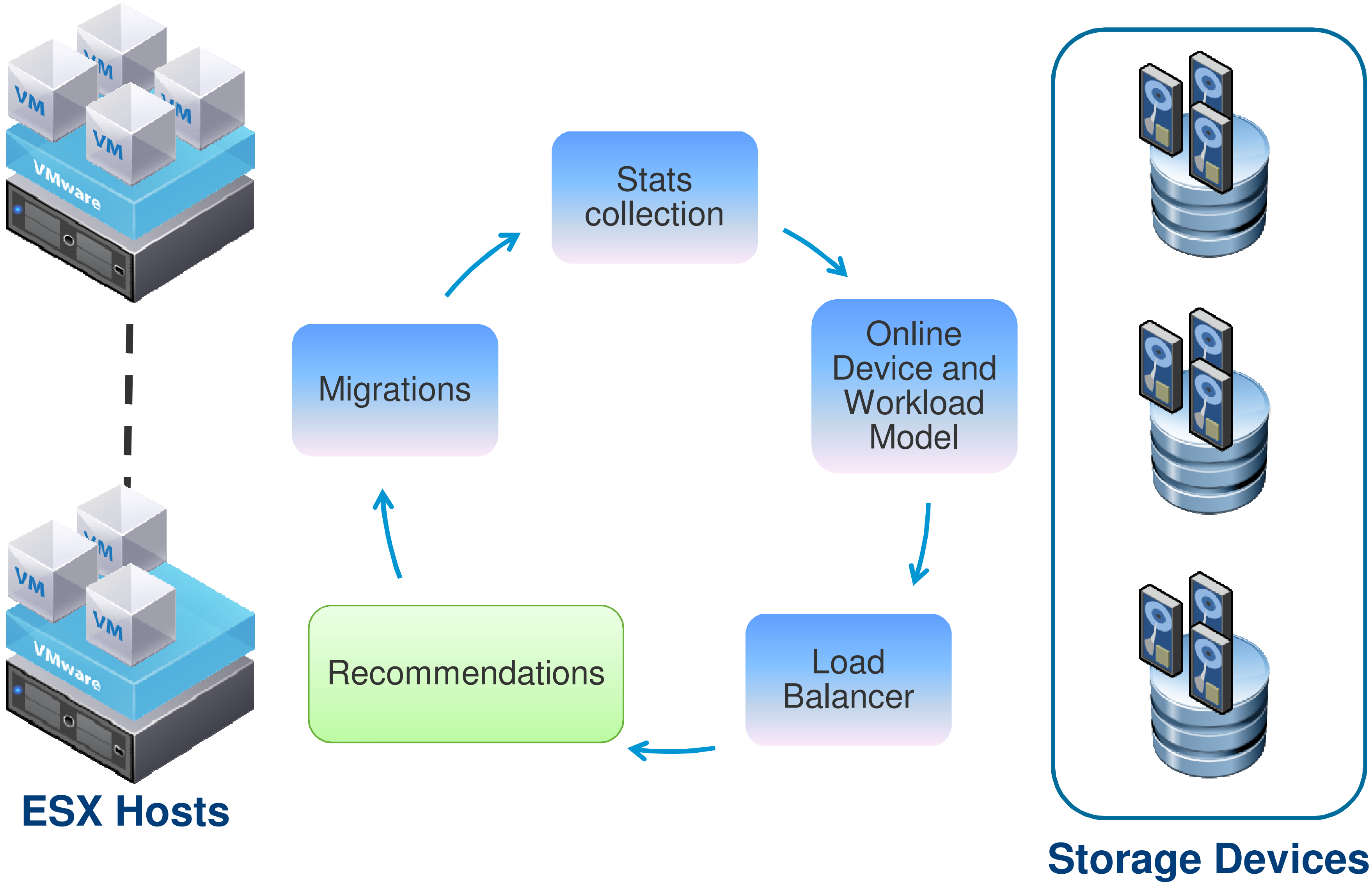
The Problem—Storage Management Not Automated



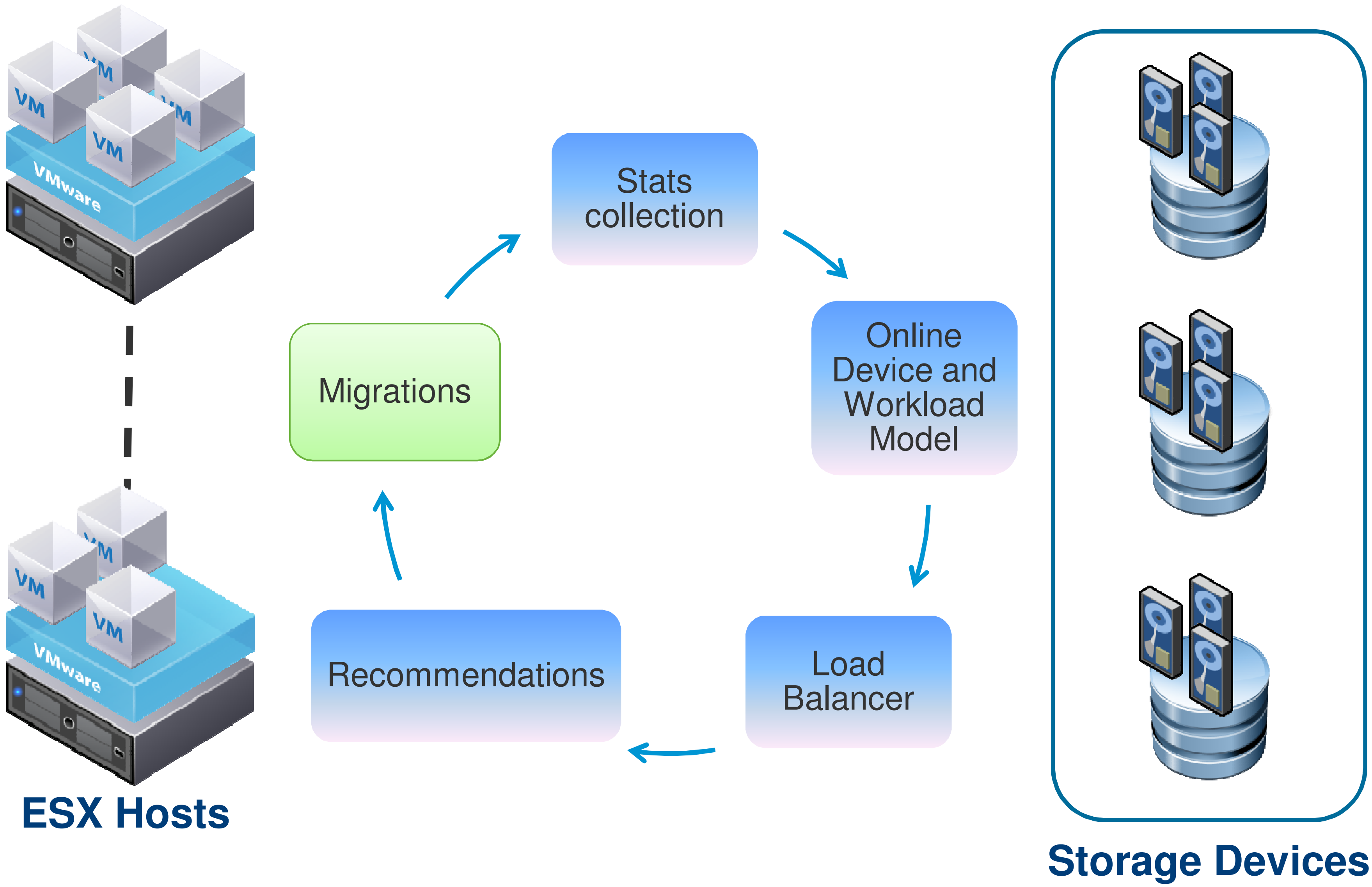
The Problem—Storage Management Not Automated



The Problem—Storage Management Not Automated



The Problem—Storage Management Not Automated



Device Models

- **Three devices for micro-benchmark experiments**

Device	#disks	Array	RAID	P= 1/slope
3diskLun	3	EMC Clariion	RAID-5	**0.6**
6diskLun	6	EMC Clariion	RAID-5	**1.4**
9diskLun	9	EMC Clariion	RAID-5	**1.8**

P: higher is better

- **Three devices for real-workload experiments**

Device	#disks	Array	RAID	P= 1/slope
EMC	6 FC	EMC Clariion	RAID-5	**1.10**
NetApp-SP	6 FC	NetApp FAS 3140	RAID-5	**0.83**
Netapp-DP	7 SATA	NetApp FAS 3140	RAID-6	**0.48**

Load Balancing

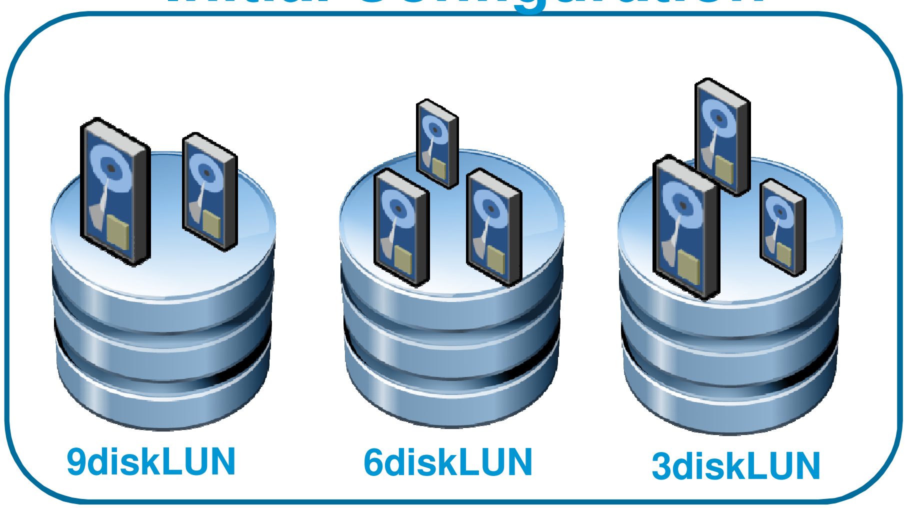
IOPS	Latency (in ms)
4172	16.7

% Change

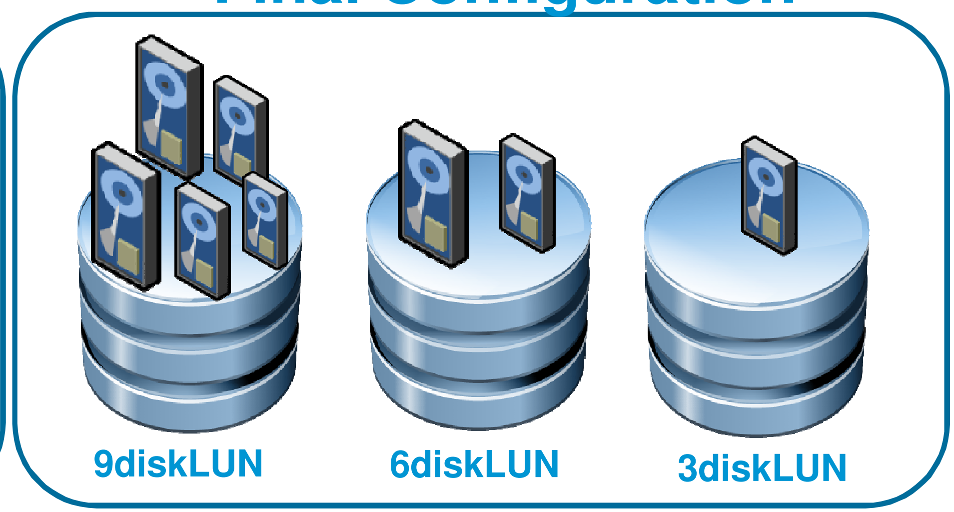
IOPS	Latency (in ms)
35%	-11%

IOPS	Latency (in ms)
5631	14.9

Initial Configuration

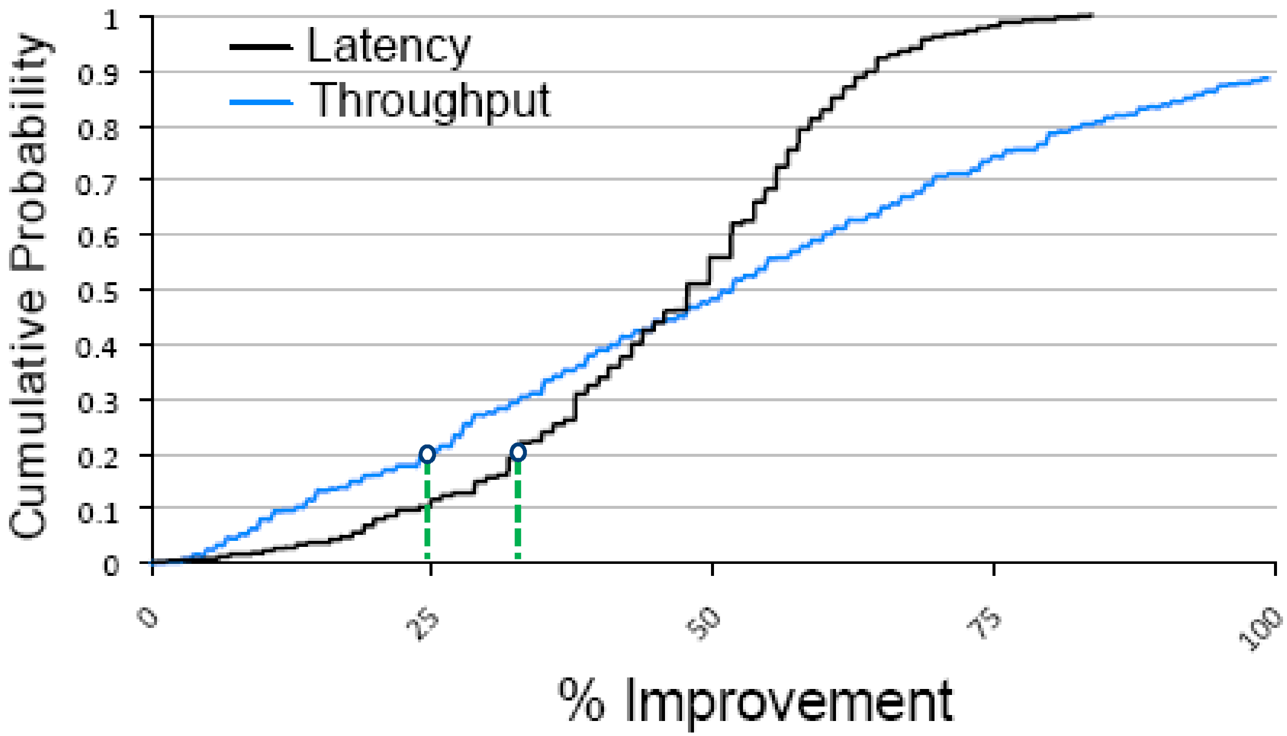


Final Configuration



Summary: 500 Runs

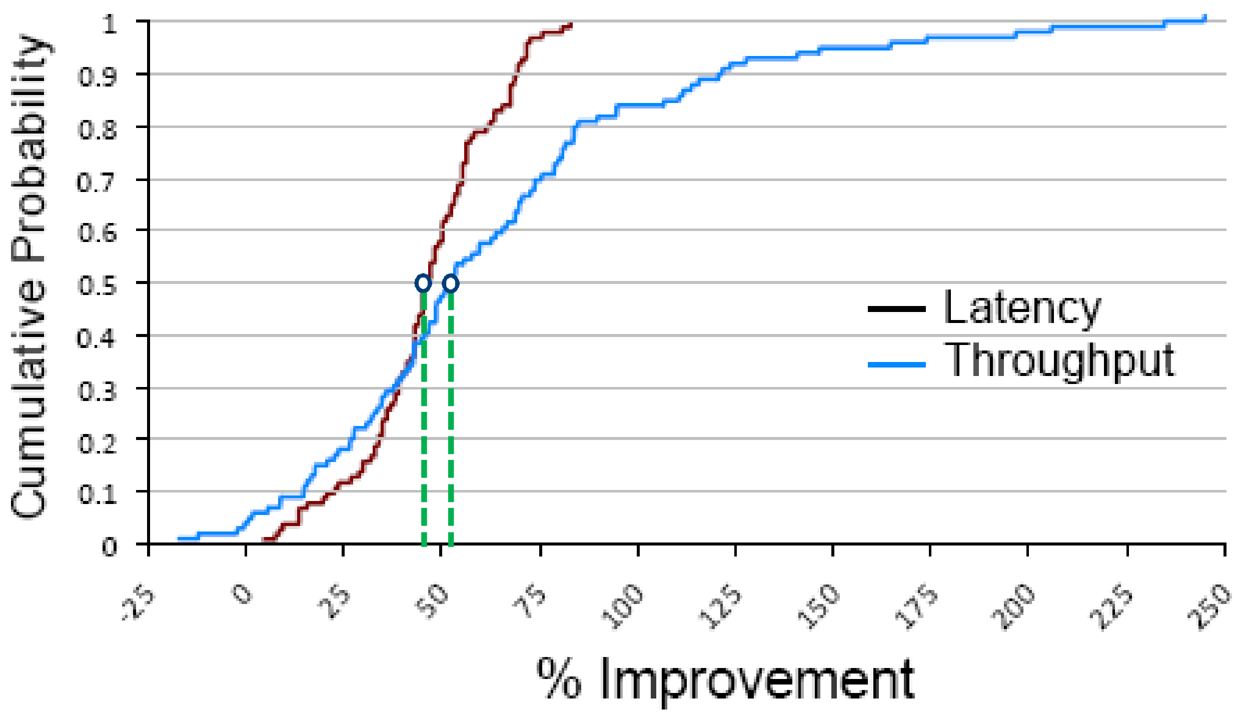
- **Random placement vs. BASIL (80th percentile values)**
  - ≥ 25% improvement in IOPS
  - ≥ 33% decrease in overall latency (computed using IOPS as weights)



vmware

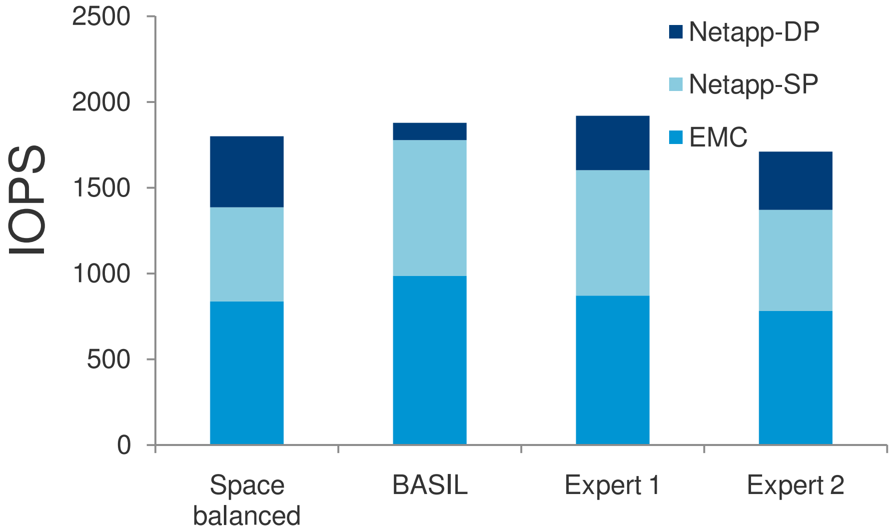
Summary: 100 Initial Placements

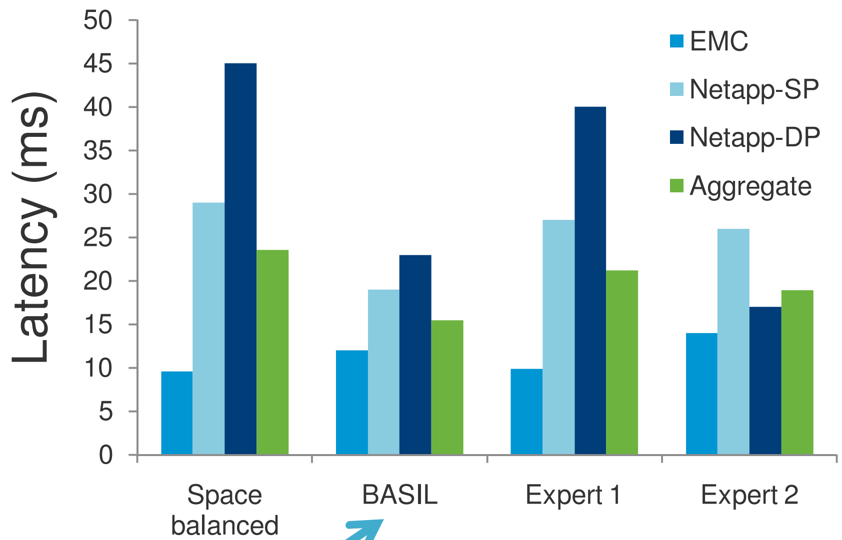
- **Random initial placement vs. BASIL (50th percentile values)**
  - ≥ 53% improvement in IOPS
  - ≥ 45% decrease in overall latency (computed using IOPS as weights)



Summary: Enterprise Workloads

- **Human Experts vs. BASIL**
  - 13 VMs: 3 DVDstore, 2 Swingbench, 4 mail servers, 2 oltp, 2 webservers
  - 2 ESX hosts, 3 storage devices





BASIL provides **lowest average latency** and similar throughput

vmware

Outline

- Problem Description & Motivation
- BASIL – Modeling & Load Balancing
- Experimental Framework & Results
- **Conclusions & Future Work**

Conclusions and Future Work

- **BASIL provides**
  - Practical online workload and device models
  - Efficient initial placement
  - Load balancing results in higher utilization, lower overall latency
- **Future Work**
  - K_i values: static vs. dynamic
  - Try out alternate workload models
  - Separate device modeling for reads & writes
  - Detailed cost-benefit metric for storage vmotions