Storage I/O Control:
Proportional Allocation of Shared Storage Resources

Chethan Kumar
Sr. Member of Technical Staff, R&D
VMware, Inc.

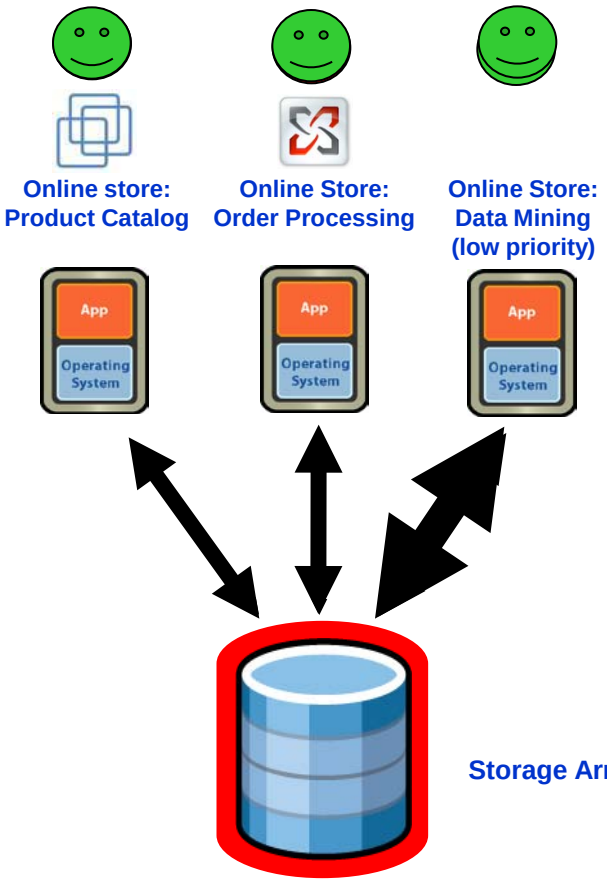
Outline

- The Problem
- Storage IO Control (SIOC) overview
- Technical Details
- SIOC in Action
  - Case study 1: Benefit of Disk Shares
  - Case study 2: Dynamic IO Prioritization
- Conclusions

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

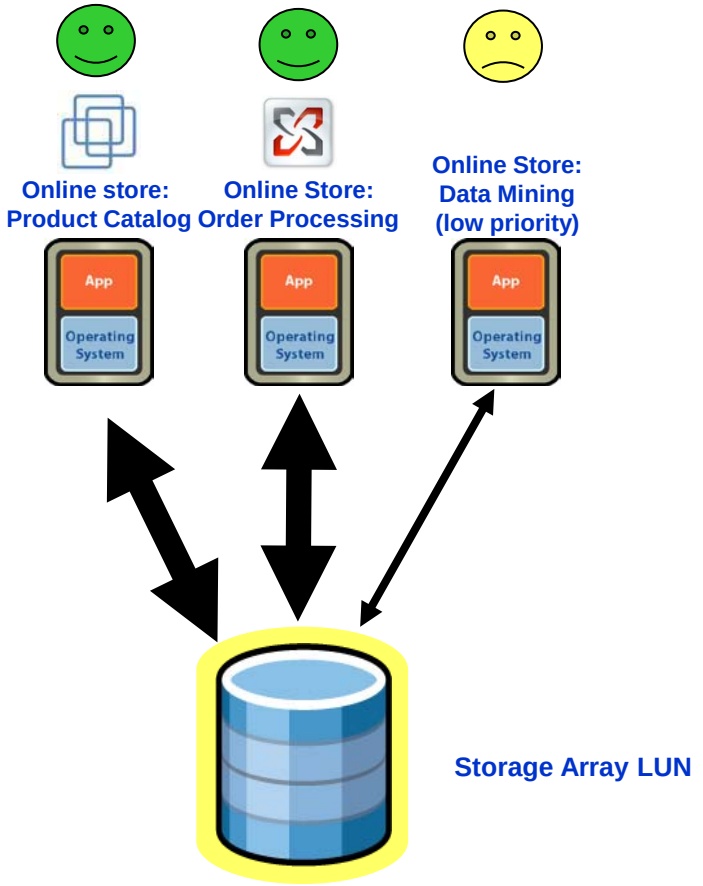
The Problem

What you see



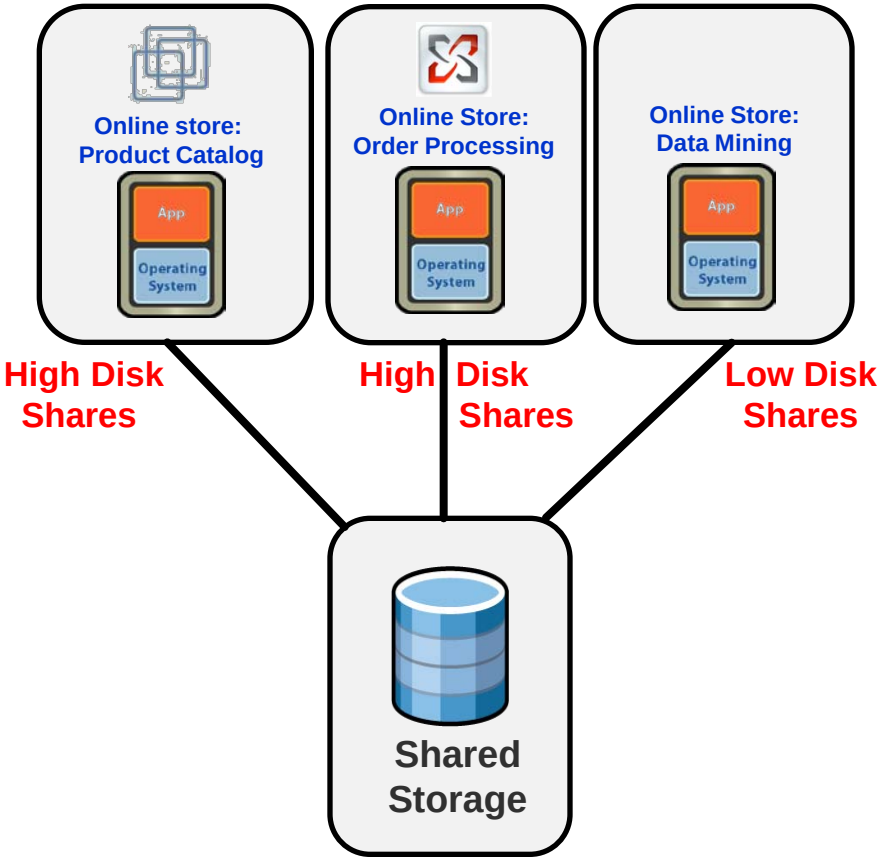
Database Server Farms

What you want to see



2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

The Solution: Resource Controls



- **Shares:** Relative priority of a virtual machine (VM)

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Outline

- The Problem
- Storage IO Control (SIOC) overview
- Technical Details
- SIOC in Action
  - Case study 1: Benefit of Disk Shares
  - Case study 2: Dynamic IO Prioritization
- Conclusions

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Typical vSphere Datacenter Architecture

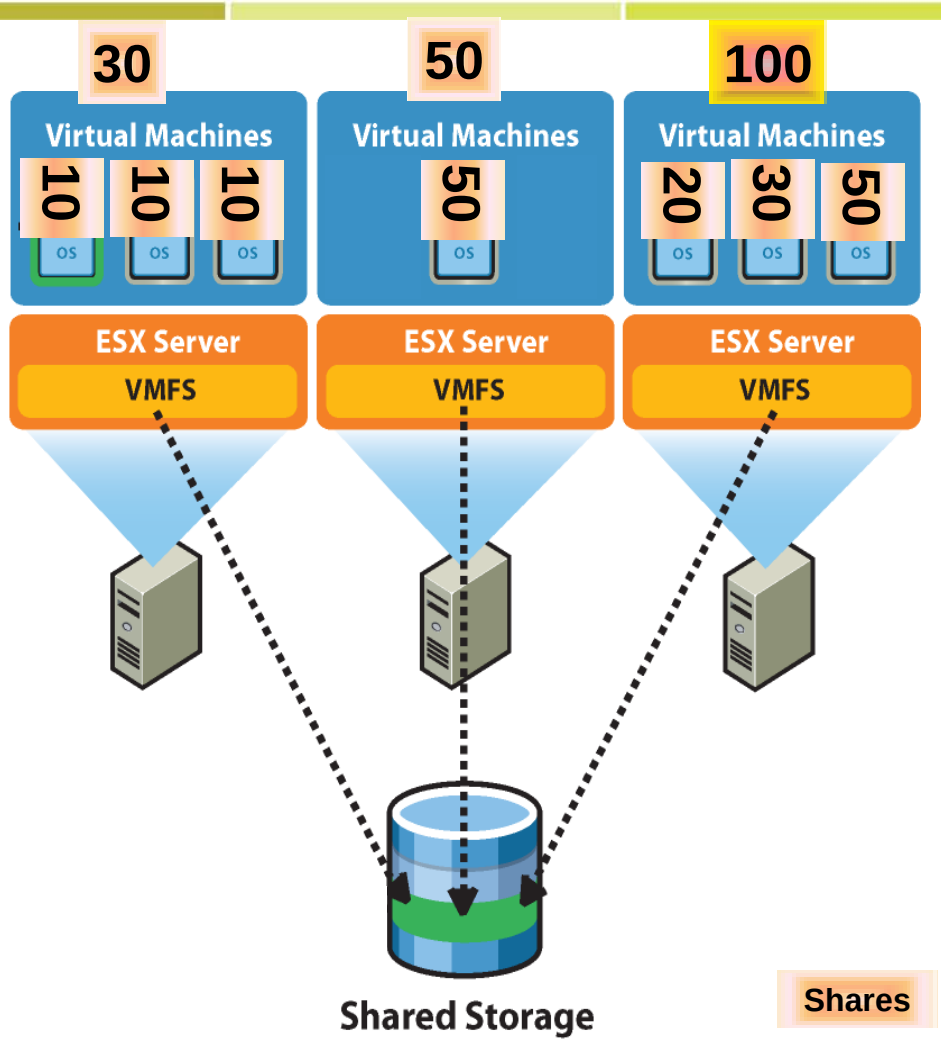
- VMs running across multiple hosts
- Hosts share LUNs using **V**irtual **M**achine **F**ile **S**ystem (a cluster file system)

Issue:

VMs interfere with each other

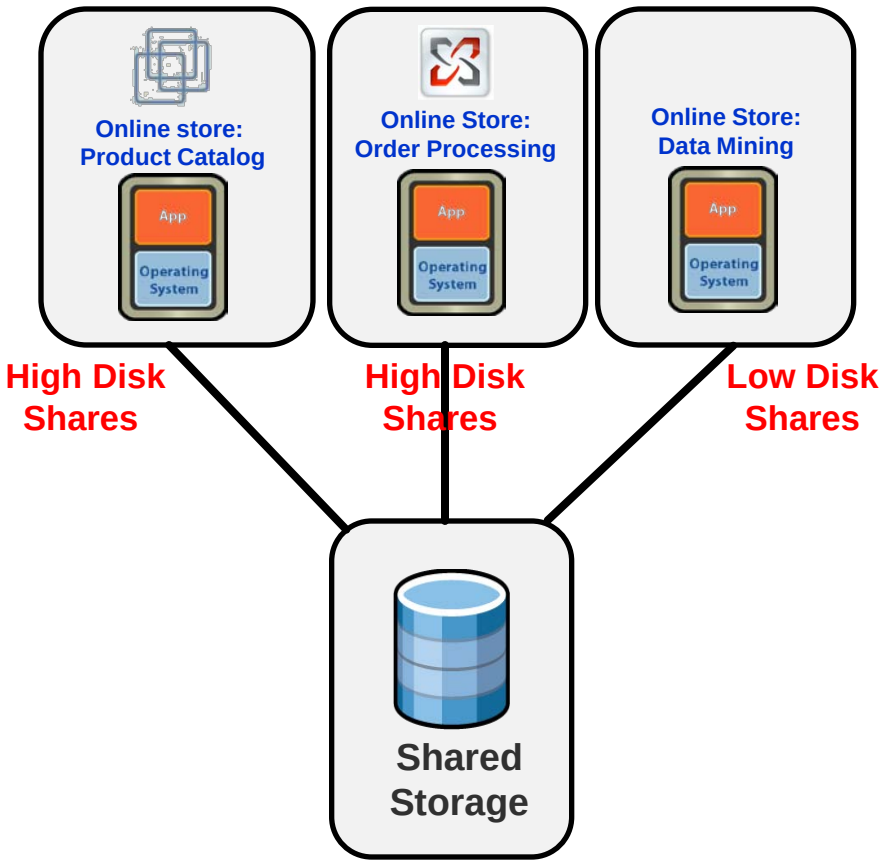
Desired Solution:

Performance isolation
while maximizing array utilization



2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Disk Shares

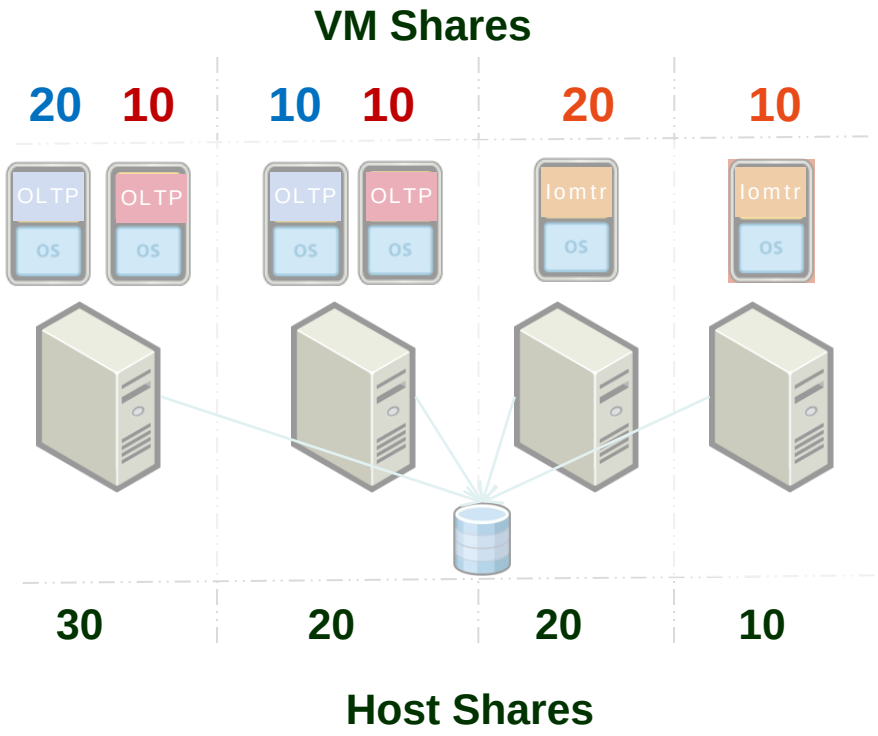


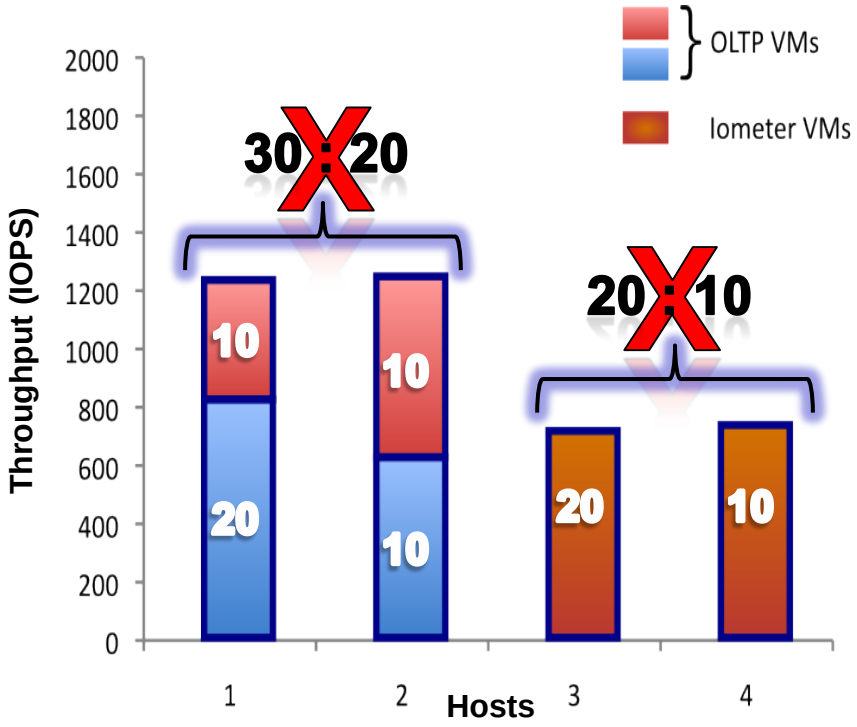
- **Shares:** Relative priority of a virtual machine (VM)

  E.g., 2x shares → 2x resource allocation **during contention**
- High, Normal, Low shares (4:2:1 ratio)
- Custom shares (numerical value)
  - Proportional weight
- Relative priority changes:
  - VMs are powered on / off
  - VMs don't utilize all the resources

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Without a Global Storage Resource Control





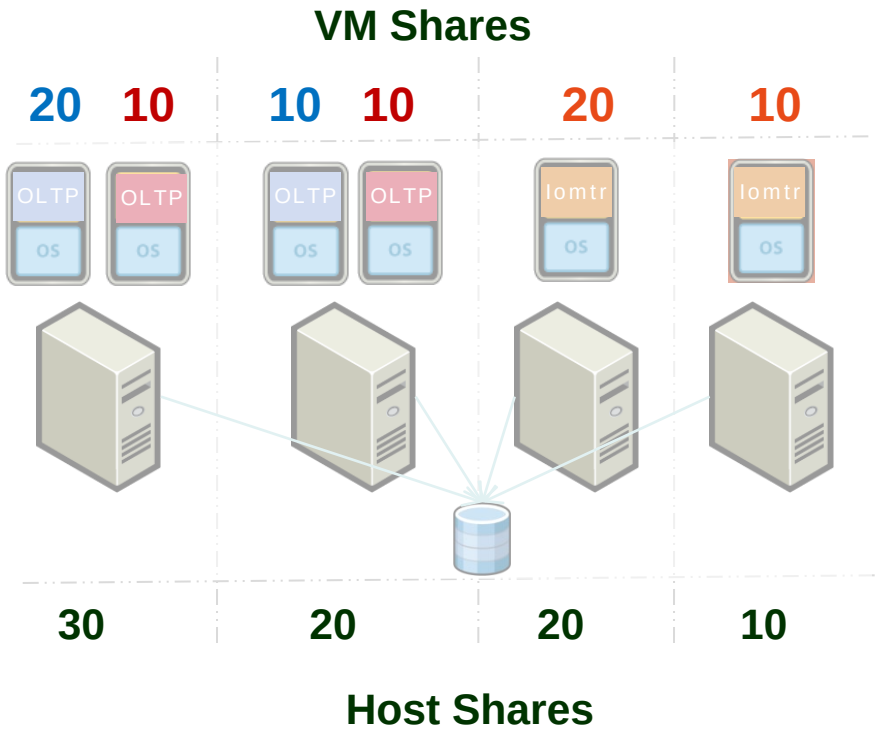
- VM shares respected only within a host
- Local scheduling helps, but not sufficient

- Hosts get equal IOPS

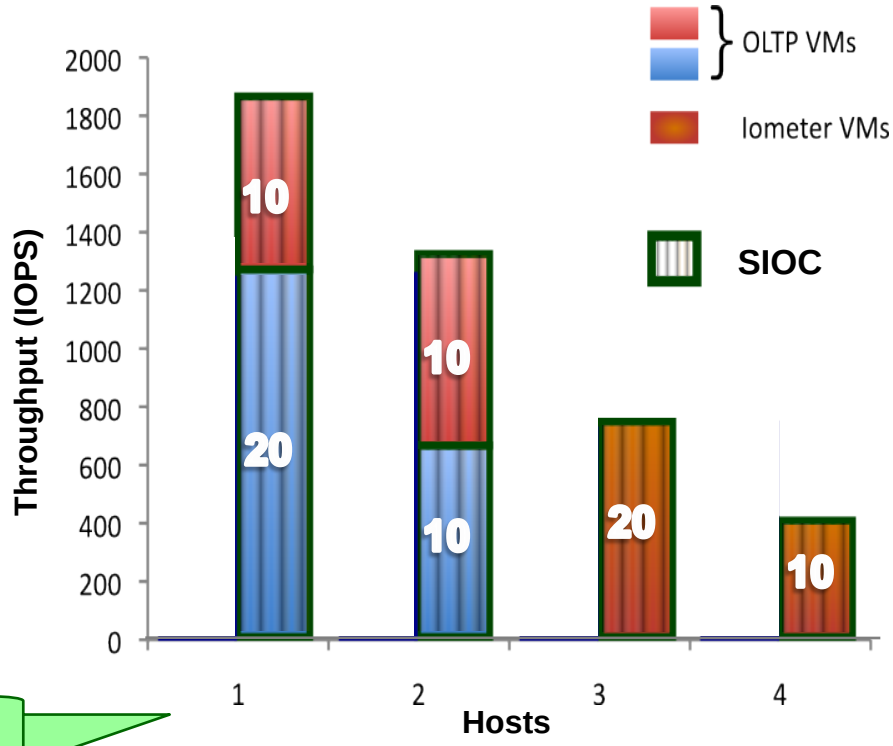
⇒IOPS dependent on VM placement!

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Optimal Storage Resource Control



- Shares should be respected across hosts
  - *Independent of VM placement*



Storage IO Control (SIOC)

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

SIOC: How To Use

- Just two steps:
  1. Enable **Storage IO Control (SIOC)** on a shared storage device (called a datastore) in ESX
  2. (Optional) Set **Disk shares** (and limit values) for virtual disks

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Outline

- The Problem
- Storage IO Control (SIOC) overview
- Technical Details
- SIOC in Action
  - Case study 1: Benefit of Disk Shares
  - Case study 2: Dynamic IO Prioritization
- Conclusions

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

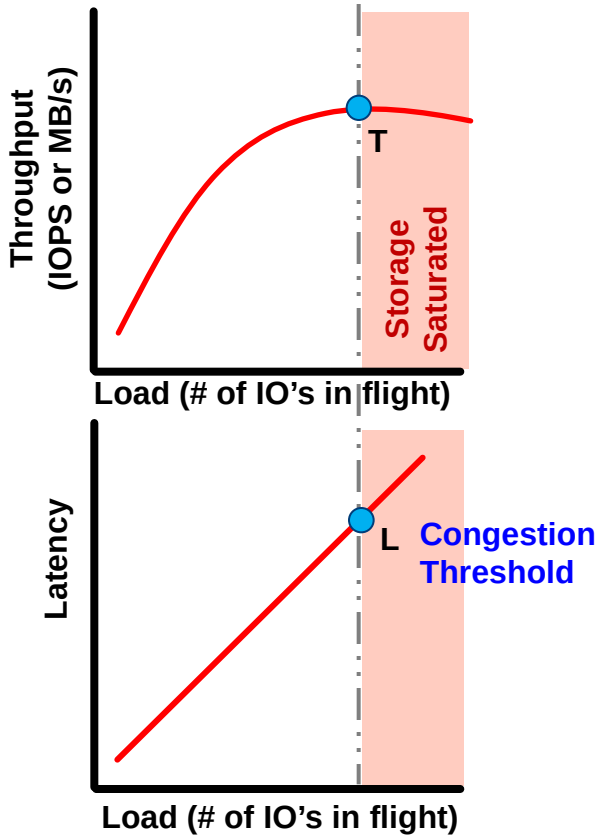
Our Approach

- Detect Congestion
  - SIOC monitors average IO latency for a datastore
  - Latency above a threshold indicates congestion (triggers SIOC)
  - *If it ain't broke, don't fix it*
- Control IOs issued per host
  - Based on VMs and their shares on each host
  - Adjust dynamically to workload
    - Idleness
    - Bursty behavior

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Congestion Detection: Setting the Threshold

- Performance suffers if datastore is overloaded
- Congestion threshold value (ms):
  - Higher is better for overall throughput
  - Lower is better for stronger isolation



- **SIOC uses a reasonable default setting**
- **Default threshold good for most cases**
  - If latency is very critical (IOPS may suffer), lower the threshold

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Per-Host Control Algorithm

Current Window size

New Delta based on current Latency

$$w(t+1) = (1-\gamma)w(t) + \gamma\left(\frac{\mathcal{L}}{L(t)}w(t) + \beta\right)$$

$\mathcal{L}$: latency threshold, operating point for IO latency

β : proportional to aggregate VM shares for host

γ : smoothing parameter between 0 and 1

Control Algorithm -

- Adjusts window (queue) size ***w(t)*** of each host using datastore-wide average latency ***L(t)***
- Runs every 4 seconds
- Motivated by FAST TCP mechanism

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Control Algorithm Features

$$w(t+1) = (1-\gamma)w(t) + \gamma\left(\frac{\mathcal{L}}{L(t)}w(t) + \beta\right)$$

- Maintain high utilization at the array
  - Overall array queue proportional to Throughput x $\mathcal{L}$
- Ability to allocate queue size in proportion to hosts' shares
  - At equilibrium, host window sizes are proportional to β
- Ability to control overall latency of a cluster
  - Cluster operates close to $\mathcal{L}$ or below

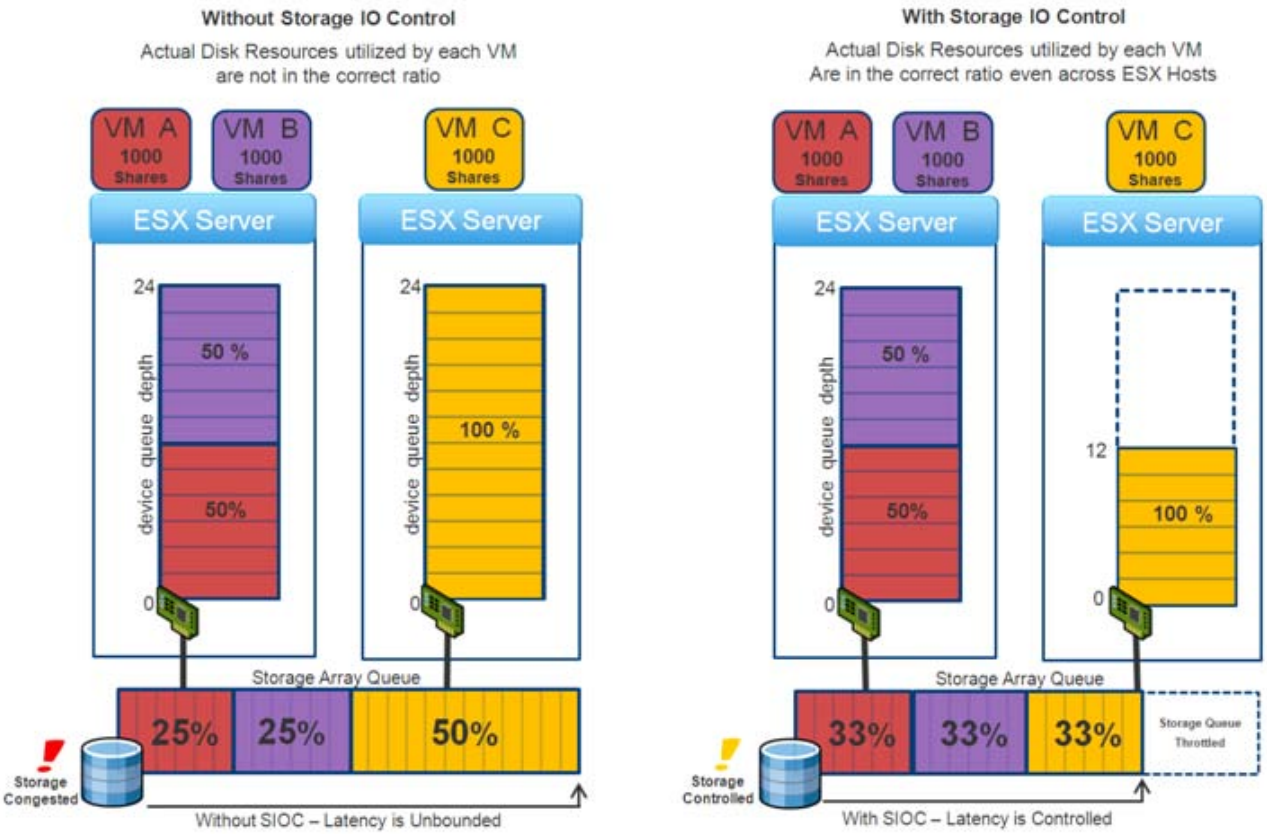
2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

What does the Priority Setting Mean?

- Two main units exist in industry
  - Bandwidth (MB/s)
  - Throughput (IOPS)
- Both have problems
  - Using bandwidth may hurt workloads with large IO sizes
  - Using IOPS may hurt VMs with sequential IOs
- SIOC: carves out array queue among VMs
  - VMs reuse queue slots faster or slower (depending on array latency)
    - e.g. Sequential streams get higher IOPS even if shares identical, similarly for workloads with high read cache hit rates
    - This is a good thing!
  - Maintains high overall throughput

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Control IOs Issued per Host (Based on Shares)



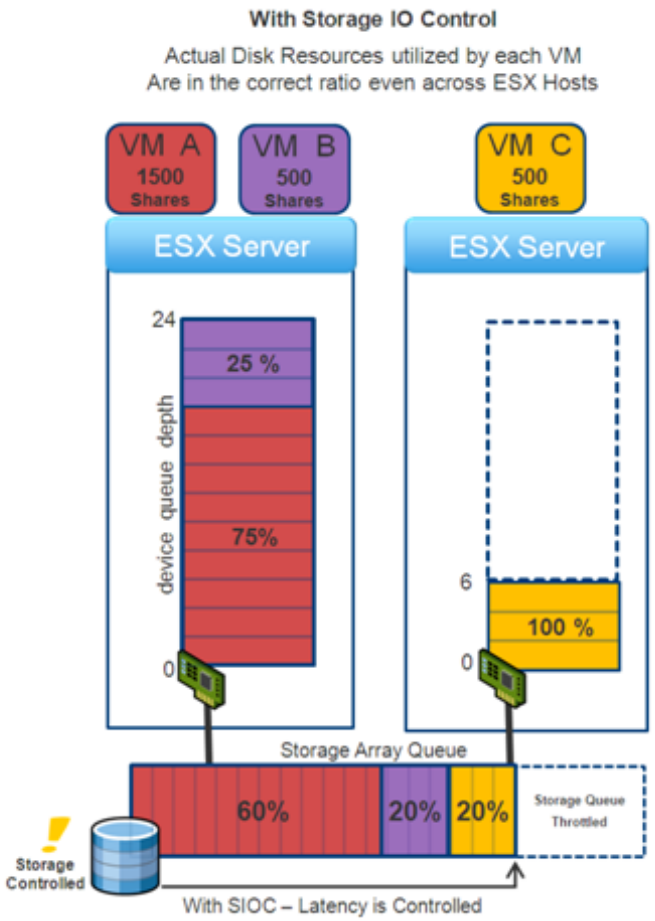
VM	Disk Shares
A	1000
B	1000
C	1000

Without SIOC: VM C gets equal queue slots as VMs A+ B

With SIOC: All VMs get equal queue slots

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Control IOs Issued per Host (Based on Shares)



VM	Disk Shares
A	1500
B	500
C	500

With SIOC: VMs get queue slots proportional to shares

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Outline

- The Problem
- Storage IO Control (SIOC) overview
- Technical Details
- SIOC in Action
  - Case study 1: Benefit of Disk Shares
  - Case study 2: Dynamic IO Prioritization
- Conclusions

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Workload

DVD Store Version 2 (DS2)

- Open Source, online E-commerce workload
- Leverages commonly used database features
- Supports SQL Server, Oracle and MySQL for backend database
- Easy to set up and run
- Supports multiple database sizes
- To download the workload:

  http://www.delltechcenter.com/page/DVD+Store

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

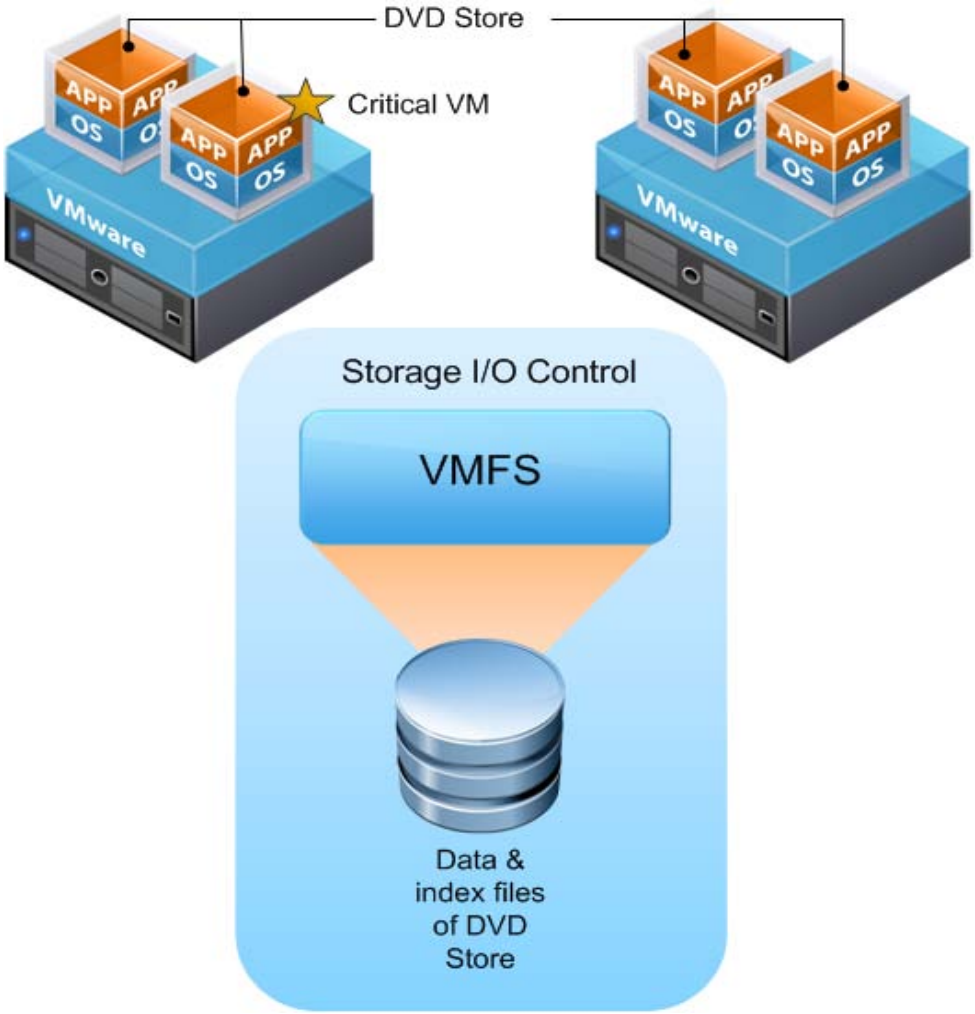
Outline

- The Problem
- Storage IO Control (SIOC) overview
- Technical Details
- SIOC in Action
  - Case study 1: Benefit of Disk Shares
  - Case study 2: Dynamic IO Prioritization
- Conclusions

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Benefit of Disk Shares

Experimental Setup

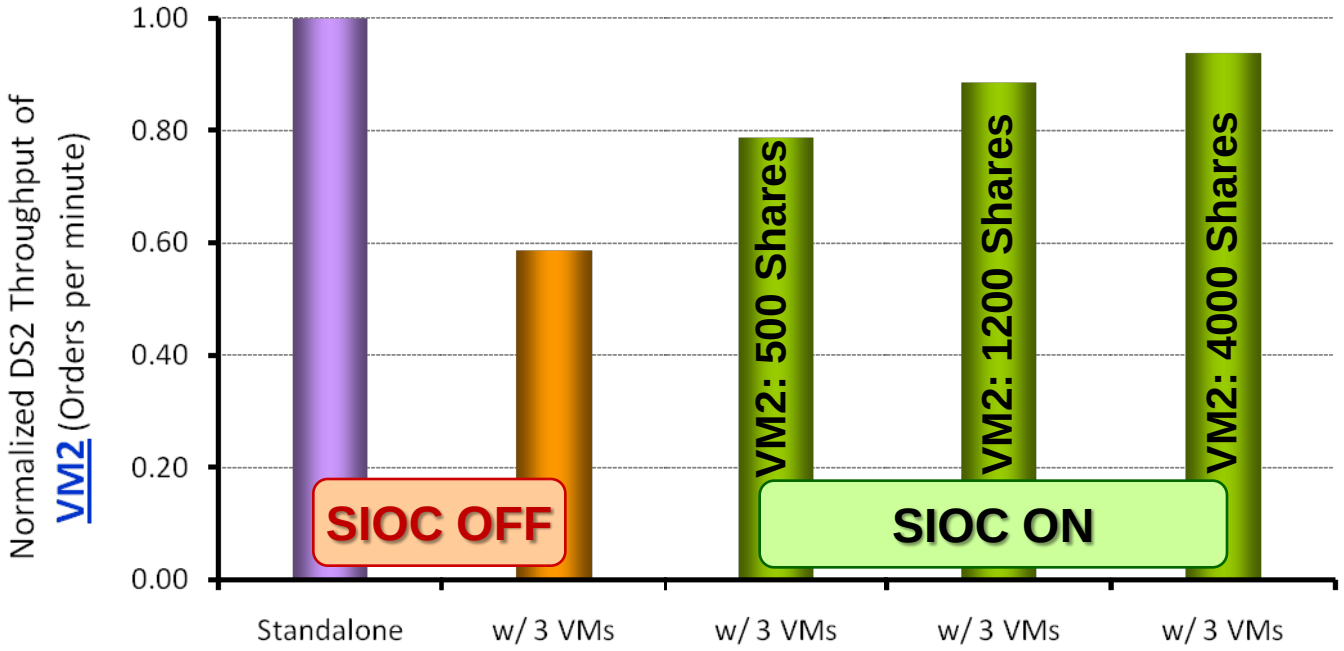


VM ID	Number of DS2 Users
1	36
★ 2	50
3	36
4	36

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Benefit of Disk Shares

- Performance of DS2 workload in **critical** VM
  - Standalone
  - When sharing datastore with other VMs: without and with SIOC enabled



With SIOC ON

VM ID	Disk Shares
1	200
★ 2	variable
3	200
4	200

Congestion Threshold: 20ms

Higher Shares ➡ Better Performance

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

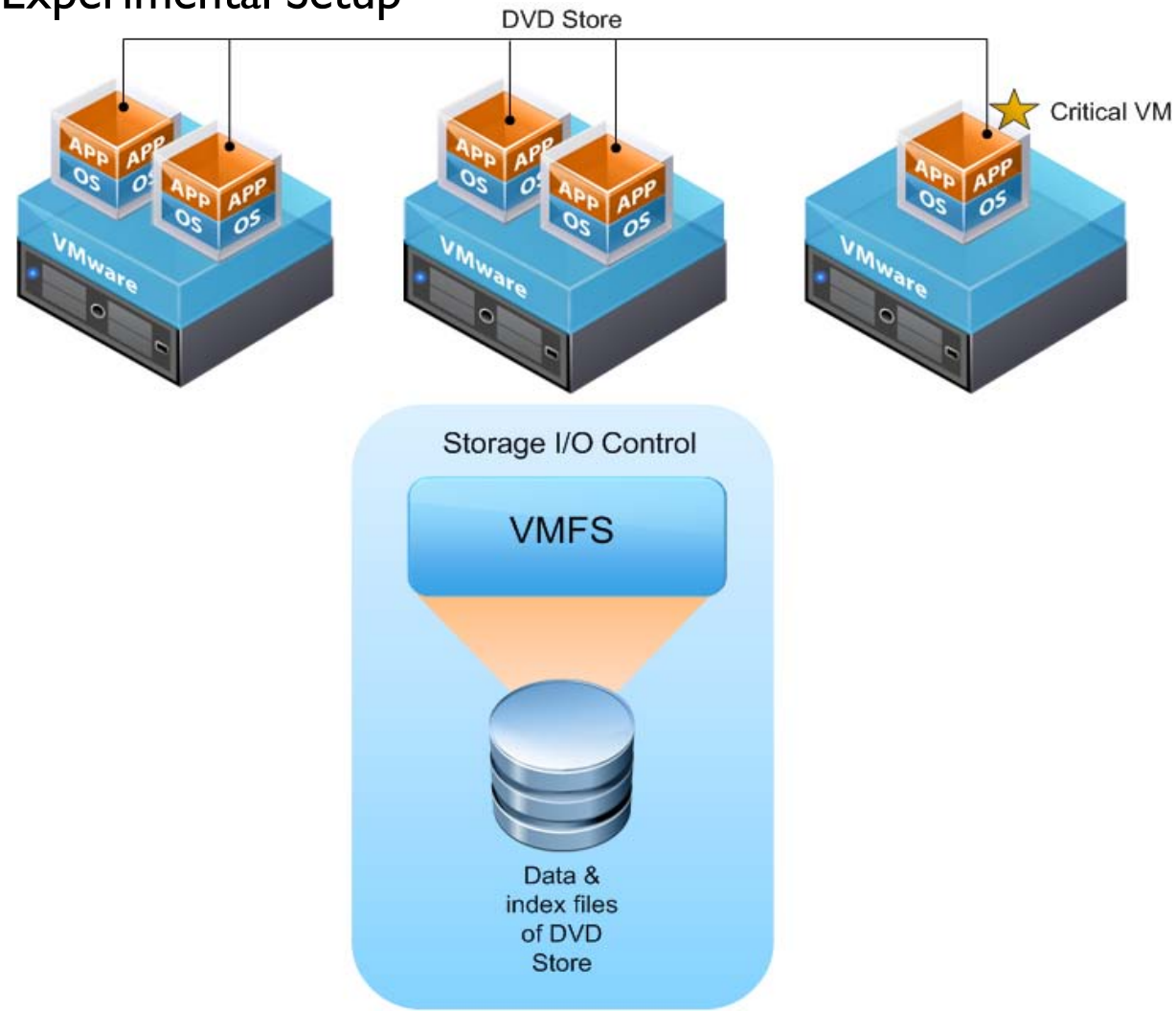
Outline

- The Problem
- Storage IO Control (SIOC) overview
- Technical Details
- SIOC in Action
  - Case study 1: Benefit of Disk Shares
  - Case study 2: Dynamic IO Prioritization
- Conclusions

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Dynamic IO Prioritization

Experimental Setup

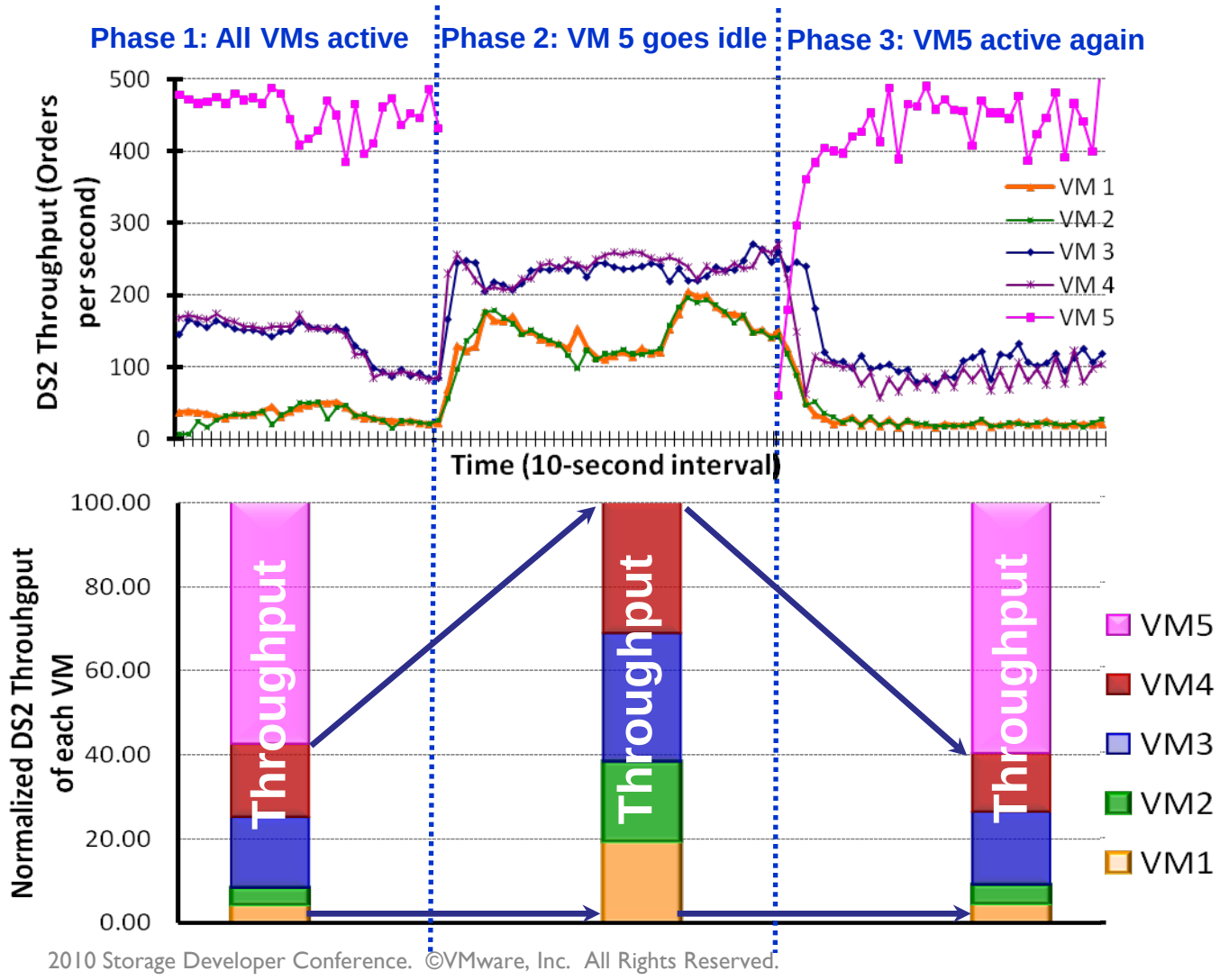


VM ID	Number of DS2 Users
1	24
2	24
3	24
4	24
★ 5	50

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Dynamic IO Prioritization

Effect of dynamic I/O prioritization on performance of DVDStore workloads



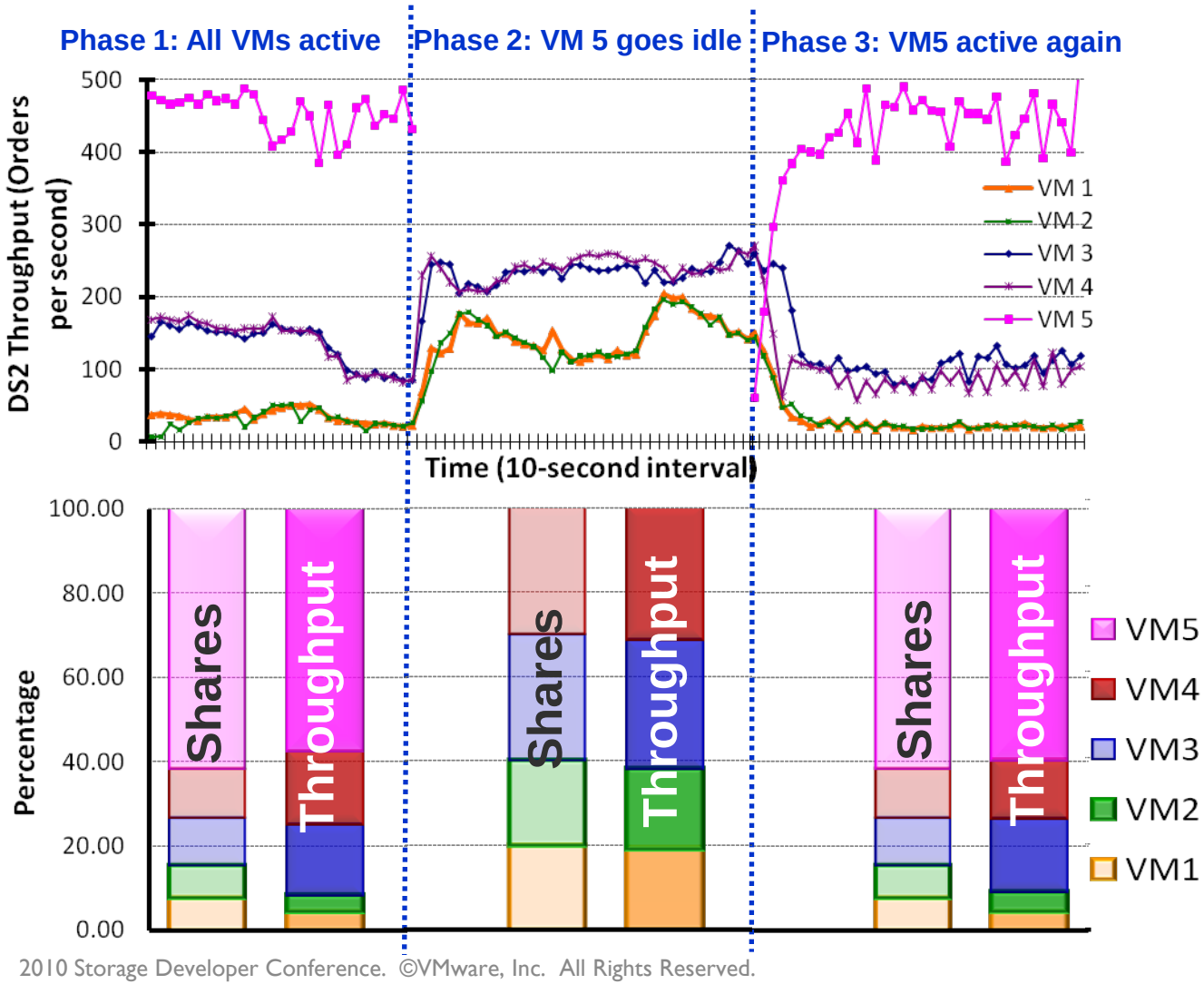
VM ID	Disk Shares
1	500
2	500
3	750
4	750
★ 5	4000

Congestion Threshold: 20ms

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Dynamic IO Prioritization

Effect of dynamic I/O prioritization on performance of DVDStore workloads

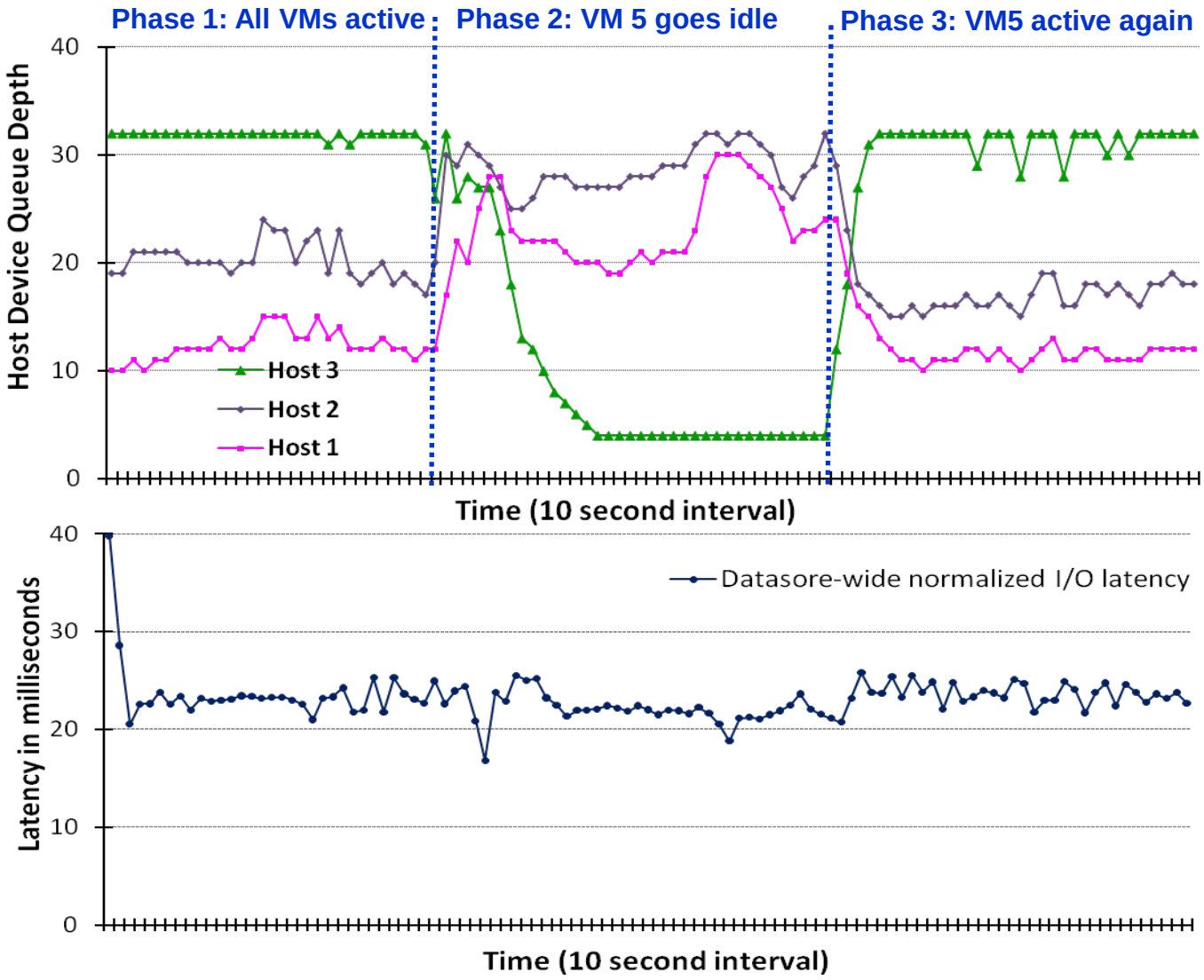


VM ID	Disk Shares
1	500
2	500
3	750
4	750
★ 5	4000

Congestion Threshold: 20ms

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Under the Hood …



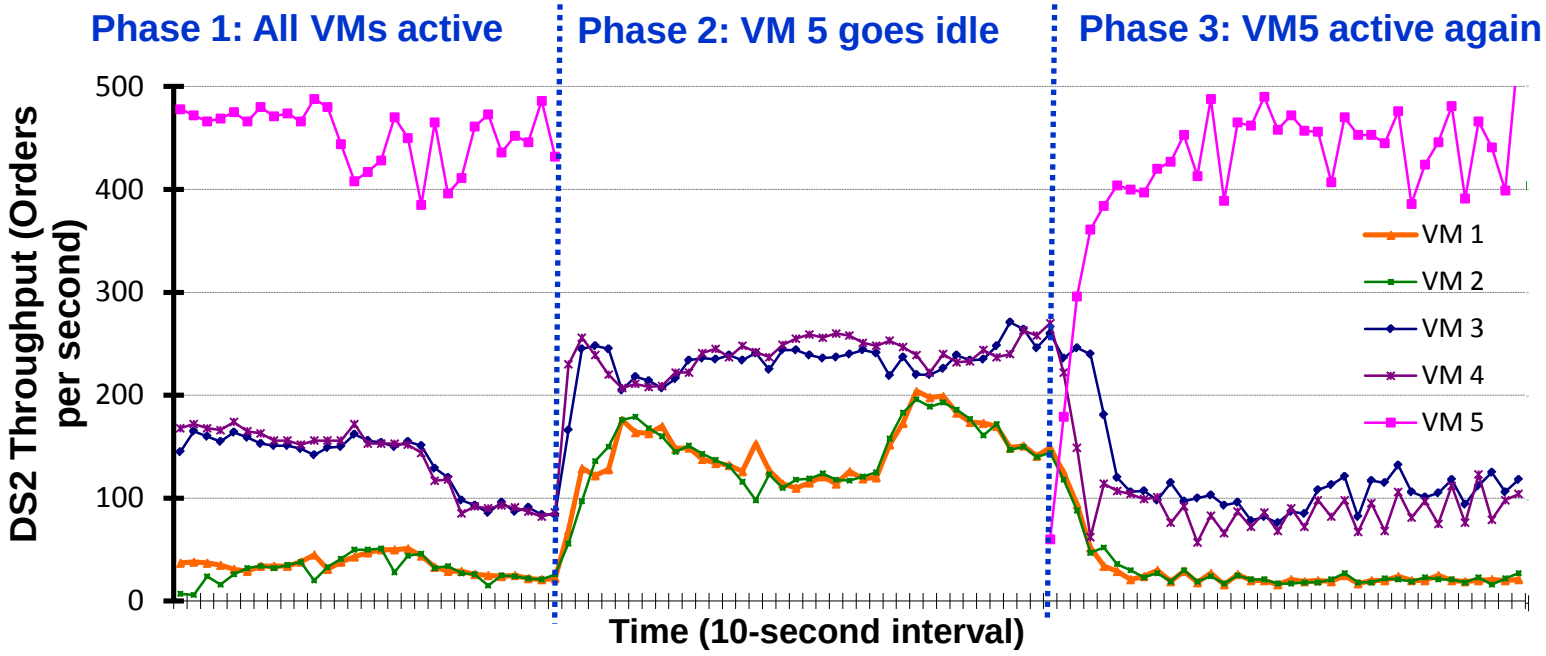
Host ID	Host disk Shares
1	1000
2	1500
★ 3	4000

Congestion Threshold: 20ms

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Dynamic IO Prioritization

- I/O Prioritization based on
  - Disk Shares (set by User)
  - Usage of allocated resources (monitored by SIOC)



SIOC maintains high utilization of storage devices

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Outline

- The Problem
- Storage IO Control (SIOC) overview
- Technical Details
- SIOC in Action
  - Case study 1: Benefit of Disk Shares
  - Case study 2: Dynamic Prioritization
- Conclusions

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Conclusions

- **A need for resource control for shared storage in virtual environments**
- **VMware's solution: Storage IO Control**
  - Control VMs access to shared storage using "Disk Shares"
  - Easy to use – just two steps
    - Enable Storage IO Control on a Datastore
    - Set Disk shares (and limit values) for virtual disks
- **SIOC is smart**
  - Automatic detection of I/O congestion
  - Dynamic decisions

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

Related Resources

- USENIX Annual Technical Conference 2009 paper

  "PARDA: Proportional Allocation of Resources for Distributed Storage Access"

  - Paper (http://www.usenix.org/events/fast09/tech/full_papers/gulati/gulati.pdf)
  - Slides (http://www.usenix.org/events/fast09/tech/slides/gulati.pdf)

- Managing Performance Variance of Applications using Storage I/O Control
  http://www.vmware.com/files/pdf/techpaper/vsp_41_perf_SIOC.pdf

- vSphere Resource Management Guide for ESX / ESXi / vCenter Server 4.1
  http://www.vmware.com/pdf/vsphere4/r41/vsp_41_resource_mgmt.pdf

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.

THANK YOU

2010 Storage Developer Conference. ©VMware, Inc. All Rights Reserved.