Removing the I/O Bottleneck in Enterprise Storage

WALTER AMSLER, SENIOR DIRECTOR

HITACHI DATA SYSTEMS

AUGUST 2013

Agenda

- Enterprise Storage
  - Requirements and Characteristics
  - Reengineering for Flash – removing I/O bottlenecks
- Measuring Performance
  - Application Performance Metrics vs Synthetic Benchmarks Numbers
- Summary

Requirements for Storage Systems

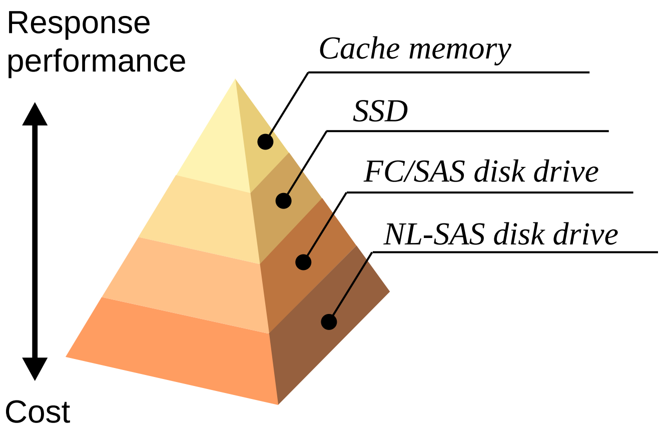
EFFECTIVELY ADDRESSING TECHNOLOGY AND BUSINESS CHALLENGES

- What Customers are demanding
  - Reduce cost, optimize service-level delivery via scale-up, dynamic provisioning and tiering across different media types
  - Management abstraction to enable ease of use, speed and automation
  - 24x7xforever application availability, eliminate planned and unplanned outages
  - Reliability, Availability Serviceability (RAS)
- Recent Trends in High-End Storage
  - Storage Subsystems are designed for the Virtual data center
  - Storage Infrastructure is transformed in Storage Services
  - Exploitation of loosely coupled vs tightly coupled Architectures

Characterizing Storage systems

ANOTHER FORM OF RAS: REDUNDANCY, ARCHITECTURE, SCALABILITY

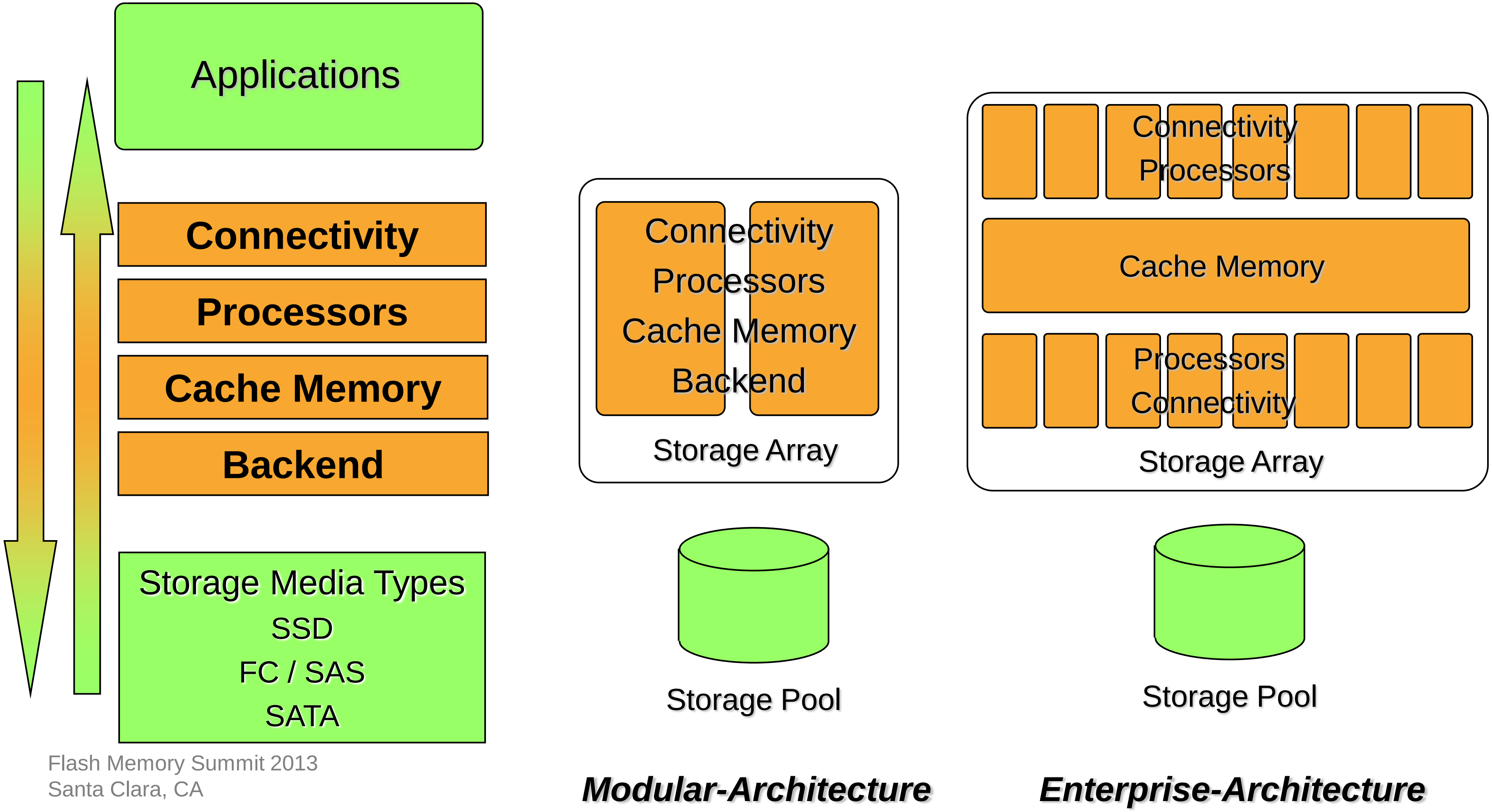
- Storage Architectures and design – different value propositions
  - Modular Architecture vs Enterprise Architecture
  - Component/Site Redundancy
- Performance
  - Time is Money – must cope with peak demands and satisfy strict SLA's



- Functionality
  - Virtualisation
  - Dynamic Tiering
  - In-System Snapshots and Clones
  - 2DC and 3DC Sync and Asynchronous Replication
  - Rich GUI/CLI Management Capabilities – Ease of Use

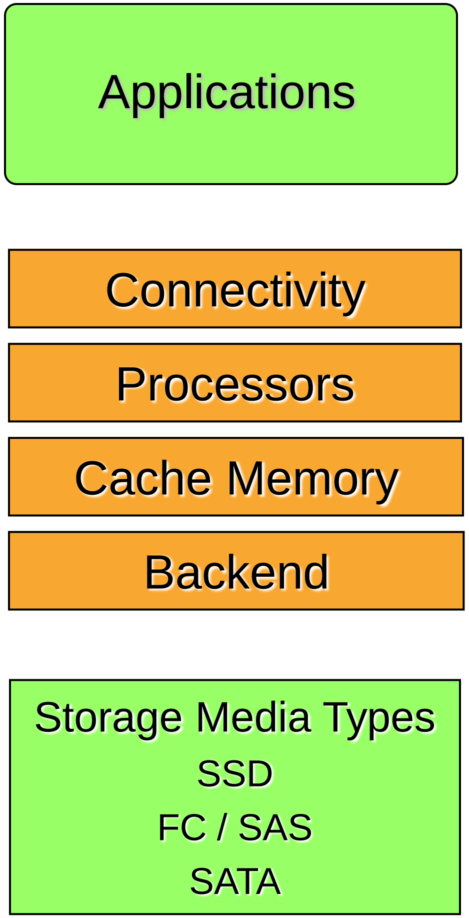
Modular vs Enterprise Architecture

BALANCING COST, SCALABILITY, PERFORMANCE AND CAPACITY



Modular storage growth – Scale-Out

ADD MORE OF THE SAME – BUT BEWARE OF ISLANDS



Applications

Connectivity

Processors

Cache Memory

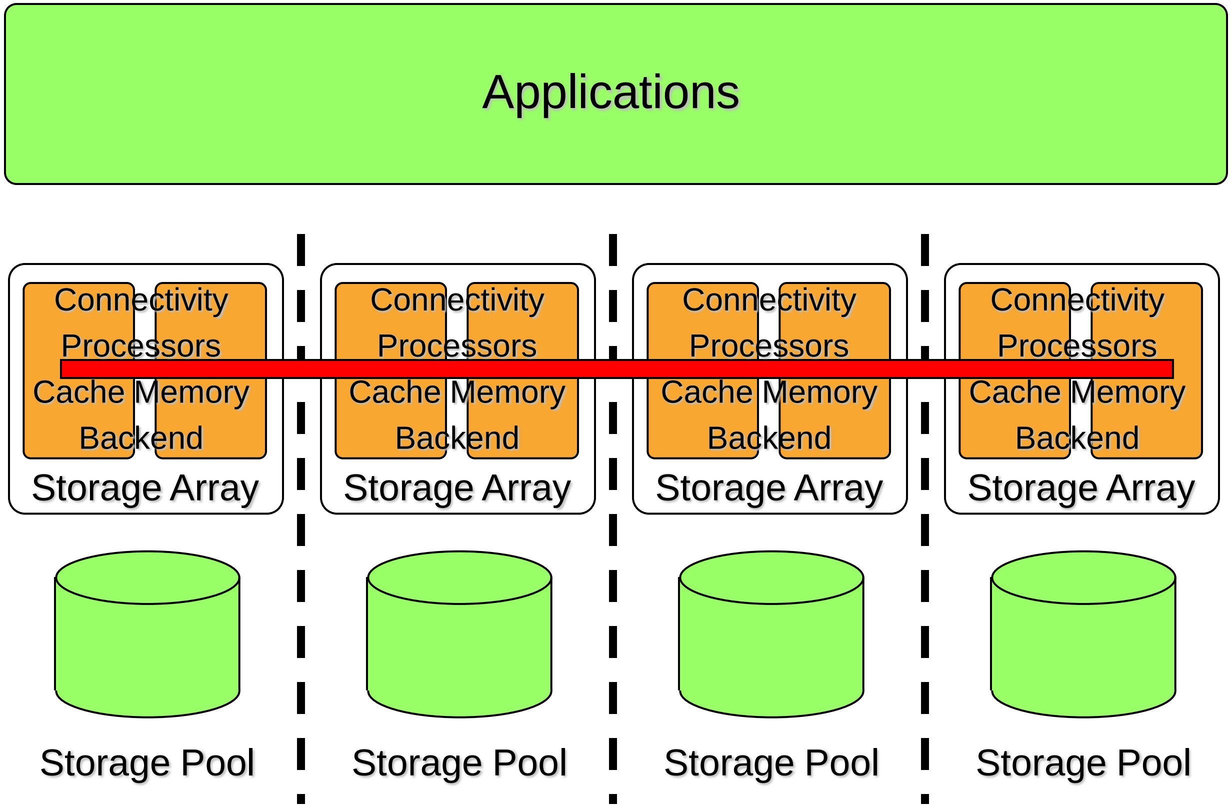
Backend

Storage Media Types

SSD

FC / SAS

SATA



Applications

Connectivity Processors Cache Memory Backend Storage Array

Connectivity Processors Cache Memory Backend Storage Array

Connectivity Processors Cache Memory Backend Storage Array

Connectivity Processors Cache Memory Backend Storage Array

Storage Pool

Storage Pool

Storage Pool

Storage Pool

Loosely coupled Architecture

Enterprise storage growth – Scale-Up

EXPAND CAPACITY, CONNECTIVITY AND PROCESSING POWER

Applications

Connectivity

Processors

Cache Memory

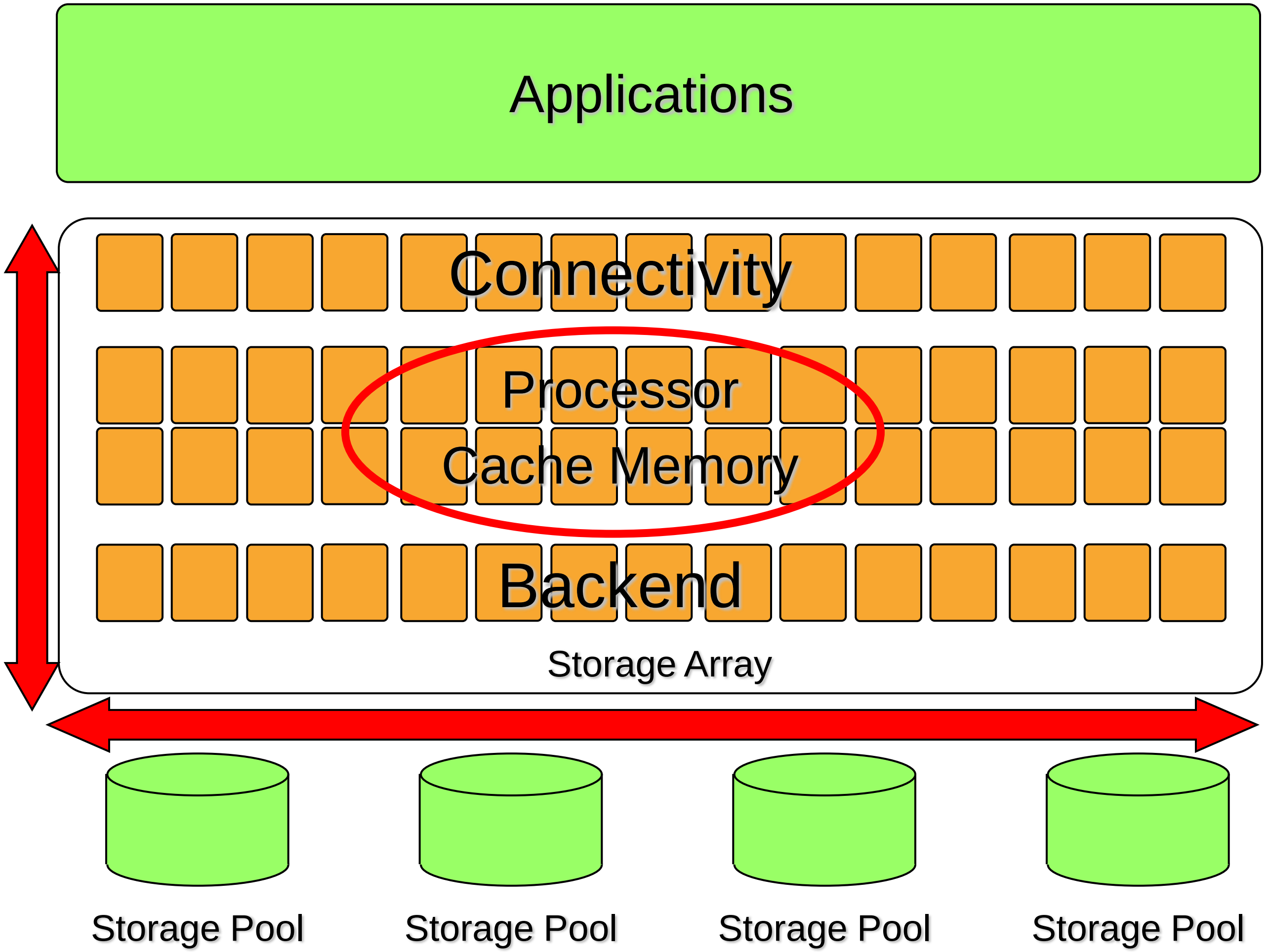
Backend

Storage Media Types

SSD

FC / SAS

SATA



Applications

Connectivity

Processor

Cache Memory

Backend

Storage Array

Storage Pool

Storage Pool

Storage Pool

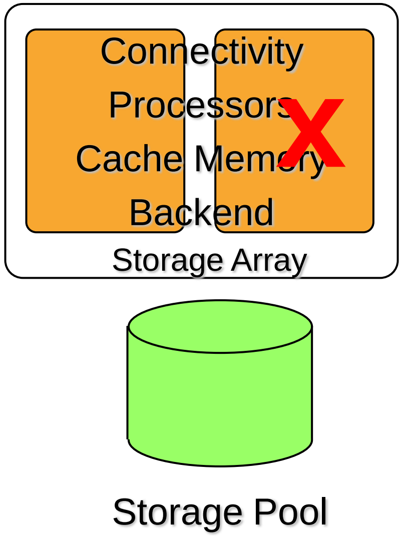
Storage Pool

Tightly coupled Architecture

When things go wrong

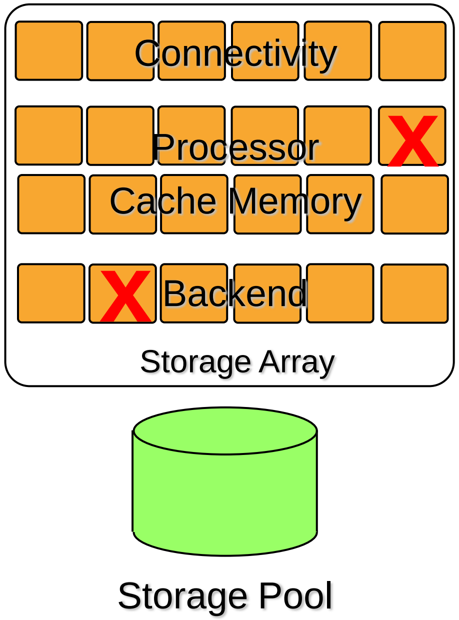
FAILURE IMPACT - GOOD ENOUGH VS BULLET PROOF

- Availability depends on failure domains and the choice of component/site redundancy options
  - Bulletproof storage array: http://www.youtube.com/watch?v=Gnjb1WVkhmU



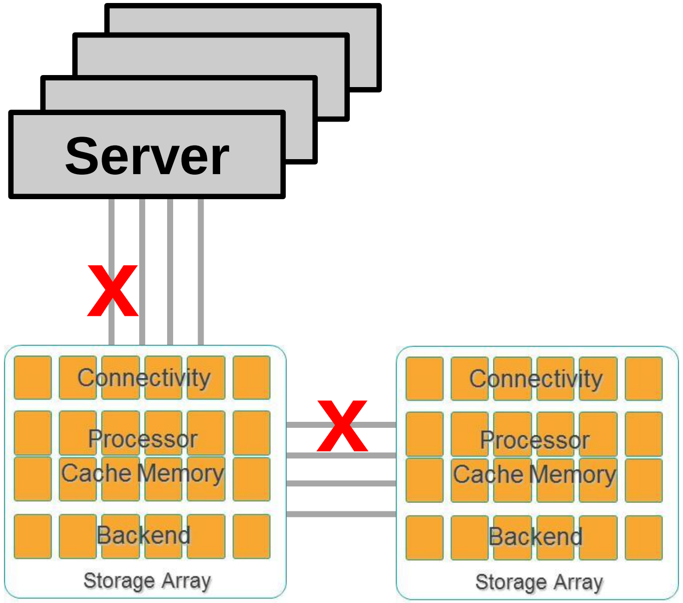
Modular-Architecture

Probably System down



Enterprise-Architecture

Relatively little impact



Component Redundancy

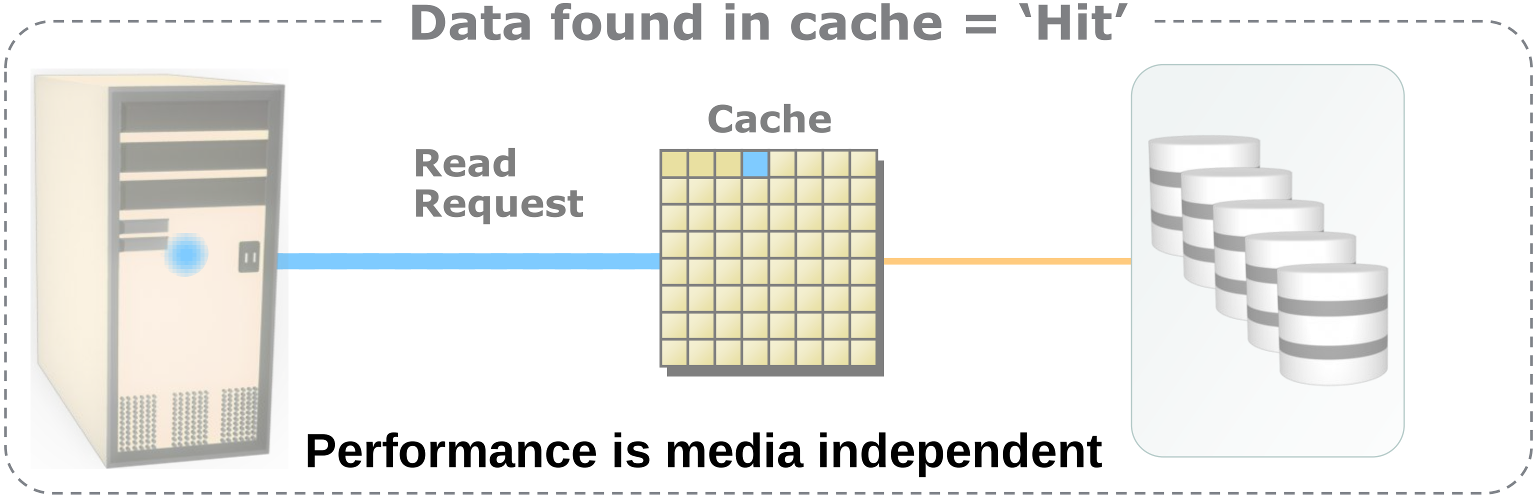
25% loss of connectivity

I/O Bottleneck in Enterprise Storage

BUILT FOR FLASH FROM THE GROUND UP VS RE-ENGINEERED

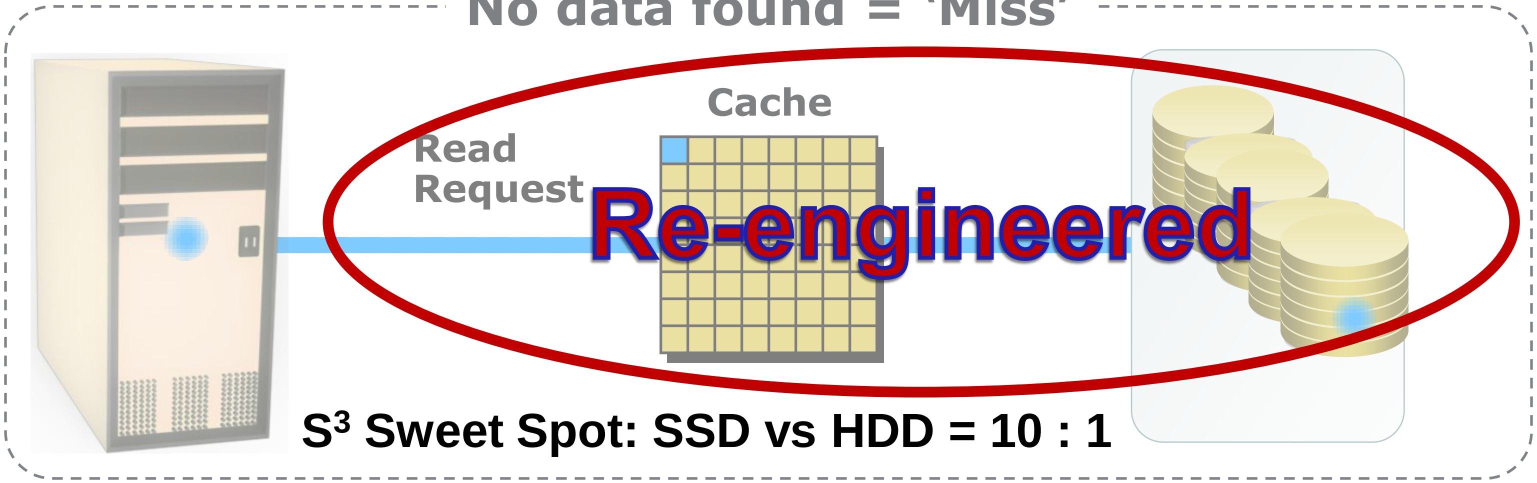
- Traditional Storage Arrays
  - originally designed for hundreds, then thousands of HDD's
  - Ever larger DRAM Cache and sophisticated Algorithms mitigate/hide HDD performance characteristics
    - Works great for sequential read/write
    - Works very well for Random I/O with good Locality of reference
- The IO Gap
  - Moore's Law - processor speed has increased dramatically
  - HDD Speed (Seek and RPM) has virtually stayed the same
  - server virtualization randomizes I/O, LOR is lost, aka «I/O Blender»
- The Emergence of Flash demands a new approach

Read IO Operation; Cache Hits and Misses



Data found in cache = 'Hit'

Performance is media independent



No data found = 'Miss'

S^3 Sweet Spot: SSD vs HDD = 10 : 1

Hitachi Innovation: Flash Acceleration

OPTIMIZING STORAGE SYSTEM SOFTWARE TO EXPLOIT FLASH

- 30+ fundamental software changes to turbo-charge performance with Flash
  - New “express” I/O processing
  - New Cache Slot Allocation method
  - Reduced ucode Overhead and path length
- Significant performance impacts
  - Up to 65% reduction in response time
  - Up to 4X Random IO scalability
- Non-disruptive installation and transparent to current applications

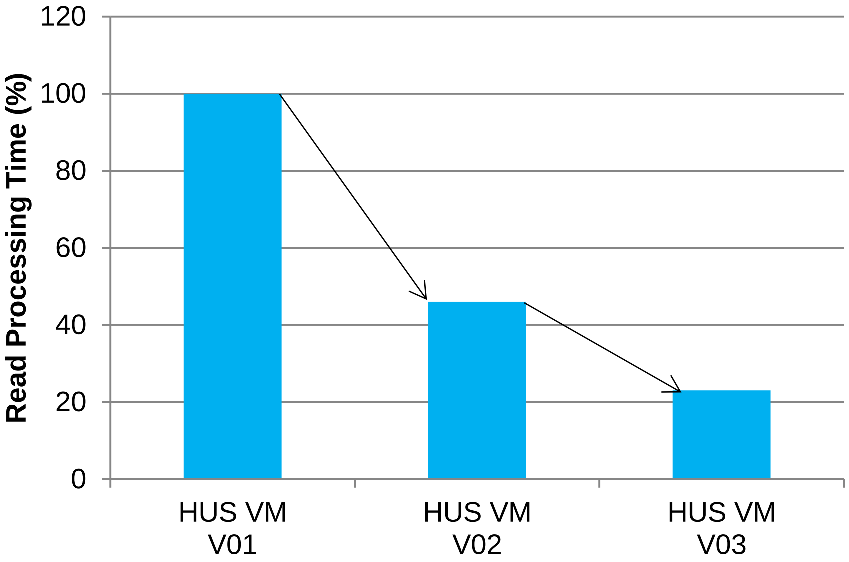
Hitachi Storage

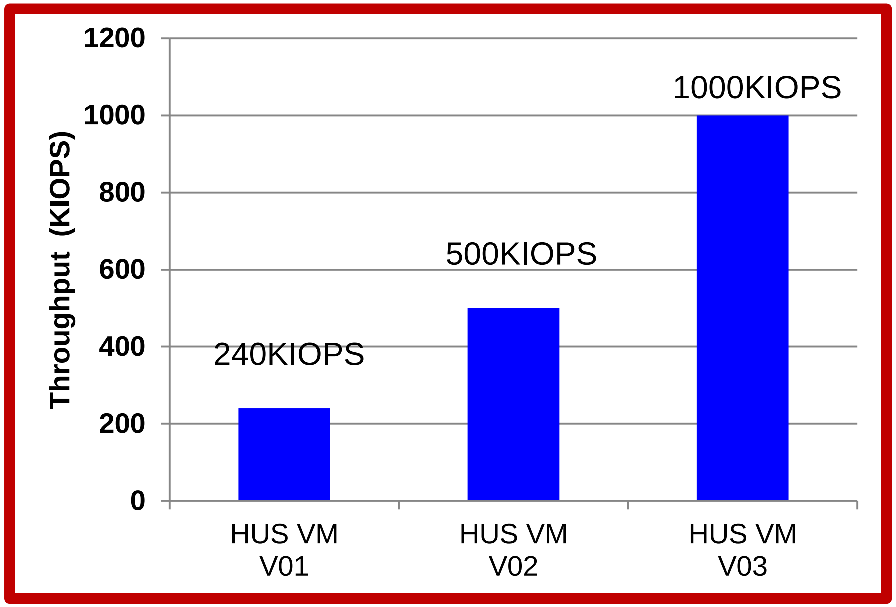
HUS VM & VSP
Optimized for
Flash
Today

Flash Acceleration Impact for all flash array

145 PATENTS RELATED TO HITACHI FLASH TECHNOLOGY

- Backend Codepath reduction, logic and ASIC optimization
  - Version 1: Basic Design for HDD – non optimized
  - Version 2: BE/FE Job Integration, Cache Buffer Slot Management
  - Version 3: use DXBF, avoid CTL to CTL communication Improve CPU L1 Cache Hit Rate for Instructions



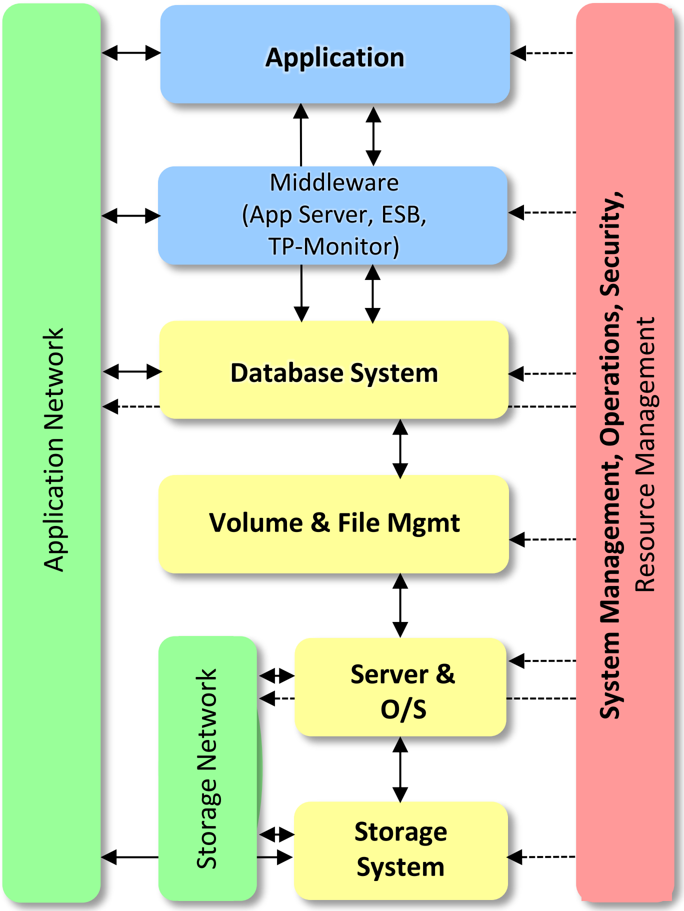


Latency and what does it really mean

MEASURING PERFORMANCE - RELEVANCE TO YOUR BUSINESS

- Vendor Provided Measurement Data
  - Objective is to show «champion numbers»
  - Customers need to have a complete understanding of what was measured and how, for example:
    - 80 usec Latency: single 512Byte Block Read measured at Fibre Channel Port with a Fibre Channel Analyzer
    - 1 Million IOPS: 4KB Random Reads measured by IOMETER
  - Interesting, but not relevant from an application perspective
- Need a different approach
  - Include and consider all the different technology layers of entire platform
  - Example: Oracle Database Platform Architecture

Oracle Database Platform Architecture



Complexity of Oracle platform

Application Network (IP-based)
Bandwidth, latency during remote database mirroring (sync, async) due to switches and sql*net and tcp/ip stack (frame size, …).

Oracle Database
Different versions, patches and options, about hundred configuration parameters.

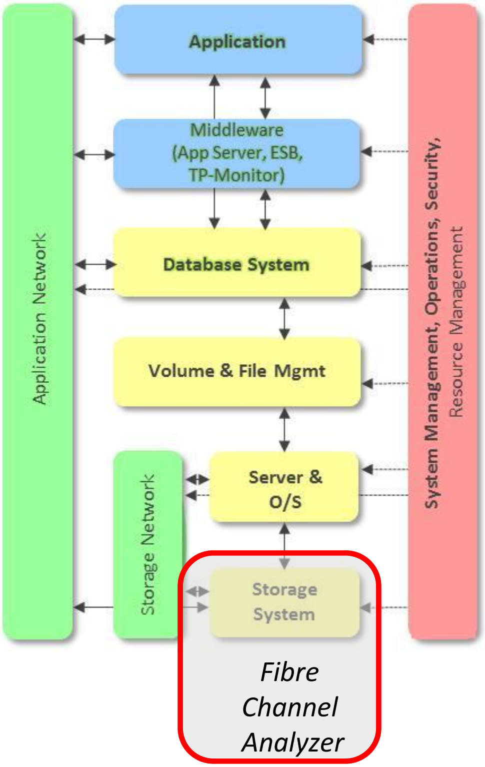
Volume & File Management
Different volume managers (VxVM, ASM) and file systems (UFS, VxFS, ext3, JFS, ZFS, raw devices), different I/O methods (async, direct), a lot of config parameters (#LUNS, queue depth, max i/o unit), software striping and/or mirroring, multipathing.

Storage Network (IB-, FC- or IP-based)
Bandwidth, latency during remote storage mirroring (sync, async) due to switches, hubs and distance.

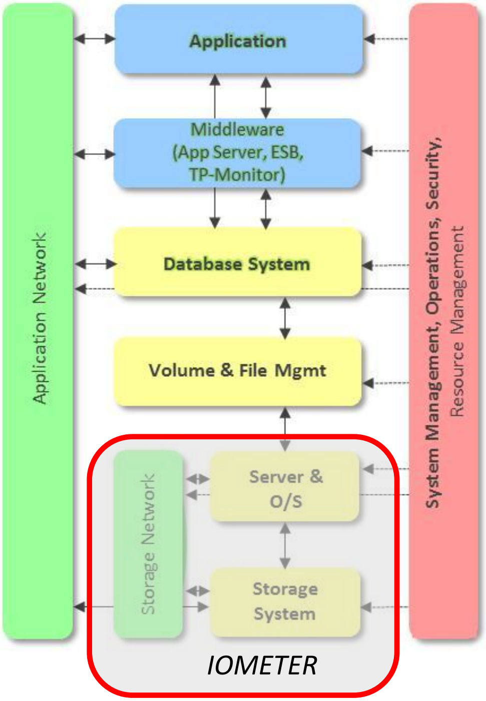
Server & Operating System
Different server systems, processors and CPU architectures, (x86, IA-64, UltraSparc, SPARC64, Power), #cores, multithreading, main memory, bus architecture. Different operating systems and patches, over hundred configuration parameters, virtualization of resources.

Storage System
Different storage systems, storage tiers and storage technology: spindle count and speed, RAID management, cache management, server interface technology, storage system options like remote copy, hardware striping and/or mirroring, virtualization of resources.

Measuring Oracle Application Performance

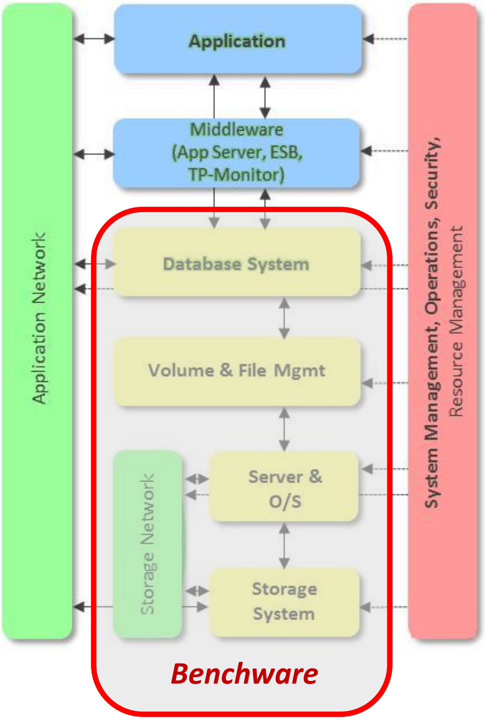


Measuring Hardware



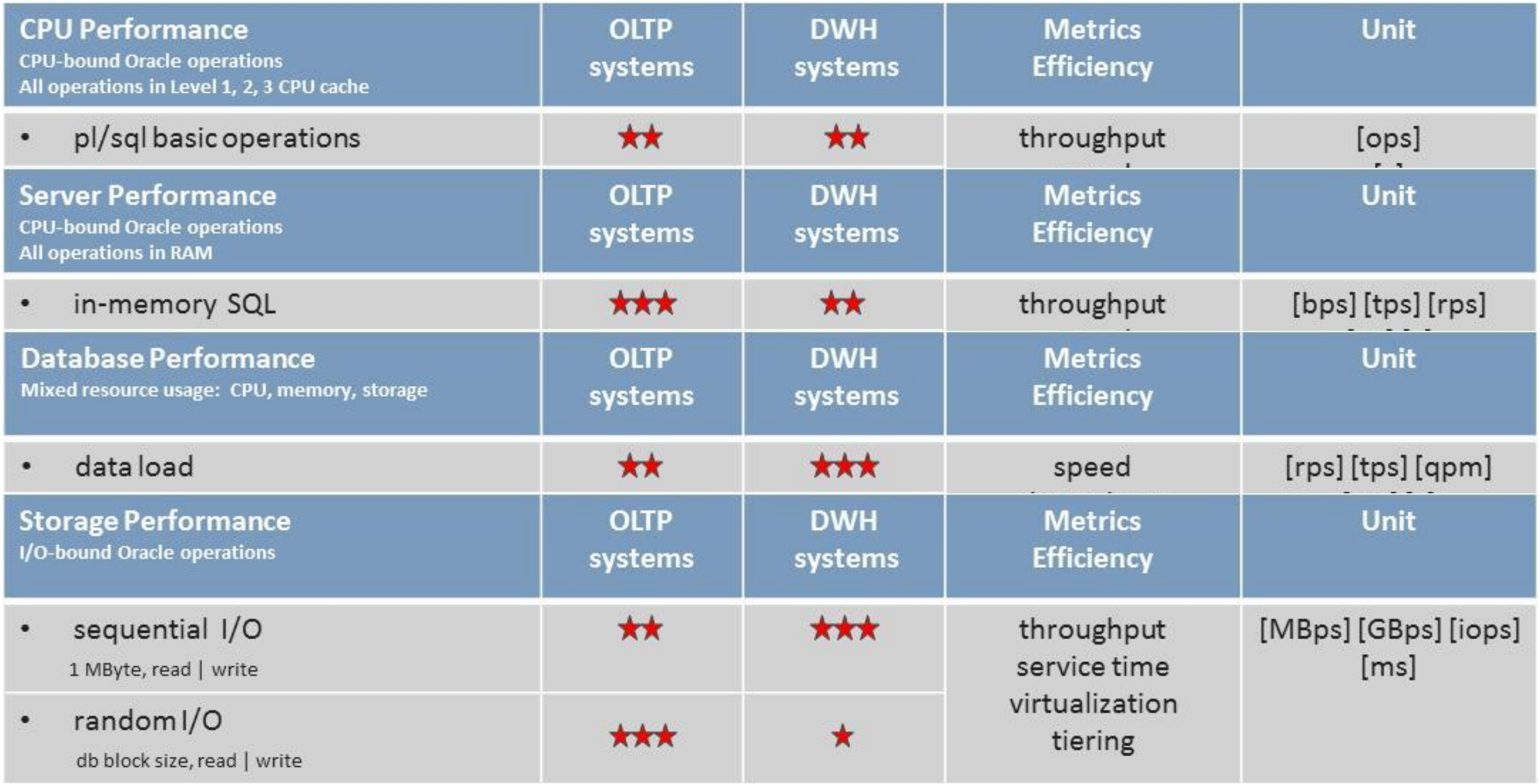
Measuring Server and Storage and «Mindless» IO

Measuring complex Application I/O or customer reality



Benchware Approach

Library of Oracle benchmark tests - implemented in PL/SQL, Java and SQL

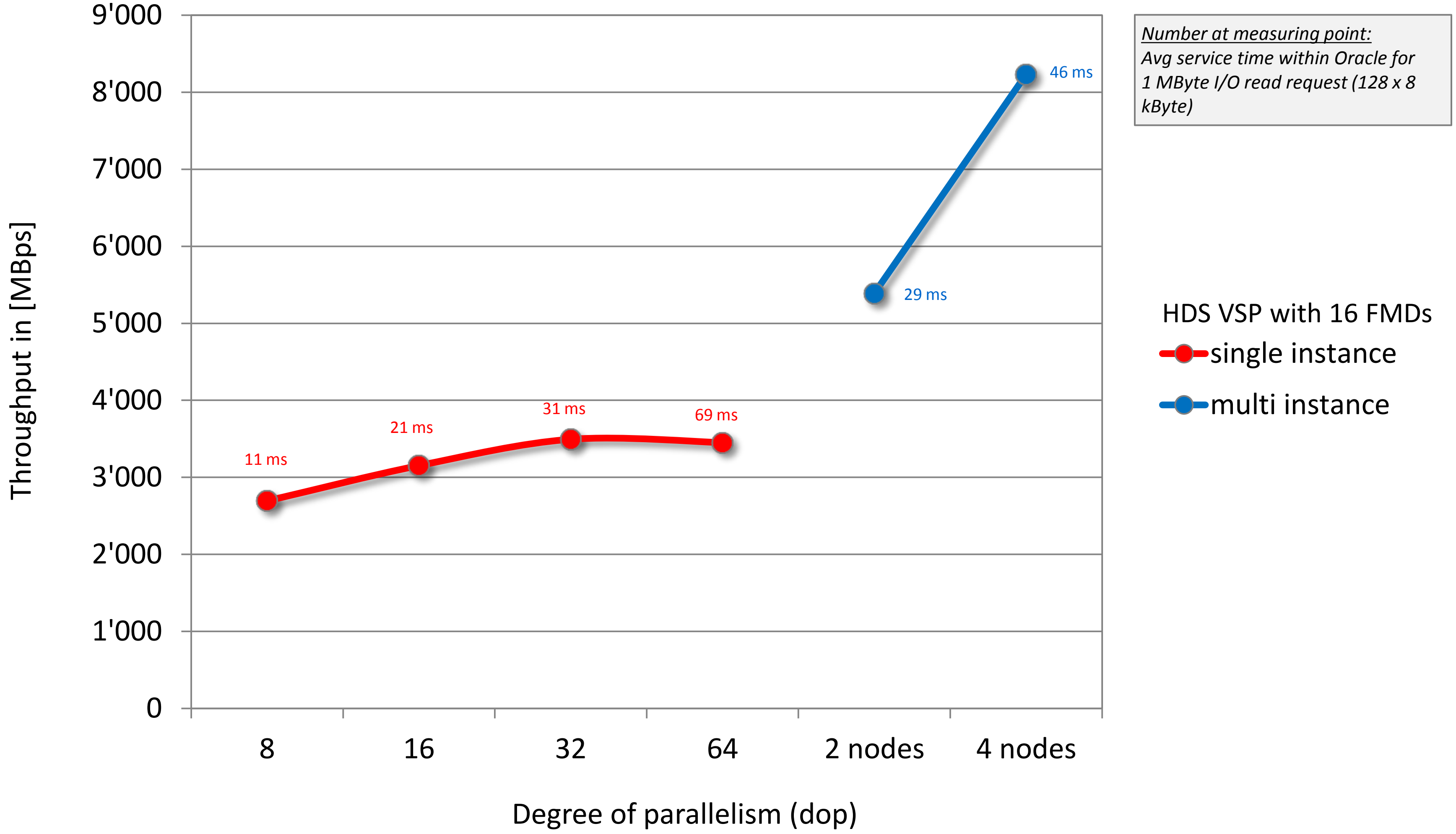


CPU Performance CPU-bound Oracle operations All operations in Level 1, 2, 3 CPU cache	**OLTP systems**	**DWH systems**	**Metrics Efficiency**	**Unit**
• pl/sql basic operations	★★	★★	throughput	[ops]
Server Performance CPU-bound Oracle operations All operations in RAM	**OLTP systems**	**DWH systems**	**Metrics Efficiency**	**Unit**
• in-memory SQL	★★★	★★	throughput	[bps] [tps] [rps]
Database Performance Mixed resource usage: CPU, memory, storage	**OLTP systems**	**DWH systems**	**Metrics Efficiency**	**Unit**
• data load	★★	★★★	speed	[rps] [tps] [qpm]
Storage Performance I/O-bound Oracle operations	**OLTP systems**	**DWH systems**	**Metrics Efficiency**	**Unit**
• sequential I/O 1 MByte, read | write	★★	★★★	throughput service time virtualization tiering	[MBps] [GBps] [iops] [ms]
• random I/O db block size, read | write	★★★	★		

[s]	seconds	[bps]	buffers per second	[MBps]	mega bytes per second	★	less important
[ms]	milli seconds (10^{-3})	[rps]	rows per second	[GBps]	giga bytes per second	★★	important
[µs]	micro seconds (10^{-6})	[tps]	transactions per second	[iops]	i/o operations per second	★★★	very important
[ns]	nano seconds (10^{-9})	[ops]	operations per second	[qpm]	queries per minute		

Measuring Datawarehouse Workload

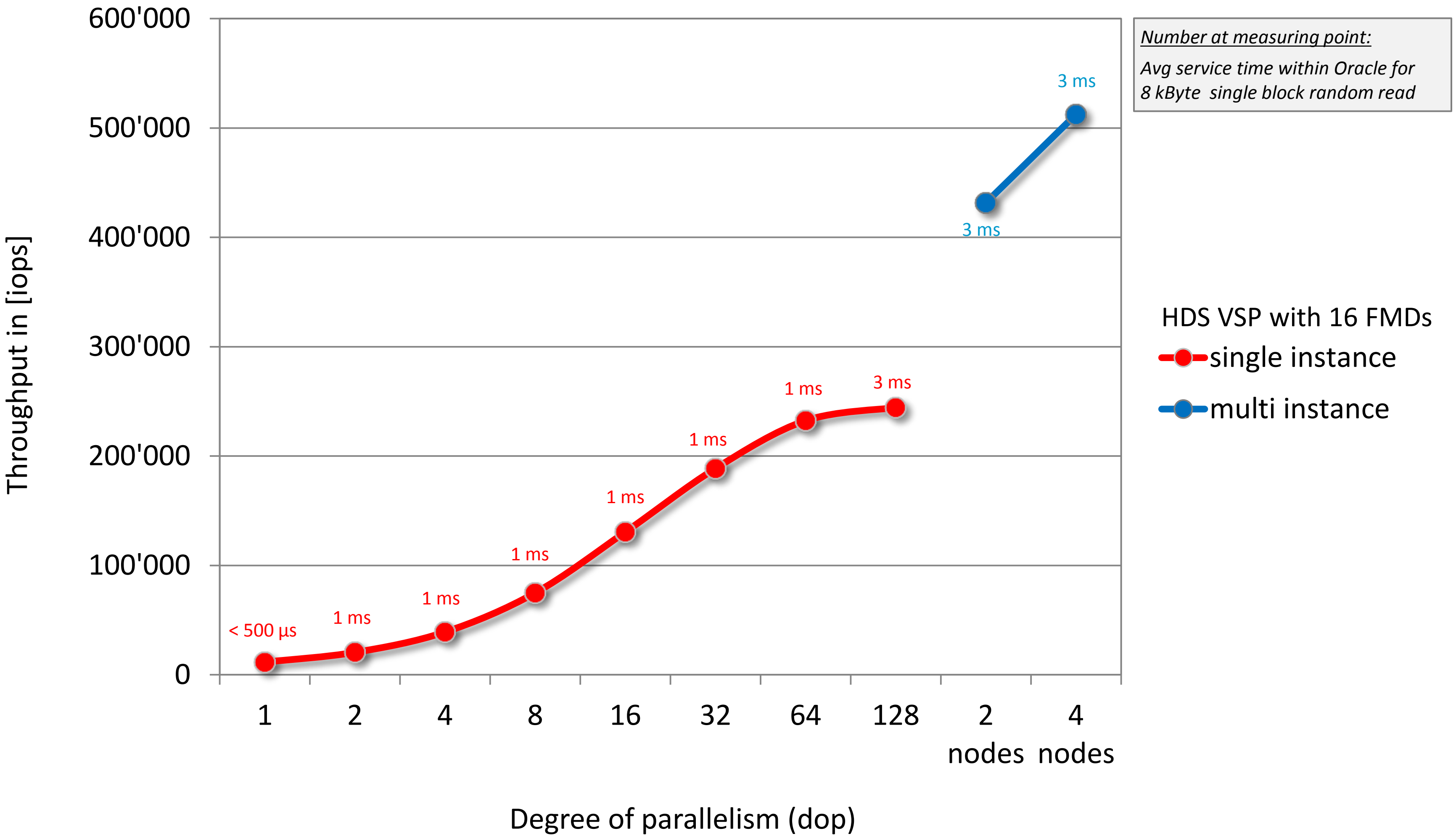
SEQUENTIAL READ, MULTIPLE PROCESSES – TYPICAL FOR DWH



__Number at measuring point:__
Avg service time within Oracle for 1 MByte I/O read request (128 x 8 kByte)

Measuring OLTP Workload

8KB RANDOM READ; 100% CACHE MISS - TYPICAL FOR OLTP



Number at measuring point:

Avg service time within Oracle for 8 kByte single block random read

What does it mean to your business?

KEY PERFORMANCE METRICS LEAD TO SERVICE LEVEL AGREEMENTS

- The measured server/storage platform will deliver:
  - 8GB/sec sequential Read throughput for your DWH
  - 250’000 8KB Random Reads with Zero Cache Hits for your OLTP application with a Response Time of less than 3 Milliseconds

- Note: Oracle Measurements for Random Read IO
  - Oracle currently does not understand «Microseconds»
  - Response Time for Random Read is reported in Milliseconds, and data is rounded e.g. 0 MS or 1 MS
- R/T for high Random Read I/O Rates generally at 1-4 MS
  - This also applies to All Flash Appliances/Arrays

Summary

- Enterprise Storage today has a lot to offer
  - RAS: Reliability, Availability, Serviceability
  - Superior Performance
  - Seamless Scale-Up Architecture
- Flash Storage Exploitation
  - Value of Re-engineering equals «Built from scratch»
  - In addition you get the functionality and EoU you need
- Performance and Latency Claims
  - Must understand what is being measured and how
  - The key is the mileage you get for your application!

Questions and discussion

Thank you