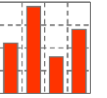


Flash Technology for Oracle Database Server

Technical Presentation

May 2014



1 Introduction to Storage Performance Tests

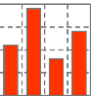
2 Storage Tiers

3 Storage Architectures

4 Benchmark Results

5 Summary

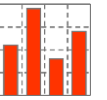
Introduction to Storage Performance Tests



Why measure Storage performance?

- Storage performance is essential not only for overall Oracle database performance, but also for system management tasks like backup, recovery and archiving
- Oracle uses all kinds of I/O patterns, but different o/s calls dependent upon the
 - operating system
 - system load (Oracle changes system call dependent on load)

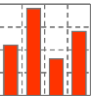
Introduction to Storage Performance Tests



Why measure Storage performance?

- Oracle sequential read
 - User processes: full table scan, full index scan
 - Temp segment
 - Backup, restore, recovery RMAN, Export, Data Pump
 - ARCH: reading online REDO logfile
- Oracle random read
 - User processes
- Oracle sequential write
 - Temp segment
 - Backup, restore RMAN, Export, Data Pump
 - LWGR process: small block size
 - ARCH processes: writing archived REDO logfile
 - RVWR process: flashback log file writer
 - CTWR process: block change tracking file
- Oracle random write
 - DBWR processes

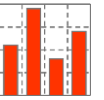
Introduction to Storage Performance Tests



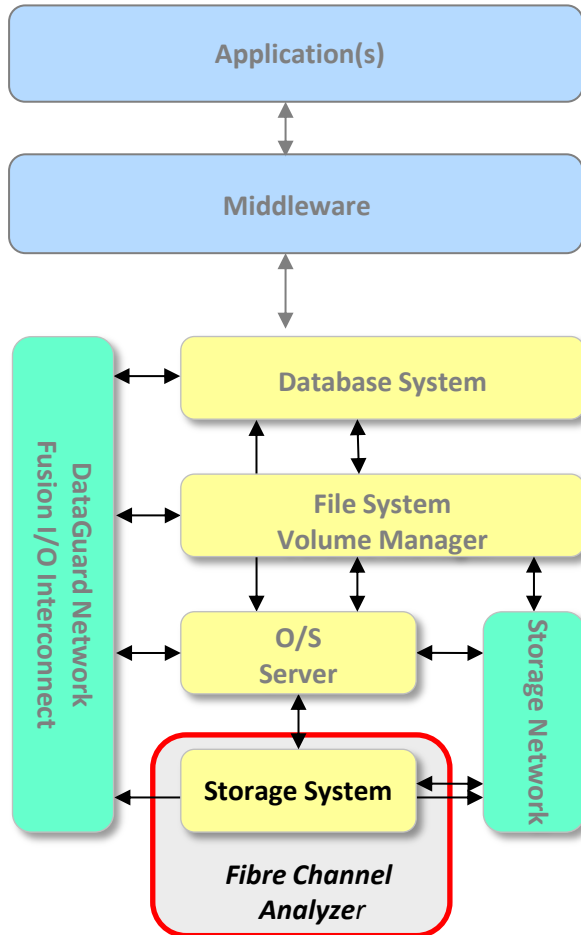
What is measured?

- Storage performance from the Oracle point of view
 - Using database block size
 - I/O service time measured within Oracle
- Throughput
 - Data transfer in mega byte per second [MBps]
 - Data transfer in database blocks per second [dbps]
 - I/O Operations in O/S system calls per second [IOPS]
- Service Time
 - For random I/O operation in [ms] or [μ s]
- Efficiency of
 - Auto-Tiering
 - RAID-level
 - Striping
 - Remote mirroring
 - Virtualization

Introduction to Storage Performance Tests



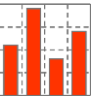
Comparison of I/O Benchmarks



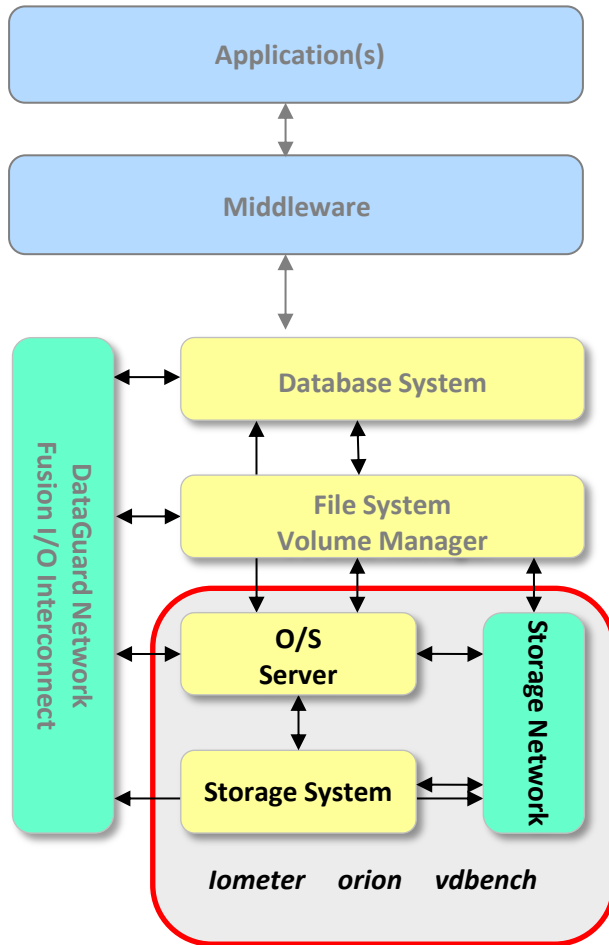
■ Storage System I/O Performance

- Useful to test storage system performance at port level
- Vendors data sheet numbers

Introduction to Storage Performance Tests



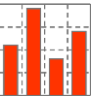
Comparison of I/O Benchmarks



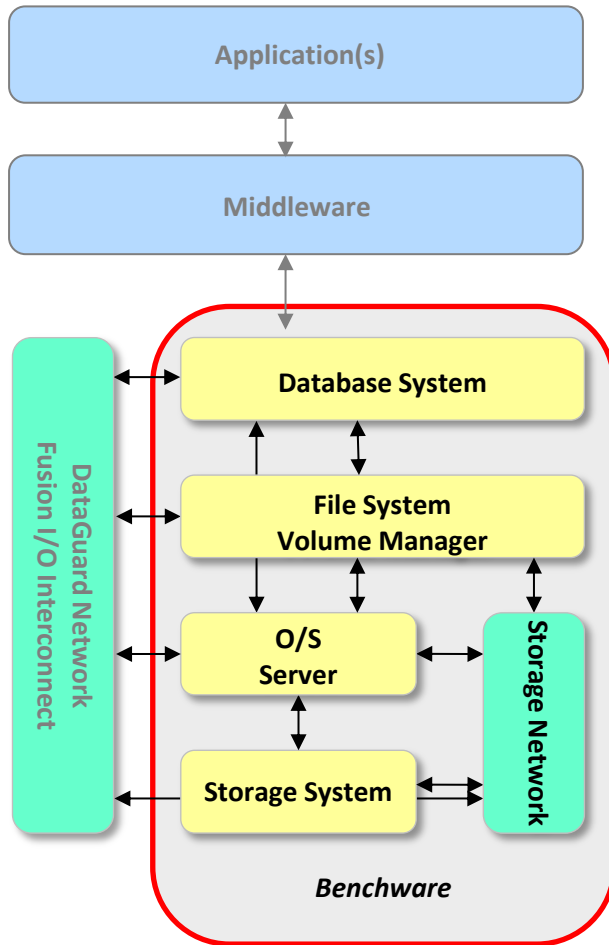
■ Server System I/O Performance

- Tools like vdbench, lometer, Orion, etc. just generate I/O system calls, but no further I/O processing
- Useful to analyze transfer performance between storage system and server system
- Unable to benchmark storage grids
- Unable to benchmark Oracle ASM infrastructure

Introduction to Storage Performance Tests

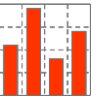


Comparison of I/O Benchmarks

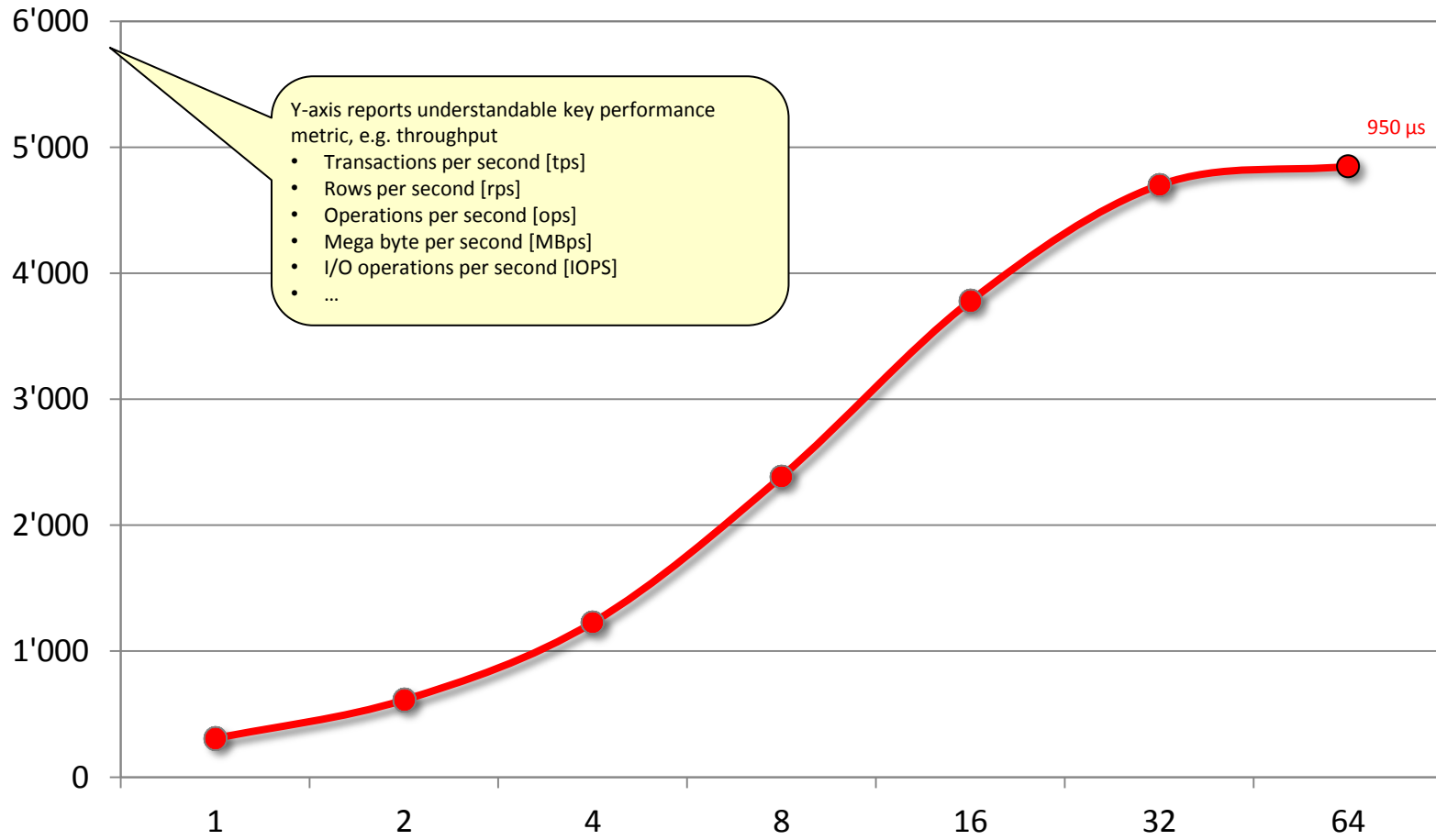


- Database System I/O Performance
 - Most complex I/O operation
 - Database buffer cache management
 - » find a free slot
 - » replace older blocks
 - » synchronize access to buffer cache
 - » database block consistency checks
 - Database I/O needs much more cpu resources than simple I/O generator
 - » Rule of thumb: 25'000 IOPS per x86 core
 - » Throughput does not scale linear
 - `dbms_resource_manager.calibrate_io` does not recognize hybrid storage systems and delivers wrong results

Introduction to Storage Performance Tests



All load profiles from single process to saturation



Y-axis reports understandable key performance metric, e.g. throughput

- Transactions per second [tps]
- Rows per second [rps]
- Operations per second [ops]
- Mega byte per second [MBps]
- I/O operations per second [IOPS]
- ...

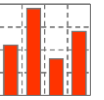
Optional additional information for each measuring point describes quality of performance, e.g.

- I/O service time in μ sec
- Operations per core
- ...

X-axis reports benchmark load, e.g. degree of parallelism

- Inter SQL or Intra SQL dependent on test
- Parallelism 1 shows best case: only one process
- Number of involved RAC nodes

Introduction to Storage Performance Tests



Monitoring

- Speed:
- Best performance for one process
 - No conflicts
 - No contention

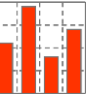
■ I/O throughput in MBps and IOPS

Run	Tst	Code	#N	#J	#T	CPU busy [%]	CPU sys [%]	Physical read [iops]	Physical read [dbps]	Physical read [MBps]	Physical write [iops]	Physical write [dbps]	Physical write [MBps]	REDO write [iops]	Hitrate db flash [%]	Hitrate exa flash [%]	Elap time [s]
2	20	STO-14	1	1	1	1	0	10082	1280024	10000	27	29	0	6	0	67	300
	21	STO-14	1	2	1	1	0	25605	3255841	25437	20	25	0	2	0	94	301
	22	STO-14	1	4	1	2	0	49250	6270027	48985	20	24	0	3	0	94	300
	23	STO-14	1	8	1	3	0	87799	11208636	87568	21	25	0	3	0	88	302
	24	STO-14	1	16	1	3	1	98534	12582808	98303	22	26	0	3	0	87	302
	25	STO-14	1	32	1	3	1	103230	13184237	103002	23	26	0	4	0	87	304
	26	STO-14	1	64	1	4	1	105525	13480546	105317	24	27	0	5	0	86	308

Legend:

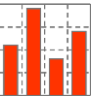
- | | | | | | | | | | |
|------|---------------------|----|----------------------------------|-------|-------------------------|--------|----------------------------|------|-----------------------|
| Run | benchmark run id | #N | number of RAC nodes | [rps] | rows per second | [iops] | i/o operations per second | [s] | time in seconds |
| Tst | benchmark test id | #J | number of load generators (jobs) | [tps] | transactions per second | [dbps] | database blocks per second | [ms] | time in milli seconds |
| Code | benchmark test code | #T | number of threads (PX) | [ops] | operations per second | [MBps] | mega byte per second | [μs] | time in micro seconds |

- Max throughput:
- Storage system saturated



- 1 Introduction to Storage Performance Tests
- 2 Storage Tiers**
- 3 Storage Architectures
- 4 Benchmark Results
- 5 Summary

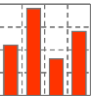
Architectures with Flash Technology



Capacity and access times

Server System	Capacity	Average Access Time	Throughput
CPU register	1 KByte	< 1 ns	
Level 1 Cache on-chip	128 KByte	1 ns	
Level 2 Cache	8 MByte	5 ns	
Level 3 Cache	32 MByte	15 ns	
Main memory	4 TByte	100 ns	Seq/Ran: > 10 GBps
PCI attached flash	n x 2.4 TByte	< 100'000 ns	Seq: > 1 GBps Ran: > 500'000 IOPS
Storage System			
Cache	<= 1 TByte	500'000 ns	
Solid state disks (SSD)	n x 1.4 TByte	1'000'000 ns	Seq: > 1 GBps Ran: > 200'000 IOPS
Hard drive disks (HDD)	n x 600 GByte, 15k rpm	8'000'000 ns	Ran: 250 IOPS
Hard drive disks (HDD)	n x 3.0 TByte, 7.2k rpm	15'000'000 ns	Ran: 120 IOPS
Tape drive	n x 2.5 TByte pro LTO Cartridge	50'000'000'000 ns	Seq: 160 MBps

Architectures with Flash Technology



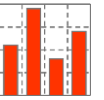
Capacity and access times

Server System	Capacity	Average Access Time	Throughput
CPU register	1 KByte	< 1 ns	
Level 1 Cache on-chip	128 KByte	1 ns	
Level 2 Cache	8 MByte	5 ns	
Level 3 Cache	32 MByte	15 ns	
Main memory	4 TByte	100 ns	Seq/Ran: > 10 GBps
PCI attached flash	n x 2.4 TByte	< 100'000 ns	Seq: > 1 GBps Ran: > 500'000 IOPS
Storage System			
Cache	<= 1 TByte	500'000 ns	
Solid state disks (SSD)	n x 1.4 TByte	1'000'000 ns	Seq: > 1 GBps Ran: > 200'000 IOPS
Hard drive disks (HDD)	n x 600 GByte, 15k rpm	8'000'000 ns	Ran: 250 IOPS
Hard drive disks (HDD)	n x 3.0 TByte, 7.2k rpm	15'000'000 ns	Ran: 120 IOPS
Tape drive	n x 2.5 TByte pro LTO Cartridge	50'000'000'000 ns	Seq: 160 MBps

x 80'000



Architectures with Flash Technology



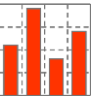
Capacity and access times

Server System	Capacity	Average Access Time	Throughput
CPU register	1 KByte	< 1 ns	
Level 1 Cache on-chip	128 KByte	1 ns	
Level 2 Cache	8 MByte	5 ns	
Level 3 Cache	32 MByte	15 ns	
Main memory	4 TByte	100 ns	Seq/Ran: > 10 GBps
PCI attached flash	n x 2.4 TByte	< 100'000 ns	Seq: > 1 GBps Ran: > 500'000 IOPS
Storage System			
Cache	<= 1 TByte	500'000 ns	
Solid state disks (SSD)	n x 1.4 TByte	1'000'000 ns	Seq: > 1 GBps Ran: > 200'000 IOPS
Hard drive disks (HDD)	n x 600 GByte, 15k rpm	8'000'000 ns	Ran: 250 IOPS
Hard drive disks (HDD)	n x 3.0 TByte, 7.2k rpm	15'000'000 ns	Ran: 120 IOPS
Tape drive	n x 2.5 TByte pro LTO Cartridge	50'000'000'000 ns	Seq: 160 MBps

x 80

x 2'000

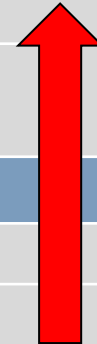
Architectures with Flash Technology

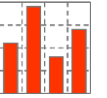


Capacity and access times

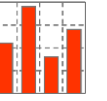
Server System	Capacity	Average Access Time	Throughput
CPU register	1 KByte	< 1 ns	
Level 1 Cache on-chip	128 KByte	1 ns	
Level 2 Cache	8 MByte	5 ns	
Level 3 Cache	32 MByte	15 ns	
Main memory	4 TByte	100 ns	Seq/Ran: > 10 GBps
PCI attached flash	n x 2.4 TByte	< 100'000 ns	Seq: > 1 GBps Ran: > 500'000 IOPS
Storage System			
Cache	<= 1 TByte	500'000 ns	
Solid state disks (SSD)	n x 1.4 TByte	1'000'000 ns	Seq: > 1 GBps Ran: > 200'000 IOPS
Hard drive disks (HDD)	n x 600 GByte, 15k rpm	8'000'000 ns	Ran: 250 IOPS
Hard drive disks (HDD)	n x 3.0 TByte, 7.2k rpm	15'000'000 ns	Ran: 120 IOPS
Tape drive	n x 2.5 TByte pro LTO Cartridge	50'000'000'000 ns	Seq: 160 MBps

x 10'000

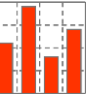




- Flash Technology currently changes architecture of computer systems dramatically
- Flash Technology may change application architecture and algorithms
- Example of a very new flash system: IBM Flash 840 (launched Q2 2014)
 - Rack space 2 units
 - Raw capacity 66 TByte
 - Usable capacity 48 TByte
 - Throughput ~ 500'000 IOPS - 8 kByte - 150 μ s (estimation at port level)
 - Power 625 Watt
- Hardware vendors offer different architectures with flash technology for Oracle platforms

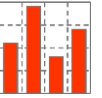


- The architectures provided by hardware vendors differ in
 - Volatility
 - Capacity
 - Throughput
 - Latency
 - Cost
 - Manageability
 - Shareability

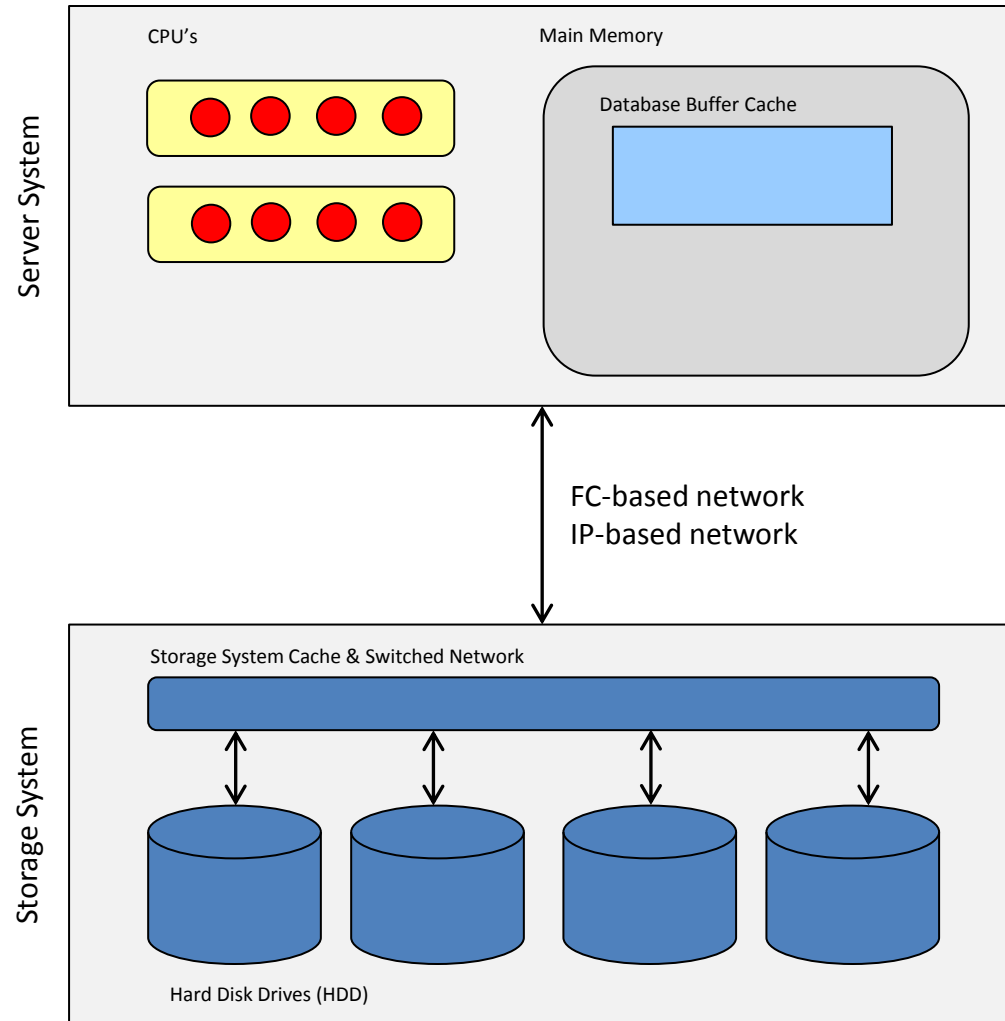


- 1 Introduction to Storage Performance Tests
- 2 Storage Tiers
- 3 Storage Architectures**
- 4 Benchmark Results
- 5 Summary

Conventional System Architecture



Architecture without any Flash Technology

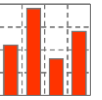


Access times (Sep 2011):

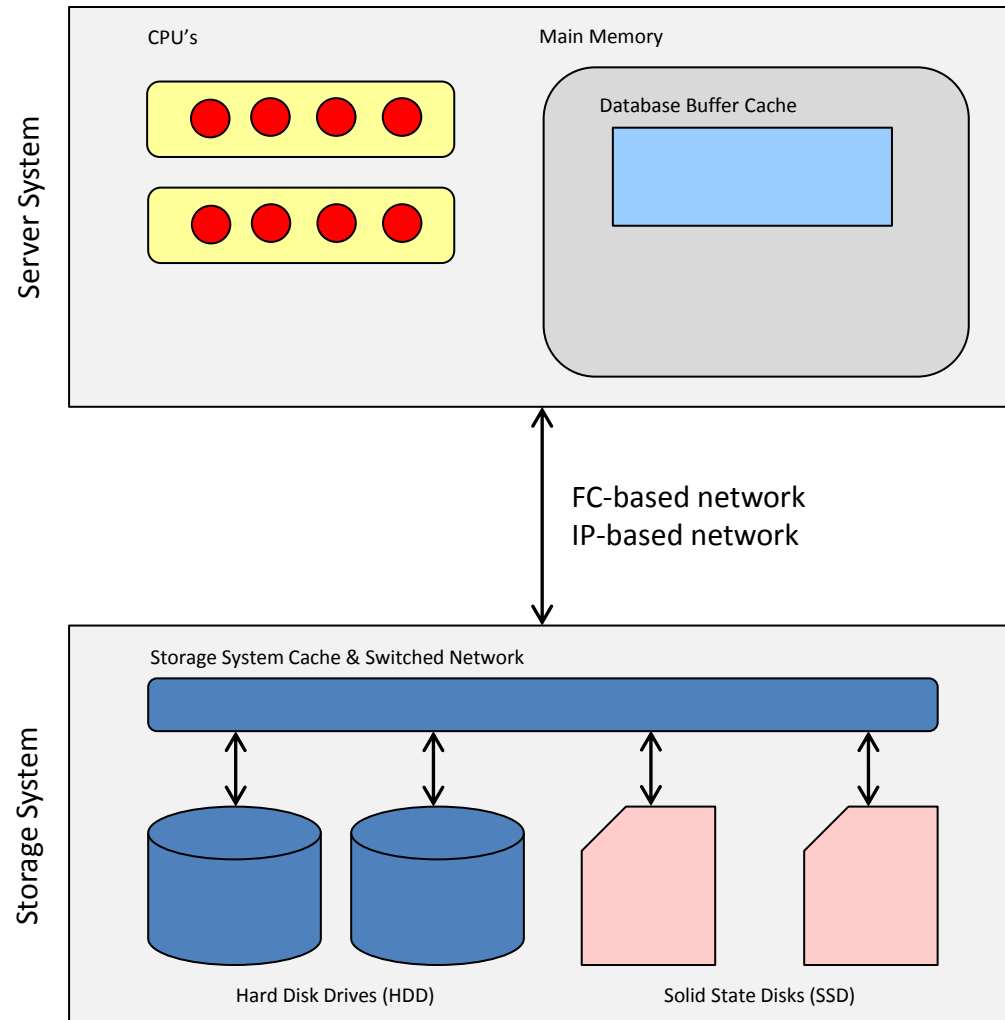
- CPU cache (SRAM) 1×10^{-9} s
- Database cache (DRAM) 100×10^{-9} s

- Storage system cache 500×10^{-6} s
- Storage system hdd 8×10^{-3} s

Storage System with mixed disks



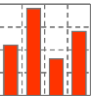
Automatic tiering, e.g. EMC, HDS or IBM storage systems



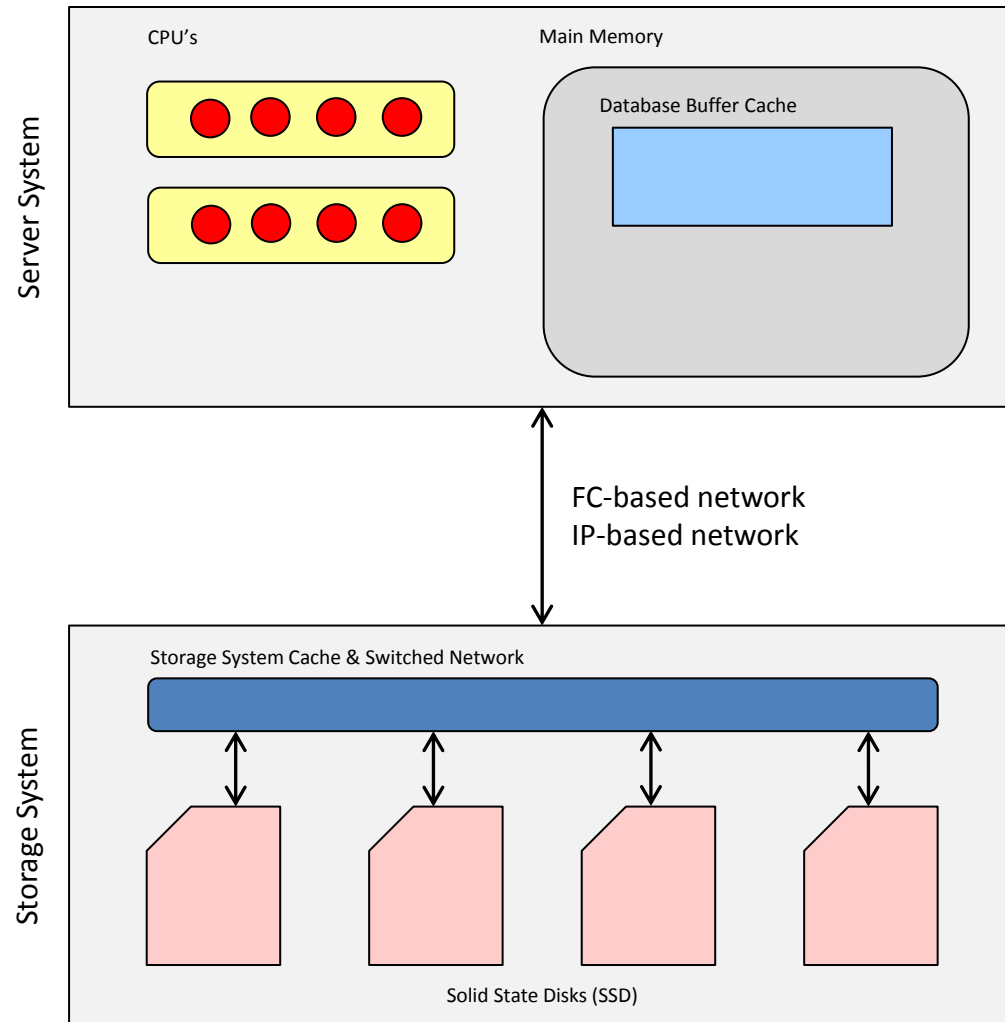
Access times (Sep 2011):

- CPU cache (SRAM) $1 \times 10^{-9} \text{ s}$
- Database cache (DRAM) $100 \times 10^{-9} \text{ s}$
- Storage system ssd $< 1 \times 10^{-3} \text{ s}$
- Storage system cache $500 \times 10^{-6} \text{ s}$
- Storage system hdd $8 \times 10^{-3} \text{ s}$

Storage System with solid state disks only



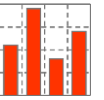
All Flash Array, e.g. HDS HUS VM AFA



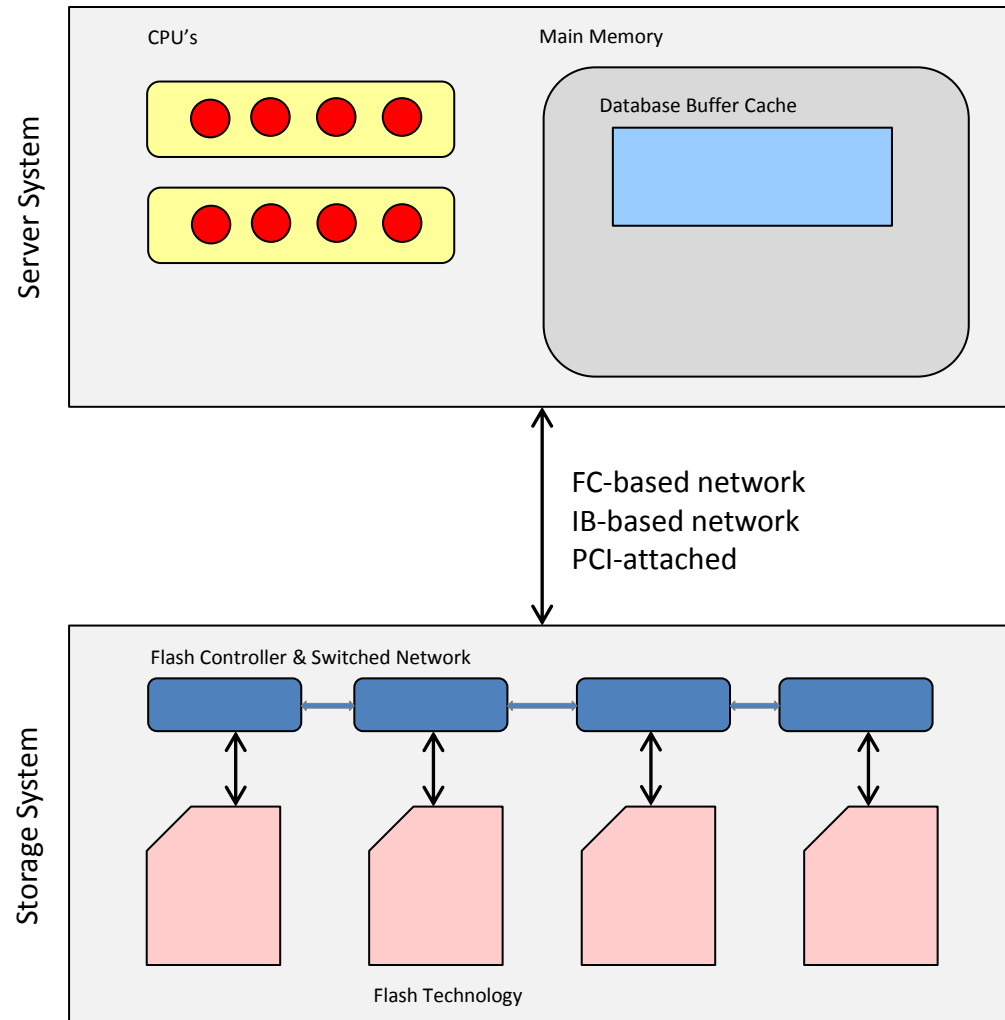
Access times (Sep 2011):

- CPU cache (SRAM) $1 \times 10^{-9} \text{ s}$
- Database cache (DRAM) $100 \times 10^{-9} \text{ s}$
- Storage system ssd $< 1 \times 10^{-3} \text{ s}$
- Storage system cache $500 \times 10^{-6} \text{ s}$

Storage System based on Flash Technology



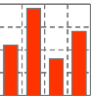
Storage Systems completely build on flash technology, e.g. Violin, IBM Flash 840



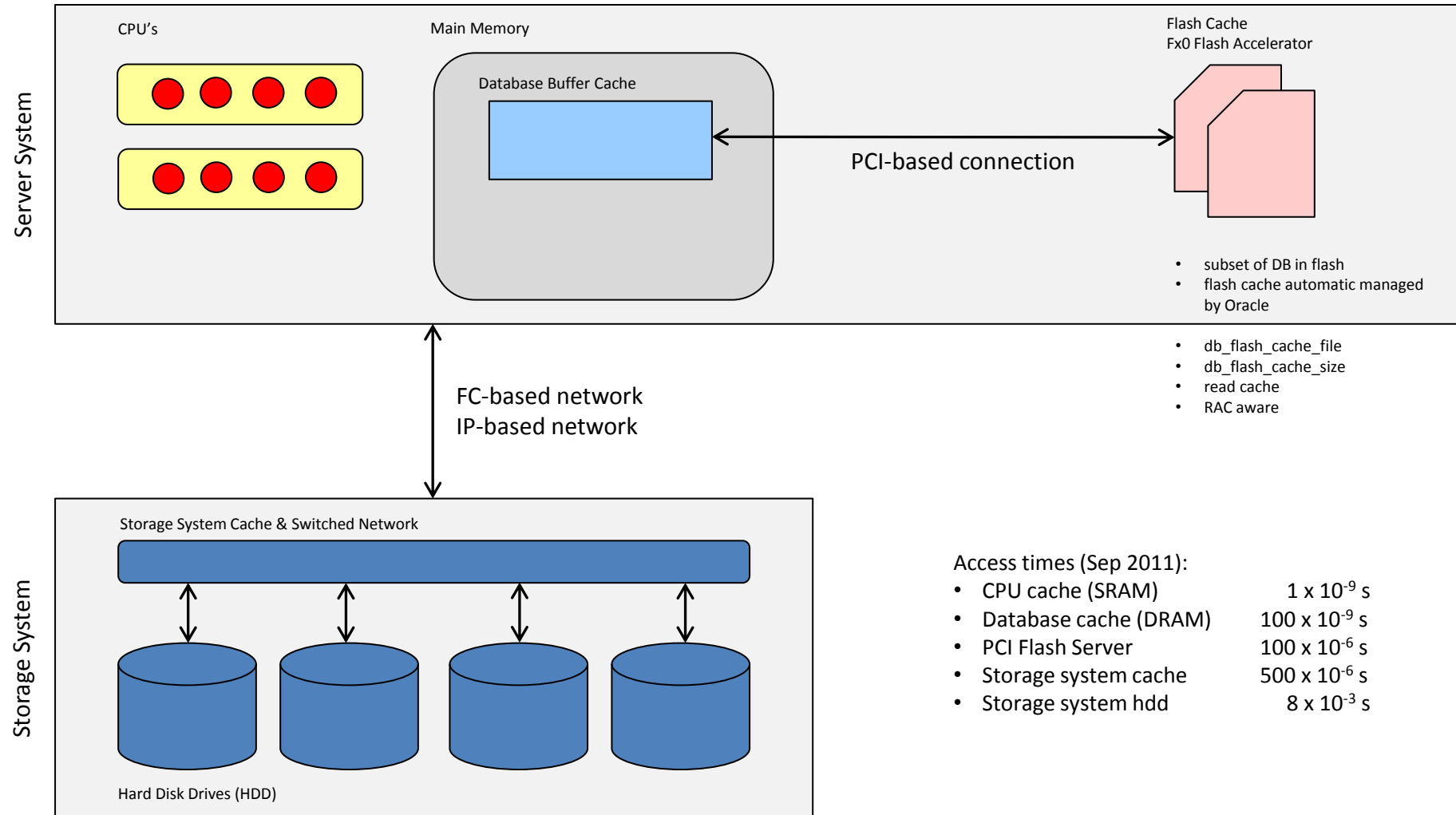
Access times (Sep 2011):

- CPU cache (SRAM) $1 \times 10^{-9} \text{ s}$
- Database cache (DRAM) $100 \times 10^{-9} \text{ s}$
- Storage system ssd $< 500 \times 10^{-6} \text{ s}$

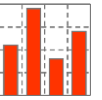
Oracle Database Smart Flash Cache Technology



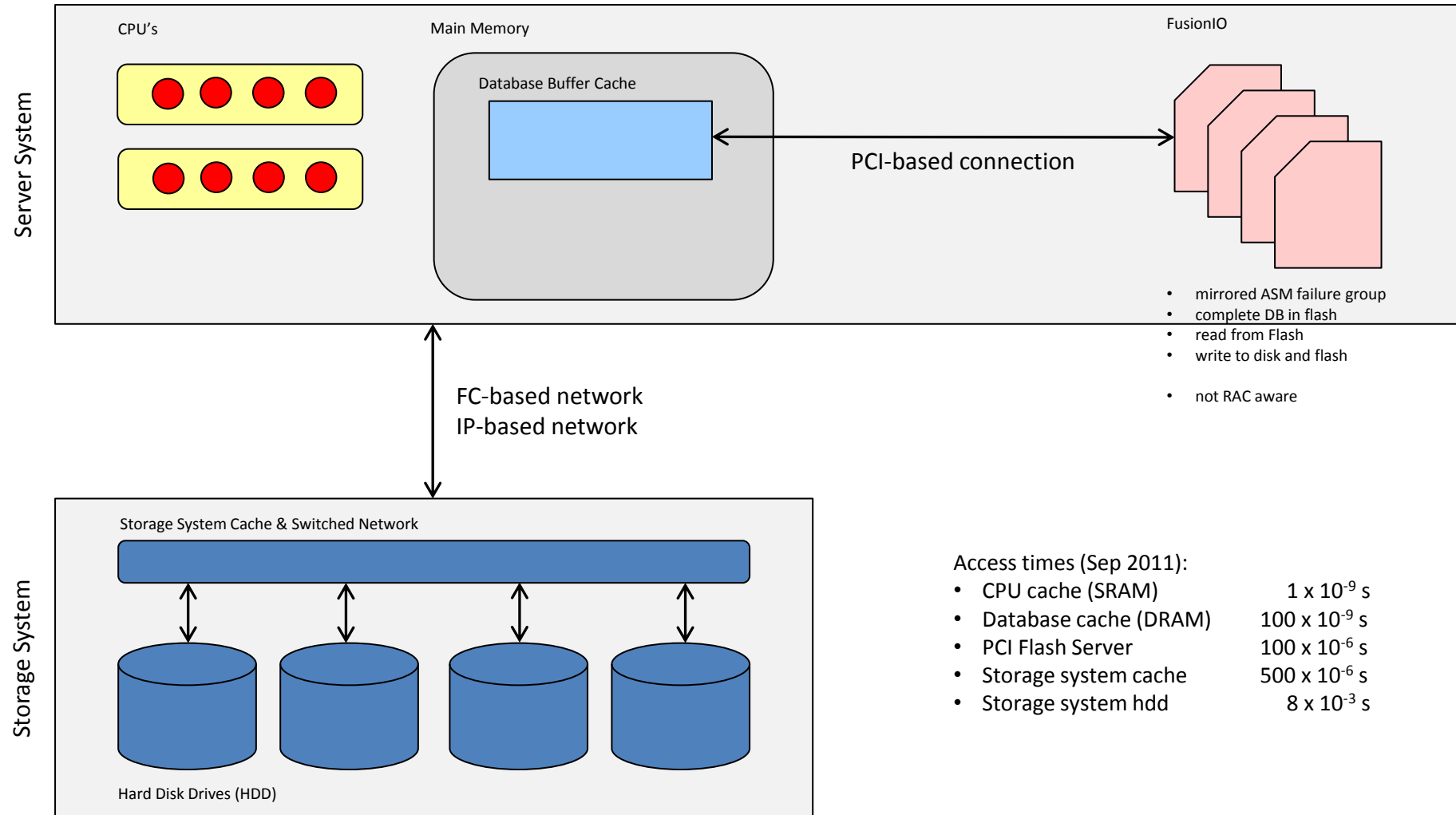
Specific solution only for Solaris and OEL and Sun Fx0 Flash



Mirrored Database in Server Fusion IO Cards



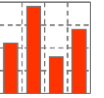
Mirrored ASM failure group, e.g. Hitachi UCP



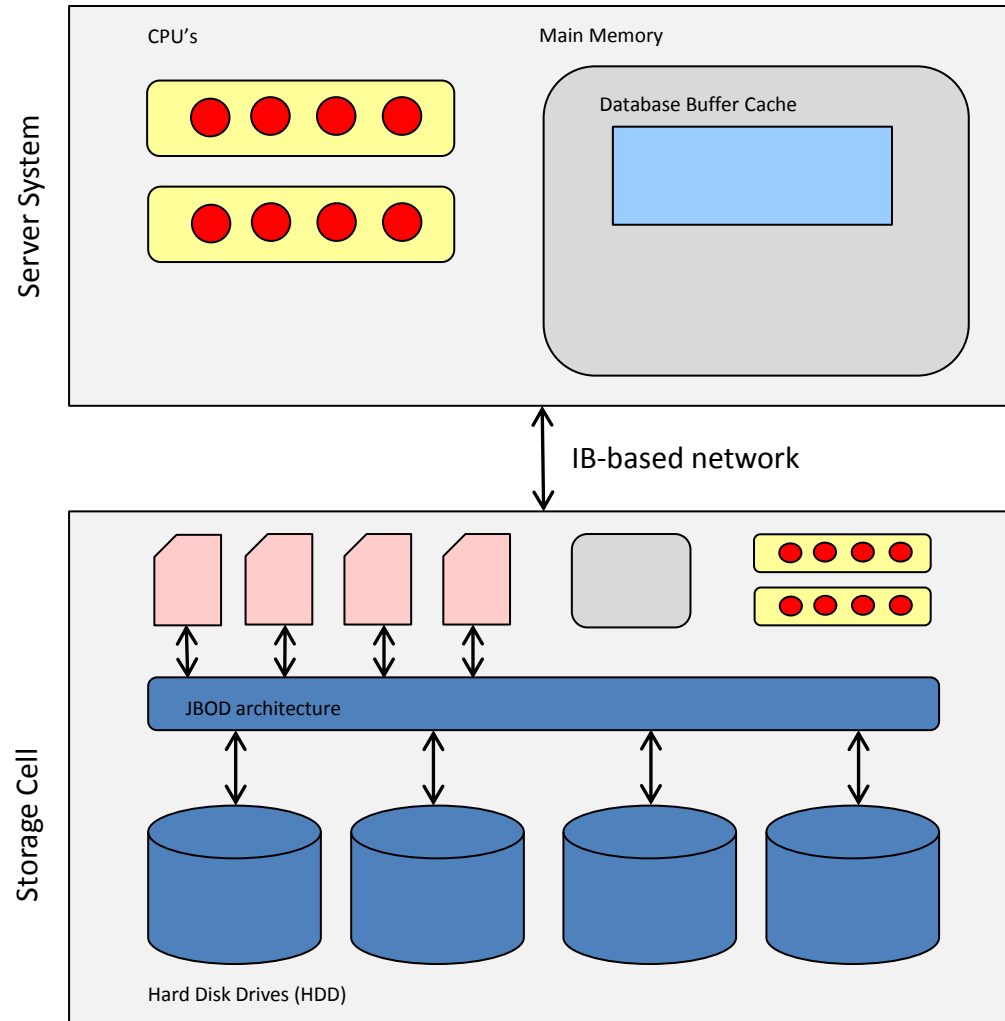
Access times (Sep 2011):

- CPU cache (SRAM) 1 x 10⁻⁹ s
- Database cache (DRAM) 100 x 10⁻⁹ s
- PCI Flash Server 100 x 10⁻⁶ s
- Storage system cache 500 x 10⁻⁶ s
- Storage system hdd 8 x 10⁻³ s

Intelligent Storage Grid



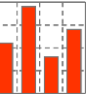
Offloaded database functions, e.g. Oracle Exadata



Access times (Sep 2011):

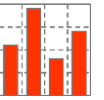
- CPU cache (SRAM) 1×10^{-9} s
- Database cache (DRAM) 100×10^{-9} s

- Storage system ssd 1×10^{-3} s
- Storage system hdd 8×10^{-3} s

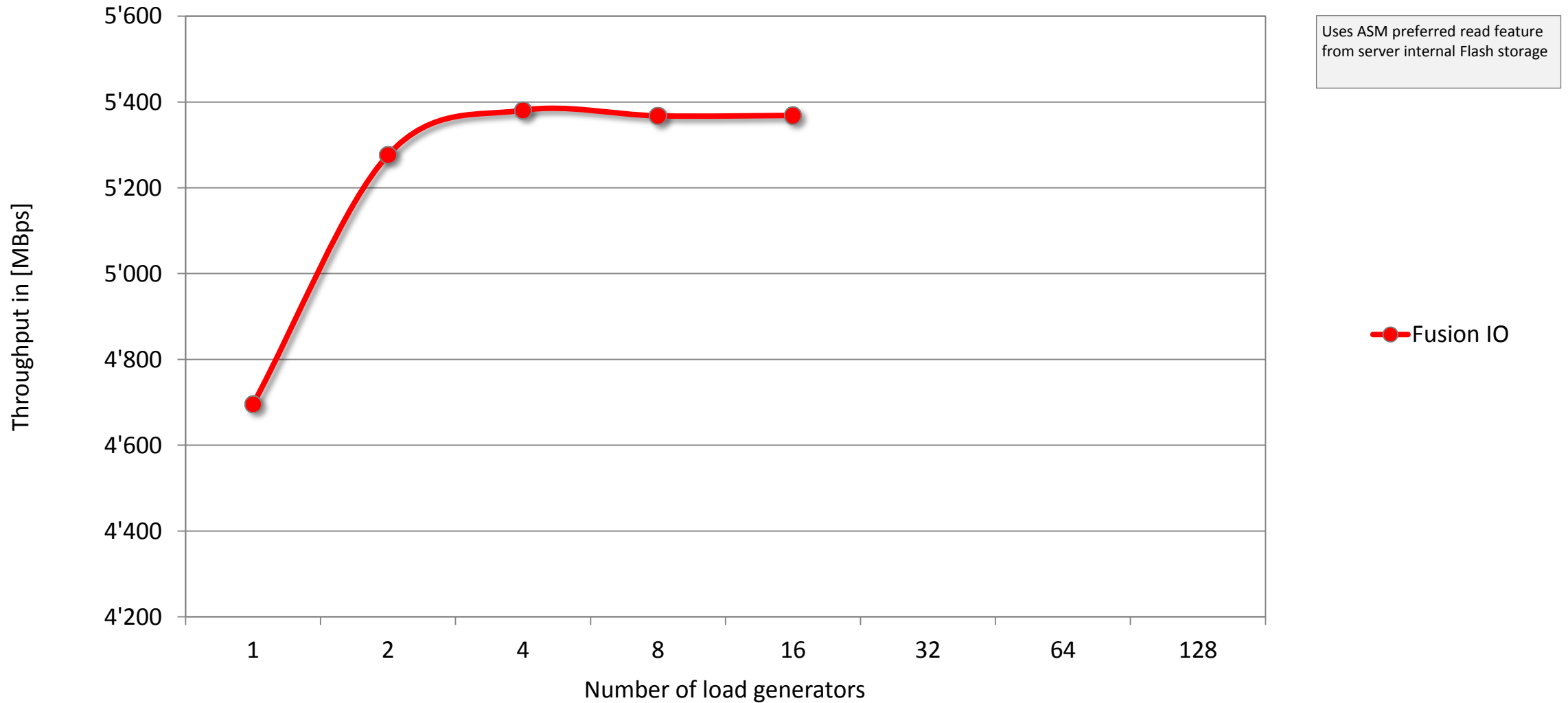


- 1 Introduction to Storage Performance Tests
- 2 Storage Tiers
- 3 Storage Architectures
- 4 Benchmark Results**
- 5 Summary

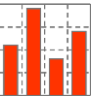
PCI attached Server Flash



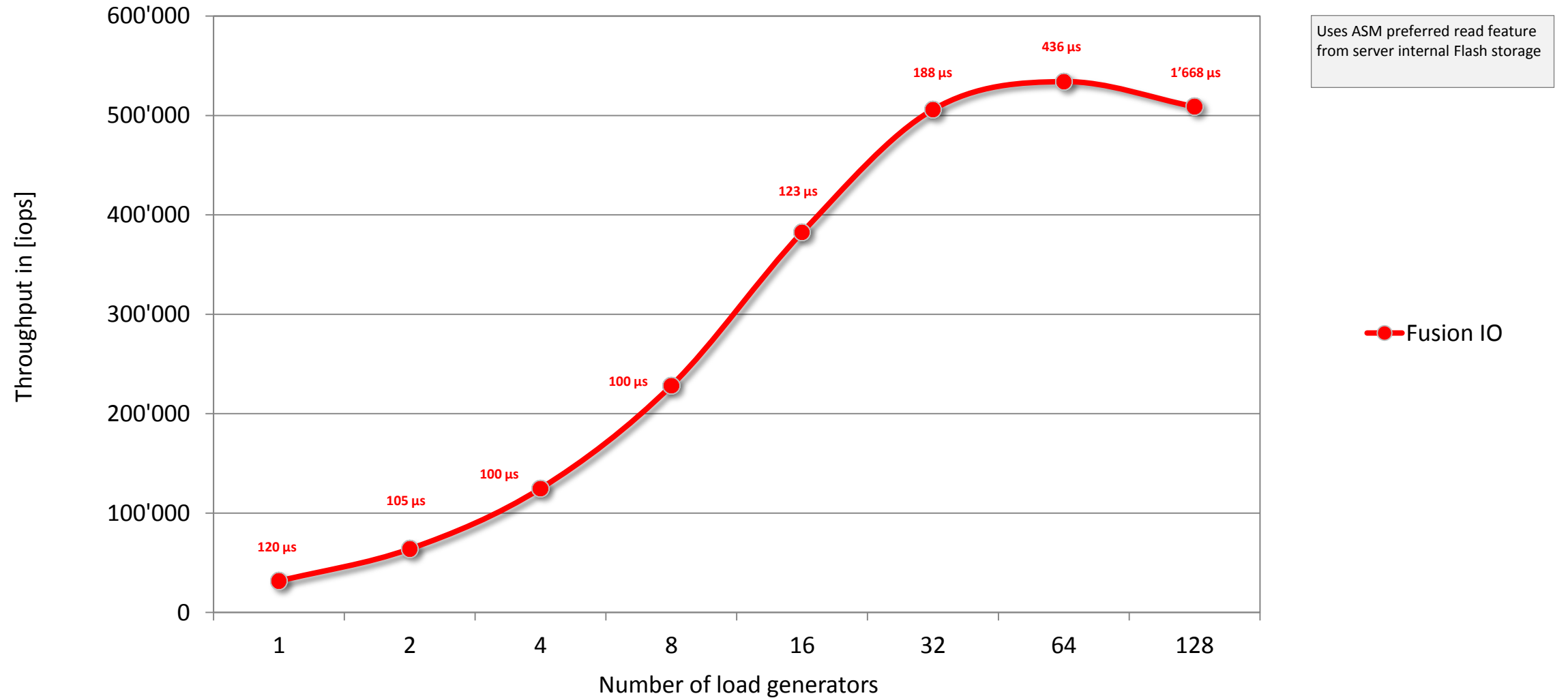
Oracle storage performance: sequential read



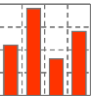
PCI attached Server Flash



Oracle storage performance: random read



PCI attached Server Flash



Oracle storage performance: random read

Fusion IO

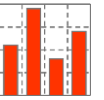
Run	Tst Code	#N	#J	#T	CPU busy [%]	CPU sys [%]	Physical read [iops]	Physical read [dbps]	Physical read [MBps]	Physical write [iops]	Physical write [dbps]	Physical write [MBps]	REDO write [iops]	Hitrates db flash [%]	Hitrates exa flash [%]	Elap time [s]
15	10 STO-62	1	1	1	3	2	31747	31745	248	11	23	0	2	0	0	51
	11 STO-62	1	2	1	6	4	64039	64037	500	59	68	1	5	0	0	51
	12 STO-62	1	4	1	12	7	124897	124895	976	164	168	1	10	0	0	52
	13 STO-62	1	8	1	23	13	228361	228359	1784	356	351	3	19	0	0	57
	14 STO-62	1	16	1	45	27	382520	382518	2988	629	607	5	31	0	0	68
	15 STO-62	1	32	1	77	45	506050	506047	3954	854	812	6	41	0	0	103
	16 STO-62	1	64	1	95	51	534207	534206	4174	879	832	7	41	0	0	196
	17 STO-62	1	128	1	98	51	508957	508962	3976	810	764	6	40	0	0	305
	18 STO-62	1	256	1	98	48	508889	508888	3976	793	730	6	61	0	0	310

Top 5 Timed Foreground Events					
~~~~~					
Event	Waits	Time (s)	Avg wait (ms)	% DB time	Wait Class
DB CPU		4,451		20.9	
db file parallel read	2,686,643	2,033	1	9.6	User I/O
db file sequential read	3,079,002	1,343	0	6.3	User I/O

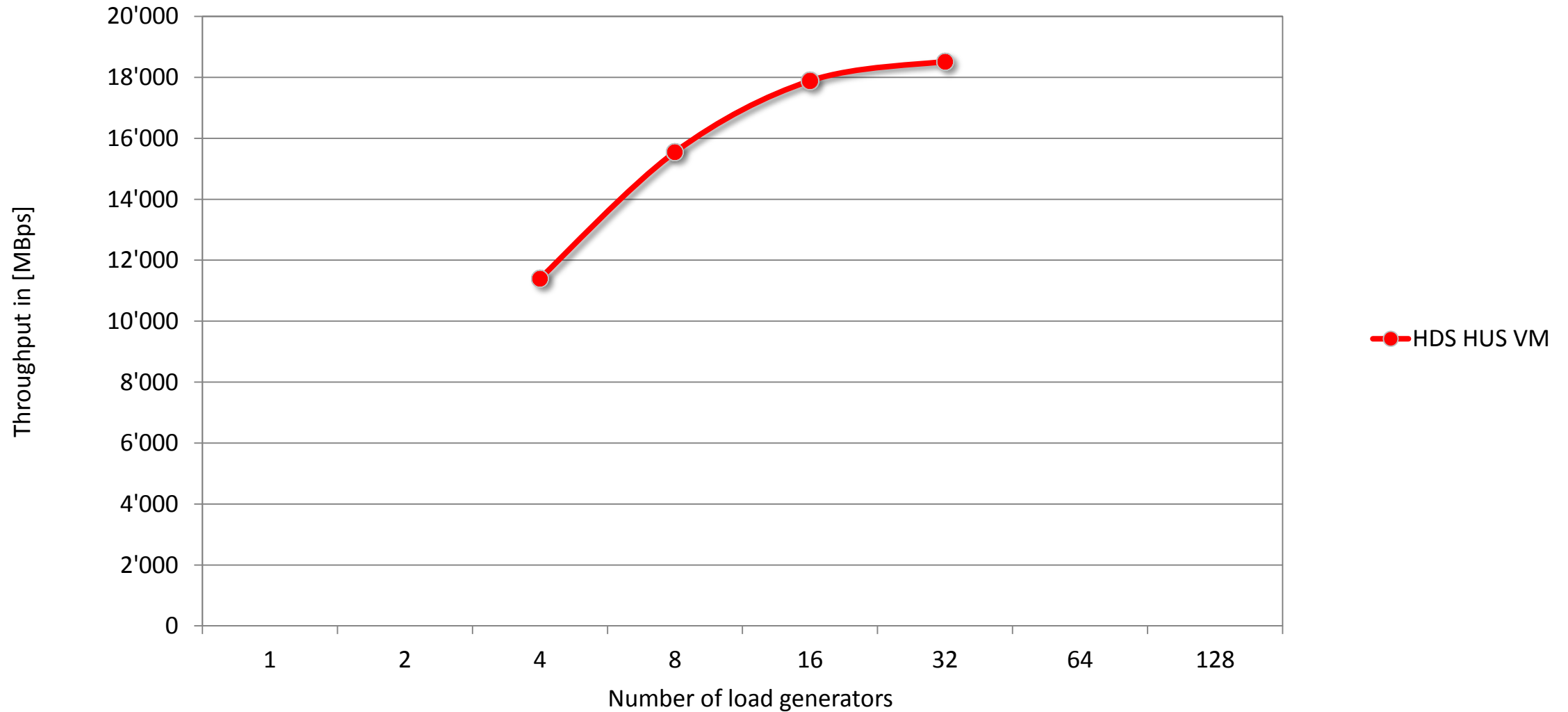
**Remarks:**

- In this test case, Oracle optimizes random I/O operations
- "db file sequential read" is a true single block read operation (one system call per database block read)
- "db file parallel read" is an optimized non-contiguous multi block random read operation (one system call for at least 2 or more database block read)

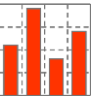
# Conventional Storage Performance AFA



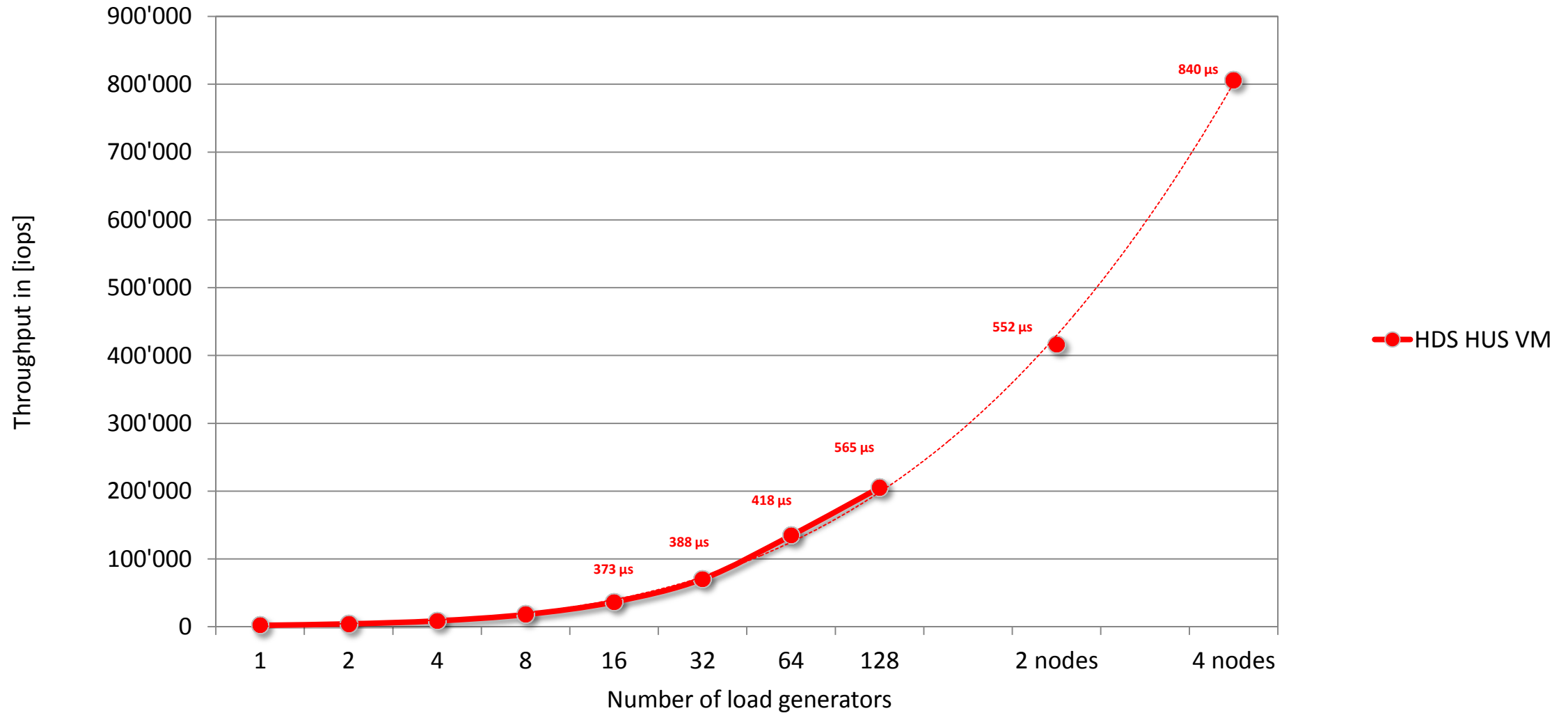
Oracle storage performance: sequential read



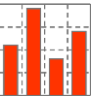
# Conventional Storage Performance AFA



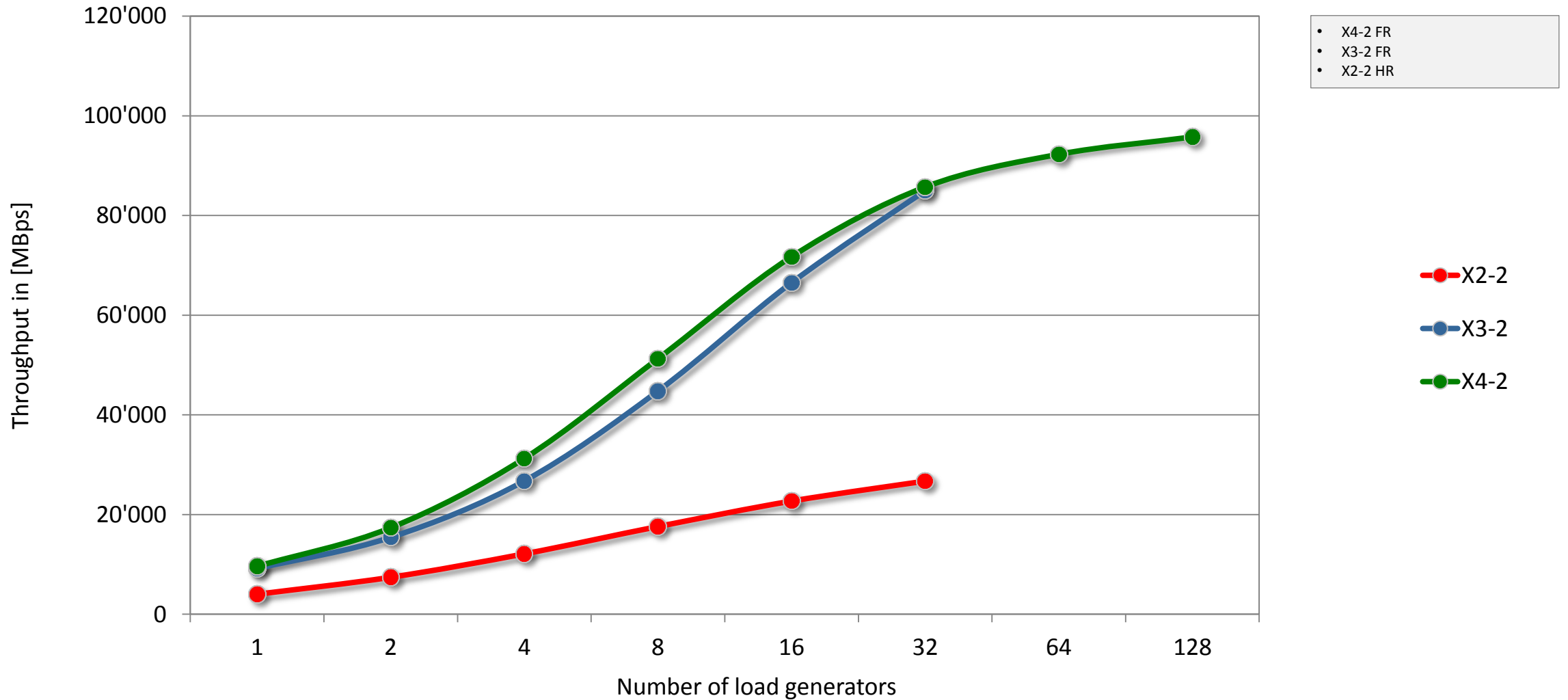
Oracle storage performance: random read



# Exadata Storage Grid

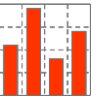


Oracle storage performance: sequential read, 1 DB Server

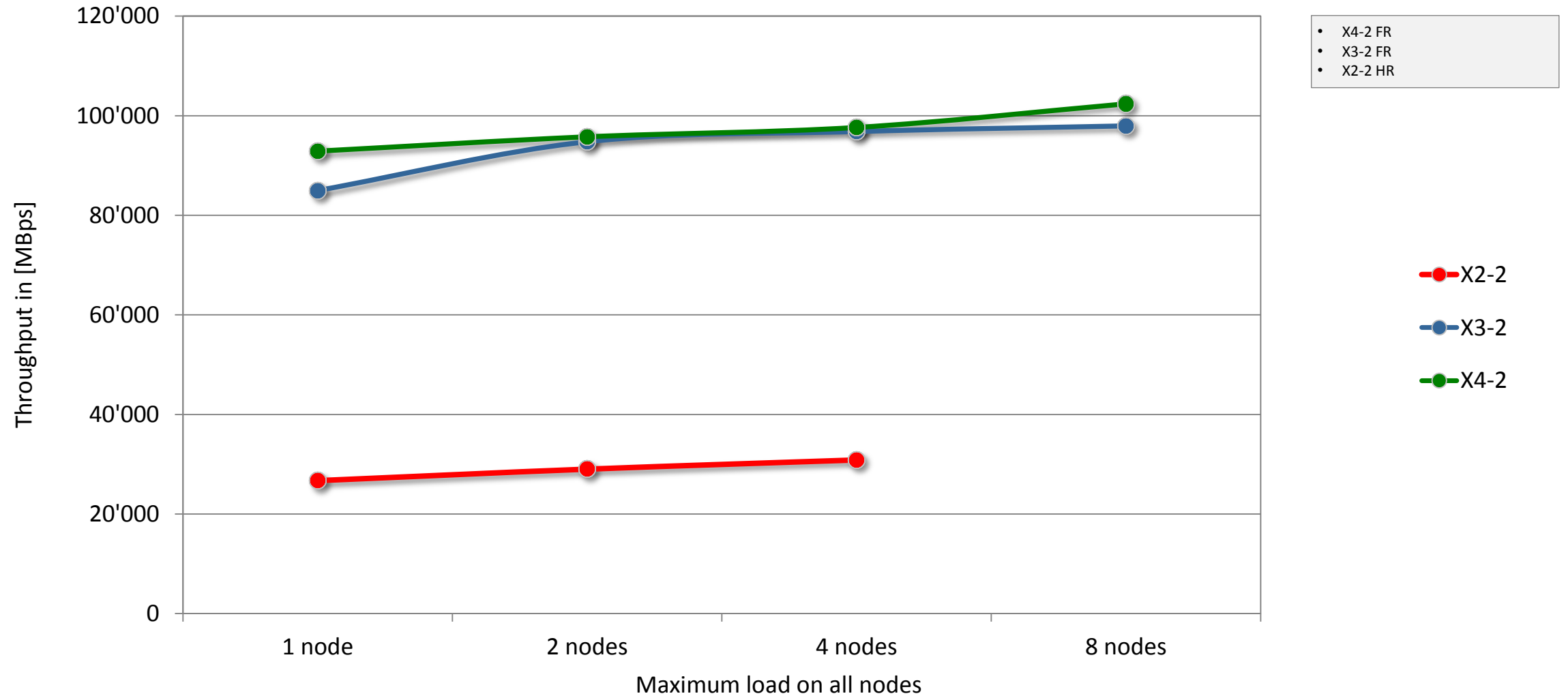




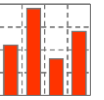
# Exadata Storage Grid



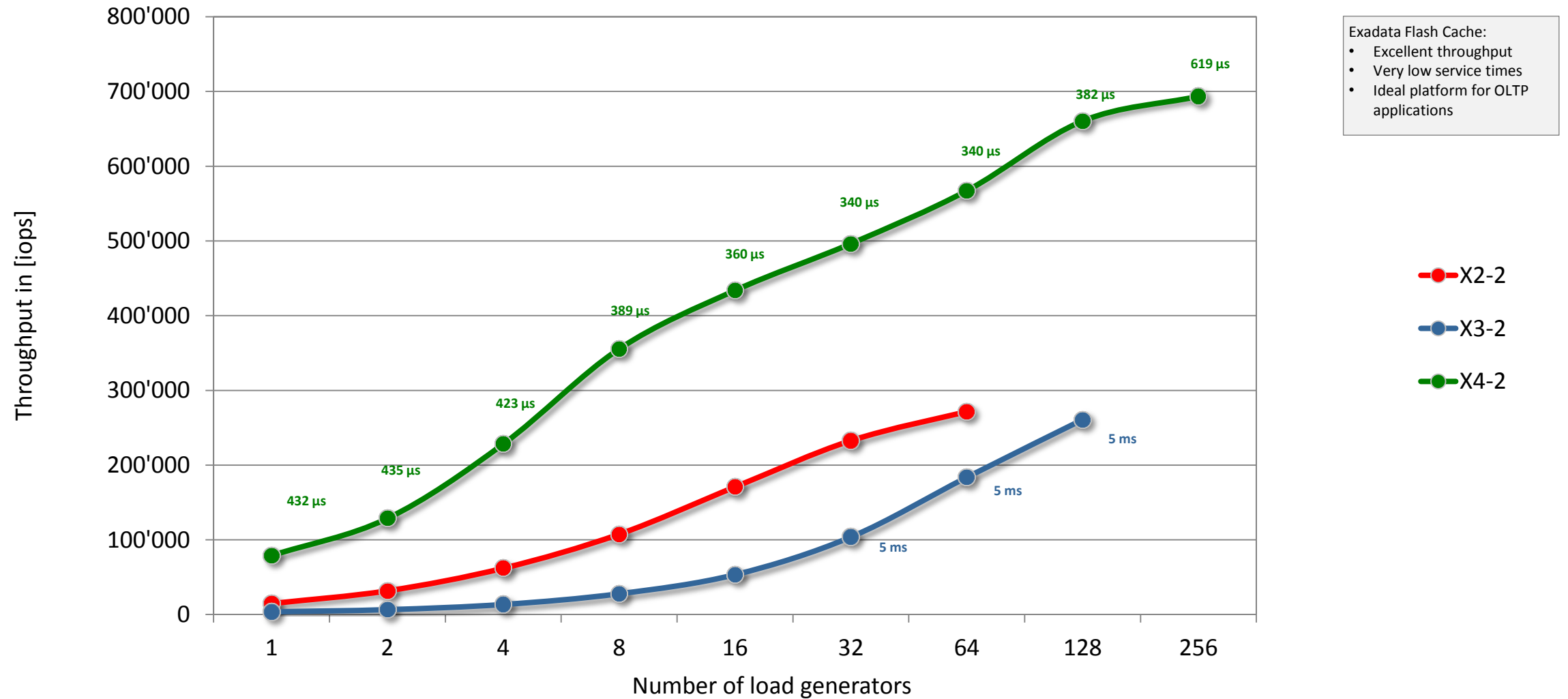
Oracle storage performance: sequential read, Cluster



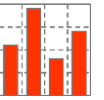
# Exadata Storage Grid



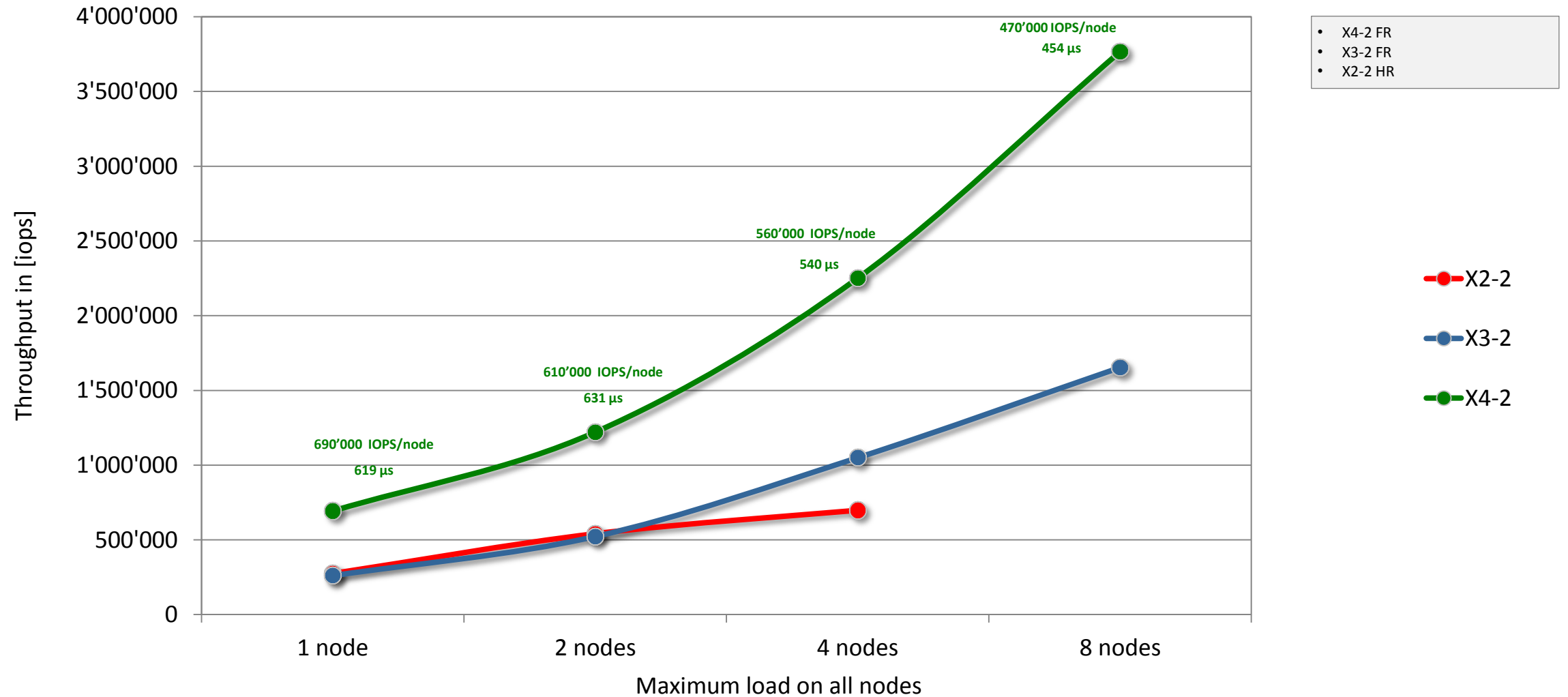
Oracle storage performance: random read, 1 DB server

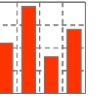


# Exadata Storage Grid



Oracle storage performance: random read, Cluster

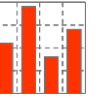




- 1 Introduction to Storage Performance Tests
- 2 Storage Tiers
- 3 Storage Architectures
- 4 Benchmark Results
- 5 Summary**

# Summary

---



- Do you need MORE storage I/O performance?
  - Avoid shared storage systems
  - Replace last century hard disk drive technology
- Where is the pain
  - Quick fix for some specific systems?
  - Overall poor performance on shared storage system?
- Smooth integration?

**BENCHWARE**

*swiss precision in performance measurement*

*[www.benchmark.ch](http://www.benchmark.ch)*

*[info@benchmark.ch](mailto:info@benchmark.ch)*