Enterprise-Class Storage

We’ve Come a Long Way, Or Have We?

Kevin Closson

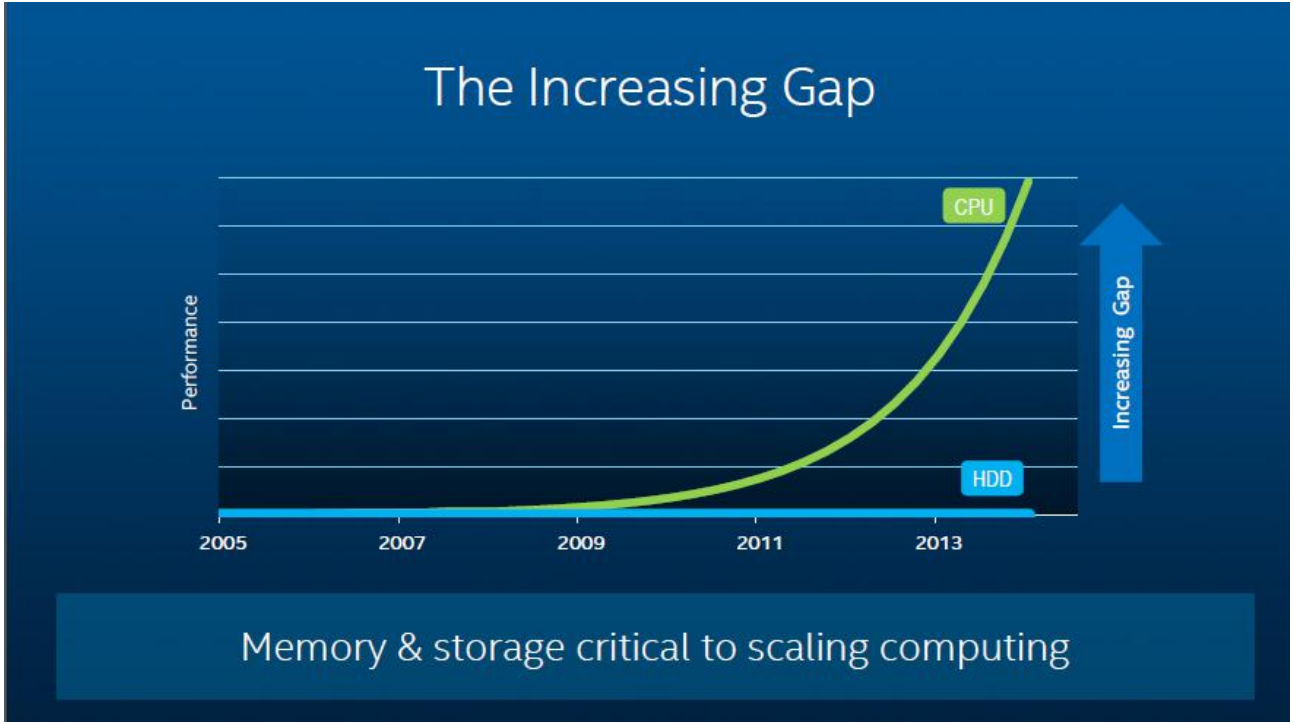
Sr. Director / Chief Performance Architect

XtremIO, CDT, EMC

ABOUT THE SPEAKER

- Performance Architect in XtremIO
- Former Performance Architect (IC6) in Oracle Exadata Development
- Oaktable Network since 2002
- Inventor of SLOB (The Silly Little Oracle Benchmark) platform testing kit
- Performance optimizations in Oracle Disk Manager Library at Veritas and PolyServe
- 10 years Database Engineering on the Sequent ports of Oracle including the first Unix port of Parallel Server and development platform for Intra-Node Parallel Query
- US Patents in high performance NUMA optimized locking primitives and database caching methods
- Lots of book collaborations and blogging at kevinclosson.net

HDD – The Root of All That Is Evil (?)



The Increasing Gap

Memory & storage critical to scaling computing

1980

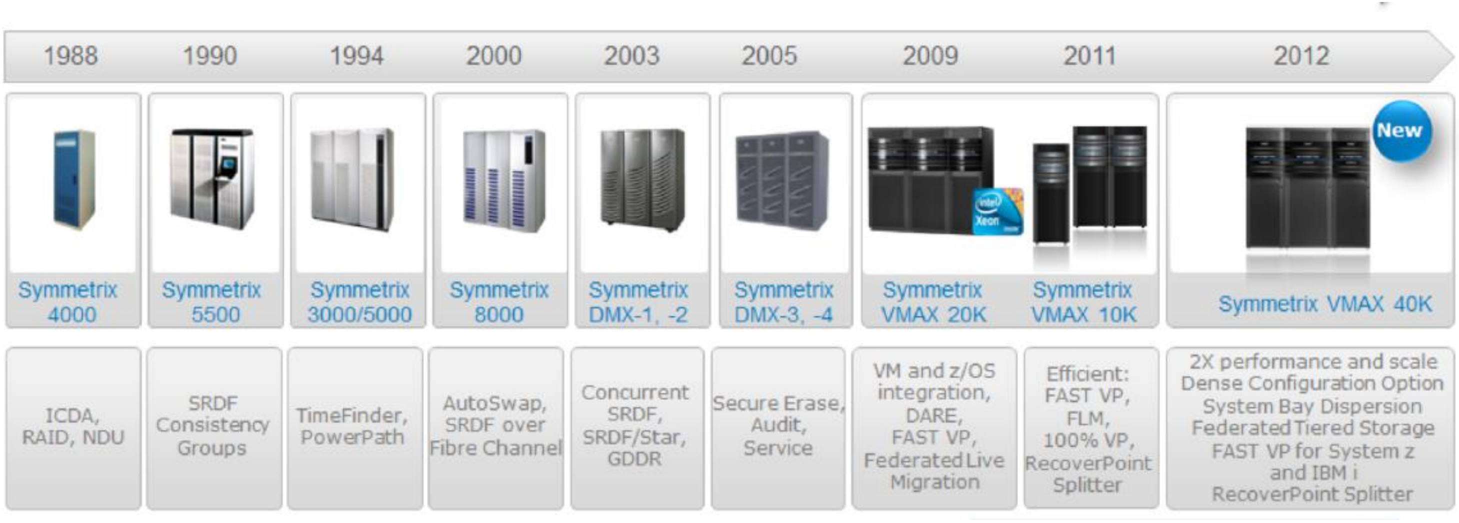
Eight 2.5GB IBM 3380 Disk Systems: 20GB
Estimated value: $648,000 - $1,137,600
Weight: 2,000,000 grams (4,400 pounds)

TAPE LIBRARY

MEQUON, WISCONSIN U.S.A.

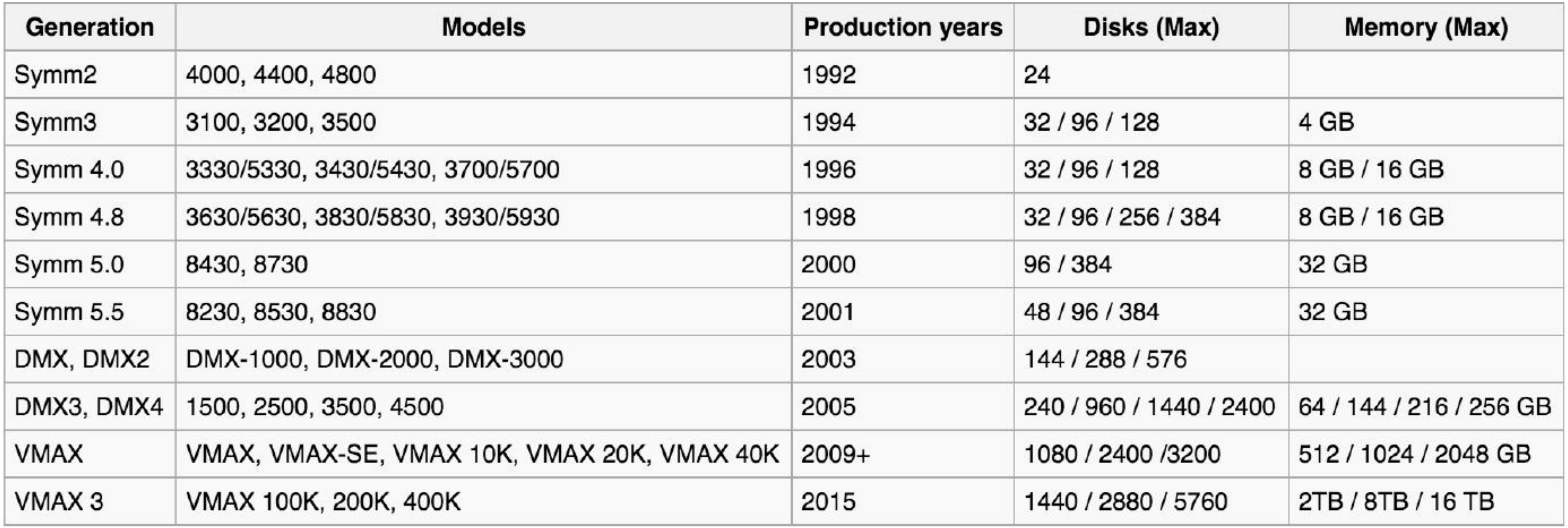
Enterprise Storage

HISTORICAL ENTERPRISE ARRAY



1988	1990	1994	2000	2003	2005	2009	2011	2012
Symmetrix 4000	Symmetrix 5500	Symmetrix 3000/5000	Symmetrix 8000	Symmetrix DMX-1, -2	Symmetrix DMX-3, -4	Symmetrix VMAX 20K	Symmetrix VMAX 10K	Symmetrix VMAX 40K
ICDA, RAID, NDU	SRDF Consistency Groups	TimeFinder, PowerPath	AutoSwap, SRDF over Fibre Channel	Concurrent SRDF, SRDF/Star, GDDR	Secure Erase, Audit, Service	VM and z/OS integration, DARE, FAST VP, Federated Live Migration	Efficient: FAST VP, FLM, 100% VP, RecoverPoint Splitter	2X performance and scale Dense Configuration Option System Bay Dispersion Federated Tiered Storage FAST VP for System z and IBM i RecoverPoint Splitter

Focus on features, host connectivity, protocols, capacity and performance



Generation	Models	Production years	Disks (Max)	Memory (Max)
Symm2	4000, 4400, 4800	1992	24	
Symm3	3100, 3200, 3500	1994	32 / 96 / 128	4 GB
Symm 4.0	3330/5330, 3430/5430, 3700/5700	1996	32 / 96 / 128	8 GB / 16 GB
Symm 4.8	3630/5630, 3830/5830, 3930/5930	1998	32 / 96 / 256 / 384	8 GB / 16 GB
Symm 5.0	8430, 8730	2000	96 / 384	32 GB
Symm 5.5	8230, 8530, 8830	2001	48 / 96 / 384	32 GB
DMX, DMX2	DMX-1000, DMX-2000, DMX-3000	2003	144 / 288 / 576	
DMX3, DMX4	1500, 2500, 3500, 4500	2005	240 / 960 / 1440 / 2400	64 / 144 / 216 / 256 GB
VMAX	VMAX, VMAX-SE, VMAX 10K, VMAX 20K, VMAX 40K	2009+	1080 / 2400 /3200	512 / 1024 / 2048 GB
VMAX 3	VMAX 100K, 200K, 400K	2015	1440 / 2880 / 5760	2TB / 8TB / 16 TB

Is Evolution Always Incremental?

THE VENERABLE ARRAY GOES **ALL** FLASH

4M IOPS, <.5ms LATENCY
150GB/s BANDWIDTH

3.8TB SSD FOR HIGHEST
IOP/TB/FLOOR TILE

APPLIANCE-LIKE PACKAGING
SOFTWARE INCLUDED

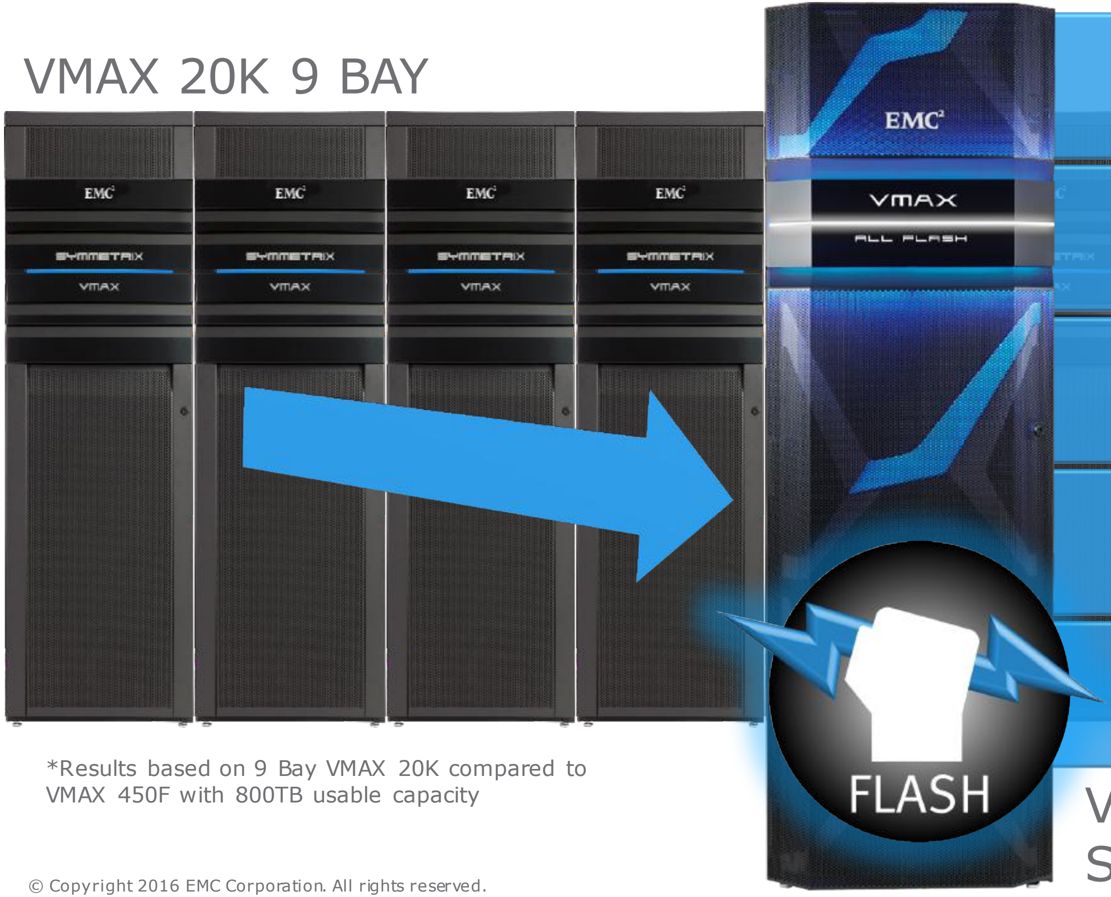
SIMPLE, SIMPLE, SIMPLE
ONE TIER, ANY SKEW, NO HDDS

* Performance numbers based on 8 Engine , RRH, OLTP2

PREVIOUS TO CURRENT GEN REFRESH

800TB USABLE

VMAX 20K 9 BAY



6X MORE PERFORMANCE

40% LOWER TCO

87% LESS ENERGY

92% LESS FOOTPRINT

98% FEWER DRIVE REPLACEMENTS

VMAX ALL FLASH SINGLE BAY

*Results based on 9 Bay VMAX 20K compared to VMAX 450F with 800TB usable capacity

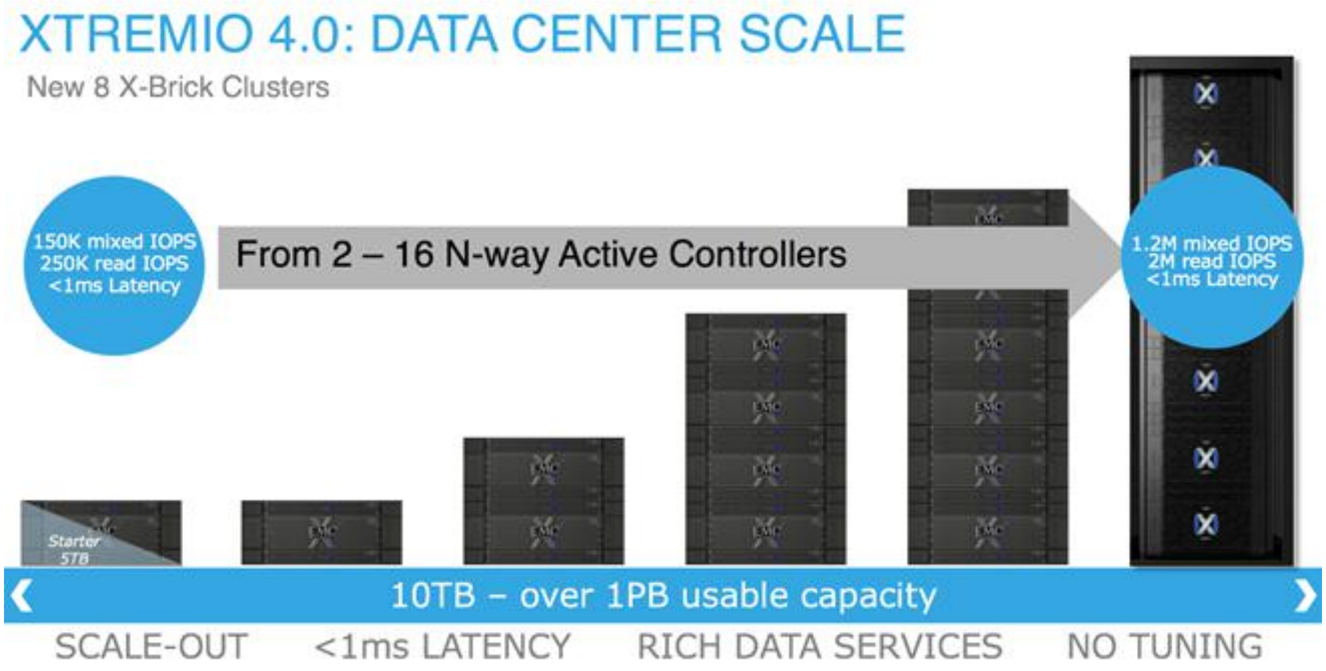
IS IT REALLY JUST THIS?

VS

HDD VS SSD

There’s More To All Flash Array Than The Simple SDD vs HDD

ALL FLASH ARRAY – NOT ALL CREATED EQUAL



XTREMIO 4.0: DATA CENTER SCALE

New 8 X-Brick Clusters

150K mixed IOPS
250K read IOPS
<1ms Latency

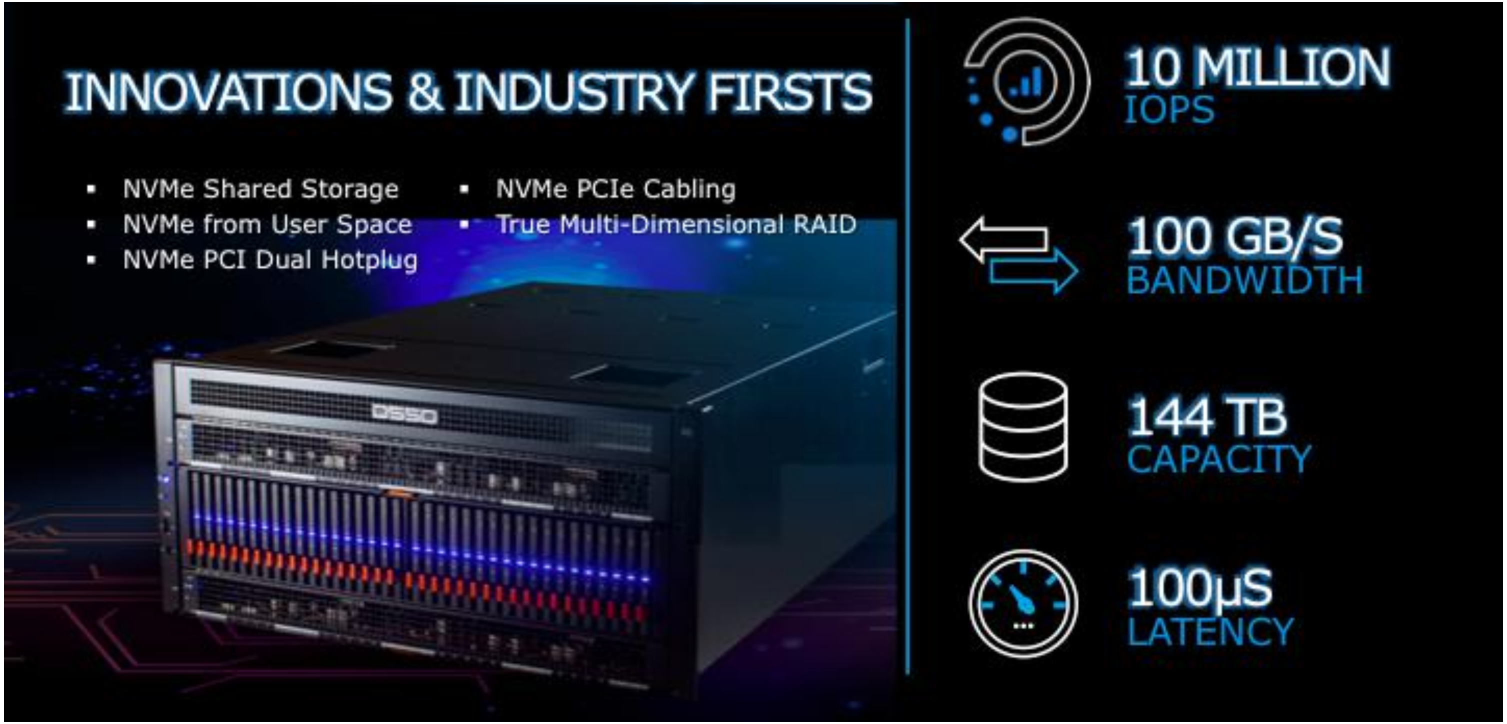
From 2 – 16 N-way Active Controllers

1.2M mixed IOPS
2M read IOPS
<1ms Latency

10TB – over 1PB usable capacity

SCALE-OUT <1ms LATENCY RICH DATA SERVICES NO TUNING

DSSD - COMPLETELY NEW CLASS!



INNOVATIONS & INDUSTRY FIRSTS

- NVMe Shared Storage
- NVMe from User Space
- NVMe PCI Dual Hotplug
- NVMe PCIe Cabling
- True Multi-Dimensional RAID

10 MILLION IOPS

100 GB/S BANDWIDTH

144 TB CAPACITY

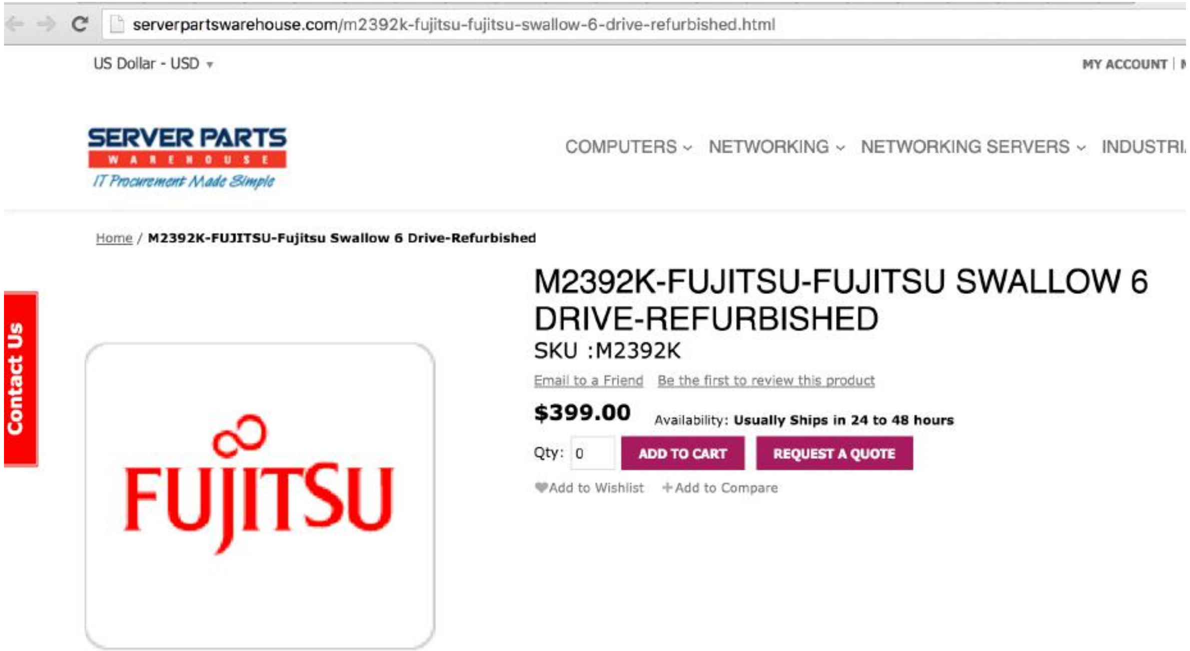
100µS LATENCY

D.I.Y. OR CONVERGED

HDD VS SSD

Let’s Go Back In Time Again …For Another Perspective

DISC INITIALIZATION CHEC



serverpartswarehouse.com/m2392k-fujitsu-fujitsu-swallow-6-drive-refurbished.html

Home / M2392K-FUJITSU-Fujitsu Swallow 6 Drive-Refurbished

M2392K-FUJITSU-FUJITSU SWALLOW 6 DRIVE-REFURBISHED

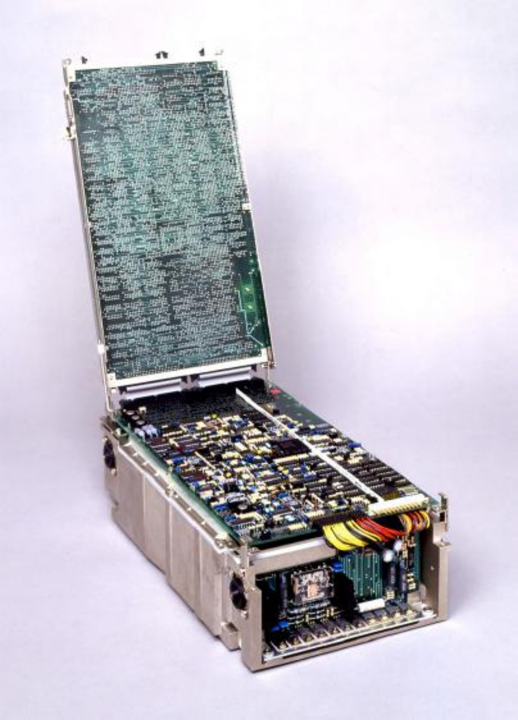
SKU :M2392K

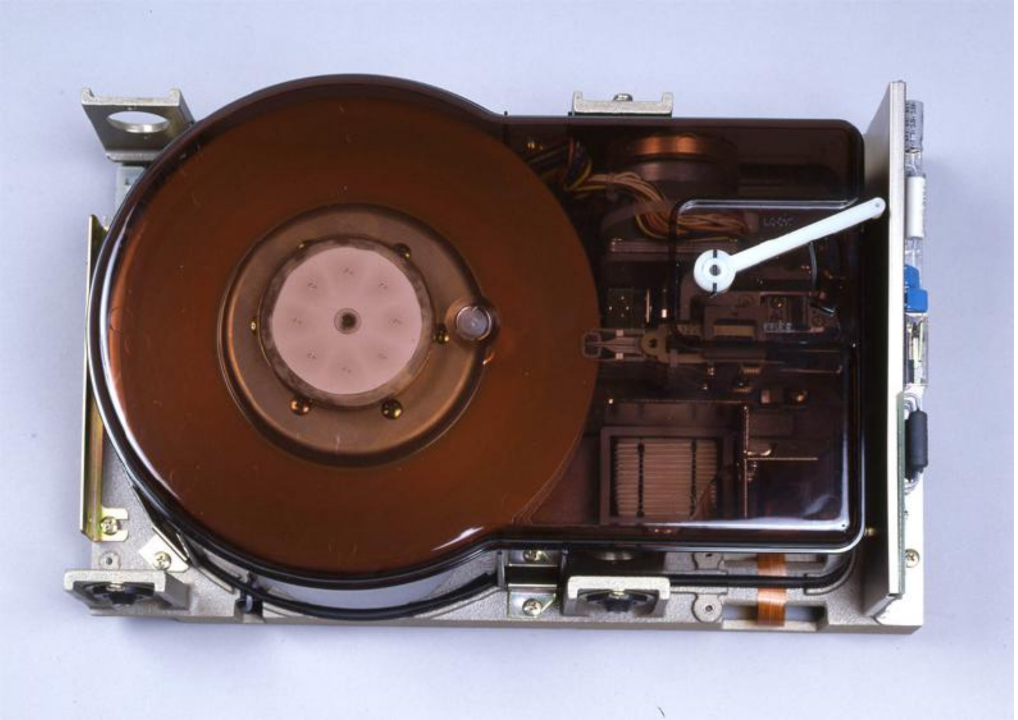
Email to a Friend Be the first to review this product

$399.00 Availability: Usually Ships in 24 to 48 hours

Qty: 0 ADD TO CART REQUEST A QUOTE

Add to Wishlist +Add to Compare





- Fujitsu Swallow-6. (SMD - Storage Module Device)
- 2GB Capacity
    - 12 8” platters
    - 12ms position time
    - Each platter delivered ~80 IOPS
    - Roughly 1,000 IOPS
    - But…the head electronics limited to 3MBs
        - 8K IOPS == 384
        - 4K IOPS == 768
        - 2K IOPS == 1,536 (platters can’t get there)
- No worries. Systems had multiples of these drives…
- Right?

- So attach multiple Fujitsu Swallow-6 to a host (as DAS)
- Example Sequent Balance (circa ’87)
  - Connect up to 8 Swallow drives
  - But, main memory (RAM) was 32MB
    - Only 4MB of that for I/O buffers
    - So, with an Oracle block size of 4K that’s a whopping 1024 concurrent I/O in flight.
  - No matter, at least there was only 53MB/s bus bandwidth
  - What happens to CPU utilization at bus saturation?

Everything is always a CPU problem

Did I Forget To Say That Everything Is A CPU Problem?

We’ve Come A Long Way

But...

In Some Key Areas So Very Little Has Changed

“It’s Simple, So It Must Be Easy. Right?”

SO LITTLE HAS CHANGED

Just because something is simple that doesn’t mean it is easy.

SO LITTLE HAS CHANGED

- In 2016 it’s simply true that platform performance for Oracle is simple.
- But is it easy?
- Are we getting in our own way?

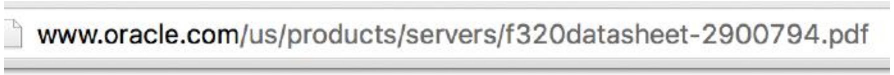
We’re Still Getting In Our Own Way

GETTING IN OUR OWN WAY

- Going out of our way to cripple data flow
  - If Fibre Channel, configure modern Oracle servers with 2 x 8GFC paths **per socket**
    - Story time
  - If dNFS, more 10GbE NICS. Period!
- Bad server choices
  - Avoid 2-hop NUMA servers at all cost unless chopping it up with virtualization
    - 4 Socket EP versus 2 Socket EP and 4 Socket EX
    - Pich the *right* Xeon SKU
- Throwing hardware at bad query plans

We Still Think Performance Is The Sum Of Components

PERFORMANCE IS NOT THE SUM OF COMPONENTS



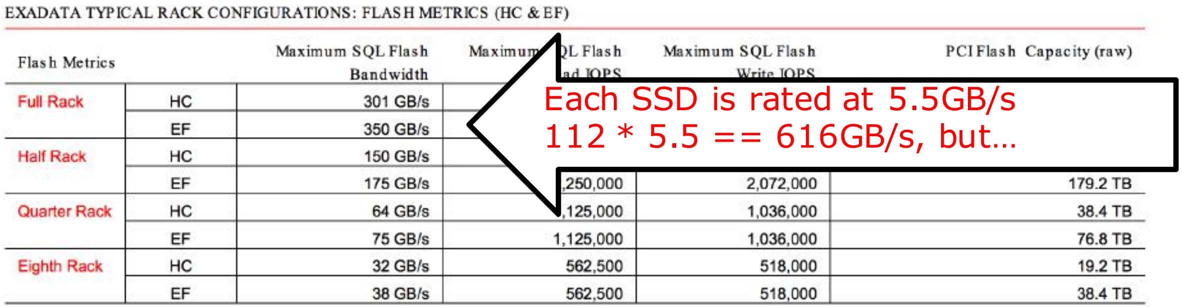
www.oracle.com/us/products/servers/f320datasheet-2900794.pdf

KEY FEATURES

- 3.2 TB NVMe device
- Eight-lane PCIe Gen 3 interface
- 520 K random IOPS (8 K), 5.5 GB/sec throughput performance *

Note: X6-2 Cells have 8 of these each. (520K * 8 == 4,160,000)

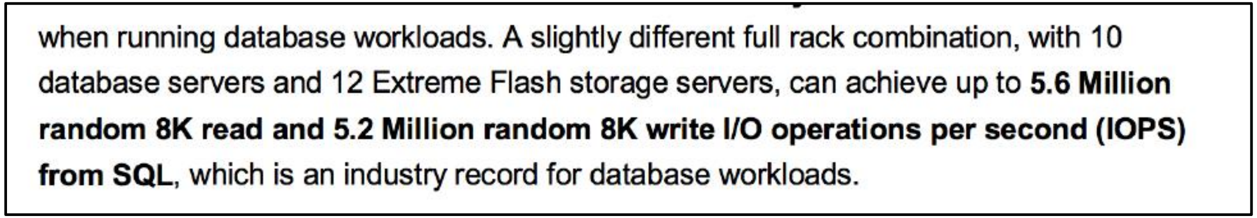
PERFORMANCE IS NOT THE SUM OF COMPONENTS



EXADATA TYPICAL RACK CONFIGURATIONS: FLASH METRICS (HC & EF)

Flash Metrics		Maximum SQL Flash Bandwidth	Maximum QL Flash ad IOPS	Maximum SQL Flash Write IOPS	PCI Flash Capacity (raw)
Full Rack	HC	301 GB/s			
	EF	350 GB/s			
Half Rack	HC	150 GB/s			
	EF	175 GB/s	,250,000	2,072,000	179.2 TB
Quarter Rack	HC	64 GB/s	,125,000	1,036,000	38.4 TB
	EF	75 GB/s	1,125,000	1,036,000	76.8 TB
Eighth Rack	HC	32 GB/s	562,500	518,000	19.2 TB
	EF	38 GB/s	562,500	518,000	38.4 TB

Each SSD is rated at 5.5GB/s
112 * 5.5 == 616GB/s, but…

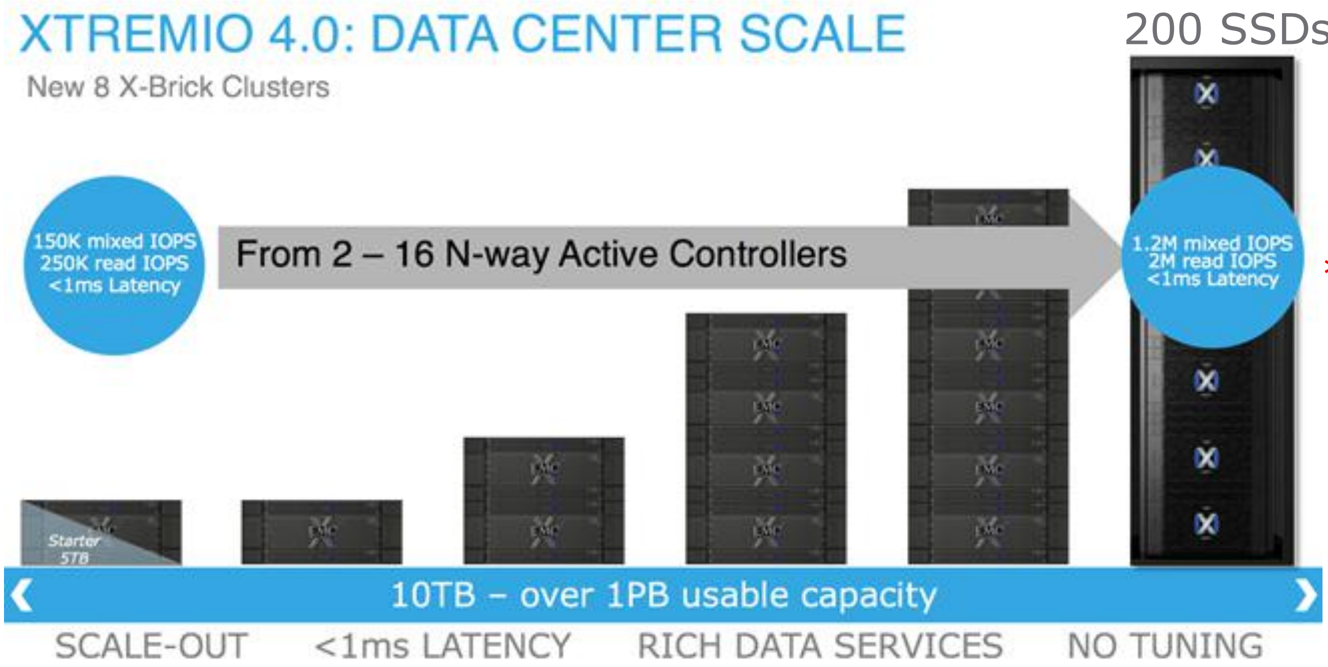


when running database workloads. A slightly different full rack combination, with 10 database servers and 12 Extreme Flash storage servers, can achieve up to **5.6 Million random 8K read and 5.2 Million random 8K write I/O operations per second (IOPS) from SQL**, which is an industry record for database workloads.

Per host GB per second:	IOPS/host:
((5600000 / 10) * 8192) / 2^30 == 4.27 ((4500000 / 8) * 8192) / 2^30 == 4.29	(5600000 / 10) == 560K (4500000 / 8) == 562K

IOPS/SSD: 4,500,000 / (8 * 14) == 40,178

PERFORMANCE IS NOT THE SUM OF COMPONENTS



XTREMIO 4.0: DATA CENTER SCALE

New 8 X-Brick Clusters

200 SSDs

150K mixed IOPS
250K read IOPS
<1ms Latency

From 2 – 16 N-way Active Controllers

1.2M mixed IOPS
2M read IOPS
<1ms Latency

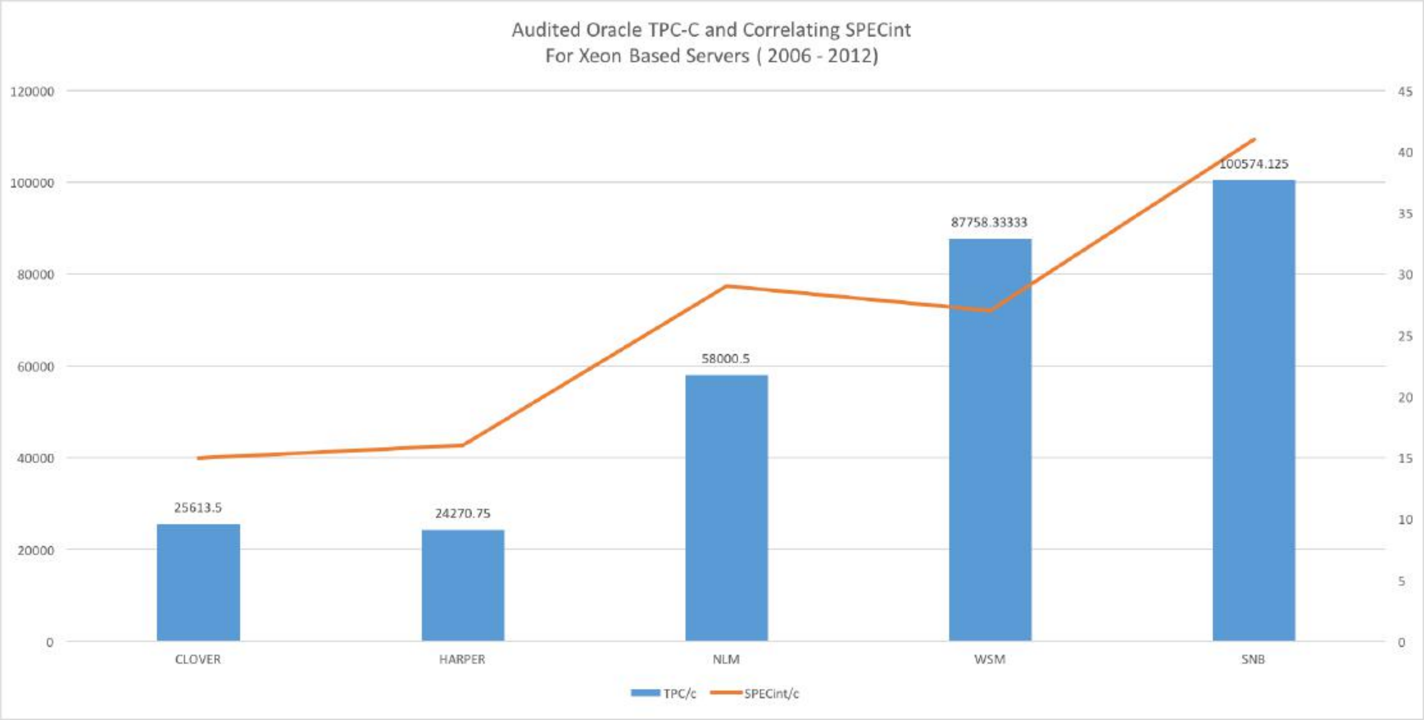
*

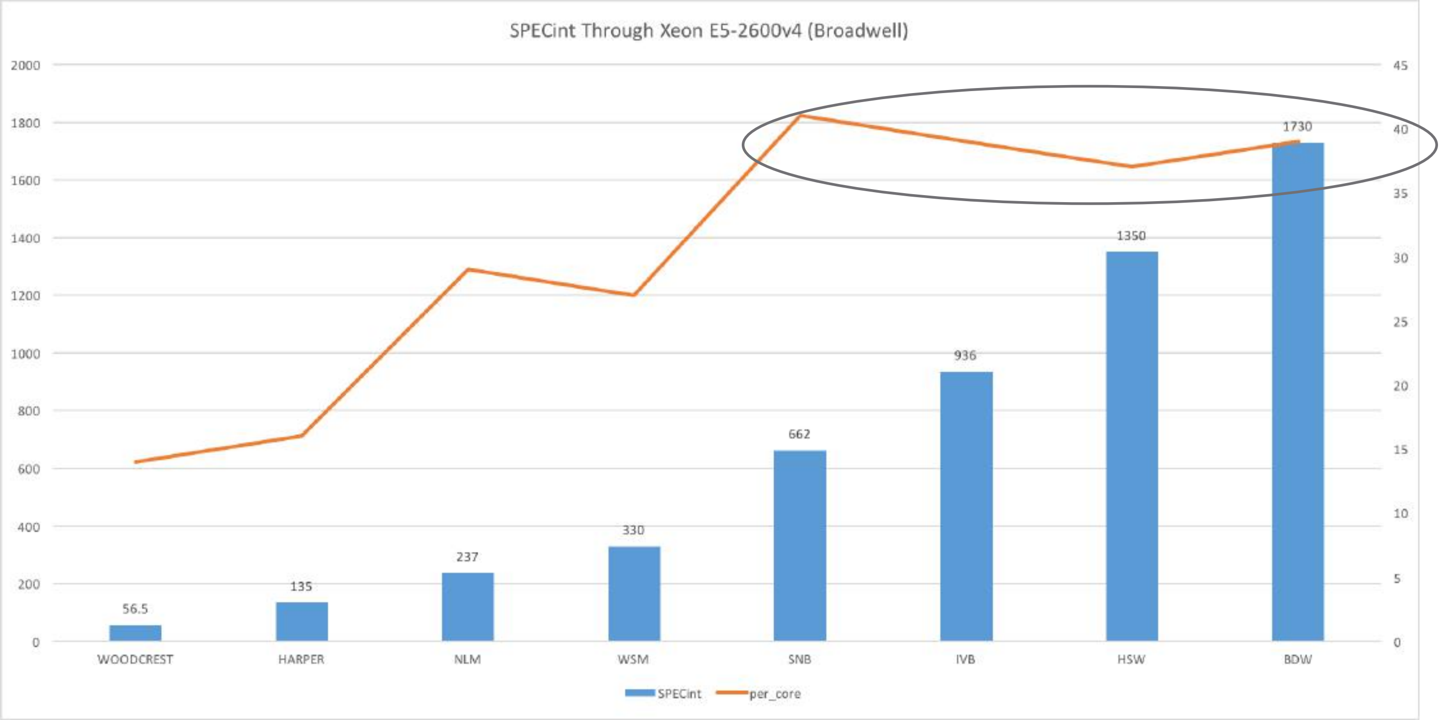
10TB – over 1PB usable capacity

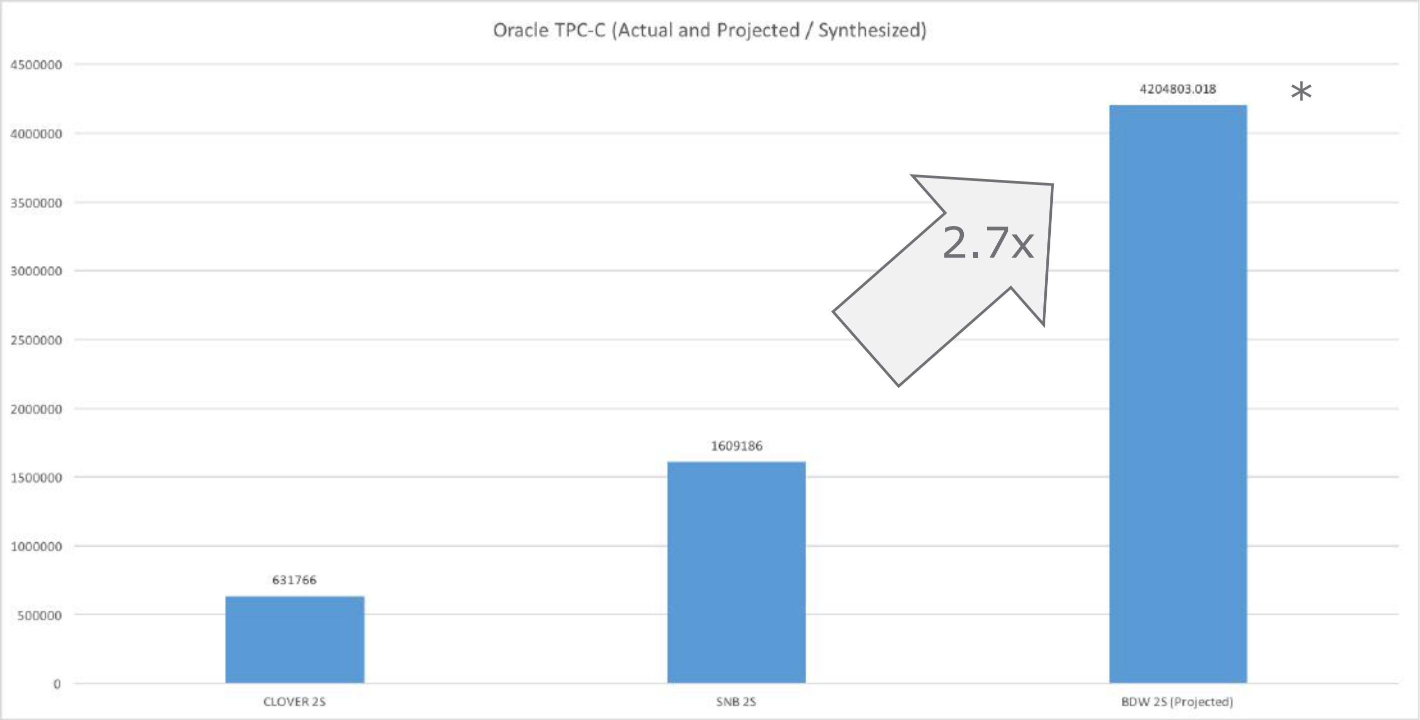
SCALE-OUT <1ms LATENCY RICH DATA SERVICES NO TUNING

* IOPS/SSD: 2,000,000 / (8 *25) ==10,000

Modern Systems Are Probably Better Than You Think


Audited Oracle TPC-C and Correlating SPECint
For Xeon Based Servers (2006 - 2012)
120000
100000
80000
60000
40000
20000
0
45
40
35
30
25
20
15
10
5
0
25613.5
24270.75
58000.5
87758.33333
100574.125
CLOVER
HARPER
NLM
WSM
SNB
TPC/c
SPECint/c


SPECint Through Xeon E5-2600v4 (Broadwell)
2000
1800
1600
1400
1200
1000
800
600
400
200
0
45
40
35
30
25
20
15
10
5
0
56.5
135
237
330
662
936
1350
1730
WOODCREST
HARPER
NLM
WSM
SNB
IVB
HSW
BDW
SPECint
per_core

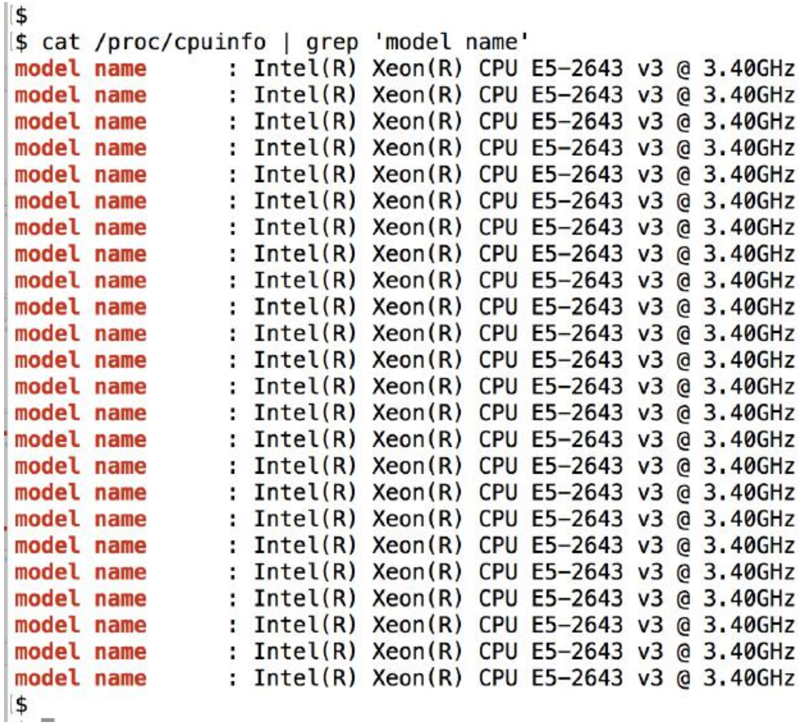


* 6/26/12: Oracle 11g 8S E7-8800 4,803,718 TpmC

Lab Example

MODERN SERVER IOPS CAPACITY

- 6-core HSW-EP Parts!

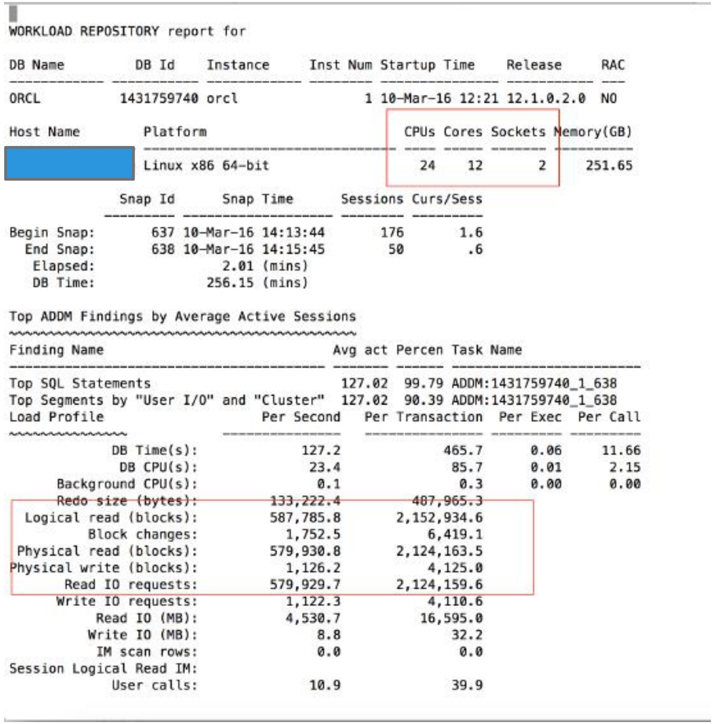


```
$
$ cat /proc/cpuinfo | grep 'model name'
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
model name      : Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz
$ _
```

MODERN SERVER IOPS CAPACITY

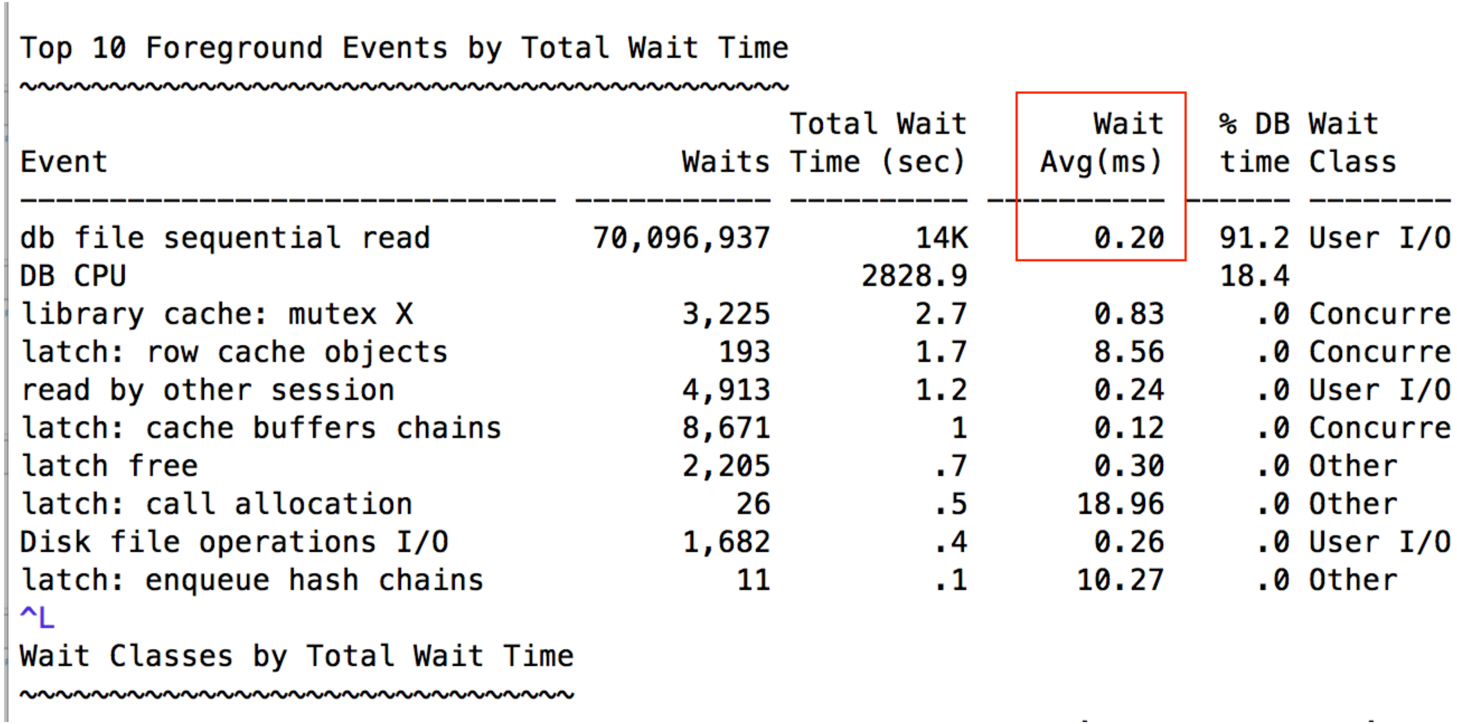
- 48,327 IOPS/c

- To put that in perspective X6-2 hosts are 22 core BDW-EP
  - 560K / 44 == 12,727



```
WORKLOAD REPOSITORY report for

DB Name         DB Id    Instance     Inst Num Startup Time    Release     RAC
------------ ----------- ------------ -------- --------------- ----------- ---
ORCL          1431759740 orcl                1 10-Mar-16 12:21 12.1.0.2.0  NO

Host Name        Platform                         CPUs Cores Sockets Memory(GB)
---------------- -------------------------------- ---- ----- ------- ----------
[redacted]       Linux x86 64-bit                   24    12       2     251.65

              Snap Id      Snap Time      Sessions Curs/Sess
            --------- ------------------- -------- ---------
Begin Snap:       637 10-Mar-16 14:13:44       176       1.6
  End Snap:       638 10-Mar-16 14:15:45        50        .6
   Elapsed:                2.01 (mins)
   DB Time:              256.15 (mins)

Top ADDM Findings by Average Active Sessions
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Finding Name                             Avg act Percen Task Name
---------------------------------------- ------- ------ ------------------------
Top SQL Statements                        127.02  99.79 ADDM:1431759740_1_638
Top Segments by "User I/O" and "Cluster"  127.02  90.39 ADDM:1431759740_1_638
Load Profile                    Per Second   Per Transaction  Per Exec  Per Call
~~~~~~~~~~~~~~~            ---------------   --------------- --------- ---------
             DB Time(s):             127.2             465.7      0.06     11.66
              DB CPU(s):              23.4              85.7      0.01      2.15
      Background CPU(s):               0.1               0.3      0.00      0.00
      Redo size (bytes):         133,222.4         487,965.3
  Logical read (blocks):         587,785.8       2,152,934.6
          Block changes:           1,752.5           6,419.1
 Physical read (blocks):         579,930.8       2,124,163.5
Physical write (blocks):           1,126.2           4,125.0
       Read IO requests:         579,929.7       2,124,159.6
      Write IO requests:           1,122.3           4,110.6
           Read IO (MB):           4,530.7          16,595.0
          Write IO (MB):               8.8              32.2
           IM scan rows:               0.0               0.0
Session Logical Read IM:
             User calls:              10.9              39.9
```

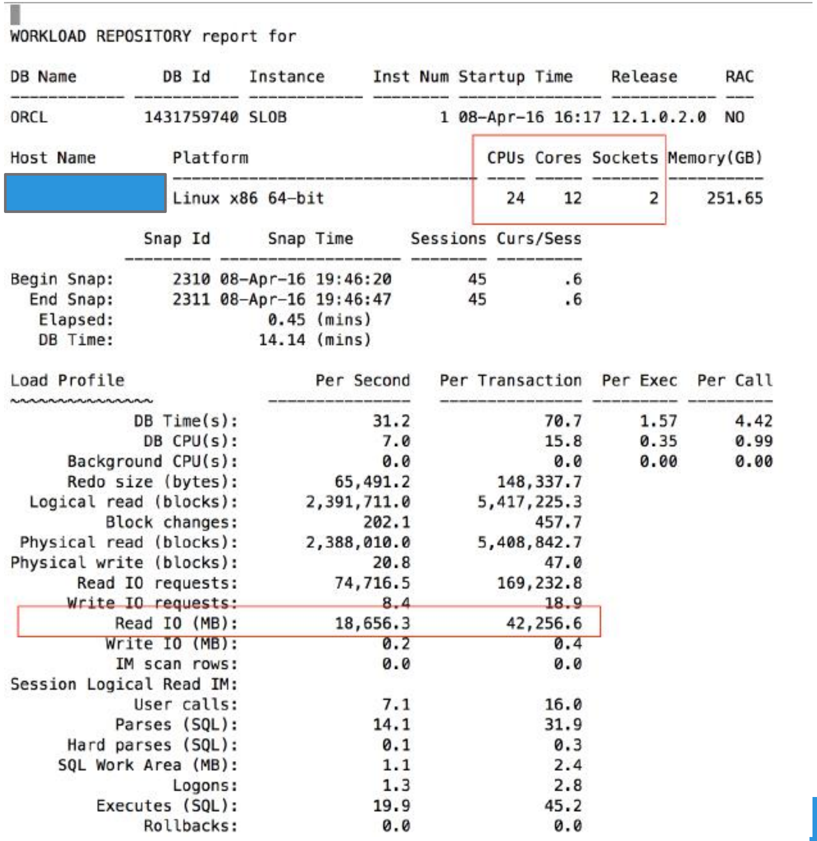
BLOCKING I/O ALWAYS REFLECTS LATENCY



```
Top 10 Foreground Events by Total Wait Time
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
                                            Total Wait       Wait   % DB Wait
Event                                 Waits Time (sec)    Avg(ms)   time Class
------------------------------ ----------- ---------- ---------- ------ --------
db file sequential read         70,096,937        14K       0.20   91.2 User I/O
DB CPU                                         2828.9              18.4
library cache: mutex X               3,225        2.7       0.83     .0 Concurre
latch: row cache objects               193        1.7       8.56     .0 Concurre
read by other session                4,913        1.2       0.24     .0 User I/O
latch: cache buffers chains          8,671          1       0.12     .0 Concurre
latch free                           2,205         .7       0.30     .0 Other
latch: call allocation                  26         .5      18.96     .0 Other
Disk file operations I/O             1,682         .4       0.26     .0 User I/O
latch: enqueue hash chains              11         .1      10.27     .0 Other
^L
Wait Classes by Total Wait Time
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

MODERN SERVER I/O BANDWIDTH

- 18+ Gigabytes/sec
- DB CPU 7
  - 10.4 GBPS/c
- 17 (71%) threads remaining for all other processing



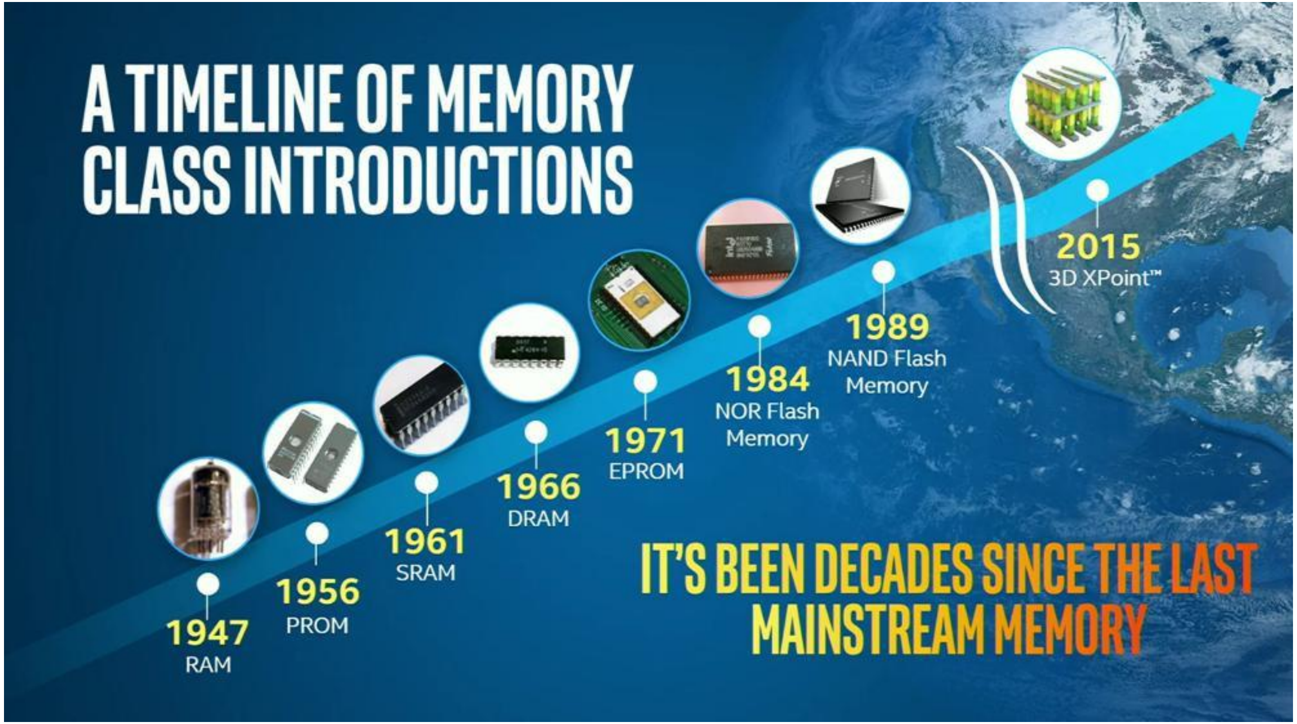
```
WORKLOAD REPOSITORY report for

DB Name         DB Id    Instance     Inst Num Startup Time    Release     RAC
------------ ----------- ------------ -------- --------------- ----------- ---
ORCL          1431759740 SLOB                1 08-Apr-16 16:17 12.1.0.2.0  NO

Host Name        Platform                         CPUs Cores Sockets Memory(GB)
---------------- -------------------------------- ---- ----- ------- ----------
[redacted]       Linux x86 64-bit                   24    12       2     251.65

              Snap Id      Snap Time      Sessions Curs/Sess
            --------- ------------------- -------- ---------
Begin Snap:      2310 08-Apr-16 19:46:20        45        .6
  End Snap:      2311 08-Apr-16 19:46:47        45        .6
   Elapsed:                0.45 (mins)
   DB Time:               14.14 (mins)

Load Profile                    Per Second   Per Transaction  Per Exec  Per Call
~~~~~~~~~~~~~~~            ---------------   --------------- --------- ---------
             DB Time(s):              31.2              70.7      1.57      4.42
              DB CPU(s):               7.0              15.8      0.35      0.99
      Background CPU(s):               0.0               0.0      0.00      0.00
      Redo size (bytes):          65,491.2         148,337.7
  Logical read (blocks):       2,391,711.0       5,417,225.3
          Block changes:             202.1             457.7
 Physical read (blocks):       2,388,010.0       5,408,842.7
Physical write (blocks):              20.8              47.0
       Read IO requests:          74,716.5         169,232.8
      Write IO requests:               8.4              18.9
           Read IO (MB):          18,656.3          42,256.6
          Write IO (MB):               0.2               0.4
           IM scan rows:               0.0               0.0
Session Logical Read IM:
             User calls:               7.1              16.0
           Parses (SQL):              14.1              31.9
      Hard parses (SQL):               0.1               0.3
     SQL Work Area (MB):               1.1               2.4
                 Logons:               1.3               2.8
         Executes (SQL):              19.9              45.2
              Rollbacks:               0.0               0.0
```

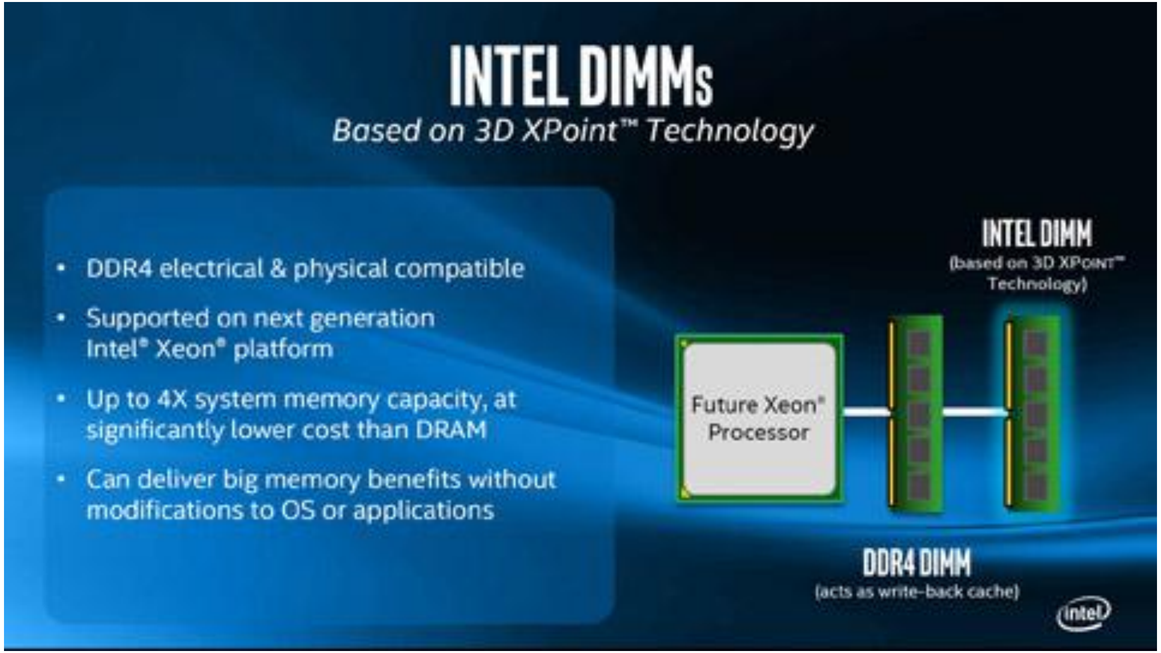
The Future Is Quite Bright And Not Too Distant

THE LESS THINGS STAY THE SAME THE MORE THEY CHANGE

- Machines are not (always) machines (VM vs physical)
- A CPU is not always a CPU (Threaded CPUs)
- Memory is not (always) predictable (NUMA)
- Clock frequency is not (always) predictable (TurboBoost)

- And soon…
- **Main memory is not (always) DRAM**


A TIMELINE OF MEMORY CLASS INTRODUCTIONS
1947
RAM
1956
PROM
1961
SRAM
1966
DRAM
1971
EPROM
1984
NOR Flash Memory
1989
NAND Flash Memory
2015
3D XPoint™
IT'S BEEN DECADES SINCE THE LAST MAINSTREAM MEMORY

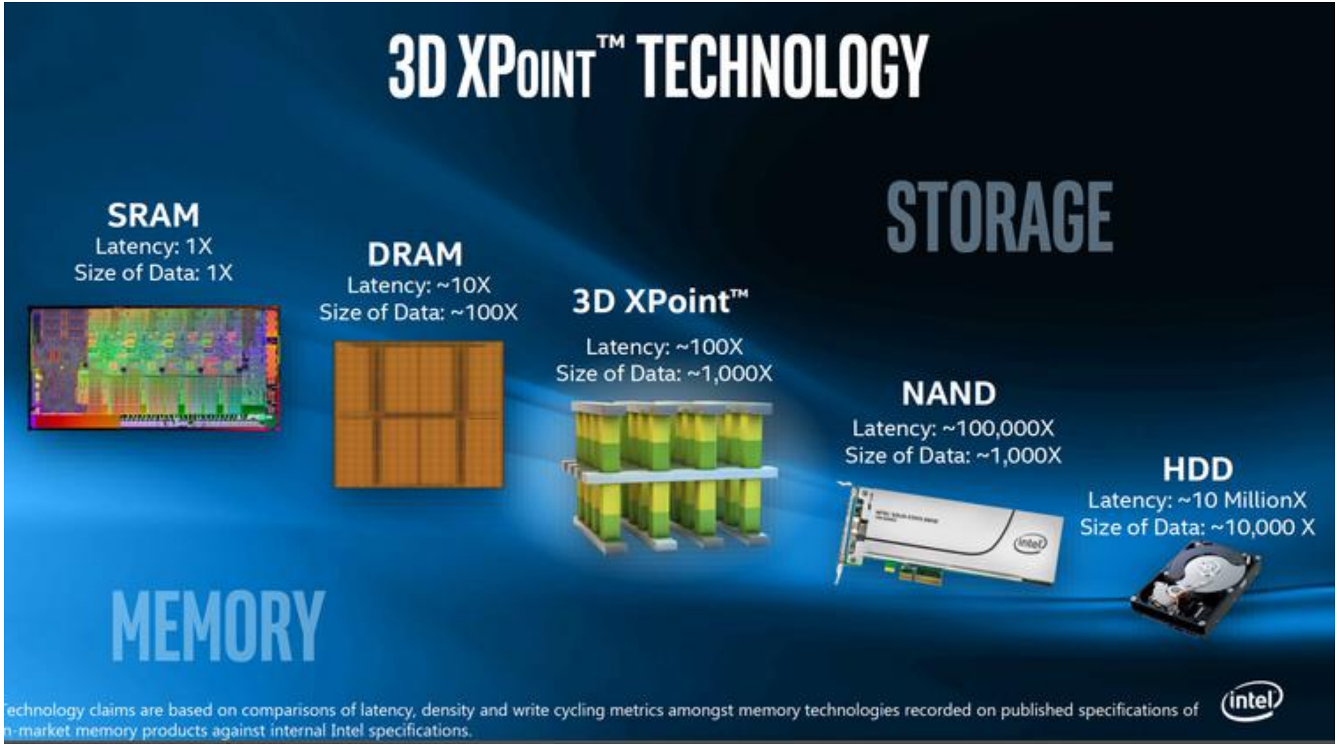


INTEL DIMMs

Based on 3D XPoint™ Technology

- DDR4 electrical & physical compatible
- Supported on next generation Intel® Xeon® platform
- Up to 4X system memory capacity, at significantly lower cost than DRAM
- Can deliver big memory benefits without modifications to OS or applications

INTEL DIMM
(based on 3D XPoint™ Technology)
Future Xeon® Processor
DDR4 DIMM
(acts as write-back cache)
intel



3D XPoint™ TECHNOLOGY

STORAGE

SRAM
Latency: 1X
Size of Data: 1X

DRAM
Latency: ~10X
Size of Data: ~100X

3D XPoint™
Latency: ~100X
Size of Data: ~1,000X

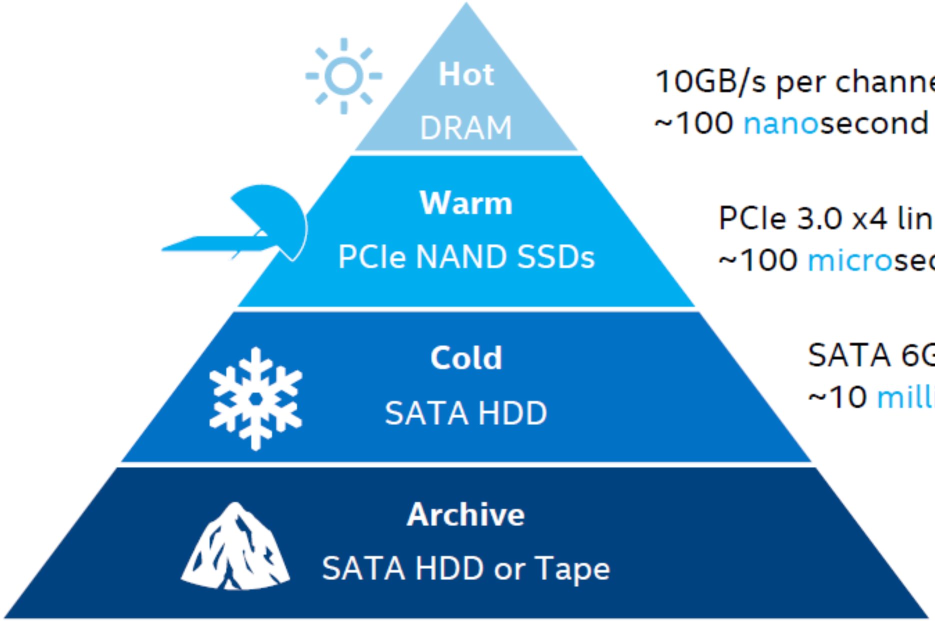
NAND
Latency: ~100,000X
Size of Data: ~1,000X

HDD
Latency: ~10 MillionX
Size of Data: ~10,000 X

MEMORY

echnology claims are based on comparisons of latency, density and write cycling metrics amongst memory technologies recorded on published specifications of n-market memory products against internal Intel specifications.

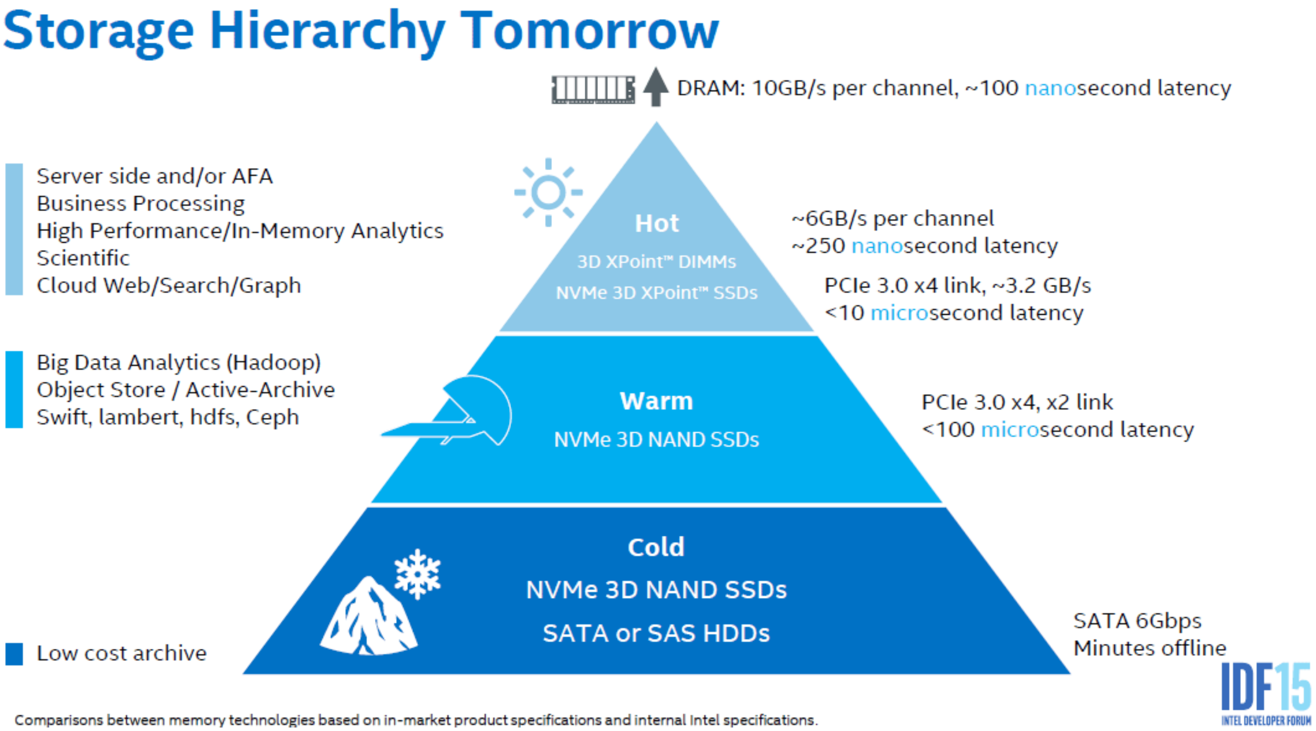
Storage and Memory Hierarchy Today



10GB/s per channel
~100 nanosecond latency

PCIe 3.0 x4 link, ~3.2 GB/s
~100 microsecond latency

SATA 6Gbps, ~540 MB/s
~10 millisecond latency



Storage Hierarchy Tomorrow

DRAM: 10GB/s per channel, ~100 nanosecond latency

- Server side and/or AFA
- Business Processing
- High Performance/In-Memory Analytics
- Scientific
- Cloud Web/Search/Graph

Hot

3D XPoint™ DIMMs

NVMe 3D XPoint™ SSDs

~6GB/s per channel
~250 nanosecond latency

PCIe 3.0 x4 link, ~3.2 GB/s
<10 microsecond latency

- Big Data Analytics (Hadoop)
- Object Store / Active-Archive
- Swift, lambert, hdfs, Ceph

Warm

NVMe 3D NAND SSDs

PCIe 3.0 x4, x2 link
<100 microsecond latency

- Low cost archive

Cold

NVMe 3D NAND SSDs

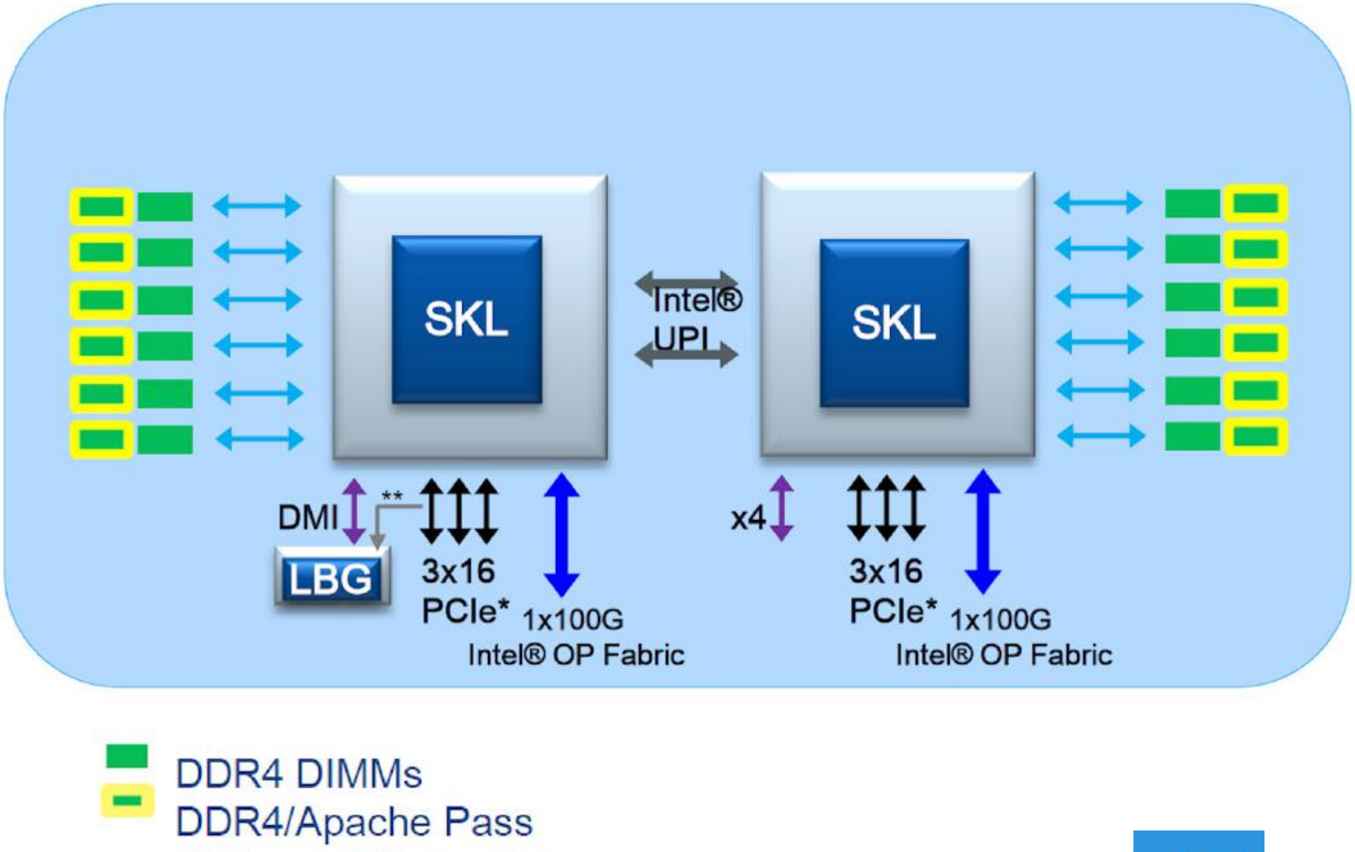
SATA or SAS HDDs

SATA 6Gbps
Minutes offline

Comparisons between memory technologies based on in-market product specifications and internal Intel specifications.

Sad it doesn't say Database or RDBMS or OLTP anywhere ☹

INTEL “SKYLAKE” XEON (SKL-EP)



So It Is Really More Than…

IS IT REALLY JUST THIS?

VS

THANK YOU

EMC^2®