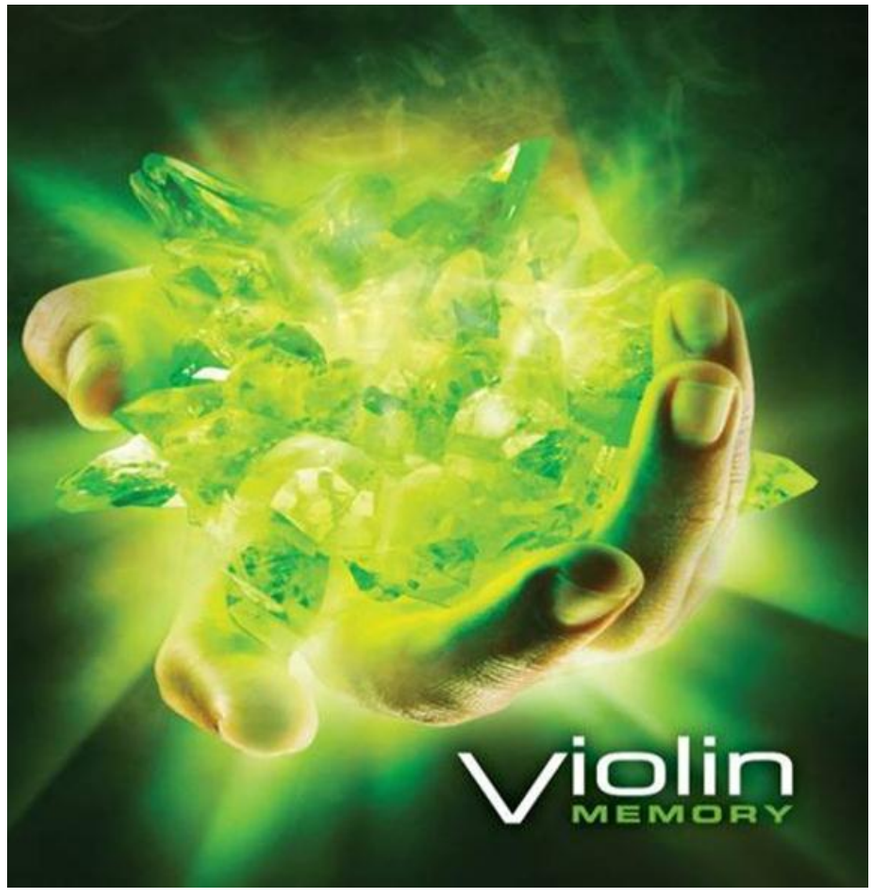
The All Flash Database – It's Not Just Cache Anymore

Session 378

Randy Cook



IOUG independent oracle users group 20th ANNIVERSARY

COLLABORATE 13
TECHNOLOGY AND APPLICATIONS FORUM FOR THE ORACLE COMMUNITY

Randy Cook

18+ years as an Oracle Database Animal Trainer – It has always been about fighting the IO

First 11 in the Denver/Boulder area as a consultant learning how to be a DBA

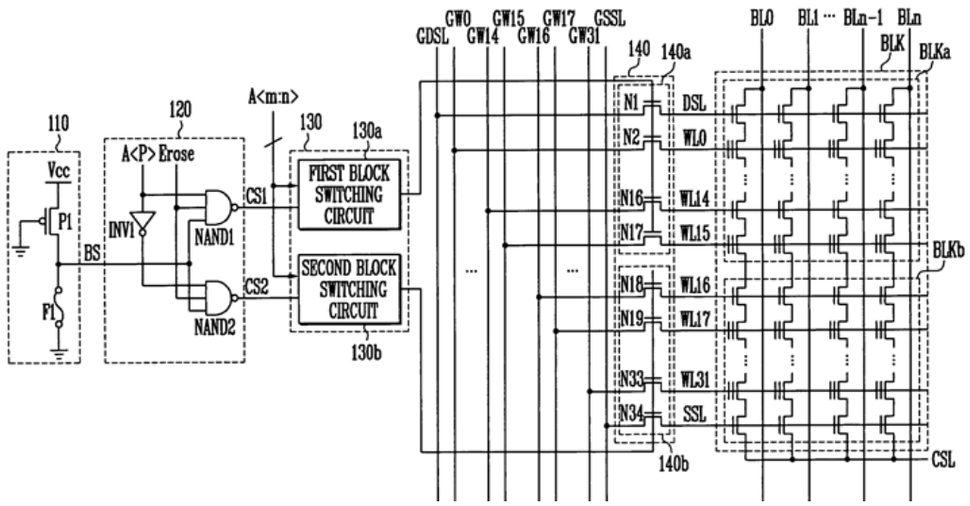
Next 5 around the country as a tuning specialist and architect with TUSC

Last 2+ with Violin Memory spreading the word of flash around the world

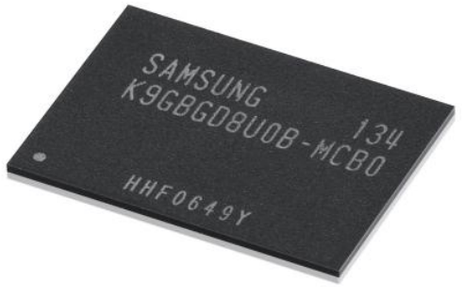
Agenda

- Flash – what, where, why, how
- Identifying the IO Bottleneck
- Adopting a New Paradigm

What is Flash???



- It's NOT DRAM
- Solid State Persistent Storage
- Invented by Dr. Fujio Masouka in 1980 while working for Toshiba



IOUG
independent oracle users group
20th ANNIVERSARY

It's Everywhere !!

65% of all NAND Flash Manufactured by Toshiba and Samsung*

http://vimeo.com/toshiba/nand-flash-deprevation-247

COLLABORATE 13
TECHNOLOGY AND APPLICATIONS FORUM FOR THE ORACLE COMMUNITY

* http://www.dramexchange.com/WEEKLYRESEARCH/Post/2/3121.html

The Nomenclature

IOPS

Latency

Bandwidth

Quiz Time

1) Web based booking system. 100,000 transactions per second…

2) Decision Support System for the COO. How many tennis balls did we sell last quarter?

IOPS

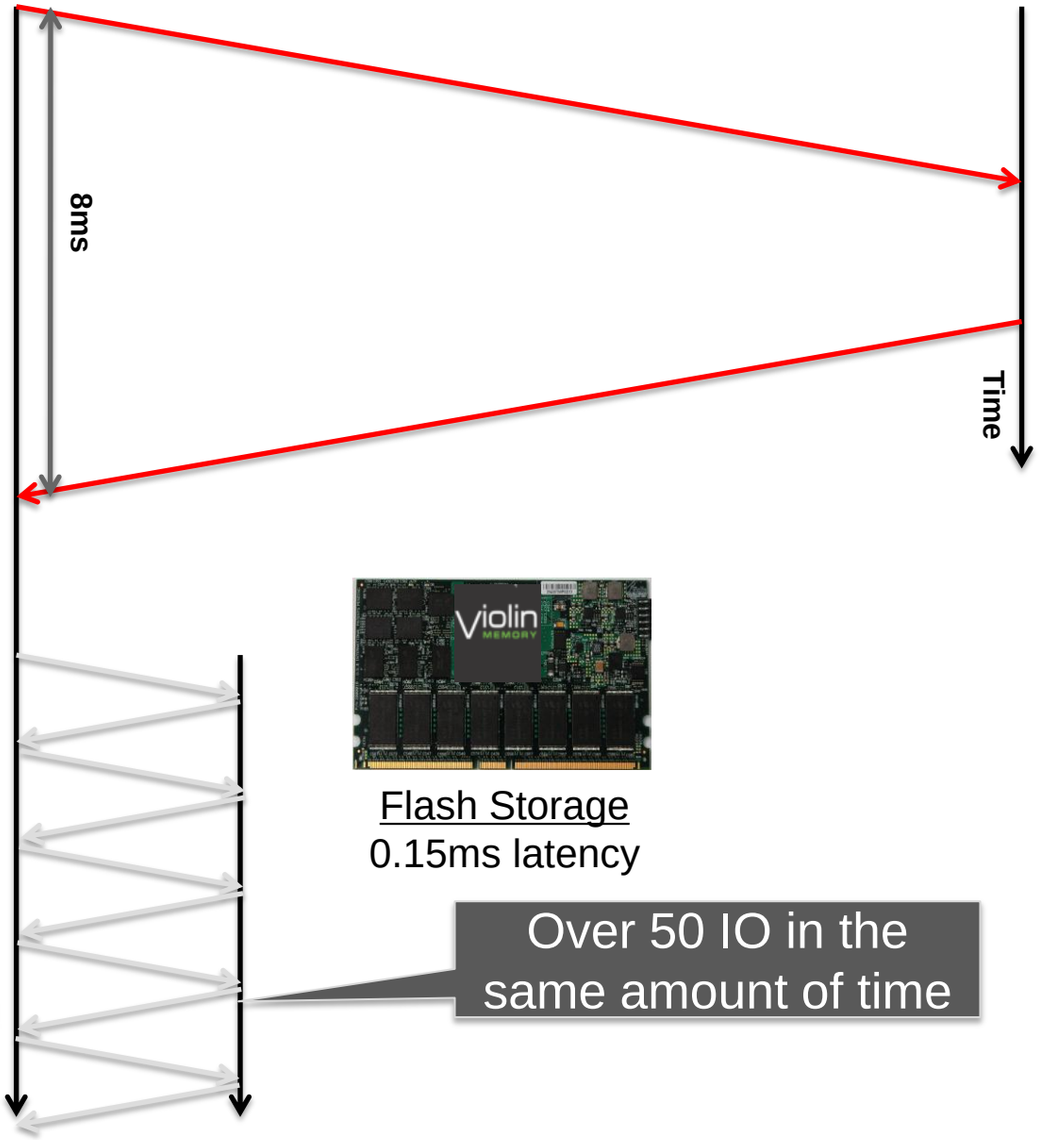
OR

Bandwidth

Visualizing Latency

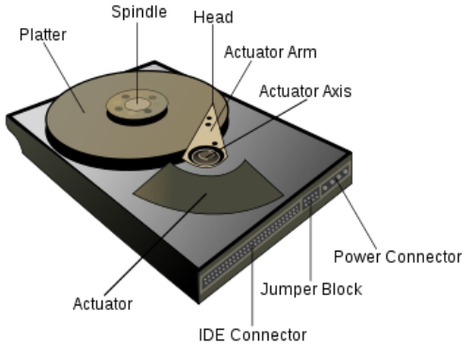
I/O Bound Apps

100K IOPS per CPU
- Oracle
- SQL Server
- Exchange
- Etc…



Flash Storage
0.15ms latency

Over 50 IO in the same amount of time



HDD Storage
8ms latency

Total latency =
seek time + rotational latency

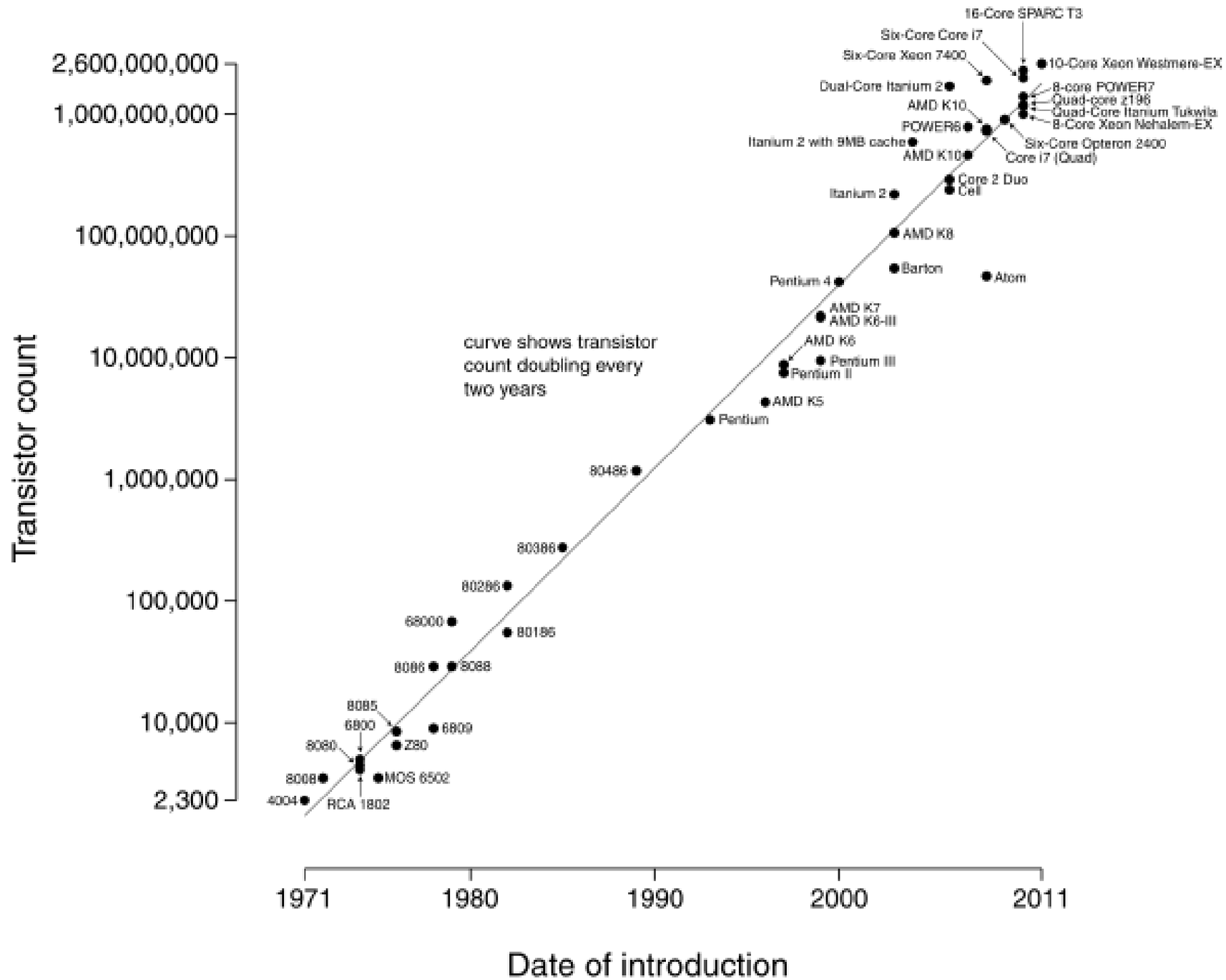
IOUG independent oracle users group 20th ANNIVERSARY

COLLABORATE 13
TECHNOLOGY AND APPLICATIONS FORUM FOR THE ORACLE COMMUNITY

Moore's Law

The observation that over the history of computing hardware, the number of transistors on integrated circuits doubles approximately every two years.

Intel Co-Founder, Gordon E. Moore

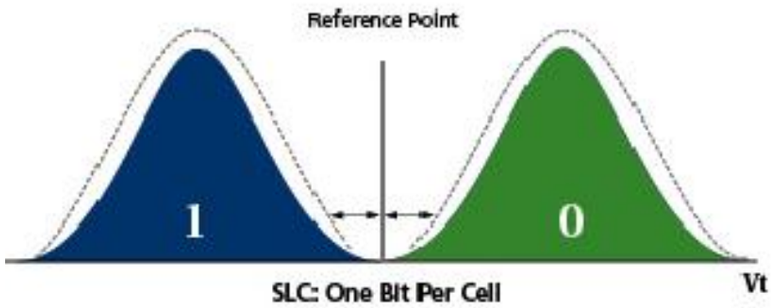


SLC versus MLC

SLC – Single Level Cell

Faster Write Speeds = High Performance
Higher Cell Endurance → 100,000 program/erase cycles per cell
Lower Power Consumption
=
A good fit for databases, embedded systems, critical applications

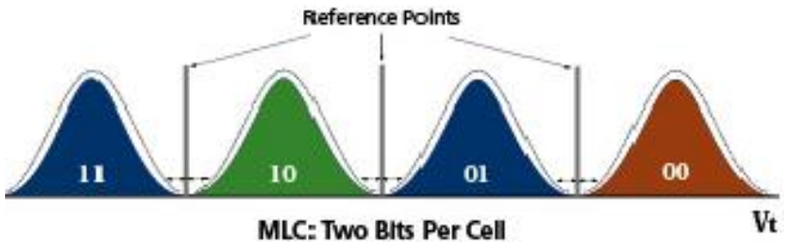
One bit per cell = More $$$/GB



MLC – Multi-Level Cell

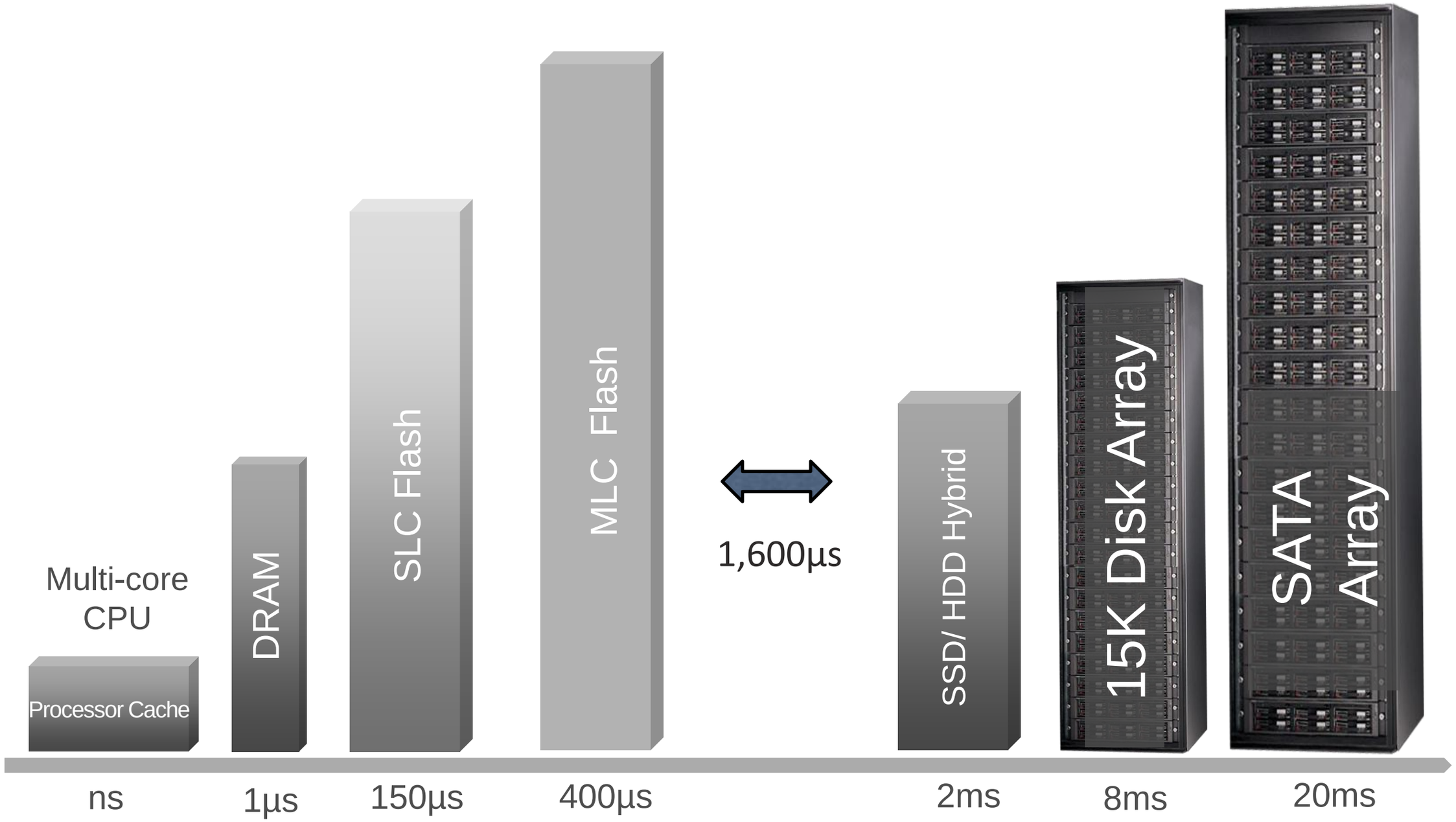
Two bits per cell = High Capacity = More GB/$$$

Lower write speeds
Lower cell endurance → 10,000 program/erase cycles per cell
Higher Power Consumption
=
A good fit for static data. Not suggested for applications which require frequent update of data.

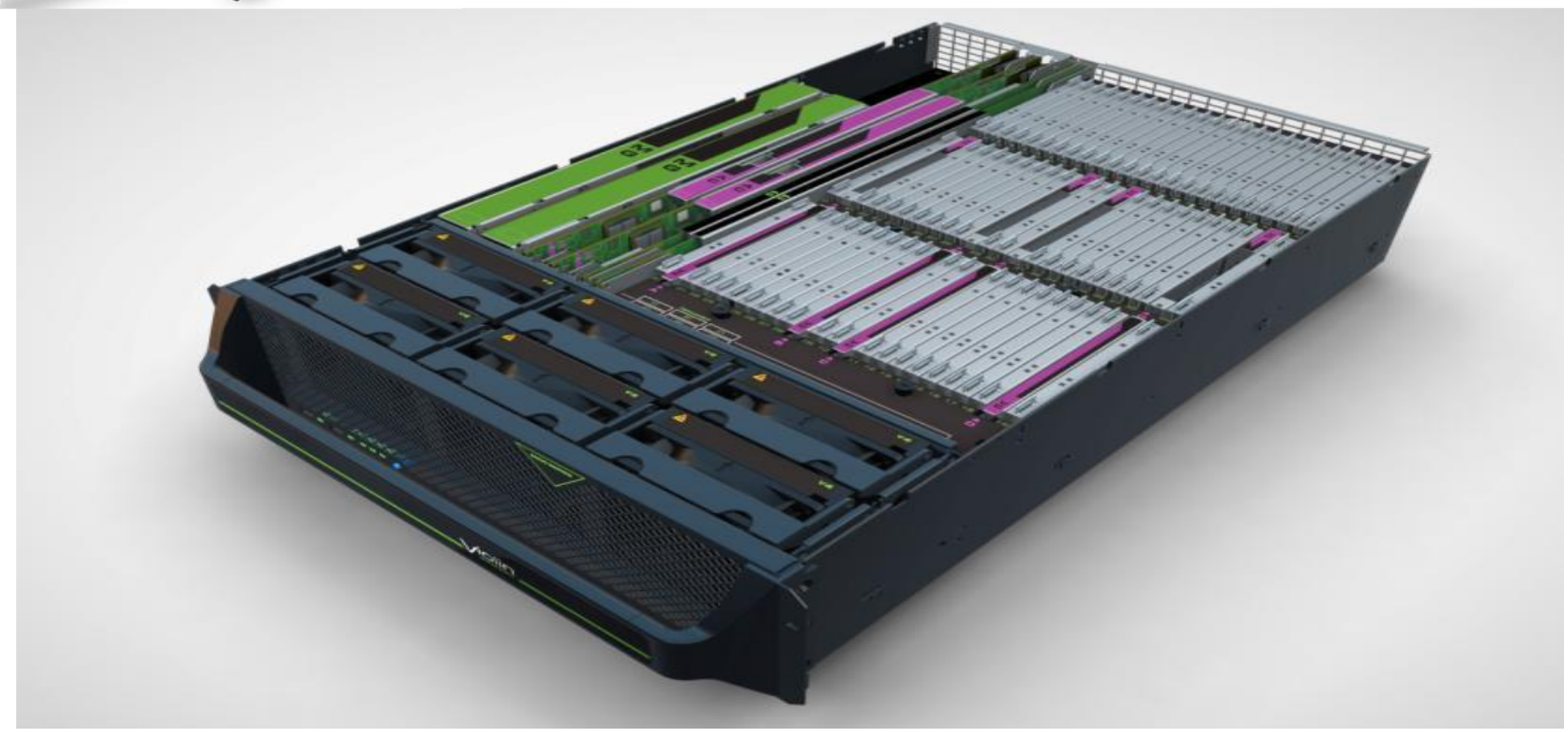


Note – eMLC is MLC designed for lower error rates at an even higher cost to performance.

Latency Comparison

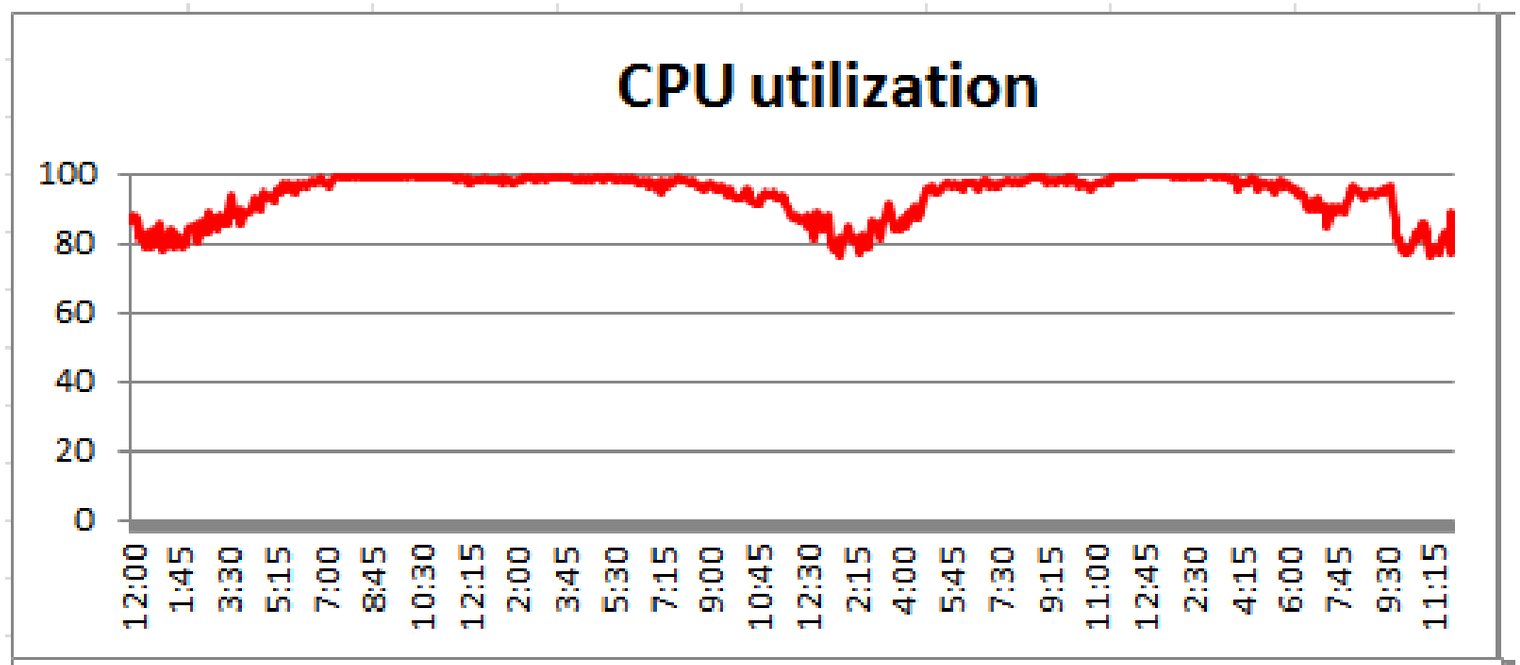


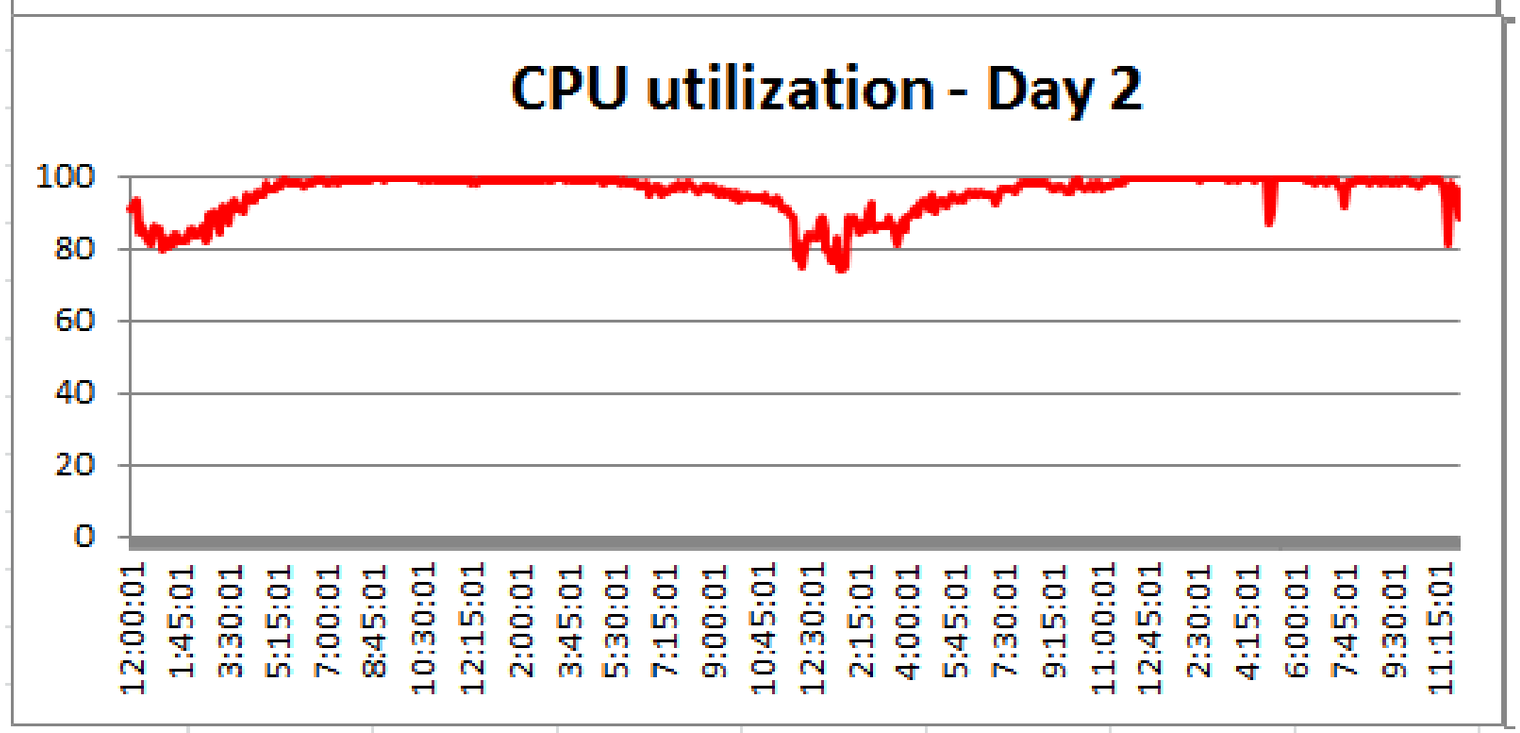
PCI Cards and Arrays



Identifying the Bottleneck

Are we CPU or IO bound?



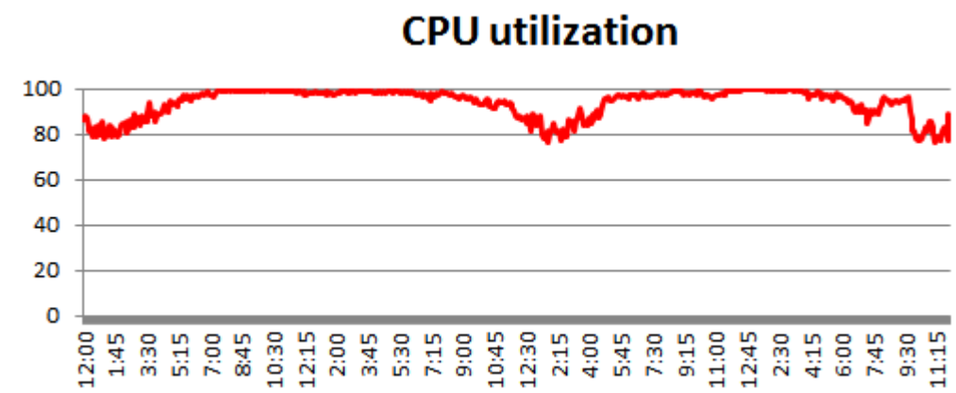


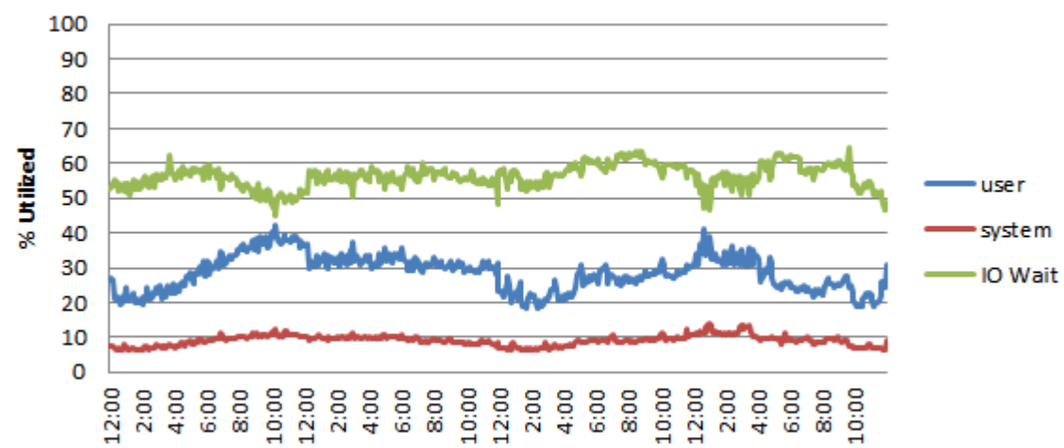
The SAR Report

[root@tme-2svr ~]# sar -u

12:00:01AM	CPU	%user	%nice	%system	%iowait	%steal	%idle
12:05:01AM	all	28.2	0	7.72	54.57	0	9.5
12:10:01AM	all	29.12	0	8.03	54.36	0	8.49
12:15:01AM	all	29.94	0	8.36	54.97	0	6.73
12:20:01AM	all	22.09	0	7.32	55.12	0	15.47
12:25:01AM	all	22.44	0	7.17	57.4	[illegible]	[illegible]
12:30:01AM	all	21.01	0	6.92	57.58	[illegible]	[illegible]
12:35:01AM	all	22.3	0	6.92	54.4	0	16.38
12:40:01AM	all	20.61	0	6.91	54.96	0	17.51
12:45:01AM	all	20.93	0	6.82	53.76	0	18.49
12:50:01AM	all	21.07	0	6.95	54.14	0	17.83
12:55:01AM	all	24.03	0	7.74	54.13	0	14.1
1:00:01 AM	all	21.69	0	7.97	55.6	0	14.74
1:05:01 AM	all	23.37	0	7.51	54.44	0	14.67
1:10:01 AM	all	19.8	0	6.61	53.56	0	20.03
1:15:01 AM	all	21.37	0	6.81	54	0	17.83
1:20:01 AM	all	21.91	0	6.96	53.92	0	17.2
1:25:01 AM	all	20.25	0	6.64	53.21	0	19.91
1:30:01 AM	all	20.36	0	6.73	54.32	0	18.59
1:35:01 AM	all	20.05	0	6.76	54.17	0	19.03
1:40:01 AM	all	22.05	0	6.99	54.58	0	16.38
1:45:01 AM	all	22.43	0	6.93	54.15	0	16.49
1:50:01 AM	all	20.92	0	6.77	54.18	0	18.13
1:55:01 AM	all	20.76	0	6.89	54.4	0	17.95
2:00:01 AM	all	21.96	0	7.04	54.14	0	16.86
2:05:01 AM	all	20.95	0	6.85	54.22	0	17.98
2:10:01 AM	all	21.49	0	6.98	54.83	0	16.7
2:15:01 AM	all	22.02	0	7.41	56.12	0	14.45
2:20:01 AM	all	22.42	0	7.12	54.5	0	15.97
2:25:01 AM	all	21.72	0	6.99	55.7	0	15.59
2:30:01 AM	all	22.32	0	7.07	54.52	0	16.08
2:35:01 AM	all	23.27	0	7.15	54.3	0	15.28
2:40:01 AM	all	22.48	0	7.27	56.57	0	13.68
2:45:01 AM	all	21.13	0	6.83	54.19	0	17.85
2:50:01 AM	all	22.31	0	7.08	54.56	0	16.05
2:55:01 AM	all	23.73	0	8.08	57.62	0	10.57
3:00:01 AM	all	24	0	8.47	56.76	0	10.77
3:05:01 AM	all	24.29	0	7.93	58.12	0	9.66
3:10:01 AM	all	23.13	0	7.24	55.82	0	13.81
3:15:01 AM	all	22.16	0	7.04	55.36	0	15.44
3:20:01 AM	all	23.52	0	7.68	57.25	0	11.55
3:25:01 AM	all	24.89	0	8.18	58.3	0	8.63
3:30:02 AM	all	23.29	0	7.84	58.16	0	10.71
3:35:01 AM	all	22.29	0	7.35	57.41	0	12.94
3:40:01 AM	all	24.16	0	8.13	60.33	0	7.38
3:45:01 AM	all	23.56	0	7.68	58.89	0	9.87
3:50:02 AM	all	25.77	0	8.46	59.25	0	6.51
3:55:01 AM	all	26.24	0	7.92	56.96	0	8.87
4:00:01 AM	all	24.71	0	7.89	58.16	0	9.24
4:05:01 AM	all	25.44	0.02	8.13	57.96	0	8.45
4:10:01 AM	all	24.26	0	7.87	58.16	0	9.71
4:15:01 AM	all	28.6	0	8.71	57.37	0	5.32
4:20:01 AM	all	25.92	0	8.33	59.24	0	6.51
4:25:01 AM	all	26.29	0	8.49	58.89	0	6.33
4:30:01 AM	all	25.74	0	8.65	59.26	0	6.35
4:35:01 AM	all	27.91	0	9.03	58.26	0	4.8
4:40:01 AM	all	28.36	0	9.01	58.5	0	4.13
4:45:01 AM	all	27.3	0	8.71	59.24	0	4.74
4:50:01 AM	all	26.28	0	8.71	59.91	0	5.1
4:55:01 AM	all	30.62	0	11.01	56.46	0	1.92
5:00:01 AM	all	29.88	0	9.58	57.6	0	2.93
5:05:01 AM	all	30.56	0	9.12	56.72	0	3.6
5:10:01 AM	all	29.56	0	9.17	57.82	0	3.46
5:15:01 AM	all	28.31	0	9.02	58.98	0	3.69
5:20:01 AM	all	33.23	0	9.81	54.81	0	2.16
5:25:02 AM	all	29.29	0	9.4	58.41	0	2.9
5:30:01 AM	all	29.98	0	9.47	57.92	0	2.63
5:35:01 AM	all	31.66	0	10.08	56.79	0	1.48
5:40:02 AM	all	32.17	0	10.33	56.2	0	1.31

%iowait --- Bingo !!





COLLABORATE 13
TECHNOLOGY AND APPLICATIONS FORUM
FOR THE ORACLE COMMUNITY

The AWR Report

IO Indicators to look for in Four Easy Steps

Load Profile

	Per Second	Per Transaction
Redo size:	2,963,178.53	1,017.42
Logical reads:	75,484.91	25.92
Block changes:	19,722.21	6.77
Physical reads:	4,799.55	1.65
Physical writes:	2,737.06	0.94
User calls:	4,635.09	1.59
Parses:	5,756.72	1.98
Hard parses:	0.00	0.00
Sorts:	2,229.37	0.77
Logons:	0.04	0.00
Executes:	9,019.64	3.10
Transactions:	2,912.45	

Physical !!

Top 5 Timed Events

Event	Waits	Time(s)	Avg Wait(ms)	% Total Call Time	Wait Class
db file sequential read	17,292,809	358,703	21	77.7	User I/O
log file sync	10,485,848	90,996	9	19.7	Commit
CPU time		32,220		7.0	
db file parallel write	4,181,646	9,475	2	2.1	System I/O
log file parallel write	1,337,917	3,284	2	.7	System I/O

Look for the I/O !!

IO Stats

- Tablespace IO Stats
- File IO Stats

Back to Top

Latency

Tablespace IO Stats

- ordered by IOs (Reads + Writes) desc

Tablespace	Reads	Av Reads/s	Av Rd(ms)	Av Blks/R	Writes	Av Wr s/s	Buffer Waits	Av Buf Wt(ms)
P1TBS	548,436	152	20.77	1.	248,729		185	17.89
P14TBS	551,286	153	20.74	1.	239,582	67	251	17.45
P5TBS	545,707	152	20.77	1.	241,888	67	531	18.32
P10TBS	545,718	152	20.78	1.	241,141	67	44	18.72
P9TBS	547,710	152	20.70	1.	239,081	66	2	18.10
P15TBS	544,350	151	20.80	1.	241,826	67	242	19.79
P29TBS	547,098	152	20.78	1.	235,515	65	463	15.96
P16TBS	547,533	152	20.77	1.	234,987	65	765	8.39
P27TBS	546,741	152	20.74	1.	233,014	65	410	1

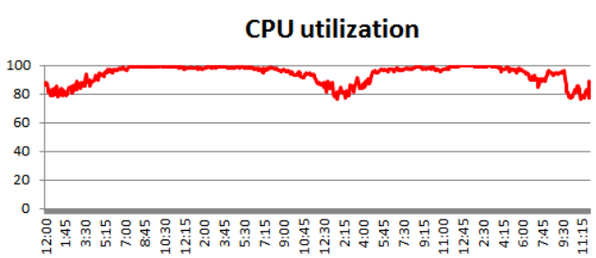
File IO Stats

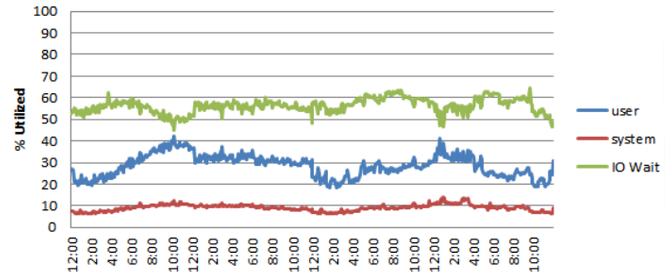
- ordered by Tablespace, File

Tablespace	Filename	Reads	Av Reads/s	Av Rd(ms)	Av Blks/Rd	Writes	Av Writes/s	Buffer Waits	Av Buf Wt(ms)
P0TBS	+DATAMMDB01/mmdb01/datafile/datatbs2.533.708266061	188,678	52	20.62	1.00	72,667	20	77	20.78
P0TBS	+DATAMMDB01/mmdb01/datafile/p0tbs.441.717171495	163,711	45	20.75	1.00	77,456	22	52	20.19
P0TBS	+DATAMMDB01/mmdb01/datafile/p0tbs.557.711215913	183,214	51	20.73	1.00	76,230	21	65	28.00
P10TBS	+DATAMMDB01/mmdb01/datafile/p10tbs.345.731971089	24,437	7	20.89	1.00	26,052	7	12	10.00
P10TBS	+DATAMMDB01/mmdb01/datafile/p10tbs.474.716437083	177,098	49	20.86	1.00	80,847	22	141	20.78

Before and after Flash

Before





Load Profile

	Per Second	Per Transaction
Redo size:	2,963,178.53	1,017.42
Logical reads:	75,484.91	25.92
Block changes:	19,722.21	6.77
Physical reads:	4,799.55	1.65
Physical writes:	2,737.06	0.94
User calls:	4,635.09	1.59
Parses:	5,756.72	1.98
Hard parses:	0.00	0.00
Sorts:	2,229.37	0.77
Logons:	0.04	0.00
Executes:	9,019.64	3.10
Transactions:	2,912.45	

Top 5 Timed Events

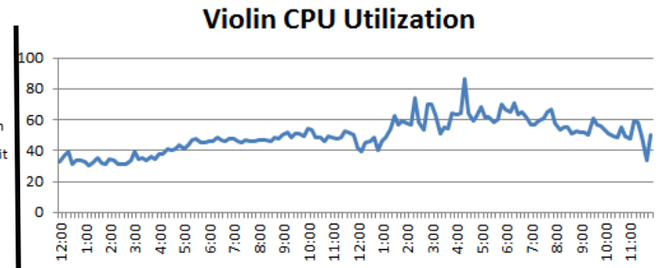
Event	Waits	Time(s)	Avg Wait(ms)	% Total Call Time	Wait Class
db file sequential read	17,292,809	358,703	21	77.7	User I/O
log file sync	10,485,848	90,996	9	19.7	Commit
CPU time		32,220		7.0	
db file parallel write	4,181,646	9,475	2	2.1	System I/O
log file parallel write	1,337,917	3,284	2	.7	System I/O

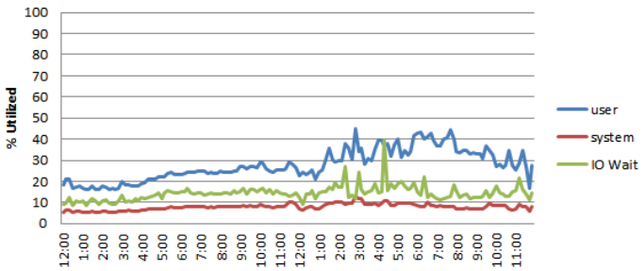
File IO Stats

- ordered by Tablespace, File

Tablespace	Filename	Reads	Av Reads/s	Av Rd(ms)	Av Blks/Rd	Writes	Av Writes/s	Buffer Waits	Av Buf Wt(ms)
P0TBS	+DATAMMDB01/mmdb01/datafile/datatbs2.533.708266061	188,678	52	20.62	1.00	72,667	20	77	20.78
P0TBS	+DATAMMDB01/mmdb01/datafile/p0tbs.441.717171495	163,711	45	20.75	1.00	77,456	22	52	20.19
P0TBS	+DATAMMDB01/mmdb01/datafile/p0tbs.557.711215913	183,214	51	20.73	1.00	76,230	21	65	28.00
P10TBS	+DATAMMDB01/mmdb01/datafile/p10tbs.345.731971089	24,437	7	20.89	1.00	26,052	7	12	10.00
P10TBS	+DATAMMDB01/mmdb01/datafile/p10tbs.474.716437083	177,098	49	20.86	1.00	80,847	22	141	20.78

After





Load Profile

	Per Second	Per Transaction
Redo size:	2,785,931.11	966.96
Logical reads:	77,542.73	26.91
Block changes:	19,583.60	6.80
Physical reads:	4,716.76	1.64
Physical writes:	2,790.56	0.97
User calls:	4,563.21	1.58
Parses:	5,686.36	1.97
Hard parses:	0.00	0.00
Sorts:	2,299.96	0.80
Logons:	0.04	0.00
Executes:	8,910.12	3.09

Top 5 Timed Events

Event	Waits	Time(s)	Avg Wait(ms)	% Total Call Time	Wait Class
CPU time		23,035		87.2	
log file sync	10,337,686	9,372	1	35.5	Commit
db file sequential read	16,980,465	9,124	1	34.5	User I/O
log file parallel write	5,403,199	2,543	0	9.6	System I/O
log file sequential read	11,899	155	13	.6	System I/O

File IO Stats

- ordered by Tablespace, File

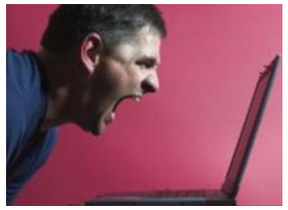
Tablespace	Filename	Reads	Av Reads/s	Av Rd(ms)	Av Blks/Rd	Writes	Av Writes/s	Buffer Waits	Av Buf Wt(ms)
P32TBS	+DATAMMDB02/mmdb02/datafile/p32tbs.400.717902943	164,647	46	0.54	1.00	76,355	21	6	0.00
P32TBS	+DATAMMDB02/mmdb02/datafile/p32tbs.510.710643793	167,579	47	0.52	1.00	73,431	20	14	0.71
P32TBS	+DATAMMDB02/mmdb02/datafile/p32tbs.511.710644043	203,248	56	0.54	1.00	95,498	27	20	0.50
P33TBS	+DATAMMDB02/mmdb02/datafile/p33tbs.416.732796149	3,108	1	0.48	1.00	5,286	1	1	0.00
P33TBS	+DATAMMDB02/mmdb02/datafile/p33tbs.445.717197343	192,211	53	0.54	1.00	81,878	23	6	3.33

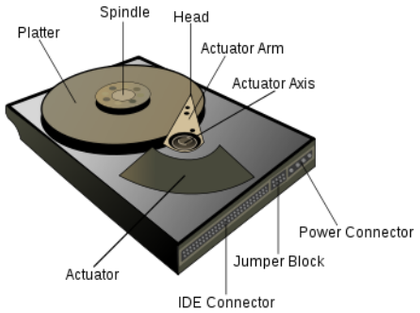
COLLABORATE 13
TECHNOLOGY AND APPLICATIONS FORUM
FOR THE ORACLE COMMUNITY

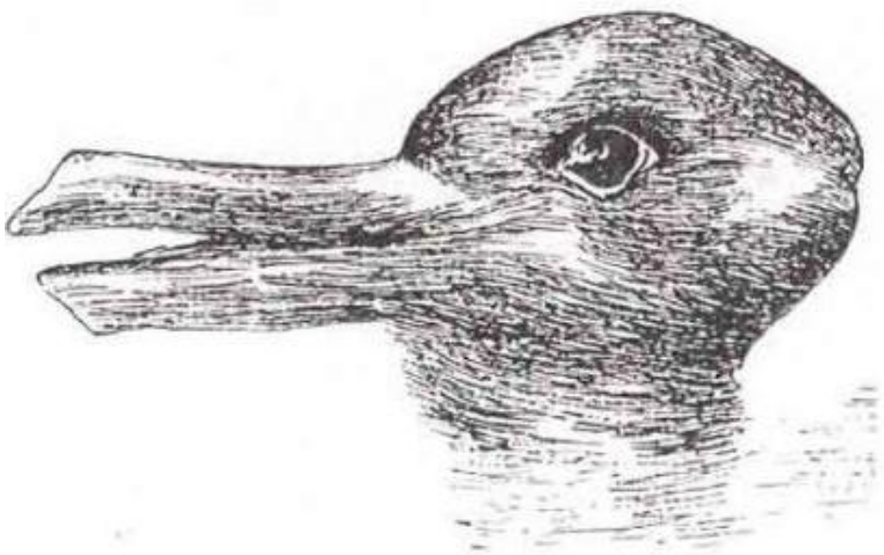
Adopting a new Paradigm

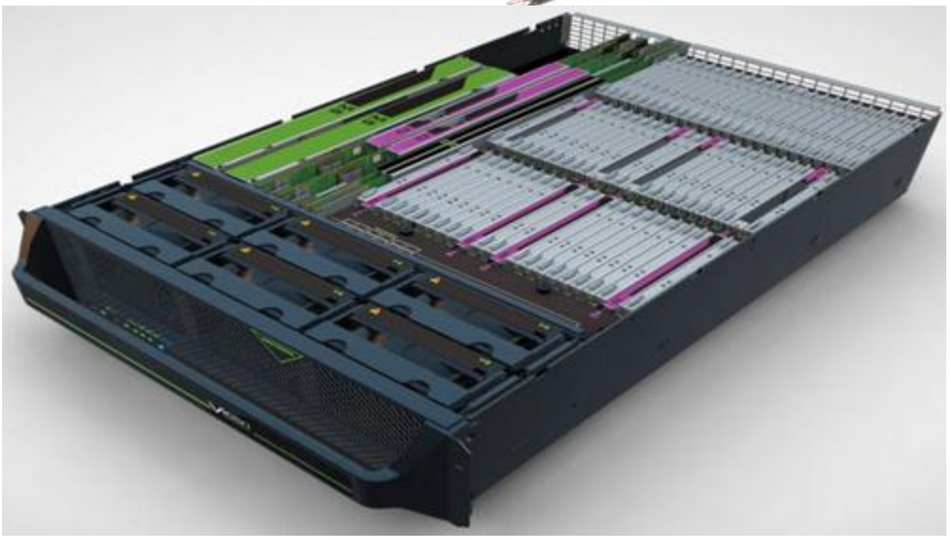
Speed of electron in nth orbit is

$$v_n = \left(\frac{e^2}{2\,\epsilon_0 h}\right)\frac{Z}{n}$$









Taking the Governor off Auto DOP within Oracle

```
SQL> desc resource_io_calibrate$;
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------
 START_TIME                                         TIMESTAMP(6)
 END_TIME                                           TIMESTAMP(6)
 MAX_IOPS                                           NUMBER
 MAX_MBPS                                           NUMBER
 MAX_PMBPS                                          NUMBER
 LATENCY                                            NUMBER
 NUM_DISKS                                          NUMBER
```

DBMS_RESOURCE_MANAGER.CALIBRATE_IO

SELECT * FROM resource_io_calibrate$;

MAX_IOPS	MAX_MBPS	MAX_PMBPS	LATENCY	NUM_PHYSICAL_DISKS
77597	1250	767	2	22

Per Process maximum throughput in MB/sec

The larger the number, the less the CBO wants to parallelize.

Oracle Installation Differences to be Aware of

IOUG independent oracle users group 20th ANNIVERSARY

11gR2 ASM 4k sector disk groups

```
CREATE DISKGROUP DATA
EXTERNAL REDUNDANCY
DISK '/dev/mapper/violin_lun1','/dev/mapper/violin_lun2'
ATTRIBUTE 'au_size'='64M',
'sector_size'='4096',
'compatible.asm' = '11.2',
'compatible.rdbms' = '11.2';

Note: au_size should be set for your environment
```

```
> select group_number, name, sector_size, allocation_unit_size as au_size,
type, total_mb from v$asm_diskgroup;

GROUP_NUMBER NAME     SECTOR_SIZE  AU_SIZE  TYPE   TOTAL_MB
------------ -------- ----------- ---------- ------ ---------
           1 DATA     4096        67108864 EXTERN   831488
           2 RECO     4096         1048576 EXTERN   102400
```

11gR2 RDBMS Online Redo Logs with 4k blocksize files

```
Create controlfile reuse set database "vmem1"
MAXINSTANCES 8
MAXLOGHISTORY 1
MAXLOGFILES 16
MAXLOGMEMBERS 3
MAXDATAFILES 100
Datafile
'&&file0',
'&&file1',
'&&file2',
'&&file3'
LOGFILE GROUP 1 SIZE 300M BLOCKSIZE 4k,
GROUP 2 SIZE 300M BLOCKSIZE 4k,
GROUP 3 SIZE 300M BLOCKSIZE 4k,
GROUP 4 SIZE 300M BLOCKSIZE 4k,
GROUP 5 SIZE 300M BLOCKSIZE 4k,
GROUP 6 SIZE 300M BLOCKSIZE 4k,
GROUP 7 SIZE 300M BLOCKSIZE 4k,
GROUP 8 SIZE 300M BLOCKSIZE 4k RESETLOGS;
```

The POC

Keep it REAL !!!

Change Your Paradigm

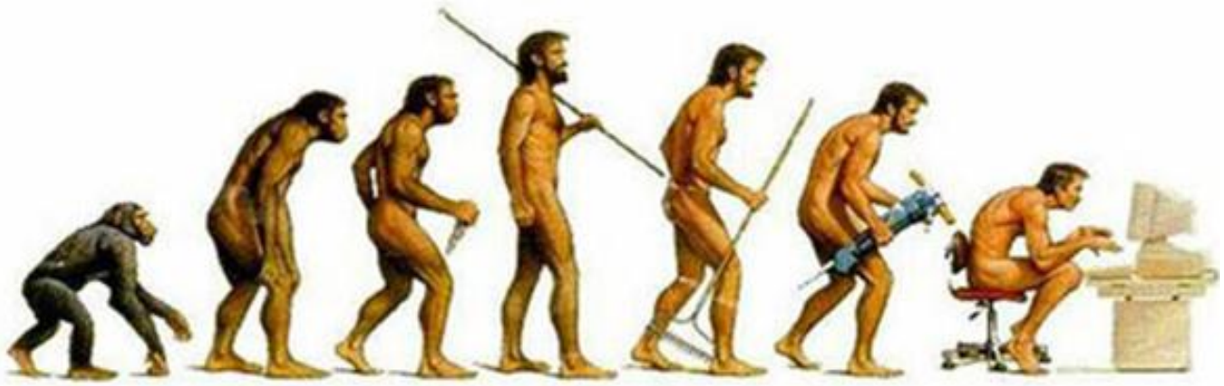
Fast Storage does not always fix BAD CODE

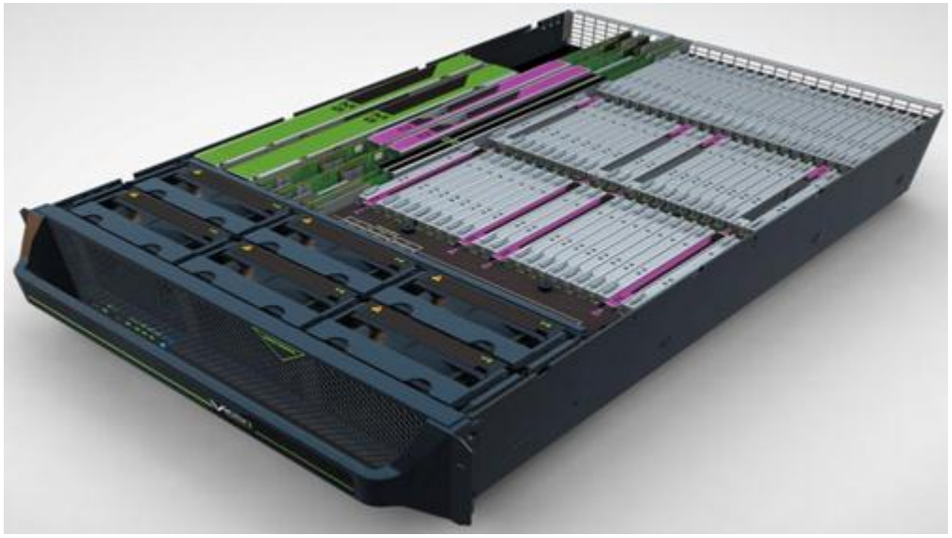
Fast Storage can Amplify Bad Architecture

IOUG independent oracle users group 20th ANNIVERSARY

COLLABORATE 13
TECHNOLOGY AND APPLICATIONS FORUM FOR THE ORACLE COMMUNITY

Evolution

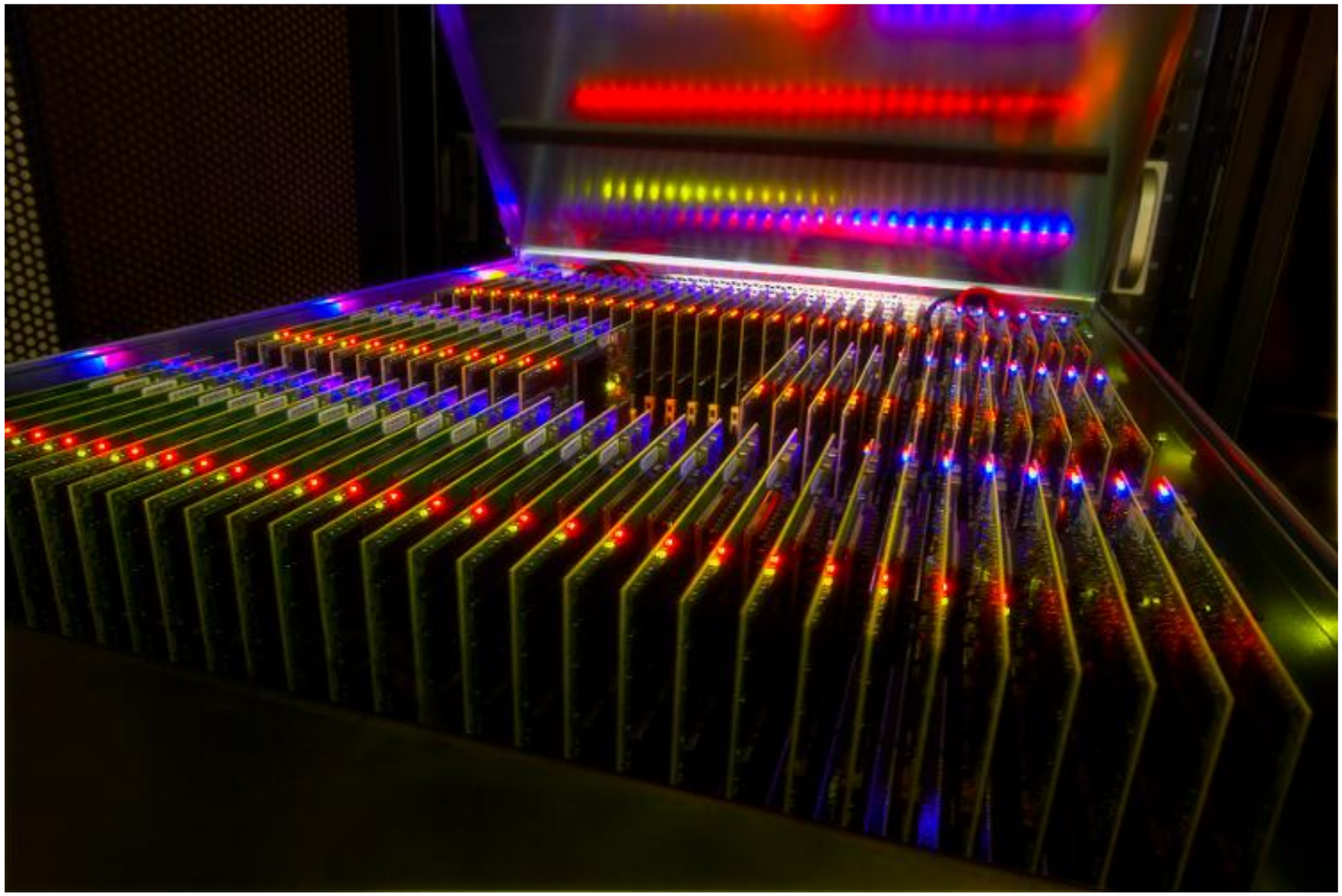




Questions?
Comments?

THANK YOU

rcook@vmem.com



IOUG independent oracle users group 20th ANNIVERSARY

COLLABORATE 13
TECHNOLOGY AND APPLICATIONS FORUM FOR THE ORACLE COMMUNITY