Bart Sjerps
Principal Systems Engineer
Oracle SME - EMEA
bart.sjerps@emc.com
+31-6-27058830
Blog: http://bartsjerps.wordpress.com

ORACLE INTEGRATION ON EMC

INNOVATIONS & BEST PRACTICES

EMC ADDED VALUE FOR ORACLE

INVESTMENT FOR CUSTOMERS

1995–Present
80,000+ joint customers

- **Basic infrastructure advantage**
  Availability, *Predictable* Performance & Scalability, Data Integrity, Freedom of connectivity & disk type)
- **ILM for Oracle & FAST**
  Reducing TCO, allowing growth, eliminating performance, manageability and scalability issues
- **Application (Database) cloning**
  For backup, firefighting, test/dev/acceptance refresh, DWH loading, app or DB upgrades
- **Business Continuity**
  D/R replication, Backup/Restore with dedupe, business landscape consistency
- **Database & App Virtualization**
  Reducing license & HW cost, improving flexibility, enabling the cloud
- **Security**
  Database & storage encryption, key management, Data Leakage Prevention (DLP)
- **Joint solutions to shorten time to value**
  Whitepapers and reference architectures, assessments, Design & Deployment Services
- **Joint support services to reduce risk**
  Joint Service Center

ORACLE Platinum Partner

EMC²

CUSTOMER CHALLENGES: PERFORMANCE

- Still an issue after 40+ years of Moore's law
- Database sizes still grow
- As do workloads
- Applications don't always behave
- "Big Data" workloads

WHAT IF…

YOUR DATABASE STORAGE WOULD PROVIDE:

- Consistent, very low, predictable response times
- High Bandwidth
- No need for:
  - Striping
  - Tiering
  - Many “spindles”
  - Separation of data types
- Equal performance for prod & non-prod
- Zero-overhead:
  - Cloning (DB copies)
  - Compression
  - De-duplication
  - Encryption
  - Thin provisioning

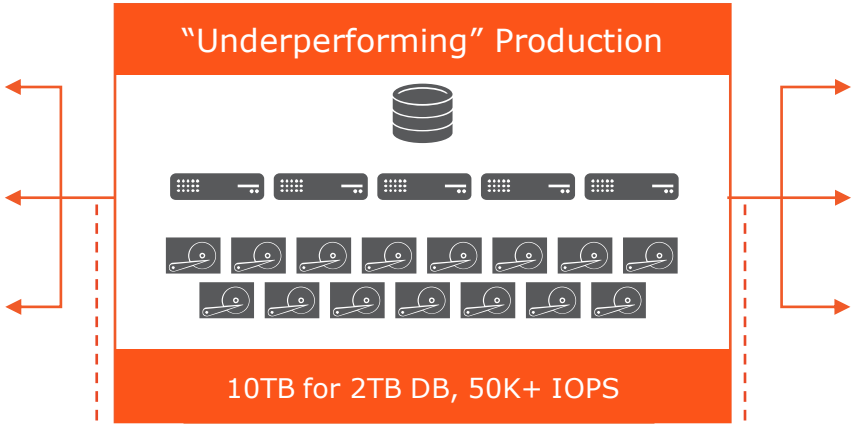
WHAT IS THE ONE THING YOU KNOW ABOUT THIS CAR?

Did you say ELECTRIC?

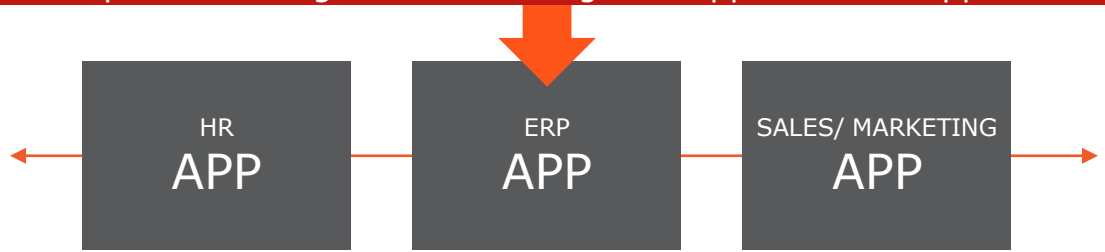
IT'S A BETTER CAR!

- Only one moving part
- No oil changes
- 0-60 in 4.4 sec
- Lowest center of gravity
- Silent
- More interior space
- No emissions
- Five star crash ratings
- Tax Rebates
- Access to carpool lanes

TRADITIONAL DATABASE INFRASTRUCTURE - COMPLEX

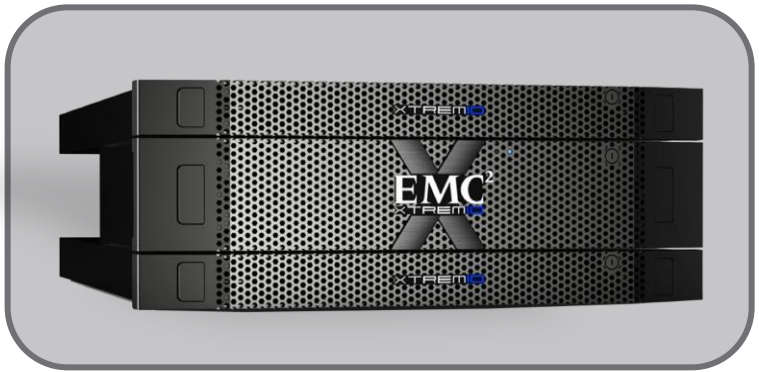


14+ full copies to manage 36TB of storage to support 2TB of application data



Application Siloes to manage hotspots

XtremIO Flash Array



One X-Brick
10TB(7.47TB[4]) /20TB (14.94TB[4]),
150K[1] IOPS

- Consistent performance
  - 150K[1] IOPS per x-brick
  - Predictable low latency
- Data Management
  - Unique Content-aware Data Protection
  - Deduplication[2] & Compression[3]
  - Space-efficient Snaphosts/clones
  - Encryption
- Simple to manage and scale
  - 3-step provisioning—no tuning
  - Scaleout architecture

[1] 8K 100% random read
[2] Works on complete copies of database
[3] Upcoming release
[4] Usable capacity without data reduction

XtremIO Scaleout Array

One Array

Single X-Brick Array
10/40TB[2], 150K IOPS[1]

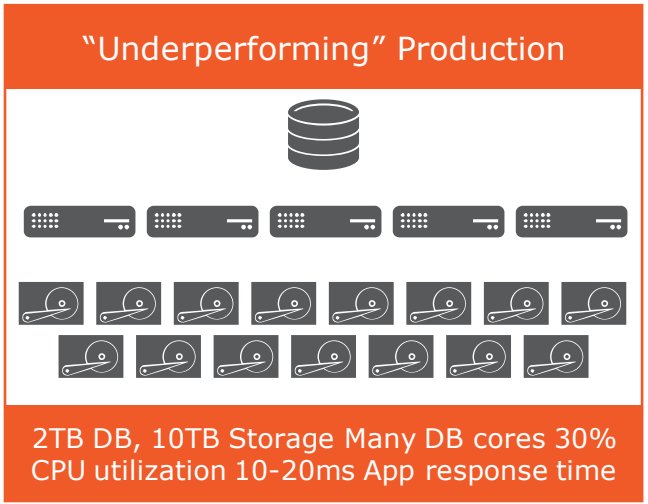
Two X-Brick Array
20/40TB[2], 300K IOPS[1]

Four X-Brick Array
40/80TB[2], 600K IOPS[1]

[1] 8K 100% random read
[2] Raw capacities

- Linear scaling
  - Add additional X-Bricks for capacity and performance
  - Capacity and performance scale linearly
- Single system
  - Managed as one array
  - Data management and reduction
- Naturally balanced
  - Uniform LUN layout
  - Load balanced IOs

WHY DON'T APPLICATIONS PERFORM?



- Application databases produce variety of IOs
  - Constant high load sustained & random/sequential I/O
  - Huge user concurrency leads to random workload
- Datasets are growing and becoming more random
  - Keeping up with data hotspots has become primary function
  - Disk/RAID types, caching, tiering can go only so far

CHASING HOTSPOTS

Table 6 Aggregate Layout

Controller	Aggregate Name	Option /RG Size	#	
FAS-3170_A_1	aggr0	RAID-DP, RG-16		
FAS-3170_A_1	AGGR_ORA_1_A_1	RAID-DP, RG-21		Database and Cluster Ware Binary
FAS-3170_A_2	aggr0	RAID-DP,	3 no's/350GB	DOT and root volume
FAS-3170_A_2			/13TB	Da
FAS-3170_B_1			s/350GB	Do
FAS-3170_B_1	AGGR_ORA_1_B_1	RAID-DP, RG-21	42 no's/13TB	Data , Redo logs, control files
FAS-3170_B_2	aggr0	RAID-DP, RG-16	3 no's/350GB	DOT and root volume
FAS-3170_B_2	AGGR_ORA_2_B_1	RAID-DP, RG-21	42 no's/13TB	Data files, Redo logs, control files, Archive Log

1. Different RAID groups

2. Different controllers

3. Different drive types

Table 7 Volume Layout

Controller	Volume name	Aggregate Name	Size	Purpose
FAS-3170_A_1	ORA_HOME	AGGR_ORA_1_A_1	100GB	Database Binary
FAS-3170_A_1	CRS_HOME	AGGR_ORA_1_A_1		
FAS-3170_A_1	OCR_CSS	AGGR_ORA_1_A_1		
FAS-3170_A_1	VOL_DATA_1_A_1	AGGR_ORA_1_A_1	12TB	Datafiles, control file
FAS-3170_A_2	V			log Files
FAS-3170_A_2	VOL_FRA_2_A_1	A_2_A_1		FRA
FAS-3170_B_1	VOL_DATA_1_B_1	AGGR_ORA_1_B_1	12TB	Datafiles, Control file
FAS-3170_B_2	VOL_DATA_2_B_1	AGGR_ORA_2_B_1	10TB	Datafiles, Control File

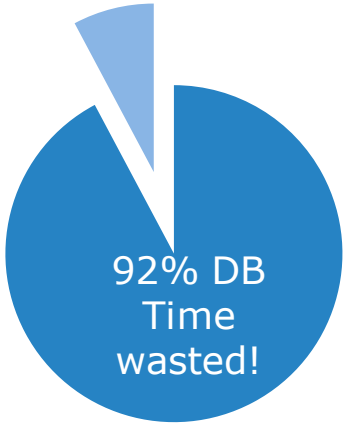
4. Different volumes

5. Different sizes

Do you have application performance issue?

A TYPICAL I/O-BOUND SERVER. WASTED CPU CYCLES

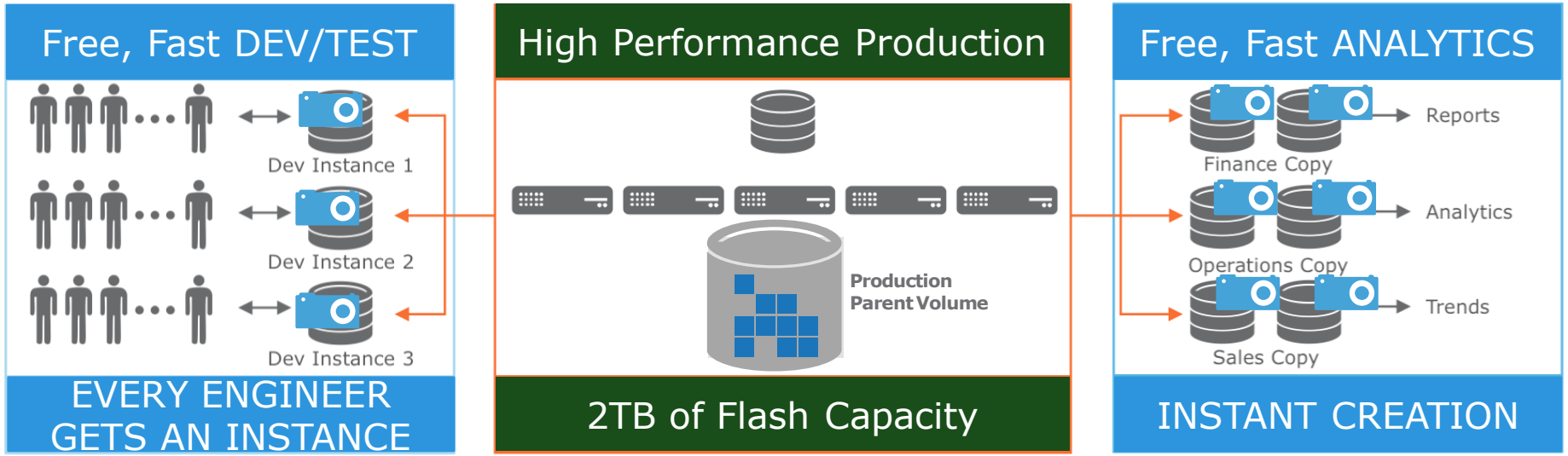
4 hours wasted by high-latency single block random reads



Top 5 Timed Foreground Events

Event	Waits	Time(s)	Avg wait (ms)	% DB time	Wait Class
db file sequential read	1,420,110	14,362	10	92.7	User I/O
Tree buffer waits	89,072	986	11	6.4	Configurat
DB CPU		158		1.0	
library cache lock	138	51	371	.3	Concurrenc
write complete waits	9	17	1919	.1	Configurat

Transform Database Infrastructure



x10 Dev/Test Instances

02 TB total XtremIO space

0 Full Copies

SCALE-OUT IOPS IN ABUNDANCE

RESULTS-AWR REPORT

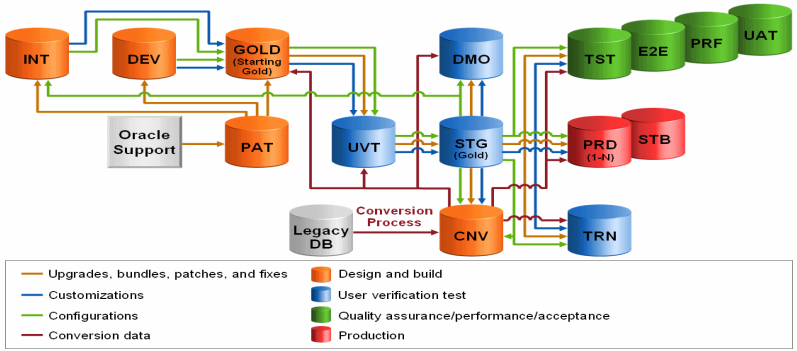
Event	Waits	%Time -outs	Total Wait Time (s)	Avg wait (ms)	Waits /txn	% DB time
direct path read	361,425	0	930	3	7,376.0	16.4
db file parallel read	365,088	0	898	2	7,450.8	15.9
db file scattered read	64,628	0	114	2	1,318.9	2.0
db file sequential read	56,748	0	84	1	1,158.1	1.5
gc cr multi block request	137,524	0	62	0	2,806.6	1.1
direct path read temp	19,942	0	44	2	407.0	.8
read by other session	17,389	0	31	2	354.9	.6

Event	Waits	Time(s)	Avg wait (ms)	% DB time	Wait Class
DB CPU		3,427		60.5	
direct path read	361,425	930	3	16.4	User I/O
db file parallel read	365,088	898	2	15.9	User I/O
db file scattered read	64,628	114	2	2.0	User I/O
db file sequential read	56,748	84	1	1.5	User I/O

- Avg. latency went down by many folds
- CPU utilization doubled

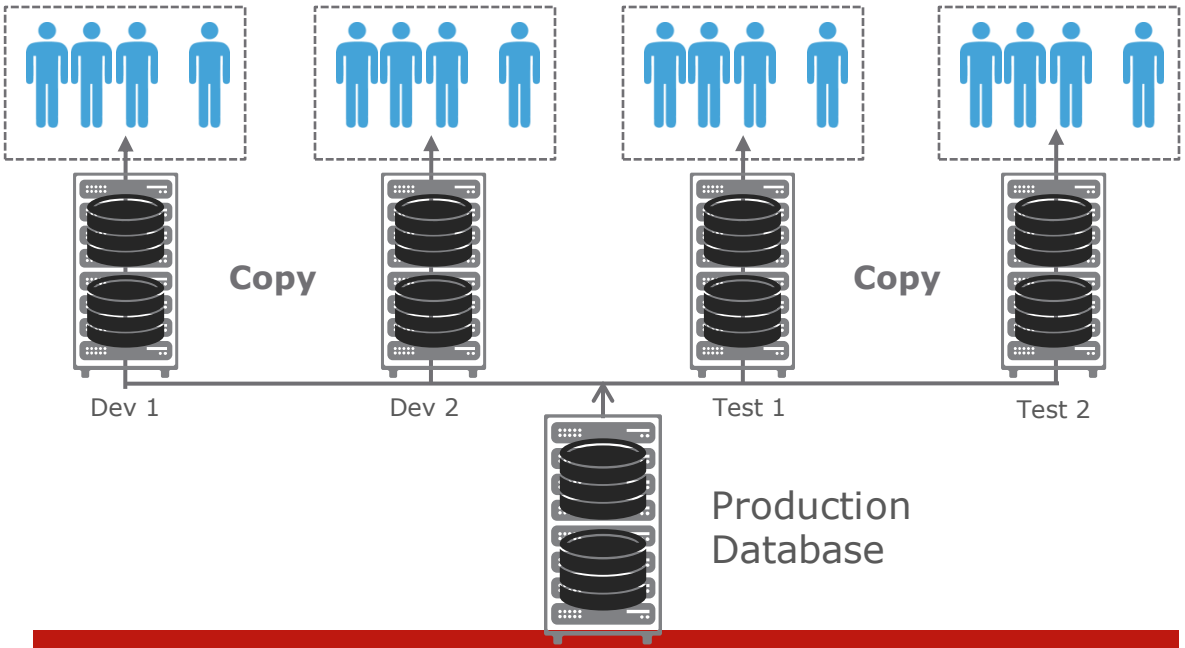
Why Database Cloning?

- "Serverless backups" & Quick restores
  - Zero production performance impact
  - RTO seconds to minutes
  - Protected & Instant restores
  - Out-of-order restores
  - Instant Restore from remote copies
- Firefighting
  - Creating a quick production copy to solve application problems
  - Without messing with production data
- Creating Test / Dev / Acceptance copies
  - Automated, no tape restores, low people effort
- Creating copies for reporting / staging
  - Datawarehouse queries can bring production performance down
  - Moving reporting workload to copy relieves production
- Application / Database Upgrades
  - Creating application "checkpoints" avoids having to fall back to starting point due to small errors
  - Easy upgrade testing



TEST & DEV - COMPLEX, RESTRICTIVE AND EXPENSIVE

- Full copies of production on separate re-purposed storage
- Copy process is long

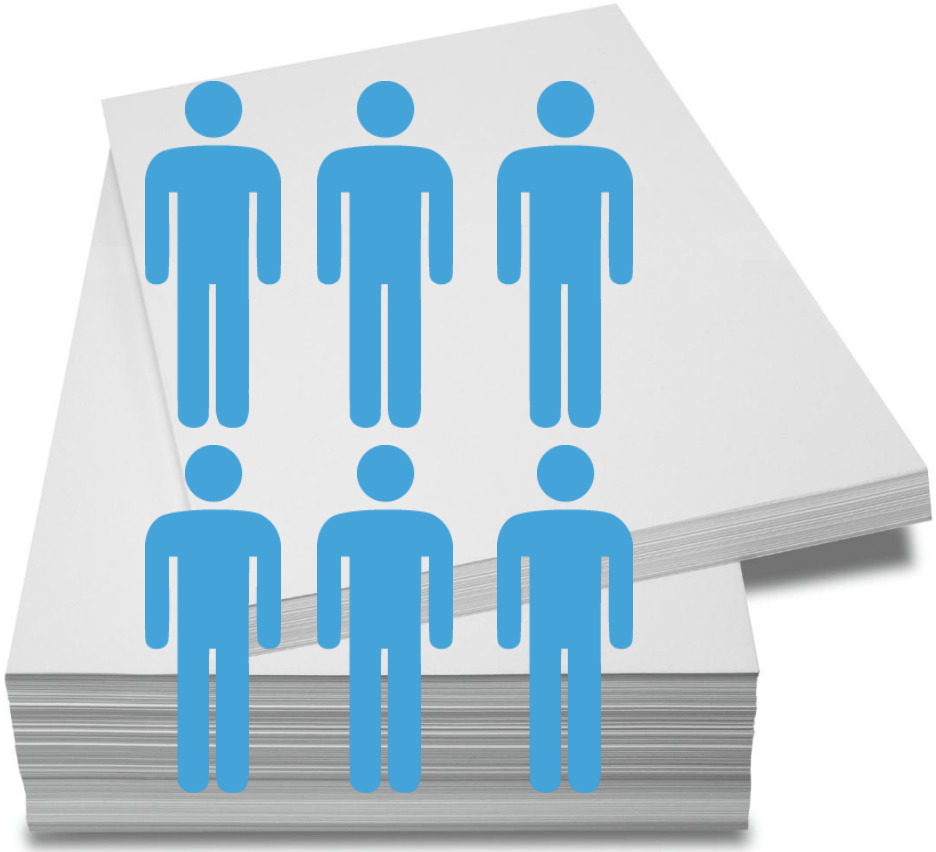


Group of engineers share the copies

Delays & compromised quality

"How many copies of DB do you keep for test and development?"

"Not as many as we like"



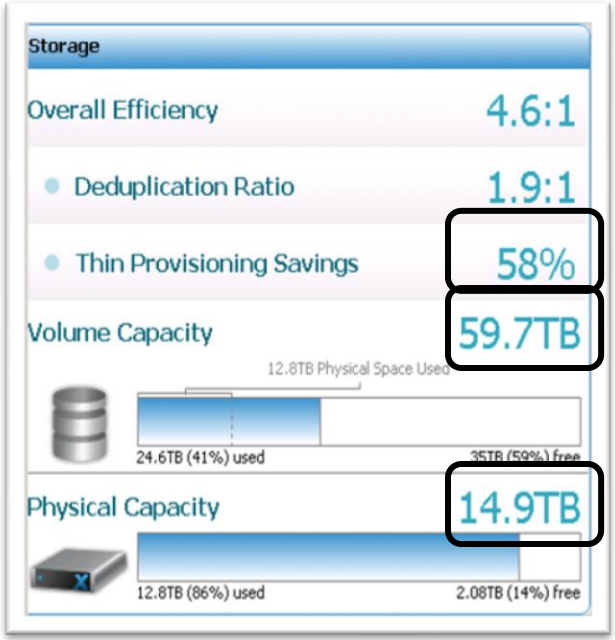
"If cost and complexity were not an issue, how many copies of your DB would you keep?"

"One for each of my engineers"

SPACE EFFICIENT SNAPSHOTS

1 OLTP Database LUN*

Logs/database on 1 LUN

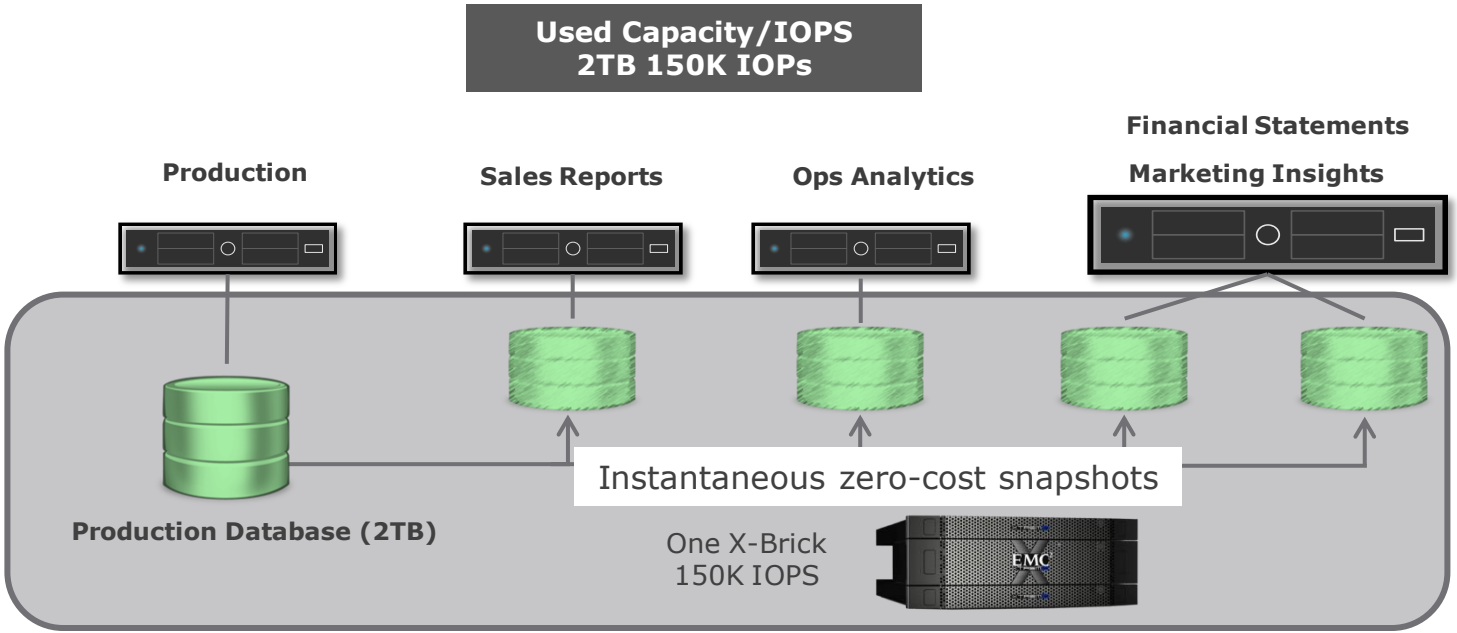


1 OLTP LUN and 6 snapshots*

1 LUN and 6 snapshots

* Internal tests

ZERO COST DATA MART COPIES ANYTIME!

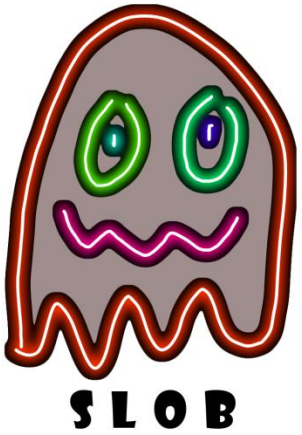


- Create data mart copies when needed
- Consolidate DW infrastructure

SILLY LITTLE ORACLE BENCHMARK

I/O PROFILING FOR ORACLE – TEST RESULTS

Info & download: XtremIO.com/slob

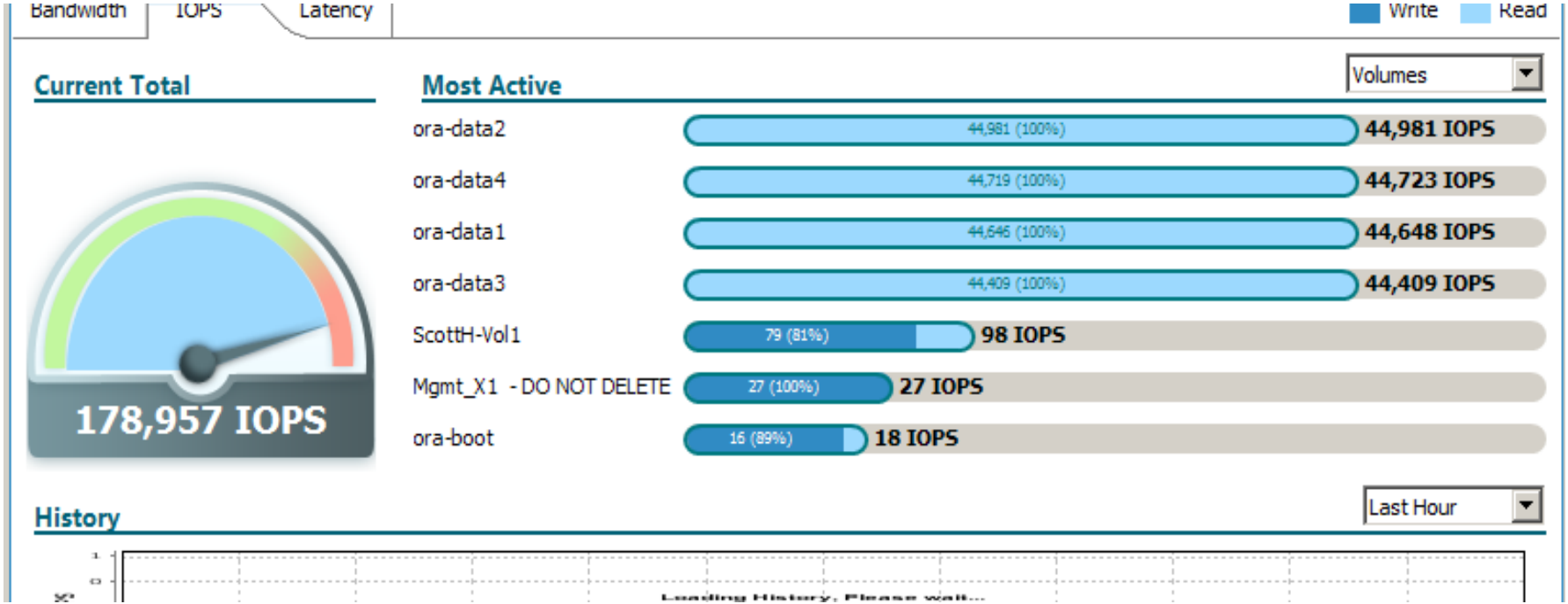


XtremIO

ORACLE DATABASE

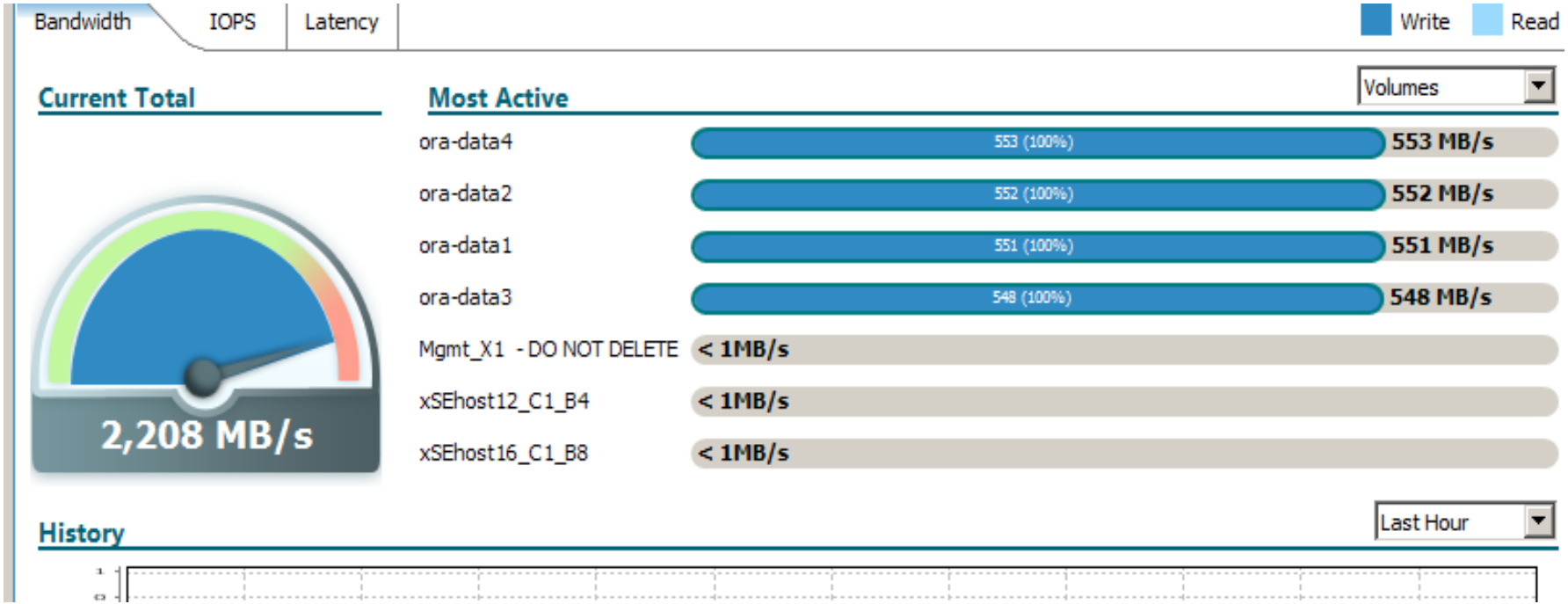
PERFORMANCE EXAMPLE WITH SLOB

LAB TEST: 1 V2.4 X-BRICK, 3 VM'S ORACLE 11.2.0.4.0, VMDKS



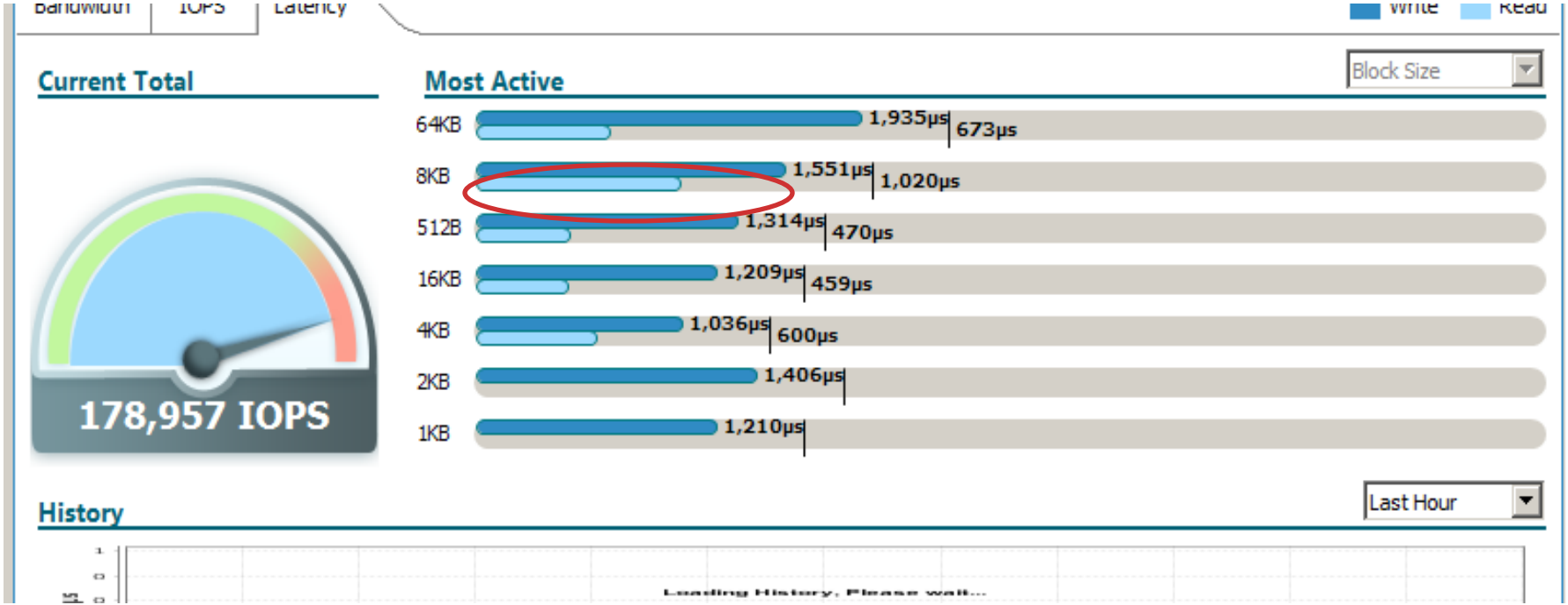
PERFORMANCE EXAMPLE WITH SLOB

BANDWIDTH



PERFORMANCE EXAMPLE WITH SLOB

LATENCY (NEARLY ALL I/O IS 8K RANDOM)




Keep Calm
and run
your
Oracle
workload
on
XtremIO

BACKUP OPTIMIZATION FOR ORACLE

Real world (extreme) example:

- Customer with several 1-2 Terabyte databases (tier 1 production)
- SLA demands 2x full backup / day, Retention: 1 Month

Tape capacity required for a 2 TB database: 2 TB * 2 copies / day * 31 days = ***124 TB tape***

Considerations:

- How much savings would be achieved when reducing DB by 20% ?
- What if we could store only 1 full copy plus 61 delta sets?
- What if I need 6 months retention?
- How fast can we recover from backup using tape?
- Is it reliable?
- What's the performance impact on production?
- What's the backup window?

The best thing about being me… There are so many "me"s

–Agent Smith, The Matrix Reloaded.

BACKUP OPTIMIZATION FOR ORACLE

- Don't backup to Tier-1 storage ("FRA")
  - Expensive, no real benefit
- Deduplicate all backups
  - Allows you to do a virtual full daily backup
  - Not all backup appliances perform well when deduplication is turned on
  - EMC Data Domain performs deduplication ***INLINE***
- Snapshot backups
  - Allow to remove 100% of the performance overhead
  - Allow you to do FULL restores in minutes
  - Make snapshot before doing full restore?
- Replication of backup data?
  - For DR purposes
  - Archiving?

- Always consider ***RESTORES***
  - RTO/RPO?
  - Reliability?
  - What if the restored backup is bad?
  - How quick can you get an older copy back online?
  - Surgical repair (single file) or full restore?
  - Can you still mount the broken database to salvage data or do root cause analysis?
- Offload the deduplication process (EMC DDBoost)
  - Speeds up backups by >50%
  - Reduce I/O & network overhead by >80%
  - Integrated with Oracle RMAN

REMOTE REPLICATION & DISASTER RECOVERY

DISASTER RECOVERY FOR ORACLE

(AND THE REST OF THE DATA)

*Nothing travels faster than the speed of light, with the possible exception of **bad news**, which obeys its own special laws*
- Hitchhiker's guide to the galaxy

- How to protect the business against disasters?
  - Physical? Logical? Large? Small?
  - Human error (accident, bug) or deliberate (virus, hacker)?
  - What about rolling disasters? Can a problem affect the DR site?
- What do you need to recover?
  - Just the database? Middleware? App servers? File servers? Domain controllers? Email? Document management? Backup environment?
  - Are transactions cross-related?
  - How fast? (RTO, including decision point, procedures)
  - How accurate? (how much dataloss, RPO)
  - How far away? ("RDO") - More than 2 sites?
- Different replication tool for each different platform / app?
  - Each with many instances?
  - Dependent on host resources? Overhead?
  - Each one sensitive to config errors, bugs, silent session failures? Everything monitored?
- ***Are you sure everything works after failover?***
  - Did you test large scale or just one application?
  - Frequently? Or just after go-live?
  - Can you test during production? Without interrupting replication?

EMC has 15++ years history in solving these issues:
One single, reliable, consolidated approach for all apps

LET'S JUST USE DATA GUARD

AS THIS IS THE ONLY METHOD WE KNOW… WAIT A SEC, WHAT ABOUT:

- Non-database data?
  - Or is that just SEP (Somebody Else's Problem)?
- Non-Oracle databases?
  - SEP too?
- Multi-database business consistency?
  - Huh what?
- Using the D/R copy for testing, reporting…
  - Active Data Guard? Requires license and no read/write, just read-only
- "Force logging" mode overhead?
  - i.e. /NOLOG transactions, direct path loads, DB reorg, rebuild index etc…
  - "Never tested that during POC"…
- What if we need to rebuild the standby?
  - Don't tell me that does not happen…

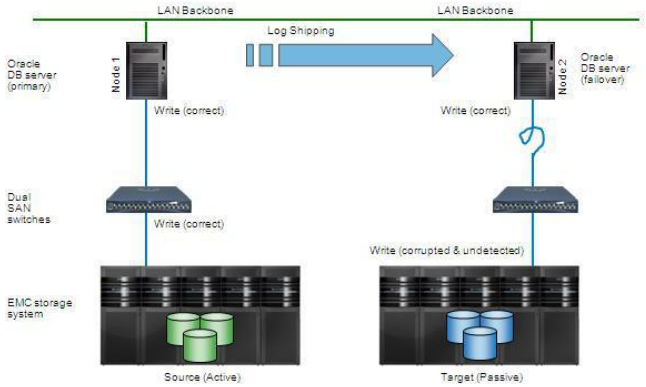
Data Guard:
Classical case of (Oracle) worshipping?

Be wise and make a rational , independent choice!

BUT….

ORACLE CLAIMS WE NEED “DATABASE AWARE” REPLICATION?

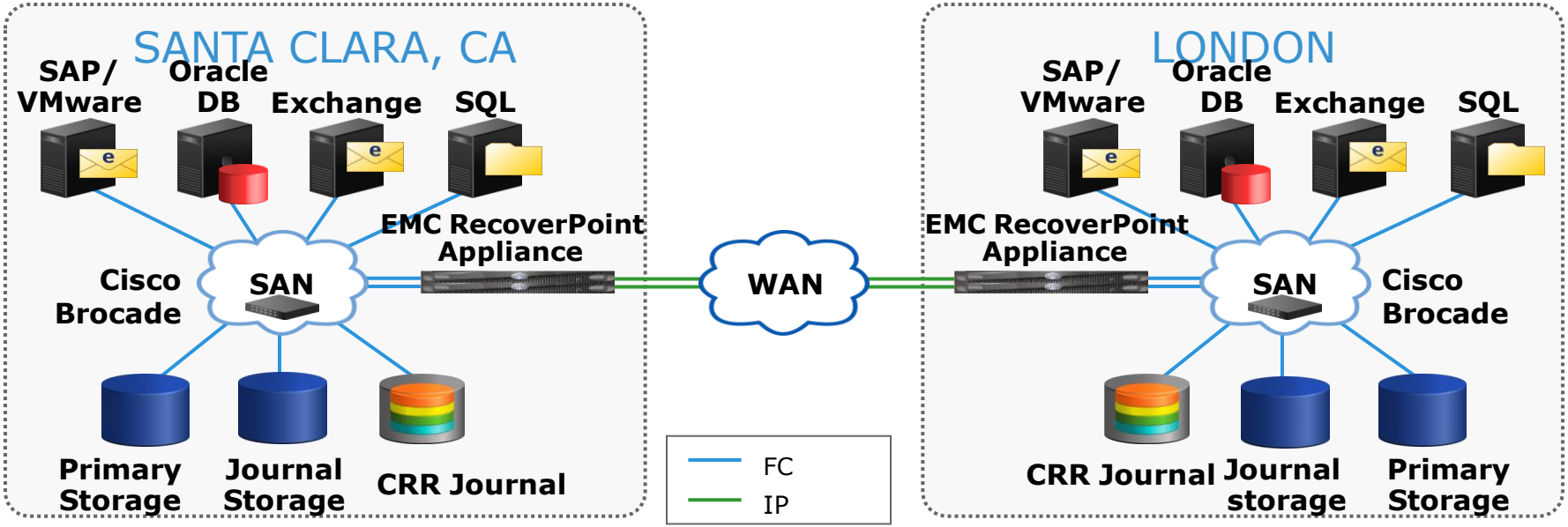
- Really? Since when?
- EMC replication has protected Oracle databases for 20+ years
  - Rock solid, always works, fire & forget
  - Can read from remote storage without failover
  - Provides READ/WRITE remote copies (but you need to refresh them every now and then… there’s no free lunch)
- SAN replication is GIGO (Garbage In, Garbage Out)
  - So yes, local corruption gets transferred to DR
  - IMHO, data integrity protection should not depend on D/R infra
  - But should be a fundamental feature of the infrastructure
  - Even if you don’t have DR replication but only rely on backups
- And…
  - Data Guard does not protect from REMOTE corruptions
  - You win some, you lose some



More info: [Data Guard protecting from EMC block corruptions?]

EMC CONTINUOUS DATA PROTECTION

AND ROLLBACK FOR NON-VIRTUAL OR VIRTUALIZED ORACLE ENVIRONMENTS



KEY TECHNOLOGY HIGHLIGHTS

Applications
- SQL and Exchange
- Oracle, SAP
- Middleware, files, etc

CDP and CRR
- Continuous Data Protection
- Continuous Remote Replication

Protects against:
- Logical corruptions
- Physical corruptions
- Site failures

Provides:
- Failover
- PIT, INSTANT restores
- Point-in-time copies (snaps)

D/R features compared

DR feature	Data Guard	EMC SRDF	EMC Recoverpoint
Method	Log shipping	Data mirroring	"Continuous" Data replication
Operation	Sync / Async	Sync / Async	Sync / Async
Requires "force logging" (performance), archivelog mode, db configuration changes	Yes	No	No
Requires remote DB server + license	Yes	No (*)	No (*)
Uses DB host CPU & network resources	Yes	No	No
Replicates non-Database data	No	Yes	Yes
Requires one instance per each database	Yes	No	No
Multi DB/Platform/App consistency groups	No	Yes	Yes
Failover time	Seconds	Minutes	Minutes
Standby DB access	Read-only	Read/write via snaps	Read/write via snaps
Allows instant restore to older checkpoint	No	Yes (via snapshots)	Yes (time shift)
Requires regular refreshes of remote snapshot	No	Yes	Yes
Incremental re-sync after link fail	Only if data was not modified at both locations	Always	Only when not modified
Transfers local data I/O corruption	No (does not transfer datafiles – only logs)	Yes (works as designed)	Yes (as designed)
Prevents remote data corruption	No (remote server can silently corrupt data)	Yes (no remote server involved)	Yes
Continues if local RAID group fails (i.e. serve I/O from remote system)	No (failover is triggered)	Yes (it's a remote "mirror" not a "copy")	No

*) Beware that Oracle requires "remote storage" to be fully licensed

If active/passive is not good enough…
Achieving True Fault Resilience

ORACLE® RAC

Continuous Availability
Of Oracle RAC Stretched Clusters
Using EMC Storage Virtualization

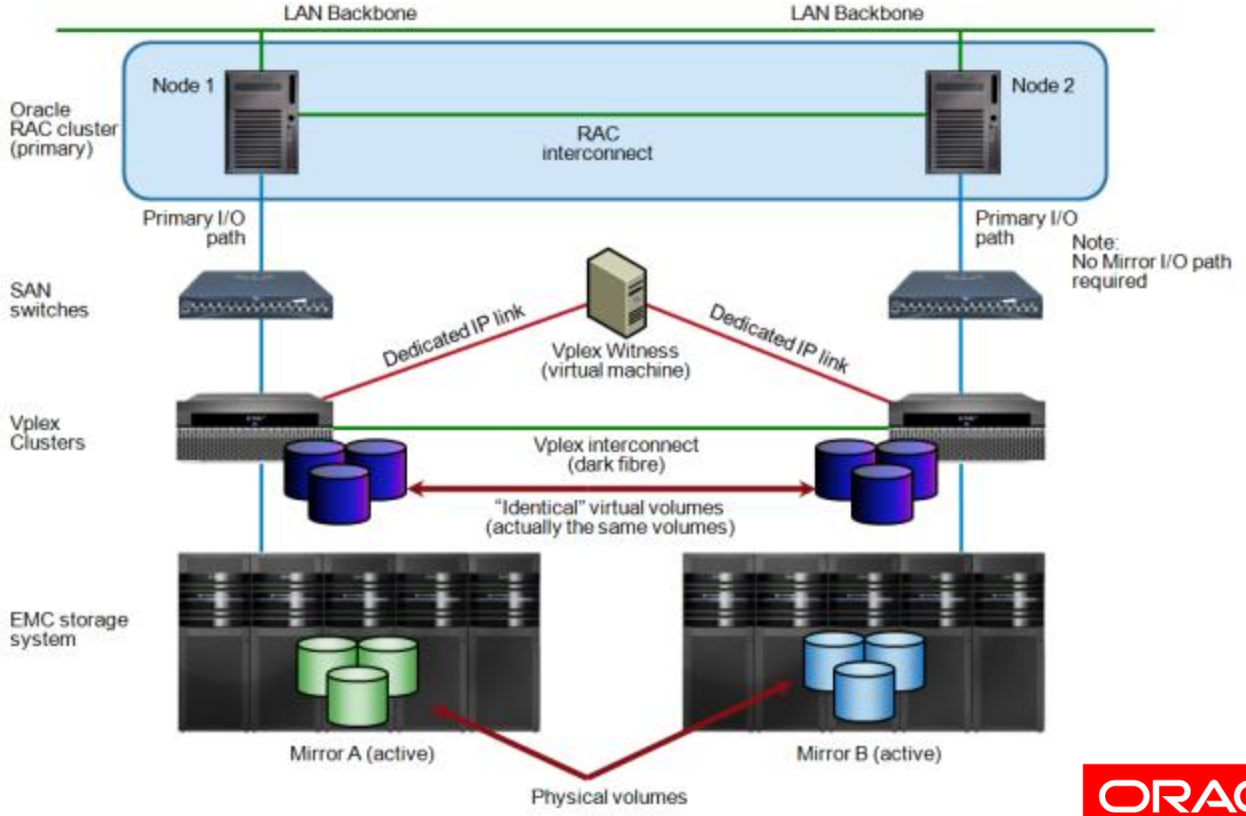
EMC²

ORACLE RAC REQUIREMENTS

- Mission critical business application availability
- Optimal performance
- Online scalability

ORACLE RAC ON EMC VPLEX SOLUTION

EMC & ORACLE CERTIFIED



ORACLE Platinum Partner

WHAT IS (DATABASE) CONSOLIDATION?

- **con·sol·i·date[1]**

- 1: to join together into one whole : unite <consolidate several small school districts>
- 2: to make firm or secure : strengthen <consolidate their hold on first place>
- 3: to form into a compact mass

- It is all about:
- Standardization
- Getting the maximum benefit from as few resources as possible (Lean & Mean)
- Improve service levels (performance, availability, reliability, …)
- Ease of management, deployment, maintenance, control, performance

- So the business goals are: **Lowering cost (TCO) and improving service levels**
- [1]) source: Merriam Webster Dictionary

DATABASE CONSOLIDATION GOALS

1. Maximize use of license investment
2. Maintain or even improve performance
3. Improve High Availability - Avoid (planned and unplanned) downtime
4. Achieve hardware independence - Avoid Vendor lock-in
5. Simplify server & storage refresh cycles
6. Speed up provisioning of new databases
7. Improve security, compliance and auditing
8. Simplify management

CALL TO ACTION

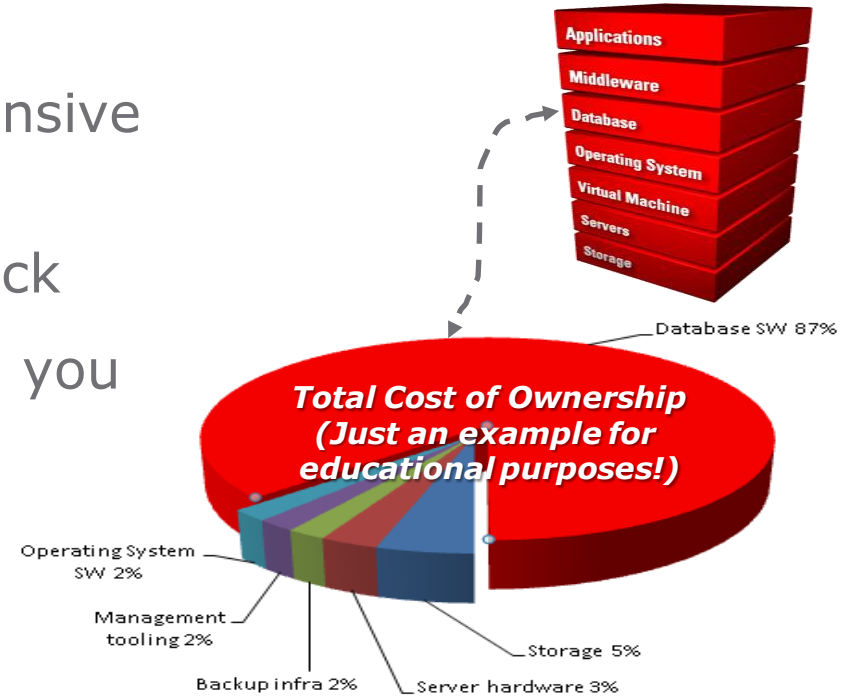
FOLLOW THE MONEY!

Question:
Where do organizations spend most money in the Business Application stack?

Storage? Servers? Networks? Management tooling?

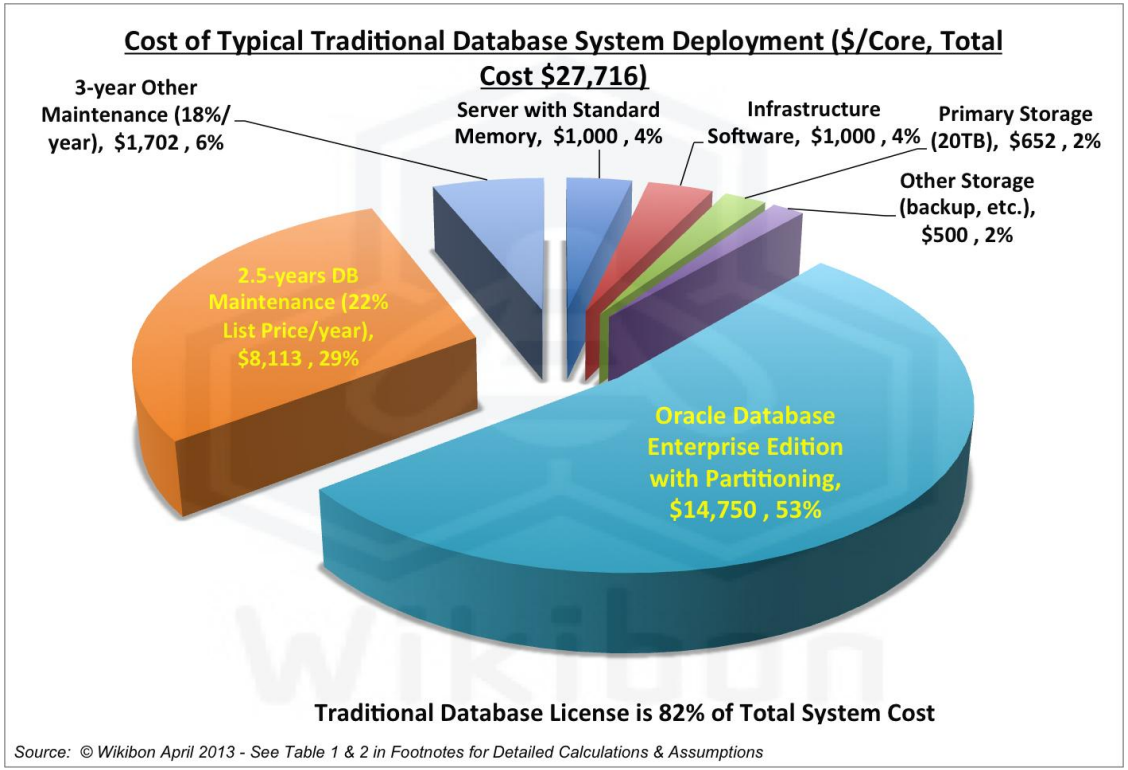
INTRODUCTION - WHY LOOK AT LICENSING?

- Oracle DB licensing is expensive
- Large part of the TCO of a database infrastructure stack
- Save 10% on licensing and you justified more expensive hardware
- Much more is possible



If we can save 10% on db licenses…
We easily justified 50% more expensive infrastructure

WIKIBON ON VIRTUALIZATION



Wikibon Article: [Virtualization of Oracle Evolves to Best Practice for Production Systems]

QUESTION

WHAT IS THE DB LICENSE COST ON A TYPICAL X86 SERVER?

CPU type	Intel x86-64
Sockets	2
Cores per socket	12
Total cores	24
RAM	512GB
DB size	4 TB (usable)
Oracle license type	Enterprise Edition
Additional options	Partitioning, Advanced Compression, Diagnostics pack, Tuning pack
Discount	50%
Maintenance	Ignore for now

$25,000?
$50,000?
$150,000?

What if I add Oracle RAC?

If you want to achieve Oracle cost savings, you have to understand licensing!

SUMMARY

License cost (ex maintenance)	12-core server	24-core server	Exadata X4-2 Full Rack
Enterprise + basic options	$ 241,500	$ 483,000	N/A
Enterprise + basic + RAC + Active Data Guard	$ 340,500	$ 680,000	$5,440,000
Standard Edition (RAC included)	$ 17,500	$ 17,500	N/A
VMware Enterprise Plus (2 sockets)	$ 3,495	$ 3,495	N/A

- Discount 50% (street prices)
- Basic options: Partitioning + Advanced Compression + Diag & tuning pack

- We don't consider user based licensing as many customers will not use this
- Also we don't consider Enterprise License Agreements (but in the end they are similar to CPU licensing
- Oracle can be very creative to win the deal
  - Transfer existing licenses
  - Disable processors on their own hardware (Oracle Database Appliance)
  - Mess around with Oracle VM settings
  - Extra discount if you buy Oracle HW (i.e. Exadata or SPARC)
- Beware of Oracle's license police! (first make sure our customer is compliant)

BEFORE WE START…

BEWARE OF THE LICENSE DEMON

Are you 100% "bet your paycheck" sure that you're compliant?™

ORACLE
LICENSE MANAGEMENT SERVICES

If needed…
Bring in the superheroes
They help you with licensing & legal issues

LC

Licenseconsulting.eu

iQuate
INVENTORY · INNOVATION · INSIGHT

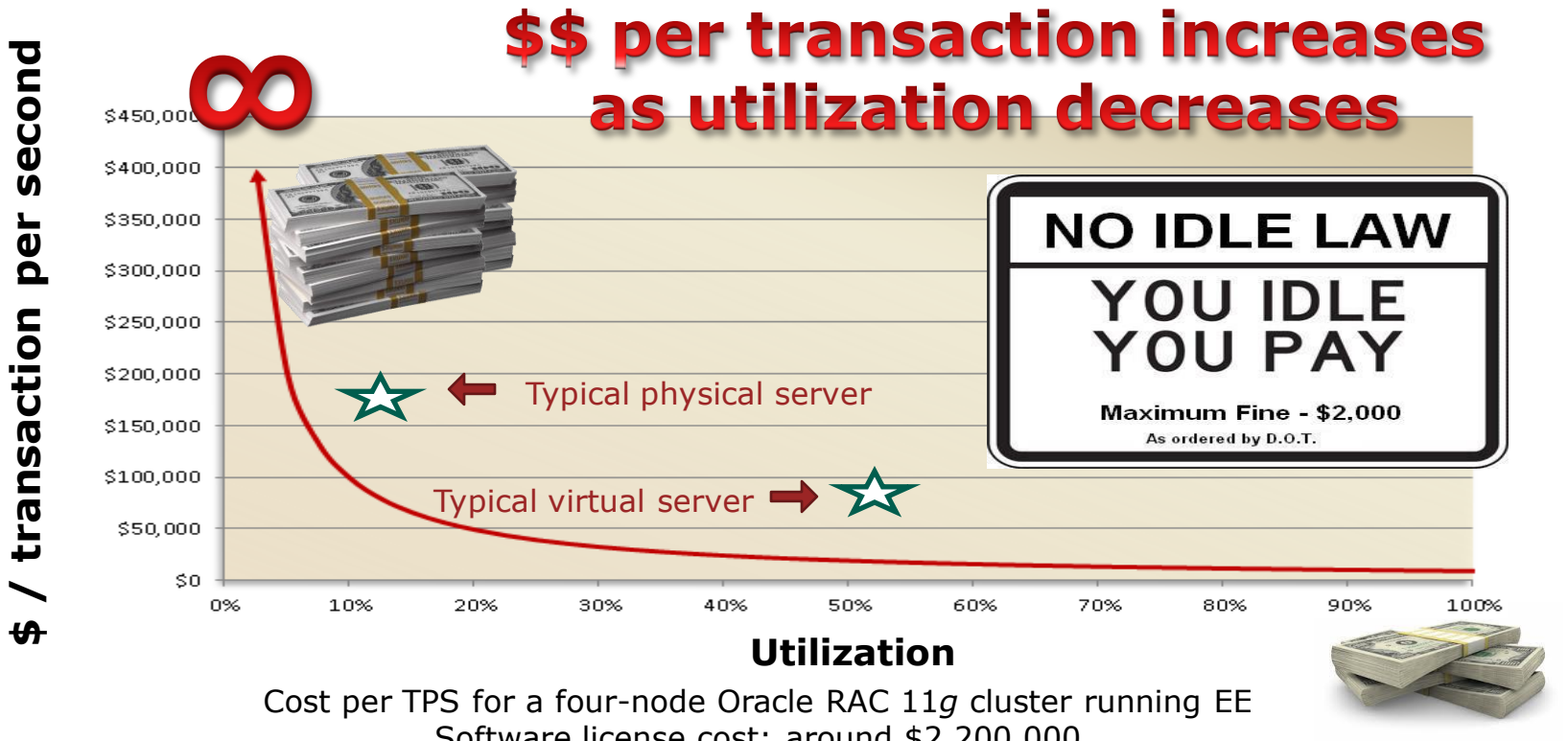
[iQuate iQSonar for Oracle]

Madora Consulting

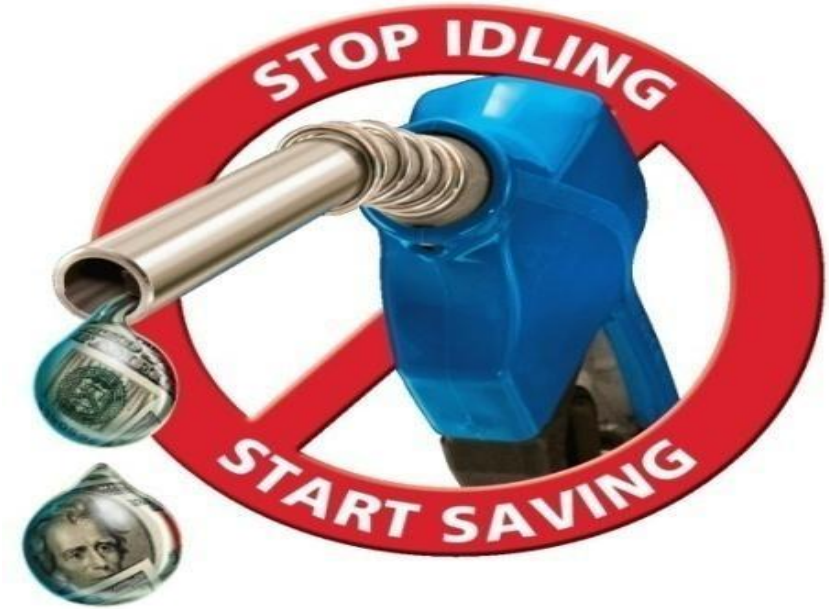
[Madora Consulting UK]

EMC²

TRANSACTION COST VS. UTILIZATION

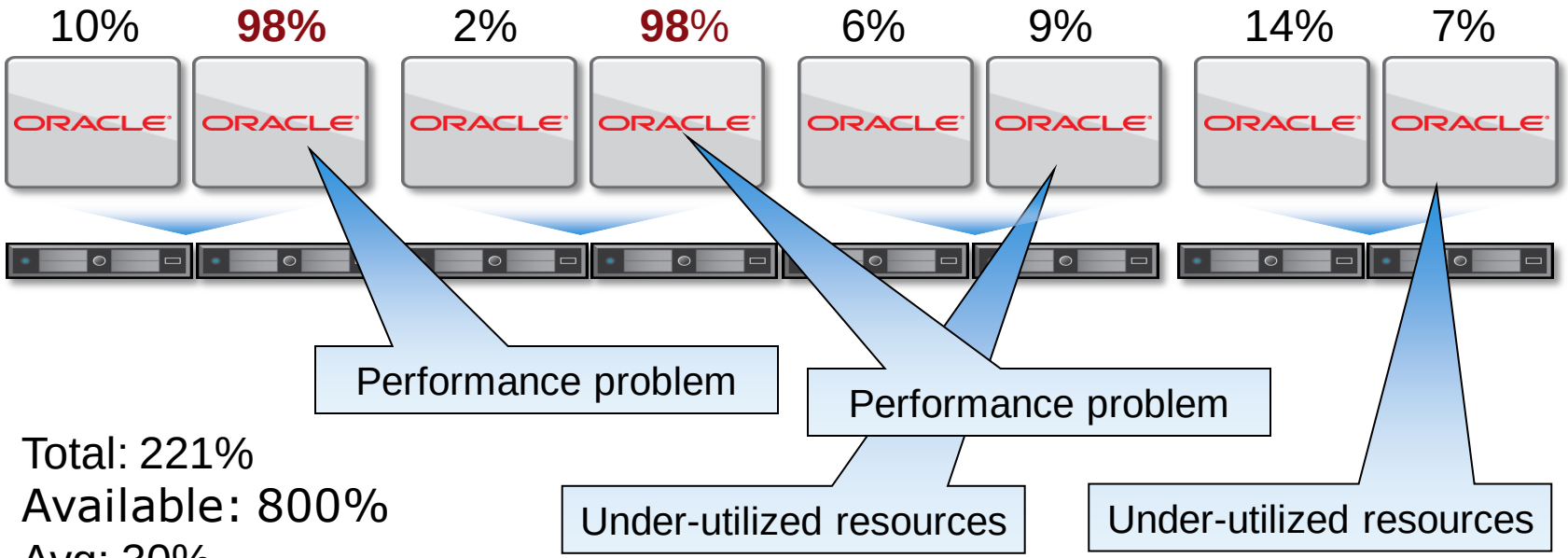


Cost per TPS for a four-node Oracle RAC 11*g* cluster running EE
Software license cost: around $2,200,000
TPS: Around 4,000 at peak utilization



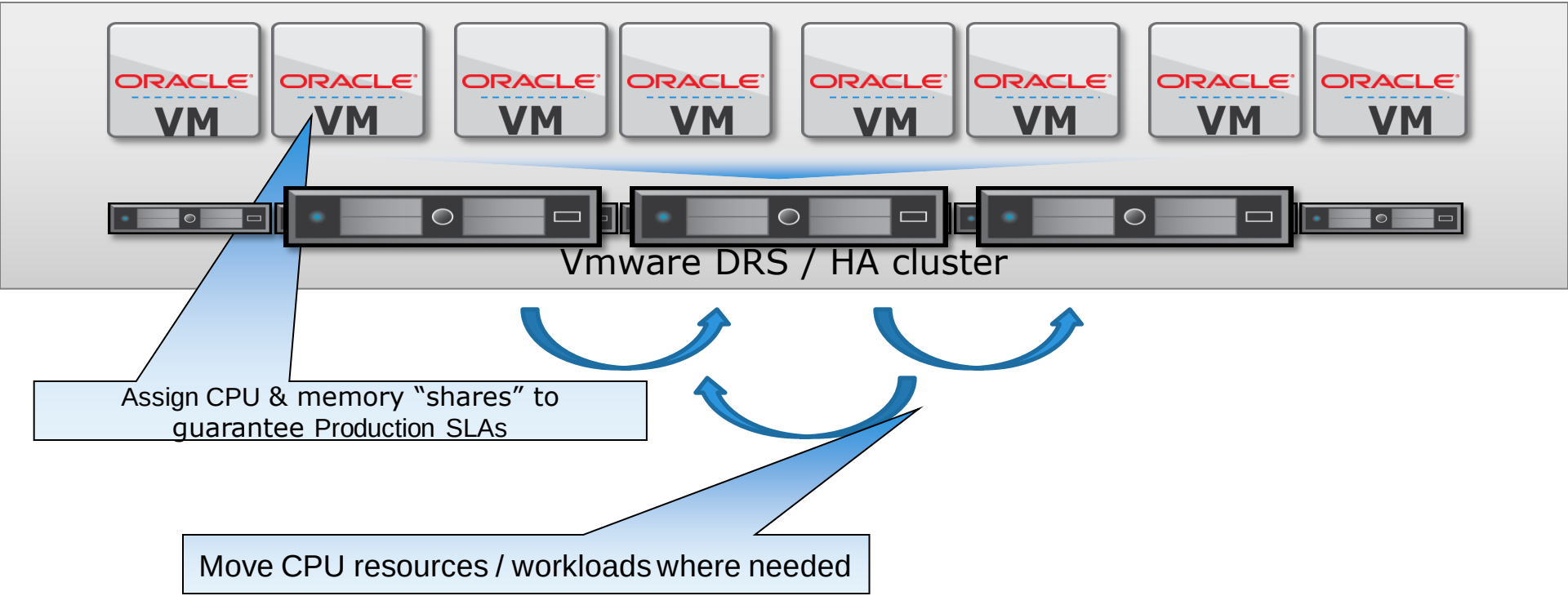
CLASSIC PROBLEM OF RESOURCE MANAGEMENT

(APPLIED TO DB PROCESSING POWER)



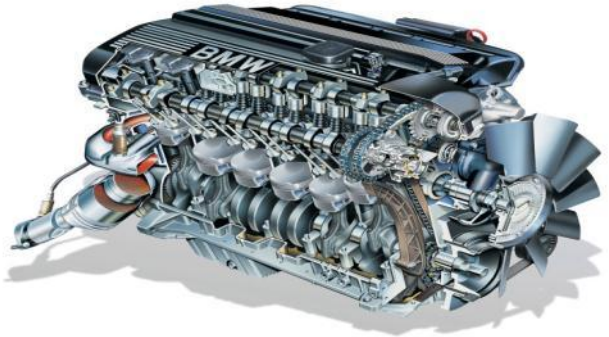
RESOURCE MANAGEMENT

“MAINFRAME STYLE”



IMPACT OF CPU POWER ON LICENSE COST

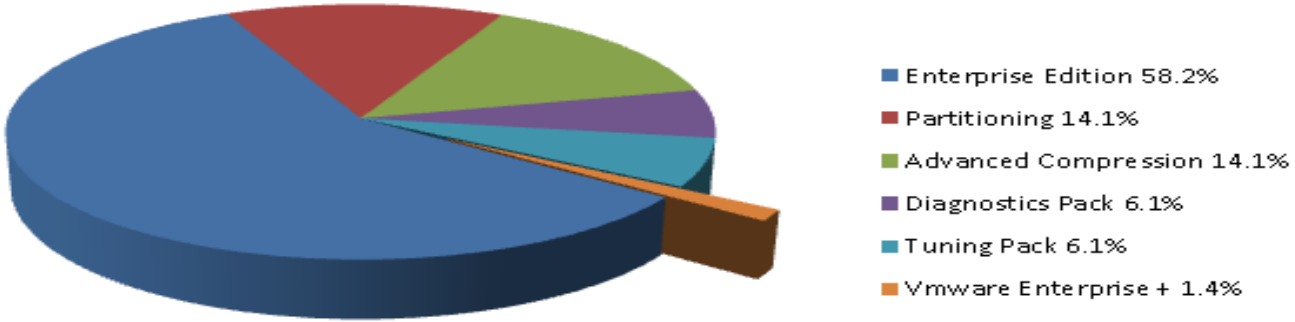
- **CPU power**
  - The more powerful the CPU is per core, the more workload you can run with the same footprint (Without adding licenses!)
- **Memory size**
  - Oracle runs better with lots of RAM (SGA)
  - More RAM allows more VM's per host
- **TPC-C benchmark for OLTP**
  - The industry standard
  - Not all servers listed (Oracle "Engineered" systems are missing ☺)
  - If you're creative you can find similar CPUs and their TPC ratings – or look at SPEC ratings to compare CPU power
- **Powerful CPU cores are more efficient**
  - High TPC-C and/or SPEC ratings will allow you to drive higher consolidation ratios
  - And provide better performance
- **Minimize overhead where possible**
  - VMware: 4% (verified by EMC)
  - Oracle RAC – 10%? (conservative estimate)



Processor types and TPC ratings	TpmC/Core
Intel X5690	87758
Intel E7-8870	63199
Intel E5-2690	100574
Intel E5-2643	100574

Note: Intel E5-2697v2 ~ 115,000 TpmC/Core (estimate)
SPARC T5 ~ 66,800 TpmC/Core (used in SPARC Supercluster T5)
IBM POWER 7+ ~ 150,000, POWER 8 200,000+ (but beware of core factor)

VMWARE – EXPENSIVE?



- Vmware licenses make up less than 2% of total SW licensing
- Will even be lower if you go to 8 cores/socket (common)
- Or if you use Oracle RAC or other additional options

Server: Dual-Socket, 12 core X64
DB licenses: Oracle Enterprise + Partitioning + Advanced Compression + Diagnostics & Tuning pack
VMware licenses: Enterprise Plus (most expensive type)

Based on publicly available list pricing - All other costs (HW&SW) ignored for simplicity

WHY DOES EMC CARE?

THE 3-STAGE ROCKET

Our opinion:

1. Virtualizing Oracle Databases brings huge cost savings and significant operational benefits for our customers
2. VMware is the best platform to make this happen
3. EMC has the best infrastructure and integration to run VMware – and Oracle for that matter

Saturn V liftoff (Apollo 15 mission)

THE REAL COST OF DATABASE TRANSACTIONS

AN OPEN CALCULATION

- Assume a system is CPU bound (removed all I/O bottlenecks)
- Assume we're only looking at license cost as this is >80% of TCO
- The CPU cores can do 35,000 TPC-C transactions per minute (TpmC)
- 10 cores in the system so **RAW** performance is 350,000 TpmC
- License cost is $28,000 per core and includes RAC and Data Guard – total license cost $280,000
- RAW license cost per TpmC $ 280,000 / 350,000 = **$0.80**
- However, there's overhead. RAC=10% (conservative) so TPMC goes down to 315,000
- Servers are utilized only 20% average so at 20% load the server will do only (average) 20% of 315,000 = 63,000
- Real Cost per real transaction per minute is now $280,000 / 63,000 = **$4.44**

OLTP License Cost – Another Example

Physically deployed vs virtually deployed (i.e Cisco UCS/Vblock)

Item	Physical cluster E7-8870	Virtual cluster E5-2690
DB nodes	2	4
Total DB cores	160	64
Clustering/Replication	RAC + Advanced Data Guard	VMware HA / DRS + EMC
Oracle licenses	$4,540,000	$1,288,000
VMware licenses	$0	$ 27,960
Theoretical TpmC @ 100%	10,111,840	6,436,736
Overhead (RAC / Hypervisor)	10%	4%
Average utilization (conservative!)	20%	50%
Effective TpmC (avg load)	1,820,131	3,089,633
Price / TpmC @ 100% load	$0.50	$0.21
Price / TpmC @ average load	**$2.49**	**$0.43**

Highest RAW performance

Highest usable performance

How to get better ROI?

- **Remove RAC & Data guard licensing – replace with VMware and EMC features**
- **Overhead for VMware is lower than RAC (this compensates the VMware licensing a bit)**
- **Utilization goes up from 20% to 50% (conservative)**
- **Because of higher utilization we need less CPU cores (160 ➔ 64) – Over $ 3M savings and still better performance!**

MY ORACLE SUPPORT NOTE 249212.1

Purpose

Explain to customers how Oracle supports our products when running on VMware

Scope & Application

For Customers running Oracle products on VMware virtualized environments. No limitation on use or distribution.

Support Status for VMware Virtualized Environments

Oracle has not certified any of its products on VMware virtualized environments. Oracle Support will assist customers running Oracle products on VMware in the following manner: **Oracle will only provide support for issues that either are known to occur on the native OS, or can be demonstrated not to be as a result of running on VMware.**

If a problem is a known Oracle issue, Oracle support will recommend the appropriate solution on the native OS. If that solution does not work in the VMware virtualized environment, the customer will be referred to VMware for support. **When the customer can demonstrate that the Oracle solution does not work when running on the native OS, Oracle will resume support, including logging a bug with Oracle Development for investigation if required.**

If the problem is determined not to be a known Oracle issue, we will refer the customer to VMware for support. When the customer can demonstrate that the issue occurs when running on the native OS, Oracle will resume support, including logging a bug with Oracle Development for investigation if required.

NOTE: Oracle has not certified any of its products on VMware. **For Oracle RAC, Oracle will only accept Service Requests as described in this note on Oracle RAC 11.2.0.2 and later releases.**

REPRODUCING PROBLEMS ON PHYSICAL SERVERS?

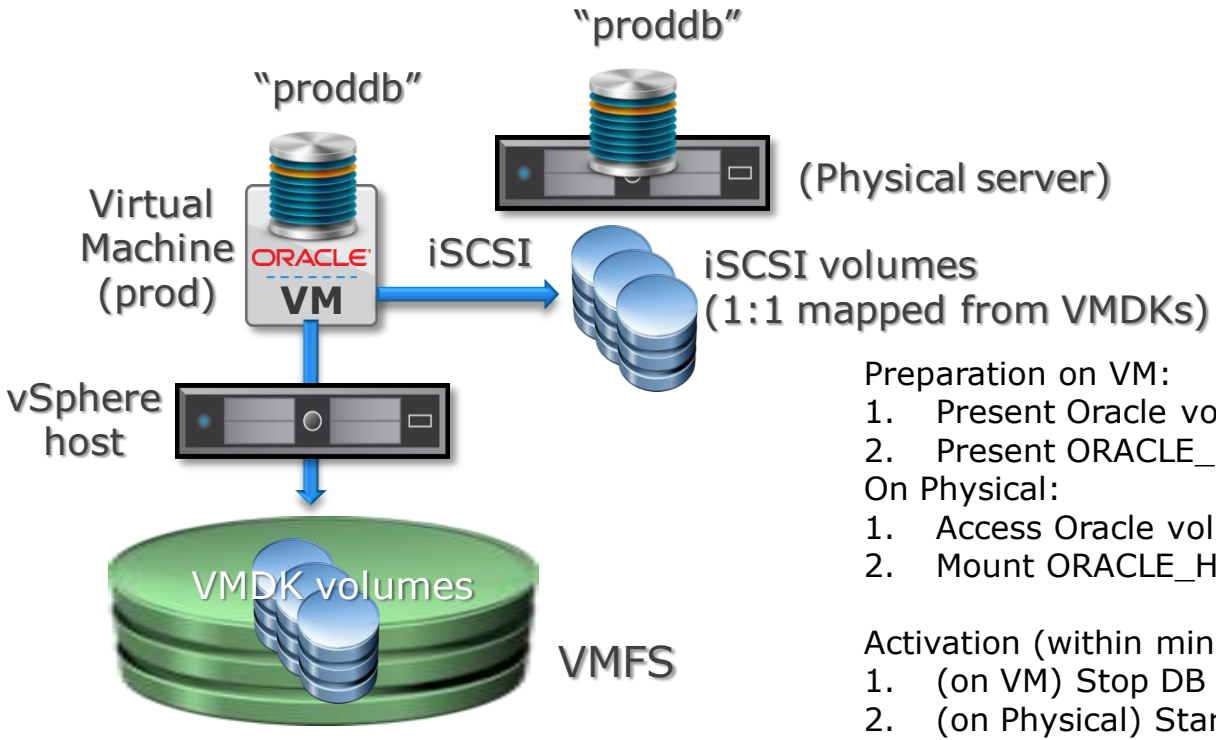
Oracle support occasionally may ask to isolate an issue on a physical server.
How? -> Depends on storage config

First step: VMware Escalation process !

Storage Access	Reproduction Method
Raw Disk Mappings (RDM)	Just mount on physical host
NFS / dNFS	Just mount on physical host
iSCSI	Just mount on physical host
VMFS / VMDK volumes	Physical host cannot mount VMFS file systems or VMDK volumes - Now What?

MOUNTING VMDK VOLUMES ON A PHYSICAL HOST

BEING CREATIVE WITH IP STORAGE



Preparation on VM:
1. Present Oracle volumes as iSCSI targets
2. Present ORACLE_HOME via NFS

On Physical:
1. Access Oracle volumes via iSCSI initiator
2. Mount ORACLE_HOME via NFS

Activation (within minutes):
1. (on VM) Stop DB (or make clone)
2. (on Physical) Start database

Blogpost: [Starting an Oracle database on physical server using VMware VMDK volumes]

VMWARE EXTENDED SUPPORT FOR ORACLE

Total Ownership

VMware Support will accept accountability for any Oracle-related issue reported by a customer. By being accountable, VMware Support will drive the issue to resolution regardless of which vendor (VMware, Oracle, or others) is responsible for the resolution. In most cases, reported issues can be resolved via configuration changes, bug fixes, or feature enhancements by one of the involved vendors.

In the rare situation that another vendor is unable or unwilling to provide a satisfactory technical resolution, VMware Support will immediately notify the customer, assist in escalation and explore other potential technical workarounds with the customer.

VMware will also assist its customers with technical issues for other Oracle software products, besides the Oracle Database and provide similar escalation assistance if needed.

Besides technical assistance, VMware Support will advocate on the customer's behalf to:

- Provide any relevant evidence that virtualization does not play a part in the Oracle product technical problem
- Engage Oracle Support in resolving the customer's technical issue, escalating management attention as appropriate

http://www.vmware.com/support/policies/oracle-support.html

TOP-10 OBJECTIONS (FUD)

#	Objection	Reply
1	VMware is not supported	Oracle Metalink note is available showing support from Oracle
2	VMware is not certified	There is no certification of any platform (OS/HW) except from Oracle/SUN. But all you need is good support
3	Performance is limited	Single VM can have 64 vCPUs, 1TB memory, 300.000 iops (ESX 5.1)
4	Requirement to reproduce problems on physical server	Rare – but if it happens (escalate to VM extended support first) then this is easy with EMC snapshot/cloning technology and offers additional benefits
5	License cost is higher on VMware	Only if you do not build separate DB processing nodes (but make sure you're compliant)
6	Performance overhead	Minimal, and less than Oracle RAC (typical)
7	No workload isolation	Both VMware and EMC have excellent workload management tools
8	No End-to-end "platinum" Customer Service	EMC offers Joint Escalation support, VMware has Oracle accountability program - Both do not require additional support contracts
9	No integrated stack	VBlock systems are completely integrated and tested in EMC's E-lab
10	Alternative hypervisors are cheaper	You need the one that has the right enterprise features and provides the highest consolidation ratios – VMware license cost is very small part of TCO

But the most common objection: we don't want to change because we are familiar with what we have today / frightened of new innovations

VIDEO

Virtualizing Oracle: Caging the Licensing Dragon
http://www.youtube.com/watch?v=FuXBMS2UwyE
Oracle's Richard Garsthagen on Oracle licensing with Vmware
http://www.licenseconsulting.nl/vmworld-richard-garsthagen-oracle-over-licenties-in-gevirtualiseerde-omgevingen/

END of PART 1

Performance Tuning

Best Practices

THE NEW STORAGE AUDIENCE: DBA

Oracle DBAs

Being Asked To Do More…

Performance

Availability

Management

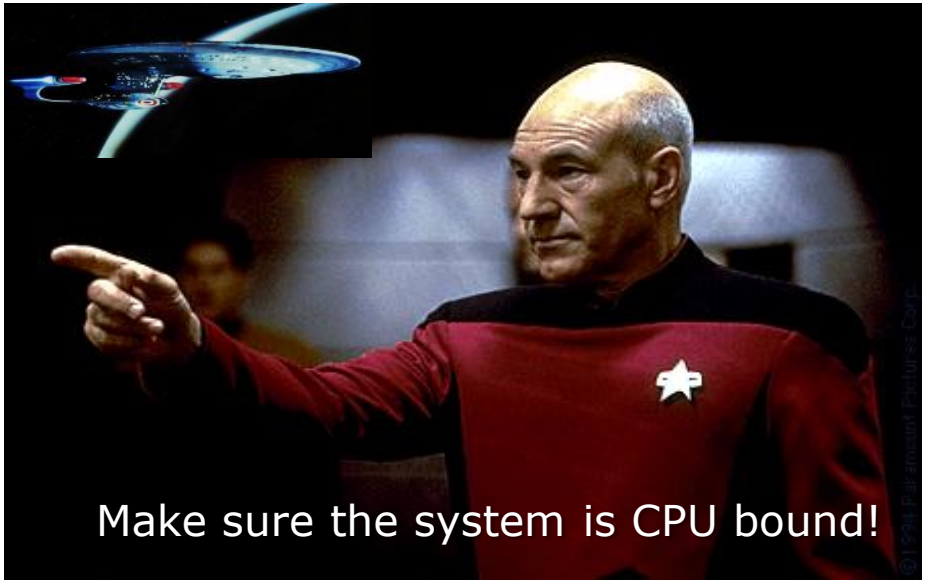
Storage Admin

Being Given More Tools…

EMC²

DATABASES SHOULDN'T HAVE HIGH I/O WAIT

- Adding CPU does not speed up I/O bottlenecks
  - Memory does somewhat
- IOPS are relatively (!) cheap
- CPU cycles are expensive
  - Because of licenses
- Databases have "hot" and "cold" regions
  - No need to make *all* storage fast
  - Modest amount of Flash will do – if applied correctly
  - Adding 5-10% Flash can boost performance by over 80%
  - YMMV ☺



STORAGE IS NO LONGER THE BOTTLENECK

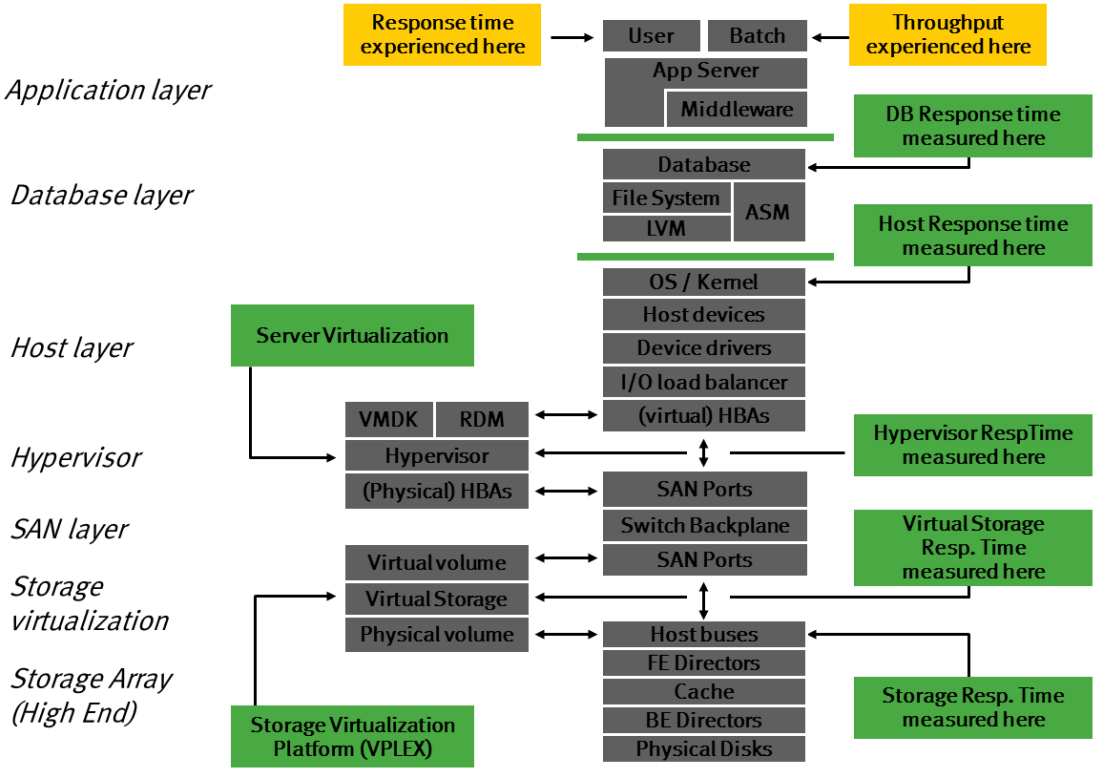
Findings from the field (1)

- DBA and storage teams don't always work well together
- Performance tuning focus on SQL and DB optimization
  - I/O and storage are underrated
  - Knowledge gap between DB and storage specialists
- Performance measured at different levels
  - But using deceivingly similar metrics (i.e. response time)
- Best practices often not honored
  - Data layout, striping, block size, alignment etc
- Limited performance tooling and capacity management in place

Findings from the field (2)

- Business expectations don't match IT
  - Undersized systems
  - Unexpected high peak loads
- Bottlenecks are not known
  - Adding CPU to avoid I/O problem
- Plain wrong architectural decisions
  - Limited up-front research, politics
  - Conservative thinking
- Storage as "black box"
  - "just give me my LUNs"
  - As per the myth told by storage vendor marketing/sales (including EMC…) "the new hardware is so fast, doesn't need tuning"
  - Ignoring storage characteristics such as striping, RAID, disk speed
  - Not using advanced storage features (i.e. snaps/clones, performance features)
  - SATA is cheap, let's put everything on large RAID-6 SATA disks!

UNDERSTANDING THE WHOLE STACK



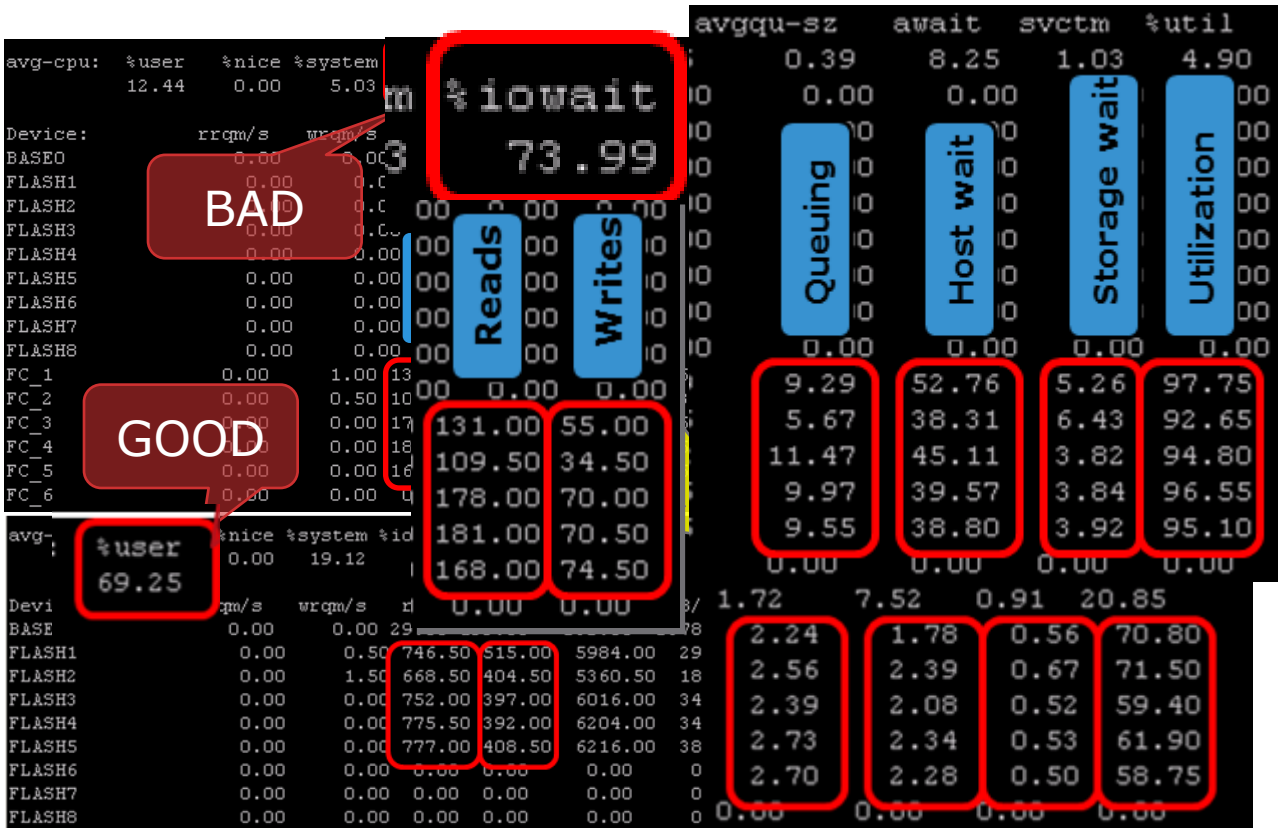
Users experience different performance than DBAs

DBAs measure different metric than storage admins (but named similar!)

- If batch runs 2 hours, is that a perf issue?
- If CPU peaks 100%, is that a perf issue?
- If I/O wait is 95%, is that a problem?

Simplified overview of layers in the database stack: know what you're talking about

UNDERSTANDING I/O WAIT



Linux:

 # iostat –xk 2 /dev/sdX /dev/sdY …

$$Host\ Wait = Service\ time + Queuing\ time$$

- Queuing happens (mostly) on the host
- Having multiple queues is common
- Utilization metric is unreliable

Goal:

Remove all I/O bottlenecks! CPU cycles are too expensive to spend waiting. Or idling.

Locality of reference

- Oracle was developed in a time where CPU and memory was expensive (thus limited)
- Disks perform well (both read and write) if you avoid disk head movements (seeks)
  - How many IOs per sec do you get from cheap SATA disk – given *sequential* 8K reads?
- Therefore database stores related data as close together as possible

➔ Locality of reference

Oracle Database I/O behavior

- Reads are not always sequential but short sequences and related I/O may happen, i.e. block offsets 1001 ➔ 1002 ➔ 997 ➔ 1004 ➔ 1005 ➔ 1009 (consider B-tree index, range scans)
- Storage caching algorithms can optimize this. Consider all of these blocks share a physical disk track – if we do a seek to get to 1001 let's then read the whole track in cache. Now the first I/O (1001) has 7ms resp. time, the rest has << 1ms ☺
  - Since 1995, EMC has invested heavily in R&D (i.e. analyze I/O traces etc.) to improve these algorithms
  - Note that tablespace and file system fragmentation, striping and other indirection mechanisms (Volume managers, write-anywhere file system schemes) can ruin your day ☹
- If you have sequential write data it could make sense to assign dedicated disks
  - REDO logs, DWH staging areas

I/O skewing

- Database objects (indexes, tables) tend to grow by appending blocks at the end
- Due to the nature of business processing, the most recently added data (rows) are likely to be retrieved more often
- The oldest data is less likely to be very active
- So we get (slowly moving) hot spots (and respectively, cold spots) in the data
- This is called “skewness” i.e. 80/20 skew means 80% of I/O happens on 20% of the data blocks
- In that case you can reduce seek time on 80% of all I/O requests to be below 1ms – by putting it on FLASH storage (but the devil is in the details)

The Performance Gap Challenge

CPU Improves 100 Times Every Decade; HDD Remains Flat

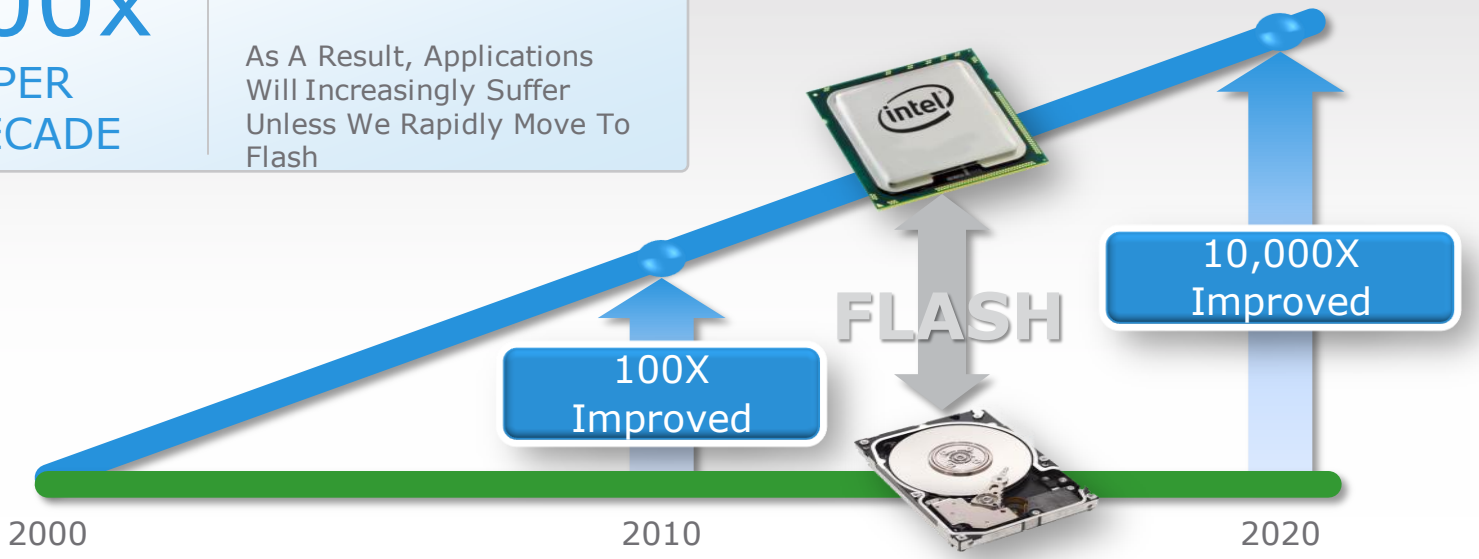
MOORE'S LAW:

100x

PER DECADE

CPU Continues To Improve While Disk Drive Performance Remains Flat

As A Result, Applications Will Increasingly Suffer Unless We Rapidly Move To Flash

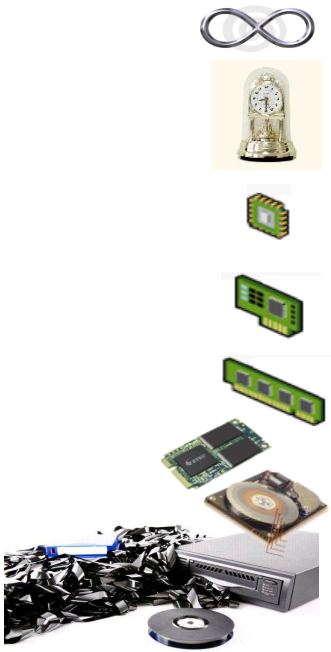


FLASH VERSUS SPINNING DISK

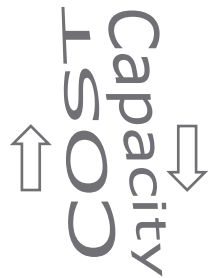
Single spinning disk	Single Flash Disk (SLC / eMLC)
One operation at a time	Parallel operations – any workload
Mechanical movements required for seeks	No mechanical parts
Cannot handle high utilization well	High utilization is fine
Reads perform like writes – no need for zero out before write	Writes require clearing out flash regions first – sustained writes may cause degraded performance - Garbage collection required
Sweet spot: sequential R/W	Sweet spot: random read
I/O directly relates to physical offset on disk	I/O offset obfuscated due to wear leveling
Typical resp. time ~ 7 ms (@ low % busy)	Typical resp. time ~ 0.5 ms (@ high % busy)
Random IOPS ~ 150	Random IOPS ~ 3000 (depends!) (* outdated)
Bandwidth ~ 70 MB/S (sequential read/write)	Bandwidth ~ 70 MB/s (sequential read)
Wears out by age, not usage	Wears out by (overwrite) usage
No wear leveling required	Needs wear leveling
Requires caching algorithms for good (random) performance	Requires caching algorithms for (good write) performance + endurance

ACCESS TIMES OF STORAGE MEDIA

TYPICAL RELATIVE SPEEDS OF COMPONENTS (2013) 1NS = 1S

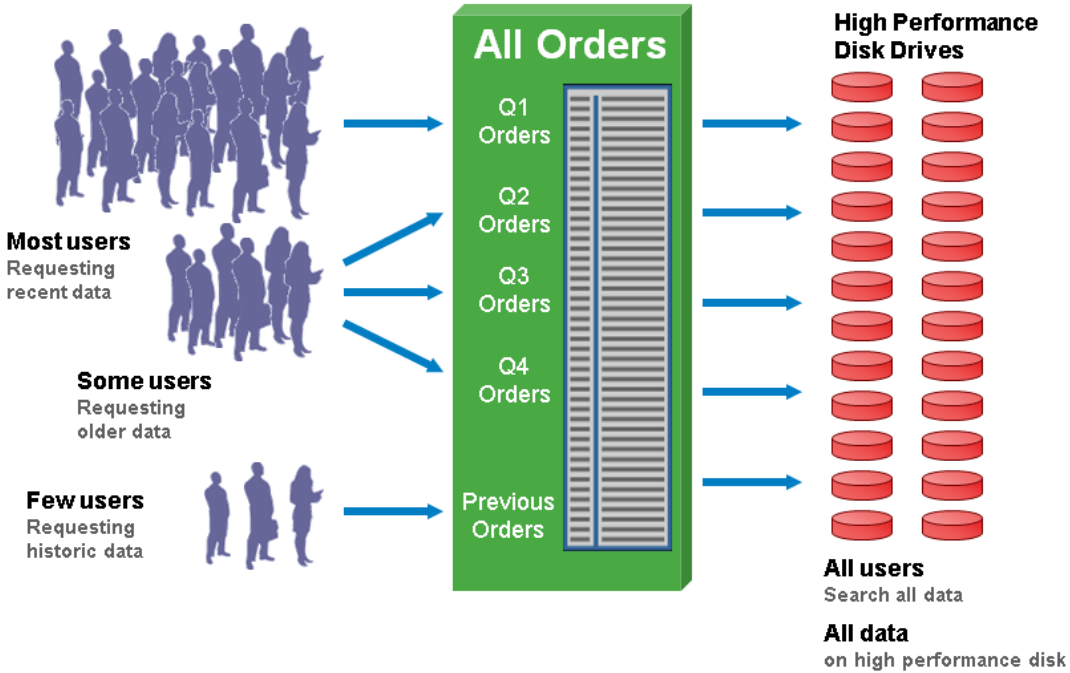


Access type	Typical Cycle Time (nanoseconds)	Cycle time (s)	Scaled Cycle Time (scale = 10^9)	Typical Capacity
Avoided IO	Zero	Zero	Zero	-
CPU clock (2.5 GHz)	0.4	4×10^{-10}	0.4 seconds	-
L1 cache	2	2×10^{-9}	2 seconds	64KB
L2 cache	4	4×10^{-9}	4 seconds	256KB
L3 cache	25	$25 * 10^{-9}$	25 seconds	4 MB
DRAM	100	100×10^{-9}	1 minute 40 sec	256 GB
Flash Memory	50,000	50×10^{-6}	14 hours	1 TB
Flash Disk	500,000	0.5×10^{-3}	5 days	10TB
Rotating Disk	7,000,000	7×10^{-3}	3 months	100TB
Tape	10,000,000,000	$1 \times 10^{+1}$	3 centuries	Petabytes



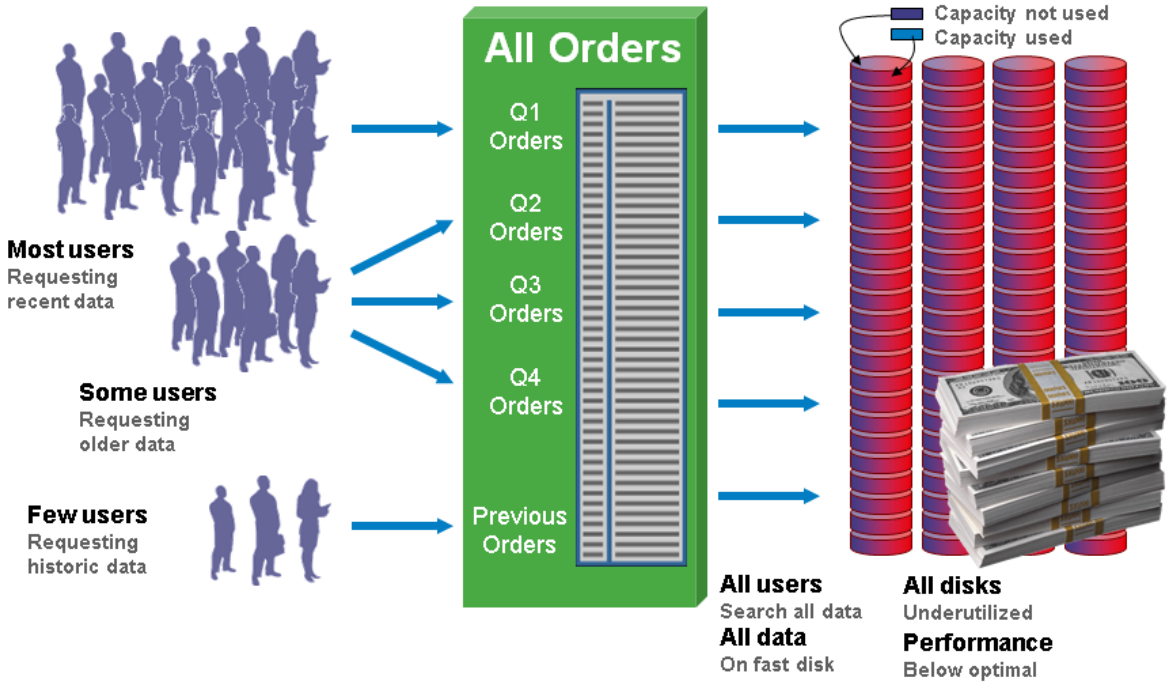
DATABASE STORAGE TIERING

TRADITIONAL DEPLOYMENT – ONE TIER – SPINNING DISK



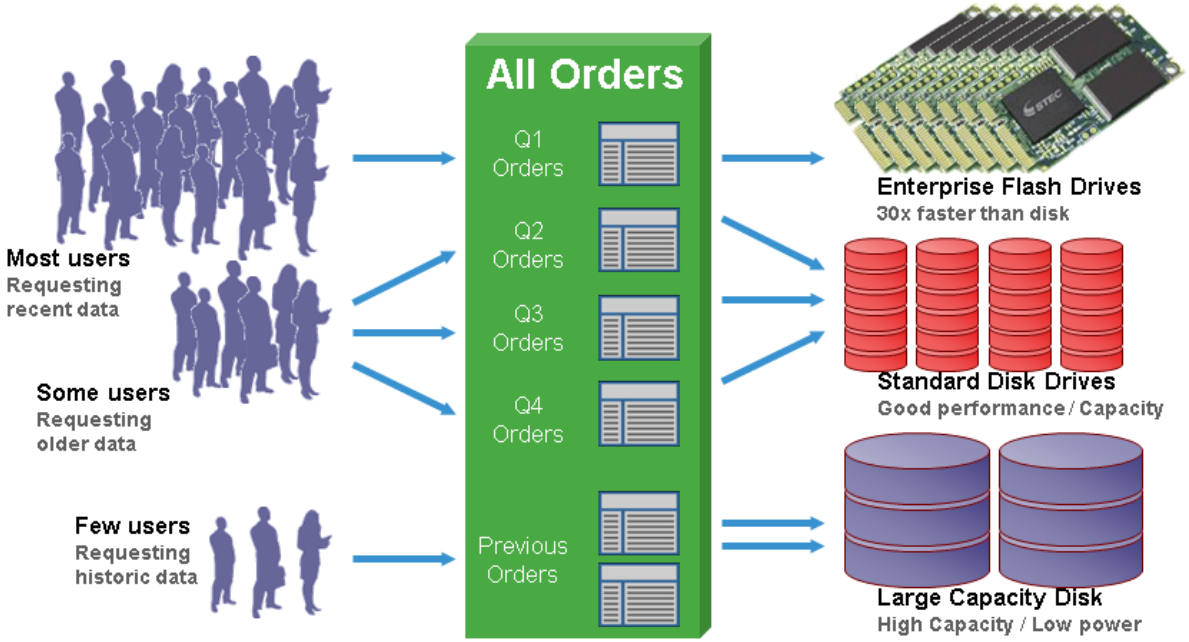
DATABASE STORAGE TIERING

GROWING SIZE AND WORKLOADS



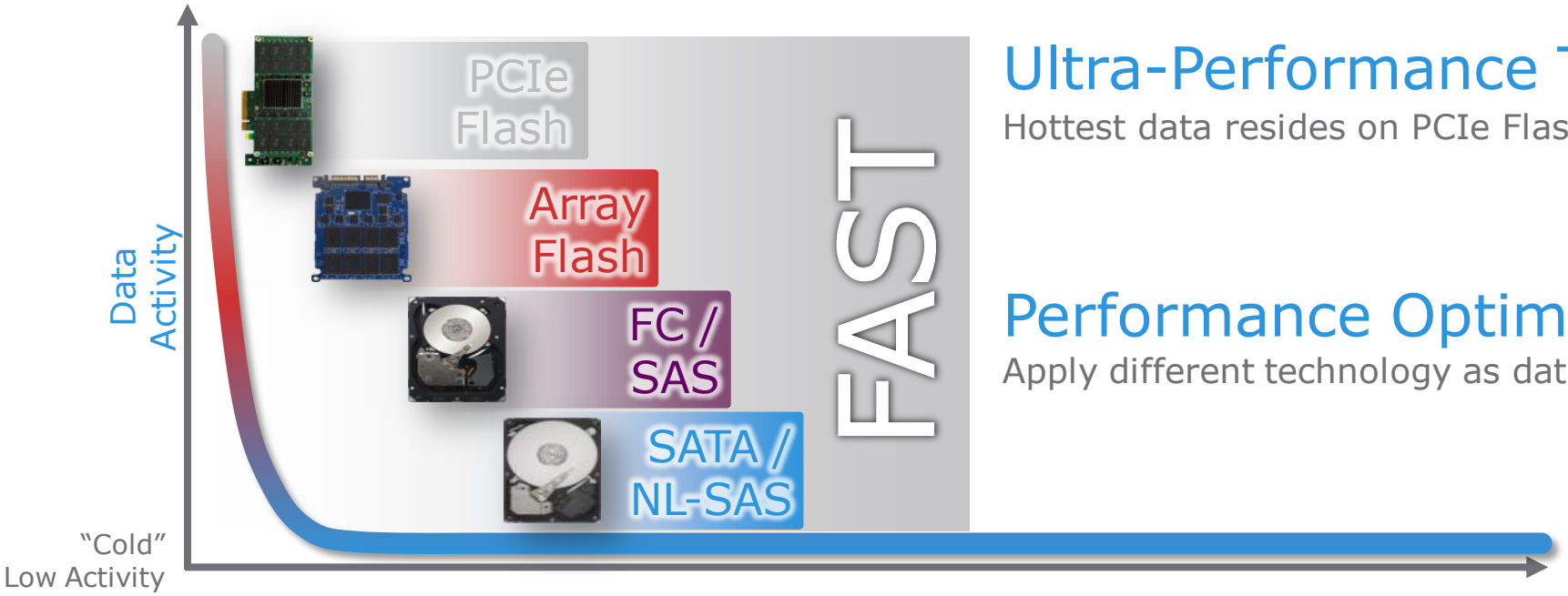
DATABASE STORAGE TIERING

IMPLEMENTING ILM (MANUAL, ~ 2008)



DATABASE STORAGE TIERING

FULLY AUTOMATED TIERING



Ultra-Performance Tier

Hottest data resides on PCIe Flash

Performance Optimized

Apply different technology as data cools

ORACLE ON EMC BEST PRACTICES

EMC RECOMMENDS VARIOUS SETTINGS FOR GOOD PERFORMANCE. EXAMPLES:

- Linux Hugepages
  - Reduces CPU overhead in managing Linux memory management
- Linux I/O scheduler
  - Elevator or deadline? Or CFQ?
  - Virtual: NOOP!
- Queue depths
  - Tradeoff between response time and throughput
  - No good "formula" available. Trial & error.
- EMC Powerpath for load balancing
  - Works better than native or 3rd party "MPIO"-style balancers
  - Linux MPIO is known to sometimes chop large I/O into 4K chunks (bad)

ORACLE ON EMC BEST PRACTICES

- Disk alignment
  - Use 64K or 1MiB (both are fine)
  - Linux "fdisk" creates 31,5K "misaligned" partitions – resulting in overhead
  - More info: http://bartsjerps.wordpress.com/2013/03/28/linux-alignment-reloaded/
- REDO logs
  - 100% sequential write
  - No duplexing required unless 3rd party vendors require this (has no benefit for protection)
  - Don't make larger REDO log groups than needed
  - ASM: External redundancy - EMC is very good at data protection, don't spend precious host CPU and I/O cycles on that
  - Where possible, dedicate physical disk groups for REDO. RAID-5 FC/SAS is fine. Sharing with other DBs is fine.
  - Where possible, dedicated I/O channels might reduce response times (avoid REDO IO having to wait for background DB writer I/O for example)

ORACLE ON EMC BEST PRACTICES

- Striping
  - Oracle 11.2: defaults to coarse striping for REDO. Change back to FINE striping (128K)
  - Avoid striping for everything else (both ASM and FAST-VP avoid hotspots anyway)
  - Really avoid double striping (can kill all prefetch / performance algorithms)
- ASM
  - External redundancy!
  - Separate ASM disk groups
  - Increase default ASM AU size to ≥ 8MB (recommended 16MB)
  - Split REDO logs, FRA/ARCH, TEMP and regular data files
  - Sometimes it makes sense to go beyond that and split some index/data
- TEMP
  - Create TEMP on dedicated FLASH/EFD if DB uses TEMP for sorting/joining etc
  - TEMP generates random read/write which is boosted by using Flash storage

ORACLE ON EMC BEST PRACTICES

- Remote Replication
  - Asynchronous SAN replication typically has ZERO performance impact but still guarantees consistency
  - And reasonable RPO for many applications (~ 5 to 10 minutes)
  - Use SYNC only where really needed (such as financial processing)
  - ZERO Dataloss is (partly) a myth: The Zero Dataloss Myth blogpost
  - No matter if you use Data Guard or SAN replication (i.e. EMC SRDF, Recoverpoint)
- Database init parameters
  - Don't modify things for performance POCs that you wouldn't modify in production
  - Such as block checksum "disabled" settings and other exotic stuff
  - We're in search of realistic, predictable, not just "breaking the record" performance numbers
  - DB block size: 8KB (DWH benefits from ≥ 16K sometimes). Never go lower than 8K !
  - Many parameters that potentially influence IO (such as MBRC)

ORACLE ON EMC BEST PRACTICES

- Queue depths
  - Large queue depth: more throughput
  - Small queue depth: better response time
  - No silver bullet / single recommendation
- Consistent, predictable “good” performance is better than unpredictable, unreliable “Guinness World Records” performance
  - Can athletes consistently achieve world records? Or once in a lifetime?
  - Should we test performance also under “special conditions” ?
  - Such as disk failures, broken cables/channels, during RAID rebuilds, with SYNC replication enabled (i.e. Data Guard or EMC SRDF), when performing DB cloning using snaps/clones, when users are submitting crazy table scans, …
  - During backups / restores (same server or same cluster / shared infra)
  - During firmware updates

ORACLE ON EMC BEST PRACTICES

- Oracle RAC?
  - Can sometimes cause more problems than improvements due to RAC interconnect traffic, locking, pinging etc
  - A workload that requires 30 CPU cores is typically better off with a 32-core single-node server than a 2-node 16-core/node cluster
  - These days a single Intel host can have 80+ processors. Why scale out? Scale up!
  - Use when you need extreme availability (mostly not performance as large single-node servers do better) - In that case, consider Oracle RAC stretched clusters (with EMC VPLEX)
- Generic HA (cluster) tools can offer quick failover times as an alternative
  - And don't forget license cost
- Beware of CPU Overhead
  - Specific hypervisors: VMware ESX overhead= 4% (as measured by EMC IT)
  - Oracle RAC: no hard numbers (but many would agree it's at least 10%)
  - Host replication (i.e. ASM redundancy, log shipping): ~ 1-2% CPU + mirrored writes
  - Don't run anything else on DB server except DB processing! (No apps, middleware, mgt agents, …)

ORACLE ON EMC BEST PRACTICES

- IP based protocols
  - (Direct) NFS as good as Fiber Channel these days
  - Provided one applies all best practices (jumbo frames, non-blocking switches, 10GigE, …)
  - Excellent alternative to ASM, dNFS = 100% NFSv3 compliant (no vendor-specific magic)
- Exotic filesystems?
  - Avoid ZFS for primary datafiles (heavy fragmentation and other issues, requires lots of tuning, see my blogposts on the matter)
  - Avoid OCFS/OCFS2 (performance, I/O chopping™ into 4K, not mainstream)
- Other filesystems: YMMV ;-)
  - Be prepared for lots of "Evil" tuning of bottlenecks
  - Filesystems often use RAM that otherwise could be allocated to SGA (use directIO etc)
  - FS prefetch is much less efficient than DB caching itself -> disable!
- Beware of heavy memory paging / thrashing

RAID LEVELS & DISK TYPES FOR ORACLE DATAFILES

- Data / Index
  - Read and Write
  - Large & small I/O
  - Both Random & sequential
  - RAID-5 is OK, RAID-1 is (a bit) better
  - Avoid RAID-6 (and RAID-6 - like)
  - Split tablespaces if you need to squeeze out that extra 5%
  - Isolate from REDO, ARCH, FRA, etc on physical disk level
  - A bit of FLASH a day keeps the performance doctor away
  - Auto-tiering (FAST-VP)!
- REDO logs
  - 100% sequential write
  - RAID-1 or RAID-5 (both are OK)
  - No need for 15K rpm (but use this if rest of system also uses 15K)
  - FC/SAS is OK (no need for EFD/Flash)
  - Preferably on dedicated physical disks (if redo I/O is high)
  - Sharing with other databases is fine
  - Tune for fast write response times of small block I/O
  - Exclude from tiering policies

RAID LEVELS & DISK TYPES FOR ORACLE DATAFILES

- Binaries
  - Any (reliable) storage is OK
- TEMP
  - Oracle's "paging space"
  - Separate if high DB TEMP usage
  - Very random I/O pattern (if used)
  - Used for joins / sorts / aggregates
  - And Index builds (+ reorg?)
  - On Flash/EFD where needed
  - Regular tier is OK if no high TEMP usage (shared with DATA)
- FRA/ARCH
  - Confusion: used for both Archive logs and backup files, and Flashback logs…
  - All three are good candidates for RAID-6 SATA (cost-effective) as performance is not very important
  - Sometimes contains control files as well (tricky with replication) – avoid!

PERFORMANCE PROOF OF CONCEPTS

PROFILING NEW SYSTEMS BEFORE YOU GO LIVE

- Always test low-level performance
  - Using a mix of "dd", "iorate" or Vdbench, etc
- Always test transactional workloads
  - Using Swingbench, HammerDB or similar TPC-C "like" tools
- Always test IOPS and throughput
  - Only one tool is good enough: SLOB
  - SLOB does IOPS only (random read and/or write)
  - "slob-fulltablescans.sql" adds sequential read (bandwidth) test (beware: single threaded for now): Slob Full table scans (blogpost)
- Only after basic tests, run your own custom queries
- Now you're confident to go live ☺

SUGGESTED WORKLOAD GENERATING TOOLS

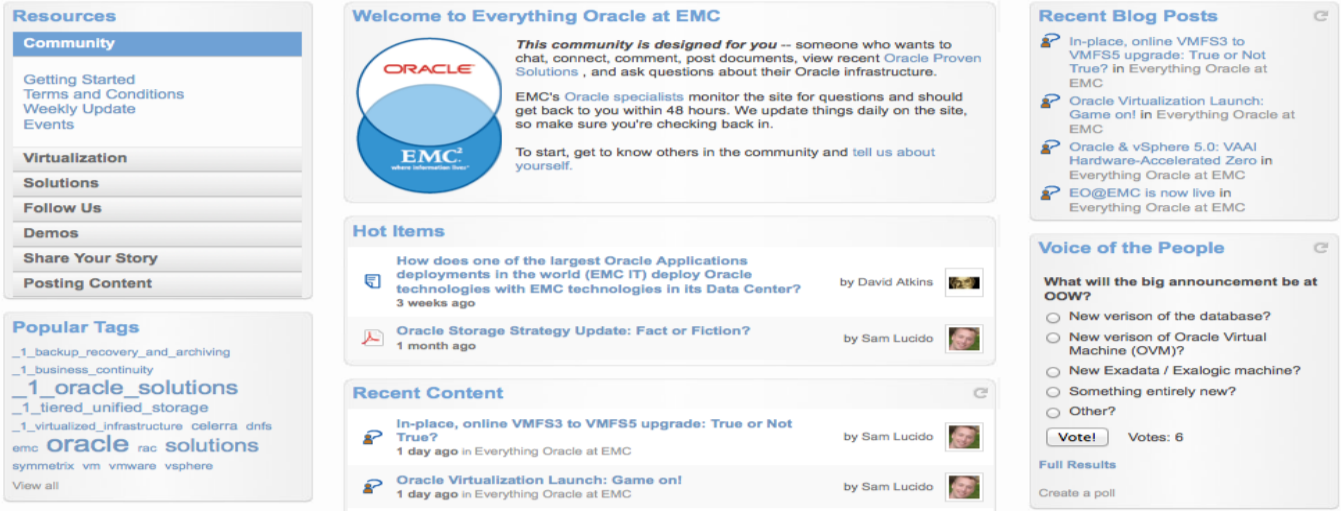
- Swingbench
  - Has become the De-facto tool to simulate OLTP workloads
  - Swingbench SOE (Sales Order Entry) has become the "unofficial" TPC-C like benchmark
  - Typically CPU bound (if infra configured to have no I/O bottlenecks i.e. use Flash where needed etc)
  - Performance may vary depending on generated data size and DB configuration (i.e. SGA, block size etc) – the detailed DB stack configuration + Swingbench setup must be documented and repeated across different tests
  - Not a good tool to drive lots of I/O
  - Very good tool to compare CPU power of platforms
  - Note that OLTP is often CPU-bound (like many DWH queries for that matter)
- SLOB
  - The "Silly Little Oracle Benchmark" created by Kevin Closson
  - Not a real benchmark but a pure Oracle I/O generator
  - Basically generates lots of database block reads and/or writes (plus redo I/O) without driving high CPU
  - Use it to profile I/O limits without depending on CPU and memory
- UNIX tools
  - dd, cp, etc: good for getting initial "feel" if the system is driving enough bandwidth
  - Not a good benchmark
- IORate
  - EMC public domain tool to generate I/O (without database)
  - Can be used for initial profiling
  - If all works well, should match SLOB results (more or less)

PERFORMANCE POC SUGGESTIONS

- Use both Swingbench and SLOB
  - Swingbench to profile TPC-C like transactions per minute
  - SLOB to profile I/O workload
- Test multiple workloads (different servers) at the same time
- Using VMware CPU shares, see how service levels are met
  - i.e. a VM "prod" with 2000 shares should get more TPM than a VM "test/dev" with 500 shares if they share the same physical host
  - See if and how VMware starts moving workloads across physical servers to balance out the workloads real-time
- Test the replication to physical server procedure
  - Oracle might occasionally ask for that when providing support
- Optional: Using and auditing CPU affinity
  - To manage license cost in some occasions

EMC COMMUNITY NETWORK

EVERYTHING ORACLE @ EMC



- Provides a focal point for all of EMC's Oracle-related activities
- EMC's Oracle-related Proven Solutions content now publicly available and searchable on Google
- Go to:http://community.emc.com/community/connect/everything_oracle

REFERENCES

My Blog

http://bartsjerps.wordpress.com

Blog post on Oracle/Vmware licencing

(and how to save money):

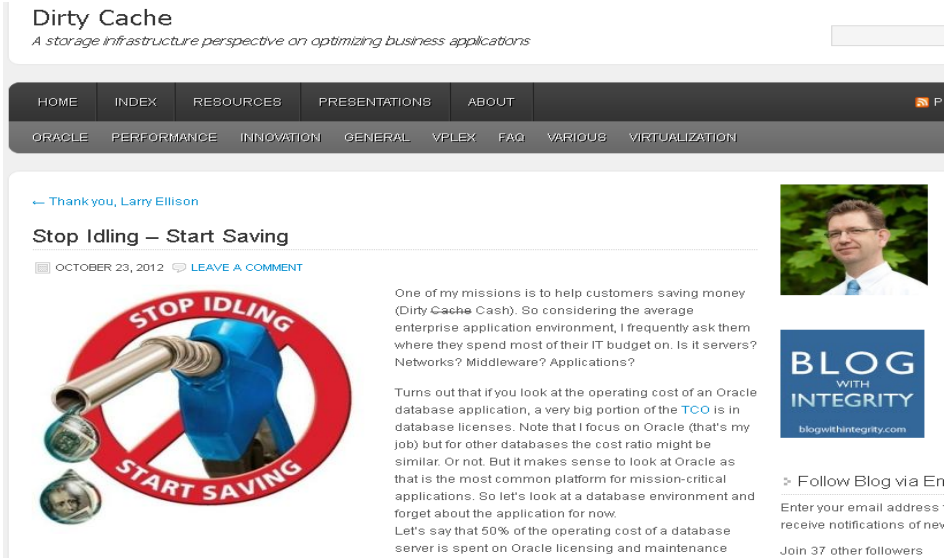
http://bartsjerps.wordpress.com/2011/11/09/oracle-vmware-licensing-cost-savings/

Everything Oracle @ EMC (community):

https://community.emc.com/community/connect/everything_oracle

Or

http://emc.com/everythingoracle




EMC²