REMEDIOS
PROFESSIONAL IT SERVICES

otpbank

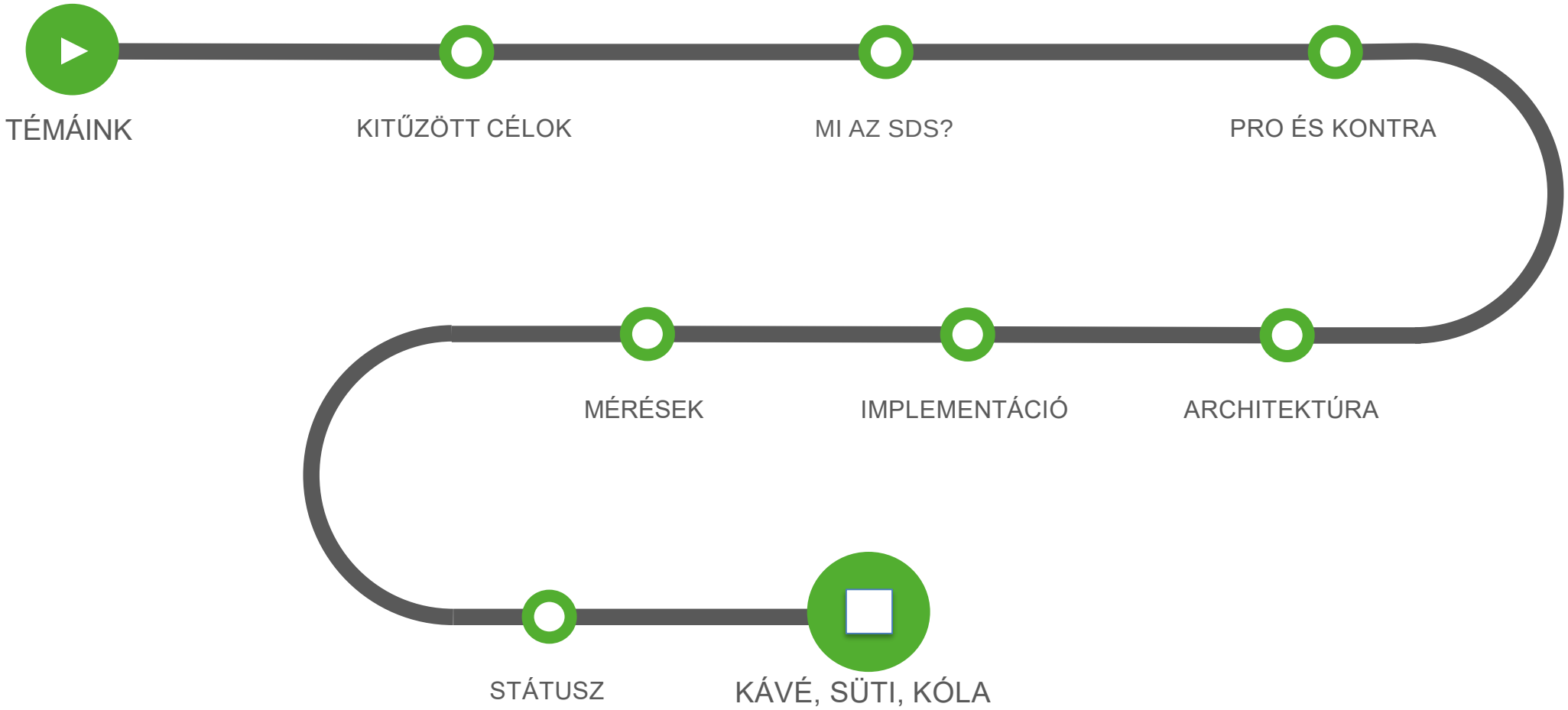
Oracle RAC tapasztalatok
Software Defined Storage-on az OTP Bankban

Banki megoldások szekció

Előadók:
Lapos Attila – OTP Bank
Koltai Róbert – Remedios

DELL EMC

Témáink


TÉMÁINK
KITŰZÖTT CÉLOK
MI AZ SDS?
PRO ÉS KONTRA
ARCHITEKTÚRA
IMPLEMENTÁCIÓ
MÉRÉSEK
STÁTUSZ
KÁVÉ, SÜTI, KÓLA

Bank céljai

Üzleti célok

- Kritikus, banki alkalmazások adatainak tárolása
- Stabilitás, magas rendelkezésre állás
- Teszt környezetek gyors készítésének támogatása
- Változó alkalmazás oldali igények flexibilis támogatása

Bank céljai

Nagy rendelkezésre állás, megbízhatóság

- Szolgáltatás kiesése nélkül bővíthető diszkekkel, node-okkal.
  Az OS, illetve VxFlex OS kiesésmentesen upgrade-elhető legyen
- iSER helyett IP alapú, teljesen titkosított, hibakezeléssel rendelkező adatátvitel
- A storage réteg karbantartása transzparens a compute node-ok felé
- Certifikált, támogatott megoldás

Bank céljai

Flexibilitás

- Korábban fix, nagy méretű LUN-ok voltak csak kialakíthatók (4,5TB, illetve 6 TB) VxFlex OS esetén a legkisebb LUN méret 8GB
- Megoldás legyen egyszerűen bővíthető, horizontálisan, illetve vertikálisan skálázható

Bank céljai

Egyszerűbb üzemeltetés

- Gyors üzembe helyezés
  (Install során egy Excel fájl megadása a gateway szervernek)
- Váljon egyszerűbbé, biztonságosabbá az ASM disk group-ok bővítése, karbantartása
- Megoldás rendelkezzen menedzsment felülettel

Bank céljai

Storage funkciók

- Lehetőség snapshot készítésre.
- QOS LUN szinten.

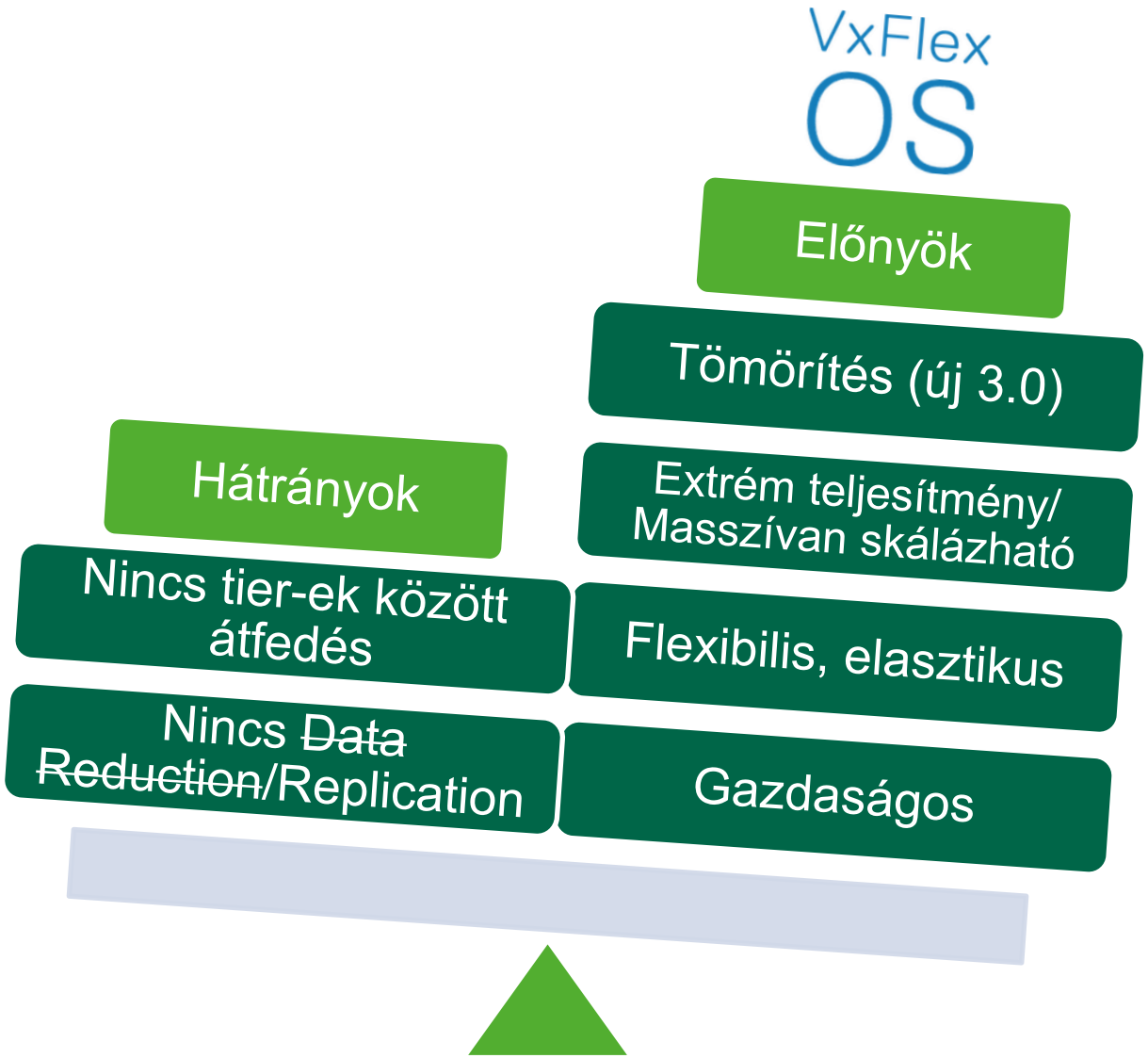
Teljesítmény

- 1 million IOPS (12 storage node)
- 28 GB/s olvasási, 8GB/s írási sebesség (128k blokkméret mellett)

SDS definíciója

- Általános célú hardware-en fut (commodity hardware)
- Lokális erőforrásokat használ (CPU, RAM, local disk)
- Általános célú operációs rendszeren fut
- Blokk alapú adattárolási funkciót lát el ethernet hálózaton
- Biztosítja az általános storage funkciókat (RAID, snapshot, tömörítés)

Pro/Con


VxFlex OS
Előnyök
Tömörítés (új 3.0)
Extrém teljesítmény/ Masszívan skálázható
Flexibilis, elasztikus
Gazdaságos
Hátrányok
Nincs tier-ek között átfedés
Nincs ~~Data Reduction~~/Replication

Architektúra

Tároló felhasználó

Storage Data Client (SDC)

- Tárolót használó szerverekre installáljuk
- Alkalmazások, filesystem-ek részére volume elérést biztosít
- Minden SDS-el P2P kapcsolatot tart fenn

Tároló szolgáltató

Storage Data Server (SDS)

- A klaszter számára tárolót szolgáltató szerverekre installáljuk
- Lokális tárolóból biztosít tárhelyet az SDC-k részére

Tároló klaszter menedzser

Metadata Manager (MDM)

- Tároló klaszter konfigurációt, monitorozást, rebalancing-et, monitoring-ot és újáépítést vezérli
- Magas rendelkezésre állású klaszter, mely 3-5 node-on van telepítve
- Futhat SDS-en, SDC-n vagy egyéb node-on
- Az adatelérésben nem vesz részt

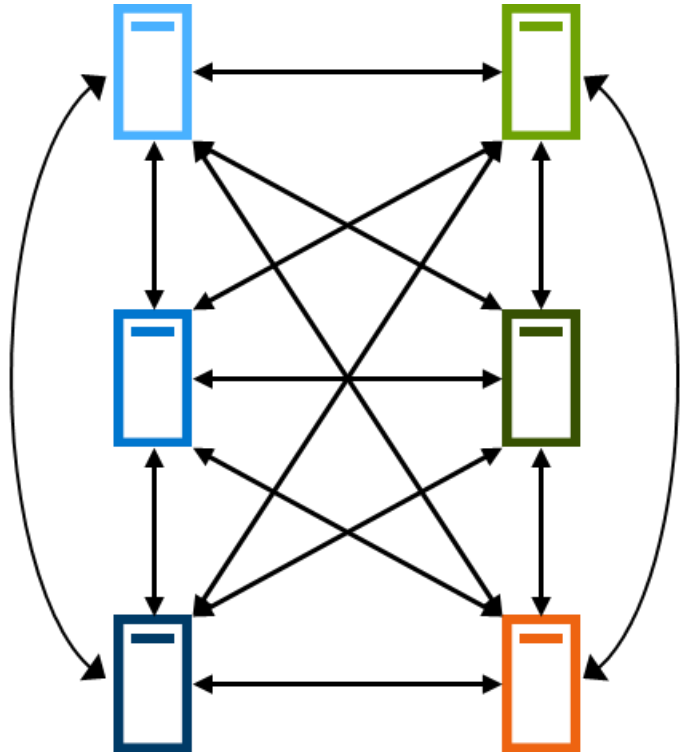
Architektúra

Nagy mértékű párhuzamosság

- IO elosztása a szerverek közti nagyszámú útvonalon
- Minimális CPU és memória használat
- Flash media tökéletes kihasználása
- Az adat chunkok és ezek tükreinek elosztott hálója

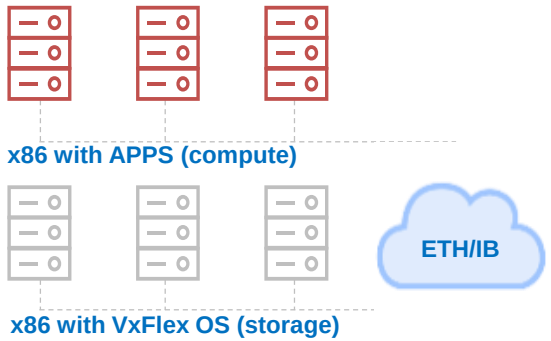
Lehetővé válik

- Nagyon nagy teljesítmény – az SDS-ek teljesítményének összege
- Média meghibásodás javítása másodpercekben mérhető
- Elasztikus kapacitás & lineáris teljesítmény
- CPU és memória túlnyomó része az alkalmazásoké marad
- Extrém flexibilitás



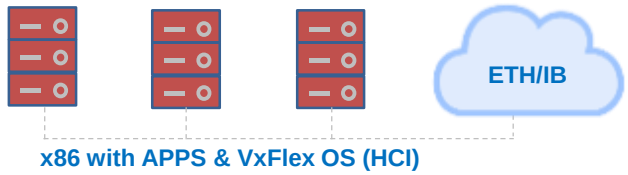
Architektúra

Hagyományos kétrétegű



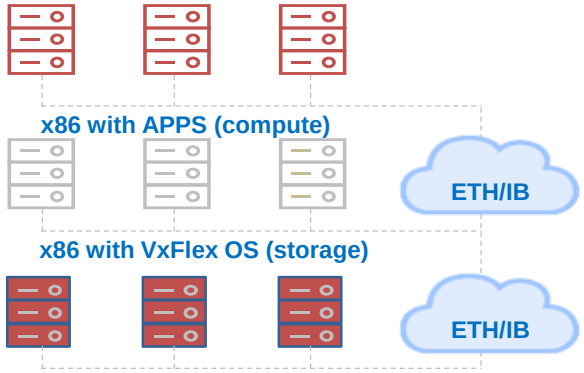
- Hagyományos SAN felépítéshez hasonló
- Szerver vagy SDS vagy SDC szerepben van
- Olyan szervezeteknél ajánlott, ahol a rétegek üzemeltetését más csapat végzi

Modern, hiperkonvergens



- Szerverek SDS és SDC szerepet megvalósítanak
- Modern megközelítés
- Maximális flexibilitást nyújt

Kevert



- Mindkét megközelítést lehetővé teszi
- IT osztályok megválaszthatják a saját megközelítésüket

Oracle RAC

Áttekintés

- RAC = Real Application Clusters
- High Availability és Scalability képességekkel terjeszti ki az Oracle adatbázist
- Ugyanazt az adatbázist egyidejűleg több hosztról használjuk
- Hoszt szintű, instance szintű meghibásodást kiesésmentesen kezel
- 2-4 node esetén majdnem lineáris skálázódás
- Patchelés kiesés nélkül támogatott
- Elérhető Standard és Enterprise Edition kiadások esetén

Mérések

KONFIGURÁCIÓ

- 3 RAC klaszter (2+4+4=10 node)
- Oracle 12.2.0.1 verzió
- Silly Little Oracle Benchmark (SLOB)
  - 512 séma adatbázisonként
  - 1GB/séma
  - 8k táblatér blokkméret

MÉRÉSEK KIÉRTÉKELÉSE

- AWR riportok segítségével történt
- Oracle Diagnostics Pack termék része

Válaszidő

- 2 node-on rövid 5 perces teszt
- 192 thread/node
- 100% olvasás
- 8k blokkméret

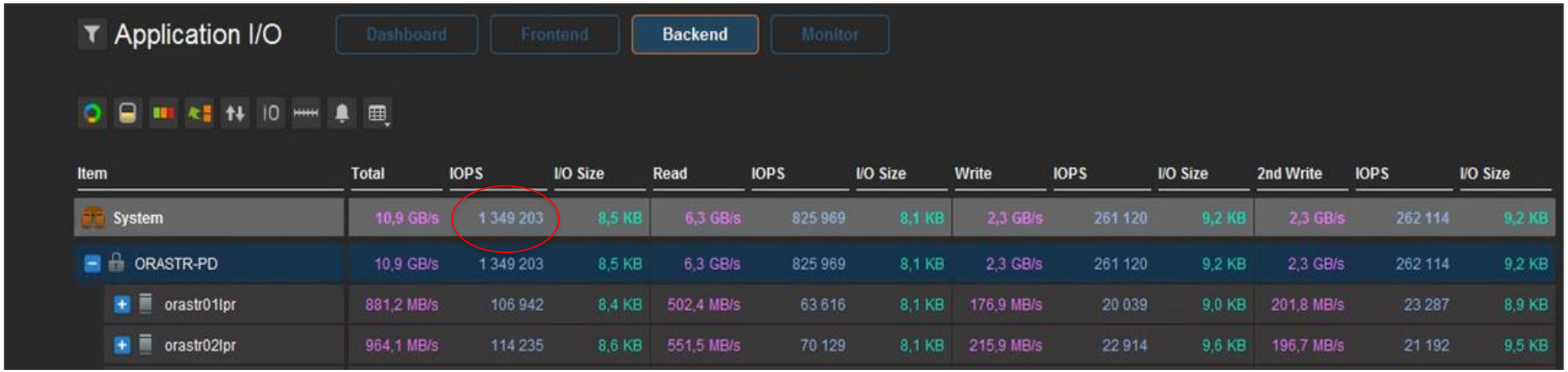
Top Timed Events

	Wait		Event		Wait Time			Summary Avg Wait Time				
I#	Class	Event	Waits	%Timeouts	Total(s)	Avg Wait	%DB time	Avg	Min	Max	Std Dev	Cnt
*	User I/O	db file sequential read	60,363,972	0.00	16,891.00	279.82us	73.20	279.95us	257.50us	302.40us	31.75us	2
		DB CPU			4,024.94		17.44					2

- 68% < 256µs
- 92% < 512µs
- 98% < 1ms

Áteresztőképesség

- 10 node-on rövid 5 perces teszt
- 640 DB session
- 70% írás + 30% olvasás
- 8k blokkméret



Item	Total	IOPS	I/O Size	Read	IOPS	I/O Size	Write	IOPS	I/O Size	2nd Write	IOPS	I/O Size
System	10,9 GB/s	1 349 203	8,5 KB	6,3 GB/s	825 969	8,1 KB	2,3 GB/s	261 120	9,2 KB	2,3 GB/s	262 114	9,2 KB
ORASTR-PD	10,9 GB/s	1 349 203	8,5 KB	6,3 GB/s	825 969	8,1 KB	2,3 GB/s	261 120	9,2 KB	2,3 GB/s	262 114	9,2 KB
orastr01lpr	881,2 MB/s	106 942	8,4 KB	502,4 MB/s	63 616	8,1 KB	176,9 MB/s	20 039	9,0 KB	201,8 MB/s	23 287	8,9 KB
orastr02lpr	964,1 MB/s	114 235	8,6 KB	551,5 MB/s	70 129	8,1 KB	215,9 MB/s	22 914	9,6 KB	196,7 MB/s	21 192	9,5 KB

- 1,35 mIOPS (825k Read, 2x260k Write
- 10,9 GB/s (6,3GB/s Read, 2x2,3 GB/s Write)

Stabilitási stressztesztt

WORKLOAD REPOSITORY REPORT (RAC)

Database Summary

9,652 nap

Database								Report Total (minutes)	
Id	Name	Unique Name	Role	Edition	RAC	CDB	Block Size	DB time	Elapsed time
268109763	SLOBG	slobg	PRIMARY	EE	YES	NO	8192	3,333,165.93	12,903.00

OS Statistics By Instance

- Listed in order of instance number, I#
- End values are diplayed only if different from begin values

	CPU			Load		% CPU				
I#	#CPUs	#Cores	#Sckts	Begin	End	% Busy	% Usr	% Sys	% WIO	% Idl
1	24	12	2	1.70	74.94	22.16	10.61	9.42	69.12	77.84
2	24	12	2	0.37	55.93	15.20	7.56	6.48	64.78	84.80
3	24	12	2	0.60	68.08	20.52	9.92	8.87	68.35	79.48
4	24	12	2	0.47	52.39	14.93	7.38	6.36	65.63	85.07
Sum										

3 RAC klaszteren összesen
240 core
6128 Xeon Gold 3.4Ghz

~70% WIO

Stabilitási stresszteszt

Foreground Wait Classes - % of DB time

- % of Total DB time - instance DB time as a percentage of the cluster-wide total DB time

				% DB time				
I#	User I/O	Sys I/O	Other	Network	Concurcy	Config	Cluster	DB CPU
1	97.60	0.00	0.05	0.00	0.00	0.01	0.02	3.48
2	97.48	0.00	0.07	0.00	0.00	0.02	0.02	3.40
3	97.47	0.00	0.07	0.00	0.00	0.02	0.02	3.50
4	97.61	0.00	0.06	0.00	0.00	0.02	0.02	3.40
Avg	97.54	0.00	0.06	0.00	0.00	0.02	0.02	3.44

Adatbázis szinten kizárólag IO aktivitás

IO Profile (Global)

Statistic	Read+Write/s	Reads/s	Writes/s
Total Requests	284,321.35	216,270.39	68,050.95
Database Requests	282,274.38	216,211.91	66,062.47
Optimized Requests	0.00	0.00	0.00
Redo Requests	502.34		502.34
Total (MB)	2,378.49	1,787.15	591.33
Database (MB)	2,311.95	1,767.47	544.48
Optimized Total (MB)	0.00	0.00	0.00

1,7MB x 12,903 x 60 =
= 1,383,459,660 MB =
= **1,3 PB** olvasás egy RAC klaszter felől

Stabilitási stresszteszt

Top Timed Events

~1ms single block olvasás (8k)

	Wait		Event		Wait Time			Summary Avg Wait Time				
I#	Class	Event	Waits	%Timeouts	Total(s)	Avg Wait	%DB time	Avg	Min	Max	Std Dev	Cnt
*	User I/O	db file sequential read	166,885,184,710	0.00	194,818,853.14	1.17ms	97.41	1.16ms	1.14ms	1.20ms	26.22us	4
		DB CPU			6,902,459.74		3.45					4
	System I/O	db file parallel write	12,465,529,860	0.00	6,063,730.11	486.44us	3.03	492.53us	458.38us	525.53us	35.47us	4

Wait Event Histogram (Global)

86% < 1ms

		% of Total Waits														
Event	Waits	<64us	<128us	<256us	<512us	<1ms	<2ms	<4ms	<8ms	<16ms	<32ms	<64ms	<128ms	<256ms	<512ms	<1s	>=1s
db file sequential read	167.1G		8.8	37.1	28.7	12.8	3.5	1.5	2.9	3.9	0.9	0.0	0.0	0.0	0.0	0.0	

Jelenlegi használat

- 10 storage, 12 compute node
- Több, mint 100 adatbázis, folyamatos migrációk
- Nettó 1 PB tervezett adat
- Stabil, hibamentes működés

Kérdések?