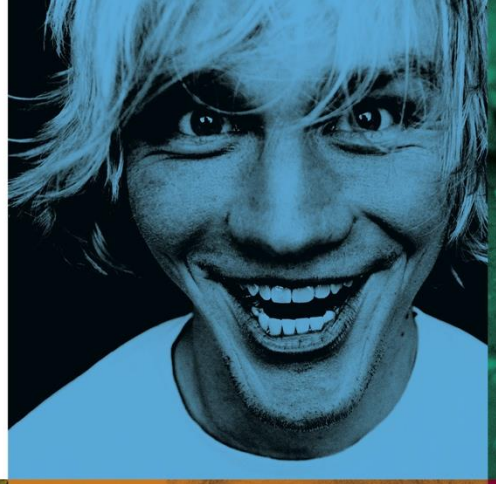
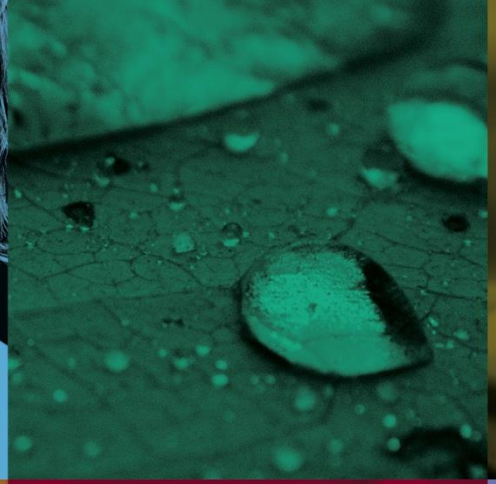
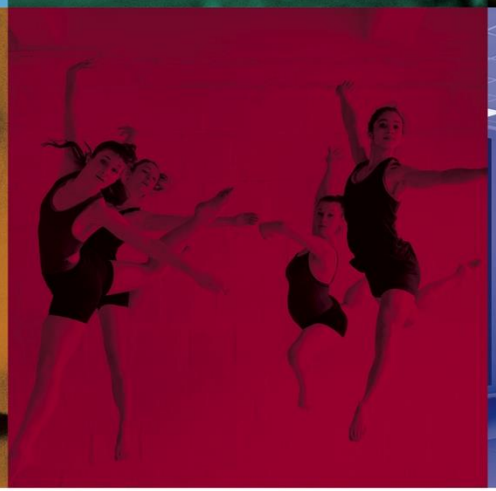
NIL





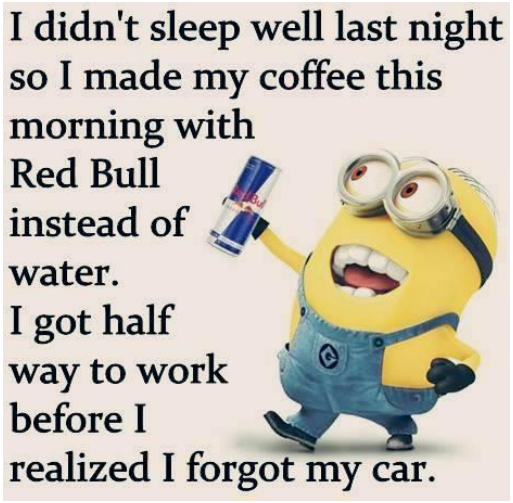

nil.com

NIL

Mitja Robas

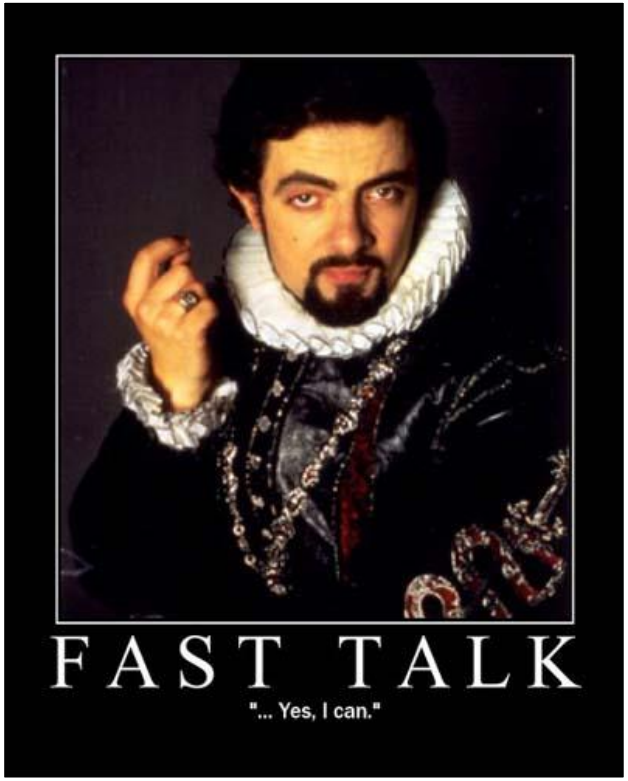
STORAGE
THE FINAL FRONTIER

Disclaimer



Ker se že dalj časa ukvarjam s tehnologijo in rešitvami SSD/flash kjer so zakasnitve minimalne ter je količina uporabnega prostora večja od količine nazivnega ...

... govorim hitro in veliko ...



NIL

Storage – The Final Frontier

„OZKO GRLO“

ZAHTEVE SE VIŠAJO

- Vedno več (zahtevnih) aplikacij
  - Podatkovne baze (MS SQL, Oracle, DB2, PostgreSQL, poštni stežniki ...)
  - Virtualna namizja (VDI, RDS)
  - Infrastrukturni sistemi - Unified Communications (Voice, Video, IM), EMM(MCM+MDM+MAM), ...
  - Aplikacijski sistemi - SAP, MS Dynamics NAV, MS Dynamics CRM, MS Exchange, Sharepoint, Lotus Domino, Websphere, analitika, ...
  - Razvojno-testno-produkcijsko okolje
- Sobivanje aplikacij !

DISKOVNI SISTEMI

- Moorov zakon => CPU, MEM
- Zahteve po kapacitetah se množijo (TB >> PB)
- Prepustnost/hitrost sistemov ne sledi kapaciteti
  - IOPS
  - Predvidljiva nizka zakasnitev pri dostopu do podatkov

	Performance	Capacity	Flexibility
CPU	✓	✓	✓
Memory	✓	✓	✓
Network	✓	✓	✓
Storage	X	✓	X

KLASIČNE REŠITVE

- Toge, zapletene, neprilagojene današnjim in prihodnjim zahtevam
- Izboljšave/dodatki zgolj prikrijejo probleme
- Niso za vse tipe aplikacij
- Rešitev lahko postane problem

Diskovni sistemi

- Medpomnilnik, autotiering, vgrajeno stiskanje podatkov
- Omejena velikost
- “Kazen” v kolikor podatki niso prisotni v hitrem nivoju
- Dodatna obremenitev (prostor, obdelava)

Strežniške kartice

- Nivo medpomnilnika v strežnikih
- Tipično ni sinhronizacije podatkov med medpomnilniki (ali pa je ta zelo “draga)
- Bralno intenzivne aplikacije z bolj statičnimi podatki

Virtualizacijske rešitve

- Virtualizacija diskovnih sistemov
- Dodaten nivo med shranevalnim(i) sistemi ter strežniki
- Programsko-strojna rešitev „medpomnenja“

TEHNOLOGIJA SSD/FLASH

MITI O TEHNOLOGIJI FLASH/SSD

- Tehnologija je
  - draga
  - nezanesljiva
  - zgolj za naslavljanje visokih zahtev
  - kapacitete so premajhne
  - ne niža stroškov (napajanje, prostor)

- Tipi flash/SSD (SLC, eMLC, MLC, TLC)

- Tehnologija flash/SSD
  - >10x hitrejši od diskov (3500+ IOPS)
  - 10x večja učinkovitost pri porabi prostora & energije

Leto 1956

Prvi disk IBM 305 RAMAC

- Teža > 1 tona
- š*g*v = 74*152*172 cm
- Kapaciteta = 5MB
- 50 * 61cm diskov
- Povprečen čas dostopa do zapisa 600 msec

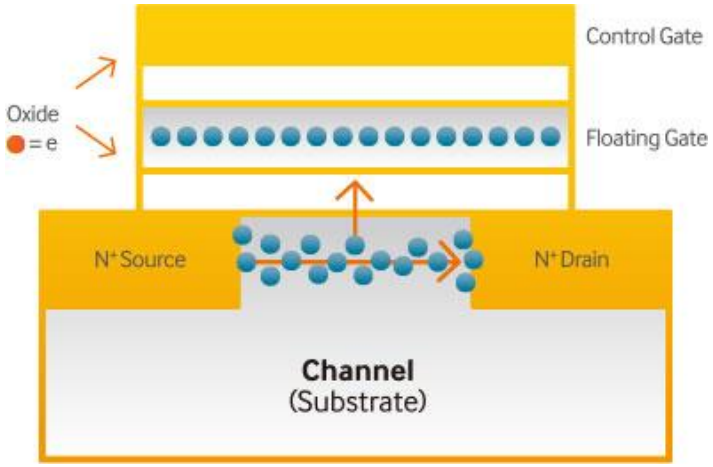
Leto 2015

Samsung Portable SSD T1

- Teža = 25,5 g
- 71mm * 53mm * 9,14mm
- Kapaciteta = 1TB (200.000* 305 RAMAC)
- 3D-vertical NAND flash

PRIMERJAVA
SSD/FLASH – DISK

- Tehnologija NAND
  - Pomnilniške celice
  - Shranjevanje informacij
    - Uporaba napetosti na CG
    - Write – približanje elektronov na FG
    - Erase – oddaljevanje elektronov od FG
  - Informacija je lahko shranjena več let
  - Dodatni procesi
    - Bad block management
    - Wear leveling
    - Garbage Collection (GC)
    - Error Correcting Code (ECC)
    - Write amplification handling
- Arhitektura trdih diskov je popolnoma drugačna
  - Plošče, glave, sektorji, bloki, ...



KARAKTERISTIKE SSD/FLASH TEHNOLOGIJE

- Kapaciteta SSD/Flash diska – GB >> TB
- Vzdržljivost, Zanesljivost, Prepustnost
  - Izraba (wear out) => Write cycles = Program/Erase cycles
  - Izraba zaradi spreminjanja stanj celic (write, erase)
  - Več bitov/celico = manj pisanj
  - Write Amplification Factor – brisanje celic pred pisanjem

SSD tip*	Bit/cell	Write Cycles**	Program time [us]	Erase time [ms]	Read time [us]	Cost/GB
SLC – Single-Level Cell	1	100,000	200-300	1 – 2.5	25	Highest
eMLC – Enterprise MLC	2	20,000 – 30,000	600-900	2.5 – 3.5	50	High
MLC – Multi-Level Cell	2	2,000 – 10,000	600-900	2.5 – 3.5	50	Low
TLC – Triple-Level Cell	3	1,000 – 5,000	~900-1350	4 – 5	~75	Lowest

* TLC je tudi 3-bit MLC
** Numbers vary

ŽIVLJENJSKA DOBA SSD/FLASH

- Vzdržljivost v praksi
  - Življenjska doba [Leta] = letni TB prirast
  - Letni TB prirast = količina podatkov pisanih na SSD/flash na leto
  - Celoten TB = celotna količina podatkov zapisanih v življenjskem ciklu
- Statični (75%) in dinamični (25%) podatki

Namig – pogooglajte „SSD endurance/torture testing“

SAMPLE TLC SSD ENDURANCE TEST*

- 97TB podatkov v 8 dneh – neprekinjeno zapisovanje
- S.M.A.R.T - pokaže da so se celice napolnile do 406
- Porabljeno 41% uradne življenjske dobe (1,000 ciklov pisanja)

@41% wearing reached

Lifespan	TB/year	GB/day
27 years	3.65	10
5 years	19.4	53
3 years	32.3	88.5
1 year	97	265.75

Avg. 40GB/day in Enterprise environment

- Rezultati testa => ~3x več ciklov pisanja

SSD	TB written in total	Write Cycles	Lifespan
Disk #1	~777	2945	17.74 years
Disk #2	~768	3247	17.53 years

Disk fails 40GB/day WAF = 3

* Samsung SSD 840 250GB TLC SSD & non-real life load

KAKO SE ZAGOTAVLJA UČINKOVITOST?

- Uporaba funkcionalnosti stiskanja podatkov
  - Kompresija
  - Data-deduplication (block size)
  - Inline (izraba izravnalnikov za stiskanje)
  - All-thin-provisioning
- Poimenovanje razpoložljivih kapacitet
  - RAW
  - Used RAW
  - Useable
  - Koncept “manj fizičnega” več “logičnega prostora”

Tip aplikacije	Povprečna stisljivost	Mogoča stisljivost
Server virt.	5:1	6:1 or more
VDI	10:1	12-25:1
Database	3:1	4:1 or more

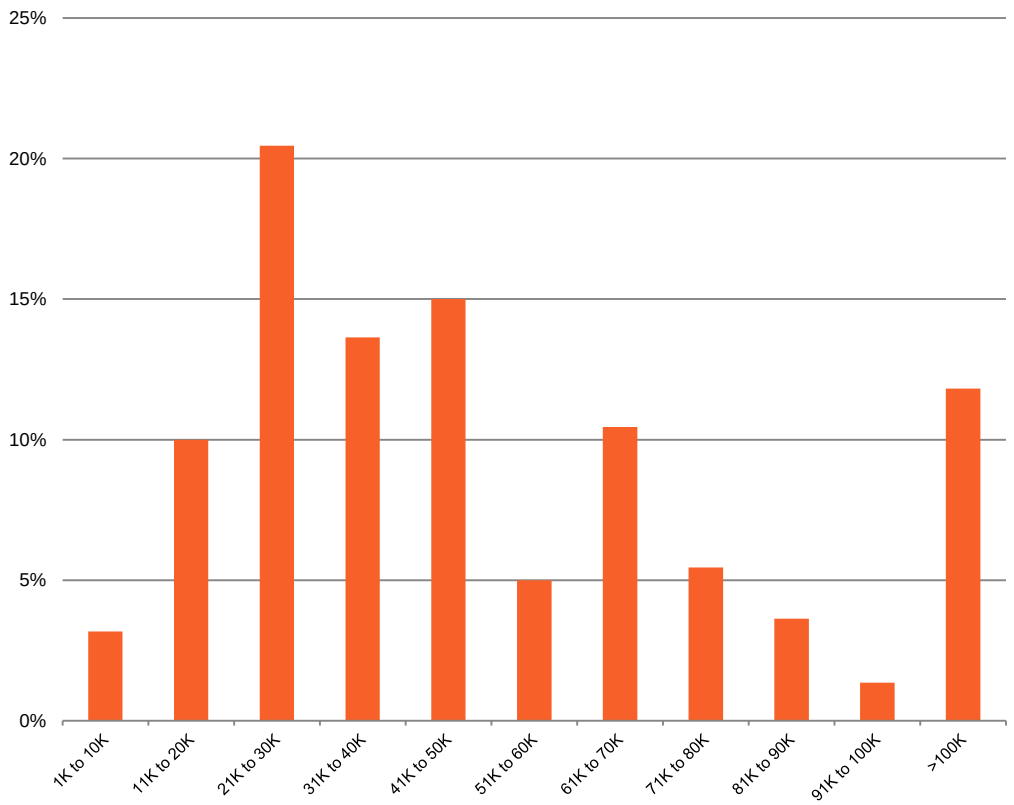
NIL

Storage – The Final Frontier

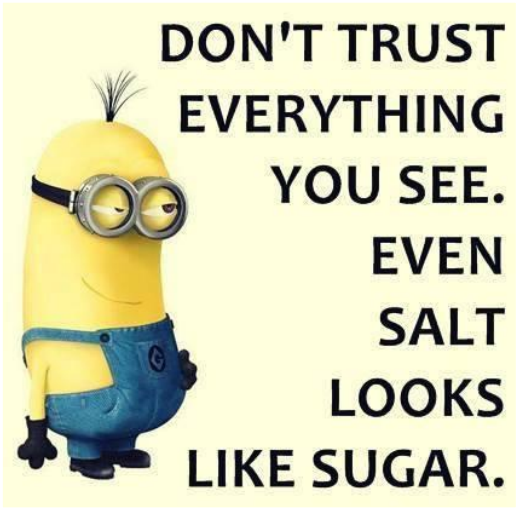
IN PRAKSA?

SHRANJEVALNI SISTEM
PREPUSTNOST/ZMOGLJIVOST

- Pomembna je povprečna velikost IO



RAZKORAK MED PODOBO IN RESNICO



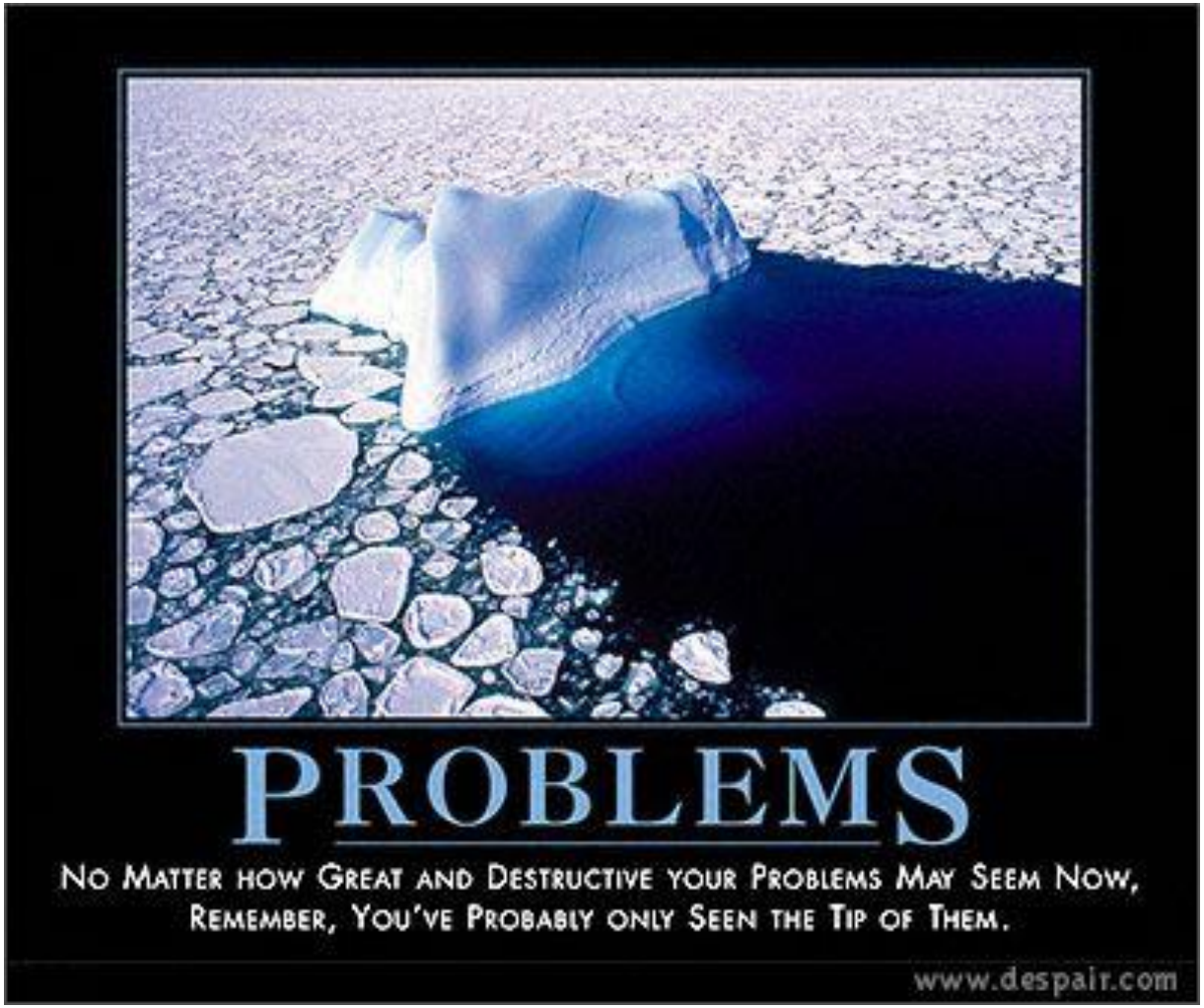
- Ne verjemite tehničnim specifikacijam (lepotno tekmovanje)
- Brošure „lažejo“ oz. prikazujejo nepomembne podatke
- Pojmi so ZELO „raztegljivi“
- Razumevanje je ZELO odvisno od interpretacije
- Pomembna je arhitektura
- Zastavite (si) prava vprašanja
  - Podatki <=> aplikacije <=> uporabniki
- Predvsem pa ne postavljajte si omejitve z obstoječo tehnološko rešitvijo

PRIMER

- Aplikacije
  - Strežniška virtualizacija = 10TB
  - Podatkovne baze = 5TB

- Zahteve
  - Prostor – uporabna kapaciteta
  - Konstantno nizka zakasnitev
  - Cena
  - Vzdrževanje
  - Življenska doba
  - Razširljivost

HIDDEN CHALLENGES



VPRAŠANJA

Pomembno

- ✓ Arhitektura sistema
- ✓ Vsota funkcionalnosti (istočasnost)
  - HA
    - Dual parity
    - Komponente
    - Zaščita „cache“
  - Kompresija
- ✓ Razširljivost
- ✓ Enostavnost uporabe
- ✓ Upravljanje

Nepomembno

- ~~× Thin provisioning~~
- ~~× Dual-controllers~~
- ~~× RAID 6~~
- ~~× Cache~~
- ~~× Scale-up vs. scale-out~~
- ~~× Fork-lift upgrade~~
- ~~× Cell wearing level~~
- ~~× Tipi diskov~~
- ~~× Število diskov~~

REŠITEV #1

Podoba

- 20 * 500GB diski = 10TB
- Data reduction (4:1)
- Visoka razpoložljivost
  - Dva kontrolerja
  - RAID-6
- Prepustnost m*100k IOPS

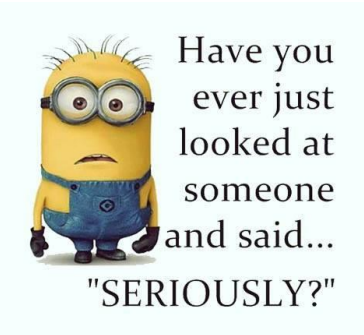
Resnica

- Data reduction max 2:1
  - Thin provisioning
  - Ni kompresije => baza
- Dve ločeni RAID-6 skupini
  - 16 * 500GB = 8TB useable RAW
  - Baza vzame 5TB useable RAW
  - Srv.virt 10TB => 3TB RAW
- Funkcionalnosti zmanjšajo prepustnost
- Izpad kontrolerja – zmanjšana prepustnost
- Max 5 letno vzdrževanje
  - „Write edurance limit“

REŠITEV #2

Podoba

- „Scale-out“ arhitektura
- 5TB RAW node
- Data reduction (5:1)
- Visoka razpoložljivost
  - Dva kontrolerja
  - RAID-6
- Razširljivost do 160TB RAW



Resnica

- Metapodatki v RAM-u
- Data reduction max 3:1
- Razširljivost „scale-out“
  - Korak +5TB, potem 10 TB
  - Max 4x 10TB
  - 160TB zahteva 4*40TB – fork-lift upgrade
- Visoka razpoložljivost
  - Odpoved node-a => izpad celotnega sistema
  - Dolgotrajna inicializacija sistema po izpadu
- Izpad kontrolerja – prekinitev ter začasna zmanjšana prepustnost

REŠITEV #3

Podoba

- Hibridni „unified“ sistem
  - „Tiered arhitektura“
  - NL-SAS,15k,SSD,cache
- IOPS 40.000
  - 95% zahtev serviranih z minimalno zakasnitvijo
- Visoka razpoložljivost
  - Dva kontrolerja
  - RAID nabor 1, 10, 5, 6, ...
- Razširljivost – dodajanje diskov
- Post-production data reduction

Resnica

- Zmogljivost
  - Preobremenjenost sistema
  - Ni narejen za različne aplikacije
  - Ni omejitve obremenitve kontrolerjev
  - Stiskanje še poslabša performance
- Razširljivost omejena s prepustnostjo in max.diski
- Izpad kontrolerja – začasna (ali celo trajna) prekinitev, zmanjšana prepustnost
- „Failure“ domena !

IZBIRA NI LAHKA

- Ne verjemite tehničnim specifikacijam (lepotno tekmovanje)
- Klasični sistem z SSD/Flash != All-flash sistem
- Mio IOPSi pri 4/8KB blokih
- Praksa = 32kB+ bloki
- Konstantna nizka zakasnitev pri dostopu do podatkov

TESTIRANJE JE OBVEZNO!

- Ne testirajte s sintetičnimi testi ter testnimi orodji proizvajalcev (~~iometer, SQLIO,FIO,...~~)
- Ne testirajte sintetičnih robnih pogojev
- POZOR - potrebno je celovito testiranje
- Pravi, živi podatki
- Prave, žive aplikacije (backup, anitvirus, VM restart, baze, ...)
- Visoka razpoložljivost z vseh vidikov
  - Dual parity, z vsemi funkcionalnostmi, izpad kontrolerja, strežnika, obremenitve, funkcionalnosti, ...
- Upravljanje (kreiranje, spremembe, nadgradnje, ...)
- !!! Testirajte odzivnost VAŠIH aplikacij !!!

"STISLJIVOST" PODATKOV PRIMER #1

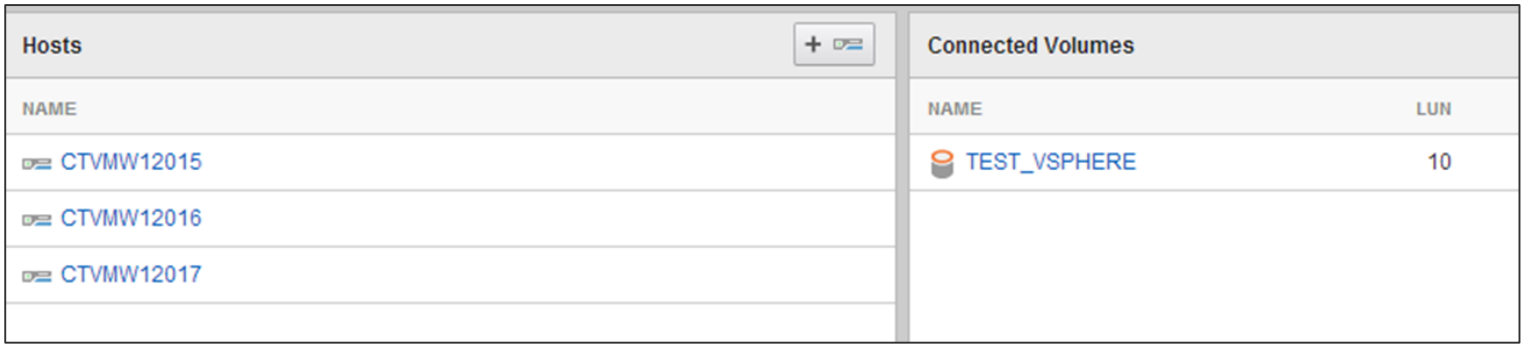
- Okolje Oracle podatkovne baze
  - Povprečno razmerje = 6,95:1 (raw vs. useable)

Volume Summary

NAME	PROVISIONED	VOLUME DATA	SNAPSHOT DATA	REDUCTION
TESTDSK_PURE_Log1	10 GB	523.00 KB	0 GB	10.1 to 1
TESTDSK_PURE_Log2	10 GB	870.00 KB	0 GB	9.1 to 1
TESTDSK_PURE_Log3	10 GB	628.00 KB	0 GB	9.7 to 1
TESTDSK_PURE_Log4	10 GB	721.00 KB	0 GB	9.6 to 1
TESTDSK_PURE_Table1	1536 GB	248.87 GB	0 GB	4.2 to 1
TESTDSK_PURE_Table2	1536 GB	246.14 GB	0 GB	4.2 to 1
TESTDSK_PURE_Table3	1536 GB	241.04 GB	0 GB	4.3 to 1
TESTDSK_PURE_Table4	1536 GB	235.02 GB	0 GB	4.4 to 1

“STISLJIVOST” PODATKOV PRIMER #2

- Okolje VMware vSphere
  - Predstavljeno = 1,27 TB, uporabljeno = 32,01 GB
  - Stisljivost (raw used vs. useable) = 8:1
  - Stisljivost + “thin provisioning” = 16.4:1



STISLJIVOST PODATKOV
PRIMER #3

FS-esxi55-clone-01	64 TB	96.49 MB	0 GB	11.9 to 1
FS-esxi55-clone-02	64 TB	50.26 MB	0 GB	6.9 to 1
FS-esxi55-ISO-01	10 TB	14.96 GB	0 GB	3.7 to 1
FS-esxi55-nfv-01	2 TB	344.48 MB	0 GB	10.9 to 1
FS-esxi55-sql-01	3 TB	56.69 GB	0 GB	7.3 to 1
FS-esxi55-sqlbkp-01	1 TB	2.31 GB	0 GB	9.9 to 1
FS-esxi55-sqllog-01	500 GB	13.00 GB	0 GB	2.8 to 1
FS-esxi55-srv-01	5 TB	20.00 KB	36.27 MB	12.0 to 1
FS-esxi55-srv-01-clone1	5 TB	73.00 KB	0 GB	12.0 to 1
FS-esxi55-srv-02	5 TB	4.77 GB	0 GB	11.1 to 1
FS-esxi55-srv-03	5 TB	196.94 GB	0 GB	3.5 to 1
FS-esxi55-vdi-01	10 TB	13.81 MB	0 GB	4.4 to 1
FS-esxi55-vdi-02	10 TB	13.80 MB	0 GB	4.4 to 1
FS-multimediaRDM-01	100 GB	61.92 GB	0 GB	1.5 to 1
FS-multimediaVMFS-02	120 GB	2.48 GB	0 GB	9.5 to 1
FS-spesxi01-boot	10 GB	69.02 MB	0 GB	4.1 to 1
FS-spesxi02-boot	10 GB	19.00 MB	0 GB	8.1 to 1
FS-spesxi03-boot	10 GB	16.46 MB	0 GB	8.4 to 1
FS-spesxi04-boot	10 GB	18.45 MB	0 GB	8.1 to 1
FS-spesxi05-boot	10 GB	14.30 MB	0 GB	8.7 to 1
FS-veeambrs-01	20 TB	59.89 MB	0 GB	2.0 to 1

Če sem govoril prehitro in preveč ...

ČE VAS ŠE KAJ ZANIMA ...

- me poiščite na NIL-u
- pa tudi na 


nil.com