Facultat d'Informàtica de Barcelona
Departament d'Arquitectura de Computadors
Centres de Processament de Dades

Activitat EBH

Emmagatzematge, *backup* i *housing*

Bru Carci, Jordi
Rojas, Carlos

Escenari ANA2

Data: 21/10/2022

1.-Descripció bàsica

TAULA 1: ESCENARI ORIGINAL: EXTRET DE L'ENUNCIAT. OMPLIU EL QUE HI HA EN GRIS.	
Nombre de Us	96U
Alçada Rack (en Us)	42U
Consum	202,8kW
Sobreprovisionament d'electricitat	7%
Nombre de servidors	22
Diners Totals	€10.000.000,00
Diners gastats	€7.500.000,00

taula 2: Elements que escolliu vosaltres	
Elements de mirror i backup	
GB a emmagatzemar al backup	76740
Dies entre 2 backups	1
Còpies senceres a mantenir	10
Opció Backup (1=M-A; 2=MS3; 3=Cintes)	3
Opció Mirror (0=NO; 1=SI)	0
Sistema de backup on-site? (0=N=; 1=SI)	1
Elements de housing	
Opció escollida (1:MOCOSA, 2: CPDs Céspedes, 3: Mordor)	3
Gestió local de *backup*? (0=No, 1=Si)	1
Monitorització? (0=NO; 1=SI)	1
Bandwidth provider	
Tipus de línia (1:10Mbps; 2:100Mbps; 3:1Gbps; 4:10Gbps; 5:100Gbps)	4

Número de línies agregades	2
Segon proveïdor? (0=NO, 1=SI)	1
SAN? (0=no, 1=8Gbps, 2=16Gbps, 3=32Gbps, 4=64Gbps, 5=128Gbps)	3
Cabina de discos	
Opció Disc principal (Entre 1 i 10)	9
Nombre de discos a comprar	12
Opció cabina de discos (Entre 1 i 6)	4
Nombre de Cabines	1
Cabina de discos 2 (cas de fer servir dos tipus)	
Opció Disc (Entre 1 i 10)	9
Nombre de discos a comprar	14
Opció cabina de discos (Entre 1 i 6)	4
Nombre de Cabines	1
Cabina de discos 3 (cas de fer servir tres tipus)	
Opció Disc (Entre 1 i 10)	0
Nombre de discos a comprar	0
Opció cabina de discos (Entre 1 i 6)	0
Nombre de Cabines	0

TAULA 3: OPEX	anual	cinc anys
Consum energètic (hardware només)	€38.738,92	€193.694,61
Empresa de Housing escollida	Mordor	
Cost Housing (inclou electricitat addicional)	€103.810,84	€519.054,19
Off-site: empresa escollida	Take the tapes and run	
Cost mirror	€0,00	€0,00
Cost backup	€97.200,00	€486.000,00
Cost Bandwidth provider	€21.168,00	€105.840,00

TAULA 4: CAPEX	Cost
Diners gastats en servers, xarxa, etc	€7.500.000,00
SAN	€361.706,00
Sistema emmagatzematge	€50.170,00

TAULA 5: AJUST AL PRESSUPOST	
Opex a 5 anys, total	€1.304.588,80
Capex a 5 anys, total	€7.911.876,00
Despeses totals a 5 anys	€9.216.464,80
Diferència respecte al pressupost	€783.535,20

2.-Anàlisi de necessitats

2.1- Número de GB a emmagatzemar (en cru).

Tenim 22 servidors de càlcul i cada servidor guarda un màxim 1TB de dades al disc centralitzat. Ademés, segons l'enunciat, necessitem 10 TB per les dades històriques. Per tant necessitarem en cru, i com a mínim **32 TB** (~32000 GB).

2.2- Velocitat requerida del sistema de disc (IOPS).

Número de Servidors = 22
Percentatge d'accessos al disc. = 25%
Tràfic mitjà = 1,5 Mbps.
Tràfic màxim= 100 Mbps.
Cada operació al disc és de 4KB.
Ens interessa, considerar sempre els valors en els pics per tenir un major control.
Velocitat màxima d'accés al disc = 0,25 * 100Mbps = (0,25 * 100 * 1024)/8 = 3200 KBps.
De màxima = 3200 KBps / 4KB per operació = **800 IOPS**.
La velocitat requerida del sistema de discs és de 800 IOPS.

2.3- Tràfic amb el client (entre servers i de server a switch de connexió a xarxa):

Considerem el tràfic màxim (pic) que és de 100Mbps i tenim 75% d'accessos entre servidors.
(0,75 * 100) => **75 Mbps.**
Podem arribar a tenir un tràfic de 75 Mbps amb el client.

2.4- Tràfic amb el disc:

Seguim considerant el tràfic màxim (pic) de 100 Mbps i tenim 25% d'accesos entre servidors.
(0,25 * 100) => **25 Mbps.**
Podem arribar a tenir un tràfic de 25 Mbps amb el disc.

2.5- Pressió sobre la xarxa (ample de banda mínim necessito per servir el tràfic de client i disc). M'arriba?:

Cal considerar que tenim una xarxa de 2 Gbps i que treballem en tràfic màxim. Per tant, agafem els valors aconseguits del tràfic amb el client i amb el disc.

Tràfic amb el disc => 25 Mbps
Tràfic amb el client => 75 Mbps
Tràfic total = **100 Mbps**
100 Mbps < 2Gbps

Quan succeeixen pics de tràfic, la xarxa no es veurà sobrecarregada ja que tenim bastant marge. No és necessari augmentar la capacitat de la xarxa.

3.-Decisions preses

3.1- Descripció dels elements d'emmagatzematge escollits, en funció de les necessitats. Quants tipus de cabines? (i perquè), RAID escollit a cadascuna d'elles. Nombre de cabines de cada tipus

										RAID 0	RAID 10	RAID 5	RAID 51	RAID 6	RAID 61
IOPS requerides	800									Dades bàsiques					
% escriptures (0-1)	0,3								IOPS necessaris	800	1040	1520	2480	2000	3440
Mida dades (GB)	32768								IOPS read	560	560	1040	1520	1280	2000
									IOPS write	240	480	480	960	720	1440
									Discs mínims	2	4	3	6	4	8

Amb els resultats obtinguts en la secció anterior podem veure que tenim bastant marge en termes de IOPS. El factor que hem de tenir present a l'hora d'escollir la quantitat de discos és la mida de les dades.
Com de moment tenim també un gran marge en pressupost podem optar a varies opcions. La que més convenç seria instal·lar dues cabines una per les dades intermitges i l'altre per les dades històriques. D'aquesta manera, per temes de logística i seguretat , no perdríem totes les dades en cas d'una fallada del sistema de disc.

Per evitar temps de downtime, creiem adient l'ús de hot spare disk. Per tant podem descartar les opcions de cabina 1 i 6.

Finalment, decidim la cabina 4 degut a que amb el capital que disposem, ens podem permetre dimensionar les instal·lacions de manera que fins i tot, tenir badies lliures, en surt rentable. D'aquesta menra podem començar a pensar en un futur creixement i assegurar-nos escalabilitat.

També hem vist la oportunitat d'implementar un raid 51 en les dues cabines, ja que segons el nostre cas, ens interessa accelerar qualsevol recuperació necessària. A més a més, en cas de corrupció de dades, el raid 51 ens beneficia totalment.

En un principi, la idea era utilitzar discos HDD però alfinal, com la nostre intenció és reduïr el temps de recuperació de backups, hem prioritzat un augment dels IOPS. És per això que hem acabat optant per SSD en les dues opcions.

Procedim a discernir entre dades intermitges i històriques:

- **Dades intermitges:** cabina de model 4 on hi instal·larem un raid 51 amb 12 discs d'opció 9 (WD Gold S768T1D0D (Enterprise)). D'aquests 12 discs, 2 d'ells adquereixen la funció de spare disk com havíem mencionat. De les 36 badies que ens proporciona la cabina, 24 resultaran lliures per un futur. Amb aquest sistema, tenim una capacitat de 42,24 TB utilitzables.

  Dades intermitges = capacitat_opció9 * (discs - disc_paritat(RAID51)) / 2 = [76880(12-1)] / 2 = 42,24 TB

- **Dades històriques:** cabina de model 4 on hi instal·larem un altre raid 51 amb 14 discs d'opció 9 (WD Gold S768T1D0D (Enterprise)). D'aquesta manera, la nostre idea és també tenir un marge sobreprovisionament amb les badies lliures. Dels 14 discos, 2 són també spare disk. Amb aquest sistema, tenim una capacitat de 34,6 TB utilitzables.

  Dades històriques = capacitat_opció9 * (discs - disc_paritat(RAID51)) / 2 = [76880(10-1)] / 2 = 34,5 TB

3.2- Es justifica la necessitat d'un SAN? Si la resposta és si, raonar si el cost és assumible o no, i cas de no ser-ho calcular l'impacte sobre el rendiment del CPD

trafc client no neecesari pero trafic backup es beneficia significativamt

Com veurem a l'apartat 3.7, el tràfic a la hora de fer backup pot arribar a saturar la nostre xarxa. És per això que creiem convenient la implementació d'una SAN al nostre sistema ja que d'aquesta manera, la càrrega del tràfic sortiria beneficiada significativament.

3.3.- Posem un *mirror*?

Finalment, hem decidit no contractar cap mirror i fer backups off-site. Hem de tenir en compte que estem tractant amb moltes dades i que fem backups diaris. Si només incloguessim les dades historiques en el mirror seria una idea, però tenir tantes TB de disc en el mirror no surt tan rentable.

3.4- Empresa de *housing* escollida i perquè (relació entre el que ofereix, el que necessito i el que costa)

En termes de housing, com a empresa, creiem oportú prioritzar la màxima fiabilitat possible i sobretot una transparència d'alt nivell per saber en tot moment l'estat del nostre sistema. És per això que escollim l'opció 3 (Mordor) ja que està certificada com a tier 3. D'aquesta manera ens assegurem la fiabilitat i transparència mencionada. A més a més, Mordor ens assegura a l'any, 1,6 hores de downtime així que ens estalviaríem uns diners que amb les altres opcions hauríem de dedicar.

3.5- Posem monitorització?

Si, ja que la monitorització ve inclosa en la nostre opció de housing. A més a més, per reduir el downtime al mínim, ens interessa que se'ns proporcioni canvi de hardware (e.g. canvi de spare disk).

3.6- Opció de backup?

Opció de backup on-site proporcionada per la nostre empresa de housing i una de dades off-site en cas de que se'ns impossibiliti la recuperació de les dades on-site.

Com a opció de backup optem per la empresa *Take the tapes and run* ja ens és interessant fer copies de seguretat en cintes degut al seu baix cost. També, la nostre idea és fer backups totals diaris i guardarem 10 copies de backups totals ja que només necesitem dades històriques en un lapse de temps de 10 dies. És per això que al haver-hi 76,74 TB a guardar i/o recuperar la opció de cintes ens és més atractiu. A més a més, la empresa que s'encarrega d'això desprèn molta fiabilitat ja que ens ens garanteix un bon funcionament i s'encarregaran de la major part de les coses. Com a empresa, és un factor important que també hem de considerar.

3.7- Tràfic amb l'exterior afegit pel sistema de *backup*/*mirror* escollit. Quin *bandwith* caldria?

Backup on site:

Si tenim present el tràfic del nostre sistema amb el exterior, el tràfic més carregat correspon a la restauració d'un backup. La càrrega de la xarxa a l'hora de fer el backup onsite és de 5 TB/hora.
Velocitat restauració de les cintes: 5TB/h.
5TB/h * 1000 * 8 * 1H/ 3600s = 40000 Gb/h /3600s = 11,11 Gbps.
La càrrega de la xarxa afegida per la restauració d'un backup serà de 11,11 Gbps

IOPS de lectura del nostre disc = 467000
467K IOPS * 4KB/IO = 1868000 KB/s = 1.87 GB/s = 14,94 Gbps.
La càrrega de la xarxa afegida al realitzar un backup serà de 14,94 Gbps.

Un cop visualitzades les dades resultats, com ja hem mencionat a l'apartat 3.2, creiem recomanable implementar al nostre sistema una SAN i augmentar la capacitat de la nostres connexions internes i externes per a que siguin suficient per a poder suportar la càrrega mencionada. Recomanaríem si és possible, augmentar la xarxa interna a 20 Gbps.

4.-Recomanacions als inversors

4.1.- Anàlisi de Riscos (*Risk Analysis*)

Quines desgràcies poden passar i com les hem cobert?
Al menys s'han de cobrir els següents casos:

• **Hi ha pèrdua d'un fitxer (per error o corrupció). De quan puc recuperar versions?**
Com fem un backup diari total, es pot recuperar la versió del dies anteriors fins a 10 dies. A més, utilitzarem snapshots degut a que tenim molt espai adicional per recuperar dades de fa menys d'un dia.

• **Es trenca un disc (es perden dades? quan trigo en recuperar-me? el negoci s'ha d'aturar?)**
Com tenim spare disk en les dues cabines, aquest problema és prevenible. D'aquesta manera, evitem haver d'aturar el negoci i així assegurar continuïtat.
En el cas de que no s'hagi pogut preveure a temps, com tenim implementat un RAID51, en la majoria de casos, fer una recuperació de disc no implicarà una reconstrucció: hi ha una copia identica del disc gràcies al Raid 1.

<u>Copiar la totalitat d'un disc serán:</u>
IOPS: capacitat_disc KB / (4KB d'operació) =
(7680 GB * 1000 * 1000)KB / (4KB/io) = 1920000000 IOPS
1920000000/65000 IOPS = 29538 segons = 8,2 hores.
Copiar la totalitat d'un disc tardaria 8,2 hores.

<u>Reconstrucció de disc:</u>
El temps de reconstrucció habitual és de 4 hores per TB de disc.
Cada disc té una capacitat de 7.68 TB.
4h/TB * 7.68TB = 30.72h.

En el cas de reconstrucció de disc, no s'atura la màquina. Per tant, per calcular el downtime que tindrem per culpa de discs trencats, assumirem els casos on realitzem una copia sense hot spare disk.

<u>Temps Downtime:</u>
El temps de downtime es calcula amb:
downtime_restaurar_disc * (1-prob_fall_detectada_smart) * prob_fall_ssd_anual * num_discs
30% fallada no previnguda per SMART
0,45% Fallada de SSD TLC anual.

Downtime dades intermitges:
8,2h * 0,3 * 0,0045 * 12 * 60 = 7,97 min per any
8,2h * 0,3 * 0,0045 * 12 * 5 anys * 60 = 39,85 min per 5 anys

Downtime dades històriques:
8,2h * 0,3 * 0,0045 * 10 = 6,64 min per any
8,2h * 0,3 * 0,0045 * 10 * 5 anys = 33,2 min per 5 anys

Total:
14,61 min en un any
1h13 min en 5 anys

• Puc tenir problemes de servei si falla algun disc?
En teoria no hi hauria d'haver cap problema en el nostre servei ja que hem implementat varies solucions en cas de fallada. L'ús de spare disk, backups i dels RAID 51, ens proporcionen seguretat en vers a fallades de disc.
Per arribar a tenir algun problema, hauriem d'imaginar l'escenari on, apart de la fallada de disc, es produeix corrupció de dades. A conseqüència, s'hauria de reconstruir i això produeix una reducció de la velocitat dels sistema en dos. Però, com excedim significativament la càrrega que rebem per clients, això no suposa cap problema pel servei.

• Cau la línia elèctrica. Què passa?
Si es produeix una caiguda de la xarxa elèctrica, es calcula una aturada de menys d'una hora un cop o cada dos anys, una de entre 1 i 6 hores cada 4 anys. Les línies d'alta tensió industrials estan preparades per garantir un màxim de 6 hores per reconnexió.
Però com tenim la idea de contractar a la empresa de housing mordor ens assegurem que es redueixi el downtime anual a 1,6 hores. A més, tindríem una segona línia d'entrada d'electricitat i ademés un generador diesel amb capacitat per aguantar la potència pic durant 72 hores.

• Cau una línia de xarxa. Què passa?
La empresa Mordor també ens proporciona una segona línia de xarxa que ens assegura una solució a una caiguda d'ella. A més a més, com tenim els suficient capital, hem decidit que contractar un segon proveïdor de la xarxa en un cas d'emergència és una bona idea. Les dues línies de xarxa contractades són de 10 Gbps i en el cas de que una caigui, afectaria en la velocitat de creació i restauració d'un backup.

- **En cas de pèrdua o detecció de corrupció de dades no ens podem permetre seguir treballant fins que recuperem les dades correctes. Calculeu temps i costos de recuperació en cas de**
  - **Pèrdua/ corrupció d'un 1% de les dades**

    Com hem sobredimensionat l'emmagatzamatge del nostre servidor de disc, cal mencionar que segons l'enunciat les dades totals del sistema són de 32 TB en lloc de 75 TB:

    - 10 TB per les dades històriques
    - < 22 TB de dades intermitges.

    Si la pèrdua té lloc al sistema Raid de dades historiques:

    10 Discs * (65000 IOPS W/4 W reals) = 162500 W/s * 4 KB/W = 650000 KBps = 650 MBps

    temps de recuperació = 10TB * 0,01 / 650 MBps = 2,56 minuts

    Si la pèrdua té lloc al sistema Raid de dades intermitges:

    12 Discs * (65000 IOPS W/4 W reals) = 195000 W/s * 4 KB/W = 780000 KBps = 780 MBps

    temps de recuperació = 22 TB * 0,01 / 780 MBps = 4,7 minuts

    Si la pèrdua és global:

    temps de recuperació total = 4,7 minuts
  - **Pèrdua/ corrupció de la totalitat de les dades**

    Si la pèrdua té lloc al sistema Raid de dades historiques:

    10 Discs * (65000 IOPS W/4 W reals) = 162500 W/s * 4 KB/W = 650000 KBps = 650 MBps

    temps de recuperació = 10 TB / 650 MBps = 4,27 hores

    Si la pèrdua té lloc al sistema Raid de dades intermitges:

    12 Discs * (65000 IOPS W/4 W reals) = 195000 W/s * 4 KB/W = 780000 KBps = 780 MBps = **6,24 Gbps**

    temps de recuperació = 22 TB / 780 MBps = 7,84 hores

    Si la pèrdua és global:

    temps de recuperació total = 7,84 hores

Aquests temps són suposant que s'ha instal·lat una SAN de per lo menys 6.24 Gbps.
Si seguissim amb la Xarxa LAN de 2 Gbps, seria el nostra bottleneck:
Trigariem 32 TB / 2 Gbps = 1.48 dies.

4.2.- Anàlisi de l'impacte al negoci (*Business Impact Analysis*)

En funció de l'anàlisi de riscos anterior i del que costa estar amb la màquina aturada o no donar el servei complert, calcular quant perdo en diners per tenir-lo aturat i quant em costaria evitar aquesta situació.

Caiguda de la xarxa de dades:

Evitar aquesta situació no suposaria un gran esforç ja que amb el housing contractat ens assegurem un baix downtime.

Mordor ens garanteix solucions eficients a possibles caigudes de la xarxa i és per això que com a màxim, només arribarem a tenir 1.6 hores de downtime a l'any. Tenint en compte que al nostre SLA, hem de pagar 150.000€ per hora de downtime, deuríem uns 240.000€ a l'any. Si ho plantejem en un període de 5 anys, resultarien ser 1.200.000€ d'euros. Encara que a primera vista semblin quantitats de diners molt elevades, altres empreses de housing ens donen pitjors resultats i ens garanteixen menys solucions.

Fallada de disc

Hem de tenir present que l'ús de SMART per preveure una caiguda sol tenir èxit el 70% dels casos i que els discos plantejats en les nostres cabines de dades, tenen un 0,45% de fallada anualment.

Evitar una fallada de disc no ens és un gran problema ja que hem plantejat varies solucions que permeten preveure o solventar una fallada de disc. En primer lloc, hem plantejat l'ús de hot spare disk per fer un canvi de disc rapid juntament amb SMART que preveu una possible caiguda. A més a més, amb el RAID 51 implementat en les dues cabines de dades, garantim una recuperació ràpida de les dades del disc que ha fallat.

En l'Apartat 4.1 hem calculat el Downtime de 5 anys per culpa de discs i estimem que seria de 1 hora i 13 min si tenint en compte el conjunt de dades del nostre sistema.

4.3.- Creixement

Si creix el nombre de clients/ màquines/ dades (depèn de l'escenari), hem d'estar preparats.

Primer de tot, caldria mencionar que com tenim un alt pressupost, de primera mà, hem preparat un sobreprovisionament amb bona escalabilitat per garantir un bon creixement. És per això que amb les implementacions documentades tenim moltes badies lliures. Tot i així amb tot el sobreprovisionament, podem veure com encara ens sobra capital per gastar.

Realment, el nostre espai de dades també està pensat per tenir un bon marge. Tenim ocupades un terç de la capacitat de dades històriques i la meitat de les dades intermitges. És per això que si pensem en creixement, la idea seria afegir discos per augmentar el número de servidors de càlcul. Com no sabem com creixen les dades històriques no podem comentar sobre les limitacions actuals, però tenim sobreprovisionament per 20TB addicionals.

Si hem de pensar quin recurs es pot esgotar abans, podem dir amb total certesa que la LAN seria la primera en tocar fons. És per això que en la nostre idea de sobreprovisionament des d'un principi ja contempla la implementació d'una SAN.

4.4.- Inversions més urgents

Caldria augmentar la capacitat de la nostra xarxa interna a per lo menys 20 Gbps per sostenir la càrrega dels nostres backups.

En el nostre cas, degut a la poca pressió de recursos i a l'alt pressupost, hem decidit plantejar-nos quina seria la millor manera de planificar un CPD que pugui suportar el necessari i a més a més preparar un sobreprovisionament per un futur creixement.

Com asumim que la empresa creixerà, hem plantejat cabines preparades amb badies lliures per a que un futur puguin ser omplertes amb els discs necessaris. D'aquesta manera, no caldrà invertir més diners en noves cabines o cabines més grans. Segon el model de negoci, creiem oportú seguir la planificació documentada en aquest informe. Per tant, si cal expandir la capacitat dels nostres sistemes, únicament s'haurà d'invertir en els discos ja documentats de cada cabina.

Els discos documentats son més que suficients, així que si es dóna el cas de que s'ha de comprar nous discos per augmentar la capacitat total, es recomana invertir els diners en omplir les badies lliures pels discos SSD documentats.