Facultat d'Informàtica de Barcelona

Departament d'Arquitectura de Computadors

Centres de Processament de Dades

Activitat EBH

Emmagatzematge, *backup* i *housing*

Cuscó Sendra, Pau

Lladó Cortés, Nacho

Escenari 08: ANA2

Data: 13/10/2023

Índex

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	14000
Dies entre 2 backups	2
Còpies senceres a mantenir	8
Opció Backup (1=M-A; 2=MS3; 3=Cintes)	2
Opció Mirror (0=NO; 1=SI)	1
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)	2
Cabina de discos	
Opció Disc principal (Entre 1 i 10)	10
Nombre de discos a comprar	22
Opció cabina de discos (Entre 1 i 6)	2
Nombre de Cabines	1
Cabina de discos 2 (cas de fer servir dos tipus)	
Opció Disc (Entre 1 i 10)	14
Nombre de discos a comprar	10
Opció cabina de discos (Entre 1 i 6)	2
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)	€37.788,72	€188.943,62
Empresa de Housing escollida	Mordor	
Cost Housing (inclou electricitat addicional)	€103.668,31	€518.341,54
Off-site: empresa escollida	MonsoonS3 MS3	
Cost mirror	€13.308,00	€66.540,00
Cost backup	€31.635,00	€158.175,00
Cost Bandwidth provider	€21.168,00	€105.840,00

TAULA 4: CAPEX	Cost
Diners gastats en servers, xarxa, etc	€7.500.000,00
SAN	€215.706,00
Sistema emmagatzematge	€57.500,00

TAULA 5: AJUST AL PRESSUPOST	
Opex a 5 anys, total	€1.037.840,16
Capex a 5 anys, total	€7.773.206,00
Despeses totals a 5 anys	€8.811.046,16
Diferència respecte al pressupost	€1.188.953,84

2.-Anàlisi de necessitats

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

22 servidors * 1 TB per server + 10 TB dades històriques = *32 TB*

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

22 servidors * 1,5 Mbps * 0,25 * 1000 Kbps * 0,125 KB = 1031,25 KB/s (velocitat de disc)

(1031,25 KB/s) / (4KB/IO) = *257,8125 IOPS ~ 258 IOPS*

Però si considerem que tenim pics de 100Mbps i que el temps de *downtime* en el nostre cas ens perjudica greument, calculem amb les dades pic:

22 servidors * 100 Mbps * 0,25 (25% operacions amb disc) * 1000 Kbps * 0,125 KB = 68750 KB/s (velocitat de disc en pic)

(68750 KB/s) / (4KB/IO) = *17187,5 IOPS ~ 17188 IOPS*

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

1,5 Mbps de tràfic mitjà del qual un 75% és entre servidors (LAN), per tant:

22 servidors * 1,5 Mbps * 0,75 = *24,75 Mbps*

En el pitjor cas, tots els servidors (22) tindran un pic de tràfic de 100 Mbps i, per tant:

22 servidors * 100 Mbps * 0,75 = *1650 Mbps*

2.4- Tràfic amb el disc:

1.5 Mbps de tràfic mitjà del qual un 25% és de disc, per tant:

22 servidors * 1,5 Mbps * 0,25 = *8,25 Mbps*

En el pitjor cas, tots els servidors (22) tindran un pic de tràfic de 100 Mbps i, per tant:

22 servidors * 100 Mbps * 0,25 = *550 Mbps*

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

La pressió total a la xarxa és de 1650 Mbps + 550 Mbps = 2200 Mbps -> 2,2 Gbps.

No ens arriba perquè la nostra xarxa és de 2 Gbps. Aquest càlcul és considerant que tots els servidors estan treballant en el tràfic de pic, el cas extrem, cosa que és força poc probable que passi. Majoritàriament, podrem abastir en tot moment la pressió sobre la xarxa amb la connexió que tenim.

L'anterior càlcul l'hem fet considerant que les transmissions amb el disc s'efectuen mitjançant la xarxa ethernet (LAN).

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

Dades volàtils:

RAID 10 -> doble de discs però doble velocitat

22TB + 22 TB = 44 TB -> +40% de sobredimensionament = 57,2 TB aprox

17188 IOPS -> +40%sobreprovisionament = 24063 IOPS

16 discs de 3,8 TB (disc 10) -> 8 en raid 0 i duplicats (raid 1), en total 16. Comptabilitzant els *spare disks,* necessitarem comprar 20 discs.

Els posarem en el model de cabina 2 que consta de 22 badies sense cache donat que els nostres discs són prou ràpids. Els configurarem en dos clústers un de 5 discs i l'altre de 4, tots ells en RAID 10, així diversifiquem la capacitat de lectures i escriptures simultànies. També omplim els 4 espais per *spare disks*. Triem RAID 10 donat que les dades que movem estan canviant constantment i ens interessa un processat ràpid. Si un disc es trenqués, és molt ràpid i eficient anar a buscar les dades al disc replicat.

Dades històriques:

10 TB + 40% de sobreprovisionament = 14 TB

14 TB + 3,8 TB (disc striping RAID5) * 2 (RAID1)

RAID 51 -> 4 discos de dades i 1 de paritat, tot duplicat (discos de 3,8 TB tipus 10)

3,8 TB * 4 = 15,2 TB tenim espai de sobres

no tenim necessitats de IOPS específics

cabina tipus 2 més que res per *spare disks* i SLA *downtime*

La nostra necessitat de dades històriques és de 10 TB i per tenir sobreprovisionament contarem 14 TB (+40%). Com que són dades molt importants que no es poden perdre en el temps que les mantenim muntarem un sistema RAID 51 amb discs de tipus 10 de 3,8 TB. D'aquesta manera amb quatre discs per dades i un de paritat podrem tenir el sistema RAID 5. Per a fer el RAID 51 només hem de duplicar el RAID 5, tenint un ús de 10 discs, també comparem 4 més per *spare disks*. No tenim cap especificació de IOPS, per tant, no ens n'hem de preocupar. Per triar cabina l'hem triat de 24 discs, ja que la de 12 no té *spare discs* i ens interessen especialment per evitar *downtime* perquè la nostra SLA és estricte.

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

Com que ens trobem que tenim una LAN de 2 Gbps i quan en fer el càlcul vam veure que el tràfic total ens sortia de 2,2 Gbps per tal de reduir el consum de línia i poder assegurar el correcte intercanvi d'informació.

De fet mirant-ho amb números veiem que si posem la LAN alliberarem 550 Mbps de tràfic de xarxa assolint així 1650 Mbps, on ja ens trobarem dins els marges imposats per la nostra línia.

Així mateix, cal destacar que aquest tràfic de xarxa es tracta d'un moment on tots els servidors es troben en un pic, però tenint en compte l'SLA que hem garantit als nostres clients, val la pena usar una SAN.

La nostra LAN és molt justa per les nostres necessitats, el problema ve quan hem de recuperar dades del *mirror*, la nostra xarxa interna es queda molt curta i el temps de recuperació és molt lent. Considerem doncs posar una SAN de 16 Gbps per tal de poder abastir els requeriments.

3.3.- Posem un *mirror*?

Creiem que sí que és necessari tenir un *mirror*. Gràcies a com funciona el sistema *mirror* sempre tindrem una còpia de les dades del disc en un espai *off-site* i això ens proporcionarà una gran estabilitat de cara a problemes que puguem tenir en què ens falli tot el sistema del disc. Així doncs, tenint en compte el contracte SLA que tenim el més segur és tenir un *mirror* de les nostres dades del disc.

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

A causa de dues raons principals, la primera que a l'SLA tenim que hem de pagar 150.000€ per cada hora de *downtime* i segon que tenim una quantitat de memòria a emmagatzemar molt reduïda i només ens ocupa dos racks, hem triat Mordor Colocation Center. Modular Containers S.A. queda descartat, ja que no ens ofereix cap Tier Certificat. CPDs és més econòmic que Mordor, però com que la seva certificació és d'un mínim de 99,749% de *uptime*, en el pitjor dels casos pagaríem 22 hores * 150.000€ = 3.300.000€, caríssim vaja. Per tant, la millor opció per a nosaltres és Mordor, ja que ens certifiquen un mínim de *uptime* d'1,6 hores l'any (240.000€ màxim).

3.5- Posem monitorització?

La monitorització del nostre sistema va inclosa en la tarifa de l'empresa de *housing* triada, Mordor.

3.6- Opció de backup?

Només farem *backup* de les dades històriques del nostre sistema. El gruix de dades del nostre escenari tenen una vida molt curta als nostres servidors i constantment estàn canviant, pel que fer-lis un *backup* seria poc útil donat que al cap de poc temps ja estaria completament desactualitzat respecte les dades del servidor. Les dades històriques, en canvi, ens demanen de mantenir-les un mínim de 10 dies. És per això que per emmagatzemar aquests, triem fer un *backup* cada dos dies i guardar-lo durant 10 dies i també guardar el *backup* dels últims 3 divendres (o dissabte, segons caigui). En total estem guardant 8 *backups*. El *backup* el farem amb l'empresa

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

El sistema de *mirror* ocuparà 550 Mbps de bandwith en pic amb l'exterior. El *mirror* pesa 46TB i ens interessa poder-nos recuperar molt de pressa en cas de pèrdua total de dades, per la qual cosa triem una connexió de 10 Gbps de dues línies agregades. Ens surt més barat pagar aquesta despesa que el cost del *downtime* donat el nostre SLA.

4.-Recomanacions als inversors

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

• Hi ha pèrdua d'un fitxer (per error o corrupció). De quan puc recuperar versions?

De fa dos dies, de fa quatre dies, de fa sis dies, de fa vuit dies, de fa deu dies, de fa 12 dies, 20 dies i 27 dies.

• Es trenca un disc, es perden dades? Quant trigo a recuperar-me? El negoci s'ha d'aturar?

En cas que se'ns trenqui un disc no perdrem les dades, ja que en el cas de les dades temporals estan en un RAID 10 i en cas de les històriques en un RAID 51. Tardo el temps de transferència de les dades d'un disc de RAID 5 o 0 al pertinent duplicat. Els discs de dades temporals tenen una capacitat màxima de 3,8 TB i uns IOPS d'escriptura quan es recupera de 41 K (la meitat), cada operació de disc és de 4 KB. Calculant 3800 GB/(41000 x 0,004 MB) = 23,17 s en recuperar un disc. No haurem d'aturar el negoci, ja que podrem fer servir l'altra còpia de les dades i la recuperació es pot fer en calent sobre un *spare disk* dels que consten a la cabina.

• Puc tenir problemes de servei si falla algun disc?

L'únic problema extrem seria si fallés el mateix disc dels dos RAIDs del sistema RAID 10. En aquest cas hauríem de tornar a efectuar les operacions per arribar a aquells resultats. Tenint en compte que la majoria de dades que movem tenen una temporalitat molt baixa, no és un problema que puguem evitar, donat que tot i que féssim *backups* freqüentment d'aquestes dades, es quedarien obsolets molt de pressa.

• Cau la línia elèctrica. Què passa?

El servei de *housing* triat, Mordor Colocation Center, consta de connexió a dues línies d'electricitat i a més a més tenen un generador Dièsel amb capacitat d'aguantar la potència pic 72h. Pel que si cau una línia no hauria de passar res. Si cauen les dues, el subministrament recauria sobre el generador. Ens garanteixen un *downtime* de màxim 1,6 h l'any.

- **Cau una línia de xarxa. Què passa?**

En cas que caigui la línia principal de xarxa tenim la línia secundària, si aquesta també fallés tenim un segon proveïdor. En cas que el problema sigui causat per la connectivitat amb el primer proveïdor, se'ns resoldria el problema. Si no, no podríem donar servei als nostres clients i pagaríem 150.000€ l'hora de *downtime* donat el nostre SLA.

- **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**

Tant si aquestes dades són dels discos de dades temporals o històriques, la seva capacitat màxima és de 3,8 TB i els IOPS d'escriptura de 82 K, cada operació de disc és de 4 KB. Calculant 3800 GB / (82000 x 0,004MB) veiem que podem recuperar el disc en 11,6 s sobre un *spare disk*. El cost de trencar el SLA durant 11,6 s és de 458.33€.

  - **Pèrdua/ corrupció de la totalitat de les dades**

En cas de perdre la totalitat de les dades, la recuperació l'hauríem de fer sobre la xarxa fent ús del *mirror* contractat. Les dades ocupen 45 TB al *mirror* considerant, quan les copiem als nostres discos les copiarem per duplicat. Tenim 9 + 4 discos, podem fer 82.000 operacions d'escriptura per 13 discos = 1.066.000 x 4KB/W = 4.264.000 KB/s → 4,26 GB/s velocitat màxima que permeten els discos. Hem de recuperar en total 45000 GB a un màxim de 4,26 GBps. La nostra xarxa LAN i SAN és el nostre coll d'ampolla, podem transmetre a màxim 130 Gbps. Triguem doncs 30800/(130/8) = 84507 s, 31 mins.

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

<u>Caiguda de la xarxa de dades:</u>

Tenint en compte el que se'ns diu a l'exemple que la probabilitat de *downtime* en el cas que tinguem dues línies contractades és d'entre 0,00034% i 0,00071%. Això s'arrodoneix en 1 h de *downtime* màxim en un termini de 5 anys. Tenint en compte el nostre SLA, pagarem un màxim de 150.000€.

Fallada de disc

Tenim dos sistemes muntats, un amb RAID 10 i l'altre amb RAID 51.

Les dades històriques les guardem amb un RAID 51, per tant: tenim 10 discos amb probabilitat de fallada 0,0284 i 30% per l'SMART = 8,5% de fallada a l'any. Com que tenim IOPS sobredimensionats, no ens passa res, ja que la recuperació és ràpida.

Les dades temporals estan configurades en RAID 10 dividit en tres clusters de 8 discs i un de 10. Fallarà quan els dos discos amb el mateix contingut d'un mateix clúster fallin alhora. Tenint en compte que el cas més probable és en el que tenim un clúster de 4 discs veiem que la probabilitat és de ¼ per a cada grup, aleshores ¼ x ¼ és 0,0625. Tenint en compte l'ús de SMART on reduïm en el 70% la probabilitat de fallada i veient que els discs tenen un 2,84% de probabilitat de fallar cada any veiem que: 0,7 x 0,0625 x 0,0284 = 0,0012425. Així doncs, aquesta és la probabilitat de fallada de disc cada any. En aquest cas anirem a buscar les dades en el mirror i com que els discs són de 3,8TB tardarem 38000/130 = 292 s tardarem gairebé 5 minuts a recuperar el disc, per tant, cada 5 anys: 0,0012425 * 43830 hores cada 5 anys = 54 hores.

4.3.- Creixement

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

Quin creixement (en nombre de clients, etc…) podem assumir sense canviar el sistema (sobreprovisionament)?

> Nosaltres hem previst fins a un mínim d'un 40% de creixement assegurat pel que fa a la capacitat de dades i de la xarxa amb l'exterior. La xarxa interna ens permet treballar amb normalitat mentre no hi hagi pics de treball. Però quan tots els servidors treballin en pic i demandin una velocitat de 2.8 Gbps sobre la xarxa LAN no es podrà abastir. És un cas poc probable, però sí que l'hem de tenir en compte.

Quin és el recurs que s'esgota abans?

> El recurs que se'ns esgota abans és l'amplada de banda de la nostra LAN, tot i tenir una SAN. El tràfic quan tots els servidors treballen en pic augmenta fins a 2.8 Gbps, velocitat que la nostra xarxa LAN no pot subministrar, ja que és de 2 Gbps. Es veuria lleugerament afectada la velocitat durant el curt temps que durés el pic.

Feu un informe de les implicacions que suposaria un increment d'un 20% en el volum de negoci (tot, clients, dades, …)

Si les dades incrementen en un 20% fent augmentar de 32 TB requerits a 38,4 TB. El nostre sistema no tindrà problemes per alocar les dades perquè pot suportar fins a 45 TB. Si les operacions sobre la xarxa augmenten un 20%, passarem d'ocupar 2,2 Gbps a 2.64 Gbps, provoca que estiguem fora de les possibilitats de la nostra xarxa LAN quan tots els servidors estan treballant en pic. Tot i que és una cosa improbable. El nostre sistema de discos i de xarxa amb l'exterior està suficientment sobredimensionat per a poder alocar el creixement de 20%.

4.4.- Inversions més urgents

Donada la naturalesa del nostre servei, tenim molts intercanvis de dades entre servidors. La connexió LAN de 2 Gbps que tenim actualment intercomunicant els nostres servidors es queda curta quan els servidors treballen en pic o quan volem recuperar dades del *mirror* o dels *backups*. Si augmentéssim considerablement les capacitats de la xarxa interna, ens podríem estalviar la contractació de la SAN, que són 220.520€ al cap de cinc anys. Si ho invertíssim en comptes en augmentar l'amplada de banda de la xarxa LAN, podríem també reduir costos a l'hora de recuperar per complet el sistema del *mirror*, ja que es podria fer molt més de pressa.

La resta de diners disponibles del pressupost es podrien destinar a ciberseguretat i a augmentar el nombre i velocitat dels servidors de càlcul.