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

Activitat EBH

Emmagatzematge, *backup* i *housing*

Donate Durán, Daniel
Malqui Cruz, Miguel Angel

Escenari 04 - CBR1

Data: 25/03/2021

1.-Descripció bàsica

TAULA 1: ESCENARI ORIGINAL: EXTRET DE L'ENUNCIAT. OMPLIU EL QUE HI HA EN GRIS.	
Nombre de Us	1100U
Alçada Rack (en Us)	42U
Consum	1233,kW
Sobreprovisionament d'electricitat	12%
Nombre de servidors	370
Diners Totals	€20.000.000,00
Diners gastats	€16.500.000,00

taula 2: Elements que escolliu vosaltres	
Elements de disc, mirror i backup	
GB a emmagatzemar	760000
Dies entre 2 backups	7
Còpies senceres a mantenir	2
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)	2
Gestió local de *backup*? (0=No, 1=Si)	0
Monitorització? (0=NO; 1=SI)	1
Bandwidth provider	
Tipus de línia (1:10Mbps; 2:100Mbps; 3:1Gbps; 4:10Gbps; 5:100Gbps)	3

Número de línies agregades	4
Segon proveïdor? (0=NO, 1=SI)	0
SAN? (0=no, 1=8Gbps, 2=16Gbps, 3=32Gbps)	0
Cabina de discos	
Opció Disc principal (Entre 1 i 10)	3
Nombre de discos a comprar	102
Opció cabina de discos (Entre 1 i 6)	5
Nombre de Cabines	3
Cabina de discos 2 (cas de fer servir dos tipus)	
Opció Disc (Entre 1 i 10)	4
Nombre de discos a comprar	476
Opció cabina de discos (Entre 1 i 6)	5
Nombre de Cabines	14
Cabina de discos 3 (cas de fer servir tres tipus)	
Opció Disc (Entre 1 i 10)	9
Nombre de discos a comprar	0
Opció cabina de discos (Entre 1 i 6)	6
Nombre de Cabines	0

TAULA 3: OPEX	anual	cinc anys
Consum energètic (hardware només)	€226.976,34	€1.134.881,72
Empresa de Housing escollida	CPDs Céspedes	
Cost Housing (inclou electricitat addicional)	€387.820,54	€1.939.102,69
Off-site: empresa escollida	Take the tapes and run	
Cost mirror	€0,00	€0,00

Cost backup	€31.514,29	€157.571,43
Cost Bandwidth provider	€3.024,00	€15.120,00

TAULA 4: CAPEX	Cost
Diners gastats en servers, xarxa, etc	€16.500.000,00
SAN	€0,00
Sistema emmagatzematge	€248.200,00

TAULA 5: AJUST AL PRESSUPOST	
Opex a 5 anys, total	€3.246.675,84
Capex a 5 anys, total	€16.748.200,00
Despeses totals a 5 anys	€19.994.875,84
Diferència respecte al pressupost	€5.124,16

2.-Anàlisi de necessitats

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

19 grups

40 TB per grup

760 TB = 760000GB

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

30 racks
8 MBps intercanvi amb el sistema d'emmagatzematge per rack
240 MBps = 240000 KBps
4 KB per operació IO
60000 IOPS

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

30 racks
4 MBps per rack
Tràfic exterior: 120 MBps = 960 Mbps

2.4- Tràfic amb el disc:
30 racks
8 MBps per rack
Tràfic interior: 240 MBps = 1920 Mbps

2.5- Pressió sobre la xarxa (ample de banda mínim necessito per servir el tràfic de client i disc). M'arriba?:
Tràfic total: 2880 Mbps
La xarxa és de 4Gbps. Sí arriba.

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

Observem que, d'alguna manera, tenim dos grups de dades: aquelles que són molt poc consultades (i que els clients dels grups de recerca accepten amb retràs), que suposen el 70% del total, i el 30% restant que, en principi, requerirà d'un servidor amb la velocitat requerida calculada anteriorment. Per tant, tindrem dues cabines, una per les dades que suposen aquest 30% del total, i altra per les del 70% (anomenem-les S_{30} i S_{70}, respectivament, per comoditat). Les dades de S_{70} les posarem en RAID5, ja que no ens podem permetre (ni ho necessitem per aquestes dades) replicar-les, ja que suposen una gran quantitat d'informació. En canvi, les dades S_{30} les posarem amb replicació, mitjançant RAID51.

Re-calculem Capacitat i IOPS del servidor S_{70} i de S_{30}

Servidor S_{70}:

- Capacitat = 0.7·(760) TB = 532 TB
- Escriptures = 0.5 (en principi el percentatge d'escriptures és igual pels dos grups de dades)
- IOPS = 0 (podem considerar que és 0, ja que amb la gran quantitat de TB que tenim obtenim una quantitat de IOPS que, sens dubte, serà suficient per complir amb les exigències d'aquest servidor).

Servidor S_{30}:

- Capacitat = 0.3·(760) TB = 228 TB
- Escriptures = 0.5
- IOPS = 30 racks x 8 MBps x (1 IO)/4KB = 60.000 IOPS (a efecte pràctic, estem actuant com si TOTA l'activitat es donés en aquest servidor).

Després d'avaluar les dades ens decidim per:

S_{70}: Opció 3 de disc (HDD Enterprise) en RAID 5. El mínim de discos requerits per RAID5 són 55 (obtenint, així, una quantitat de 0.5x(55x710)+0.5x(55x710/4) = 24.406 IOPS!), però per cobrir només les dades actuals, sense creixement.

Tenint en compte que hem de fer clústers pel RAID 5, suposem que fem clústers de 6 discos (5 de dades + 1 degut al RAID 5). Fariem 11 clústers de 5+1 discos, el que ens dona 66 discos. La cabina més gran és de 36 badies i voldré afegir discos de *spare* (puc afegir fins a 4). Es pot observar que es necessiten, almenys, dues cabines, però amb dues no tenim creixement possible. Per tant, escollirem tenir tres.

El mínim són 55 discos, mirem de tenir un marge de creixement, donat que el preu per disc no és tan alt. Una opció serien 80 discos, però farem clústers de 6 discos (5+1). Per tant, fem 15 clústers de 6 discos (90 discos, 75 per dades i 15 degut al RAID), o sigui, 5 clústers per cabina.

Escollim cabines de 36 badies amb SSD (per si en un futur proper, com es diu a l'enunciat, s'ha d'adaptar el disseny). Com ocupem 30 de les 36 badies, posem 4 discos de *spare*.

Total: 3 cabines tipus 5 i 102 discos de tipus 3 en RAID 5.

S_{30}: Opció 4 de disc (HDD Enterprise) en RAID 51. Seria ideal poder fer servir discos SSD, però tenim una quantitat tan gran de dades, que elevaria massa el pressupost, de manera que ens tornem a decantar per un model HDD Enterprise. Ens calen un mínim de 230 discos (el que ens resulta en una quantitat de IOPS de 0.5 x (230x3360) + 0.5 x (230x3360/8) = 434700 IOPS!). Novament, procurem tenir un marge de creixement, de manera que decidim crear 42 clústers de 10 discos (només 4 de cada 10 són de dades no-duplicades, o sigui, 168 discos per dades).

Per tant, ara tenim 3 clústers per cabina.

Tornem a escollir cabines de 36 badies amb SSD, posant 4 discos *spare* per cabina.

Total: 14 cabines tipus 5 i 476 discos de tipus 4 en RAID 51.

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

El tràfic total (extern + disc) és de 2,88 Gbps i la xarxa que tenim és de 4 Gbps. Per tant, només podem créixer un 138% el tràfic total sense saturar la xarxa, de manera que seria molt raonable, en particular si, com en aquest cas, "les condicions poden canviar en pocs mesos", estudiar el desviament del trànsit de la xarxa mitjançant una SAN, però com no tenim prou pressupost, hem de prescindir de la SAN.

3.3- Posem un *mirror*?

El cost del *mirror* és molt alt (més de 500.000€/any), però a canvi, lògicament, permet reduir molt el temps de recuperació de dades. Donat el fet que no tenim clients i, per tant, no tenim un SLA que ens obligui a pagar-los (encara que els grups de recerca poden perdre diners per hores d'inactivitat), considerem que la xifra és massa alta per fer un *mirror* del 100% de les dades.

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

Tenint en compte que les nostres dades son de grups de recerca i no tenim penalització economica ja que no depenem de clients, pero tenim un downtime màxim de 24h l'any, escogim Cespedes SL que és més barat que l'altre opció posible que és Mordor.

3.5- Posem monitorització?

Encara que només sigui perquè comprem cabines de disc amb spare disc, ens interessa monitorització i que ens vigilin quan la tècnica SMART avisi i ens canviïn els discos. Per tant, escollim monitorització.

3.6- Opció de backup?

Considerem que per la natura de les nostres dades no necessitem mirror, llavors podem triar entre qualsevol de les tres opcions. La opció 3 (Cintes) es la més barata i per tant es la que més s'adequa al nostre pressupost.

Pel que fa els dies entre dos backups, en ser el nostre pressupost reduït decidim fer una copia de seguretat cada 7 dies. Guardarem dues còpies (l'última i una de fa sis mesos) i decidim tenir un backup on-site i no mirror perquè és el que podem asumir amb el nostre pressupost.

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

Amb una xarxa interna de 4 Gbps, contracten quatre lines de 1Gbps. Però seria molt interessant considerar la possibilitat d'una segona línia.

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?

Fem un backup cada 7 dies, per tant recuperar els fitxers pot ser un problema si hi ha corrupció o esborrat erroni. Pero com tenim els discos sobredimensionats (al voltant de x1.5) implementarem snapshots per tal de solventar aquest problema de manera eficient.

• Es trenca un disc (es perden dades? quan trigo en recuperar-me? el negoci s'ha d'aturar?)

Les dades a les que s'accedeixen més, i per tant és més fàcil que ocorri algun error, estan en RAID51. El trencament d'un disc en aquest cas no afectaria ja que tenim replicació. La recuperació és fàcil gràcies a la copia. A més tampoc passa res si es trenquen dos discos perquè també esta en RAID5.

Les altres dades estan en RAID5 i la seva recuperació es pot fer per mitjà de recuperació de RAID a costa de rebaixar l'ample de banda amb disc, que no comporta cap problema ja que aquestes dades no s'accedeixen quasi mai.

- **Puc tenir problemes de servei si falla algun disc?**

En el cas de les cabines de S_{30}no hi ha problema al haver-hi RAID51. En les cabines de S_{70} tampoc tenim problema, perquè és a on menys s'accedeix (i els clients són conscients que aquestes dades poden trigar més temps en ser accessibles).

- **Cau la línia elèctrica. Què passa?**

Cespedes té un sistema alternatiu de diesel, però no especifica quant temps dura, pero si diu que té un màxim de 22h a l'any de downtime, cosa que ens garanteix que, com a molt, estem un dia sense poder treballar.

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

Tenim contractades quatre línies de 1Gbps, llavors si cau una ens queda una velocitat de 3Gbps que, en principi, és més del que necessitem (2.88 Gbps).

- **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**

    1% de S_{70} és 5.32TB

    Velocitat de recuperació proporcionada per les cintes es 5TB/h

    Velocitat de recuperació proporcionada per un disc:

    IOPS en RAID5 = 90 x 710 x (0.5+0.5/4) = 39937.5

    Velocitat de recupearció: 39937.5 x 4kB/OP x 3600s/h = 0.5751TB/h.

    Clarament, la velocitat de recuperació esta limitada per la velocitat dels discos.

    Tardariem 5.32/0.92016=9.2505h.

    1% de S_{30} és 2.28TB

    Les dades de S_{30}estan en RAID51. Per tant, és molt difícil que hi hagi necessitat de recuperació d'aquest petit volum de dades, ja que és poc probable que es corrompeixi simultàniament un disc i el seu mirror.

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

    Volum de dades de S_{70}són 532TB

    Temps de recuperació és 532TB/0.5751TB/h = 925.056 hores (uns 38.54 dies)

    Volum de dades de S_{30}són 228TB

    IOPS en RAID51 = 420 x 3360 x (0.5+0.5/8) = 793800

    Velocitat de recupearció: 793800 x 4kB/OP x 3600s/h = 11.43TB/h.

    En aquest cas estem limitats pel sistema de cintes. Tardarem 228/5 = 45.6h = 1.9 dies.

Temps de recuperació total és 38.54 dies + 1.9 dies = 40.44 dies. Lògicament, com les dades més prioritàries són les de S_{30}, ens centraríem, primer, en recuperar aquestes dades i així, possiblement, podríem tenir el nostre sistema en funcionament.

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 quan em costaria evitar aquesta situació.

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

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 quan em costaria evitar aquesta situació.

Caiguda de la xarxa de dades:

Segons les dades que tenim (apèndix 6) , una línia cau 1 hora cada 18 mesos, entre 1 i 3 hores (o sigui, 2±1h) cada 3 anys (36 mesos) i entre 3 i 9 hores (6±3h) cada 6 anys (72 mesos)

Per tant la probabilitat de caiguda mensual és de 1h/18m + (2h±1h)/36 + (6±3h)/72m = (4+4±2+6±3 h)/72m= 14±5 hores/ 72 mesos

Un mes té en mitja 365,25 / 12 dies de 24 hores (compto l'any de traspàs) = 730,5 hores 72 mesos x 730,5 =52.596 hores

Una línia està penjada entre 9 i 19 hores (14±5 hores) de cada 52.596, o sigui que la possibilitat de downtime es de entre 9/52596 (0,017%) i 19/52596 (0,036%).

En el nostre cas tenim quatre línies, la caiguda de les quatre simultàniament és 1/(18x18x18x18) + 2±1 /(36x36x36x36) + (6±3/72x72x72x72) = 294±19h de cada 72^4 mesos, o sigui 20±7 hores de cada $19.63x10^{10}$h. Per quatre línies, la probabilitat d'estar downtime està entre 0.0000014% i 0.0000015%.

En cas de tenir una línia, en 5 anys (43830 hores) tenim entre un 0.0000014% i 0.0000015% de probabilitat de downtime. Notem que nosaltres no pagaríem res en cas de caiguda de la xarxa, ja que no tenim SLA, però els grups de recerca, poden arribar a perdre, en conjunt, entre 20.000x19x0.00061 i 20.000x19x0.00066, o sigui entre 231.80€ i 250.80€.

<u>Fallada de disc</u>

Per reconstruccions de disc puc tenir problemes d'accés en IOPS.

S_{30}: estem en RAID 51. Si falla un disc puc copiar-lo del mirror, no cal reconstruir.

Cal, però, fer la còpia del disc. Si es pot predir la fallada per SMART (70%) es farà la còpia quan el clúster estigui inactiu. Si falla es podrà atendre els accessos sense problema.

Tenim 420 discos (168 de dades i 252 per RAID51). És un disc HHD Enterprise de 10000rpm, per tant, la probabilitat de fallada és un 2.84% anual. De les fallades, només el 30% requereixen

reconstrucció, la resta es preveuen per SMART. Per tant, 420 discos x 2.84% fallada x 30% fallades amb reconstrucció, els discos tenen un 357.84% de probabilitat de fallar per any, un 1789.20% de probabilitat de que falli un disc en 5 anys, és a dir, uns 18 discos cada 5 anys.
420x3360=1411200 IOPS
Com estem en RAID51 els IOPS requerits son 270000 en comptes de 60000.
De totes formes, els IOPS estan sobredimensionats x5.23. Per tant, si necessito copiar un disc en un altre és ràpid.

S_{70}: El disc escollit falla cada 2,84% anual (HDD Enterprise <10000 rpm), tinc
90 discos. Fallen 90 x 0.0284 = 2.556 discos per any x 5 anys = 12.78 discos.
El 70% de les fallades es poden predir per SMART. Per tant, les que s'han de reconstruir són el 30%, o sigui, 12.78 x 0.3 = 3.834 discos en tots (posem 4).
Cada cop que reconstrueixo un disc en RAID 5 triguem 4 hores per TB. Els discos són de 10TB = 40 hores de reconstrucció. Durant aquest temps els discos d'aquell clúster van a la meitat de velocitat. Tinc 15 clústers de 6 discos, 14 a tot funcionament i 1 a mig funcionament. Cada clúster té 6 discos a 710 IOPS = 4260 IOPS per clúster x 14.5 clústers en funcionament = 61770 IOPS. L'escenari demana 60000 IOPS (en el més pesimista dels casos) ja que aquest servidor no s'utilitza quasi mai. Per tant, hauríem d'anar sense problemes.

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)? Quin és el recurs que s'esgota abans? Feu un informe de les implicacions que suposaria un increment d'un 20% en el volum de negoci (tot, clients, dades, …)

Tant per S_{70}com per S_{30}tenim la capacitat sobredimensionada al voltant d'un 48%.
Pel que fa als IOPS, en el cas de S_{70}tenim 39937.5 disponibles i com no s'accedeix quasi mai (entenem que és un 10% o menys d'accessos), sempre estarà sobredimensionat.
En el cas de S_{30}tenim 793.800 IOPS, ja que estem en RAID51, que representa un 1323% del que necessitem.
La pressió sobre la xarxa és de 2.88Gbps i tenim una de 4Gbps. És un 138.89% el que necessitàvem.
Per tant, incrementar un 20% el volum de negoci no implica, en principi, realitzar més inversions.

4.4.- Inversions més urgents

Donat el CPD resultant és possible que no haguem escollit la millor opció per manca de diners. El CPD no és nostre, nosaltres només ho dissenyem, així que al final s'hauria de fer un informe als que posen els diners de en què valdria la pena invertir per millorar rendiment, seguretat o…

Com ja havíem anunciat anteriorment, creiem que seria convenient desviar tràfic de la xarxa mitjançant una SAN, especialment de cara a una futura introducció d'un mirror en el nostre sistema. També seria interessant augmentar la nostra capacitat d'emmagatzematge, ja que és un dels factors més limitants que tenim de cara al creixement del nostre CPD. Una altra millora podria ser incrementar el nombre de backups que tenim, que ara només són dues còpies.