CPD - EBH

Segona entrega

Escenari 11: VM-2

Albert Bernal i Blai Ortuño

08/04/2021

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

1.- Descripció bàsica

TAULA 1: ESCENARI ORIGINAL: EXTRET DE L'ENUNCIAT. OMPLIU EL QUE HI HA EN GRIS.	
Nombre de Us	112U
Alçada Rack (en Us)	42U
Consum	350,kW
Sobreprovisionament d'electricitat	6%
Nombre de servidors	50
Diners Totals	€10.000.000,00
Diners gastats	€7.500.000,00

taula 2: Elements que escolliu vosaltres	
Elements de disc, mirror i backup	
GB a emmagatzemar	700000
Dies entre 2 backups	2
Còpies senceres a mantenir	3
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)	0
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

[1]

[1] El valor (700000) de la captura de pantalla està obsolet, ja que a les etapes finals del treball vam decidir sobredimensionar-lo lleugerament, fins a 800000GB.

Bandwidth provider	
Tipus de línia (1:10Mbps; 2:100Mbps; 3:1Gbps; 4:10Gbps; 5:100Gbps)	5
Número de línies agregades	2
Segon proveïdor? (0=NO, 1=SI)	0
SAN? (0=no, 1=8Gbps, 2=16Gbps, 3=32Gbps)	1
Cabina de discos	
Opció Disc principal (Entre 1 i 10)	8
Nombre de discos a comprar	387
Opció cabina de discos (Entre 1 i 6)	4
Nombre de Cabines	11
Cabina de discos 2 (cas de fer servir dos 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
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)	€66.538,36	€332.691,82
Empresa de Housing escollida	Mordor	
Cost Housing (inclou electricitat addicional)	€121.980,75	€609.903,77
Off-site: empresa escollida	Take the tapes and run	
Cost mirror	€0,00	€0,00
Cost backup	€70.100,00	€350.500,00
Cost Bandwidth provider	€151.200,00	€756.000,00

TAULA 4: CAPEX	Cost
Diners gastats en servers, xarxa, etc	€7.500.000,00
SAN	€230.106,00
Sistema emmagatzematge	€198.964,00
TAULA 5: AJUST AL PRESSUPOST	
Opex a 5 anys, total	€2.049.095,59
Capex a 5 anys, total	€7.929.070,00
Despeses totals a 5 anys	€9.978.165,59
Diferència respecte al pressupost	€21.834,41

2.-Anàlisi de necessitats

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

El nostre sistema constarà de dues parts diferenciades, ja que les màquines virtuals es trobaran al servidor i les dades de l'usuari estaran al disc centralitzat., per tant no tindrem en compte els 4 GB que ocupa cada VM si no migra:

Disc central → Cada VM hi tindrà reservat 1TB per a dades → 500VM * 1TB = 500TB.

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

Disc central → Aquests discos no hauran de tenir en compte el tràfic exterior, per tant els seus requeriments seran els següents:

Tràfic amb disc servidor: 6000 Kbps d'escriptura i 4000 Kbps de lectura

Lectura → 500 * (4000) = 2 Gbps → 2 Gbps / 8 bits / 4KB = 62.500 IOPS

Escriptura → 500 * (6000) = 3 Gbps → 3 Gbps / 8 bits / 4KB = 93.750 IOPS

IOPS totals = 62.500 + 93.750 = 156.250 IOPS

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

En aquest cas, tindrem tant el tràfic degut a les migracions com el tràfic amb l'exterior. Per tant, obtindrem el següent resultat:

4000 Kbps * 500 VM = 2000 Mbps de tràfic exterior

(5 * 4) GB / min * 1 min/60 seg * 1000MB / 1GB = 333,33 MBps * 8 bits / 1 byte = 2,67 Gbps

En total, el tràfic amb el client és de 2 Gbps + 2,67 Gbps = 4,67 Gbps

2.4- Tràfic amb el disc:

En aquest cas, el tràfic serà el que fem mitjançant les lectures i escriptures a disc:

4000 Kbps * 500 VM = 2 Gbps de lectures

6000 Kbps * 500 VM = 3 Gbps d'escriptures

En total: 3 Gbps + 2 Gbps = 5 Gbps de tràfic amb el disc

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

La xarxa del nostre escenari és de 6 Gbps, per tant, tenint en compte que el tràfic total del nostre sistema (exterior + disc) és de 4,67 Gbps + 5 Gbps = 9,67 Gbps, no tenim suficient amb els 6 Gbps que ens proporciona la 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.

En un principi, el dispositiu d'emmagatzematge que vam escollir era el 6, Samsung 860 EVO, ja que oferia les prestacions necessàries en quant a preu per GB i sent SSD comptava amb moltes més IOPS de lectura i escriptura de les que necessitàvem. Tot i estar bastant segurs de que era una bona opció, més endavant vam veure que teníem una petita part del pressupost restant que podíem destinar a millorar la fiabilitat del sistema canviant uns discos *consumer* per uns *enterprise*, i així passar de QLC NAND a TLC NAND. Entre les tres opcions restants de SSDs *enterprise*, ens vam decidir per la 8, Kingston SEDC100M, simplement pel seu cost per GB i, tot i que no ens fossin de molta utilitat, unes IOPS més altes.

En quant a les cabines que allotjaran aquests discos, aquestes seran d'un sol tipus, donat que només tenim un tipus de client amb unes necessitats concretes. Ens hem decidit pel model 4, ja que ens oferia el major nombre de badies per cabina, tenia suport per a la configuració de RAID que hem escollit i admetia *spare disks* per a complementar la fiabilitat del sistema. Hem triar la cabina 4 i no la 5 ja que només trobem afegit el suport per a SSDs i el considerem un aspecte inútil amb la nostra configuració.

Després de valorar les diferents configuracions, hem decidit que al nostre sistema utilitzarem RAID 6, ja que creiem que és un bon balanç entre percentatge d'emmagatzematge útil i protecció contra errors. El nostre plantejament va ser que amb RAID 5 no reduïm el cost suficient com per a justificar la rebaixa en fiabilitat, i amb sistemes com 51 i 61 el cost de duplicar no es podia justificar amb l'augment en fiabilitat ja que comptàvem amb altres mesures de protecció com un *backup* i *spare disks*.

Com hem calculat a l'apartat 2.1, per al nostre escenari necessitàvem com a mínim 500TB d'emmagatzematge útil per a les dades dels clients. Hem pensat que era una bona decisió sobredimensionar el sistema per a un futur amb més clients, i comptant també amb el cost afegit que comporta la paritat de RAID 6, utilitzarem un total de 800TB. A més, haurem de tenir en compte els *spare disks*, 2 per cabina. En total, necessitarem 13 cabines, 443 discos.

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

Hem escollit utilitzar una SAN ja que considerem que ens aporta una velocitat necessària per a la nostra infraestructura, ja que realitzem *backups* cada dos dies en cintes, i si volem recuperar aquestes dades en un temps viable i no en 20 hores ens caldrà un molt bon ample de banda. A més, també ens proporciona un extra de seguretat, protecció contra fallades i prevenció de *bottlenecks* en un pic de demanda, entre d'altres, aspectes que creiem importants i que s'han de tenir en compte.

En quant al cost, la idea era triar un sistema que pogués recuperar ràpidament les dades dels *backups*, i la diferència de preus entre un SAN o NAS era asumible en el nostre cas. Per tant, vam optar per la SAN encara que era l'opció més cara.

3.3.- Posem un mirror?

En quant al mirror, hem decidit no utilitzar-ho per diverses raons. La primera és l'elevat cost que té, el qual no podíem assumir amb el nostre pressupost sense treure diners d'altres recursos que consideràvem més importants, com per exemple la qualitat i la quantitat dels discs, el temps entre *backups* o el *housing*.

La segona raó és que tampoc crèiem que fos necessari l'ús del mirror, ja que les dades que guardem són les dades dels usuaris en les VMs, i per tant no estem parlant de dades crítiques com podrien ser dades d'un banc. Tenint en compte això, vam prioritzar la velocitat de recuperació de les dades i la realització d'un *backup* cada dos dies per sobre del mirror.

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

Entre les tres opcions, ens vam decantar per les dues últimes ja que s'adaptaven més a les nostres necessitats. En quant a aquestes dues, vam optar per la tercera opció, Mordor, per diverses raons. Primerament, ens oferia la monitorització dins del preu, i teníem clar que en volíem tenir. A més, també oferia un millor sistema de seguretat en general que la segona.

Tot i això, la principal raó va ser que aquesta opció ens donava la possibilitat de contractar un servei de gestió de *backup* de cintes, on s'inclou una copia off-site d'aquestes juntament amb la seva recuperació en cas d'haver-hi algun problema, sent per tant perfecte per a nosaltres amb el *backup* realitzat en cintes.

En conclusió, encara que la segona opció es una mica més econòmica i tampoc està malament, opinem que la tercera és una millor opció en general, ja que ens ho podem permetre i ens aporta tranquil·litat, comoditat i seguretat, aspectes molt a valorar de cara al correcte funcionament del CPD a llarg termini.

3.5- Posem monitorització?

Com ja hem esmentat anteriorment, creiem que una bona monitorització del sistema és essencial per a un CPD. Aquesta ha estat una de les raons per les quals hem triat l'empresa de *housing* Mordor, ja que venia inclosa en els seus serveis. A més, havent decidit utilitzar *spare disks*, és necessari tenir algú per a canviar-los.

3.6- Opció de backup?

Primerament, teníem pensat usar HDDs per als backups, i així trobar un equilibri entre rapidesa, cost, eficiència i capacitat d'emmagatzematge. Tot i això, amb el pressupost que teníem només podíem realitzar backups amb HDDs cada 15 dies, i encara i així ens reduïa els diners que podíem utilitzar en altres aspectes de manera substancial. És per això que vam decidir utilitzar cintes, de manera que el preu era molt més baix i teníem la capacitat de realitzar *backups* cada dos dies.

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

Com hem esmentat a l'anterior apartat, farem backups cada dos dies en cinta. Això ens consumirà molt *bandwidth*, sobretot tenint en compte que són un total de 800TB, de les quals 724.7TB (742077.8GB) seran emmagatzematge útil. Hem decidit posar dues línies agregades de 100 Gbps per a que el procés de

backup i de recuperació de les dades en cas de fallida sigui el menys costós possible en quant a *downtime* de l'empresa. Els càlculs han estat els següents:

742077.8GB * 8bits / (214-9.67)Gbps / 3600s =8.07 hores

Per tant, tenint en compte que fem un *backup* cada dos dies, és un temps considerable però raonable.

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?

La podré recuperar de l'últim backup realitzat, i ja que els realitzem cada dos dies llavors serà del dia anterior o l'anterior a aquest.

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

Si es trenca un disc, en principi tenim 2 *spare disk* amb tecnologia SMART que haurien d'haver detectat aquest deteriorament del disc. En aquest moment, poc a poc es van traspassant les dades del disc que es trencarà cap a l'*spare disk*, de manera que quan el procés hagi finalitzat, aquest *spare disk* passarà a ser el nou disc en funcionament i l'altre disc serà retirat. Aquests *spare disks* es poden canviar un cop han estat utilitzats, de manera que sempre tinguem (en el nostre cas 2) *spare disk* en la cabina i en cas de trencar-se un disc no perdem dades.

Per tant, si es trenquen discs de manera habitual (sense cap incident con inundacions, incendis, etc), el temps de recuperació serà 0 i el negoci no s'hauria d'aturar. I en el cas poc probable en el que la tecnologia SMART no ho detecti, llavors sí que haurem de recuperar les dades a partir de la paritat, sent el temps de recuperació el temps que trigui el procés de recuperació per paritat. En aquest cas, caldrà aturar el negoci per a les persones que tinguin les seves dades en aquell disc.

Puc tenir problemes de servei si falla algun disc?

Com hem esmentat en la pregunta anterior, si falla un disc només podrem tenir problemes de servei si no hem arribat a utilitzar l'*spare disk*, ja sigui perquè no l'hem detectat o alguna altra raó.

Cau la línia elèctrica. Què passa?

En el nostre cas, ja que hem escollit la opció de housing 3 (Mordor), tindrem un generador dièsel capaç d'aguantar 72 hores la potència pic, temps en teoria suficient per a poder solucionar els problemes que tinguem amb la línea.

En cas d'haver passat algun accident que afecti a la línea general (tsunami, terratrèmol, etc.), llavors podrem aguantar aquestes 72 hores i esperar que la línea es recuperi aviat si comptem que el CPD ha aguantat aquest accident i no s'han perdut els discs ni el generador.

Cau una línia de xarxa. Què passa?

Com que tenim més d'una línia, tenim connectivitat inclús si una cau i s'ha de restablir, amb el temps que això comporta. Si cau el proveïdor, com que hem decidit que no tenim un segon, ens quedaríem sense poder fer res més que pagar als nostres clients els diners acordats al SLA fins que el proveïdor solucionés el seu problema. Tot i això, aquest últim també hauria de complir el seu contracte amb nosaltres i pagar-nos.

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.

En el cas de perdre només un 1% de les dades, haurem de reconstruir-les utilitzant la paritat que ens aporta la configuració de RAID 6, que hauria de poder solventar el problema tenint en compte que és una petita porció del total.

Pèrdua/corrupció de la totalitat de les dades.

En aquest cas l'impacte serà molt major i no podrem reconstruir, per tant, haurem de recuperar les dades del *backup* més recent que tinguem, que com a molt serà de fa 2 dies. Com hem calculat abans, això ens costarà, arrodonint, unes 7h.

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:

En aquest cas, segons el nostre SLA pagaríem 250€ l'hora per client, dels quals en tenim 500, sumant un total de 500 * 250 = 125.000 € l'hora.

Fallada de disc:

En cas de no haver pogut canviar el disc a temps amb la tecnologia SMART, també pagarem 250€ per hora per cada client afectat, de manera que hauríem de pagar només als clients afectats per la caiguda d'aquest disc. Ja que un disc pertany a una cabina i pot estar dins d'un *cluster* de discs en aquesta cabina, tindrem *x* clients treballant en aquest disc. Per tant pagarem 250 * *x*, sent *x* el número de clients que treballen sobre aquest disc i que veuran el seu servei interromput fins que es solucioni el problema.

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, ...)

El recurs que s'esgotarà abans és l'emmagatzematge. Amb un 20% més de volum de negoci, el nombre de clients ascendiria a 600. Ja que tenim aproximadament un 45% de sobreprovisionament en emmagatzemtage, podríem créixer aquest 20% sense problemes en aquest aspecte, però estaríem apropant-nos al límit, que, repetim, està al voltant d'un 45%. Hem valorat el fet d'augmentar l'espai, però el cost en hores de *backup* i la conseqüent ocupació de l'ample de banda de la nostra xarxa ens ha semblat prohibitiu.

En quant a les IOPS, havent triat un disc amb un rendiment extraordinari i sense trobar-nos en un escenari molt exigent, no tindrem cap tipus de problema. Per a posar una mica en context amb dades, els nostres requeriments d'IOPS de R/W eren de 62.500/93.750 i un disc del tipus escollit sol proporcionava 540.000/205.000.

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...

Si haguéssim tingut més pressupost, hauríem fet canvis en les següents àrees:

- Xarxa: Tot i que els requeriments dels nostres clients no ens ocupen quasi res del *bandwidth* total (al voltant d'un 5%), creiem que reduir el temps de *backup* i de recuperació d'aquest ens evitaria temps de *downtime* que, tot i que sigui poc probable gràcies a la RAID i els *spare disks*, pot succeir.
- *Backup* diari: En un principi teníem pensat fer un *backup* diari per a tenir còpies més recents i que es perdessin menys dades, però el cost de duplicar la freqüència ens sortia del pressupost.
- Còpies senceres a mantenir: La nostra idea era seguir un model tradicional de quatre còpies a mantenir, però les restriccions econòmiques només ens han permès tenir-ne tres.
- Emmagatzematge: Com hem vist a l'apartat anterior, l'espai per a dades és un aspecte que ens limitaria el creixement de l'empresa en un futur relativament pròxim i, per tant, seria un aspecte que necessitaria millores.