Activitat 2: Enunciat general

Introducció:

L'objectiu d'aquesta activitat és simular la construcció d'un CPD a partir d'unes especificacions i restriccions preestablertes. Cada grup d'estudiants dissenyarà un CPD amb restriccions diferents. Podeu (i us animem) a col·laborar entre vosaltres, però donat que els treballs són diferent i que les decisions a prendre depenen de les restriccions, no assumiu directament que la solució d'un company pot ser la que us vingui bé a vosaltres. Més endavant teniu explicat com escollir pràctica.

En aquest document trobareu una descripció restriccions tècniques i econòmiques de cadascun dels escenaris a considerar, i que seran assignats als diferents estudiants. Parlarem de quines combinacions hi ha, com escollir el CPD que us toca fer, quina documentació s'ha de lliurar i quan, i com i quan fer es auditories. Llegiu-vos curosament tot l'enunciat per tenir clares les regles del joc. Recordeu que aquesta activitat és un 40% de la nota i que no acceptarem "despistes" per manca de planificació.

Requeriments tècnics (variacions a escollir):

Hem preparat un conjunt de plecs de requeriments tècnics que definiran el CPD que he de dissenyar. Cada grup d'estudiants n'haurà d'escollir un. Les podeu trobar a la taula següent. En les següents seccions d'aquest document es desgrana el significat de cada opció.

Id/11	Id/12	Storage	Workload	Budget
11/1	12/1	NAS/SAN	Web 2.0	2.5M€
11/2	12/2	NAS/SAN	VM hosting	2.5M€
11/3	12/3	NAS/SAN	HPC	2.5M€
11/4	12/4	HDFS/GPFS-SNC	MapReduce	2.5M€
11/5	12/5	NAS/SAN	HPC	12M€
11/6	12/6	NAS/SAN	Web 2.0	12M€
11/7	12/7	NAS/SAN	VM hosting	12M€
11/8	12/8	HDFS/GPFS-SNC	MapReduce	12M€
11/9	12/9	NAS/SAN	HPC	25M€
11/10	12/10	NAS/SAN	Web 2.0	25M€
11/11	12/11	NAS/SAN	VM hosting	25M€
11/12	12/12	HDFS/GPFS-SNC	MapReduce	25M€

Aspectes tècnics:

Tipus de emagatzematge:

- La primera opció és que el CPD té un emmagatzematge centralitzat, en forma de NAS o SAN. En aquesta configuració podreu assumir que cada node del CPD té un sol disc local per al sistema operatiu, però que no serà directament usable pels usuaris.
- La segona opció és aquella en què el sistema d'emmagatzematge del CPD es troba totalment distribuït, significant això que cada node del CPD aporta la seva capacitat de disc local al sistema d'emmagatzematge global. Assumirem que en aquest escenari es fa servir un sistema de fitxers distribuït del tipus HDFS de Hadoop o GPFS-SNC de IBM. També assumirem que cada node té un disc local per al sistema operatiu que no pot ser usat pels usuaris, i una capacitat addicional d'emmagatzematge que caldrà determinar i que serà la que s'utilitzarà per al sistema de fitxers distribuït i per a l'ús local dels processos.

NOTA: En l'activitat 2 veurem l'emmagatzematge com una "caixa negra" on les dades van a parar. Ens importarà el seu ample de banda, local (per node) i agregat en la xarxa, però no mirarem els detalls de com funcionarà internament.

Tipus de Workload: El requisit de memòria general serà que es necessiten per tots els workloads **4GB de RAM per cada core** que tingui un node físic.

MapReduce:

- Característiques:
  - 1 tasca per core (assumirem que l'aplicació només té *maps* i no *reduces*)
  - Per cada 100Mhz de cada core, s'obté un MB/s de capacitat de processat de la tasca
  - Les dades que processa un *map* són un 30% dades remotes (que venen d'un altre DataNode) i un 70% dades locals (es troben en un DataNode que coexisteix en el mateix node on corre el map).
  - Les dades a processar ja es troben en el HDFS a l'inici (no cal llegir-les de fora, ni exportar-les en acabar l'execució).
  - Dades remotes perfectament distribuïdes
  - Cal tenir un mínim de 1Gpbs de connectivitat cap a Internet
- Mesures de capacitat:
  - Quin serà el rate de processat de dades?
    - Unitat: MB/s
  - Quin serà el working set més gran que es podrà tenir?
    - Unitat: MB

Web 2.0:

- Característiques:
  - Escalabilitat horitzontal perfecta (tant cores com nodes)
  - Es tracta d'una aplicació tipus Flickr o Picassa, en què el contingut principal són fitxers (fotografies per exemple) i no complexes transaccions a una Base de Dades. Per tant, aquí obviarem la presència de cap servido de BD.
- Ús de recursos:
  - Per cada 800Mhz d'un core de capacitat, es pot processar una petició HTTP dins d'un temps de resposta de 100ms (i aquest és el SLA compromès).
  - Cada petició ocupa 600 bytes en mitjana
  - Cada resposta ocupa 180KB en mitjana
  - Cada petició genera 5 accessos a disc de 1KB cadascun, que es realitzen dins dels 100ms de resoldre la petició
  - Mesures de capacitat:
    - Quin serà el màxim nombre de peticions HTTP servides concurrentment sense augmentar el temps de resposta?
      - Unitat: peticions per segon
    - Quin serà el working set més gran que es podrà tenir?
      - Unitat: MB

VM Hosting:

- Característiques:

    - Totes les VMs tenen una vCPU, i més d'una vCPU es pot mapejar sobre un mateix core físic. Com a restricció, 1Ghz de cada core queda reservat per cada vCPU que s'hi mapeja.
    - Munten el disc local per iSCSI
- Ús de recursos:
    - Per cada VM, s'observa en mitjana 950KB/s de dades llegides i escrites a el storage (bidireccional, per tant 950KB/s llegit i 950KB/s escrit)
    - Per cada VM, s'observa en mitjana 50KB/s de comunicació amb l'exterior
- Mida de la VM:
    - 4GB de RAM
    - 40GB de disc
- Mesura de capacitat:
    - Quin serà el màxim nombre de VMs que es podran suportar concurrentment
        - Unitat: VMs concurrents en marxa, i nombre de vCPU per core físic
    - Quin serà el nombre més gran de VM que es podrà tenir (corrent o suspeses)?
        - Unitat: Nombre

HPC:

- Característiques:
    - Un job MPI vol córrer sobre tot el CPD
- Ús de recursos:
    - Cada tasca paral·lela MPI corre sobre un core d'una CPU
    - En el cas d'utilitzar GPUs, assumirem que cada GPU va associada a una CPU (és a dir, podem tenir com a màxim tantes GPUs com xips per node, no tantes com cores).
    - En les CPUs:
        - Cada tasca paral·lela MPI corre en un **core** diferent, i repeteix un bucle d'execució en què fa una ràfega de CPU de 100ms (tota a base d'operacions en coma flotant a la màxima velocitat que la CPU permeti) abans de comunicar-se amb un altre procés d'un altre node, enviant un missatge de 64KB. No es pot començar la següent ràfega de CPU fins que no s'ha acabat la transferència, i per tant, la latència de la transferència impacta directament sobre el temps d'execució de les aplicacions.
    - En les GPUs:
        - Cada tasca paral·lela MPI corre en un **node** diferent, i és l'encarregada de passar a les GPUs del node les dades a processar. Cada GPU repeteix un bucle d'execució en què fa una ràfega de GPU de 100ms (tota a base d'operacions en coma flotant a la màxima velocitat que la GPU permeti) abans d'acabar, retornar les dades a la CPU i que la tasca MPI iniciï una comunicació amb un altre procés d'un altre node, enviant un missatge de 64KB. No es pot començar la següent ràfega de CPU fins que no s'ha acabat la transferència, i per tant, la latència de la transferència impacta directament sobre el temps d'execució de les aplicacions.
    - Per cada 1000 ràfegues de CPU/GPU, llegeix un fitxer de disc 125MB. Aquesta lectura es pot repartir durant el temps d'execució de les 1000 ràfegues.

    - El model de màquina a muntar ha de ser basada en la computació o bé en CPUs o bé en GPUs, però no una barreja de les dues coses. Barrejar-les, causaria un greu desbalanceig que alentiria tota l'aplicació.
    - Es poden sol·lapar les lectures a disc amb la comunicació i processat
- Mesura de capacitat:
    - Quin serà el nombre màxim nombre d'operacions en coma flotant que podrà realitzar el CPD?
        - Unitat: FLOPS[1]
    - Quin serà el working set més gran que es podrà tenir?
        - Unitat: MB

- **Budget a 5 anys** (incloent compra de HW i OPEX), IVA inclòs:
    - 2.500.000€
    - 12.000.000€
    - 25.000.000€

Com escollir pràctica:

Si us fixeu, hi ha 12 escenaris diferents i sou 38 estudiants matriculats a l'assignatura. Fareu la pràctica per parelles, per tant, i dues parelles podran escollir el mateix enunciat. Després usarem aquestes parelles dobles per a auditar-vos el treball.

Podeu escollir la pràctica que vulgueu en funció dels vostres interessos (quina combinació us crida més l'atenció), però és molt important saber que no es podran fer canvis a posteriori.

Per escollir quina combinació fareu, crearem un doodle per a cada subgrup de l'assignatura. Es podrà escollir en el **doodle** que s'indicarà al Racó.

Lliuraments:

Heu de fer un lliurament setmanal del progrés del vostre treball, una documentació final i una auditoria d'un treball final d'un altre grup.

- Dates pels lliuraments setmanals: cada dijous, abans de mitjanit. Hi ha d'haver un lliurament el 23 i 30 de novembre, i el 7 i 14 de desembre. Els fitxers han de ser PDF i el nom seguir el següent format: *GrupA2.Cognom.Nom.Data.pdf*. Per exemple, si els dos professors de l'assignatura fossin el grup "11/3" i lliuréssin l'informe del 22 de novembre, el fitxer s'hauria d'anomena 11-3.*Carrera.David.López.David.22.novembre.pdf*. Podeu fer el lliurament amb anticipació, però mantingueu al nom el dia de lliurament OFICIAL, no el que lliureu realment.
- Lliurament de la memòria final: tots teniu la mateixa data límit per lliurar la memòria final. La data és el divendres **dia 21 de desembre** abans de la mitjanit. El nom del fitxer ha de seguir el següent format: *GrupA2.Cognom.Nom.Final.pdf*, així el nostre lliurament d'exemple seria: 11-3.*Carrera.David.Lopez.David.Final.pdf*. Si voleu lliurar els fulls de càlcul desenvolupats, aleshores feu un zip amb tots els fitxers i lliureu un fitxer amb el mateix nom, però extensió zip.

[1] http://www.intel.com/content/www/us/en/support/processors/000005755.html

- Lliurament de les auditories. Abans del dia **18 de gener** caldrà lliurar les auditories de les pràctiques que s'us hagin assignat. El nom del fitxer i el format de l'auditoria s'us penjaran al racó en el seu moment.

Tots els lliuraments es faran al racó. Hi ha dos llocs actius on fer els lliuraments, un pels lliuraments setmanals, un altre pel final, i un per a les auditories.

Avaluació del treball:

El treball representa el 40% de la nota de l'assignatura. Al seu temps, la nota del treball consta de dues parts: la nota tècnica (entre 0 i 5) i la nota de qualitat de la comunicació (entre 1 i 2), que inclou la qualitat de presentació del treball i la seva redacció. La nota final és el producte de la nota tècnica i la nota de comunicació, de manera que les dues parts tenen molta influència: una bona part de comunicació pot pujar molt la nota tècnica, però si aquesta és dolenta, no tindràs una bona nota final. Si per altra banda fas una part de comunicació molt dolenta (sería tenir una nota igual a 1) llavors la nota màxima a la que podeu aspirar (sobre 10) és de 5, cal de fer un treball tècnic impecable.

Les **auditories** serviran per a ajustar la nota tècnica dels altres grups, i la seva no presentació implicarà una **baixada de nota significativa en la nota tècnica de qui no la lliuri**.

La nota tècnica serà la mateixa pels dos membres del grup (si es fa en parelles), excepte si algun membre del grup es queixa als professors de que l'altre no fa la seva part. Si teniu conflictes feu-los arribar als professors.

Els lliuraments setmanals no són tan importants per la nota final. Més aviat és per controlar-vos i fer-vos notar que si ho deixeu pel final suspendreu. El treball requereix **MOLTES** hores i heu de treballar a partir d'ara mateix i sense afluixar. Intentar-lo fer les dues últimes setmanes és suspendre (excepte dedicant-vos en exclusiva a la pràctica durant aquests 15 dies).

Enllaços útils:

Eina per a dibuixar esquemes de RACKs

- *http://www.edrawsoft.com/rack-diagram.php*
- *https://www.lucidchart.com/pages/examples/server-rack-diagram*

Webs per configurar equipament de xarxa:

- Configuració d'equips de xarxa Cisco:
  - *https://apps.cisco.com/ccw/cpc/guest/estimate/create*

Webs per configurar servidors:

- Servidors pre-muntats (configurat per euros / espanya sempre):
  - **Servidors ThinkMate, SuperMicro i Intel**
    http://www.thinkmate.com

Webs per trobar preus de components (xarxa/servidors):

- Components:
  - *http://www.lambda-tek.com/componentshop/index.pl?region=ES*
  - *http://www.senetic.es*
  - *http://www.pccomponentes.com*

Informació addicional:

Al llarg dels propers dies anirem penjant més informació de suport, així com fulls de càlcul que us ajudin a prendre les decisions. Tota aquesta documentació anirà apareixent al racó.

Recordeu que totes les setmanes dedicarem una estona a discutir els problemes que trobeu. És important, però, que treballeu primer. La pregunta no hauria de ser del tipus "no se com fer això", sinó més aviat del tipus "necessito saber aquestes dades per fer la meva elecció" o "no se veure la diferència entre aquestes dues solucions". A més, us encoratgem a discutir amb els vostres companys. Hem obert un fòrum a l'assignatura dedicat només a l'intercanvi d'informació entre vosaltres.

Recordeu també que l'empresa D&D (de la que som propietaris) està oberta a que pregunteu especificacions que no tingueu, a que us aclarim dubtes i a que ens pugueu plantejar modificacions (com demanar més diners de manera justificada, o buscar altre centre de col·locació, si heu trobat un, o...) Estem oberts a tota mena de suggeriments.

Igualment, ens reservem el dret de canviar alguna de les condicions de l'enunciat si ens convé com a part de la política d'empresa (el que, desgraciadament, és un escenari força realista).

Algunes suggerències d'equipament a prendre com a punt de partida:

Noteu que només són alguns punts per iniciar la cerca, i que ni són les úniques opcions ni tothom les necessitarà en la seva solució! Recordeu que ni tots els estudiants tenen un SAN, ni esteu obligats a accedir al SAN amb cap tecnologia particular (FC, FCoE – CNAs, Infiniband)

- Servidors SAN/NAS:
  - *http://www.thinkmate.com/systems/storage*
    Noteu que accepten dispositius de xarxa HBA, HCA, NIC, CNA.
- Switch Top of Rack:
  - *http://www.cisco.com/c/en/us/products/switches/campus-lan-switches-access/index.html*
  - *Cisco Catalyst 2960X*
  - *Cisco Catalyst 3850X*
- Switch Agregació:
  - *http://www.cisco.com/c/en/us/products/switches/data-center-switches/index.html*
  - *Cisco Nexus 5000*
  - *Cisco Nexus 6000*
  - *Cisco Nexus 7000*
- Switch i HCA Infiniband (per LAN i/o SAN):
  - *http://www.mellanox.com/page/switch_systems_overview*
  - *http://www.mellanox.com/page/infiniband_cards_overview*

- Consultar els GFLOPS dels processadors Intel *skylake:*
  - *https://www.microway.com/knowledge-center-articles/detailed-specifications-of-the-skylake-sp-intel-xeon-processor-scalable-family-cpus/*

Housing i Backups:

En aquesta Activitat assumirem que no tenim restriccions de housing (on posar les màquines) ni de backups, donat que això ja s'ha treballat abans. Per tant, aquests costos els podeu obviar.

Connnexió amb l'exterior:

El Network Service Provider (NSP) dels nostres colocation services ha muntat una xarxa 100G a partir d'equips de diversos fabricants[2], el que significa que usant la tecnologia Dense Wavelength Division Multiplexing (DWDM), pot transportar 100Gbps per cada longitud d'ona que viatja per la fibra. En funció del fabricant, cada equip permet transportar entre 88 i 96 longituds d'ona (canals) per cada fibra, i per tant, l'ample de banda que pot oferir per fibra és de 8.8Tbps a 9Tbps. Per tant, tenim disponible un ample de banda molt gran.

[2] *Els equips muntats són una combinació de Cisco ONS 15454 Trunk Cards, Nokia Siemes Networks hiT 7300, Fujitsu FLASHWAVE 9500 i Alcatel-Lucent 1830 Photonic Service Switches.*

Dades rellevants per a l'activitat 1

Internament, les empreses de hosting que us hem proposat, posen a la nostra disposició, per un preu detallat a la següent taula, la interconnexió cap a l'exterior a través d'interfícies Ethernet.

Nosaltres només ens hem de preocupar de tenir en el nostres switches d'agregació el connector indicat a la taula, i ells s'encarreguen de fer-nos arribar el cable i donar connectivitat IP cap a l'exterior. És possible agregar el tràfic a través de diferents canals dels llistats (per exemple aconseguir 200Gbps de sortida a partir de dos connector de 100Gbps cadascun) i, òbviament, pagant el cost de totes les línies individuals que haguem usat.

Lectura recomanada:
http://www2.telegeography.com/hubfs/2017/presentations/telegeography-ptc17-pricing.pdf

Velocitat de sortida	Tipus de connector que té muntat el centre de colocation	Preu mensual (€)
10Mbps	1GbE RJ45	0,63 €
100Mbps	1GbE RJ45	6,30 €
1Gbps	1GbE RJ45	63,00 €
10Gbps	SFP-10G-SR	630,00 €
100Gbps	CFP-100G-SR10	6.300,00 €