EXPEDIENT DE CONTRACTACIÓ D'UN CLUSTER COMPUTACIONAL I D'EMMAGATZEMAMENT PER ALS SERVEIS CIENTÍFICO-TÈCNICS DEL DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

Expedient 03-05-14

Plec de prescripcions tècniques

Sumari

1. Objecte del contracte

L'objecte del contracte és l' adquisició d'un sistema en 'clúster' composat per nodes de càlcul i emmagatzematge per al Departament de Ciències Experimentals i de la Salut de la Universitat Pompeu Fabra.

2. Justificació de la necessitat

El Departament de Ciències Experimentals i de la Salut disposa ja d'un sistema d'aquest tipus, però és necessari augmentar-ne les capacitats, atès que la incorporació de grups d'investigació al DCEXS amb necessitats addicionals, tant computacionals com d'emmagatzematge, ha provocat que la infraestructura actual estigui arribant al seu punt de saturació i sigui necessari un increment de la capacitat de la mateixa.

El motiu d'aquest concurs, és doncs, la creació d'un nou sistema d'emmagatzematge i còmput (clúster) per tal d'augmentar les prestacions existents, la provisió d'unes eines de gestió de clústers més modernes que les actuals, i la elaboració d'una proposta sobre com integrar en el nou clúster el sistema d'emmagatzemament i càlcul de que ja disposa el Departament (només l'elaboració de la proposta, no la seva execució).

3. Descripció de l'equipament existent

Actualment el departament de Ciències Experimentals i la Salut de la Universitat a través del seu servei Científico-Tècnic d'Informàtica Científica disposa d'un clúster d'altes prestacions de càlcul composat per una part de nodes de càlcul, una part de servei d'emmagatzemament i una de connectivitat.

No és finalitat d'aquest concurs el manteniment de maquinari i programari del clúster existent.

a) Servei d'emmagatzemament.

Disposa d'un sistema de servidors de fitxers distribuït amb el programari GPFS (actualment a la versió 3.5), el qual està composat per 4 servidors connectats a 4 cabines (no DAS). Les cabines disposen de doble controladora i 48 discs de 3TB. (veure llistat de maquinari)

Cada servidor de fitxers disposa de 2 CPUs d'arquitectura basada en Intel x86 de 64 bits. Cada CPU disposa de sis cores, 12MB de memòria cache i 2,66GHz de velocitat de rellotge. Cada servidor també disposa de 64GB de memòria RAM, dos discs interns de 320GB configurats en raid 1, 2 interfícies Ethernet 1Gbps, una interfície Infiniband QDR de 40 Gbps per comunicar amb els nodes de càlcul i 3 plaques dualSAS de quatre canals cada un, pels quals es connecta a les cabines de disc. Els servidors de fitxers utilitzen el sistema operatiu GNU/Linux.

La interconnexió del sistema GPFS és duu a terme a través de la xarxa Infiniband, mentre la informació de senyalització i backup utilitza una mateixa xarxa ethernet.

b) Nodes de càlcul

Es disposa d'un total de 28 nodes, cada node amb 2 CPUs Intel x86 de 64 bits. Cada CPU disposa de 6 cores, 12MB de memòria cache i 2,40GHz de velocitat de rellotge. Cada node també disposa de 96GB de memòria RAM, un disc intern de 500GB, dues interfícies Ethernet 1Gbps per comunicació amb el node de gestió del clúster i una interfície Infiniband QDR de 40 Gbps. Aquests nodes tenen format 'blade' i estan distribuïts en dos xassís. El sistema operatiu instal·lat als nodes és GNU/Linux.

El sistema de càlcul també disposa d'un node Frontal i login amb 2 CPUs d'arquitectura basada en x86 de 64bits de 6 cores cada una, 12MB de cache i 2,66Ghz de velocitat de rellotge. Aquest node disposa de 2 discs de 320GB configurats en Raid 1, 2 interfícies Ethernet d'1Gbps i una interfície Infiniband QDR de 40 Gbps per comunicació amb els nodes de càlcul i amb el sistema d'emmagatzemament. El sistema operatiu instal·lat es un GNU/Linux. Aquest node gestiona la configuració i parametrització del clúster mitjançant el sistema de gestió de clúster "Rocks Cluster" (versió 6.1). El sistema de gestió Rocks Cluster porta les eines necessàries per a gestionar i configurar les cues de processos, programari per a la instal·lació de nodes, sistemes de monitorització, etc.

c) Connectivitat

Tal i com s'ha mencionat els nodes tenen format 'blade' i estan distribuïts en dos xassís. Cada xassís disposa de 2 switch Ethernet amb 14 ports interns de 1Gbps, 3 ports externs de 10Gbps i 6 ports externs de 1Gbps, així com un switch Infiniband QDR de 40 Gbps amb 30 ports, 14 d'ells interns de comunicació dels blades.

A més, es disposa d'un switch Infiniband mellanox IS5030 amb 14 ports ocupats dels 36 que té disponibles.

d) Programari

Es disposa de llicències GPFS suficients per al correcte funcionament del sistema. (Veure llistat de programari i maquinari existent)

4. Descripció de l'equipament a adquirir

Dividim les necessitats en 4 apartats: emmagatzemament, nodes de càlcul, connectivitat i integració

a) Emmagatzemament:

Espai de disc en un mínim de 250TB bruts. La tecnologia mínima d'aquests discs ha de ser Near-Line SAS i l'accés a la informació ha de ser mitjançant un sistema de fitxers distribuït tipus GPFS. L'equipament existent, descrit en el punt anterior, s'ha de poder integrar amb la solució proposada.

Per tal de millorar el rendiment a l'accés de les metadades del sistema d'arxius, es proposa adquirir un mínim de 8TB bruts en discos d'estat sòlid (SSD). Aquest espai caldrà que sigui accessible per un mínim de dos servidors o estar duplicat. S'admetran solucions alternatives sempre i quan ofereixin rendiments equivalents.

Els servidors de fitxers oferts han de tenir les característiques tècniques mínimes iguals a la dels servidors existents. (veure annexos)

Per tal d'optimitzar les I/O del sistema és necessari complir les següents premisses:

- Les CPUs dels servidors de fitxers han de tenir un mínim de 8 cores.
- Les CPUs han de disposar d'un mínim del 20MB de cache.
- Una relació mínima de 4GB de cache per cada 180TB de disc.
- Un mínim de 4 GB de memòria per core
- 2 interfaces de xarxa ethernet a 1 Gbps
- 1 interface infiniband de baixa latència
- 500 GB de disc local

Per tal de garantir la continuïtat del servei el sistema ha de disposar dels següents discs de recanvi en calent (*hot-spare*) que s'activaran automàticament en cas de fallada d'un disc en servei:

- Un disc de *hot-spare* per cada 30 discs SSD d'un mateix tipus o fracció.
- Un disc de *hot-spare* per cada 30 discs SAS d'un mateix tipus o fracció.
- Un disc de *hot-spare* per cada 15 discs NL-SAS o fracció.

b) Nodes de càlcul

Adquisició d'un clúster de computació d'altes prestacions (HPC), que en un primer moment serà independent de l'actual, però que ha de permetre la incorporació dels nodes actuals.

Característiques tècniques:

- Conjunt de nodes de càlcul amb unes característiques tècniques mínimes equivalents a les dels nodes existents. Un o dos d'aquests nodes faran la funció de node de '*login*'.
- Un mínim de 192 nuclis de càlcul en servidors de 2 CPUs. Per tal de garantir la interoperabilitatdel microcodi de les aplicacions i evitar dobles compilacions dels programes i els possibles errors provocats per les mateixes és imprescindible que la solució presentada utilitzi CPUs del mateix fabricant que la solució de clúster existent.
- Cada servidor precisa un mínim de:
    - Un mínim de 8 GB de memòria per core
    - 2 interfaces de xarxa ethernet a 1 Gbps
    - 1 interface infiniband de baixa latència
    - 500 GB de disc local
- Com a mínim un dels nodes de càlcul ha d'incorporar
    - Un mínim 1 d'un coprocessador de càlcul tipus GPU

Tal i com s'ha comentat, el cluster actual utilitza la distribució Rocks Cluster 6.1.0 per al desplegament i control dels nodes de còmput. El licitador ha de proposar un programari de gestió del cluster de prestacions iguals o superiors a l'existent.

c) Connectivitat

Cal que la solució proposada inclogui tots els elements de connectivitat necessaris per tal que el sistema global pugui funcionar complint tots allò especificat en aquest Plec de Prescripcions Tècniques. Per assolir aquest objectiu es podran utilitzar les interfícies actualment disponibles a l'equipament existent i que s'han especificat a l'apartat 3c).

A més, el sistema ha de proporcionar un mínim de 4 enllaços ethernet de 1 Gbps en coure per a la connexió a la xarxa troncal de comunicacions de la Universitat Pompeu Fabra.

d) Integració

Forma part d'aquest concurs la instal·lació i parametrització del nou sistema.

Cal presentar també una proposta de migració de la instal·lació actual cap a la nova instal·lació de manera que s'acabi convergint el màxim del maquinari actual al nou sistema. L'execució d'aquesta proposta no forma part de la licitació, només la presentació de la mateixa.

Cal incloure, si s'escau, el llicenciat de qualsevol nous producte de programari que la solució incorpori o l'ampliació del llicenciat existent per fer front a la solució final.

5. Característiques tècniques de la solució a proposar

- Els equips hauran d'anar instal·lats en armaris rack estàndard de 42U i d'amplada normalitzada a 19 polzades segons la norma DIN 41494
- S'haurà d'optimitzar l'ús del corrent elèctric en tot el sistema.
- Tots els equipaments hauran de tenir doble font d'alimentació per garantir el seu funcionament en cas de caiguda d'una de les fonts.
- El sistema proposat haurà d'optimitzar la velocitat de transmissió de dades entre els diversos components del mateix.
- Es valorarà la relació de TB per servidor de fitxers, la compactació i les possibilitats d'ampliació en vertical de la solució proposada.
- No s'acceptaran solucions amb un "únic punt de fallada" que pugui provocar l'aturada del sistema. Cal que el sistema proposat tingui una robustesa tal que l'avaria d'un qualsevol dels seus components no provoqui l'aturada de tot el sistema.
- Es farà la instal·lació i configuració de tots els elements dels que consti la solució proposada en el CPD del Campus de Ciutadella. Totes les instal·lacions es realitzaran en horari laboral. Aquestes instal·lacions es duran a terme seguint les indicacions del responsable del contracte i comprendran com a mínim:
    - La configuració de les connexions de gestió remota i la seva verificació.
    - Un test basic del maquinari i de la seva configuració.
    - L'elaboració i el subministrament de la documentació sobre la solució implantada, la qual haurà d'incloure un informe de rendiment.

6. Garantia

El contracte inclourà un termini de garantia mínim de 3 anys.

En cas de detecció d'una avaria per part dels serveis tècnics de la UPF, aquests contactaran amb l'empresa adjudicatària per reportar-la en horari 24x7 (qualsevol dia de l'any a qualsevol hora).

Els temps d'atenció a les avaries per al material instal·lat al CPD seran els següents:

- Sistemes d'emmagatzematge i connectivitat: 13x7 NBD per a tots aquells components amb més d'un xassís. 24x7 per a components amb un únic xassís.
- Sistema de càlcul: 13x5 NBD (dies laborables de 8 a 17 hores, següent dia laborable fora d'aquest horari)

El recanvi de qualsevol element avariat es realitzarà sempre utilitzant material original del mateix fabricant. Inclourà despeses en material, ports (fins al CPD en tots els cassos), mà d'obra, desplaçament, configuració i posta en marxa, així com qualsevol altre cost que pugui aparèixer.

Apart de la substitució i reparació de l'equipament malmès, la garantia inclourà també les següents facilitats:

- Disponibilitat de les noves versions de programari que el fabricant alliberi per al maquinari instal·lat a la UPF.
- Suport d'un mínim de 15 hores anuals on-site per a configuració i suport del programari.
- Obertura d'incidències sobre mal funcionament dels equips. Possibilitat d'escalar les incidències al fabricant en cas que sigui necessari.
- Obertura de consultes sobre el funcionament o la implementació de noves funcionalitats en els equips i els programaris.

Posta en marxa de l'equipament:

El contratista, un cop instal·lada i configurada la solució, haurà de fer una presentació als tècnics que utilitzaran l'equipament sobre el seu ús i característiques, amb la finalitat que l'ús que se'n faci sigui l'òptim, tant pel que fa al maquinari com al programari. Aquestes presentacions seran com a mínim de 21 hores per a l'optimització del sistema i 21 hores més per a la parametrització i consultes sobre el programari de gestió del clúster.

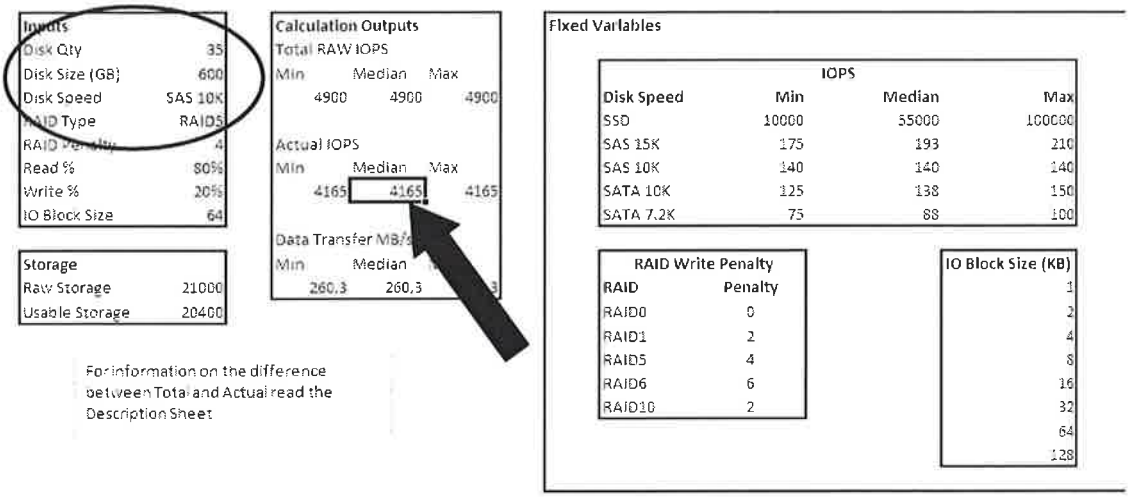
7. Metodologia de càlcul dels criteris automàtics

a) Rendiment de l'emmagatzematge

S'utilitzarà el full de càlcul que s'inclou com document annex, utilitzant-se els RAIDs disponibles una vegada descomptats els necessaris per emmagatzemar el programari del sistema.

Per utilitzar el full de càlcul només es modificaran els camps 'Disk Qty', 'Disk Size', 'Disk Speed' i 'RAID Type', la resta de cel·les del full de càlcul no es poden modificar. Com mesura del rendiment s'utilitzarà el valor 'Actual IOPS-Median'

Exemple: si una proposta inclou un apartat amb discs SAS de 600 GB i 10Krpm configurats com 7 RAIDs 5 (4+1), la proposta haurà d'incloure un total de 37 discs (cal afegir 2 discs de *hot-spare*) i el rendiment que s'obtindria és de 4165 IOPS, tal i com es mostra a la figura:



En cas que la solució proposada utilitzi discs o configuracions no contemplats en el full de càlcul el licitador haurà de fer una adaptació del mateix als models de disc que proposa. Aquesta adaptació haurà de ser validada pels tècnics de la Universitat.

b) Capacitat útil de l'emmagatzematge

S'utilitzarà el disc brut (abans de particionar) disponible en base 10 una vegada descomptats els discs de paritat, els de *hot-spare*.

Exemple: si una proposta inclou un apartat amb discs de 600 GB configurats com 7 RAIDs 5 (4+1), la proposta haurà d'incloure un total de 37 discs (cal afegir 2 discs de *hot-spare*) i es considerarà que aquest apartat proporciona un total de 7x4x600GB=16,8TB de capacitat.

c) Rendiment del sistema de còmput

El rendiment es mesurarà mitjançant la mediana del rendiment del test **SPECfp_rate_base2006**[1], pertanyent al conjunt de mesures de rendiment SPEC CPU2006, de tots els equips reportats a spec.org fins a la data de valoració, per part de la Universitat Pompeu Fabra, que comparteixin les següents característiques amb els servidors de còmput oferts (tot excloent-hi el node de login):

- Fabricant, sèrie i model del processador.
- Nombre de processadors per node.
- Nombre de cores per processador.

La xifra obtinguda serà multiplicada pel total de nodes oferts per a obtenir la puntuació final. La fórmula serà llavors:

- R = S(NP, NC) * N
- NP: Nombre de processadors
- NC: Nombre cores
- S(...): Mediana de les puntuacions al test SPECfp_rate_base2006 dels processadors que comparteixin les característiques tècniques especificades com a arguments.
- N: Nombre de nodes ofert

Només es valoraran les mesures de rendiment obtingudes mitjançant tests reals. Tot i que la normativa de SPEC CPU2006 permet estimar el rendiment, cap de les mesures reportades com a estimacions serà tinguda en compte.

Exemple. Suposem una oferta de 10 nodes, cadascun d'ells amb 4 processadors Intel Xeon E5-4617 cores per processador:

- Accedim al formulari de cerca del test CPU2006/SPECfp_rate_base2006 seguint els següents passos:
- Accedim a http://www.spec.org/cgi-bin/osgresults?conf=rfp2006&op=form
- Omplim l'opció **# Chips** amb el valor "4", l'opció **# Cores per Chip** amb el valor "6" i l'opció **Processor** amb el valor "Intel Xeon E5-4617" i pitgem **Fetch Results**.
- Calculem la mitjana dels rendiments reportats a la columna **Baseline**. Al moment de la redacció d'aquest annex, les puntuacions eren de 733, 708, 733, 731, 734 i 735. La seva mediana és 733.
- Multipliquem el valor obtingut pel nombre total de nodes oferts per obtenir una valoració final del rendiment del sistema de 7330 punts.

[1] El raonament per utilitzar aquest criteri queda reflectit a:
http://www.cisco.com/en/US/solutions/collateral/ns340/ns517/ns224/ns944/SPECCPU2006_overview.pdf

8. Pressupost

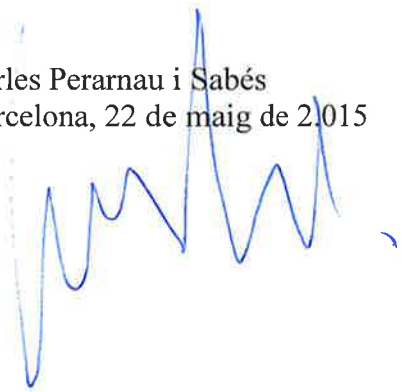
Per a la determinació del preu de sortida, s'ha realitzat una estimació d'altres equipaments de característiques similars existents al mercat.

- Cost aproximat del sistema d'emmagatzematge: ..... 80.000 €
- Cost aproximat del sistema de còmput: ..... 80.000 €
- Cost aproximat del programari: ..... 20.000 €

Atesa la diversitat de possibles ofertes, cadascuna de les quals amb composició d'equipament diferent, no és possible estimar un pressupost en base a preus unitaris de l'equipament.

El responsable del contracte,

Carles Perarnau i Sabés
Barcelona, 22 de maig de 2015



9. Annexos

Components de còmput

Model	Unitats	Components
IBM Bladecenter H (8852FT4)	2	2 x Ethernet SM 44W4406 InfiniBand HHS 46M6008
IBM HS22 (7870FT2)	27	2 x CPU: Intel(R) Xeon(R) E5645 @ 2.40GHz RAM: 96GB
IBM HS22 (7870K4G)	2	2 x Intel(R) Xeon(R) E5620 @ 2.40GHz RAM: 48GB
IBM X3650M3 (7945FT1)	1	2 x Intel(R) Xeon(R) X5650 @ 2.67GHz RAM: 48GB Mellanox Technologies MT27500 Family [ConnectX-3] 2 x Broadcom Corporation NetXtreme II BCM5709 Gigabit Ethernet (rev 20)

Components d'emmagatzemament

Model	Unitats	Components
DELL PowerEdge R720	4	2 x Intel(R) Xeon(R E5-2650 0 @ 2.00GHz RAM: 64GB 3 x LSI Logic / Symbios Logic SAS2008 PCI-Express Fusion-MPT SAS-2 [Falcon] (rev 03) Mellanox Technologies MT26428 [ConnectX VPI PCIe 2.0 5GT/s - IB QDR / 10GigE] (rev b0) 4 x Broadcom Corporation NetXtreme BCM5720 Gigabit Ethernet PCIe
DELL DELL Storage MD 3200	4	Doble controladora 12 x disk 3GB SAS NL
DELL PowerVault MD 1200	12	12x disk 3TB SAS NL

Components de connectivitat

Model	Unitats	Components
MELLANOX MIS5030Q	1	14 ports ocupats de 36

Llicenciat de programari

Model	Unitats	Components
Llicències client GPFS	2.184	IBM General Parallel File System on x86 Architecture Client 10 Processor Value Units (PVUs) License
Llicències Server GPFS	252	IBM General Parallel File System on x86 Architecture Server 10 Processor Value Units (PVUs) License