Treball Final de Grau

GRAU D'ENGINYERIA INFORMÀTICA, ENGINYERIA DE COMPUTADORS.

Universitat Oberta de Catalunya

DISSENY I IMPLEMENTACIÓ D'UN CLÚSTER HIGH PERFORMANCE COMPUTING

Javier Ortega Martín

Director: Francesc Guim Bernat
Lliurament: PAC3
Realitzat a: Arquitectura de Computadors i Sistemes Operatius.

Barcelona, 25 de juny del 2014

Índex

1 INTRODUCCIÓ.

Un clúster és tracta d'un conjunt de sistemes informàtics que treballen dins d'una mateixa xarxa amb un objectiu comú. Aquests tipus de sistemes informàtics denominats Sistemes paral·lels s'utilitzen principalment en les organitzacions destinades a la defensa federals i/o laboratoris d'investigació, tot i així estan disponibles comercialment des de fa molts anys.

Aquests sistemes generalment son tractats com a gegants difícils d'utilitzar on el programa sovint requereixen coneixements especialitzats depenent de l'arquitectura del sistema.

La indústria de la Supercomputadora[1] ha arribat fins a la recent tendència de construir Supercomputadors de clústers d'estacions de treball o fins i tot amb computadores personals amb el maquinari bàsic.

Aquests tipus d'agrupacions d'ordinadors son denominades Grids[2] o Clústers[3] (més endavant s'explica la diferencia entre aquestes), on tots els seus ordinadors integrants treballen amb una fi comuna. Aquests ordinadors agrupen maquinari de procés, maquinari de xarxa de comunicació i programari; tots els integrants han de tenir el mateix programari per tal de treballar conjuntament com si fossin un únic sistema. El principal motiu per realitzar aquestes agrupacions és poder efectuar el processament de la informació de forma més eficient i ràpida.

Cal destacar que en aquests tipus de sistemes existeixen diverses premisses fonamentals, la primera és l'Alt rendiment per tal d'aprofitar els recursos adientment i evitar els colls d'ampolla. La segona l'Alta disponibilitat, els clústers estan dissenyats per tal de garantir l'ús de recursos tot i assumint la caiguda o errada d'un element. La tercera l'equilibri de la carrega on és fa un repartiment i distribució de les tasques equitatiu. I finalment l'escalabilitat per tal de poder assumir el creixement de la infraestructura sense haver de fer canvis que afectin a la resta del clúster.

Un clúster d'alt rendiment és un conjunt d'ordinadors que està dissenyat per donar altes prestacions en quant a capacitat de càlcul. Els principals motius per utilitzar un clúster d'aquest tipus són la grandària del problema per resoldre i el preu de la màquina necessària per resoldre-ho. Per mitjà d'un clúster es poden aconseguir capacitats de càlcul superiors a les d'un ordinador més car que el cost conjunt dels ordinadors del clúster.

Fent un repàs a la recent historia dels sistemes informàtics aquests eren molt costosos, grans, lents, pesats i destinats a l'usuari expert. Poc a poc és van anar afegint millores com per exemple les Interactive Shells[4], el Time Sharing[5], Creació de terminals i la considerable reducció de la grandària del maquinari. A partir dels anys vuitanta es millora notablement les prestacions i volum dels anomenats Microcomputadors[6]. En aquesta línia les estacions de

[1] Equip en la primera línea de la capacitat de transformació contemporània.
[2] Tecnologia innovadora que permet utilitzar de forma coordinada tot tipus de recursos.
[3] Conjunts d'ordinadors construïts mitjançant la utilització de maquinaris comuns.
[4] Tasques interactives, s'utilitzen per a dos tasques o per al manegar les comades del terminal del sistema.
[5] Temps compartit, compartició dels recursos entre usuaris a través de la multiprogramació i la multitasca.
[6] Un microordinador és un barat i petit ordinador amb un microprocessador com a CPU.

treball han marcat un factor fonamental en el desenvolupament del maquinari de baix preu, ja que és important remarcar que econòmicament un conjunt de estacions de treball pot aconseguir una potencia major de còmput mitjançant el pas de missatges, tot i que tecnològicament existeixin limitacions.

Per norma general les màquines d'un clúster és troben dins d'un espai físic proper, per tal de poder treballar dins d'una xarxa LAN[7] i així garantir la millor comunicació entre els diferents elements que conformen el clúster. Per altra banda el concepte Grid tracta aquests sistemes en agrupacions d'ordinadors units per xarxes WAN[8], i aquests en alguns casos son tractats com un clúster de clústers. Si ben cada vegada més la tecnologia i els costos permeten aquestes aproximacions, els esforços i la complexitat d'utilització de desenes, centenars o milers és molt gran.

Les sigles HPC[9] van relacionades a l'alt rendiment, és considera que un ordinador és d'alt rendiment si empra múltiples processadors i a més a més aquests estan connectats entre si per una xarxa, de tal manera que aconsegueix millorar el rendiment d'un únic processador. És habitual considerar l'ús de múltiples processadors connectats alhora, aquesta manera s'anomena computació paral·lela.

Cal destacar que tot i aquestes tècniques d'interconnexió de processadors milloren moltísim el rendiment, el desenvolupament de nous microprocessadors ha evolucionat de manera exponencial des de fa molts anys.

Finalment, fent un passeig per la historia, mirant al present i segons aquesta experiència obtinguda fent una predicció al futur, trobem estudis que diuen que les màquines de major rendiment en l'actualitat (des del 2010) utilitzen centenars de milers de nuclis de processament i són capaços de fer 1 PetaFLOP[10] d'operacions (10^{15} operacions de punt flotant per segon).

En màquines 10 anys enrere això és 1.000 vegades inferior, el que vol dir que una simulació en una màquina Peta[11]FLOP hauria necessitat gairebé 3 anys en ordinador d'una dècada enrere.

Segons els experts que han realitzat la predicció pel 2020, apareixeran ordinadors de tipus EXA[12]escala (capaços de fer 10^{18} operacions de punt flotant per segon) o 1.000 vegades més ràpid que els ordinadors de l'actualitat. El futur del HPC s'enfronta a una sèrie de desafiaments on aquestes màquines EXAescala es possible que donin a conèixer en diferents formats d'arquitectures, el que si es pot predir és que segurament hi haurà un impacte important sobre el programari i s'hagi de tornar a escriure.

[7] LAN, Xarxa d'àrea local
[8] WAN, Xarxa d'àrea extensa
[9] High Performance Computing
[10] Un operació en Coma Flotant és una operació matemàtica amb un número decimal com a exponent.
[11] Prefixe del sistema internacional que indica un factor del 10^{15}
[12] Prefixe del sistema internacional que indica un factor del 10^{18}

1.1 Àmbit del projecte

Els problemes computacionals en ciències i enginyeria requereixen els ordinadors més potents disponibles en l'actualitat. L'ampli desplegament de múltiples nuclis i arquitectures de diversos nuclis ha posat en relleu la necessitat d'aprofitar les tècniques de computació paral·lela. Tal i com s'esmenta en la introducció, aquest tipus de sistemes s'utilitzen diàriament a l'entorn de la recerca, la computació d'alt rendiment és gestionada activament i ofereix servei a professors, investigadors, estudiants i afiliats. Amb centenars d'usuaris registrats, catalogats i agrupats per grups de recerca, el servei de HPC dona suport a moltes característiques i programari per beneficiar als usuaris en el desenvolupament dels projectes.

Actualment la recerca és base en àrees amb arquitectures heterogènies, algoritmes paral·lels, l'optimització del rendiment, i una varietat d'aplicacions de computació intensiva, com la del desenvolupament del cervell, anàlisi de dades, robòtica, intel·ligència artificial, desenvolupament de jocs digitals, transport, estudi del so i la música, tractament d'imatges, biomedicina, entre d'altres.

Donat que l'entorn és molt heterogeni i degut a que les capacitats dels usuaris son molt diferents, a causa de que treballen en diferents grups de recerca i en diversos àmbits de la multiprogramació; el clúster ha de trobar el punt entremig per tal de satisfer totes les necessitats, i fer una distribució del seu ús de manera equitativa. L'alt rendiment (HPC) s'estén per algoritmes, programari, eines i aplicacions que aprofiten aquestes plataformes d'arquitectures multinucli.

Així doncs és té en compte el número total de processadors disponibles, memòria física, els canals de comunicació i el sistema d'emmagatzemament (remot o local) estigui a l'abast de tothom.

1.2 Marc del projecte

El clúster actual està format per una màquina principal HEAD Node més 19 Nodes de càlcul de la gama SGI[13] Altix[14] XE Servers and Clusters, unides en una topologia de màster subordinat, amb un total de 40 processadors Dual Core i aproximadament 160Gb de RAM.

Tots els nodes disposen de connectivitat Gigabit Ethernet i tenen per defecte instal·lat un sistema operatiu Novell SuSE Enterprise Server SP1, com a interfície d'administració s'empra Platform Manager 5.7

El servidor master (head node) que actua com a Servidor NFS de fitxers per a tots els nodes, disposa d'un gestor de cues que actualment està discontinuat i sense suport, és tracta de la versió 1.5.6 de Fura. Fura és un middleware[15] Grid autònom que permet l'habilitació i la dis-

[13] Silicon Graphics, Inc.
[14] Nom de la generació de servidors dedicats a la informàtica tècnica i d'alt rendiment.
[15] Capa d'abstracció de programari per interactuar o comunicar-se amb altres aplicacions.

tribució d'aplicacions de recursos computacionals heterogenis. Fura compta amb una instal·lació de GUI basat en un web assistent guiat per la seva configuració.

Aquest clúster bàsicament està destinat a:

- Escombrada paramètrica, treballs de computació en paral·lel que consisteixen en l'execució de múltiples iteracions de la mateixa comanda utilitzant diferents valors d'entrada i els arxius de sortida.
- Tasques MPI, programació concurrent per aportar sincronització entre processos i permetre l'exclusió mútua,
- Simulacions ANSYS, Programari de simulació que permet a les organitzacions predir amb confiança com els seus productes funcionaran en el món real.
- Execucions atòmiques, són execucions d'un sol shell script o programa en C sense paral·lelitzar, que normalment requereix molt de temps d'execució i elevades prestacions de CPU.

1.2.1 Servidor Head node

- SGI Altix XE 240
- 2 procesadors Xeon Dual-Core
- 8 Gb de RAM
- Disc intern de 210Gb en Raid-1
- 2 x RJ45 10/100/1G Ethernet (Intel® 82563EB)
- PCI card, Fiber Channel PCIX-FC-2POPT-B

1.2.2 Servidor de càlcul

- 19x SGI Altix XE XE210
- 2 processadors Xeon Dual-Core
- 8 Gb de RAM
- Disc intern de 210Gb
- 2 x RJ45 10/100/1G Ethernet (Intel® 82563EB)

1.2.3 Sistema d'emmagatzemament

- SGI InfiniteStorage IS4000
- 5Tb bruts nets de capacitat.
- 4Gb Fibre Channel.

1.2.4 Electrònica de xarxa

- SMC SMC 8848M Tiger Stack II
- 48x 10/100/1000M Base-TX LAN RAJ-45
- 2x active fiber (SFP) slot

1.3 Motivació

La principal motivació d'aquest projecte és la de substituir l'actual clúster de computació pels següents motius:

- Renovació del maquinari, els antics servidors no disposen de manteniment i tecnològicament estan obsolets, actualment el sistema RAID de la safata de discos està offline a causa de la pèrdua de dos dels setze discos.

- Renovació del programari, l'empresa que gestiona el programari del clúster ja no existeix i aquest no està mantingut davant de qualsevol cas de suport o d'incidència. Cal destacar que la gestió del programari i instal·lació de nous paquets era gestionada única i exclusivament pel contracte de suport amb el fabricant, que especificava explícitament la obligatorietat d'utilització els seus repositoris oficials.

- Renovació del programari del gestor de cues que ocasiona tasques descontrolades d'usuaris que l'han sobrecarregat, provocant la penjada dels nodes i la pèrdua de dades i de temps.

- Reducció i aprofitament de la despesa l'energia, l'antic clúster consumeix un volum anual excessiu d'energia pel rendiment obtingut.

- Sistema no escalable, és tracta d'un sistema que no permet l'escalabilitat a causa de l'arquitectura i maquinari emprat.

- Integració de tots els usuaris en un únic sistema.

- Estudi del rendiment per identificar els punts forts i febles de la nova instal·lació.

2 DESCRIPCIÓ DEL PROJECTE.

En aquest apartat és descriuen els diferents punts per tal de exposar els diferents problemes dels projecte.

2.1 Resum

La computació d'alt rendiment juga un paper importantíssim en un entorn com la recerca i es duu a terme mitjançant supercomputadors, aquests s'agrupen en un conjunt d'ordinadors de potència i arquitectura similar, els quals preferiblement poden ser de baix cost (com estacions de treball o ordinadors d'ús empresarial). Aquests computadors estan interconnectats per una xarxa d'altes prestacions que constitueixen un clúster.

La xarxa que conforma un clúster està pensada per obtenir una major potència de còmput, i així estalviar costos, així mateix utilitzar el pas de missatges entre processos per comunicar una màquina amb un altre, s'ha de disposar del maquinari (generalment molt costós) per dur-lo a terme.

Els sistemes Memòria Compartida tenen molts processadors amb diversos nuclis on cadascun té molta memòria compartida entre tots els processadors.
En una Màquina amb memòria compartida es poden executar processos amb moltíssims threads[16] aconseguint gran capacitat de càlcul gràcies a la paral·lelització de memòria compartida.

S'ha de tenir en compte que aquest tipus de programació presenta el problema del maquinari que interacciona en el pas de missatges, el qual és molt costós i complex. Així doncs per poder aprofitar tots el nuclis a nivell local al 100% o tenir concurrència entre processament; la entrada/sortida pot crear un problema de la latència en les comunicacions. Per altra banda com a punt favorable destaquen l'escalabilitat i cost.

2.2 Explicació del/s problema/es

El disseny del clúster no presenta gaires problemes greus, si més no l'experiència juga un paper molt important davant d'aquest tipus de sistemes. Segons aquesta premissa ens podem avançar a diversos problemes o limitacions relacionats amb la xarxa, les dades i les còpies de seguretat.

Com que el clúster disposa d'interfícies de xarxa d'1Gigabit, és previsible que el rendiment del protocol de pas de missatges com MPI[17] és veuran greument afectat. Aquest protocol va ser dissenyat dissenyat per a ser emprat en programes que exploten l'existència de múltiples processadors. En aquest context el protocol MPI pot arribar a fer un ús complet i intensiu d'estructures de xarxa, com a conseqüència d'una comunicació intensiva, les aplicacions gasten molt temps intercanviant informació entre els processos, provocant problemes tant en el sistema de comunicació, com a l'entrada i sortida I/O.

[16] Fils d'execució.
[17] Sistema de pas de missatges estandarditzats i portàtil dissenyat per un grup d'investigadors.

La solució de backup definida i basada en Snapshots del sistema de fitxers, no té en compte la recuperació d'un determinat fitxer per un usuari autònomament. Si és vol recuperar un fitxer l'administrador ha de muntar el snapshot i buscar aquest.

2.3 Anàlisi de riscos

Per tal de garantir l'ample de banda les targetes dels nodes són d'1Gigabit i aquestes és connecten directament al Switch de 48 ports, i que alhora aquest és connectat amb la cabina de discos amb un enllaç de 10G. Com que la infraestructura només disposa de 11 nodes el problema no és gaire greu. N'obstant si tenim en compte des del punt de l'escalabilitat que el clúster pot créixer entre 20 o 30 nodes més, ens podem trobar amb un problema d'ample de banda en la connexió 10G. En aquest punt l'Stacking de Switchos pot supossar un problema.

Per tal de cobrir al 100% la viabilitat i estabilitat del projecte és contracta el servei de suport 13x5 del maquinari a 5 anys, per cadascun dels elements dels clúster (servidors + Electrónica de xarxa). El temps de resposta per part del proveïdor serà del dia següent laborable amb desplaçament de tècnic sense cost afegit.

2.4 Conclusions prèvies a l'inici del projecte

El muntatge del maquinari és relativament senzill, s'utilitzaran dos armaris de tipus Rack[18] i és cablejaran les xarxes definides a la fase de disseny. Per altra banda la configuració de RAIDs, configuració de les BIOS i IPMI ben fetes donaran un valor afegit al clúster i facilitarà la tasca d'administració en el dia a dia.

Per normal general el programari s'instal·la directament dels repositoris estables del Sistema Operatiu, per tal de assegurar l'estabilitat de les aplicacions i garantir la consistència dels serveis.

Per tal d'aprofitar els recursos al màxim i garantir la seguretat és segmenta la comunicació en diferents xarxes de comunicació.

Cada usuari disposarà del seu propi espai en disc totalment protegit per els permisos del sistema per deixar les seves simulacions. A més a més disposaran d'un directori compartit en mode lectura amb programari que no estigui disponible als repositoris. Tots dos directoris seran accessibles en tots els nodes de còmput.

L'objectiu del projecte és satisfer i facilitar als investigadors d'un entorn ràpid, accessible, fiable, estable i robust.

Per motius de simplicitat a priori només s'estableix un únic punt d'accés al sistema per l'usuari, d'aquesta manera és garanteix la seguretat i el control d'accés al clúster mitjançant el directori d'usuaris de l'organització.

[18] Recinte estandarditzat per muntar múltiples mòduls d'equips, cada mòdul té un panell frontal que és de 19 polzades

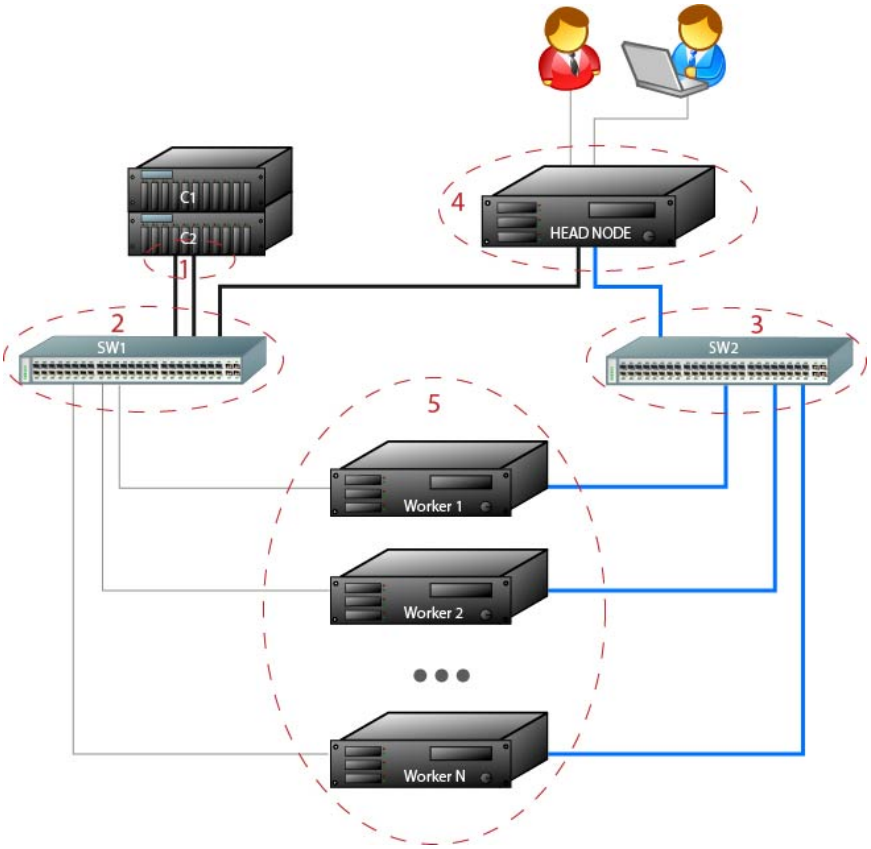
Així mateix la formació de l'usuari ha de ser primordial, és per això que aquest disposarà de la documentació bàsica de la infraestructura i a més a més de guies i recomanacions de bones pràctiques per fer un bon ús del clúster.

3 PROPOSTA TÈCNICA.

S'identifiquen tots els objectes que conformen el projecte, així com els recursos involucrats que representen la plataforma tecnològica del clúster. A més és defineixen les fases a alt nivell més importants a assolir.

3.1 Descripció breu del projecte

El present document contempla la implantació d'un nou sistema de càlcul que estarà format per 11 servidors físics amb una capacitat de càlcul suficient per tal de poder satisfer els requeriments dels usuaris. A banda d'aquests nodes[19], s'afegeix un node físic més per tal d'hostatjar les màquines virtuals que donaran servei per dur a terme les tasques de gestió del clúster. A més a més, mitjançant l'electrònica de xarxa els nodes es connectaran al sistema d'emmagatzemament per un enllaç d'alt rendiment.



Esq. 1 Esquema general dels elements principals.

A la figura és mostra la configuració general del tots els elements que conformen el clúster. Principalment és destaca la cabina de discos (1) la qual és connecta al Switch 1 (2) mitjançant una interconnexió de 10G, tots els nodes (5) és connectaran a aquest Switch 1 mitjançant ethernet 1000 Mb/seg Full Duplex per tal d'assegurar el major ample de banda amb l'emmagatzemament. Per altra nada el Switch 2 (3) és destinarà per tal de crear les

[19] Nodes de còmput on s'executen totes les tasques dels usuaris.

diferents xarxes de gestió i la comunicació entre les nodes de càlcul i el head node (4). La tasca principal d'aquest últim és la d'hostatjar màquines virtuals per tal de dur a terme tota la gestió del clúster com poden ser la gestió de tasques i cues, desplegament de paquests, monitorització i passarel·la de autenticació pels usuaris.

3.2 Elements de maquinari

A continuació és descriuen tots els elements de maquinari descrits a la figura anterior, d'aquesta manera és dona una visió més acurada del maquinari emprat i les funcions de cadascun dels elements.

3.2.1 Sistema d'emmagatzemament

El clúster s'haurà d'integrar amb la tecnologia de emmagatzemament NetApp, concretament amb una FAS de la sèrie 3100 amb el sistema Data ONTAP, és pot satisfer alhora les diverses necessitats SAN i NAS, emmagatzematge primari i secundari alhora que proporciona un alt nivell de disponibilitat de les aplicacions, des de les operacions crítiques del negoci a les aplicacions tècniques.

El nivell de flexibilitat, disponibilitat i rendiment d'aquest fabricant aporta al clúster un valor afegit molt gran. Permet connectar a l'entorn una gama de sistemes heterogenis (incloent Windows, UNIX i servidors Linux); l'alta disponibilitat i recuperació davant desastres proporcionen al sistema de la protecció de dades que li permet protegir les dades crítiques mitjançant altres mitjans d'emmagatzematge. Disposa de la capacitat de fer Snapshots[20] per reduir els temps de còpia de seguretat en minuts. El sistema de fitxers té la peculiaritat de paritat dual (RAID-DP), és tracta d'un RAID 6 d'alt rendiment que ofereix una millor protecció de dades que la utilització de la capacitat de RAID 5 o 1+0.

Amb aquesta tecnologia és pot combinar Fibre Channel d'alt rendiment i de la gran capacitat de les unitats de disc SATA, en els nivells d'emmagatzematge de rendiment i cost òptim. Dins del mateix sistema, és pot consolidar la relació bloc i emmagatzematge d'arxius amb FCP[21], iSCSI[22], NFS[23] i protocols d'emmagatzematge CIFS[24] mitjançant ethernet o Fibre Channel.

El clúster emprarà el protocol de compartició de fitxers NFS fent servir la cabina de discos com a servidor NFS i els nodes de càlcul com a clients. Els clients accedeixen de forma remota a les dades que es troben emmagatzemats en el servidor.

Els usuaris no necessiten disposar d'un directori HOME[25] en cadascun dels nodes dels clúster. Els directoris HOME estaran en la cabina de discos per posteriorment poder accedir a aquests des de qualsevol node mitjançant la infraestructura de xarxa. A banda d'això

[20] Snapshot, estat d'un sistema en un moment determinat
[21] Fibre Channel o FC, és una tecnologia de xarxa d'alta velocitat per a connectar l'emmagatzematge de dades.
[22] iSCSI, Éstandar que permet l'ús del protocol SCSI mitjançant xarxes TCP/IP
[23] NFS, protocol de sistema de fitxers en xarxa, permet a una client accedir a fitxers a través de xarxa..
[24] CIFS,estàndard compartició d'arxius a través d'una xarxa. Protocol natiu per compartir arxius en Windows2K.
[25] HOME, és un directori en un sistema operatiu multi-usuari que conté els arxius dels usuaris del sistema.

també disposaran d'un directori a tots els nodes de càlcul de només lectura amb programari compilat a mida per ús comú.

3.2.2 Switch de dades

El switch de dades Netgear GS752TXS-100 EUS s'encarregarà de connectar-se directament amb la cabina de discos. És tracta d'un switch de 48 ports ethernet i 4 ports 10GB/SFP+ especial per la connexió Attach Direct de NetApp.

Gestionable a nivell Capa2+ que permet principalment la gestió Spaning Tree Protocol, creació i gestió de VLAN[26], GVRP[27], 802.1v[28], Quality of Service[29], Jumbo Frames[30] entre d'altres.

Aquest tipus de switch està dissenyat per una xarxa d'entre 50 i 200 nodes per un alt rendiment, escalabilitat de la xarxa i fiabilitat sense complexitats.

Fig. 2 Vista general del Switch de dades.

3.2.3 Switch de gestió

El switch de gestió Netgear GS724TS s'encarregarà de la gestió IPMI[31], és tracta d'un switch de 24 ports 10/100/1000 Mbps que permet fer stacking amb altres switchos per dur a terme futures ampliacions. La funcionalitat bàsica que s'utilitzarà és VLAN.

Com l'anterior és gestionable a nivell Capa 2+ que permet principalment la gestió Spaning Tree Protocol, creació i gestió de VLANs, GVRP, 802.1v, Quality of Service, Jumbo Frames entre d'altres.

Fig. 3 Vista general del Switch de gestió.

[26] Mètode de crear xarxes lògicament independents dins d'una mateixa xarxa física.
[27] És defineix una aplicació en l'estàndard IEEE 802.1Q que permet el control de les VLAN
[28] Classificació VLAN per protocol i per port
[29] És el rendiment global d'una xarxa d'ordinadors, en particular el rendiment vist pels usuaris de la xarxa.
[30] Estén les trames de xarxa ethernet a 9.000 bytes.
[31] Interfície estàndard d'equips utilitzada per a l'administració fora de la xarxa dels sistemes informàtics

3.2.4 Servidor Head Node

El servidor en format de 2U amb font d'alimentació redundant[32] de 1400W 80PLUS GOLD amb certificat d'eficiència energètica.

El servidor o node head, disposa de dos processadors Opteron Abu Dhabi 6344 a 2,6Ghz (3,2 en mode turbo), dotze nuclis d'arquitectura x86_64 amb un controlador de memòria integrat al xip i amb una interfície HyperTransport 3.0[33] amb un consum mitjà entre 80W i 115W d'eficiència energètica.

Incorpora 2MB de cache per nucli i cada processador té 4 canals de BUS de 6,4G transaccions per segon a cadascun sense colls d'ampolla.

La placa base dual H8DGU-F amb chipset AMD SR5690 + SP5100 Chipset que suporta 2 processadors Opteron Abu Dhabi de nova generació de fins 16 nuclis. Incorpora 4 canals d'accés a la memòria DDR per processador amb un total de 16. 7 ports USB i 2 PCI Express 2.0x8 i x16.

32GB de memòria RAM ECC a 1600Mhz de baix consum.

4 discos durs de tecnologia SAS[34] de 15.000rpm i 6Gb/s de 300GB Seagate Cheetah en RAID5.

Lector/Regrabador DVD-RW a x8 lectura/escritura

2 ports ethernet de xarxa integrats en placa 10/100/1000 Full Duplex
4 ports ethernet de xarxa en targeta Gigabit.
1 port ethernet de xarxa per la gestió IPMI per l'administració KVM[35] over LAN.

Fig. 4 Vista principal del head node.

[32] Font d'alimentació redundant conté dos (o més) unitats de subministrament d'energia al seu interior. Cada font d'alimentació és capaç de subministrar energia a tot l'equip, només una funciona durant el seu ús.
[33] Tecnologia de comunicacions bidireccional, que funciona tant en sèrie com en paral·lel.
[34] Serial Attached SCSI, una evolució de SCSI paral·lel en una interfície de perifèrics de punt a punt en sèrie.
[35] És una infraestructura de virtualització per al nucli de Linux que la converteix en un hipervisor.

3.2.5 Servidor de càlcul

El servidor en format de 4U amb font d'alimentació redundada de 1400W 80PLUS GOLD amb certificat d'eficiència energètica. Aquest chasis disposa de 4 ventiladors per tal de evitar el sobreescalfament i millorar el rendiment de la capacitat de càlcul.

El servidor o node head, disposa de dos processadors Opteron Abu Dhabi 6378 a 2,4Ghz (3,2 en mode turbo), setze nuclis d'arquitectura x86_64 amb un controlador de memòria integrat al xip i amb una interfície HyperTransport 3.0 amb un consum mitjà entre 80W i 115W d'eficiència energètica. La peculiaritat d'aquest processador està en el SPECfp_rate_base2006 de 738 per node.

Incorpora 2MB de cache per nucli i cada processador té 4 canals de BUS de 6,4G transaccions per segon a cadascun sense colls d'ampolla.

La placa base dual H8QG7-LN4F amb chipset AMD SR5690 + SP5100 Chipset que suporta 2 processadors Opteron Abu Dhabi de nova generació de fins 16 nuclis. Incorpora 4 canals d'accés a la memòria DDR per processador amb un total de 16. 7 ports USB i 2 PCI Express 2.0x8 i x16.

256GB de memòria RAM DDR3 ECC a 1600Mhz de baix consum.

1 disc dur de tecnologia SSD[36] de amb una lectura/escriptura de 555/510MB/s i 6Gb/s de 240GB.

2 ports ethernet de xarxa integrats en placa 10/100/1000 Full Duplex
4 ports ethernet de xarxa en targeta Gigabit.
1 port ethernet de xarxa per la gestió IPMI per l'administració KVM over LAN.

Fig. 5 Vista d'un node de càlcul.

[36] Una unitat d'estat sòlid és un dispositiu d'emmagatzematge de dades que utilitza una memòria no volàtil.

3.3 Objectius a assolir

Els principals objectius són els de proveir a l'usuari d'un entorn de càlcul d'alt rendiment i millorar la disponibilitat per damunt de la proveïda per un sol ordinador, que el permeti executar les seves tasques per tal de dur a terme els seus projectes de recerca.

Així doncs per tal d'implementar i posar en marxa tots els elements esmentats a l'apartat anterior cal establir una sèrie d'objectius que s'enumeren a continuació:

1. Prèvia instal·lació física de tot el maquinari.
2. Definició de les xarxes que estableixen la comunicació dels elements dels clúster.
3. Configuració del servidor head node:
    a. Configuració BIOS[37], RAID[38] i IPMI.
    b. Instal·lació Sistema Operatiu.
    c. Instal·lació màquines virtuals de login, desplegament de paquets, gestor de cues i monitorització i els seus serveis.
    d. Configuració de les xarxes dels servidors virtuals.
    e. Configuració de la monitorització.
4. Configuració dels servidors de càlcul:
    a. Configuració BIOS i IPMI.
    b. Instal·lació Sistema Operatiu
    c. Instal·lació llibreries desenvolupament i repositoris externs.
5. Configuració gestor de cues Open Grid Engine[39]
    a. Instal·lació al Servidor Head Node.
    b. Instal·lació dels agents als nodes de càcul.
6. Configuració servidor PXE[40] i TFTP[41] per DHCP[42].
7. Documentació d'usuari i administració.
8. Estudi del rendiment del sistema.

[37] Primer programa que s'executa a una computadora quan s'engega, carrega i inicialitza el sistema operatiu.
[38] Es basa en un sistema d'emmagatzemament de la informació que combina diversos discs durs
[39] Programari de sistema de clúster Computing Grid (conegut com a sistema de lots o cues), desenvolupat Sun Microsystems i Oracle després i diferents versions de codi obert.
[40] Entorn per arrencar els ordinadors per una interfície de xarxa independent d'emmagatzematge de dades.
[41] És un protocol de transferència molt simple similar a una versió bàsica d'FTP
[42] Protocol de xarxa que permet als nodes d'una xarxa IP obtenir els seus paràmetres automàticament.

4 PLANIFICACIÓ.

Un cop plantejat el problema i sabent tot el que cal tenir en compte pel funcionament del sistema, es procedeix a la planificació del projecte. Els següents punts d'aquest capítol consisteixen en la descripció de l'anàlisi, especificacions, fites i tasques del projecte.

4.1 Fases del projecte

En total s'han identificat 7 fases, aquestes han estat distribuïdes en tres tipus de professionals, el cap de projecte que supervisarà l'evolució d'aquest, l'enginyer de sistemes que durà a terme totes les tasques específiques com el disseny i configuració dels serveis crítics i finalment el tècnic que realitzarà totes les tasques d'instal·lació de maquinari i configuració bàsica dels sistemes.

A continuació es detallen una mica cadascuna d'aquestes fases identificades al diagrama de Gantt.

Fase	0	1	2	3	4	5	FINAL	Total
Hores	56	32	16	160	64	120	24	472

Taula resum de la dedicació d'hores a cada fase.

FASE 0: Disseny infrastructural

- Disseny de les xarxes.
- Disseny Serveis i Programari dels servidors.
- Disseny del sistema de fitxers compartit.
- Preparació del programari a instal·lar.

FASE 1: Preparació de l'entorn

- Trasllat del maquinari.
- Desembalar el maquinari i fer inventari.
  - Instal·lació del maquinari als Racks.
  - Instal·lació dels servidors.
  - Instal·lació de l'electrònica de xarxa.
  - Instal·lació del cablejat de xarxa.

FASE 2: Configuració inicial

- Configuració de les BIOS.
- Configuració dels RAID's dels Servidors.
- Configuració targetes IPMI.
- Posada en marxa del maquinari.
- Configuració inicial de les xarxes als switchos.
- Instal·lació del Sistema Operatiu Base al node Head.
- Instal·lació del Sistema Operatiu al primer node de càlcul.

FASE 3: Configuració dels Serveis

- Instal·lació del Servidor Host Xen al Head node:
  Configuració de les xarxes del Servidor de màquines virtuals.
  Instal·lació d'un Servidor Virtual amb el SO Base.
  Definició del fitxer hosts.
  Creació de l'usuari d'administració
  Securització del sistema base
  Instal·lació de l'agent de monitorització nagios (*Servidor corporatiu Nagios*)

- Clonació del Guest per crear el Servidor login:
  Configuració del client LDAP.
  Configuració de la xarxa NFS de dades.
  Muntatge dels homes dels usuaris.
  Muntatge del directori de programari compartit.
  Instal·lació de les eines gridengine-client.

- Clonació del Guest per crear el Servidor Open Grid Scheduler:
  Instal·lació del programari gestor de cues GE
  (definició de les cues, prioritats, grups de hosts i grups de recerca)

- Clonació del Guest per crear el Servidor de paquets i proxy:
  Instal·lació del Servidor Proxy Squid.
  Instal·lació Servidor PXE, TFTP i DHCP.
  Instal·lació del programari de desplegament de paquets.
  Instal·lació dels nodes mitjançant el Servidor de desplegament de paquets.
  Instal·lació i configuració de llibreries de desenvolupament comuns.

- Clonació del Guest per crear el Servidor de monitoratge
  Instal·lació de Ganglia.
  Instal·lació de PHPQstat.

FASE 4: Estudi rendiment del sistema

- Execució de les bateries de proves.

FASE 5: Documentació

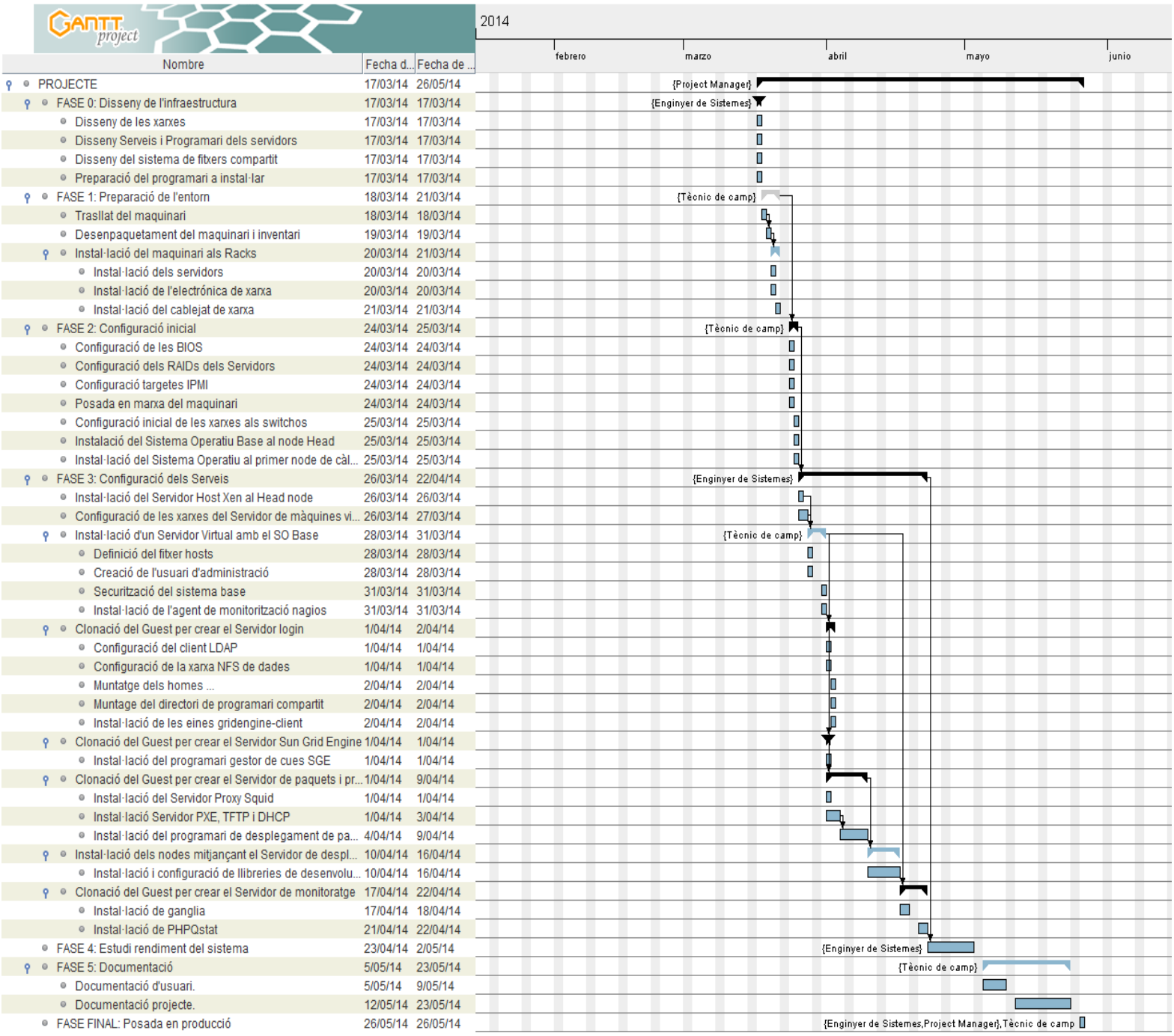
- Documentació d'usuari.
- Documentació projecte.

FASE FINAL: Posada en producció

- Lliurament del projecte.

Seguint el model d'arbre, al següent diagrama es poden visualitzar les tasques i subtasques dividides en fases del projecte. La documentació es durà a terme des del començament fins la finalització del projecte.

Al diagrama de Gantt podem visualitzar les diferents fases en les que s'ha dividit el projecte.



Nombre	Fecha d...	Fecha de ...
PROJECTE	17/03/14	26/05/14
FASE 0: Disseny de l'infraestructura	17/03/14	17/03/14
Disseny de les xarxes	17/03/14	17/03/14
Disseny Serveis i Programari dels servidors	17/03/14	17/03/14
Disseny del sistema de fitxers compartit	17/03/14	17/03/14
Preparació del programari a instal·lar	17/03/14	17/03/14
FASE 1: Preparació de l'entorn	18/03/14	21/03/14
Trasllat del maquinari	18/03/14	18/03/14
Desenpaquetament del maquinari i inventari	19/03/14	19/03/14
Instal·lació del maquinari als Racks	20/03/14	21/03/14
Instal·lació dels servidors	20/03/14	20/03/14
Instal·lació de l'electrónica de xarxa	20/03/14	20/03/14
Instal·lació del cablejat de xarxa	21/03/14	21/03/14
FASE 2: Configuració inicial	24/03/14	25/03/14
Configuració de les BIOS	24/03/14	24/03/14
Configuració dels RAIDs dels Servidors	24/03/14	24/03/14
Configuració targetes IPMI	24/03/14	24/03/14
Posada en marxa del maquinari	24/03/14	24/03/14
Configuració inicial de les xarxes als switchos	25/03/14	25/03/14
Instalació del Sistema Operatiu Base al node Head	25/03/14	25/03/14
Instal·lació del Sistema Operatiu al primer node de càl...	25/03/14	25/03/14
FASE 3: Configuració dels Serveis	26/03/14	22/04/14
Instal·lació del Servidor Host Xen al Head node	26/03/14	26/03/14
Configuració de les xarxes del Servidor de màquines vi...	26/03/14	27/03/14
Instal·lació d'un Servidor Virtual amb el SO Base	28/03/14	31/03/14
Definició del fitxer hosts	28/03/14	28/03/14
Creació de l'usuari d'administració	28/03/14	28/03/14
Securització del sistema base	31/03/14	31/03/14
Instal·lació de l'agent de monitorització nagios	31/03/14	31/03/14
Clonació del Guest per crear el Servidor login	1/04/14	2/04/14
Configuració del client LDAP	1/04/14	1/04/14
Configuració de la xarxa NFS de dades	1/04/14	1/04/14
Muntatge dels homes ...	2/04/14	2/04/14
Muntage del directori de programari compartit	2/04/14	2/04/14
Instal·lació de les eines gridengine-client	2/04/14	2/04/14
Clonació del Guest per crear el Servidor Sun Grid Engine	1/04/14	1/04/14
Instal·lació del programari gestor de cues SGE	1/04/14	1/04/14
Clonació del Guest per crear el Servidor de paquets i pr...	1/04/14	9/04/14
Instal·lació del Servidor Proxy Squid	1/04/14	1/04/14
Instal·lació Servidor PXE, TFTP i DHCP	1/04/14	3/04/14
Instal·lació del programari de desplegament de pa...	4/04/14	9/04/14
Instal·lació dels nodes mitjançant el Servidor de despl...	10/04/14	16/04/14
Instal·lació i configuració de llibreries de desenvolu...	10/04/14	16/04/14
Clonació del Guest per crear el Servidor de monitoratge	17/04/14	22/04/14
Instal·lació de ganglia	17/04/14	18/04/14
Instal·lació de PHPQstat	21/04/14	22/04/14
FASE 4: Estudi rendiment del sistema	23/04/14	2/05/14
FASE 5: Documentació	5/05/14	23/05/14
Documentació d'usuari.	5/05/14	9/05/14
Documentació projecte.	12/05/14	23/05/14
FASE FINAL: Posada en producció	26/05/14	26/05/14

Diag. 1 Diagrama de Gantt del projecte.

4.2 Anàlisi del temps requerit

El projecte començarà el primer dilluns laborable després del lliurament del present document amb la proposta, la data especificada és per al proper 17 de març del 2014. A partir d'aquesta data és durà terme la Fase 0 que consta de la definició del programari i disseny conceptual de les xarxes.

A la Fase 1 amb una duració de 4 dies és prepararà l'entorn. La següent Fase 2 és durà a terme la configuració inicial dedicant dues jornades. La següent Fase 3 és una de les més laborioses i és dedicarà gairebé 4 setmanes per tal de instal·lar i configurar tot el programari dels sistemes. Seguidament és dedicaran 7 jornades a fer l'estudi de rendiment i finalment 10 jornades per tal de construir la documentació final del projecte.

Tot i que s'han ajustat els temps de finalització de cadascuna de les fases i tasques, es calcula que la data aproximada de la implantació i configuració del clúster serà per la darrera setmana del més de març, donant un marge per qualsevol tipus d'incidència que pugui sorgir durant tot el procés. A partir de la primera setmana de maig és destinarà el temps a elaborar un estudi de rendiment i revisar i/o modificar la documentació referent. S'han inclòs 5 jornades de més a la Fase 5 per tal de compensar qualsevol improvis.

4.3 Fites

Segons les fases especificades a l'apartat de fases del projecte queden definides les següents fites per dur a terme el projecte, aquestes venen definides genèricament per els següents punts:

- Conèixer el problema.
- Disseny de la infraestructura.
- Preparació de l'entorn.
- Configuració inicial.
- Configuració específica dels serveis.
- Estudi de rendiment i proves.
- Recollida de documentació.

4.4 Data de lliurament

Segons el diagrama de Gantt es preveu que la data de finalització del projecte és l'26 de maig de 2014 que coincideix amb la data de lliurament; si més no, s'inclourà la fase de proves i rendiment si no és donen interferències durant el procés d'implantació.

4.5 Recursos a emprar

- Sistema Operatiu GNU/Linux Debian 7.0.
- Eina d'instal·lació centralitzada del clúster i automatització del clonatge dels nodes.
- Eines de monitorització Ganglia i PHPQstat.
- Seguretat iptables.
- Servidor gestor de cues Grid Engine.

- Servidor LDAP corporatiu.
- Servidor de Backup.
- Servidor de fitxers NFS
- Eines de desenvolupament x86 Open64, GCC, Python
- Llibreries matemàtiques Scipy, Atlas, ACML
- Repositoris externs Neurodebian
- High Performance Linpack

4.6 Valoració econòmica

La següent taula mostra el total de jornades de dedicació dels professionals, aquestes estan dividides en les diferents fases:

CONCEPTES	Jornades (x8h)							
	Anàlisi i disseny		Implantació			Rendiment i proves		
	Proposta	Fase 0	Fase 1	Fase 2	Fase 3	Fase 4	Fase 5	Fase F
Project Manager	5	-	-	-	-	-	-	1
Enginyer de Sistemes	2	1	-	-	18	8	5[43]	1
Tècnic de camp	-	-	4	2	2		10	1
SubTotal	**7**	**1**	**4**	**2**	**20**	**8**	**15**	**3**
TOTAL (jornades)	**8**		**25**			**26**		

Taula 1. Jornades de dedicació.

A continuació és mostra una taula dels conceptes basats en les fases descrites en el diagrama de Gantt, es mostra el número total d'hores per concepte i % del temps dedicat.

CONCEPTES	hores (h)	% hores
Anàlisi i disseny	64	13,5
Implantació	200	42,37
Rendiment i proves	208	44,06
Total	**472**	**100%**

Taula 2. Temps dedicat a cada concepte.

[43] És reserva aquest temps per qualsevol inconvenient durant el procés d'implantació.

Taula resum dels preus aplicats a cadascun dels conceptes esmentats a la Taula 2.

CONCEPTES	hores (h)	Preu hora (€/h)	Preu Total (€)
Anàlisi i disseny	64	60	3.840
Implantació	200	45	9.000
Testeig, proves i documentació	208	30	6.240
		TOTAL	**19.080**

Taula 3. Preus dels conceptes.

** S'han exclòs les hores de redacció de la documentació.*

Taula resum del preu del maquinari, consta de tots els elements principals enrackables:

Maquinari	Quantitat	Preu unitari(€)	Preu Total (€)
SUPERMICRO 24 CORE 2.6Ghz AMD OPTERON 6344 H8DGU-F	1	3.263,58	
SUPERMICRO 64 CORE 2.4Ghz AMD OPTERON 6378 H8QG7-LN4F	11	5.980,10	65.781,10
Netgear GS752TXS-100 EUS	1	933,78	
Netgear GS724TS	1	386,94	
		TOTAL	**70.365,4**

El preu final total del muntatge del maquinari, més els serveis professionals ascendeix a **89.445,40€**.

5 DESCRIPCIÓ DELS RECURSOS DE PROGRAMARI.

Al següent apartat és descriuen els diferents productes emprats i les funcionalitats principals que s'aprofitaran d'aquest. Així mateix és fa un breu descripció de l'abast de cada producte programari i els diferents requeriments.

5.1 Debian

Debian és una distribució de GNU/Linux[44] de programari lliure, no comercial. Creada pel projecte Debian l'any 1993, l'organització responsable de la creació i del manteniment de la distribució. Aquest també manté i desenvolupa sistemes GNU basats en altres nuclis.

Nasqué com una aposta per separar en les seves versions el programari lliure del privatiu.

El model de desenvolupament és independent d'empreses, creat pels seus propis usuaris no depèn de necessitats comercials. Debian no ven directament el seu programari, sinó que el posa disposició de tothom per Internet, tot i que sí que permet a persones i empreses distribuir comercialment aquest programari mentre se'n respectin el termes de la llicència.

Debian és un sistema operatiu complet, que es pot utilitzar en diverses plataformes, generalment utilitzat juntament amb el kernel o nucli Linux (tot i que hi ha versions experimentals amb altres nuclis com Hurd[45] i kFreeBSD[46]), les utilitats GNU i un gran nombre d'aplicacions i utilitats d'orígens diversos. El sistema Debian és desenvolupat per programadors i col·laboradors d'arreu del món, d'acord amb un "contracte social", que garanteix que tot el programari inclòs en aquesta distribució sigui lliure.

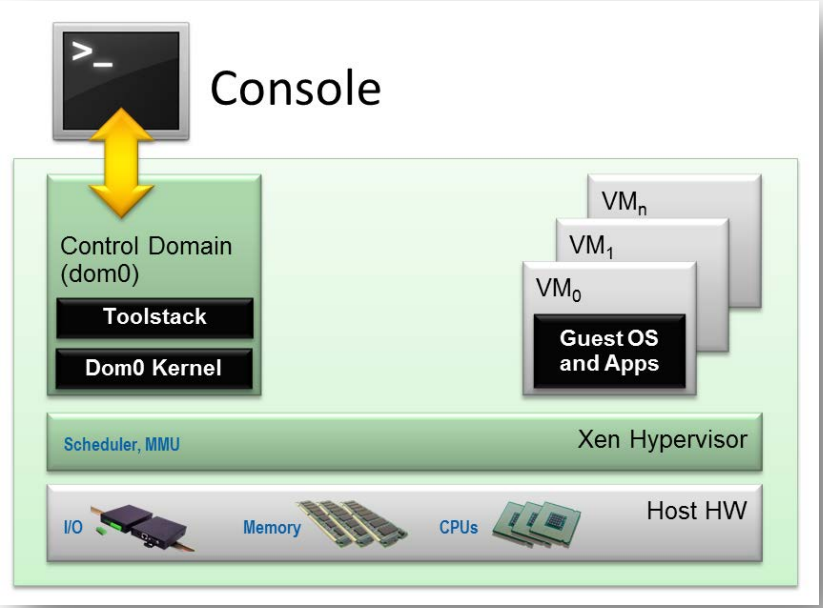
5.1.1 Funcionalitats i característiques principals

- Distribució lliure i gratuïta sense cost afegit, del Sistema Operatiu i actualitzacions
- Sistemes Operatius robust i estable; multiusuari i multiplataforma per tal de dur a terme la gestió de diversos usuaris i diferents arquitectures de processador del mercat.
- Lliure de virus i malware que no existeix en aquest tipus de sistemes.
- Milers de paquets i repositoris de programari totalment gratuït i lliure per la comunitat d'usuaris.
- El seu nucli és pot configurar a mida del sistema, així com el programari extern. Per tal d'optimitzar el funcionament a nivell de Processador o d'alguna característica en particular.
- Modular, per tal de corregir el problemes de seguretat és poden aplicar els paquets d'actualitzacions mitjançant internet.

[44] Sistema operatiu format pel nucli o kernel Linux, juntament amb les utilitats GNU.
[45] És un nucli tipus Unix que constitueix la base del sistema operatiu GNU.
[46] És un sistema operatiu que va treure el projecte Debian per les arquitectures d'ordinadors compatibles amb i486.

5.2 Xen Project

Xen és una màquina virtual de codi obert desenvolupada per la Universitat de Cambridge. El seu objectiu és poder executar instàncies de sistemes operatius amb totes les seves característiques, de forma completament funcional en un equip senzill.
Xen proporciona un aïllament segur, control de recursos, garanties de qualitat de servei i migració de màquines virtuals en viu. Els sistemes operatius han de ser modificats explícitament per córrer a Xen (encara que mantenint la compatibilitat amb aplicacions d'usuari). Això permet a Xen assolir virtualització d'alt rendiment sense un suport especial de maquinari.



5.2.1 Funcionalitats i característiques principals

- Està basat en Hypervisor[47], Xen és situa entre el maquinari i les màquines virtuals (domini U) donant una porció dels recursos físics a cadascuna d'elles.
- Gestiona la prioritat d'accés als recursos físics de la màquina mitjançant Scheduling, Quotas.
- Utilitza un sistema operatiu privilegiat de caràcter administratiu per gestionar les màquines virtuals (domini 0[48]).
- Necessita un kernel modificat tant per al domini 0 com para al domini U[49].
- En casos de virtualització maquinari HVM[50] utilitza un emulador QEMU.
- Proveeix un conjunt d'eines i una API per interactuar amb les màquines virtuals.
- Permet característiques avançades com PCI-Passtrough[51] i Live-Migration[52]

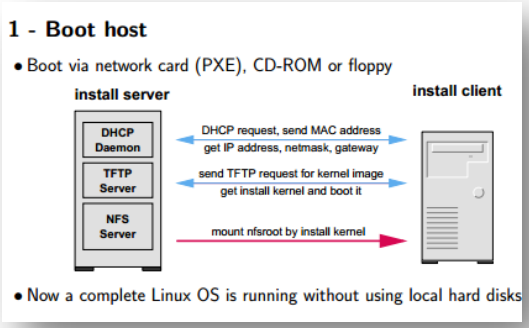
[47] Tros de programari, firmware o maquinari que crea i posa en màquines virtuals.
[48] Dom0, o domini zero és el domini inicial iniciat per l'hipervisor Xen en l'arrencada.
[49] DomU és la contrapartida de Dom0; domini sense privilegis de root que no tenen accés al maquinari.
[50] Xen HVM és de domini maquinari emulat, on el sistema operatiu no s'ha modificat de cap manera.
[51] S'utilitza per assignar un dispositiu a una màquina virtual per donar-li accés complet i directe.

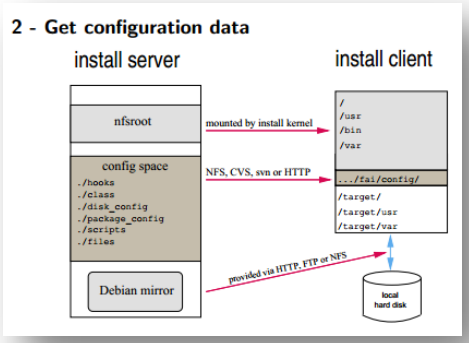
5.3 FAI Software

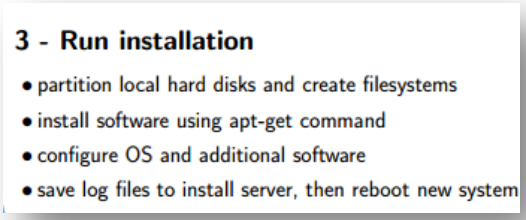
FAI va ser iniciat al 1999 a la Universitat de Colònia (Universität zu Köln), quan Thomas va haver d'organitzar una instal·lació d'un clúster Linux amb un servidor i 16 clients. Avui en dia, FAI s'utilitza regularment en entorns diferents, que van des d'una dotzena de màquines de fins a diversos milers de màquines.

FAI és un sistema no interactiu per instal·lar, personalitzar i administrar sistemes Linux i configuracions de programari en els ordinadors, així com màquines virtuals i chroot[53], des de petites xarxes d'infraestructures a gran escala, com els clústers i entorns de núvol.

És una eina per a la implementació massiva desatesa de Linux. És utilitzar a partit d'una màquina completament nova, connectant l'alimentació, i després d'uns minuts, els sistema operatiu està instal·lat i completament configurat per a les seves necessitats específiques, sense cap interacció necessària.







5.3.1 Funcionalitats i característiques principals

- Instal·lacions i actualitzacions dels sistemes operatius Debian, Ubuntu, CentOS, RHEL, SUSE, ...
- Gestió de la implementació i configuració centralitzades.
- Instal·la màquines virtuals utilitzant KVM o Xen.
- És ràpid, només es triga uns pocs minuts per a una instal·lació completa .
- Escalable, els usuaris de FAI poden gestionar les seves infraestructures informàtiques

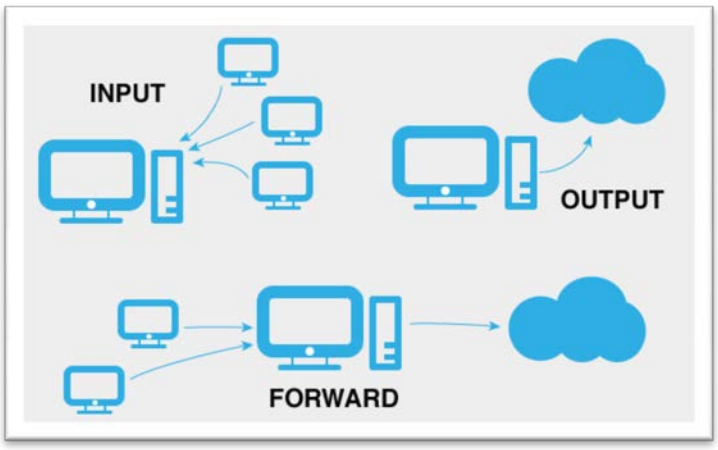
[52] Procés de moure una màquina virtual en execució o aplicació entre diferents màquines físiques.
[53] És una operació que invoca un procés canviant per a aquest i els seus fills el directori arrel del sistema.

a partir d'uns pocs ordinadors fins a diversos milers de màquines .
- És pot establir diferent maquinari i diferents requisits de configuració , no cal repetir la informació que es comparteix entre diverses màquines.
- FAI és lleuger, no hi ha dimonis especials executant i no es necessita cap configuració de base de dades. És independent de l'arquitectura ja que només depèn d'una shell , Perl i CFengine[54] scripts.
- És pot personalitzar cada petita part de la configuració a les seves necessitats locals mitjançant hooks[55].

5.4 Iptables

És un sistema de tallafocs vinculat al kernel de linux que ens permet configurar les regles de filtratge de paquets.

El funcionament de l'iptables és simple se li especifiquen unes regles amb unes determinades característiques que ha de complir un paquet, per cada regla s'especifica una acció o target. Les regles tenen un ordre i quan es rep o s'envia un paquet aquestes es recorren en ordre fins que les condicions d'una d'aquestes regles es compleixi en el paquet i la regla s'activa realitzant sobre el paquet l'acció que s'havia especificat.



[56]

5.4.1 Funcionalitats i característiques principals

- Filtrat de paquets.
- Interpretació de protocols, ports, IPs,...
- Estat dels paquets.
- Network Address translation (NAT).

[54] Cfengine és un sistema de gestió de la configuració que permet automatzitzar la administració de sistemes des d'un únic punt. És tracta d'un llenguatge d'alt nivell amb polítiques de configuració de sintaxi pròpies.
[55] Als Hooks s'especifica quines funcions s'hauran d'executar a un procés en un determinat pas.
[56] FILTER: És la taula per defecte pels paquets que es refereixen a la màquina local
INPUT: S'apliquen als paquets que tenen com a destí la màquina local.
OUTPUT: S'apliquen als paquets generats en el sistema i que són enviats a l'exterior
FORWARD: S'apliquen a paquets destinats a altres màquines que han de travessar la local.

- Infraestructura flexible i comprensible.
- Capacitat d'afegir funcionalitats mitjançant pegats.

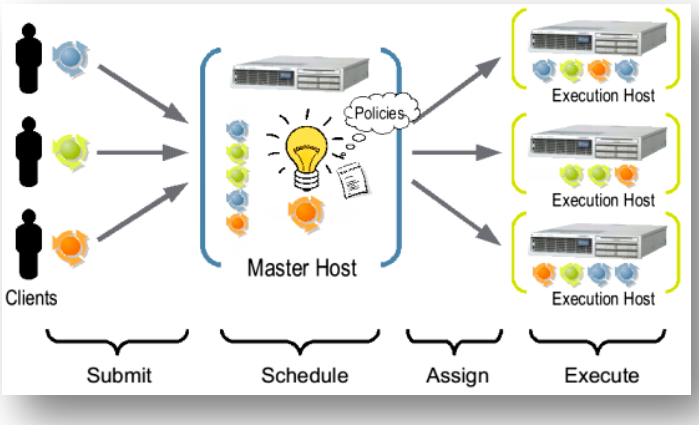
5.5 Grid Engine

Grid Engine, anteriorment conegut com Sun Grid Engine (SGE[57]), Codine (Computació Distribuïda en entorns de xarxa) o GRD[58] (Director Global de Recursos), és un sistema de clúster de computació Grid, també conegut com a sistema de lots de coles. Aquest va ser desenvolupat i recolzat per Sun Microsystems i després per Oracle. Existeixen versions de codi obert i múltiples versions comercials d'aquesta tecnologia, inicialment SUN, després d'Oracle i després d' UNIVA[59] Corporation.

El projecte original Grid Engine de codi obert va tancar la seva web al 2010, tot i que existeixen diferents versions d'aquesta tecnologia que encara són sota la seva llicència original de Sun Industry.

A partir del codi original s'han creat projectes dels quals son coneguts com 'Son of Grid Engine' i 'Open Grid Scheduler'.

Grid Engine s'utilitza típicament en entorns de clúster, granjes d'ordinadors o computadors d'alt rendiment (HPC), i és responsable d'acceptar, programar, despatxar i gestionar l'execució remota i distribuïda d'un gran nombre de tasques autònomes, paral·leles o sessions d'usuari interactives. També gestiona i programa l'assignació de recursos distribuïts, com ara processadors, memòria, espai en disc i les llicències de programari.



[57] Sun Grid Engine, la funció principal és la gestió d'un sistema de recursos computacionals o processos distribuïts en ambients heterogenis, de manera que s'utilitzin aquests recursos de la manera més eficient.
[58] GRD és un avançat recurs d'administració per a entorns heterogenis i distribuïts de computació, proporciona una sèrie de politiques per la gestió d'entorns UNIX amb diversos recursos compartits.
[59] UNIVA Corporation és dedica a gestionar mitjançant el seu programari la càrrega de treball per maximitzar el valor dels recursos informàtics existents, i compartir de manera eficient les càrregues de treball a través de milers de servidors.

5.5.1 Funcionalitats i característiques principals

- Múltiples algoritmes de planificació avançades que permeten la potent assignació de recursos basat en polítiques.
- Cues de clúster
- Treball i programador de tolerància a fallades, Grid Engine segueix funcionant com sempre que hi hagi un o més hosts disponibles.
- Checkpointing de treball
- Matrius de treball i les tasques del treball
- DRMAA[60] (API Job)
- Reserva de recursos
- Informes d'estat XML mitjançant interfícies web (qstat i Qhost), i el xml-qstat.
- Treballs paral·lels (MPI[61], PVM, OpenMP[62]), i l'inici de treball paral·lel escalable amb qrsh.
- Comptabilitat d'ús.
- Comptabilitat i Presentació d'Informes de la consola (ARCO).
- make[63] paral·lel: distmake, dmake (Sun Studio), i propi qmake de SGE.
- Integració FLEXlm i multi-clúster de gestió de llicències de programari amb License Juggler.

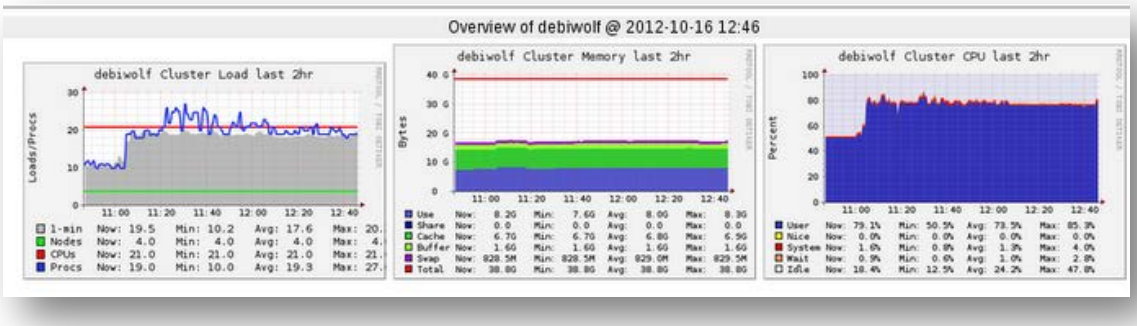
5.6 Ganglia

És tracta d'una eina de monitoratge escalable i distribuïda per a sistemes de computació d'alt rendiment, com els clústers i grids. Permet a l'usuari veure de forma remota estadístiques en temps real o històrics (com ara mitjanes de càrrega de la CPU o d'utilització de la xarxa) per a totes les màquines que s'estan supervisant.

Ganglia està basat en un disseny jeràrquic dirigit a agrupacions de maquinari com poden ser clústers. Es basa en protocol (LISTEN / ANNOUNCE) basat en multicast per monitoritzar l'estat dins dels grups. Aquest empra un arbre punt a punt de connexió entre tots els nodes del clúster per afegir el seu estat.

Aprofita tecnologies àmpliament utilitzades com XML[64] per a la representació de dades, XDR[65] per la compactació, protocol de transport de dades i RRDtool[66] per a l'emmagatzematge de dades i la seva visualització. S'empren estructures de dades i algoritmes acuradament dissenyats per aconseguir baixes despeses per al node monitoritzat.

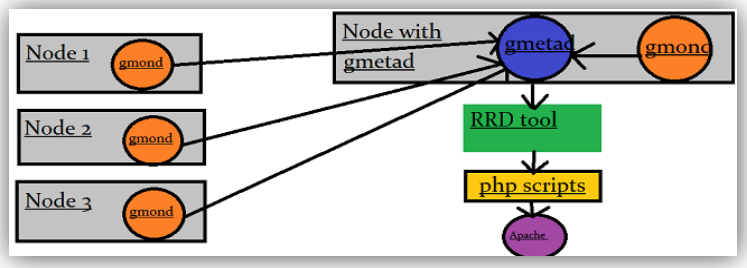
[60] És una especificació API d'alt nivell per a la presentació i control dels treballs.
[61] Estàndard de pas de missatges, defineix la sintaxi i la semàntica d'un nucli de rutines per programes de pas de missatges en Fortran o el llenguatge de programació C.
[62] Interfície de programació d'aplicacions (API) que suporta programació multiprocés amb memòria compartida multiplataforma en C/C++ i Fortran a moltes arquitectures.
[63] Make obté el seu coneixement de com compilar el programa des d'un arxiu anomenat el makefile.
[64] Extensible Markup Language, és tracta d'un metallenguatge i permet la gramàtica de llenguatges específics.
[65] External Data Representation és un estàndar de serialització de dades per a protocols de xarxes.
[66] RRDtool és una eina per la col·lecció i graficació de dades .



5.6.1 Funcionalitats i característiques principals

- Sistema escalable distribuït.
- Orientat per la monitorització de clúster i grids.
- Basat en multicast.
- Empra estàndards oberts.

Aquesta eina de monitorització està distribuïda en diferents parts que a continuació s'enumeren.



5.6.2 GMOND

És tracta d'un petit de servei instal·lat a cada node que ha de ser monitoritzat, aquest recull tota la informació mètrica i l'envia a través d' XML a través de TCP. Gmond té la capacitat de recollir molts nombre de mètriques com CPU, memòria, càrrega i molts més de personalitzades.

5.6.3 GMETAD

Aquest és el dimoni que recull dades d'altres dimonis Gmetad i tots els dimonis Gmond, Gmetad emmagatzema les dades que recull en forma d'un RRD

5.6.4 RRD Tool

Ganglia utilitza l'eina RRD per emmagatzemar les seves dades i fer la visualització d'aquestes. RRD és la abreviació per a l'eina de Base de Dades Round Robin. És tracta d'una útil eina de base de dades de codi obert.

Les dades dades s'emmagatzemen seqüencialment en la línia de temps, per exemple RRD emmagatzema tots els valors de la càrrega de la CPU en un interval de temps determinat i després gràfica i mostra les dades en funció del temps.

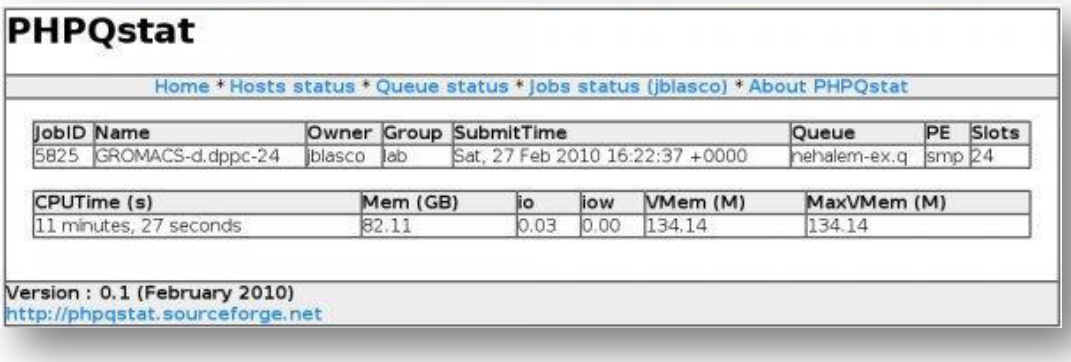
5.6.5 Apache i PHP

Ganglia empraper defecte el servidor web Apache amb PHP per representar els gràfics traçats per l'eina RRD. Apache és compatible amb una varietat de característiques, moltes implementades com a mòduls compilats que estenen la seva funcionalitat. Aquests poden anar des del suport llenguatge de programació i/o el sistema d'autenticació. Algunes interfícies comuns dels llenguatges compatibles amb Perl[67], Python[68], Tcl[69], i PHP.

El llenguatge de programació PHP s'interpreta per el servidor web mitjançant un mòdul de processament de PHP, el que genera la pàgina web resultant. Les ordres de PHP poden ser incloses directament en un document font HTML en lloc de trucar a un arxiu extern per processar les dades. S' inclou la capacitat d'interaccionar en una línia d'ordres de comandes i es pot utilitzar en aplicacions gràfiques independents.

5.7 PHPQstat

PHPQstat és un programari de codi obert que li permet realitzar tasques d'agrupament. És gratuït tant per a ús personal i comercial. És una interfície web que permet connectar els comandaments útils del sistema de cues Sun Grid Engine (SGE). Amb aquesta interfície és pot supervisar l'estat d'un treball i la seva salut de les cues en temps real.



5.7.1 Funcionalitats i característiques principals

- Execució de comandes SGE mitjançant la interfície web.
- Comptabilitat de les cues en temps real.
- Llistat de tasques per usuaris i grups.

[67] Perl és un llenguatges de programació inspirat en els llenguatges AWK, C, SED, SHELL entre d'altres.
[68] Python és un llenguatge de programació d'alt nivell de propòsit general.
[69] S'utilitza per al desenvolupament ràpid de prototips, aplicacions "script", interfícies gràfiques i proves.

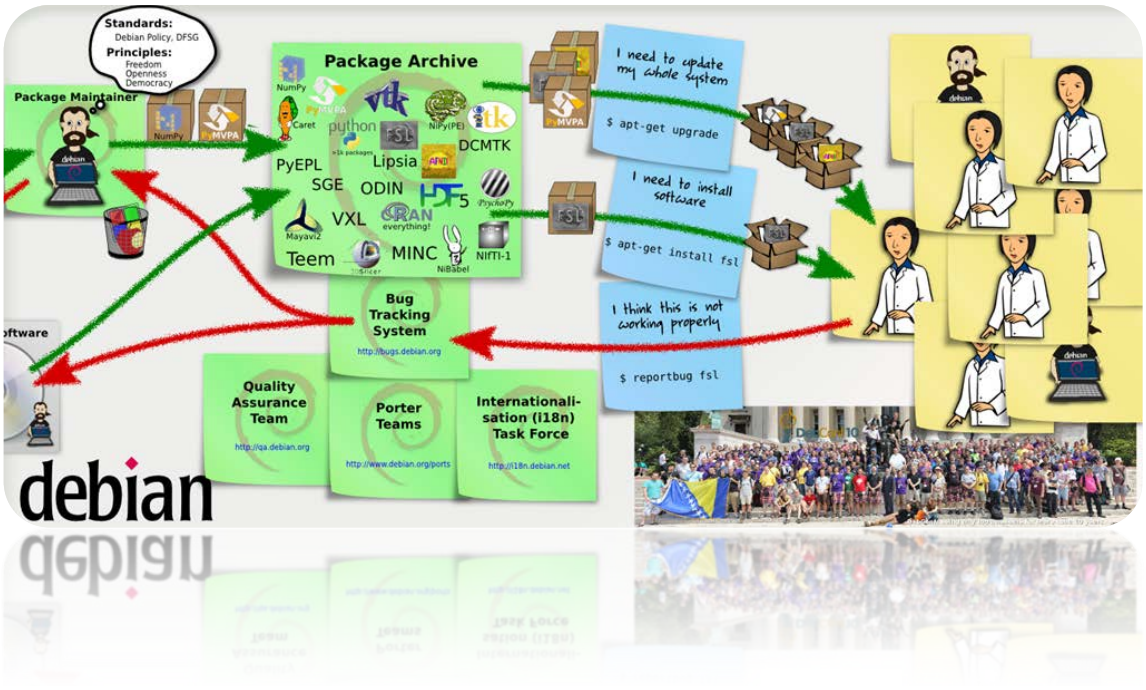
- Descripció dels recursos emprats per tasca (submission time[70], wait time[71], walltime[72], cputime[73], efficiency[74]=(cputime/(walltime*slots)

5.8 Neurodebian repositories

NeuroDebian ofereix una gran col·lecció de programari popular per la investigació de la neurociència, fet a mida per al sistema operatiu Debian i Ubuntu i altres derivats. Paquets com AFNI[75], PyMVPA[76] i molts altres és troben en diferents versions dins d'aquest repositori. És tracta d'un repositori gratuït i mantingut per la comunitat és amb un alt nivell de qualitat.

Amb aquesta finalitat, el projecte ofereix una font de paquets que complementa el principal del Projecte Debian i Ubuntu. NeuroDebian no és altra distribució de Linux, sinó més aviat un esforç dins del projecte Debian en si. Els paquets de programari estan totalment integrats al sistema Debian.

Amb NeuroDebian, instal·lar i actualitzar el paquets de neurociència no és diferent de qualsevol altra part del sistema operatiu. El manteniment d'un entorn de programari de recerca arriba a ser tan fàcil com instal·lar un qualsevol altre programari dels repositoris oficials de la distribució.



[70] Submission Time: És l'interval de temps que triga un procés des de la seva execució i la finalització.
[71] Wait time: És el temps que triga en executar-se una tasca a la Unitat Central de procés.
[72] Wall Time: És el temps que percep l'humà desdel moment que executar la tasca i el seu final.
[73] CPU Time: és el total de temps que està un procés executant-se a la unitat central de procés.
[74] Efficiency: Descriu el grau en que temps, esforç o cost ha fet servir una determinada tasca.
[75] Analysis of Functional NeuroImages, entorn de codi obert per al processament de dades dels humans.
[76] Paquest Python destinat a facilitar l'aprenentatge estadístic de grans conjunts de dades.

6 DISSENY

A continuació és defineixen, analitzen i dissenyen totes les fases del projecte descrites anteriorment. És dur a terme una relació entre la fase i la tecnologia i programaris emprats, també és fa referència als diferents annexos afegits al final d'aquest document.

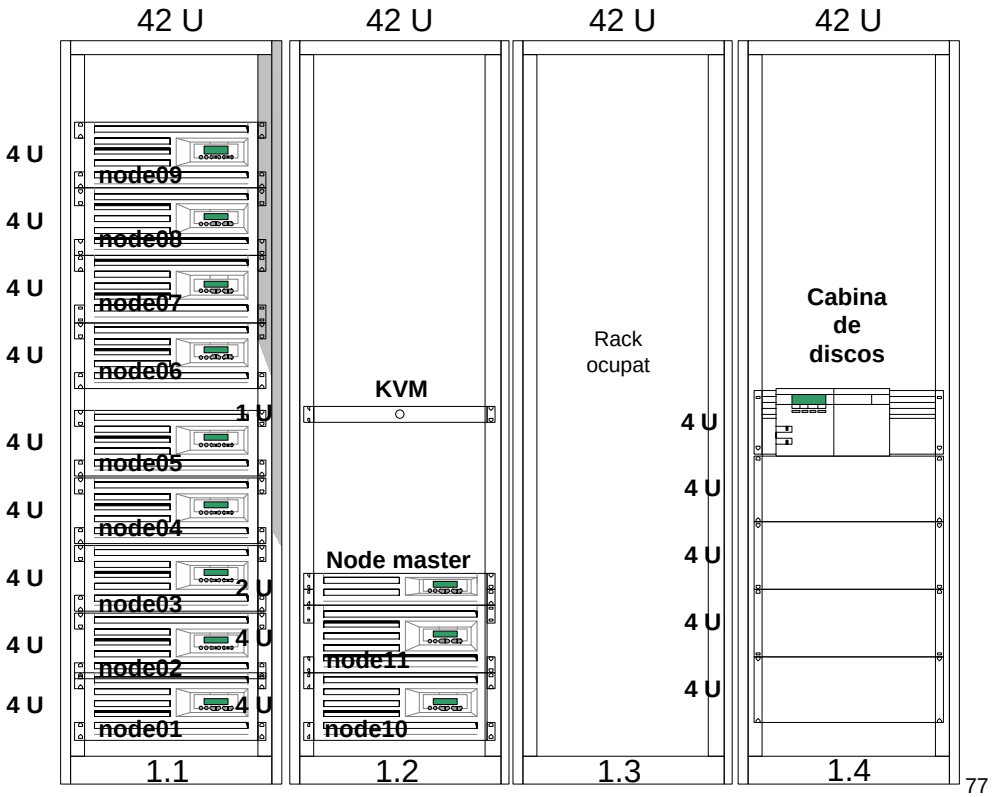
6.1 Disseny del muntatge del maquinari.

Aquest apartat està distribuït en dues parts, la vista frontal i posterior dels bastidors de 42U. Un cop distribuït l'espai dels armaris és durà a terme la tasca muntatge i enrackat dels diversos elements.

- **ANNEX-I**, Vista final del muntatge.

6.1.1 Vista frontal dels armaris.

En aquesta vista és poden identificar la part frontal dels servidors de càlcul més el màster, el switch KVM i les diferents safates de discos NetApp més la controladora que les gestiona.

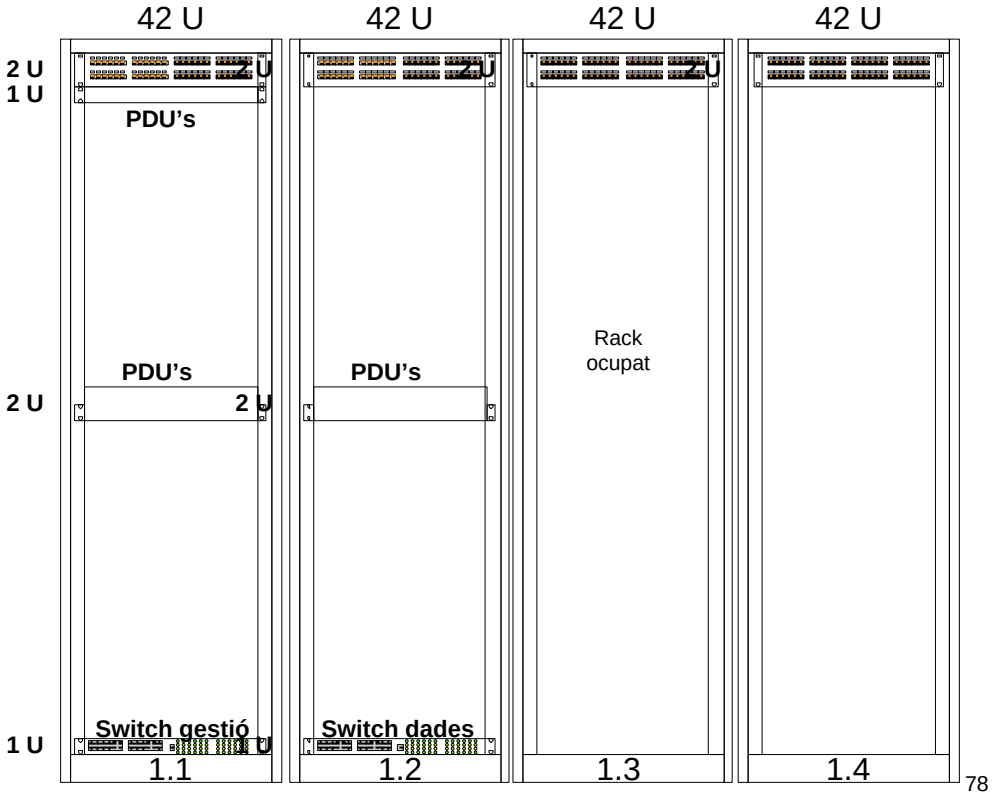


Disseny de muntatge frontal del maquinari als bastidors.

[77] Un switch KVM (Keyboard Video Mouse) és un dispositiu de commutació que permet el control de diferents equips amb només un monitor, un teclat i un ratolí.

6.1.2 Vista posterior dels armaris.

En aquesta vista és poden veure principalment els elements de xarxa, les unitats de distribució d'energia i els patch pannels de xarxa.



[78]

Disseny de muntatge posterior del maquinari als bastidors.

6.1.3 Distribució i consum eléctric del maquinari.

En aquest apartat és tracta la organització de la infrastructura eléctrica i la refrigeració del clúster. Aquests dos aspectes son els més importants per garantir el correcte funcionament del maquinari.

L'alimentació del tot el maquinari ha de assolir la protecció contra la manipulació accidental o intencional d'aquest, ja sigui per una fallada de maquinari o error humà. A més a més cal salvaguardar el funcionament davant de influències externes i proporcionar una refrigeració adient.

La redundància és un altre aspecte fonamental que protegeix contra les fallades tècniques electriques. El centre de processament de dades està connectat a dos sectors diferents de la xarxa de serveis i compta amb bateries que poden cobrir les apagades o alteracions electriques de curta durada. A més, els generadors dièsel proveeixen d'electricitat necessaria

[78] (PDU) o unitat de distribució de xarxa (MDU) és un dispositiu equipat amb múltiples sortides dissenyades per distribuir energia elèctrica, especialment als bastidors d'equips.

en cas d'emergència per garantir el funcionament autónom durant 6 hores. La doble línia de subministrament elèctric (A+B) és ideal per a equips de doble font d'alimentació, cada rack ésta provist de la instal·lació de 16A trifàsica. Cada línia disposa del seu diferencial immunitzat 300 mA, magnetotèrmic i comptador digital independents.

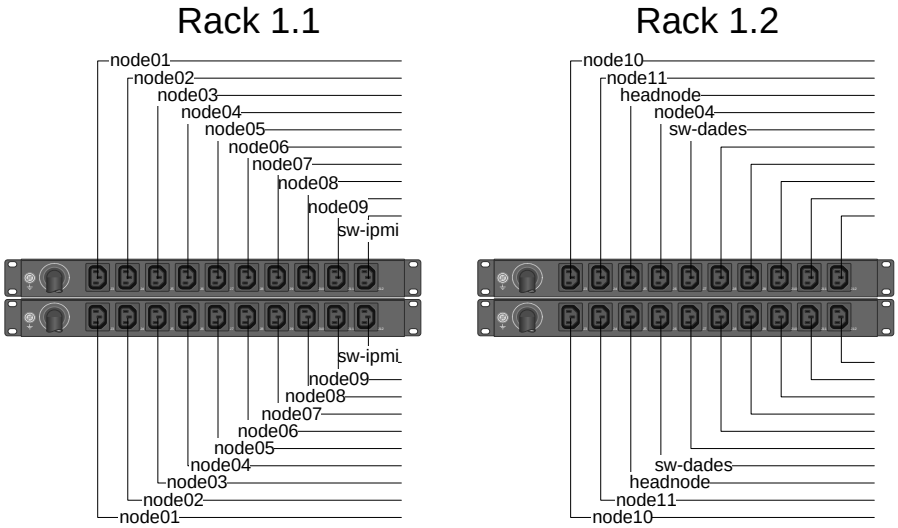
Per altra banda el sistema de refrigeració està construït de tal manera que és garanteix una nivell de temperatura i un grau d'humitat eficients. Per dur a terme aquest tasca el centre de procés de dades té instal·lat equips de climatització específics per a sales informàtiques, capaç de produir fred, calor i humidificar o deshumidificar de forma automàtica dins dels marges de ±1ºC i ±2% d'humitat relativa (HR[79]) per a valors de funcionament previstos de 21ºC i 60% HR.

Taula resum del consum a ple rendiment del maquinari:

Quantitat	**Maquinari**	**Consum màx. (W)**	
		Unitari	**Total**
1	Head màster node	600W	600 W
11	Node de còmput	670W	7.370W
2	Electrónica de xarxa	55W	110W
	TOTAL	**1.325W**	**8.080W**

Taula resum consum total (W) .

Diagrama de connexionat del maquinari:



Disseny del muntatge del cablejat d'alimentació.

[79] La humitat o humiditat és la quantitat de vapor d'aigua present a l'aire, el percentatge de la humitat total que pot contenir l'aire a la temperatura a què ens trobem és defineix com (humitat relativa o grau d'humitat).

6.2 Disseny del servidor màquines virtuals

A continuació és detallen les característiques més bàsiques per tal de planificar el disseny del hypervisor Xen del node principal màster.

- **ANNEX-III**, Instal·lació del sistema operatiu del head màster.
- **ANNEX-IV**, Instal·lació i configuració del servidor del Hypervisor Xen.

A la següent taula és mostra el llistat de programari necessari per instal·lar el node primari.

Programari	Versió	Descripció
Debian	7.3(wheezy) x86_64	Sistema operatiu
xen-system	4.1.4-3+deb7u1	Xen system with Linux for 64-bit PCs
xen-hypervisor	4.1.4-3+deb7u1	Xen Hypervisor on AMD64
xen-tools	4.3.1-1	Tools to manage Xen virtual servers

Distribució de l'espai en disc local:

Dispositiu	Ruta de muntatge	Format	Espai
/dev/sda1	/boot	ext2	512MB
/dev/mapper/vg-root	/	ext4	12GB
/dev/mapper/vg-tmp	/tmp	ext4	1GB
/dev/mapper/vg-var	/var	ext4	4GB
/swap	---	swap	4GB

La distribució del filesystem és realitza de tal manera que en cas que s'ompli /boot, /tmp o /var no afecti al comportament del servidor.

Per altra banda és destina l'espai lliure del disc (en aquest cas el disponible al 'Physical Volume'[80]) per tal de crear els volums de dades lògics LV[81] al Volume Group[82] per la creació de les màquines virtuals, cada volum nou s'assignarà al determinat guest on s'instal·larà el seu sistema operatiu.

<table>
<tr><td>/boot</td><td>swap</td><td>/tmp</td><td>/var</td><td>--</td><td>--</td><td>--</td><td>--</td><td>--</td><td>Path mount</td></tr>
<tr><td rowspan="3">sda1</td><td rowspan="3">swap</td><td>tmp</td><td>var</td><td>proxy</td><td>sge</td><td>login</td><td>monitor</td><td>deploy</td><td>LV</td></tr>
<tr><td colspan="7">Volume Group (vg)</td><td>VG</td></tr>
<tr><td colspan="7">Physical Volume</td><td>PV</td></tr>
<tr><td colspan="9">Hard Disk (900GB)</td><td>Device
(RAID5)</td></tr>
</table>

[80] Un volum físic és típicament un disc dur, encara que bé podria ser només un dispositiu que consta d'un conjunt de disc durs.
[81] Els grups de volums es poden dividir en volums lògics, on s'assignen els punts de muntatge.
[82] Un grup de volums és un conjunt de volums físics (locals o remots) a partir del qual un volum lògic (essencialment una partició) pot ser creat.

6.3 Disseny de la infraestructura de comunicacions.

En aquest punt és mostren les diferents interconnexions de xarxa (físiques i virtuals) que necessita el clúster per al seu funcionament.

- **ANNEX-V**, Configuració de les xarxes físques i virtuals.
- **ANNEX-II**, Configuració targetes IPMI.

Segons l'indicat a la proposta del projecte s'empraran dos switchos (gestió i dades), per altra banda el node màster que gestionarà les màquines virtuals Xen tindrà al seu càrrec una sèrie d'interfícies de xarxes amb links virtuals per tal d'interconnectar els 'Domain guests'.

Així doncs, les xarxes requerides son les següents:

- Xarxa de servei corporativa (PRIV).
- Xarxa de dades per la gestió interna (DATA).
- Xarxa per la gestió remota (IPMI).
- Xarxa de dades per la comunicació amb la safata de discos (NFS).
- Xarxa desmilitaritzada (DMZ).
- Xarxa d'intercanvi de missatges (MPI).

Xarxa física:

- El Switch de 24 ports GS724TS dedicat a la gestió IPMI, tindrà una xarxa plana i s'emprarà la VLAN per defecte per a tots els ports.

- El Switch de 48 ports GS752TXS tindrà definides 3 VLANS (PRIV, MPI i NFS) amb una connexió de 10G amb la cabina de discos.

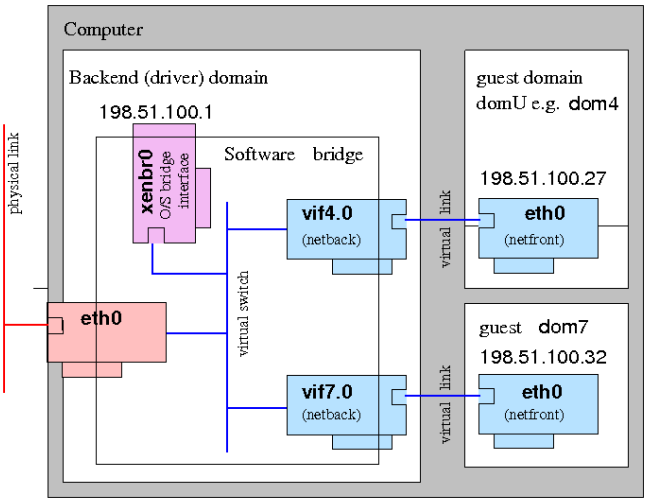
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	M	M	M	M	M	M	M	M	M
25	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
M	M	M	M	M	M	M	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	C	C

D: Xarxa DATA.
M: Xarxa MPI.
N: Xarxa NFS.
C: Connexió 10G amb la cabina de discos.

Xarxa Virtual:

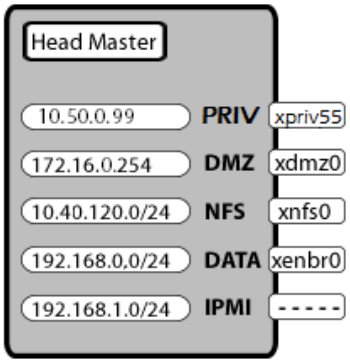
Mitjançant el mode bridged[83] de Xen és defineixen els diferents elements de xarxa virtuals a partir d'una interfície virtual, permet complet accés a tots els domU per la xarxa d'àrea local.

[83] Mode bridged descriu l'acció realitzada pels equips de xarxa per permetre dues o més xarxes de comunicació o diferents segments de xarxa.



Exemple de configuració de Xen en mode bridged.

La següent imatge mostra les interfícies de xarxa del node màster que hostatjarà totes les màquines virtuals.



Definició de les xarxes al node head màster.

Definició dels serveis de les xarxes:

Servei	Descripció
Privada Corporativa	Xarxa privada corporativa accessible per als usuaris i administradors
DATA	Xarxa per la comuniació interna entre el Head Màster, màquines virtuals i els nodes.
IPMI	Xarxa de comunicació per la gestió mitjançant IPMI.
NFS	Xarxa de dades dedicada a NFS per la comunicació amb la cabina de discos.
DMZ	Xarxa DMZ per la comunicació exclusiva de totes les màquines virtuals que resideixen al Head Màster.
MPI	Xarxa de comunicació del protocol de intercanvi de missatges *'Message Passing Interface'*

Definició de les xarxes per servei:

Servei	Xarxa
Privada Corporativa	10.55.0.0/24
IPMI	192.168.1.0/24
DATA	192.168.0.0/24
MPI	192.168.4.0/24
NFS	10.40.120.0/24
DMZ	172.16.0.0/24

Definició de les interfícies de xarxa al node màster :

Dispositiu de xarxa	Nom del bridge	Servei
eth0	xpriv55	Privada Corporativa
eth1	xenbr7	DATA
eth2	---	IPMI
eth3	xnfs0	NFS
dummy0	xdmz0	DMZ

Definició de les interfícies de xarxa als nodes de càlcul :

Dispositiu de xarxa	Servei
BMC	IPMI
eth0	DATA
eth1	MPI
eth2	NFS

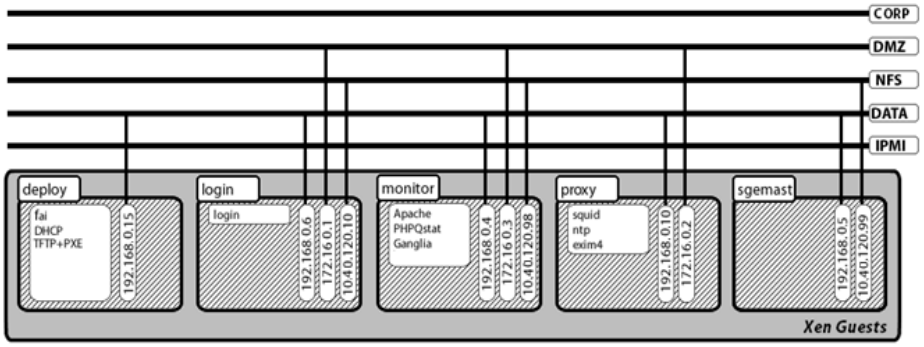
Definició de les assignacions d'adreçament IP:

Nom del servidor	Dispositiu de xarxa	IP Address	Servei
head-master	xpriv55	10.55.0.99	Privada corporativa
	xdmz0	172.16.0.254	DMZ
	xenbr0	192.168.0.251	DATA
	bmc	192.168.1.1	IPMI
deploy	eth0	192.168.0.15	DATA
login	eth0	192.168.0.6	DATA
	eth1	10.55.0.100	Privada corporativa
	eth2	10.40.120.100	NFS
monitor	eth0	192.168.0.4	DATA
	eth1	172.16.0.3	DMZ
	eth2	10.40.120.102	NFS
proxy	eth0	192.168.0.10	DATA
	eth1	172.16.0.2	DMZ
sgemaster	eth0	192.168.0.5	DATA

	eth1	10.40.120.101	NFS
nodeXX	bmc	192.168.1.1XX	IPMI
	eth0	192.168.0.1XX	DATA
	eth1	192.168.4.1XX	MPI
	eth2	10.40.120.1XX	NFS
switchint	1	192.168.0.245	DATA
template	eth0	192.168.0.199	DATA
	eth1	172.16.0.199	DMZ
	eth2	10.40.120.199	NFS

A la següent imatge és defineixen les màquines virtuals i les connexions a les diferents interfícies virtuals que disposa els domain0.

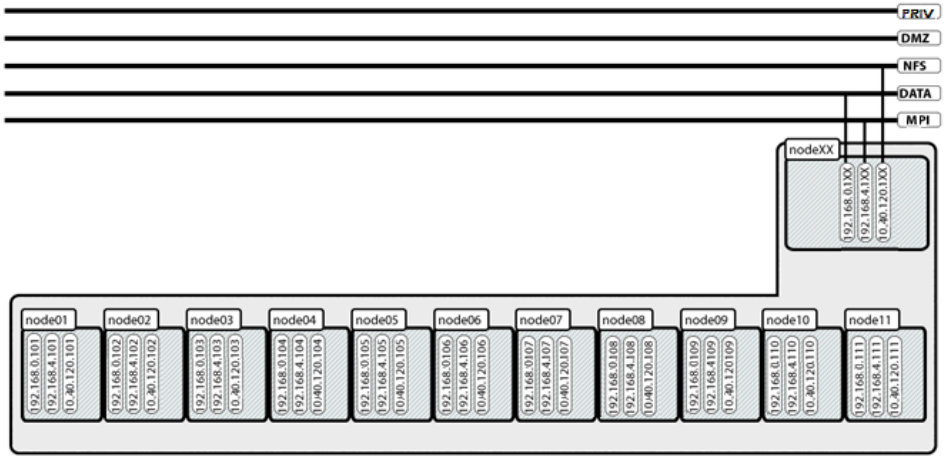
S'inclou dins de cada màquina virtual una petita descripció dels serveis primaris de cadascuna més l'adreçament IP.



Definició de les interconnexions de les màquines virtuals.

Finalment és descriu com és connectaran els 11 nodes de còmput a les diferents xarxes, per tal de simplificar a la següent imatge no és mostren els elements de l'electrònica de xarxa.

A la següent imatge és mostra el *nodeXX* que indica genericament quines interfícies seran emprades i a quines xarxes és connectaran els nodes físics. A sota és mostren totes les màquines físiques amb les seves respectives IP's.



Definició de les interconnexions de les màquines físiques.

6.4 Disseny del sistema base Guest.

És dur a terme el procés de creació de la màquina virtual base, a partir d'aquesta és clonaran la resta de màquines virtuals productives. Totes s'executaran en el head màster node sota el servidor Xen.

- **ANNEX-VI**, Instal·lació màquina virtual base de template.

A continuació és defineix la fitxa de requeriments de recursos de maquinari que necessita el nou guest, sota d'aquesta taula és detallen alguns d'aquests recursos:

Dispositiu	Quantitat
CPU's virtuals	1
Memòria RAM	1024MB
Discs virtual	1
Targetes de xarxa	3

Distribució de l'espai en disc local:

Dispositiu	Ruta de muntatge	Format	Espai
/dev/xda1	/	ext3	5GB

Connexions de xarxa:

Dispositiu de xarxa	Servei
eth0	DATA
eth1	DMZ
eth2	NFS

A la següent taula és mostra el llistat de programari específic requerit per instal·lar la màquina virtual base.

Programari	Versió	Descripció
Debian	7.5(wheezy) x86_64	Sistema operatiu base
libldap	2.4.31-1	Llibreries openldap
libpam-lap	184-8.6	Mòdul d'autenticació LDAP
libnss-ldap	264-2.5	Servei de noms LDAP
nfs-common	1.2.6-4	Client NFS
openssh	6.0p1-4	Servidor i client SSH

6.5 Disseny del sistema Login.

És dur a terme el procés de creació de la màquina virtual servidor de login, amb accés directe per la xarxa corporativa d'usuaris. En aquest servidor és connectaran els usuaris per executar les seves tasques al clúster de computació.

- **ANNEX-VII**, Clonar màquina virtual mitjançant el template.
- **ANNEX-VIII**, Instal·lació màquina virtual Login.

A continuació és defineix la fitxa de requeriments de recursos de maquinari que necessita el nou guest, sota d'aquesta taula és detallen alguns d'aquests recursos:

Dispositiu	Quantitat
CPU's virtuals	4
Memòria RAM	8096MB
Discs virtual	1
Targetes de xarxa	3

Distribució de l'espai en disc local:

Dispositiu	Ruta de muntatge	Format	Espai
/dev/xda1	/	ext3	4GB

Distribució de l'espai en disc remot:

Font	Ruta de muntatge	Protocol	Espai
10.40.120.21:/hpc_homes	/homes	NFS	6TB
10.40.120.21:/hpc_soft	/soft	NFS	
10.40.120.21:/hpc_sge	/sge	NFS	

Connexions de xarxa:

Dispositiu de xarxa	Servei
eth0	DATA
eth1	Xarxa corporativa
eth2	NFS

A part del programari instal·lat al sistema base és mostra el llistat de programari específic requerit per instal·lar la màquina virtual base.

Programari	Versió	Descripció
SGE user tool set		Set d'eines de Sun Grid Engine per executar, esborra, verificar i obtenir informació de les cues i l'entorn d'usuari normal
PHPQstat	0.2.0a	Eina web per visualitzar les tasques d'usuari al gestor de cues.

6.6 Disseny del sistema Proxy.

És dur a terme el procés de creació de la màquina virtual servidor de proxy, amb accés directe per la xarxa gestió i permet que els nodes de computació tinguin sortida a internet, per enviament de correus, accés a repositoris o descàrrega de fitxers.

- **ANNEX-VII**, Clonar màquina virtual mitjançant el template.
- **ANNEX-IX**, Instal·lació màquina virtual Proxy.

A continuació és defineix la fitxa de requeriments de recursos de maquinari que necessita el nou guest, sota d'aquesta taula és detallen alguns d'aquests recursos:

Dispositiu	Quantitat
CPU's virtuals	1
Memòria RAM	512MB
Discs virtual	1
Targetes de xarxa	2

Distribució de l'espai en disc local:

Dispositiu	Ruta de muntatge	Format	Espai
/dev/xda1	/	ext3	5GB

Connexions de xarxa:

Dispositiu de xarxa	Servei
eth0	DATA
eth1	DMZ

A part del programari instal·lat al sistema base és mostra el llistat de programari específic requerit per instal·lar la màquina virtual base.

Programari	Versió	Descripció
squid	2.7.STABLE9-4.1	Internet object cache (WWW proxy)
squid-common	2.7.STABLE9-4.1	

6.7 Disseny del sistema Deploy.

El procés de creació de la màquina virtual deploy durà a terme les instal·lacions automàtiques dels nodes de còmput, aquest només tindrà accés a la xarxa de gestió per fer les seves tasques com la configuració automàtica d'ips dels nodes, cache del repositori principal *'main'* de debian i desplegament d'imatges mitjançant el servidor TFTP[84]

- **ANNEX-VII**, Clonar màquina virtual mitjançant el template.
- **ANNEX-X**, Instal·lació màquina virtual Deploy.

A continuació és defineix la fitxa de requeriments de recursos de maquinari que necessita el nou guest, sota d'aquesta taula és detallen alguns d'aquests recursos:

Dispositiu	Quantitat
CPU's virtuals	1
Memòria RAM	512MB
Discs virtual	1
Targetes de xarxa	1

Distribució de l'espai en disc local:

Dispositiu	Ruta de muntatge	Format	Espai
/dev/xda1	/	ext3	5GB

Connexions de xarxa:

Dispositiu de xarxa	Servei
eth0	DATA

A part del programari instal·lat al sistema base és mostra el llistat de programari específic requerit per instal·lar la màquina virtual base.

Programari	Versió	Descripció
fai-server	4.0.8	Servidor d'instal·lacions automatitzades FAI
fai-doc	4.0.8	Documentació FAI
fai-quickstart	4.0.8	Paquet quick start de configuració per al servidor
fai-client	4.0.8	Paquet client FAI
ipmitool	1.8.11-5	Eines de gestió remota per la interficie de xarxa
isc-dhcp-server	4.2.2.dfsg.1-5	Servidor de configuració automàtica d'ip's
apt-cacher-ng	0.7.11-1	Servidor cau de repositoris de programari

[84] TFTP és un protocol de transferència molt simple similar a una versió bàsica d' FTP, sovint s'utilitza per a transferir petits arxius entre ordinadors en una xarxa, utilitza el port UDP/69

6.8 Disseny del sistema GE Màster.

És dur a terme el procés de creació de la màquina virtual sgemaster, de la mateixa manera que el servidor de desplegament de paquests només tindrà accés a la xarxa de gestió, la qual comunica els nodes de còmput amb les màquines virtuals. La tasca principal d'aquest servidor serà gestionar les tasques que l'usuari envia i monitoritzar la seva execució als node de computació.

- **ANNEX-VII**, Clonar màquina virtual mitjançant el template.
- **ANNEX-XI**, Instal·lació màquina virtual GE Màster.

A continuació és defineix la fitxa de requeriments de recursos de maquinari que necessita el nou guest, sota d'aquesta taula és detallen alguns d'aquests recursos:

Dispositiu	Quantitat
CPU's virtuals	1
Memòria RAM	4096MB
Discs virtual	1
Targetes de xarxa	2

Distribució de l'espai en disc local:

Dispositiu	Ruta de muntatge	Format	Espai
/dev/xda1	/	ext3	5GB

Distribució de l'espai en disc remot:

Font	Ruta de muntatge	Protocol	Espai
10.40.120.21:/hpc_homes	/homes	NFS	6TB
10.40.120.21:/hpc_sge	/sge	NFS	

Connexions de xarxa:

Dispositiu de xarxa	Servei
eth0	DATA
eth1	NFS

A part del programari instal·lat al sistema base és mostra el llistat de programari específic requerit per instal·lar la màquina virtual base.

Programari	Versió	Descripció
qmaster	ge2011.11	Sistema servidor base
execd	ge2011.11	Sistema d'execució

** S'instal·la la versió Open Grid Scheduler basada en Sun Grid Engine, i mantinguda pel mateix grup de desenvolupadors que van començar a contribuir en el codi des de l'any 2001.*

6.9 Disseny del sistema de Monitor.

És dur a terme el procés de creació de la màquina virtual servidor de monitorització, amb accés directe per la xarxa de gestió i a la xarxa de dades NFS. Aquest hostatjarà totes les eines de monitorització per tal de conèixer quin és l'estat de salut dels clúster o consultar quines tasques executa un determinat usuari mitjançant servidor web.

- **ANNEX-VII**, Clonar màquina virtual mitjançant el template.
- **ANNEX-XII**, Instal·lació màquina virtual Monitor.

A continuació és defineix la fitxa de requeriments de recursos de maquinari que necessita el nou guest, sota d'aquesta taula és detallen alguns d'aquests recursos:

Dispositiu	Quantitat
CPU's virtuals	1
Memòria RAM	1024MB
Discs virtual	1
Targetes de xarxa	3

Distribució de l'espai en disc local:

Dispositiu	Ruta de muntatge	Format	Espai
/dev/xda1	/	ext3	5GB

Connexions de xarxa:

Dispositiu de xarxa	Servei
eth0	DATA
eth1	DMZ
eth2	NFS

A part del programari instal·lat al sistema base és mostra el llistat de programari específic requerit per instal·lar la màquina virtual base.

Programari	Versió	Descripció
apache2	2.2.22-13	Servidor web.
ganglia	3.6	Servidor monitorització ganglia.
PHPQstat	0.2.0	Interfície d'execució de comandes de SunGrid Engine.

6.10 Disseny del sistema del node físic de càlcul.

La instal·lació dels nodes de càlcul les farà el servidor d'instal·lacions FAI, totes les configuracions s'afegiran als perfils definits, com el particionament del disc, instal·lació del programari, configuracions, interfícies de xarxa, etc,...

- **ANNEX-XIII,** Configuració FAI del perfil d'instal·lació.

El maquinari de cada node consta de:

Dispositiu	Descripció	Quantitat
Socket	Opteron Abu Dhabi 6378 a 2,4Ghz	4
Cores		16
Memòria RAM	DDR3 ECC; 1600Mhz	256GB
Discos	SSD R/W:555/510MB/s; 6Gb/s; 240GB	1
Targetes de xarxa	Intel Corporation I350 Gigabit Network	4

Distribució de l'espai en disc local:

Dispositiu	Ruta de muntatge	Format	Espai
/dev/sda1	/	ext3	20GB
/dev/sda5	/var	ext3	4GB
/dev/sda6	/tmp	ext3	1GB
/dev/sda8	/scratch	ext4	195GB

Distribució de l'espai en disc remot:

Font	Ruta de muntatge	Protocol	Espai
10.40.120.21:/hpc_homes	/homes	NFS	6TB
10.40.120.21:/hpc_soft	/soft	NFS	
10.40.120.21:/hpc_sge	/sge	NFS	

Connexions de xarxa:

Dispositiu de xarxa	Servei
eth0	DATA
eth1	MPI
eth2	NFS

Paquets bàsics de programari que instal·larà el mòdul Software Update de FAI:

Programari	Versió	Descripció
g++	4.7.2-1	Compilador GNU C++, forma part de GCC
gcc	4.7.2-1	Compilador GNU Compiler, inclou C, C++, Fortran
isc-dhcp-client	4.2.2	Client de configuracions automàtica d'IP's
openssh-server	6.0p1	Protocol per al Servidor Secure Shell

nfs-common	1.2.6-4	Suport NFS per als clients.
gfortran	4.7.2-1	Compilador que forma part de GCC
python3	3.2	Llenguatge de programació de propòsit general.
initramfs-tools	0.109.1	Inici del kernel a la memòria RAM durant l'arrencada.
exim4	4.0	Enviament de correu
locales-all	2.13	Determina la llengua local del sistema operatiu
ntp	4.2.6	Client de servidor de temps

Dependències de Matlab:

libatk1.0-data
libcairo-gobject2
libck-connector0
libcolord1
libdbus-glib-1-2
libdconf0
libexif12
libfile-copy-recursive-perl
libgd2-xpm
libgphoto2-2
libgphoto2-l10n
libgphoto2-port0
libgtk-3-0
libgtk-3-0-dbg
libgtk-3-bin
libgtk-3-common
libgudev-1.0-0
libgusb2
libieee1284-3
libpam-ck-connector
libpolkit-agent-1-0
libpolkit-backend-1-0
libpolkit-gobject-1-0
libsane
libsane-common
libsane-extras
libsane-extras-common
libv4l-0
libv4lconvert0
libxcomposite1
libxdamage1
libxinerama1
libxrandr2
policykit-1
sane-utils
update-inetd
libgtk-3.0:amd64
libgtksourceview-3.0-common
libgtksourceview-3.0-0:amd64
libasound2
libgtk2.0-bin
libgtk2.0-common
libgtk2.0-0:amd64
python-numpy
python-scipy
python-xlrd
python-xlwt
python-txosc
python-sklearn
python-nibabel

Llibreries de python:

python-central
python-dateutil
python-dicom
python-fuse
python-gamin
python-glade2
python-gobject
python-gobject-2
python-gtk2
python-imaging
python-matplotlib
python-matplotlib-data
python-newt
python-nose
python-openssl
python-pam
python-pip
python-pyparsing
python-serial
python-setuptools
python-simplejson
python-tk
python-tz
python-virtualenv
python
python-apt
python-apt-common
python-chardet
python-debian
python-debianbts
python-fpconst
python-minimal
python-reportbug
python-soappy
python-support
python2.6
python2.6-minimal
python2.7
python2.7-minimal

El llistat de paquets afegits al servidor FAI anirà augmentant a mida que els desenvolupadors demanin la instal·lació de les llibreries i programari.

A banda d'aquest programari els usuaris del clúster disposen d'un directori compartit per NFS amb programari a mida compilat a aquest. Aquest disposa de compiladors, simuladors, llibreries matemàtiques entre d'altres.

Programari	Versió	Descripció
jdk	1.7.0_40	Java developer kit.
openmpi	1.4.3 1.6	Interfície de pas de missatges.
x86_open64	4.5.2.1 5.0	Eina de generació de codi de qualitat de producció d'alt rendiment dissenyat per a càrregues de treball d'alt rendiment de computació paral·lela.
python	2.7.5 3.3.2	Llenguatge de programació d'alt nivell de propòsit general, combina una potència remarcable amb una sintaxi clara i entenedora.
MATLAB	R2013b	Entorn de computació numèrica i un llenguatge de programació, permet manipular fàcilment matrius, dibuixar funcions i dades, implementar algorismes, crear interfícies d'usuari, i comunicar-se amb altres programes en altres llenguatges
lmod	5.1.5	Gestiona fàcilment els ModulePath mitjançant una estructura jeràrquica. Els mòduls d'entorn proporcionen una manera convenient per canviar dinàmicament l'entorn de l'usuari a través modulefiles.
openblas	0.2.8-0	Versió optimitzada de les llibreries BLAS
psblas	2.4.0 3.1.2	Parallel Sparse Basic Linear Algebra Subroutines (PSBLAS) proporciona un marc per permetre implementacions senzilles, eficients i portàtils de sol·lucionadors iteratius per a sistemes lineals, alhora protegeix a l'usuari de la majoria dels detalls de la seva paral·lelització.
acml	5.3.1	AMD Core Math Library (ACML) és una biblioteca de desenvolupament de programari llançat per AMD. Aquesta biblioteca proporciona rutines matemàtiques útils optimitzades per als processadors d'AMD.
atlas	3.11.11	ATLAS proporciona eines que permeten a l'usuari localitzar, codi, i anotar els resultats en matèria de dades primàries, per mesurar i avaluar la seva importància. A més és pot visualitzar les complexes relacions entre elles.

7 PROVES.

En el marc del projecte és duran a terme una sèrie de proves per fer un anàlisi del potencial de clúster instal·lat i així validar i confirmar les seves qualitats. Les proves son bàsicament un conjunt d'activitats que és duran a terme mitjançant dues eines que faran un ús intensiu dels recursos mitjançant llibreries matemàtiques.

A totes les proves realitzades és compara el rendiment del sistema de còmput amb 4 processadors físics de 16 cores cadascú, un total de 64 cores.

Seguidament és fa una breu descripció de les especificacions del processador

- Generació: AMD Opteron 6300 Series.
- Model: 6378
- Nom: Abu Dhabi
- Cores: 16
- Freqüència: 3.3 GHz
- Memòria CAU L1 = 512KB (code) / 256KB (data)
- Memòria CAU L2 = 16MB
- Memòria CAU L3 = 16MB
- Consum elèctric: 115W

7.1 High Performance Linpack

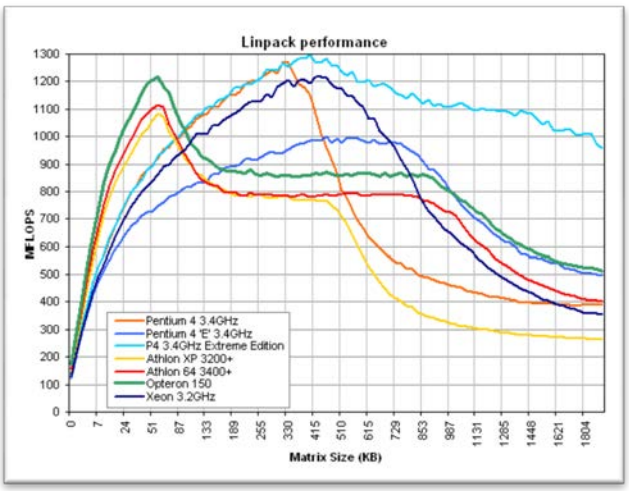
LINPACK és una llibreria software per a realitzar àlgebra lineal de manera numèrica en ordinadors digitals. Fou desenvolupat en el Argone National Laboratory per Jack Dongarra l'any 1976 i és un dels més usats en sistemes científics i d'enginyeria.

També pot esser usat com a benchmark[85], és a dir, com a mesura de potència de càlcul de dos ordinadors. El seu ús com benchmark va ser accidental, ja que originalment va ser una extensió del programa Linpack, el propòsit era resoldre sistemes d'equacions que atorgava el temps d'execució del programa en 23 màquines diferents. Després van ser agregant cada vegada més gran quantitat de màquines, segons els seus autors més com un passatemps que una altra cosa.

La característica principal de Linpack és que fa un ús molt intensiu de les operacions de coma flotant, pel que els seus resultats són molt dependents de la capacitat de la FPU que tingui el sistema. A més passen la major part del temps executant unes rutines anomenades BLAS[86].

[85] Tècnica utilitzada per mesurar el rendiment d'un sistema o component del mateix.
[86] Basic Linear Algebra Subroutines o Subrutines d'Àlgebra Lineal Bàsica.

Els resultats s'expressen en MFLOPS[87], sempre referides a operacions de suma i multiplicació de doble precisió (64 bits, encara que per alguns sistemes aquesta quantitat de bits és la precisió simple).



7.1.1 Funcionalitats i característiques principals

- Resol problemes de matrius generalment denses a tres nivells.
- Unitat de mesura en escala de GFlops[88]
- Disposa de tres benchmarks, d'ordre a 100 (matrius 100x100), d'ordre a 1000 (matrius 1000x1000) i computació altament paral·lela per analitzar les característiques d'un multiprocessador.
- Precisió de variables simple de 4bytes o doble de 8bytes.

7.1.2 Execució al clúster

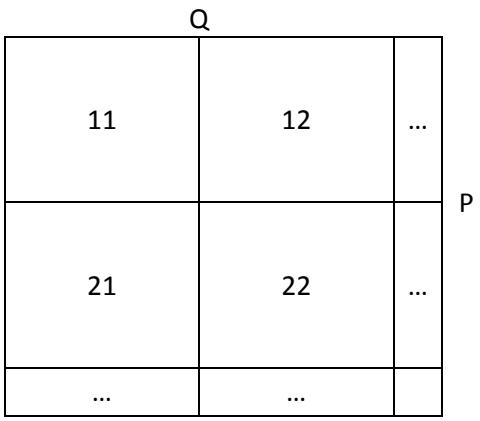
És mesura el rendiment que es pot obtenir utilitzant operacions amb un vector en coma flotant de 64 bits, el rendiment és totalment dependent de la freqüència del nucli; aquest augmenta generalment en petits problemes, i creix més lentament en problemes grans, de tal manera que la mida de la memòria és fa un ressò important que incrementa la mida del problema a resoldre.
Per altra banda, la mida de la memòria cau, la velocitat de la memòria i el rendiment de la xarxa afecten més lleugerament sobre aquest índex de referència.

Així mateix les eines Linpack (HPL) s'han enllaçat i compilat amb la versió 5.0 d'Open64, és tracta d'un compilador open source de codi obert, aquest està optimitzat per arquitectures de microprocessador x86-64.
Linpack utilitza un algorisme per resoldre un sistema d'equacions, en un sistema paral·lel mitjançant processos gestionats pel sistema de pas de missatges MPI, per tal de donar suport a la major varietat d'arquitectures.

[87] Milions d'operacions de coma flotant per segon
[88] Unitat de mesura informàtica que s'utilitza com a mesura de rendiment d'un ordinador. 1Gigaflop = 10^9

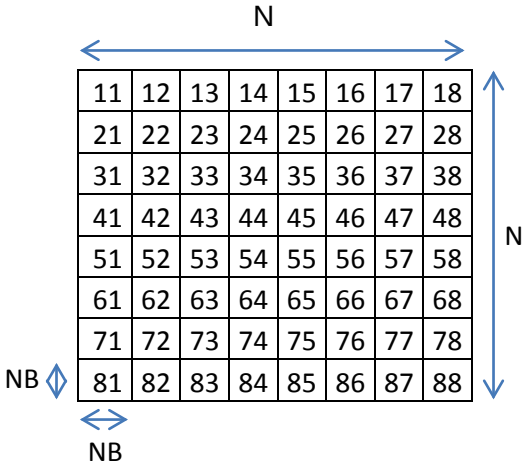
Aquest algoritme està organitzat sobre una graella, on les dades queden distribuïdes sobre ella; per altra banda els elements d'aquesta graella és representen com a processos MPI que alhora queden assignats als processadors d'1 a 1; resumint, cada processador obtindrà un sol procés MPI.

La graella de processos és pot definir a la configuració del Linpack mitjançant les variables P i Q, així doncs en un sistema amb 64 processadors és poden organitzar en graelles de 1*64; 2*32; 4*16; 8*8; 16*4; 32*2 i 64*1.



Graella de processos

Un altre indicador que juga un paper molt important és la mida del problema, aquesta bé definida per N, així mateix aquesta graella és divideix en diferents porcions o blocs determinades per la variable NB que s'utilitza per cada una de las dos dimensions de la graella.



Graella de dades

La configuració dels experiments s'ha emprat de la següent manera:

N = 9
Ns = 10000 12500 15000 17500 20000 22500 25000 27500 30000
NB = 4
NBs = 128 256 512 1024
(P x Q) = 1
Ps = 4
Qs = 4

7.1.3 Resultats obtinguts

Els experiments s'han dut a terme en dos grups de proves per diferents mides del problema, per tal d'analitzar el rendiment en diferents contexts (només els millors resultats és recullen en aquest informe).

Primer grup:

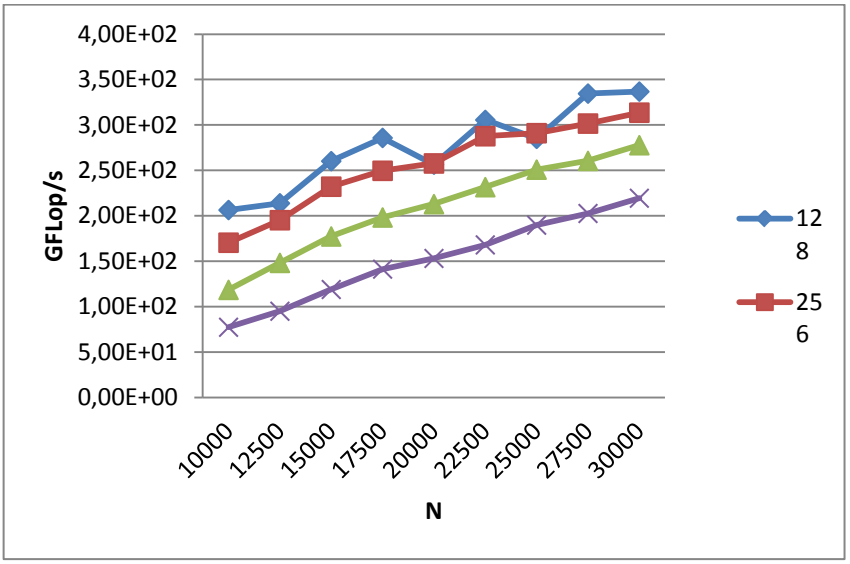
1.1. Execució a un node amb 64 nuclis.
1.2. Execució a dos nodes amb 64 nuclis (32 i 32).
1.3. Execució a dos nodes amb 128 nuclis (64 i 64).

Segon grup:

2.1. Execució a un node amb 16 nuclis a un mateix processador.
2.2. Execució a un node amb 16 nuclis amb diferents processadors.

Les gràfiques mostrades és representen en 4 mides de bloc (NB) diferents (128, 256,512 i 1024), és important veure si existeix algun tipus d'alteració al rendiment depenent de la mida del problema. Als eixos (X,Y) és troben el nivell de rendiment en format de GFLOPS i la mida total del problema, respectivament.

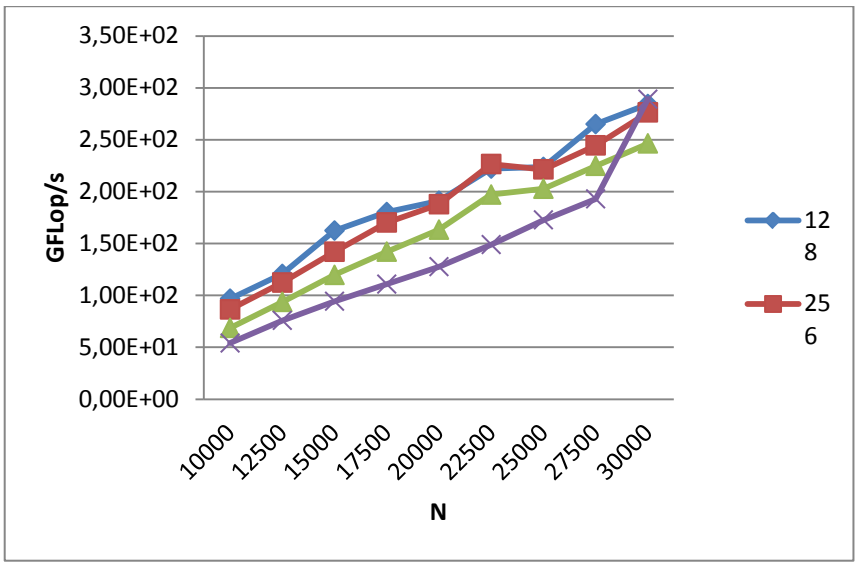
A la primera prova (1.1) s'avalua el rendiment d'un node amb 4 processadors on cada un té 16 nuclis, així doncs un total de 64 nuclis és veuen implicats.



1.1. Gràfica execució HPL a 64 nuclis d'un mateix node.

Els valors d'NB entre 100 i 256 semblen que treballen de manera molt similar, amb valors per sobre de 256; 512 o 1024 com és el cas, el rendiment cau proporcionalment.

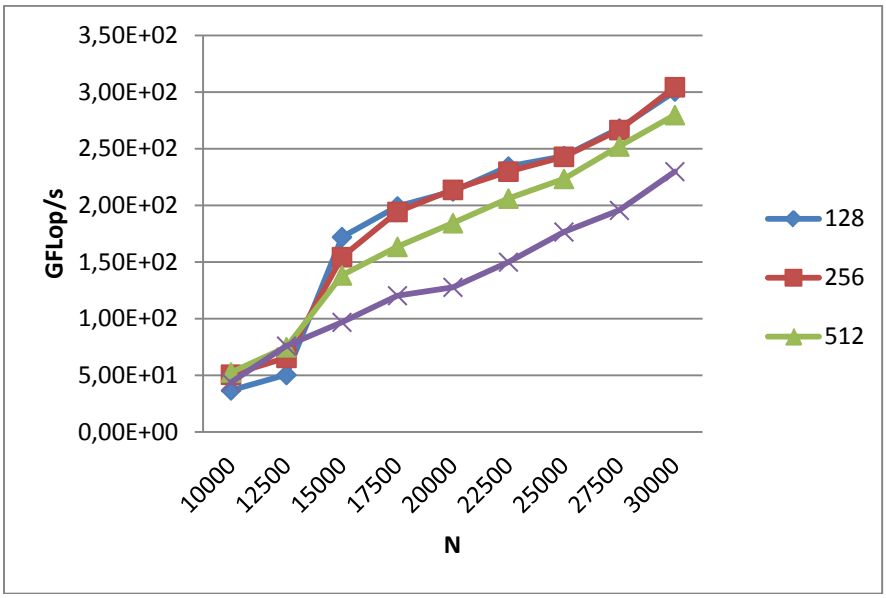
A la segona prova (1.2) s'avalua el rendiment d'execució a dos nodes, la configuració del protocol de pas de missatges és fa de tal manera que s'empraran 32 nuclis de cada node de còmput per dur a terme l'execució, d'aquesta manera els missatges MPI fan ús de la interfície de xarxa Gigabit i s'avalua l'impacte que causa; així doncs un total de 64 nuclis és veuen implicats.



1.2. Gràfica execució HPL a 64 nuclis a dos nodes.

Tal com és reflexa a l'anterior prova a mesura que incrementa NB, el rendiment cau. No obstant si és comparen les dos gràfiques en aquesta el rendiment és molt més inferior. Per últim cal destacar que amb un coeficient de la matriu alt el rendiment és dispara.

A la tercera i última prova (1.3) d'aquest primer grup s'empra la mateixa configuració que a l'anterior, a diferència que ara s'utilitzen tots els nuclis d'ambdós sistemes; un total de 128 nuclis és veuran implicats, 64 nuclis a un node i 64 nuclis a l'altre.

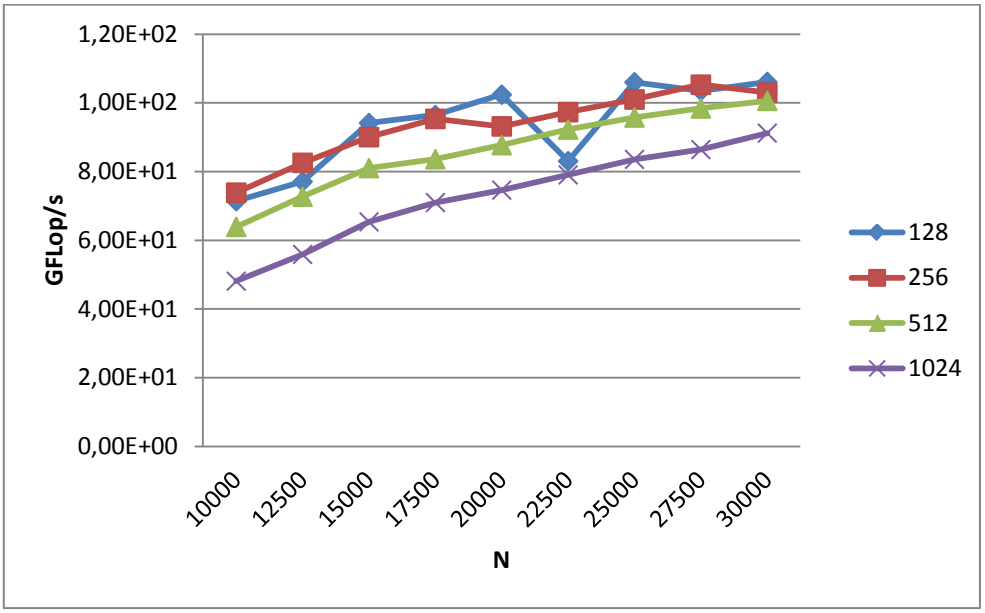


1.3. Gràfica execució HPL a 128 nuclis a dos nodes.

A l'inici del programa la millora de rendiment és imperceptible tot i que és dobleguen el total de nuclis, no obstant el sistema s'estabilitza i millora el resultat respecte l'anterior.

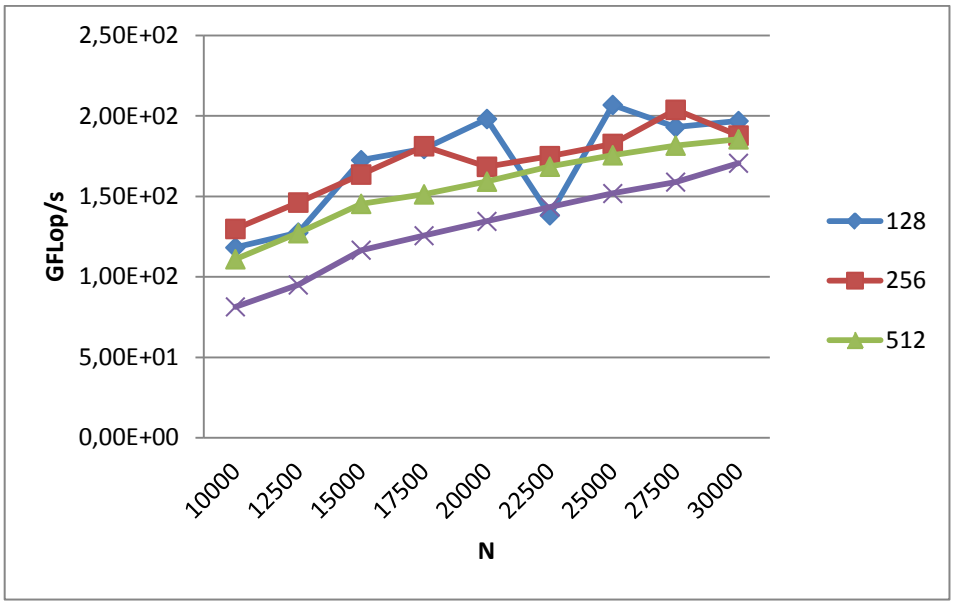
Per al segon Grup de proves s'estudia quin és l'impacte que causa l'execució quan només s'empra un processador amb 16 nuclis a un node, i quan s'utilitzen diferents processadors d'un mateix node amb el mateix número de nuclis.

A la quarta prova (2.1) s'estudia el rendiment d'un processador en tots els seus nuclis, el programa només s'executa en els 16 primers nuclis del primer processador, aquests estan reservats consecutivament del 0 al 15 .



2.1. Gràfica execució HPL a 16 nuclis consecutius a un node.

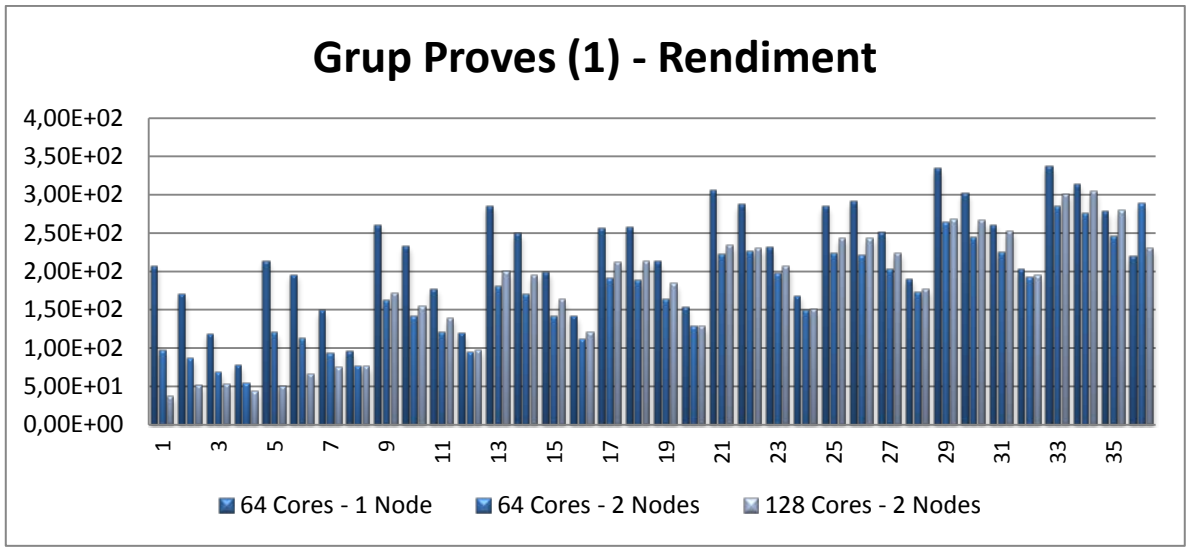
La quinta prova (2.2) avalua el rendiment respecte l'anterior per saber quina és la repercussió de deixar al sistema que esculli els processadors aleatòriament. Així doncs dins d'un mateix node s'empren 16 nuclis per executar el programa.



2.2. Gràfica execució HPL a 16 nuclis no consecutius a un node.

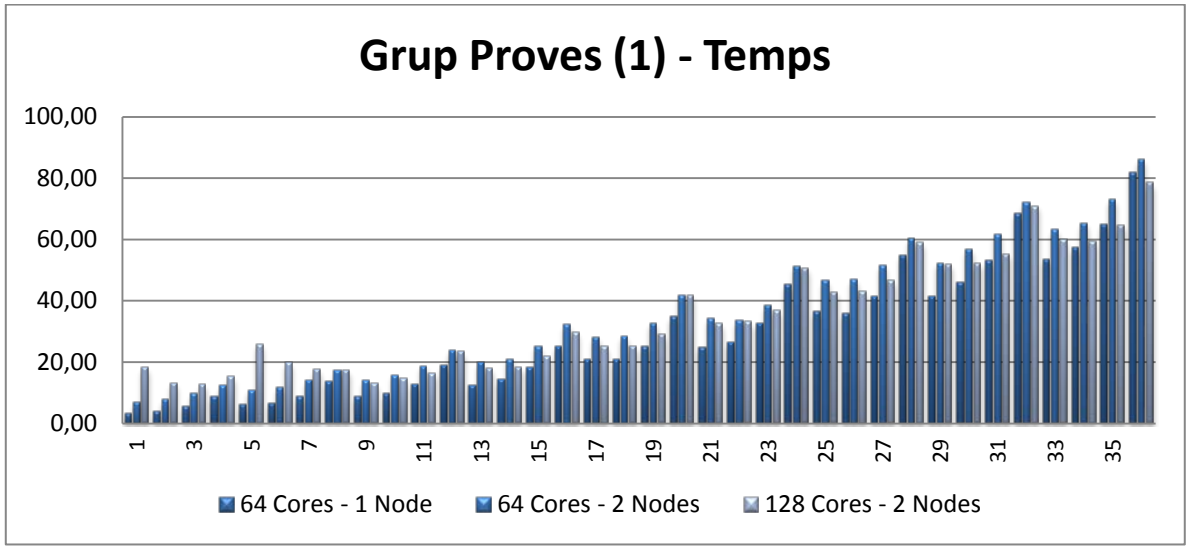
Per contrastar els diferents grups de proves, a continuació és mostren els resultats de rendiment mitjançant diferents unitats (temps i rendiment). Aquests queden agrupats en grups de tres per cadascuna de les 36 proves que genera el Solver.

Per al Primer grup de proves, s'avalua el rendiment segons l'execució per resoldre el sistema linear de l'algoritme.



Gràfica comparativa del rendiment.

la següent gràfica representa el temps (en segons) que ha trigat cadascuna de les proves en executar-se.



Gràfica comparativa del temps.

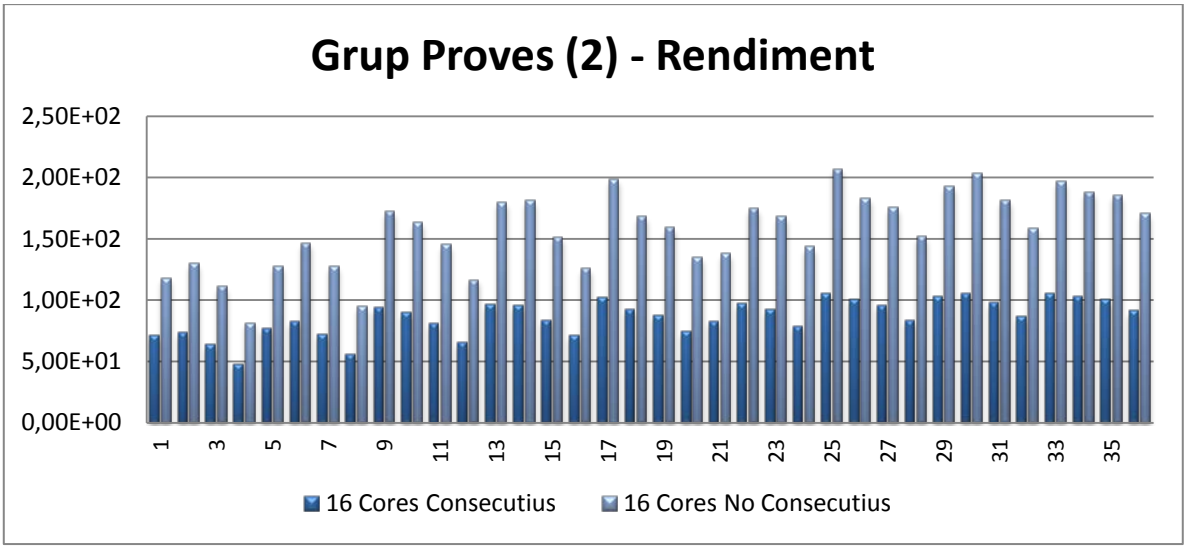
És veu clarament que el factor temps i el rendiment obtingut està ben lligat, els resultats amb un alt rendiment generalment triguen poc temps en executar-se.

Així mateix, a la gràfica de temps s'observa un creixement progressiu a mida que el problema és més gran; quan més gran és el problema més temps triga en resoldre's. Amb un coe-

ficient de la matriu baix (N=10.000 o N=12.500) i un gran nombre de nuclis en diferents sistemes el temps d'execució es duplica, no obstant a mida que N augmenta el rendiment s'equipara.

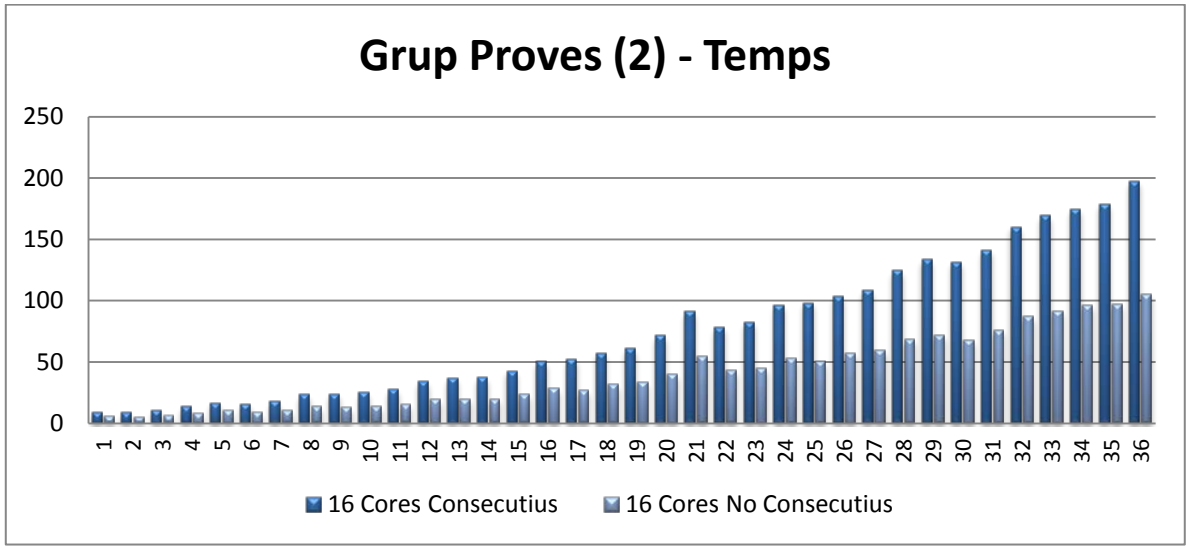
A la gràfica del rendiment i a l'igual que a la del temps amb N's petites emprant l'execució mitjançant la xarxa en dos nodes el rendiment no és tan bo. Cau fins i tot un 60% als primers valors obtinguts, no obstant per valors grans de N i amb mida de bloc (NB = 1024) grans (darrers resultats de la gràfica) és veu com influeix una petita millora i com beneficia repartir la càrrega entre dos nodes amb 64 nuclis.

Al Segon grup de proves, s'avalua de la mateixa manera el rendiment i el temps per cada configuració realitzada al node de còmput, 16 nuclis consecutius i 16 nuclis escollits aleatòriament.



Gràfica comparativa del rendiment.

La següent gràfica mostra la comparativa de temps entres les dues configuracions.



Gràfica comparativa del temps.

La primera fase de proves s'executa sobre el primer nucli del sistema, i la segona sobre els quatre processadors d'un mateix node de còmput. Ambdues amb un total de 16 nuclis. En aquest cas la gràfica de rendiment demostra com empitjora el rendiment quan s'empra només un sol processador (tot i que el número total de nuclis és el mateix), i com influeix greument sobre el rendiment. És demostra que tot i que el nombre de recursos de càlcul és el mateix, influeix moltíssim com estan distribuïts al sistema. Per altra banda la gràfica de temps confirma la pèrdua de rendiment, els temps d'execució en aquest cas son un ±45% més òptims, el que reafirma que emprant molts nuclis d'un mateix processador empitjoren molt els resultats.

Al següent taula mostra un resum dels resultats obtinguts numèricament:

				Primer Grup						Segon Grup			
				8*8 (1 NODE)		8*8 (2 NODES)		8*16 (2 NODES)		4*4 (1 NODE)C.		4*4 (1 NODE)NC.	
TEST	**T/V**	**N**	**NB**	**Time**	**Gflops**	**Time**	**Gflops**	**Time**	**Gflops**	**Time**	**Gflops**	**Time**	**Gflops**
1	**WR01R2R4**	**10000**	**128**	3,23	2,06E+02	6,89	9,68E+01	18,15	3,67E+01	9,34	7,14E+01	5,64	1,18E+02
2	**WR01R2R4**	**10000**	**256**	3,91	1,70E+02	7,72	8,64E+01	13,18	5,06E+01	9,03	7,39E+01	5,14	1,30E+02
3	**WR01R2R4**	**10000**	**512**	5,63	1,19E+02	9,72	6,86E+01	12,69	5,26E+01	10,43	6,39E+01	6	1,11E+02
4	**WR01R2R4**	**10000**	**1024**	8,63	7,73E+01	12,34	5,40E+01	15,30	4,36E+01	13,86	4,81E+01	8,19	8,14E+01
5	**WR01R2R4**	**12500**	**128**	6,09	2,14E+02	10,80	1,21E+02	25,87	5,03E+01	16,9	7,71E+01	10,22	1,28E+02
6	**WR01R2R4**	**12500**	**256**	6,67	1,95E+02	11,59	1,12E+02	19,90	6,54E+01	15,78	8,26E+01	8,91	1,46E+02
7	**WR01R2R4**	**12500**	**512**	8,79	1,48E+02	13,90	9,37E+01	17,41	7,48E+01	17,92	7,27E+01	10,24	1,27E+02
8	**WR01R2R4**	**12500**	**1024**	13,72	9,50E+01	17,17	7,59E+01	17,16	7,59E+01	23,3	5,59E+01	13,71	9,50E+01
9	**WR01R2R4**	**15000**	**128**	8,65	2,60E+02	13,85	1,62E+02	13,09	1,72E+02	23,91	9,41E+01	13,04	1,73E+02
10	**WR01R2R4**	**15000**	**256**	9,70	2,32E+02	15,85	1,42E+02	14,58	1,54E+02	24,99	9,00E+01	13,76	1,64E+02
11	**WR01R2R4**	**15000**	**512**	12,69	1,77E+02	18,80	1,20E+02	16,28	1,38E+02	27,77	8,11E+01	15,48	1,45E+02
12	**WR01R2R4**	**15000**	**1024**	18,93	1,19E+02	23,86	9,43E+01	23,25	9,68E+01	34,42	6,54E+01	19,31	1,17E+02
13	**WR01R2R4**	**17500**	**128**	12,50	2,86E+02	19,83	1,80E+02	17,95	1,99E+02	37,05	9,65E+01	19,91	1,80E+02
14	**WR01R2R4**	**17500**	**256**	14,32	2,50E+02	20,99	1,70E+02	18,40	1,94E+02	37,48	9,53E+01	19,72	1,81E+02
15	**WR01R2R4**	**17500**	**512**	18,03	1,98E+02	25,15	1,42E+02	21,87	1,63E+02	42,73	8,36E+01	23,61	1,51E+02
16	**WR01R2R4**	**17500**	**1024**	25,29	1,41E+02	32,26	1,11E+02	29,70	1,20E+02	50,33	7,10E+01	28,46	1,26E+02
17	**WR01R2R4**	**20000**	**128**	20,80	2,56E+02	27,91	1,91E+02	25,12	2,12E+02	52,1	1,02E+02	26,92	1,98E+02
18	**WR01R2R4**	**20000**	**256**	20,70	2,58E+02	28,40	1,88E+02	24,97	2,14E+02	57,25	9,32E+01	31,68	1,68E+02
19	**WR01R2R4**	**20000**	**512**	25,06	2,13E+02	32,66	1,63E+02	28,94	1,84E+02	60,82	8,77E+01	33,52	1,59E+02
20	**WR01R2R4**	**20000**	**1024**	34,83	1,53E+02	41,80	1,28E+02	41,76	1,28E+02	71,48	7,46E+01	39,67	1,35E+02
21	**WR01R2R4**	**22500**	**128**	24,85	3,06E+02	34,20	2,22E+02	32,41	2,34E+02	91,45	8,31E+01	54,9	1,38E+02
22	**WR01R2R4**	**22500**	**256**	26,40	2,88E+02	33,53	2,27E+02	33,05	2,30E+02	78,04	9,73E+01	43,39	1,75E+02
23	**WR01R2R4**	**22500**	**512**	32,80	2,32E+02	38,53	1,97E+02	36,86	2,06E+02	82,32	9,23E+01	45,07	1,69E+02
24	**WR01R2R4**	**22500**	**1024**	45,20	1,68E+02	51,02	1,49E+02	50,66	1,50E+02	96,08	7,91E+01	52,96	1,43E+02
25	**WR01R2R4**	**25000**	**128**	36,58	2,85E+02	46,57	2,24E+02	42,79	2,44E+02	98,23	1,06E+02	50,38	2,07E+02
26	**WR01R2R4**	**25000**	**256**	35,80	2,91E+02	47,06	2,21E+02	42,89	2,43E+02	103,17	1,01E+02	57,04	1,83E+02
27	**WR01R2R4**	**25000**	**512**	41,51	2,51E+02	51,40	2,03E+02	46,64	2,23E+02	108,77	9,58E+01	59,33	1,76E+02
28	**WR01R2R4**	**25000**	**1024**	54,91	1,90E+02	60,33	1,73E+02	59,03	1,77E+02	124,67	8,36E+01	68,63	1,52E+02
29	**WR01R2R4**	**27500**	**128**	41,45	3,35E+02	52,32	2,65E+02	51,76	2,68E+02	133,92	1,04E+02	71,85	1,93E+02
30	**WR01R2R4**	**27500**	**256**	45,98	3,02E+02	56,71	2,45E+02	52,03	2,67E+02	131,66	1,05E+02	68,03	2,04E+02
31	**WR01R2R4**	**27500**	**512**	53,23	2,61E+02	61,69	2,25E+02	55,02	2,52E+02	140,92	9,84E+01	76,35	1,82E+02
32	**WR01R2R4**	**27500**	**1024**	68,43	2,03E+02	71,86	1,93E+02	70,89	1,96E+02	160,44	8,64E+01	87,24	1,59E+02
33	**WR01R2R4**	**30000**	**128**	53,46	3,37E+02	63,32	2,84E+02	59,91	3,01E+02	169,59	1,06E+02	91,49	1,97E+02
34	**WR01R2R4**	**30000**	**256**	57,41	3,14E+02	65,15	2,76E+02	59,15	3,04E+02	174,82	1,03E+02	95,88	1,88E+02
35	**WR01R2R4**	**30000**	**512**	64,80	2,78E+02	73,06	2,46E+02	64,36	2,80E+02	178,95	1,01E+02	97,05	1,86E+02
36	**WR01R2R4**	**30000**	**1024**	82,06	2,19E+02	86,19	2,89E+02	78,34	2,30E+02	197,36	9,12E+01	105,53	1,71E+02

7.1.4 Conclusions

Tot i que les diferents proves s'han dut a terme amb la mateixa arquitectura de processador 64 bits, i donat que per problemes de disponibilitat dels recursos només s'han emprat dos node de còmput, els resultats obtinguts amb les diferents configuracions mostren quin és el comportament del processador en cada cas.

El rendiment incrementa a mida que el problema creix, per tant amb problemes més grans el problema triga més en resoldre's.

La configuració dels paràmetres del programari Linpack és fonamental per mostrar un bon rendiment. El resultat de multiplicar P i Q dóna el nombre de processos de MPI, aquests paràmetres han de ser tan a prop com sigui possible a la igualtat, quan no és dona el cas s'ha de especificar P inferior a Q per obtenir una graella prou simètrica per no obtenir resultats erronis. A més tal com s'esmentava anteriorment amb valors d'NB superiors a 256 el rendiment cau, no obstant aquest fet només s'agreuja en sistemes que necessiten el 100% dels recursos de còmput. Un bon valor de NB requereix una gran bateria de proves.

Un altre punt a destacar són els valors d' N inferiors a 10.000, Linpack no és prou acurat com per donar una exactitud fiable del rendiment del sistema, per aquest motiu no és representen els resultats per sota d'aquest valor.

La freqüència del processador juga un paper molt important en quan al rendiment, tot i la homogeneïtat d'arquitectures del clúster on tots els nodes disposen de la mateixa generació de processador; en entorns heterogenis generalment la freqüència de rellotge més alta és la que obté millor rendiment.

Per altra banda a les proves realitzades amb diferents nodes de còmput, el rendiment del clúster és sensibilitza al rendiment que proporciona la xarxa, aquesta sensibilitat disminueix a mesura que augmenta la mida del problema.

Finalment, totes les proves s'han realitzat seguint el document d'instal·lació i configuració del fabricant, aquest document figura a la biografia i tracta el procés amb una generació de processador AMD Opteron 6200 Series. Tanmateix els resultats obtinguts son ben semblants.

7.2 Nasa Parallel Benchmarks

El Paral·lel Benchmarks NAS (NPB[89]) són un petit conjunt de programes dissenyats per ajudar a avaluar el rendiment dels supercomputadors paral·lels. Els punts de referència es deriven de les aplicacions de dinàmica de fluids computacional (CFD[90]) i consisteixen en cinc nuclis i tres pseudoaplicacions en la original especificació (NPB 1).

Aquest conjunt de proves s'ha ampliat per incloure nous punts de referència per a la malla adaptativa no estructurada, en paral·lel d'E/S[91], les aplicacions multizona, i xarxes computacionals.

La mida del problema a NPB estan predefinits i es van indicar com diferents classes. Les Implementacions de referència de NPB estan disponibles en models de programació d'ús general com MPI i OpenMP (NPB 2 i NPB 3).

7.2.1 Funcionalitats i característiques principals

- Gradient Conjugat: Aquesta prova calcula una aproximació al valor propi més petit de la matriu simètrica definida positiva.
- Embarrassingly Parallel: S'estima que els límits assolibles superiors de rendiment de punt flotant d'un computador en paral·lel.
- Transformada Ràpida de Fourier: Aquest punt de referència resol una equació diferencial parcial 3D utilitzant un mètode espectral basat en FFT[92].
- Integer Sort: Aquest punt de referència és un programa de classificació paral·lel basat en el cub espècie.
- Multigrid: Aquest punt de referència utilitza un mètode multimalla en cicle V per calcular la solució de l'equació de Poisson[93] escalar 3-D.
- Prova de diferents tècniques en paral·lel d'E/S

[89] Nasa Parallel Benchmarks és un paquet desenvolupat a la NASA a 1991 per avaluar supercomputadors d'alta gama.
[90] La mecànica de fluids computacional (CFD) és una de les branques de la mecànica de fluids que utilitza mètodes numèrics i algorismes per resoldre i analitzar problemes sobre el flux de substàncies.
[91] Les aplicacions utilitzen els dispositius per realitzar l' entrada i sortida(E/S) o (I/O), aquests transfereixen sincrona o asincronament les dades, i poden ser de només lectura o lectura i escriptura.
[92] La transformada ràpida de Fourier (FFT), és una forma molt ràpida i eficient de calcular la transformada discreta de Fourier (DFT) d'un senyal discret i la seva inversa.
[93] L'equació de Poisson és una equació diferencial parcial de tipus el·líptic amb una àmplia utilitat en electrostàtica, l'enginyeria mecànica i la física teòrica.

7.2.2 Execució al clúster

A continuació és fa una breu descripció dels 4 problemes emprats al clúster per tal de comparar i avaluar com escala el sistema.

> **CG.** Mètode de gradient conjugat s'utilitza per calcular una aproximació al valor propi més petit d'una matriu gran. Aquest nucli és típic dels càlculs de la quadrícula no estructurats, ja que posa a prova la comunicació de llarga distància irregular, emprant una estructurada matriu de vectors de multiplicació.
>
> **IS.** Operació d'enters de tipus que és important en determinats codis d'un "mètode de partícules". Posa a prova tant la velocitat de càlcul d'enters i el rendiment de la comunicació.
>
> **EP.** Com l'acrònim indica, es tracta d'un nucli “Embarrassingly Parallel". En contrast amb altres, pràcticament no requereix d'acoblament, només la coordinació de la generació de nombres pseudo-aleatoris a l'inici
> i la recollida dels resultats al final.
>
> **BT.** Realitza un problema sintètic CFD[94] mitjançant la resolució de múltiples sistemes en bloc d’equacions tridiagonals dominants amb (5 x 5) de mida de bloc.

Per dur a terme les proves, els programes s’han compilat emprant dos paradigmes de computació paral·lela, OpenMP i MPI.

El primer proporciona el model d'execució “fork-and-join” on el programa comença la seva execució com un sol procés o subprocés, aquest subprocés executa seqüencialment fins que es trobi una directiva paral·lelització d'una regió paral·lela. En aquest moment, el fil crea un grup de fils i es converteix en el fil principal del nou equip.

El segon MPI, proporciona a l'usuari un model de programació on els processos es comuniquen amb altres processos trucant a rutines de biblioteca per enviar i rebre missatges. L'avantatge del model de programació MPI és que l'usuari té un control complet sobre la distribució de dades i sincronització de processos, el que permet l'optimització de la localitat de les dades i la distribució del flux de treball.

Per aquesta fase de proves només és necessita el recursos d’un node de càlcul, així mateix totes les comparacions és realitzen en primera instancia des de l’execució mínima de recursos requerits (1 procés o 1 fil d’execució) fins arribar a l’ús total dels nuclis del sistema (64 processos o 64 fils d’execució).

[94] CFD o Dinàmica de fluids computacional, és una branca de la mecànica de fluids que utilitza mètodes numèrics i algoritmes per resoldre i analitzar problemes que involucren fluxos de fluids.

Les proves s'avaluen segons els següents escenaris, cal destacar que no tots els programes permeten la seva compilació per alguna determinada classe amb un número específic de fils o processos.

	Class	Threads & Procs					
		1	4	8	16	32	64
<program>	S						
	A						
	B						
	C						
	D						

Model de taula emprada per fer el registre de proves.

Les escales o 'Class' indiquen la mida del problema, on S son els programes més lleugers fins la D amb programes de més grandària.

Tots els resultats s'expressen i és comparen en 'segons', al final de l'execució dels programes és retorna el total de temps que ha trigat.

```
...
Verification Successful
BT Benchmark Completed.
Time in seconds = 2.83
```

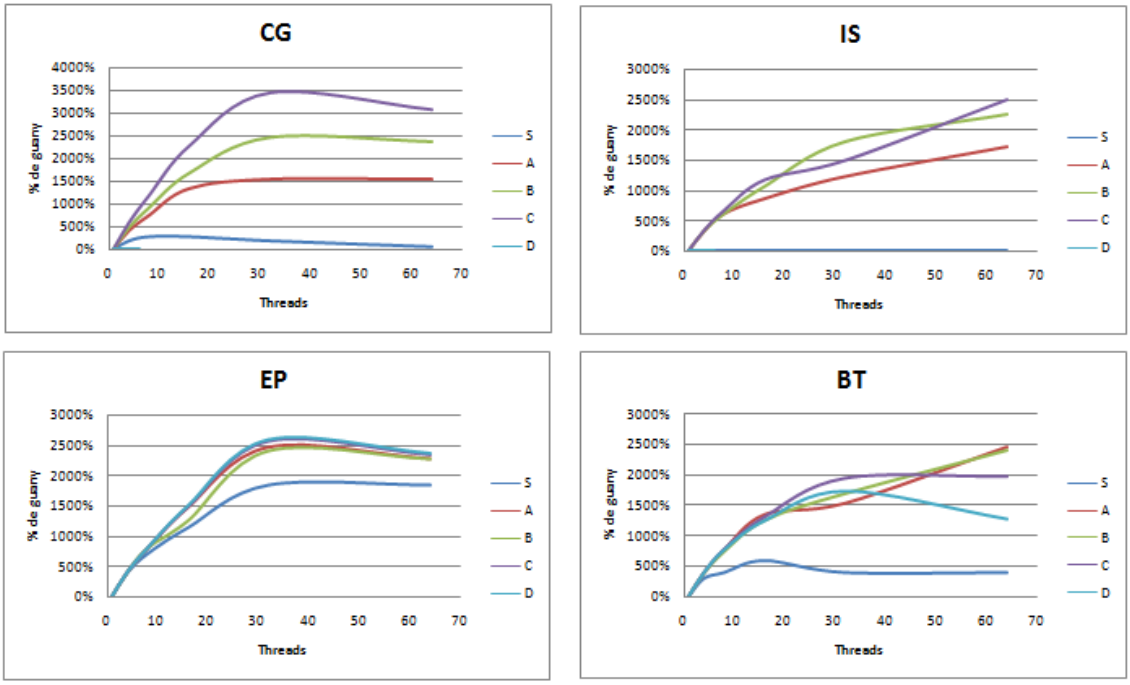
7.2.3 Resultats obtinguts

Les proves és divideixen en dues parts per tal d'avaluar quin benefici existeix alhora d'executar un programar mitjançant fils (OpenMP) o processos (MPI).

La vista de les gràfiques s'ha construït a partir del temps que ha trigat en executar-se l'aplicació, així mateix és calcula el percentatge de guany en funció dels resultats obtinguts de l'execució respecte un únic procés o fil.

7.2.3.1 OpenMP

És mostren els resultats obtinguts en format de gràfica per veure quina és la tendència de cadascuna de les execucions realitzades. L'eix de les Y's mostra el percentatge de guany respecte de l'execució del programa amb un fil, mentre que l'eix de les X's indica el total de fils implicats a l'execució.



A la següent taula és mostren els resultats en temps obtinguts.

Benchmark	Class	Threads					
		1	4	8	16	32	64
CG	S	0,06	0,03	0,02	0,02	0,03	0,11
	A	2,03	0,50	0,27	0,15	0,13	0,13
	B	106,67	22,78	11,77	6,39	4,35	4,52
	C	383,16	66,18	32,42	16,77	11,08	12,42
	D	X	X	X	X	X	X
IS	S	0,00	0,01	0,01	0,01	0,01	0,01
	A	0,86	0,27	0,14	0,10	0,07	0,05
	B	3,63	1,12	0,59	0,34	0,20	0,16
	C	16,71	5,04	2,52	1,43	1,12	0,67
	D	X	X	X	X	X	X

EP	S	1,47	0,38	0,21	0,13	0,08	0,08
	A	23,26	5,83	2,95	1,58	0,94	1,02
	B	92,92	23,47	11,66	7,45	3,86	4,09
	C	371,55	93,06	46,46	25,00	14,41	15,80
	D	5933,27	1486,74	745,80	397,74	229,21	251,10
BT	S	0,12	0,04	0,03	0,02	0,03	0,03
	A	82,29	21,04	10,75	6,10	5,35	3,34
	B	410,56	108,48	55,85	32,82	24,26	17,07
	C	1401,35	360,95	181,17	107,33	72,05	71,09
	D	29572,35	7564,05	3856,82	2371,58	1704,93	2318,73

(Els resultats marcats amb una X no estan disponibles per la compilació.)

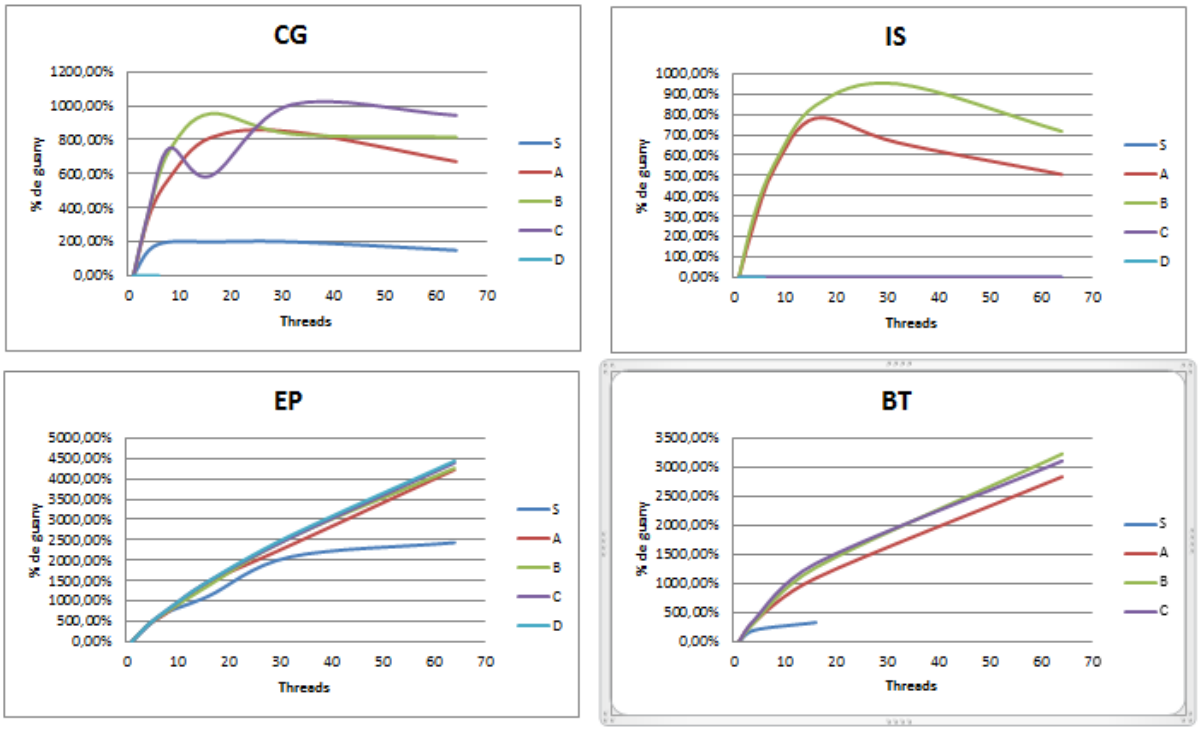
Segons totes les proves realitzades i analitzant les gràfiques és pot identificar que generalment a CG i a EP l'increment s'estabilitza a partir dels 30 fils d'execució; no obstant en el cas de les CG amb la mida de classe més gran 'C' (unes 150.000 cel·les) el rendiment no és molt bo.

Per altra banda tant les IS com les BT tenen un creixement progressiu fins arribar al número màxim de fils. De la mateixa manera que les CG en el cas de la classe 'C', les BT té un comportament similar; a diferencia aquestes computen una malla de mida 162x162x162.

Aquests resultats demostren que amb una mida del problema petita el rendiment no és molt bo, però a mida que és va incrementant aquesta millora substancialment. Observant les gràfiques és pot veure que generalment a partir dels 32 fils el rendiment s'estabilitza o fins i tot cau un mica fins arribar als 64.

7.2.3.2 MPI

De la mateixa manera que les anteriors proves és mostren les gràfiques amb el percentatge de guany obtingut respecte l'execució d'un procés.



A la següent taula és mostren els resultats en temps obtinguts.

Benchmark	Class	Procs					
		1	4	8	16	32	64
CG	S	0,06	0,04	0,03	0,03	0,03	0,04
	A	1,95	0,56	0,34	0,24	0,23	0,29
	B	84,54	22,29	11,50	8,86	10,13	10,36
	C	233,86	60,92	31,16	39,82	23,19	24,76
	D	X	X	X	842,36	420,99	1287,01
IS	S	0,00	0,00	0,00	0,00	0,01	0,04
	A	0,86	0,32	0,16	0,11	0,13	0,17
	B	3,80	1,24	0,68	0,45	0,40	0,53
	C	X	5,37	2,95	1,98	2,08	2,08
	D	X	X	X	X	29,21	39,30
EP	S	1,46	0,39	0,20	0,13	0,07	0,06
	A	23,23	5,87	3,20	1,59	0,98	0,55
	B	93,23	23,42	12,45	6,76	3,64	2,19
	C	371,81	94,03	47,00	25,05	14,56	8,47
	D	5962,95	1504,29	753,24	400,96	228,90	134,40
BT	S	0,10	0,05	X	0,03	X	X
	A	94,97	27,81	X	8,68	X	3,35
	B	409,53	124,39	X	32,09	X	12,69
	C	1647,30	437,37	X	122,30	X	53,01
	D	X	X	X	2320,46	X	1351,74

(Els resultats marcats amb una X no estan disponibles per la compilació.)

Analitzant les gràfiques és denoten alteracions segons sigui el programa, en el cas de les CG existeix un increment a mida que augmenten els número processos en execució, tot i així s'estabilitza al arribar als 16 processos.

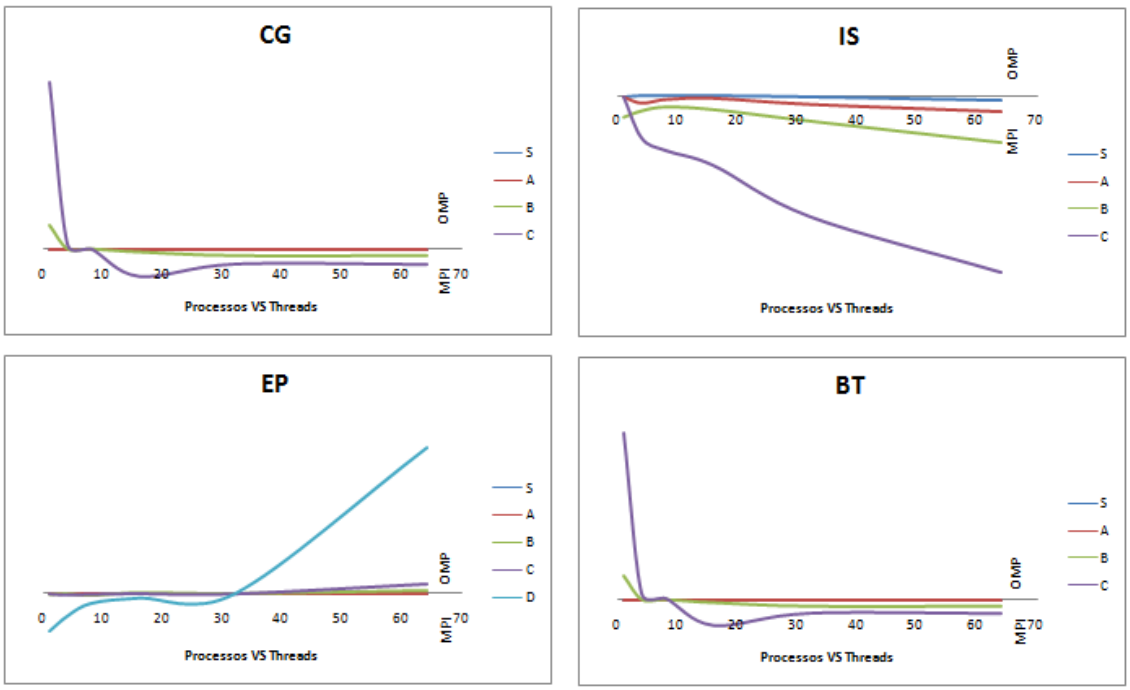
Les EP i BT tenen un guany de rendibilitat molt progressiu, donant valors molt bons que arriben a ser un 3.000% superior respecte l'execució inicial.

Finalment el rendiment de les IS és molt similar entre ell, ja que els processos triguen molt pocs temps en finalitzar el càlcul és molt difícil arribar a una conclusió prou acurada.

7.2.4 Conclusions

Finalment per comparar quin dels dos paradigmes (OpenMP o MPI) és millor per cada programa, és realitza un estudi comparatiu en format gràfic per tal de treure conclusions segons sigui el temps d'execució.

Les comparacions s'han realitzat en format de temps en segons (tal com indiquen les taules anteriors) i mostren en quin àmbit de diferencia treballa cada programa (CG, IS, EP i BT).



CG, l'execució amb una mida del problema gran (classes B i C) i amb un número de fils petit és més eficient, pel contrari a partir de 16 fils el programa funciona més eficientment per intercanvi de missatges MPI donant un rendiment lleugerament superior.

IS, en tots els casos la opció més encertada és emprar MPI, especialment en casos on la mida del problema és relativament gran. Les diferencies de temps son gairebé imperceptibles, ja que no arriben a ser superiors als dos segons de diferencia.

EP, el programa és molt canviant per una mida del problema gran (classe D), en aquest cas comença amb una notable millora de temps de 29 segons per l'execució amb MPI; i va empitjorant fins als 32 processos fins que arriba als 116 segons de diferencia amb 64 fils d'execució per OpenMP.

BT, petites diferencies molt estables de rendiment que atorguen l'eficiència a l'execució mitjançant pas de missatges, no obstant sobte veure com el rendiment és millor amb poc fils a les classes B i C arribant fins als 200 segons de diferencia en aquest últim.

8 CONCLUSIONS I FUTURES ACTUACIONS

És defineixen les conclusions del projecte i les línies d'actuació futures que s'identifiquen durant el procés.

8.1 Conclusions.

Aquest projecte ha contribuït de manera molt important per identificar i satisfer les necessitats dels diferents departaments de la empresa. Portar a terme aquest projecte beneficia molt positivament en la manera que els usuaris investigadors porten a terme l'execució dels projectes als quals estan assignats.

En aquesta línia, els punts més importants per un projecte d'aquestes característiques son; la detecció de les necessitats reals de les persones que treballen directament amb els sistemes, comprometre a l'usuari i la seva experiència a facilitar les seves necessitats, més la definició d'una manera evident dels beneficis que comporta una inversió d'aquests tipus; son unes de les raons que donen pes al projecte.

Després de conèixer amb una mica de profunditat els aspectes més teòrics i poder contrastar l'experimentació dels antics sistemes, s'extreu com a conclusió principal la renovació i actualització del maquinari seguint les actuals tendències de l'àmbit tecnològic la redacció d'una proposta interessant.

El desenvolupament e implantació d'aquest projecte s'ha centrat en l'aprofitament dels recursos al màxim, de tal manera que s'agrupen els sistemes de gestió mitjançant la tecnologia de virtualització basada en Xen, mentre que és dediquen els sistemes amb més recursos purament al processament i computació de dades; i pel que fa a les comunicacions, la segmentació de la xarxa facilita i millora notablement aquest servei.

Un dels principals valors afegits d'aquest projecte és la instal·lació, configuració i desplegament de les instal·lacions dels servidors de còmput, aquest és realitza mitjançant un servidor centralitzat que emmagatzema tota la informació necessària per executar aquesta tasca; cal destacar que tot aquest procés és fa gràcies a que els servidors arrenquen des d'un nucli extern.

Per altra banda els usuaris disposen d'un repositori comú de programari, l'escalabilitat d'aquestes aplicacions fan possible un repartiment de la càrrega entre diversos nodes, on amb un major número de nodes s'obté un major rendiment en temps de resposta per als problemes més complexos. Tot i així i depenent de les característiques de l'aplicació, aquestes son molt sensibles als paràmetres de la xarxa com la latència o l'ample de banda; és per això que les interfícies emprades al clúster poden suposar un problema segons sigui

el cas.
El sistema SGE és un gestor de recursos distribuïts (DRM) que administra la distribució de les càrregues de treball dels usuaris tenint en compte la disponibilitat d'aquests recursos. Aquest sistema de gestió de tasques permet executar treballs de manera controlada i ordenada, per tant segons sigui la prioritat i depenent de la disponibilitat de les cues, les tasques s'executaran de manera equitativa entre tots els usuaris.

8.2 Futures actuacions.

Els següents punts descriuen breument les futures línies a seguir en aquest projecte:

- Instal·lació d'una xarxa de baixa latència Infiniband per maximitzar l'ample de banda i accelerar l'intercanvi de dades, tal com s' esmentat a les conclusions segons sigui el cas una xarxa de baixa latència pot suposar un benefici a l'execució de les tasques.

- Instal·lació de xips de processament gràfic GPU, optimitzats per dur a terme càlcul de valors en coma flotant més ràpids i eficients, tot i que encara no és dona la casuística s'estudiarà la possibilitat d'implantació d'aquest tipus de tecnologia.

- Distribució dels espais d'usuari mitjançant quotes d'ús, per dur a terme un control d'ús de l'espai utilitzat pels usuaris.

- Migració de les màquines virtuals al sistema d'emmagatzemament en xarxa, per tal de facilitar i aprofitar les diferents tecnologies de còpies de seguretat de la cabina de discos NetApp.

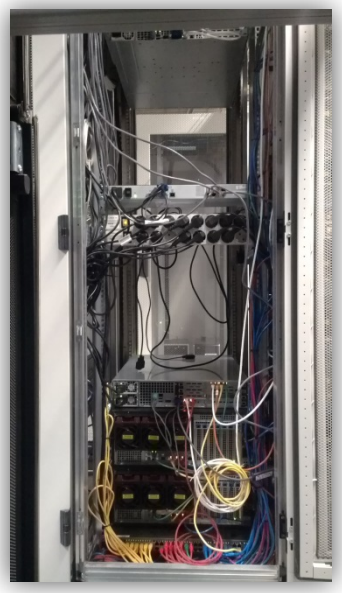
9 ANNEXOS

A partir d'aquest punt és defineixen els annexos de tots els processos descrits a la memòria per tal d'alleugerir el contingut d'aquesta.

Annex I. Vista final del muntatge del maquinari.

Les següents imatges mostren el muntatge final del maquinari als racks, per motius de seguretat no és mostra el rack que conté la cabina de discos.


Vista frontal del maquinari als bastidors 1.1 i 1.2 (segons disseny).

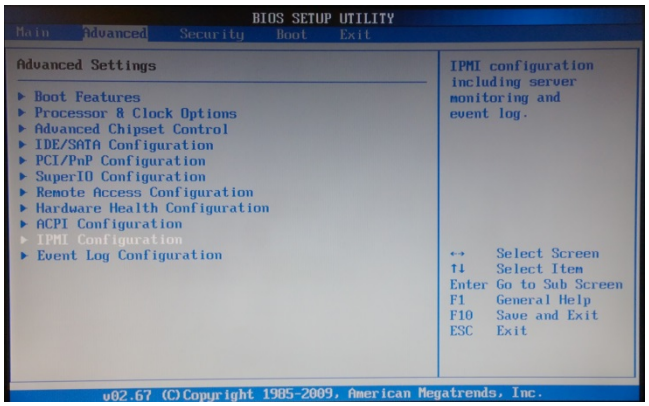



Vista posterior del maquinari als bastidors 1.1 i 1.2 (segons disseny).

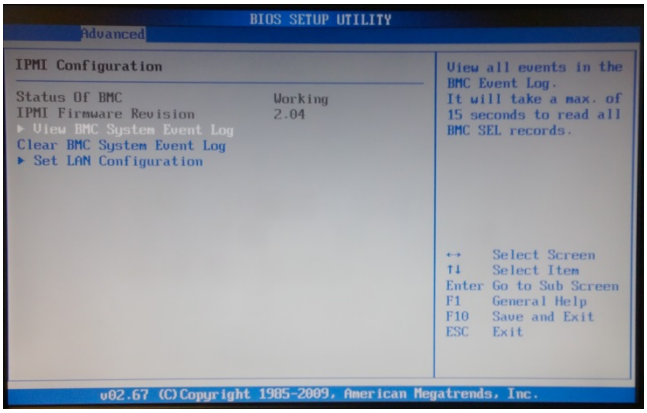
Annex II. Configuració targetes IPMI.

La configuració de les targetes IPMI és fa mitjançant la BIOS de la placa base del sistema, així doncs, durant la primera arrencada cal prémer [Supr] per obrir el menú.

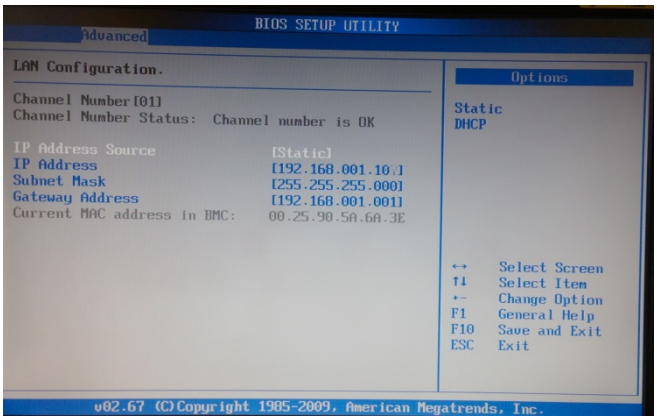
1. Accedir al menú de configuració IPMI.



2. Accedir al menú 'Set LAN Configuration' per establir la configuració de xarxa.



3. Fixar una adreça IP del rang 192.168.1.0/24 reservada per la xarxa de gestió remota a cadascun dels nodes.



Annex III. Instal·lació del sistema operatiu del node primari.

**La instal·lació d'aquest annex s'ha reproduït en una màquina virtual sota VMware Workstation[95] per tal de fer la documentació i captures de pantalles adients.*

En aquest apartat s'han obviat les pantalles del procés d'instal·lació que no tenen prou rellevància.

1. Descarregar imatge .iso del web oficial de debian (https://www.debian.org/CD/http-ftp/#stable) i seleccionar amd64.

Imatges oficials de CD/DVD de la versió «stable»»

Per instal·lar Debian en una màquina sense connexió a Internet, és possible usar imatges de CD (650 MiB cadascuna) o imatges de DVD (4.4 GiB cadascuna). Descarregueu el primer fitxer de la imatge de CD o DVD, cremeu-la fent servir una gravadora de CD/DVD, i torneu a iniciar des d'aquesta.

El **primer** disc de CD/DVD conté tots els fitxers necessaris per instal·lar un sistema Debian estàndard.
Per evitar descarregues innecessàries, si us plau **no** descarregueu altres fitxers d'imatges de CD o DVD a menys que sapigueu que necessiteu paquets trobats en aquestes.

CD
Els següents enllaços apunten a fitxers d'imatges amb mides de fins a 650 MiB, fent-los adequats per a cremar en medis normals de CD-R(W):

amd64, armel, armhf, i386, ia64, kfreebsd-i386, kfreebsd-amd64, mips, mipsel, powerpc, sparc, s390, s390x, source, multi-arch

DVD
Els següents enllaços apunten a fitxers d'imatges amb mides de fins a 4.4 GiB, fent-los adequats per a cremar en suports normals de DVD-R/DVD+R i suports similars:

amd64, armel, armhf, i386, ia64, kfreebsd-i386, kfreebsd-amd64, mips, mipsel, powerpc, sparc, s390, s390x, source, multi-arch

2. Iniciar la instal·lació amb la nova iso copiada a un USB.

3. Seleccionar:

Idioma: [english]
Territori: [Europe]
Teclat: [Catalan]
Regió geogràfica [Spain]

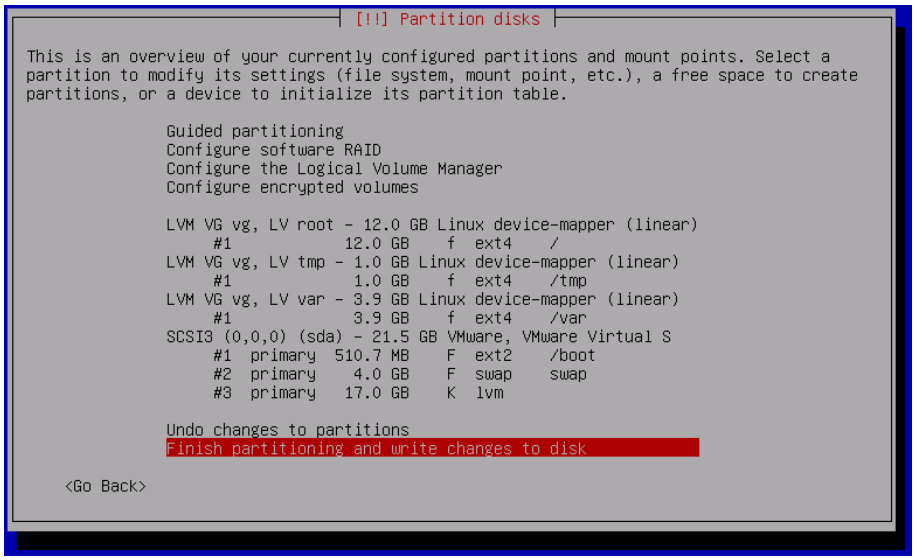
[95] VMware Workstation és un sistema de virtualització per programari ja que permet l'emulació en plataformes personals X86, això permet que qualsevol usuari amb una computadora d'escriptori o portàtil pugui emular tantes màquines virtuals com els recursos de maquinari ho permetin.

4. Configurar la xarxa definint els següents paràmetres:

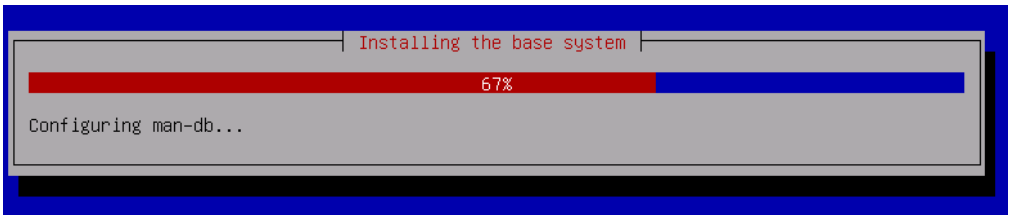
Hostname: head-master
Domain: localnet
New username: sysadmin

Password: ?????
Time zone: Berlin

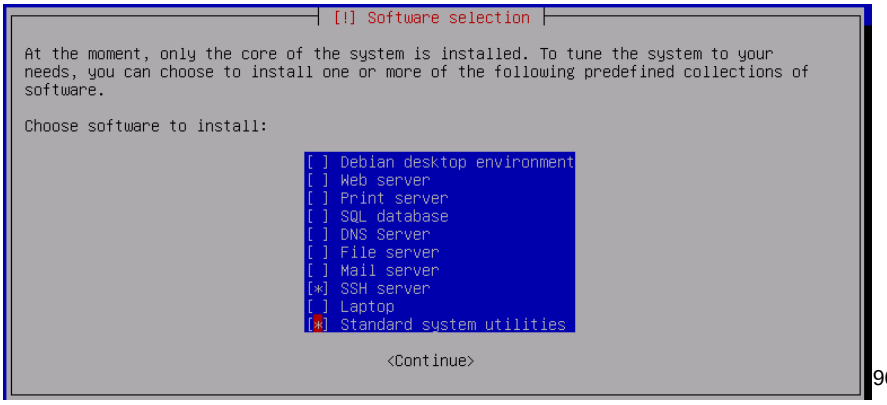
5. Particionament del disc.



6. Instal·lació del sistema base de debian.



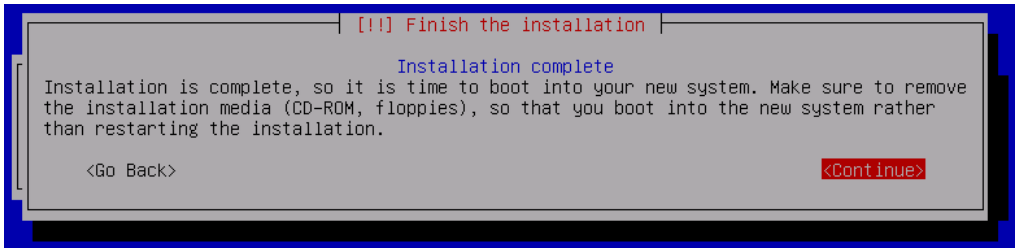
7. Seleccionar el programari requerit per al sistema base, gestió remota més utilitats.



[96]

[96] SSH server: Programa que utilitza el protocol Secure Shell per acceptar connexions des d'equips remots.

8. Reiniciar el servidor un cop instal·lats els paquets.



9. La primera arrencada del sistema, és mostra la sortida de la consola de comandes.

Annex IV. Instal·lació Xen Hypervisor.

Instal·lació del configurador de xarxa:

```
$ apt-get install bridge-utils
```

Instal·lació sistema Xen més les eines:

```
$ apt-get install xen-linux-system xen-tools
```

[BUG] Carregar el xen hypervisor abans que el genèric de linux:

```
$ mv /etc/grub.d/10_linux /etc/grub.d/50_linux
$ update-grub2
$ ls /etc/grub.d/
00_header  05_debian_theme  20_linux_xen  30_os-prober 40_custom  41_custom  50_linux README
$ reboot
```

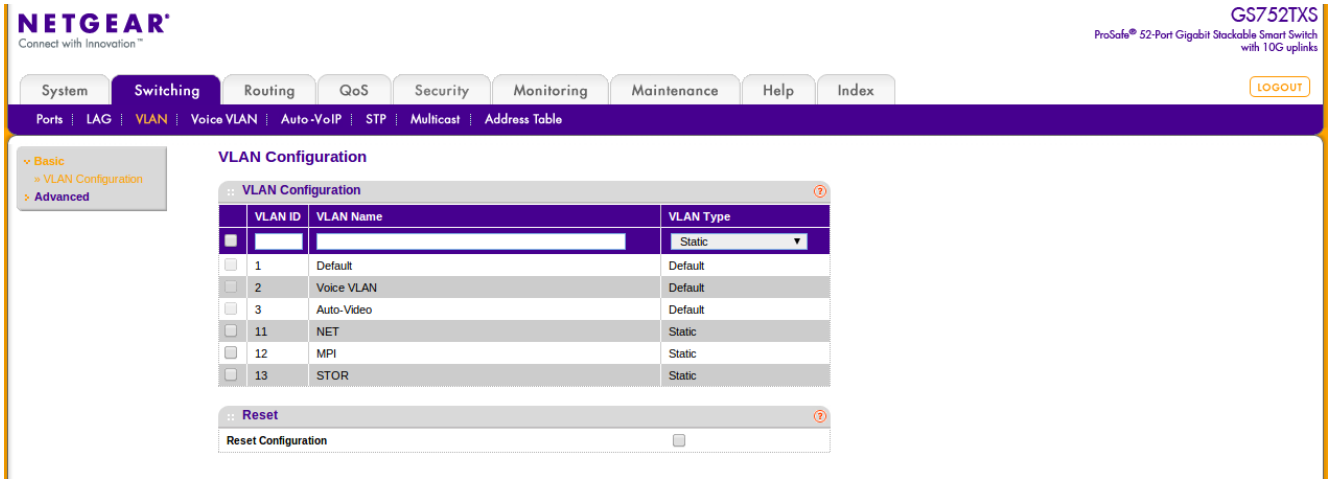
Verificació del domai0 carregat:

```
$ xm list
Name                        ID   Mem VCPUs      State   Time(s)
Domain-0                     0 14046    24      r-----     25.7
```

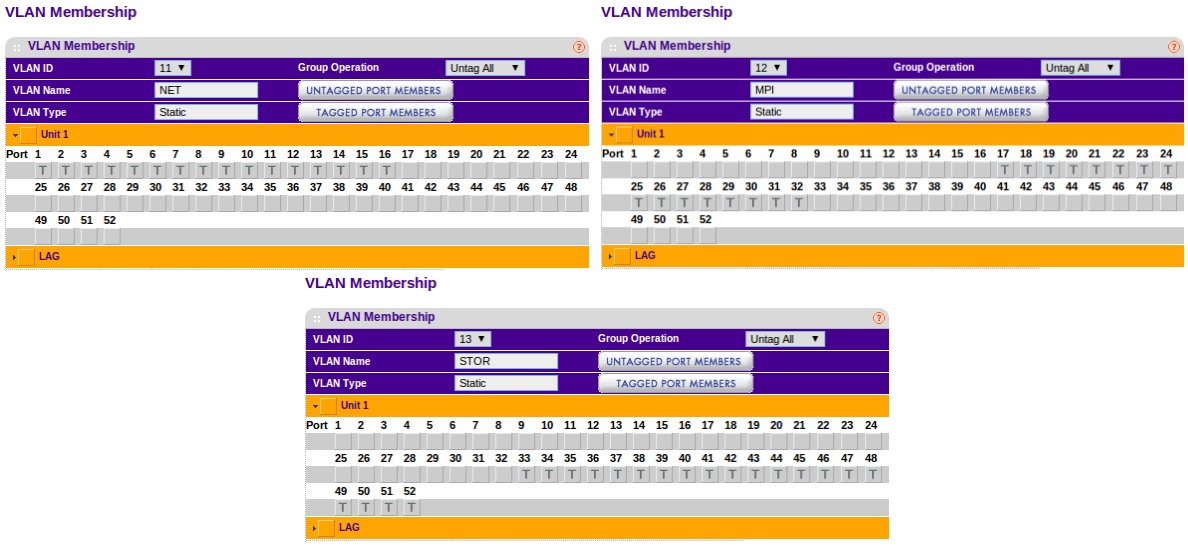
Annex V. Configuració xarxes físiques i virtuals.

Xarxes Físiques del Switchos:

És defineixen tres VLAN per tal de separar físicament la xarxa de gestió DATA, xarxa d'intercanvi de missatges MPI i la xarxa NFS.



L'assignació de ports és realitza equitativament per les tres xarxes.



S'assignen dos ports per la connexió 10G amb la cabina de discos per la xarxa NFS [VLAN:13]

☐	1/xg49			Enable	10G Full	10G Full	Link Up	Enable
☐	1/xg50			Enable	10G Full	10G Full	Link Up	Enable

Xarxes Virtuals del node Xen:

Segons els disseny de xarxes i assignacions d'adreçament IP és configuren els dispositius de pont per les màquines virtuals Xen, tota la configuració és defineix al fitxer interfícies de sistema. [/etc/network/interfaces]

Interface eth0: Creació de la interfície primària per la gestió del servidor Xen. S'habilita el POSTROUTING per tal de habilitar l'accés als serveis corporatius com LDAP, DNS i servidors de llicencies que venen de 'xenbr0'.

```
# The primary network interface
auto xpriv55
iface xpriv55 inet static
        bridge_ports eth0
        address 10.55.0.99
        netmask 255.255.255.0
        network 10.55.0.0
        broadcast 10.55.0.255
        gateway 10.55.0.245
        dns-nameservers 10.80.89.149
        dns-search s.localdomain
        pre-up iptables -t nat -A POSTROUTING -s 172.16.0.0/24 -d 172.16.1.0/24 -j ACCEPT
        pre-up iptables -t nat -A POSTROUTING -s 172.16.0.0/24 -j MASQUERADE
        pre-up brctl addbr xpriv55
        pre-up brctl stp xpriv55 off
        pre-up brctl addif xpriv55 eth0
        pre-up ifconfig eth0 0.0.0.0
        pre-up echo '1' > /proc/sys/net/ipv4/ip_forward
        post-down brctl delif xpriv55 eth0
```

Interface eth2: Definició de la interfície de xarxa per tal que les màquines virtuals puguin accedir a la cabina de discos per NFS mitjançant POSTROUTING.

```
# BRIDGE NFS
auto xnfs0
iface xnfs0 inet static
        bridge_ports eth2
        address 0.0.0.0
        pre-up iptables -t nat -A POSTROUTING -s 10.40.120.0/24 -d 10.40.120.0/24 -j ACCEPT
        pre-up iptables -t nat -A POSTROUTING -s 10.40.120.0/24 -j MASQUERADE
```

Interface eth3: Definició de la interfície de xarxa per la xarxa de gestió que connectarà els nodes de còmput amb les màquines virtuals. En aquest és defineixen els POSTROUTING per tal d'accedir als serveis corporatius.

```
# BRIDGE XEN
auto xenbr0
iface xenbr0 inet static
    bridge_ports eth3
    address 192.168.0.251
    netmask 255.255.255.0
    network 192.168.0.0
    broadcast 192.168.0.255
    # NAT LDAP
        pre-up iptables -t nat -A POSTROUTING -s 192.168.0.0/24 -d 10.80.89.100 -o xpriv55 -p tcp --dport 636 -j SNAT --
to-source 10.55.0.99
```

```
        pre-up iptables -t nat -A POSTROUTING -s 192.168.0.0/24 -d 10.80.89.100 -o xpriv55 -p tcp --dport 389 -j SNAT --
to-source 10.55.0.99

        # NAT MATLAB
    pre-up iptables -t nat -A POSTROUTING -s 192.168.0.0/24 -d 10.60.4.208 -o xpriv55 -p tcp --dport 27000 -j SNAT --to-
source 10.55.0.99
    pre-up iptables -t nat -A POSTROUTING -s 192.168.0.0/24 -d 10.60.4.208 -o xpriv55 -p tcp --dport 1047 -j SNAT --to-
source 10.55.0.99

        # DNS Rules
        pre-up iptables -t nat -A POSTROUTING -s 192.168.0.0/24 -d 10.80.89.149 -o xpriv55 -p tcp --dport 53 -j SNAT --
to-source 10.55.0.99
    pre-up iptables -t nat -A POSTROUTING -s 192.168.0.0/24 -d 10.80.89.149 -o xpriv55 -p udp --dport 53 -j SNAT --to-
source 10.55.0.99
```

Interface eth4: Definició de la interfície de xarxa per de dedicar un dispositiu de xarxa físic a la xarxa corporativa privada. Per aquesta és connectaran els usuaris mitjançant SSH al servidor virtual 'login'.

```
# BRIDGE privada login
auto xlogin0
iface xlogin0 inet static
    bridge_ports eth4
    address 0.0.0.0
```

Interface dummy0: És defineix la interfície de xarxa dummy[97] per tal de comunicar els servidors virtuals internament.

```
# BRIDGE DMZ
auto xdmz0
iface xdmz0 inet static
        address 172.16.0.254
        netmask 255.255.255.0
        network 172.16.0.0
        broadcast 172.16.0.255
        pre-up brctl addbr xdmz0
        pre-up brctl stp xdmz0 off
        pre-up brctl addif xdmz0 dummy0
        pre-up ifconfig dummy0 0.0.0.0
        post-down brctl delif xdmz0 dummy0
```

Finalment, amb la comanda 'brctl' és poden verificar els 'ponts' creats:

```
$ brctl show
bridge name     bridge id               STP enabled     interfaces
xdmz0           8000.9e0e868ce1cc       no              dummy0
xenbr0          8000.002655d9f355       no              eth3
xlogin0         8000.002655d9f354       no              eth4
xnfs0           8000.001e0b1f9cac       no              eth2
xpriv55         8000.002655da8d24       no              eth0
```

[97] Els dispositius de xarxa virtual *dummy* tenen les mateixes funcionalitats que una interfície de xarxa física, són usats per crear xarxes privades que no tinguin accés a una xarxa física.

Annex VI. Instal·lació màquina virtual base de template.

Des del servidor de màquines virtuals Xen:

1. Instal·lar LVM

```
$ apt-get install lvm
```

2. Crear una partició sense format a l'espai disponible dels disk.

3. Crear un volum físic:

```
$ pvcreate /dev/cciss/c0d0p3
  Writing physical volume data to disk "/dev/cciss/c0d0p3"
  Physical volume "/dev/cciss/c0d0p3" successfully created
```

4.Crear grup de volums:

```
$ vgcreate vg /dev/cciss/c0d0p3
  Volume group "vg" successfully created
```

5. Configurar xen-tools, és defineixen paràmetres bàsics per tal de crear una imatge de disc mitjançant els repositoris de debian, el sistema és descarregarà la darrera versió 'wheezy' d'aquesta distribució mitjançant un bootstrap[98]:

```
install-method = debootstrap
lvm = vg
install-method = debootstrap
size   = 5Gb      # Disk image size.
memory = 1Gb    # Memory size
swap   = 1Gb    # Swap size
noswap = 1      # Don't use swap at all for the new system.
fs     = ext3     # use the EXT3 filesystem for the disk image.
dist   = wheezy
image  = sparse   # Specify sparse vs. full disk images.
dhcp = 1
nameserver = 10.80.89.149
bridge = xdmz0
kernel = /boot/vmlinuz-`uname -r`
initrd = /boot/initrd.img-`uname -r`
mirror = `xt-guess-suite-and-mirror --mirror`
ext3_options     = noatime,nodiratime,errors=remount-ro
ext2_options     = noatime,nodiratime,errors=remount-ro
xfs_options      = defaults
reiserfs_options = defaults
btrfs_options    = defaults
```

6. Crear nova imatge:

```
$ xen-create-image --hostname=template
```

[98] Bootstrapping es refereix al procés de càrrega del programari de base a la memòria d'un ordinador, després de reiniciar-lo aquest sistema operatiu ha de fer-se càrrec de la càrrega del sistema i d'un altre programari, segons sigui necessari.

7. Configurar guest.cfg:

```
kernel      = '/boot/vmlinuz-3.2.0-4-amd64'
ramdisk     = '/boot/initrd.img-3.2.0-4-amd64'

vcpus       = '1'
memory      = '1024'

#  Disk device(s).
#
root        = '/dev/xvda1 ro'
disk        = [
                'phy:/dev/vg/template-disk,xvda1,w',
              ]

#  Physical volumes
#
#  Hostname
#
name        = 'template'

#  Networking
#
vif         = [
'mac=00:16:3E:97:0D:1A,bridge=xenbr0','mac=00:16:3E:97:0D:2A,bridge=xdmz0','mac=00:16:3E:97:0D:3A,bridge=xnfs0' ]

#  Behaviour
#
on_poweroff = 'destroy'
on_reboot   = 'restart'
on_crash    = 'restart'
```

8. Muntar sistema de fitxers del template:

```
$ mount /dev/mapper/vg-template--disk /mnt/
```

9. Configurar les targetes de xarxa de la plantilla amb les ip's prefixades:

```
auto eth0
iface eth0 inet static
  address 192.168.0.199
  netmask 255.255.255.0
  network 192.168.0.0
  broadcast 192.168.0.255

auto eth1
iface eth1 inet static
  address 172.16.0.199
  netmask 255.255.255.0
  network 172.16.0.0
  broadcast 172.16.0.255

auto eth2
iface eth2 inet static
  address 10.40.120.199
  netmask 255.255.255.0
  network 10.40.120.0
  broadcast 10.40.120.255
```

10. Muntar /mnt com a chroot

```
$ chroot /mnt/
```

11. Actualitzar la contrasenya de root mitjançant chroot:

```
$passwd
Enter new UNIX password:
Retype new UNIX password:
```

810. Desmuntar sistema de fitxers:

```
$ umount /mnt
```

11. Arrencar màquina virtual:

```
$ xm create /etc/xen/template.cfg
```

12. Verificar estat màquina virtual:

```
$ xm list
Name                        ID   Mem VCPUs      State   Time(s)
Domain-0                     0 14046    24       r-----     25.7
template                     1  1024     1       -b----      4.4
```

13. Agafar la consola:

```
$ xm console template
```

14. Connectar-se per SSH:

```
$ ssh root@template
 The authenticity of host 'template (192.168.0.199)' can't be established.
 ECDSA key fingerprint is 43:df:c2:1b:f8:eb:c6:62:8b:fc:98:d9:75:6e:2b:9d.
 Are you sure you want to continue connecting (yes/no)? yes
 Warning: Permanently added 'template,192.168.0.199' (ECDSA) to the list of known hosts.
 root@template's password:
 Linux template 3.2.0-4-amd64 #1 SMP Debian 3.2.54-2 x86_64

 The programs included with the Debian GNU/Linux system are free software;
 the exact distribution terms for each program are described in the
 individual files in /usr/share/doc/*/copyright.

 Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
 permitted by applicable law.
 Last login: Tue Dec 25 03:14:55 2007
 root@template:~#
```

15. Copiar el fitxer de hosts del head master al template:

```
$ scp -rp /etc/hosts root@template:/etc/
```

16. Instal·lar cau de servei de noms:

```
$ apt-get install libnss-ldap
```

17. Configurar servei nss editant el fitxer /etc/libnss-ldap.conf:

```
# Your LDAP server. Must be resolvable without using LDAP.
uri ldap://10.80.89.149

# The distinguished name of the search base.
base dc=dtldap
```

18.Afegir servei d'autenticació ldap al fitxer /et/nsswitch:

```
passwd:         compat  ldap
group:          compat  ldap
shadow:         compat  ldap
```

19.Instal·lar el paquet d'autenticació ldap:

```
$ apt-get install libpam-ldap
```

20. Configurar els següents fitxers:

/etc/pam.d/common-account:

```
account    required      pam_unix.so
account    sufficient    pam_succeed_if.so uid < 1000 quiet
account    [default=bad success=ok user_unknown=ignore] pam_ldap.so
account    required      pam_permit.so
```

/etc/pam.d/common-auth:

```
auth    sufficient      pam_unix.so nullok_secure
auth    requisite       pam_succeed_if.so uid >= 1000 quiet
auth    sufficient      pam_ldap.so use_first_pass
auth    required        pam_deny.so
```

/etc/pam.d/common-password:

```
password    sufficient    pam_unix.so md5 obscure min=4 max=8 nullok try_first_pass
password    sufficient    pam_ldap.so
password    required      pam_deny.so
```

/etc/pam.d/common-session**:**

```
session    required      pam_limits.so
session    required      pam_unix.so
session    optional      pam_ldap.so
session    required      pam_mkhomedir.so skel=/etc/skel umask=0022
```

/etc/ldap/ldap.conf:

```
# Your LDAP server. Must be resolvable without using LDAP.
uri ldap://10.80.89.149

# The distinguished name of the search base.
base dc=dtldap
```

/etc/pam_ldap.conf:

```
uri ldap://10.80.89.149
base dc=dtldap
ldap_version 3
```

20. Configurar accés ssh per a usuaris administradors, editar /etc/ssh/sshd_config i afegir la següent configuració :

```
PermitRootLogin no
AllowGroups root sysadmin srv_info_hpcdt
```

20. Afegir el grup d'administradors al fitxers de sudo:

```
srv_info_hpcdt          ALL=(ALL:ALL)      ALL
```

21. Atura màquina virtual:

```
$ shutdown -h now
```

Annex VII. Clonar màquina virtual mitjançant el template.

1. Crear nou volum de 5 GB per a cada màquina virtual:

```
$ lvcreate -L 5G -n deploy vg
  Logical volume "deploy" created

$ lvcreate -L 5G -n login vg
  Logical volume "login" created

$ lvcreate -L 5G -n monitor vg
  Logical volume "monitor" created

$ lvcreate -L 5G -n proxy vg
  Logical volume "proxy" created

$ lvcreate -L 5G -n sgemaster vg
  Logical volume "sgemaster" created
```

2. Verificar els volums creats:

```
$ lvscan
  ACTIVE            '/dev/vg/template-disk' [5.00 GiB] inherit
  ACTIVE            '/dev/vg/deploy' [5.00 GiB] inherit
  ACTIVE            '/dev/vg/login' [5.00 GiB] inherit
  ACTIVE            '/dev/vg/monitor' [5.00 GiB] inherit
  ACTIVE            '/dev/vg/proxy' [5.00 GiB] inherit
  ACTIVE            '/dev/vg/sgemaster' [5.00 GiB] inherit
```

3. Fer una copia del sistema de fitxers del template (exemple deploy):

```
$ dd bs=64k if=/dev/vg/template-disk of=/dev/vg/deploy
  81920+0 records in
  81920+0 records out
  5368709120 bytes (5.4 GB) copied, 135.561 s, 39.6 MB/s
```

4. Copiar fitxer de configuració del template amb el nom de la màquina virtual a crear:

```
$ cp -rp /etc/xen/template.cfg /etc/xen/deploy.cfg
```

5. Editar i modificar els següents paràmetres marcats en negreta (vcpu, memory, disk, name i vif):

```
#  Kernel + memory size
#
kernel      = '/boot/vmlinuz-3.2.0-4-amd64'
ramdisk     = '/boot/initrd.img-3.2.0-4-amd64'

vcpus       = '1'
memory      = '512'

#  Disk device(s).
#
root        = '/dev/xvda1 ro'
disk        = [
                'phy:/dev/vg/deploy,xvda1,w',
```

```
          ]

#  Physical volumes
#
#  Hostname
#
name       = 'deploy'

#  Networking
#
vif        = [ 'mac=00:16:3E:97:1D:1A,bridge=xenbr0' ]

#  Behaviour
#
on_poweroff = 'destroy'
on_reboot   = 'restart'
on_crash    = 'restart'
```

4. Arrencar la màquina virtual:

```
$ xm create /etc/xen/deploy.cfg
  Using config file "/etc/xen/deploy.cfg".
  Started domain deploy (id=6)
```

5. Agafar la consola i modificar el fitxer d'interfaces si escau (en aquest cas només és disposa d'una targeta de xarxa eth0):

```
$ xm console deploy
```

Annex VIII. Instal·lació màquina virtual Login.

Recursos de xarxa NFS:

1) Muntar unitats de xarxa NFS de la cabina de discos, editar el fitxer fstab:

/etc/fstab

```
10.40.120.21:/hpc_homes  /homedtic  nfs  defaults   0       0
10.40.120.21:/hpc_sge    /sge       nfs  defaults   0       0
10.40.120.21:/hpc_soft   /soft      nfs  defaults   0       0
```

2) Muntar unitats:

```
$ mount -a
```

Instal·lar eines SGE:

Com el gestor de cues està instal·lat en un directori compartit només cal activar les variables d'entorn dins del perfil d'usuari per afegir els binaris a l'entorn d'usuari.

1) Copiar Script per carregar les variables d'entorn:

```
$ sudo cp -rp /sge/default/common/settings. /etc/profile.d/gridengine.csh
$ sudo cp -rp /sge/default/common/settings.sh /etc/profile.d/gridengine.sh
```

Un cop inciciada la sessió d'usuari, aquest trobarà al seu entorn les següents variables:

```
MANPATH=/sge/man:/usr/man
LD_LIBRARY_PATH=/sge/lib/linux-x64
PATH=/homes/jortega/bin:/sge/bin/linux-x64:/usr/local/bin:/usr/bin:/bin:/soft/MATLAB/R2013b/bin
SGE_ROOT=/sge
```

Annex IX. Instal·lació màquina virtual Proxy.

1. Instal·lar Squid com proxy dels nodes:

```
$ sudo apt-get install squid
```

2. Modificar fitxer /etc/squid/squid.conf per reconfigurar el port d'escolta del proxy:

```
http_port 192.168.7.10:8080
```

3. Restringir l'accés del proxy a la xarxa de gestió, crear el fitxer /etc/squid/allowed

```
192.168.0.0/24
```

4. Crear la ACL[99] per la llista de permesos:

```
acl allowed  src "/etc/squid/allowed"
```

5. Assignar acl als protocols permesos:

```
http_access allow allowed  allowsites
```

6. Reiniciar servei:

```
$ /etc/init.d/squid restart
```

[99] Una ACL especifica quins usuaris o processos del sistema es concedeix l'accés als objectes, així com les operacions que es permeten en aquests.

Annex X. Instal·lació màquina virtual Deploy.

1) Instal·lar apt-cacher:

```
$ apt-get install apt-cacher-ng
```

2) Editar el fitxer de configuració i configurar el port i el servidor proxy per sortir a internet:

/etc/apt-cacher-ng/acng.conf

```
Port:9999
BindAddress: deploy
Proxy: http://proxy:3128
```

3) Iniciar servei apt-cacher:

```
$ /etc/init.d/apt-cacher-ng start
[ ok ] Starting apt-cacher-ng: apt-cacher-ng.
```

4) Verificar que el port del apt-cacher està escoltant:

```
$ netstat -ant | grep 9999
tcp        0      0 192.168.0.15:9999       0.0.0.0:*               LISTEN
```

5) Instal·lar servidor de configuració d'ips automàtiques:

```
$ apt-get install isc-dhcp-server isc-dhcp-client
```

6) Configurar servidor DHCP i predefinir les ip's dels nodes:

```
deny unknown-clients;
option dhcp-max-message-size 2048;
use-host-decl-names on;

subnet 192.168.0.0 netmask 255.255.255.0 {
   option routers 192.168.0.251;
   option domain-name "localnet";
   option domain-name-servers 10.80.89.149;
   option time-servers proxy;
   option ntp-servers proxy;
   server-name deploy;
   next-server deploy;
   filename "fai/pxelinux.0";
}

host node01 {
            hardware ethernet 00:26:55:DA:8A:F4;
            fixed-address 192.168.0.101;
}
...
```

7) Iniciar servei dhcpd:

```
$ /etc/init.d/isc-dhcp-server start
[ ok ] Starting ISC DHCP server: dhcpd.
```

8) Instal·lar clau dels repositoris de FAI:

```
$ wget -O - http://fai-project.org/download/074BCDE4.asc | sudo apt-key add –
OK
```

9) Instal·lar servidor TFTP:

```
$ apt-get tftpd-hpa
```

10) Afegir servei TFTP a la configuració inet.d:

/etc/inetd.conf

```
tftp dgram udp wait root /usr/sbin/in.tftpd /usr/sbin/in.tftpd -s /srv/tftp
```

11) Iniciar servei TFTP:

```
$ /etc/init.d/tftpd-hpa start
[....] Starting HPA's tftpd: in.tftpdroot
```

12) Verificar funcionament servidor TFTP:

```
$ netstat -antu | grep 69
udp        0      0 0.0.0.0:69              0.0.0.0:*
```

13) Instal·lar servidor FAI:

```
$ apt-get install fai-client fai-doc fai-quickstart fai-server
```

14) Configurar variables per als logs del servidor:

```
LOGUSER=fai
LOGMETOD=ssh
```

15) Configurar variables per fer les instal·lacions mitjançant NFSROOT:

/etc/fai/nfsroot.conf

```
FAI_DEBOOTSTRAP="wheezy http://deploy:9999/debian"
FAI_ROOTPW='$1$kkklWcO.E$djxBghU7dMkrltJHggf6d1'
NFSROOT_ETC_HOSTS="192.168.0.15 deploy"
NFSROOT=/srv/fai/nfsroot
TFTPROOT=/srv/tftp/fai
NFSROOT_HOOKS=/etc/fai/nfsroot-hooks/
FAI_DEBOOTSTRAP_OPTS="--exclude=info"
FAI_CONFIGDIR=/srv/fai/config
```

16) Configurar el sistema base que s'instal·larà als nodes, aquest és descarregarà dels repositoris oficials i s'instal·larà a NFSROOT:

```
$ fai-setup -vf
Using configuration files from /etc/fai
Creating FAI nfsroot in /srv/fai/nfsroot
Creating base system using debootstrap version 1.0.48+deb7u1
```

```
Calling debootstrap --exclude=info wheezy /srv/fai/nfsroot http://deploy:9999/debian
I: Retrieving Release
I: Retrieving Release.gpg
I: Checking Release signature
I: Valid Release signature (key id ED6D65271AACF0FF15D123036FB2A1C265FFB764)
I: Retrieving Packages
I: Validating Packages
I: Resolving dependencies of required packages...
I: Resolving dependencies of base packages...
I: Found additional required dependencies: insserv libbz2-1.0 libdb5.1 libsemanage-common libsemanage1 libslang2
libustr-1.0-1
I: Found additional base dependencies: libept1.4.12 libgcrypt11 libgnutls26 libgpg-error0 libidn11 libnfnetlink0 libp11-kit0
libsqlite3-0 libtasn1-3 libxapian22
I: Checking component main on http://deploy:9999/debian...
I: Retrieving libacl1
...
I: Base system installed successfully.
Creating base.tar.xz
ainsl: appending to /srv/fai/nfsroot/etc/hosts: 192.168.0.15 deploy
ainsl: appending to /srv/fai/nfsroot/etc/hosts: 192.168.0.15    deploy
`/etc/resolv.conf' -> `/srv/fai/nfsroot/etc/resolv.conf-installserver'
`/etc/resolv.conf' -> `/srv/fai/nfsroot/etc/resolv.conf'
Upgrading /srv/fai/nfsroot
...
install_packages: executing chroot /srv/fai/nfsroot apt-get clean
install_packages: executing chroot /srv/fai/nfsroot dpkg --configure --pending
install_packages: executing chroot /srv/fai/nfsroot dpkg -C
install_packages: executing chroot /srv/fai/nfsroot apt-get clean
install_packages exit code: 0
`/srv/fai/nfsroot/boot/vmlinuz-3.2.0-4-amd64' -> `/srv/tftp/fai/vmlinuz-3.2.0-4-amd64'
`/srv/fai/nfsroot/boot/initrd.img-3.2.0-4-amd64' -> `/srv/tftp/fai/initrd.img-3.2.0-4-amd64'
TFTP environment prepared. Enable DHCP and start the TFTP daemon on root /srv/tftp/fai.
FAI packages inside the nfsroot:
fai-client        4.0.8~deb7u1
fai-nfsroot       4.0.8~deb7u1
fai-setup-storage  4.0.8~deb7u1
FAI related packages inside the nfsroot:
dracut            020-2
dracut-network    020-2
Waiting for background jobs to finish
[1]   Done                    nice rm -rf $deldir/../.will-now-be-deleted  (wd: /srv/tftp)
[2]+  Done                    nice xz -q $NFSROOT/var/tmp/base.tar  (wd: /srv/fai/nfsroot)
fai-make-nfsroot finished properly.
Log file written to /var/log/fai/fai-make-nfsroot.log
Re-exporting directories for NFS kernel daemon....

   You have no FAI configuration space yet. Copy the simple examples with:
   cp -a /usr/share/doc/fai-doc/examples/simple/* /srv/fai/config
   Then change the configuration files to meet your local needs.
Please don't forget to fill out the FAI questionnaire after you've finished your project with FAI.

FAI setup finished.
Log file written to /var/log/fai/fai-setup.log
```

17) Instal·lar servidor NFS:

```
$ apt-get install nfs-kernel-server
```

18) Afegir la configuració de FAI i el NFSROOT amb la imatge de debian al fitxer exports:

/etc/exports

```
/srv/fai/config  192.168.0.0/24(async,ro,no_subtree_check)
/srv/fai/nfsroot 192.168.0.0/24(async,ro,no_subtree_check,no_root_squash)
```

19) Muntar nfsroot amb chroot per modificar configuracions:

```
$ chroot /srv/fai/nfsroot/
/
```

20) [en mode chroot] Canviar la contrasenya de root:

```
$ passwd
Enter new UNIX password:
Retype new UNIX password:
passwd: password updated successfully
```

21) [en mode chroot] Verificar que sortida a internet fent una actualització de la base de dades apt:

```
$ apt-get update
Get:1 http://fai-project.org wheezy Release.gpg [836 B]
Get:2 http://fai-project.org wheezy Release [5003 B]
Get:3 http://fai-project.org wheezy/koeln amd64 Packages [6878 B]
Get:4 http://ftp.us.debian.org wheezy Release.gpg [1672 B]
...
```

22) Sortir del chroot:

```
$ exit
exit
```

23) Configurar plantilla per a que els clients puguin arrencar per PXE:

/srv/tftp/fai/pxelinux.cfg/pxe.tmpl

```
default fai-generated

label fai-generated
kernel vmlinuz-3.2.0-4-amd64
append initrd=initrd.img-3.2.0-4-amd64 ip=eth0:dhcp root=/dev/nfs nfsroot=192.168.0.15:/srv/fai/nfsroot:vers=3 aufs
monserver=deploy LOGSERVER=deploy FAI_FLAGS=verbose,sshd,createvt,reboot
FAI_CONFIG_SRC=nfs://192.168.0.15/srv/fai/config FAI_ACTION=install
```

24) Crear els fitxers d'instal·lació de cada client a partir del template de PXE:

```
$ fai-chboot -c pxe.tmpl  node01
$ fai-chboot -c pxe.tmpl  node02
$ fai-chboot -c pxe.tmpl  node03
...
```

25) Crear el fitxer PXE per que els nodes s'inicïin directament del disk (post-install):

```
$ fai-chboot -o default
```

26) Els fitxers creats als punts 24 i 25 és troben a:

/srv/tftp/fai/pxelinux.cfg/

```
-rw-r--r-- 1 root root  462 May 10 01:43 C0A80065
-rw-r--r-- 1 root root  457 May 12 17:51 C0A80066
-rw-r--r-- 1 root root  457 May 12 17:52 C0A80067
...
```

NOTA: En aquest punt és pot fer un incís i personalitzar la instal·lació dels node de càlcul indicada a l'Annex XIII

27) Configurar els nodes a la BIOS per a que s'iniciïn directament per la primera interfície de xarxa mitjançant PXE.

28) Arrencar el node i verificar que agafa el corresponent fitxer de configuració del servidor TFTP, en el cas del node01. El nom del fitxer correspon amb l'adreça hexadecimal de la seva ip de gestió:

/srv/tftp/fai/pxelinux.cfg/**C0A80065**

Hexadecimal	**Decimal**
C0	192
A8	168
00	0
65	101

29) Un cop finalitzada la instal·lació dels nodes aquests és renombren automàticament i és desactiven. Si el PXE de cada node no troba el fitxer arrecarà amb el disc dur loca (fitxer default del TFTP server):

/srv/tftp/fai/pxelinux.cfg/

```
-rw-r--r-- 1 root root  462 May 10 01:43 C0A80065.disable
-rw-r--r-- 1 root root  457 May 12 17:51 C0A80066.disable
-rw-r--r-- 1 root root  457 May 12 17:52 C0A80067.disable
...
```

30) Si és vol reinstal·lar un node, s'ha de executar a següent comanda:

```
$ fai-chboot –e node01
```

31) Per fer la operació inversa al punt 39 i desactivar una instal·lació:

```
$ fai-chboot –d node01
```

Annex XI. Instal·lació màquina virtual GE Màster.

Instal·lació del gestor de cues:

1) Descarregar Open Grid Scheduler de:

```
http://dl.dropbox.com/u/47200624/respin/ge2011.11.tar.gz
```

2) Descomprimir fitxer i copiar contingut al directori compartit /sge

3) Crear usuari i grup sgeadmin (aquest usuari ID ha d'existir al servidor NFS):

```
groupadd -g 500 sgeadmin
useradd -m -d /sge/ -s /bin/bash -u 500  -g sgeadmin -c "SGE Admin User" sgeadmin
```

4) Instal·lar qmaster amb l'script i seguir el procés d'instal·lació:

```
/sge/install_qmaster
```

4.1) Welcome to the Grid Engine installation, Grid Engine qmaster host installation:

```
The qmaster installation procedure will take approximately 5-10 minutes.
Hit <RETURN> to continue >>
```

4.2) Grid Engine admin user account

```
The current directory
   /sge
is owned by user
   sgeadmin
```

4.3) Installing Grid Engine as admin user >sgeadmin<

4.4) Checking $SGE_ROOT directory

4.5) Grid Engine TCP/IP communication service

```
sge_qmaster service set to port 6444
```

4.6) Grid Engine TCP/IP communication service

```
sge_execd service set to port 6445
```

4.7) Grid Engine TCP/IP communication service

```
sge_execd
```

4.8) Grid Engine cells

```
Enter cell name [default] >>
```

4.9) Unique cluster name

```
Enter new cluster name or hit <RETURN>
to use default [p6444] >>
```

4.10) Grid Engine qmaster spool directory

```
Enter a qmaster spool directory [/sge/default/spool/qmaster] >>
```

4.11) Windows Execution Host Support

```
Are you going to install Windows Execution Hosts? (y/n) [n] >>
```

4.12) Verifying and setting file permissions

```
Did you install this version with >pkgadd< or did you already verify
and set the file permissions of your distribution (enter: y) (y/n) [y] >>
```

4.13) Default domain for hostnames

```
Do you want to configure a default domain (y/n) [y] >>
```

4.14) Grid Engine JMX MBean server

```
Do you want to enable the JMX MBean server (y/n) [n] >>
```

4.15) Setup spooling

```
Please choose a spooling method (berkeleydb|classic) [berkeleydb] >>
```

4.16) Berkeley Database spooling parameters

```
Default: [/sge/default/spool/spooldb] >>
```

4.17) Grid Engine group id range

```
Please enter a range [20000-20100] >>
```

4.18) Grid Engine cluster configuration

```
Default: [/sge/default/spool] >>
```

4.19) The following parameters for the cluster configuration were configured:

```
   execd_spool_dir      /sge/default/spool
   administrator_mail   none
```

4.20) Creating local configuration

```
Creating >act_qmaster< file
Adding default complex attributes
```

```
Adding default parallel environments (PE)
Adding SGE default usersets
Adding >sge_aliases< path aliases file
Adding >qtask< qtcsh sample default request file
Adding >sge_request< default submit options file
Creating >sgemaster< script
Creating >sgeexecd< script
Creating settings files for >.profile/.cshrc<
```

4.21) qmaster startup script

```
cp   /sge/default/common/sgemaster /etc/init.d/sgemaster.p6444   /sbin/insserv /etc/init.d/sgemaster.p6444
```

4.22) Grid Engine qmaster startup

```
Starting qmaster daemon. Please wait ...
```

4.23) Adding Grid Engine hosts

```
Do you want to use a file which contains the list of hosts (y/n) [n] >>
```

4.24) Adding admin and submit hosts

```
Host(s): login node01 node02 node03 node04 node05 node06 node07 node08 node09 node10 node11
```

4.25) If you want to use a shadow host, it is recommended to add this host to the list of administrative hosts.

```
Do you want to add your shadow host(s) now? (y/n) [n] >>
```

4.26) Creating the default <all.q> queue and <allhosts> hostgroup

```
root@sgemaster added "@allhosts" to host group list
root@sgemaster added "all.q" to cluster queue list
```

4.27) Scheduler Tuning

```
Configurations

1) Normal
      Fixed interval scheduling, report limited scheduling information,
      actual + assumed load
2) High
      Fixed interval scheduling, report limited scheduling information,
      actual load

3) Max
      Immediate Scheduling, report no scheduling information,
      actual load

Enter the number of your preferred configuration and hit <RETURN>!
Default configuration is [1] >>
```

4.28) Using Grid Engine

```
Hit <RETURN> to see where Grid Engine logs messages >>
```

4.29) Grid Engine messages

```
Grid Engine messages can be found at:

  Qmaster:     /sge/default/spool/qmaster/messages
  Exec daemon: <execd_spool_dir>/<hostname>/messages

Grid Engine startup scripts can be found at:

   /sge/default/common/sgemaster (qmaster)
   /sge/default/common/sgeexecd (execd)
```

4.30) Your Grid Engine qmaster installation is now completed

```
You may verify your administrative hosts with the command
   # qconf -sh
and you may add new administrative hosts with the command
   # qconf -ah <hostname>
```

4.31) Copiar Script per carregar les variables d'entorn a tots els nodes de càlcul i al login

```
$ sudo cp -rp /sge/default/common/settings. /etc/profile.d/gridengine.csh
$ sudo cp -rp /sge/default/common/settings.sh /etc/profile.d/gridengine.sh
```

Configuració del gestor de cues:

Un cop instal·lat el gestor de cues és procedeix a verificar les configuracions més bàsiques per tal executar una tasca. (En negreta és troben les comandes a executar).

1) Verificar llistat de hosts de la llista @allhosts

```
$ qconf -shgrp "@allhosts"
group_name @allhosts
hostlist node01 node02 node03 node04 node05 node06 node07 node08 node09 node10 node11
```

2) Verificar hosts que admeten tasques per a la seva execució:

```
$ qconf -sel
node01
node02
node03
node04
node05
node06
node07
node08
node09
node10
node11
```

3) Verificar els hosts des d'on és pot administrar el clúster:

```
$ qconf -sh
login
sgemaster
node01
node02
node03
node04
node05
node06
node07
node08
node09
node10
node11
```

4) Verificar els nodes que poden executar tasques al gestor de cues:

```
$ qconf -sh
login
node01
node02
node03
node04
node05
node06
node07
node08
node09
node10
node11
```

5) Afegir grup ldap 'info_users' a la llista d'usuaris info_users:

```
$ qconf -su info_users
name    info_users
type    ACL
fshare  0
oticket 0
entries @info_users
```

6) Modificar la cua per defecte all.q i afegir els següents paràmetres, com que és tracta d'una cua per executar tot tipus de tasques no és definirà cap tipus de restricció, només per ús administratiu.

```
$ qconf -sq all.q
qname               all.q
hostlist            @allhosts
qtype               BATCH INTERACTIVE
...
slots               1, [@abudhabi=64]
tmpdir              /scratch
shell               /bin/bash
...
user_lists          info_users
...
```

Execució d'una tasca de prova:

La tasca de prova executarà la comanda sleep a un dels nodes de càlcul que el gestor de cues trobi convenient:

1) Crear script:

./Scripts/sleep.sh

```
#!/bin/bash

sleep 330
```

2) Executar tasca amb la comanda qsub:

```
$ qsub Scripts/sleep.sh
Your job 32 ("sleep.sh") has been submitted
```

3) Verificar estat, per defecte a l'inici la tasca està 'w' esperant a la disponibilitat d'un node:

```
$ qstat
job-ID  prior   name       user         state submit/start at     queue                          slots ja-task-ID
-----------------------------------------------------------------------------------------------------------------
 32     0.00000 sleep.sh   jortega      qw    05/19/2014 10:43:41                                 1
```

4) Una vegada el gestor de cues assigna la tasca a un node, l'estat canvia a **'r'**

```
$ qstat
job-ID  prior   name       user         state submit/start at     queue                          slots ja-task-ID
-----------------------------------------------------------------------------------------------------------------
 32     0.87500 sleep.sh   jortega      r     05/19/2014 10:43:54 all.q@node03                    1
```

5) És verifica l'estat del procés a nivell de sistema, cal connectar-se al node03 per ssh:

```
$ ps -ef | grep jortega
jortega  25636 25635  0 10:43 ?        00:00:00 -bash /sge/default/spool/node03/job_scripts/32
jortega  25668 25636  0 10:43 ?        00:00:00 sleep 330
```

Annex XII. Instal·lació màquina virtual Monitor.

Instal·lació PHPQstat:

Aquest eina s'ha d'instal·lar a un servidor que amb accés a les eines de gestió del clúster, pot ser en el node 'deploy' o com és el cas 'login'

1) Descarregar PHPQstat de sourceforge http://sourceforge.net/projects/phpqstat/

2) Descomprimir fitxer:

```
tar zxvf phpqstat-0.2.0a.tar.gz
```

3) Renombrar directori:

```
mv HPCKP-PHPQstat-0203739/ PHPQstat
```

4) Moure directori al públic d'apache:

```
mv PHPQstat/ /var/www/
```

5) Configurar variables d'instal·lació SGE:

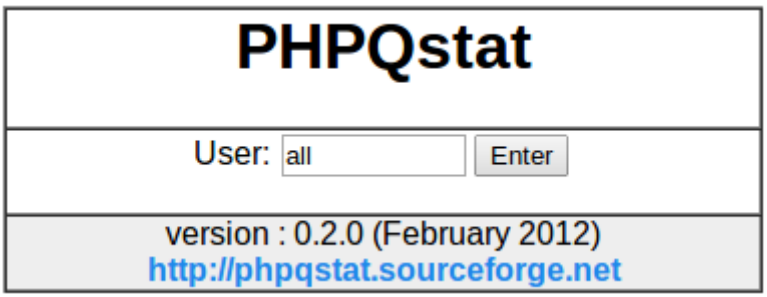
/var/www/PHPQstat/phpqstat.conf

```
SGE_ROOT=/sge
RRD_ROOT=/var/www/PHPQstat/rrd
WEB_ROOT=/var/www/PHPQstat
```

6) Afegir crontab:

```
*/3 * * * * /var/www/PHPQstat/accounting.sh > /dev/null 2>&1
```

L'eina web és molt simple, amb el nom d'usuari és pot visualitzar els processos en execució:



Instal·lació Ganglia:

1) Instal·lar els següents paquets al servidor monitor:

```
$ sudo apt-get install ganglia-monitor rrdtool gmetad ganglia-webfrontend
```

2) Configurar el servidor Master:

2.1) Copiar la configuració per al servidor web apache:

```
$ sudo cp -rp /etc/ganglia-webfrontend/apache.conf to /etc/apache2/sites-enabled/
```

2.2) Editar el fitxer /etc/ganglia/gmetad.conf i modificar la següent línia:

```
data_source "HPC Cluster" 172.16.0.2:8649
```

2.3) Editar el fitxer /etc/ganglia/gmond.conf i canviar/verificar els paràmetres en negreta:

```
cluster {
  name = "HPC Cluster"  ## Name assigned to the client groups
  owner = "TFG"
  latlong = "unspecified"
  url = "unspecified"
}

udp_send_channel   {
#mcast_join = 239.2.11.71 ## Comment
  host = 192.168.0.101   ## Master node IP address
  port = 8649
  ttl = 1
}

udp_recv_channel {
  port = 8649
}

tcp_accept_channel {
  port = 8649
}
```

Aquests canvis faran que el servidor monitor recolli dades per la IP i Port del datasource del 2.2.

2.4) Reiniciar serveis per aplicar canvis:

```
$ sudo /etc/init.d/ganglia-monitor start
$ sudo /etc/init.d/gmetad start
$ sudo /etc/init.d/apache2 restart
```

3) configurar els clients:

3.1) Instal·lar el paquet client:

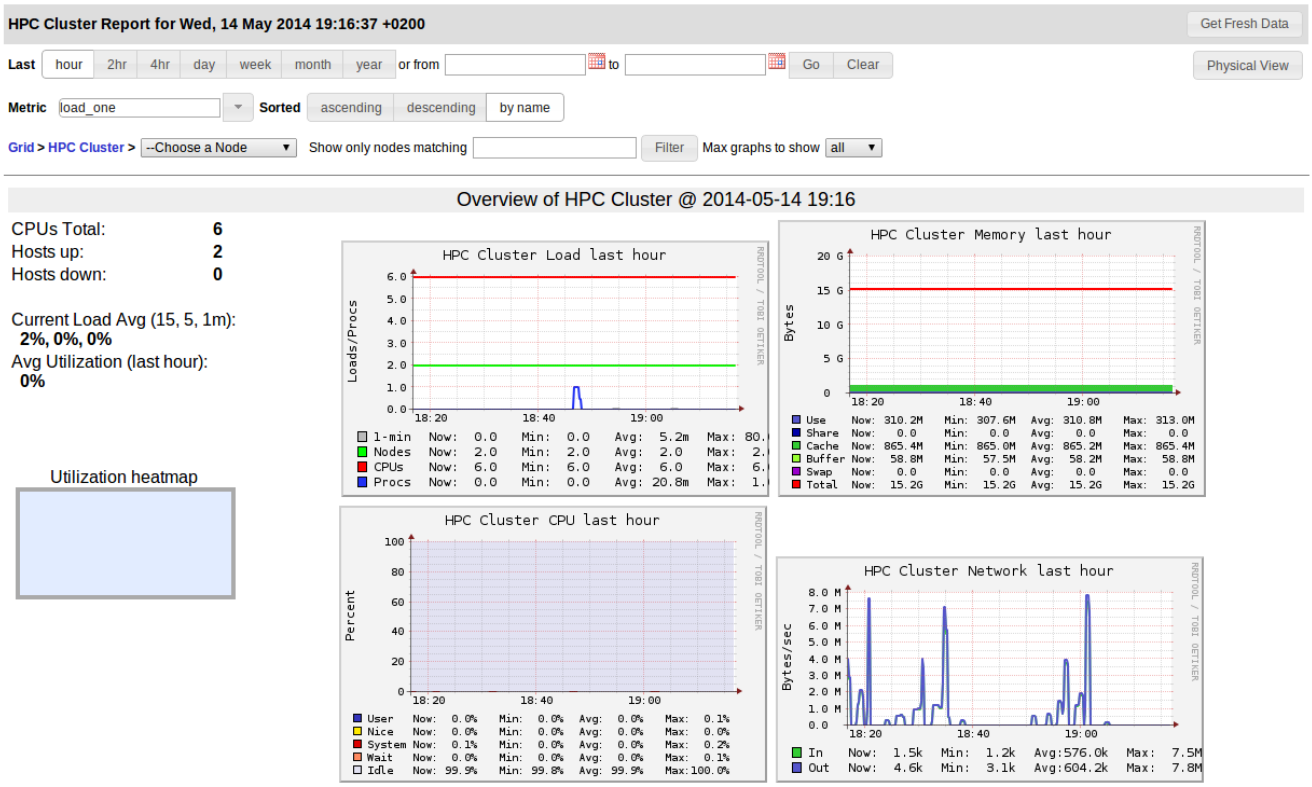
```
$ sudo apt-get install ganglia-monitor
```

3.2) Fer el pas 2.3 d'aquest Annex.

3.3) Reiniciar servei per aplicar canvis:

```
$ sudo /etc/init.d/ganglia-monitor start
```

Finalment, la interfície web de ganglia és visualitza amb l'activitat dels nodes que conformen els clústers.



NOTA: Per tal d'automatitzar la instal·lació de l'agent de ganglia als node de còmput, és poden afegir les configuracions al servidor d'instal·lacions FAI (Mirar el punt 4 de l'Annex XIII)

Annex XIII. Configuració FAI del perfil d'instal·lació dels nodes.

NOTA: Prèviament s'han de fer els passos de l'Annex X.

Les següents configuracions s'executaran i modificaran els fitxers i configuracions dels clients durant la instal·lació del node:

1) [SCRIPT] Afegir script de configuració de les interfícies de xarxa (s'ha de tenir en compte que s'ha de configurar el nom al dhcp):

/srv/fai/config/scripts/DEFAULT/15-networking

```
#! /bin/bash

error=0; trap 'error=$(($?>$error?$?:$error))' ERR # save maximum error code

#Install script
#exit in case of "softupdate"

if [ $FAI_ACTION == "softupdate" ]; then
        exit $error
fi

# Configurar la xarxa

cat <<EOF >> /target/etc/network/interfaces
# The primary network interface
auto eth0
iface eth0 inet dhcp

auto eth1
iface eth1 inet static
    address 192.168.4.1`echo $HOSTNAME |cut -c 5,6`
    netmask 255.255.255.0
    network 192.168.4.0
    broadcast 192.168.4.255

auto eth2
iface eth2 inet static
    address 10.40.120.1`echo $HOSTNAME |cut -c 5,6`
    netmask 255.255.255.0
    network 10.40.120.0
    broadcast 10.40.120.255

EOF

exit $error
```

2) [SCRIPT] Afegir script per als punts de muntatge NFS de la cabina de discos:

/srv/fai/config/scripts/DEFAULT/15-nfsmounts

```
ainsl -v /etc/fstab "10.40.120.21:/hpc_homes  /homedtic  nfs  defaults   0      0"
ainsl -v /etc/fstab "10.40.120.21:/hpc_sge    /sge   nfs  defaults   0      0"
ainsl -v /etc/fstab "10.40.120.21:/hpc_soft   /soft  nfs  defaults   0      0"
```

3) [SCRIPT] Afegir script per la configuració Open Grid Scheduler:

```
#! /bin/bash
fcopy -iMBv /etc/init.d/sgeexecd.p6444
chmod -c 755 ${target}/etc/init.d/sgeexecd.p6444
$ROOTCMD /sbin/insserv /etc/init.d/sgeexecd.p6444
$ROOTCMD ln -s /sge/default/common/settings.csh /etc/profile.d/gridengine.csh
$ROOTCMD ln -s /sge/default/common/settings.sh /etc/profile.d/gridengine.sh

ainsl -v /etc/passwd "sgeadmin:x:500:500::/home/sgeadmin:/bin/sh"
ainsl -v /etc/group "sgeadmin:x:500:"
$ROOTCMD usermod -p $1$DnLTfNL1$8jqwqorQGqoiwD9tTRfPp/ sgeadmin

exit $error
```

4) [SCRIPT] Afegir script ganglia per afegir la seva configuració:

/srv/fai/config/scripts/DEFAULT/20-ganglia

```
#! /bin/bash

mkdir $target/usr/local/lib64
tar -xvf /var/lib/fai/config/files/ganglia-usr-local-lib64-files.tar -C $target/usr/local/lib64
fcopy -iMBv /etc/init.d/gmond
#fcopy -iMBv /etc/init.d/gmond.0
fcopy -iMBv /usr/local/etc/gmond.conf
fcopy -iMBv /usr/local/sbin/gmond
chmod -c 755 ${target}/etc/init.d/gmond
#chmod -c 755 ${target}/etc/init.d/gmond.0
chmod -c 755 ${target}/usr/local/sbin/gmond
$ROOTCMD /sbin/insserv /etc/init.d/gmond
#$ROOTCMD /sbin/insserv /etc/init.d/gmond.0
#$ROOTCMD /usr/sbin/update-rc.d gmond defaults
#$ROOTCMD /usr/sbin/update-rc.d gmond.0 defaults

exit $error
```

5) [SCRIPT] Afegir script per copiar les configuracions de sistema:

```
#! /bin/bash

echo $TIMEZONE    > $target/etc/timezone
cp -f /usr/share/zoneinfo/${TIMEZONE} $target/etc/localtime

$ROOTCMD sed -i 's/#LOGUSER=fai/LOGUSER=fai/g' /etc/fai/fai.conf
ainsl -av /etc/fai/fai.conf "FAI_CONFIG_SRC=$FAI_CONFIG_SRC"

# softupdates directory
mkdir -p $target/var/lib/fai/config

exit $error
```

Les següents configuracions és poden copiar directament del servidor deploy ja que son comunes per a tot el clúster:

6) [FILES] Afegir fitxer de hosts, copiar del servidor deploy i reanomenar:

```
$ cp -rp /etc/hosts /srv/fai/config/files/etc/hosts/DEFAULT
```

7) [FILES] Copiar fitxers amb la configuració del client ldap:

```
$ cp -rp /etc/pam_lap.conf /srv/fai/config/files/etc/pam_ldap.conf/DEFAULT
$ cp -rp /etc/nsswitch.conf /srv/fai/config/files/etc/nsswitch.conf/DEFAULT
$ cp -rp /etc/libnss-ldap.conf.conf /srv/fai/config/files/etc/libnss-ldap.conf/DEFAULT
$ cp -rp /etc/ldap/lap.conf /srv/fai/config/files/etc/ldap/ldap.conf/DEFAULT
```

8) [FILES] Copiar el fitxer amb la configuració del protocol de xarxa de temps:

```
$ cp -rp /etc/ntp.conf /srv/fai/config/files/etc/ntp.conf/DEFAULT
```

Als següent fitxer és troba el llistat de paquets a instal·lar al node de càlcul:

9) [PACKAGES] Configuració de paquests a instal·lar:

/srv/fai/config/package_config/DEFAULT

```
PACKAGES aptitude
initramfs-tools
linux-image-amd64
grub-pc
isc-dhcp-common
isc-dhcp-client
openssh-server
nfs-common
acpid
at
exim4
locales
locales-all
ntp
lua5.1
liblua5.1-filesystem0
liblua5.1-posix1
gcc
g++
gcc-multilib
gfortran
make
python-scipy
python3-scipy
```

És defineix el fitxer amb la configuració de particionat del sistema d'emmagatzemament:

10) [DISK] Fitxer amb la configuració de les particions del disc de cada node:

/srv/fai/config/disk_config/DEFAULT

```
#
# <type> <mountpoint> <size>   <fs type> <mount options> <misc options>

disk_config disk1 disklabel:msdos bootable:1 fstabkey:label

primary /         20G      ext3  errors=remount-ro createopts="-L /"
logical /var      4G       ext3  defaults    createopts="-L /var"
logical /tmp      1G       ext3  defaults    createopts="-L /tmp"
logical -         256M     ext4_journal -
logical /scratch  150G-    ext4:journal=/dev/sda7  rw,relatime,journal_checksum,journal_async_commit createopts="-L
/scratch"
```

10 BIBLIOGRAFIA.

- Megías Jiménez, David; Mas, Jordi; Jorba Esteve, Josep; Suppi Boldrito, Remo; Administración avanzada de GNU/Linux, 2004 [Primera edición: marzo 2004]. Disponible a: http://www.uoc.edu/masters/oficiales/img/871.pdf

- Jiménez González, Daniel; Guim, Francesc; Rodero, Iván; Arquitectura de computadors avançades, 2012 [Promera edició: febrero 2012].

- BULL CEDOC; HPC BAS4 – Administrator's Guide ,2007 [December 2007] disponible a: http://support.bull.com/documentation/byproduct/infra/sw-extremcomp/sw-extremcomp-bas4/g/86Y230ER10/86A230ER10.pdf/at_download/fil

- IBM Corporation; using Intelligent platform Management Interface (IPMI); [2008] disponible a: http://pic.dhe.ibm.com/infocenter/lnxinfo/v3r0m0/topic/liaai.ipmi/liaaiipmi_pdf.pdf

- Chris Takemura, Luke S. Crawford; The Book of Xen (A Practical Guide for the System Administrator) [2009].

- Raphaël Hertzog, Roland Mas; The Debian Administrator's Handbook [2003-2013] disponible a: http://debian-handbook.info/get/

- LeRoy D. Cressy; Iptables [2004] disponible a: http://mirror7.meh.or.id/Network%20n%20Security/IPTABLES_book.pdf

- Paul Cobbaut; Linux Servers [2014] disponible a: http://linux-training.be/files/books/LinuxSrv.pdf

10.1 Referències

- http://fai-project.org/
- http://support.netgear.com/product/GS752TXS
- http://support.netgear.com/product/GS724TS
- http://ganglia.sourceforge.net/
- http://www.planethpc.eu/index.php?option=com_content&view=article&id=48
- http://www.sgi.com/products/remarketed/servers/altixxe210240310_datasheet.pdf
- http://www.ansys.com/
- http://www.netapp.com/us/system/pdf-reader.aspx?m=netapp-fas3100-series-datasheet.pdf&cc=us
- http://sourceforge.net/projects/fura/
- http://www.netgear.co.uk/business/products/switches/stackable-smart-switches/GS752TXS.aspx#
- http://www.supermicro.com/aplus/motherboard/opteron6100/sr56x0/h8dgu-f.cfm
- http://www.amd.com/us/products/server/processors/6000-series-platform/6300/pages/6300-series-processors.aspx
- http://www.squid-cache.org/
- https://help.ubuntu.com/community/Apt-Cacher-Server
- http://sourceforge.net/projects/phpqstat/
- http://gridscheduler.sourceforge.net
- http://amd-dev.wpengine.netdna-cdn.com/wordpress/media/2012/10/linpack_wp_bd_2.pdf