Informe per al reemplaçament de la potència de càlcul de la Universitat

F. Alted* J. Andrés† J. A. López‡ R. Mayo§

13 de Maig de 1994

Resum

Aquest informe neix com a conseqüència de la necessitat exposada durant la reunió convocada pel director tècnic d'Investigació i Doctorat el 26 de Gener passat de suplir **part** de la potència que té actualment la nostra ferramenta de càlcul científic, o siga, el cluster d'estacions de treball "vents". D'aquesta manera el *cluster vents* quedaria doncs disponible per a les tasques que inicialment se li havien reservat, i.e., les de suport a docència, encara que durant les hores de no utilització docent, hauria de continuar contribuint en l'execució de certs càlculs no pesats. L'equip informàtic que aquí se suggereix quedaria reservat, doncs, als treballs de càlcul científic més exigent des del punt de vista de recursos informàtics.

Índex

*Centre de Processament de Dades
†Departament de Ciències Experimentals
‡Departament de Matemàtiques
§Departament d'Informàtica

1 Perspectiva històrica

Des del bell començament, la Universitat Jaume I ha disposat d'una sèrie de 14 màquines HP 9000/730 destinades inicialment a suport a pràctiques docents. Tot i això, aquestes màquines han estat dedicades des d'un principi fonamentalment a càlcul científic (llevat d'una de les 14 màquines què és usada des de ja fa dos anys per a pràctiques docents). Això ha estat així degut fonamentalment a 2 fets bàsics. El primer és que no mai ha existit a la Universitat una ferramenta destinada "oficialment" a tasques de càlcul científic, i com a conseqüència de lo qual, els usuaris van començar a usar aquestes màquines amb aquests propòsits mentre s'esperaven les màquines oficials. El segón factor és que la demanda de màquines per a docència no ha superat durant els primers dos anys l'exigència d'una sola màquina.

Aquesta situació ha canviat, però, radicalment des que un grup de treball de la Universitat ha demanat l'ús simultani de 10 màquines per a impartir pràctiques amb el programa de CAD EUCLID (vegeu informe : "Estudi de l'impacte de la instal.lació del paquet de CAD EUCLID al cluster *vents*", disponible al CPD). Degut a les especials característiques d'aquestes pràctiques, s'ha arribat a la conclusió de que aquestes són incompatibles amb l'execució d'una part important de treballs de càlcul científic, en concret, els que demanden més recursos per part de la màquina. Aquesta mena de treballs són fonamentals per a la tasca investigadora desenvolupada per diferents grups de treball de la Universitat.

Una mesura que pot contribuir a alleujar l'impacte de les esmentades pràctiques és el de l'adquisició per part de la Universitat, d'una facilitat de càlcul que suplisca la pèrdua de disponibilitat, per a treballs amb grans demandes de recursos, que es preveu en les màquines que fins ara s'estan usant a tal efecte. Cal insistir en què la solució que en aquest informe es presenta no preten en cap manera donar una solució completa al tema del càlcul científic a la Universitat, sinò que s'ha d'interpretar a sols com una solució **temporal** que permeteixca continuar dignament amb la tasca d'investigació actual, a l'espera d'aconseguir una solució a mitjà o llarg plaç més definitiva, i que puga satisfer **totes** les necessitats de càlcul científic de la Universitat i no com aquesta que simplement vol ser un intent de que el suport que el càlcul numèric està donant actualment a l'activitat investigadora, no pateixca, repetim una vegada més, modificacions substancials a curt término.

2 Elecció del sistema *ideal*

La nova ferramenta de càlcul científic deuria de complir la característica fonamental de proporcionar la màxima potència per al pressupost que se li pense assignar. Unes altres característiques han de ser la generalitat (ha d'adaptar-se a les necessitats de tota la comunitat científica de la nostra Universitat), i la standarització (han de ser màquines amb el sistema operatiu UNIX, per ser el sistema operatiu que nosaltres tenim actualment com a plataforma científica).

Si atenem al principi fonamental d'una bona relació potència/preu, ens adonarem que les plataformes ideals són les estacions personals de treball. Això és degut a que, encara que no siguen màquines pensades inicialment per a treballar amb molts usuaris, el seu preu relativament barat i la seua flexibilitat de configuració, fan possible la construcció de ferramentes de

càlcul fàcilment escalables (es pot passar a models superiors sense més que canviar una placa electrònica) d'una banda, i a més amb un ampli espectre de possibles usos.

Per exemple (i açò ja s'està aplicant al cluster *vents*), es pot configurar una de les màquines amb més memòria que les demés i dedicar-la a processos amb grans demandes de memòria, afegir més disc a d'altres, per a processos amb grans requeriments d'entrada/eixida, etc...

A més, amb una distribució racional dels recursos, i amb unes ferramentes adequades (com puga ser el sistema de batch NQS), es pot ajudar a l'usuari a no tenir la impressió d'enfrontar-se a una varietat inconnexa de màquines, sinò a una estructura homogènia pròpia d'una super-màquina englobant-les a totes.

Caldria advertir, però, de la irrupció relativament recent en el mercat de màquines SMP, o siga, màquines amb processadors múltiples (no es freqüent que passen dels 16 processadors, sent un número normal 4 o 6) compartint la mateixa memòria a través d'un bus d'alta velocitat, i amb un sistema operatiu preparat per a traure el màxim rendiment de tots els processadors. Encara que d'aparició recent i per tant, no se sap encara ben bé el resultat que poden donar, el que està clar a primera vista és que aquestes màquines també suposen una molt bona relació (potència/preu), i haurien d'entrar també en les consideracions sobre l'elecció.

Plataformes més grans com ara *mainframes*, no són tant flexibles i la seua relació potència/preu eix considerablement més reduïda. A més els costos d'una possible actualització a una màquina més potenta es fan molt elevats. També és prou elevat el preu de manteniment de hardware i software, i fins i tot el consum d'energia elèctrica. Com a exemple, podem esmentar el superordinador IBM 9021 del Servei d'Informàtica de la Universitat de València. Aquella màquina fou adquirida en 1991 i costà 400 milions de pessetes (realment era un *upgrade* de l'antiga màquina IBM 3090). El manteniment hardware costa anyalment 30 milions de pessetes, i 30 milions més en consum d'energia. Actualment (Febrer del 94), amb els 60 milions es podria estar comprant granjes d'estacions de treball amb una potència fins a 25 vegades superior (en el cas de càlcul científic de tipus química computacional), i tot això anyalment!.

L'altre conjunt de plataformes que ens queda és el de les màquines MPP, ço és, màquines massivament paral.leles, amb molts processadors (típicament des d'unes poques decenes fins a milers) amb memòria que sol ser distribuida o be amb un paradigma mixte entre memòria distribuida i compartida. Aquestes màquines a banda de què tenen un preu molt elevat, la seua dificultat de programació inherent (programació paralel.la amb paradigma de memòria distribuida), fa desaconsellable la seua consideració com a màquina de càlcul científic general a la nostra Universitat.

Com a conseqüència de lo esmentat anteriorment, la inversió més apropiada per a muntar una infraestructura de càlcul científic ha de ser una granja o *cluster* d'estacions de treball o màquines SMP.

3 Elecció de la granja d'estacions de treball o màquines SMP

Avui en dia, el món de la informàtica continua evolucionant de manera tan espectacular com sempre. Continuen éssent normals creiximents exponencials en la potència de les màquines (típicament, es duplica la potència de les màquines cada any), i els preus continuen baixant, esperonats sens dubte, per la competència tan forta entre les principals marques de la in-

formàtica. Recordem que aquesta competència és conseqüència en certa mesura de l'accessibilitat a una gran potència computacional que ha obert les portes de la informàtica d'altes prestacions a grups de treball amb pressupostos relativament modestos. Com es veu, això és un peix que es mossega la cua, i per tant, cal esperar sorpreses significatives relatives al preu i potència dels sistemes en els pròxims mesos. Amb tot això es pretén dir que és arriscat fer una predicció sobre quins són els models que en el moment de la nostra possible adquisició seran més competitius.

Tot i això, el què sí que es pot fer és preveure quines hauran d'ésser les característiques principals de les màquines, les quals haurien de tenir prestacions de gama alta degut a que no hem de perdre de vista que la nova facilitat de càlcul està enfocada cap a treballs amb molta exigència de recursos, recursos que no podrà continuar donant l'actual cluster que es compartirà amb usos de docència.

Per tant, les característiques principals hauran de ser : tenir una acceptable capacitat de memòria central (a partir de 128 MB), un mínim de disc local (2 GB), un bus de sistema ràpid (superior als 350 MB/s i mínim de 64 bits), un bus d'accés a disc potent (a partir de 10 MB/s i 32 bits), i finalment, si es preveu la possibilitat d'efectuar càlcul en paral·lel, un bon ample de banda en les interconnexions entre elles (a partir de 10 MB/s). Per descomptat és clar que a totes aquestes característiques s'ha d'afegir la que pot ser és la més important de totes si parlem d'aplicacions científiques, com és la velocitat de processament de números en coma flotant.

Hem de fer èmfasi en què el paràmetre de la velocitat de processament de números en coma flotant (o potència en MegaFlops) és el ha de ser el decissiu, perquè és la característica més normalment i intensament usada per les aplicacions científiques. Realment aquest paràmetre no és una cosa aïllada en un sistema de càlcul numèric. Una mesura d'aquest implica de manera més o menys directa a la *bondat* de tots els demés. Hi ha molts tests (anomenats benchmarks en l'argot informàtic) que el mesuren i hauriem d'usar-los com una guia per a determinar la nostra plataforma ideal.

A més a més, es proposa que l'elecció de la plataforma es dirigeixca cap a workstations (o màquines SMP, que no són més que workstations unides a través d'un bus d'alta velocitat per tal de compartir la memòria) d'altes prestacions, per a poder fer front a tots aquells treballs que exigeisquen més recursos hardware, i que en el nostre cluster actual, compartit entre les classes pràctiques i càlcul científic, serien complicats de ser duts a cap. Evidentment, els treballs amb menys requeriments es podrien continuar executant-se en el cluster actual, reservant la nova ferramenta per als treballs més exigents des del punt de vista de potència de càlcul.

Pero l'elecció definitiva ha de tenir, a més, molt en compte les necessitats concretes de computació de la nostra Universitat. Per tant, la millor manera de determinar la plataforma de càlcul seria agafar els programes que estan executant actualment els nostres usuaris, provarlos en diferentes plataformes, traure la relació potència/preu i amb aquesta dada en la mà, decidir.

4 Testejant el sistema *ideal*

Per a trobar la millor plataforma de càlcul científic per a la nostra Universitat, la primera solució que ens pot venir al cap és la de fer un recull de tots els programes de caire científic que els usuaris fan servir ara mateix al nostre cluster, provar-los tots en les plataformes escollides com a candidates, traure una mitja de la relació potència/preu dels programes en les diferents plataformes, per tal d'escollir finalment la plataforma que obté una mitjana millor.

Aquest mètode presenta, però, un incovenient. El fet de fer un test d'un sol programa en una sola plataforma és ja de per si una tasca molt costosa per al personal que fa les proves : conseguir plataforma a testejar, resoldre possibles problemes durant la compilació i execució del programa, fer diverses proves amb diferents inputs del programa, i finalment la, a vegades, controvertida interpretació final dels resultats. Per tant, una prova massiva de tots els programes en les distintes plataformes, pot arribar a ser una tasca inabordable a no ser que es dispose d'un equip prou complet de persones dedicades únicament i exclusiva a aquests propòsits.

La solució passa doncs per fer una selecció de les aplicacions que cada grup de calculistes considera com més representativa dels seus treballs. Val a dir que aquesta tasca de recull d'aplicacions representatives ja s'ha iniciada dins dels grups que més ús fan ara mateix de les facilitats de càlcul a la Universitat, i que en el futur es pensa fer extensiva a qualsevol altre grup o persona que es considere un usuari potencial de les facilitats futures.

I a l'igual que s'ha de fer una selecció de les aplicacions a testejar, també se n'ha de fer el mateix amb la selecció de les possibles plataformes a testejar. Ací és on han de jugar un paper decisiu els *benchmarks*, o sèrie de tests que un equip de gent elabora per a mesurar diferents aspectes de les prestacions de les màquines. Els números resultants d'aquests tests després es publiquen periòdicament per entitats que, en principi, són independents de les empreses fabricants d'ordinadors. Per tant, aquests tests possibiliten una visió general de les prestacions de cada plataforma, i extraure'n així els possibles candidats. Veure l'apendix A, per a un exemple concret d'aquestes proves fet a petició del grup de química computacional de la Universitat.

Una altre aspecte en l'elaboració de les relacions potència/preu finals haurà d'ésser el de donar *pesos* (en el sentit estadístic del terme) als resultats de les aplicacions concretes. Aquesta qüestió no és en cap manera òbvia, i sempre sol ser font de fortes discrepàncies entre els diferents grups de treball. Una primera aproximació a aquest problema pot ser atorgar a les aplicacions un pes proporcional als recursos que aquestes han consumit al llarg d'un periode llarg de temps, com puga ser, per exemple, els últims darrers 12 mesos, de tal manera que una aplicació que s'ha usat intensivament pese molt més que no una altra que se'n haja fet un ús esporàdic. En aquest tema, però, seria desitjable arribar-hi a un consens entre els representants dels diferents grups de calculistes. En particular, aquest problema es torna a tractar més a sota en el paràgraf *Clusters mixtes*.

Es podria pensar que aquest hauria de ser el procés millor en la determinació de la plataforma de càlcul adequada, però ho serà només si tenim en compte uns altres aspectes importants del tema.

5 Altres aspectes de l'elecció

A més de testejar les aplicacions en les diferentes plataformes per a escollir-ne la millor, hi ha que tenir present una sèrie d'aspectes que ens hauran de permetre de gaudir d'un aprofitament òptim de la nova ferramenta de càlcul.

5.1 Clusters mixtes

Hem parlat anteriorment sobre l'elecció del sistema *ideal*, i sobre el problema de quins tests han de pesar més en l'elecció final.Imaginem-nos, però, que l'aplicació més usada s'execute més eficientment en plataformes que són relativament cares, i que la seua relació potència/preu, encara que beneficiosa per a ella, resulte molt desavantajosa per a d'altres aplicacions no tan intensament usades, pero si importants a l'hora de la decisió final. En aquest cas, s'adquiriria una ferramenta molt apta per al tipus de càlcul més corrent, és cert, però molt ineficient per a altra mena de càlculs també importants.

És ací on ens podem plantejar la possibilitat d'adquirir el anomenats *clusters mixtes*. Per exemple, en el cas anterior, es podrien adquirir una sèrie de màquines adequades per als càlculs més usats, però es podria destinar una part del pressupost a adquirir unes altres màquines més adaptades a la part d'aplicacions que no s'adapten bé al cas anterior. D'aquesta manera s'asseguraria una inversió òptima per a les nostres necessitats.

Caldria, però, preguntar-se com afectaria l'elecció d'aquest cluster mixte a la seua eficàcia d'ús per banda de l'usuari i també de l'administrador del sistema. Per una banda, es perdria certa facilitat d'administració de les màquines al no poder compartir totes el mateix sistema operatiu. Els usuaris també notarien certes incomoditats, com per exemple, les xicotetes diferències entre els diferents *sabors* del sistema operatiu UNIX, la incompatibilitat a nivell de codi binari executable entre les màquines, etc...

Aquests inconvenients, però, es poden esmorteïr considerablement. En el cas de l'administració de les màquines, hi ha solucions adaptades per a aquesta mena de clusters mixtes, com ara compartició de discos entre màquines amb diferents sistemes operatius (NFS, Network File System), sistemes de gestió de treball en batch adaptats a sistemes mixtes (NQS, Network Queuing System), etc... En el cas de l'usuari, el fet d'enfrontar-se amb un cluster mixte tampoc no és tan problemàtic com podria paréixer en un principi; a la cap i a la fí, cada usuari té normalment uns requeriments molt concrets que li fan escollir inicialment la plataforma més adequada per a ell, la qual cosa farà que l'usuari treballe normalment en ella i ignorant, per tant, les demés. Inclús en el cas de que hajen usuaris interessats en córrer algoritmes de manera paralela entre les diferents plataformes, això no representa cap problema, perquè hi ha ferramentes d'accés públic que permeteixen el càlcul en paral·lel entre plataformes mixtes. Un exemple d'aquest software és el PVM (Parallel Virtual Machine) que actualment ja està disponible al cluster *vents*.

Considerem doncs, que la solució d'un cluster mixte, en el cas de que incremente de forma apreciable la relació potència/preu global, pot ser considerada com a perfectament possible i desitjable.

5.2 Capacitat d'interconnexió

Aquest és un aspecte molt important en una granja d'estacions de treball. Per norma general, quant més flexible i ràpida siga la interconnexió entre les diferentes màquines, més sensació d'homogeneitat, i això, a la fi, es reflecteix en un increment de la productivitat. No sols això, sinò que amb uns bons recursos d'interconnexió, s'obri el camp a la important àrea de la paralelització d'aplicacions, àrea aquesta que s'està perfilant com una important tècnica en el càlcul científic d'altes prestacions.

En un primer estadi, es podria pensar a sols en una granja d'estacions unides per un cable ethernet de 10 Mbits/s, amb la qual cosa es lograria unes interconnexions de prestacions mitjanes. Caldria pensar, però, en la possibilitat d'augmentar les prestacions d'intercomunicació. Les solucions actuals són bàsicament dues : o bé augmentar l'ample de banda de l'únic camí d'interconnexió, o bé mantenir l'ample de banda però augmentar el número de canals d'interconnexió. Un exemple del primer cas ja està tecnològicament resolt, i comercialitzat des de ja fa uns quants anys : el FDDI. El FDDI és bàsicament una solució anàlega a Ethernet, sols que amb un ample de banda 10 vegades superior (100 Mbits/s). Aquesta solució, però, és actualment prou cara. Pel contrari, hi ha un exemple del segon cas que és relativament barat, i que pot eixir prou efectiu. És el cas de la configuració en estrella de les màquines mijantçant un concentrador central, coneguda també com ethernet commutada. Amb aquesta configuració, l'ample de banda efectiu de comunicació entre les màquines augmenta gràcies a que cadascún dels nodes és com si disposara d'un segment de comunicació particular (no el comparteix com en el cas anterior) per a la seua comunicació amb els demés nodes.

Per últim, es podria esperar a la comercialització de la anomenada Ethernet ràpida, que en la pràctica resulta igual que la solució FDDI, encara que presumiblement molt més barata. La disponibilitat comercial de la Ethernet ràpida es podria produir, segons les revistes especialitzades, a principis de l'any 95.

5.3 Software

Naturalment, tanta potència de càlcul no ens val de res si no tenim les ferramentes software necessàries per a usar-la. En un camp com el del càlcul d'altes prestacions, són necessàries aquestes tres ferramentes bàsiques :

1. **Sistema operatiu :** Ve normalment en cada màquina, i s'inclou normalment en el preu.

2. **Compilador de FORTRAN i accessoris :** Aquesta peça no s'inclou normalment en el preu de la màquina, i s'ha de comprar a banda. És essencial perquè gairebé totes les aplicacions científiques estan escrites en FORTRAN. No menys importants són els accessoris com ara un depurador de codi, un preprocessador que puga millorar l'eficiència del codi i, si es considera important la possibilitat d'una futura activitat en càlcul paral·lel, una ferramenta que facilite la programació i depuració d'algoritmes paral·lels. Aquestes últimes ferramentes, encara que no molt comunes, ja estan apareixent de manera comercial, entre les que poden destacar el paquet anomenat Forge 90 de Applied Parallel Research.

3. **Sistema de gestió de treballs en batch :** Existeixen moltes ferramentes d'ús públic (i.e. gratis) a tal efecte (NQS, DQS, BQS, Condor, Vienna Queuing System, etc..) que són suficientment potents com per a fer la tasca. Tot i això, existeixen també productes comercials que permeteixen més flexibilitat en la gestió dels treballs, com ara el Load-Leveler de IBM (disponible per a diverses plataformes, no sols per a IBM).

6 Conclusions i pressupost necessari

L'adquisició d'aquestes eines de càlcul científic de la Universitat s'ha de fer tenint en compte les necessitats de la nostra comunitat científica. S'han de provar les plataformes que els benchmarks ens suggereixen com a millors amb una selecció d'aplicacions usades a la Universitat. Les plataformes més adequades es considera que són les granjes d'estacions de treball en contraposició als superordinadors. L'elecció de la plataforma no ha de ser necessàriament màquines del mateix fabricant, sinò que s'ha que tenir en compte que una plataforma mixta podria dur a una millor satisfacció de les necessitats.

Per a fer una estimació del pressupost necessari per a suplir la potència de càlcul, hem de calcular aproximadament la potència disponible amb una unitat de mesura (benchmark) que siga representativa de la potència de càlcul cientific i alhora, ampliament disponible sobretot per a màquines actuals. Considerem que un dels millors benchmarks que acompleix aquestes condicions es el Specfp92.

La potència de càlcul actual disponible a la Universitat és d'aproximadament 1400 Specfp92 (14x100=1400). Per suplir aquesta potència amb màquines actuals, podem aprofitar l'experiència recent d'adquisició de hardware de càlcul científic per part del grup de química computacional, on s'han testejat màquines de marques punteres, i es pot dir (veure appendix A) que 35 milions de pessetes seria una quantitat suficient per a la substitució d'una **part** important dels 1400 Specfp92 actuals.

Cal afegir que es podria intentar fer revertir a la Universitat el 15% d'IVA corresponent als 35 milions (uns 4,5 milions de pessetes) i aprofitar aquesta quantitat per a adquisició del software necessari per a traure el màxim rendiment al sistema (Sistema operatiu amb suficients licències, compilador de Fortran, paquets de paral.lelització de codi, sistemes de gestió de treballs en batch, etc...).

Naturalment, hem de ser conscients de que aquesta adquisició també ha de comportar aspectes addicionals com ara el manteniment de les màquines el qual es pot xifrar aproximadament en un 7% anyal sobre el cost total. També s'han de tenir en compte les despeses en la infraestructura necessària per a la instal.lació de la nova facilitat, com ara el canvi d'acomeses per a la alimentació, la possible instal.lació d'una mampara que separe les noves màquines de l'actual cluster, adequació del sistema d'aire condicionat, etc... També hem de recordar que el personal encarregat dels sistemes físics del CPD ja està suficientment saturat (cosa que s'ha fet constar repetidament) com per a fer-se càrrec d'una manera eficient de la nova facilitat, recomanant-se fortament el reforçament del personal d'una manera addient.