Academia Română
Institutul de Geodinamică *"Sabba S.Ştefănescu"*
Laboratorul Dinamica Globului Terestru
Str. Jean-Louis Calderon, Nr. 19-21, Bucureşti-37, România, R-020032,
fax:(4021) 317.2120, tel. (4021) 317.2126; e-mail: inst_geodin@geodin.ro
http://www.geodin.ro/~prezentare/

DIRECTOR,

Dr. Crişan DEMETRESCU
Membru corespondent al Academiei Române

Programul Operațional Sectorial Creşterea Competitivității Economice
Axa prioritară 2: Competitivitate prin Cercetare, Dezvoltare Tehnologică şi Inovare
Operațiunea: O.2.1.2 „Proiecte CD de înalt nivel ştiințific la care vor participa specialişti din străinătate"
Proiect: Infrastructură cibernetică pentru studii geodinamice relaționate cu zona seismogenă Vrancea: ID-593, cod SMIS-CSNR 12499
Etapa 1: Construirea si testarea sistemului hardware (HPCC, HPVC si GeoWall)
Perioada: 18 iunie 2010 - 17 iunie 2011

STUDIU

CERCETARI DE SPECIALITATE PRIVIND CONFIGURAREA OPTIMA A SISTEMELOR HPCC, HPVC SI GEOWALL

Director de proiect,

Dr. Vlad Constantin Manea

A U T O R I :

Dr. Vlad Constantin Manea

Dr. Marina Manea

Drd. Mihai Pomeran

Alexandr Hapitchii

CUPRINS:

1. INTRODUCERE

Scopul principal al proiectului CYBERDYN este construirea unei infrastructuri cibernetice in cadrul Institutului de Geodinamica al Academiei Romane, pentru studierea evolutiei geodinamice pe termen lung a zonei seismogene active Vrancea. Aceasta infrastructura cibernetica este formata dintr-un HPCC (High Performance Computing Cluster – Grup de Servere pentru Calcule de Inalta Peformanta), un HPVC (High Performance Visualization Cluster – Grup de Servere pentru Vizualizare de Inalta Performanta) si un sistem de Vizualizare Stereo in 3D (GeoWall).

Noua infrastructura cibernetica va ajuta la crearea unui corp de cercetatori format din experti cu inalta pregatire obtinuti prin antrenarea tinerilor oameni de stiinta in campul geodinamicii computationale, permitand, in felul acesta, generarea primului centru de excelenta in domeniu din Romania. Activitatea acestui centru de excelenta se va extinde si dupa finalizarea ultimei etape a proiectului prin formarea tinerilor specialisti si prin participarea in proiecte nationale/internationale bazata pe capacitatea si performanta oferite de o asemenea tehnologie.

Acest studiu are ca scop principal alegerea arhitecturii HPCC-HPVC-GeoWall adecvate studiilor de modelare numerica care intentioneaza sa descifreze originea geodinamica a zonei Vrancea. Studiul este alcatuit din urmatoarele capitole principale:

-Principii generale de realizare si proiectare a clusterelor. In acest capitol se face o evaluare generala a modului de constructie a sistemelor de calcul/vizualizare rapide tip cluster si stereo, sau HPCC/HPVC/GeoWall.

-State of the art (stadiul actual). In acest capitol vom discuta stadiul actual al celor mai rapide clustere din lume, si vom prezenta cateva exemple de sisteme HPCC din domeniul stiintelor pamantului care studiaza fenomene asemanatoare din punct de vedere geodinamic cu fenomenele din zona seismogena Vrancea.

-Cerinte tehnice (performanta). In acest capitol o sa subliniem performantele tehnice optime pe care o infrastructura HPCC-HPVC-GeoWall trebuie sa le indeplineasca in vederea studierii evolutiei geodinamice a zonei Vrancea.

-Arhitectura de Calcul/Vizualizare optima. In cadrul acestui capitol vom prezenta arhitectura optima a unui sistem HPCC-HPVC-GeoWall care sa satisfaca cerintele proiectului CyberDyn.

-Diferente intre arhitectura optima si cea propusa in cererea de finantare a proiectului CyberDyn. In final dorim sa explicam/justificam decizia de a modifica arhitectura preliminara prevazuta in cererea de finantare (CF) cu scopul principal de a imbunatati performatele intregului sistem de calcul.

-Referinte bibliografice.

2. PRINCIPII GENERALE DE REALIZARE SI PROIECTARE A CLUSTERELOR

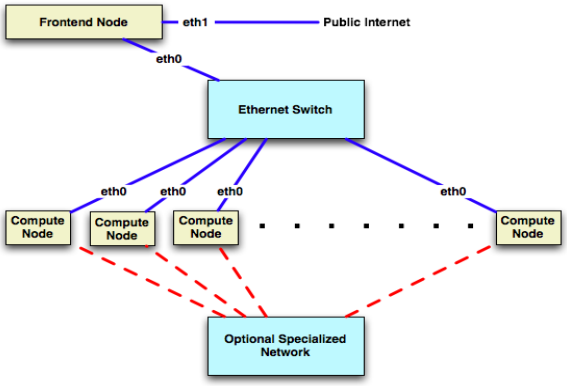
Scopul principal al proiectului CYBERDYN este construirea unei infrastructuri cibernetice in cadrul Institutului de Geodinamica al Academiei Romane, pentru studierea evolutiei geodinamice pe termen lung a zonei seismogene active Vrancea. Aceasta infrastructura cibernetica va fi formata dintr-un HPCC (High Performance Computing Cluster – Grup de Servere pentru Calcule de Inalta Peformanta), un HPVC (High Performance Visualization Cluster – Grup de Servere pentru Vizualizare de Inalta Performanta) si un sistem de Vizualizare Stereo in 3D (GeoWall).

2.1. SISTEMELE DE CALCUL TIP HPCC

In cele ce urmeaza vom prezenta o descriere, pe scurt, a partilor componente pricipale aferente sistemelor de calcul tip HPCC.

Un element fundamental din cadrul unui cluster este reprezentat de Grupul de Servere pentru Calcul de Inalta Performanta sau Sistemele de Calcul de Inalta Performanta. Acestea vor fi mentionate in continuare folosind acronimul HPCC (High Performance Computer Clusters).

Un cluster reprezinta un grup de calculatoare care lucreaza impreuna pentru realizarea unui scop final (*Lehmann, 2009; Lucke, 2005*). In general, un cluster se refera la un set de sisteme de calcul legate impreuna si care din punct de vedere fizic sunt instalate foarte aproape unele de altele, in scopul rezolvarii cu eficienta marita a diferitelor probleme de calcul. Aceste tipuri de clustere sunt cunoscute sub numele de HPCC sau simplificat clustere de calcul. Un cluster este compus din minum trei elemente de baza: o colectie de calculatoare individuale sau noduri, o retea care conecteaza aceste calculatoare si un soft care permite unui calculator sa imparta incarcatura de lucru cu celelalte calculatoare, prin intermediul retelei (*Mao, 2010*). In principiu un HPCC trebuie sa contina cel putin doua noduri, un nod principal (master node) si un nod secundar sau de calcul (slave node). Nodul principal reprezinta server-ul cu care utilizatorii interactioneaza direct si care prezinta un programator de rutine de calcul.



Schema de baza a configuratiei unui cluster de inalta performanta (sursa: *www.rocksclusters.org*).

Conform *Morton (2007)* exista o serie de beneficii in folosirea clusterelor pentru efectuarea calculelor numerice in paralel:

-Costul scazut: desi alcatuite din servere de inalta performanta legate printr-o retea specializata de mare viteza, costul unui cluster este mult mai mic in comparatie cu alte sisteme rapide de calcul, ca de exemplu masinile de tip vector.

-Scalabilitatea: clusterele asigura o posibilitate relativ usoara de marire a resurselor de calcul pe masura cresterii cerintelor de calcul si a volumului de lucrari in timp.

-Independenta fata de furnizor: chiar daca este recomandabila pe cat posibil folosirea unor componente identice/asemanatoare in cadrul serverelor ce formeaza un cluster, este posibil ca puterea de calcul sa se mareasca adaugand servere care apartin unui spectru amplu de furnizori.

-Adaptabilitatea: este relativ usoara schimbarea modului de conectare a nodurilor de calcul pentru un cluster care raspunde cel mai bine cerintelor aplicatiilor care vor fi rulate in centrul de calcul.

- Durata de Functionare: In cazul clusterelor, cedarea unui singur component poate afecta doar portiuni mici din totalul resurselor de calcul. Intr-un cluster mai mic se poate asigura intretinerea sau repararea unui component fara intreruperea intregului sistem. De asemenea, se pot adauga resurse suplimentare de calcul chiar in tipul executiei programelor fara afectarea proceselor de calcul.

Arhitectura unui cluster HPCC dicteaza viteza sa de calcul. Prin efectuarea de calcule matematice rapide, HPCC-urile si-au dovedit utilitatea in diverse domenii de activitate, cum ar fi: prezicerea climei, modelarea numerica pentru astronomie sau stiintele pamantului. Pentru a rezolva probleme cu carater stiintific (si nu numai) clusterele folosesc calculele distribuite pe mai multe calculatoare, sau noduri de calcul. In acest caz este nevoie de coduri care sa fie paralelizate, adica sa execute un set de instructiuni simultan. Astfel exista posibilitatea de a

imparti calculele intre diferite masini din retea, cu fiecare calculator efectuand o parte din calcule, fiecare lucrand in acelasi timp.

Exista mai multe tipuri de clustere in functie de aranjamentul si tipul serverelor folosite. Poate cea mai comuna arhitectura este cunoscuta sub numele de arhitectura asimetrica, foarte comuna in sistemele HPCC de dimesiuni reduse. Pentru clusterele asimetrice (Buyya, 1999), un server reprezinta nodul principal care realizeaza legatura intre restul nodurilor si utilizatori. Nodurile de calcul au de obicei sisteme de operare minimale, si sunt dedicate exclusiv pentru cluster. Dezavantajul principal al acestei arhitecturi provine din limitarile de performanta impuse de nodul principal. Din acest motiv se recomanda folosirea unui server cu putere sporita care va fi folosit pentru nodul principal. Pentru clusterele de dimensiuni medii sau mari este important sa se incorporeze servere aditionale in cluster. De exemplu, unul dintre noduri poate functiona ca un server NFS, al doilea ca o statie de administrare ce monitorizeaza starea de sanatate a clusterului, precum si o serie de servere dedicate traficului de date (I/O).

Proiectarea retelei reprezinta o problema importanta de rezolvat. Din fericire pentru clusterele mici si medii, o retea prevazuta cu un singur switch de mare viteza este suficienta (pentru procesarea si accesul serverelor I/O), si alt switch ce permite administrarea si accesul utilizatorilor printr-o retea mai lenta.

Desi clusterele au foarte mult de oferit, exista o limita cu privire la cantitatea de calculatoare pe care o adaugam pentru rezolvarea cat mai rapida a unei probleme. Intr-o situatie ideala, utilizatorul poate sa-si doreasca ca un calcul sa fie de doua ori mai rapid folosind doua masini decat daca ar folosi una. Din pacate, in realitate lucrurile nu se intampla in acest fel. Un fragment de cod poate fi paralelizat si impartit apoi intre doua sau mai multe masini, insa anumite fragmente de cod nu pot fi paralelizate si trebuiesc executate intr-o anumita ordine seriala. In plus, coordonarea comunicatiilor intre diverse procese va necesita un cod suplimentar. Aceasta se adauga la timpul total de executie.

In general, proiectarea unui cluster tip HPCC implica urmatoarele etape principale (*Lehmann, 2009*):

a) Determinarea de la inceput a scopului clusterului.

b) Selectionarea arhitecturii generale a clusterului.

c) Selectionarea sistemului de operare, a softului pentru cluster si a celorlaltor softuri de sistem care vor fi folosite.

d) Selectionarea hardware-ului pentru cluster.

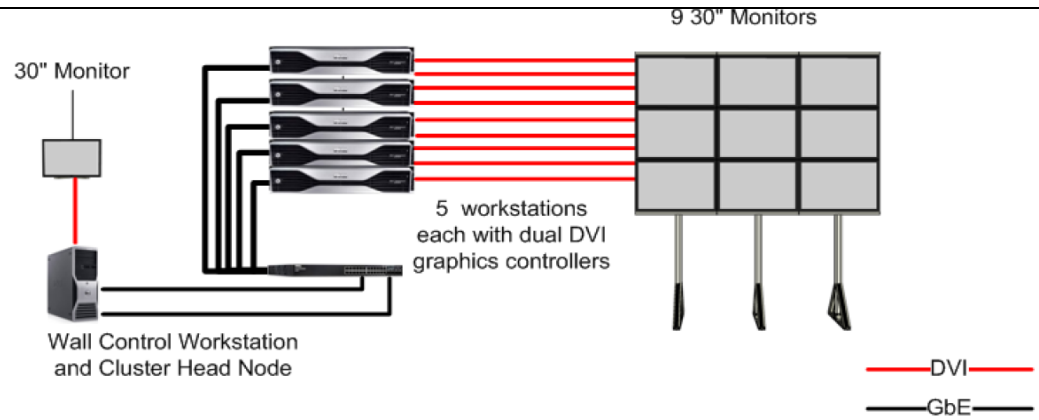
e) Proiectarea incintei (daca este cazul) unde se va instala HPCC-ul.

2.2. SISTEMELE DE VIZUALIZARE TIP HPVC

In continuare, Grupul de Servere pentru Vizualizare de Inalta Performanta sau Sistemele de Vizualizare de Inalta Performanta vor fi mentionate folosind acronimul HPVC (High Performance Visualization Clusters)

HPC devine o parte integranta a cercetarii stiintifice iar stiinta vizualizarii seturilor de date furnizate de catre HPC a devenit un camp important de studiu. Una dintre solutiile de excelenta o reprezinta HPVC-ul, care se aplica din ce in ce mai mult in diferite domenii. Adoptarea HPC-ului a permis cercetatorilor sa proceseze seturi de date multivariate masive care genereaza atat de multe date la iesire incat analiza acestora se dovedeste de cele mai multe ori imposibila. Una dintre metodele eficiente de a vizualiza acest tip de date in rezolutie completa consta in construirea unui HPVC care foloseste solutii multi-CPU, multi-GPU, multi-monitor, pentru a obtine performante de rezolutie si rendering mult peste capacitatile unui sistem simplu.

Desi exista o multitudine de posibilitati legate de modul de generare al imaginii pentru vizualizare (rendering), in continuare vom descrie pe scurt cea mai comuna metoda si tehnologie, folosite in acest scop in multe domenii de activitate stiintifica in general, precum si in domeniul stiintelor Pamantului in particular. Acest sistem este un sistem combinat de perete de vizualizare si cluster de rendering. Acesta consta dintr-un HPC standard cu adaugarea unor placi grafice si monitoare aranjate intr-un grid. In mod normal, este nevoie de o statie grafica aditionala pentru controlarea peretelui, asa cum se poate vedea in figura de mai jos.



Exemplu de HPVC si perete de vizualizare alcatuit din 3x3 monitoare.

Vizualizarea pe mai multe monitoare este condusa de mai multe noduri si necesita software specializat. In cazul particular al simularilor numerice ce se vor efectua in cadrul proiectului CyberDyn, datele de iesire se postproceseaza in format VTK/VTS, format care poate

fi importat de catre softul specializat ParaView (*www.paraview.org*). ParaView reprezinta o multi-platforma open-source si este folosit pentru vizualizarea de seturi mari de date. Utilizatorii ParaView pot construi rapid vizualizari si pot analiza la fel de rapid datele folosind tehnici calitative si cantitative. Interogarea datelor se poate face interactiv in 3D. ParaView a fost conceput astfel incat sa fie folosit pentru vizualizarea unor seturi de date de dimensiuni foarte mari folosind resurse de memorie RAM distribuite pe nodurile de vizualizare. Poate rula pe supercomputere pentru a analiza seturi foarte mari de date (TB), dar poate fi folosit si pe laptop-uri pentru vizualizarea unor seturi mai mici de date. ParaView ruleaza pe platforme Windows si Linux, si reprezinta un excelent mediu pentru interogarea rezultatelor obtinute in urma simularilor numerice efectuate pe sistemele de calcul paralele tip HPCC.

2.3. SISTEMUL GeoWall

Sistemul GeoWall (Geological Wall) reprezinta un sistem interactiv de proiectare stereoscopica 3D ieftina si eficienta a imaginilor, hartilor, filmelor precum si a modelelor numerice avansate. Sistemele de vizualizare 3D tip GeoWall sunt folosite pentru a explora seturi de date complexe din diverse domenii de cercetare, incluzand bineînteles stiintele Pamantului.

In general, un sistem GeoWall foarte raspandit este alcatuit dintr-o statie grafica cu o placa video performanta duala, doua proiectoare si un rack special pentru pozitionarea lor, filtre polarizante, un ecran de proiectie special, ochelari polarizanti si optional doua monitoare. Desi acest tip de GeoWall este comun in multe Universitati si centre de cercetare, prezinta dezavantajul de a fi in general static, sau foarte dificil de mutat si calibrat. O solutie la o asemenea problema ar fi folosirea unui display tip DLP cu capabilitati 3D. Aceasta are avantajul ca se elimina problema alinierii celor doua proiectoare, oferind un grad de mobilitate ridicat.

Exista o colectie bogata de pachete software care functioneaza pe GeoWalls pentru vizualizare, cum ar fi:

-Wallview: un 'Powerpoint pentru sistemele GeoWall' care ofera posibilitatea vizualizarii stereo a imaginilor

-Immersaview: afiseaza date 3D in format IV (Open Inventor) si VRML, atat date statice cat si serii care variaza cu timpul.

-Partiview: afiseaza seturi de date (statice si dinamice) de mari dimensiuni.

-PokeScope. Este folosit in principal pentru alinierea imaginilor stereo.

-WalkAbout. Este folosit la explorarea modelelor tip VRML specifice stiintelor pamantului.

-ArcGIS: un sistem commercial care permite vizualizarea 3D a oricarei imagini tip ArcScene.

-TriDef: program de vizualizare stereo.

3. STATE OF THE ART (STADIUL ACTUAL).

In vederea alegerii unei solutii de supercalculator care sa deserveasca scopul proiectului CyberDyn, am analizat pe scurt care este stadiul actual al celor mai rapide supercalculatoare din lume. O lista cu cele mai rapide supercalculatoare se publica de doua ori pe an pe site-ul *www.Top500.org*. Clasificarea are la baza testul Linpack prin care se masoara timpul de calcul necesar rezolvarii unui sistem de ecuatii liniare. Performanta acestor supercalculatoare se masoara in Tflops/s in termeni de performata teoretica (Rmax) si performata reala (Rpeak) (testul Linpack). In continuare prezentam o situatie la zi (iunie 2010) a celor mai rapide 5 supercomputere din lume (sursa *www.Top500.org*):

1) Jaguar - Cray XT5-HE (Oak Ridge Natioal Laboratory – SUA). Acest supercomputer are un numar record de nuclee de calcul, si anume 298,592 si o memorie RAM totala de 598,016 GB. Conexiunea intre nodurile de calcul se face printr-o retea de mare viteza tip Cray Gemini, iar sistemul operativ este bazat pe **Linux**. Pentru a obtine o asemenea performanta de calcul (petaflop), Jaguar foloseste **procesoare de tip AMD** cu 6 nuclee la o viteza de 2.6 GHz. Eficienta de calcul (Rmax/Rpeak) a acestui supercomputer este de ~75%.
2) Pe locul doi se afla supercalculatorul Nebulae – Dawning TC3600 Blade (National Supercomputing Centre in Shenzhen – China). Acest supercomputer foloseste procesoare Intel Xeon cu 6 nuclee (in total 120,640) la o frecventa de calcul de 2.66 GHz. Interconexiunea intre noduri se face printr-o retea de mare viteza **Infiniband QDR**. In plus, acest sistem foloseste si carduri grafice tip NVIDIA 2050. Sistemul operativ folosit este de asemenea **Linux**. Eficienta de calcul (Rmax/Rpeak) a acestui supercomputer este de ~42%.
3) Locul trei este detinut de supercalculatorul Roadrunner – BladeCenter QSss/LS21 (DOE/NNSA/LANL – SUA). Acesta este contruit de IMB si detine un numar de 122,400 nuclee de calcul si procesoare de tip PoweXCell 8i la 3.2 GHz in combinatie cu procesoare **AMD Opteron** DC 1.8 GHz. Reteaua de mare viteza este de tip **Infiniband** Voltaire. Ca si primele doua clasate, acesta utilizeaza ca sistem de operare **Linux**. Eficienta de calcul (Rmax/Rpeak) a acestui supercomputer este de ~75%.
4) Pe locul patru se afla supercalculatorul Kraken XT5-HE de la Cray (National Institute for Computational Sciences/University of Tennessee – SUA). Acesta foloseste tot procesoare **AMD Opteron** cu 6 nuclee la 2.6 GHz cu un total de 98,928 nuclee de calcul. Reteaua de

mare viteza este tip «Proprietary» iar sistemul de operare este **Linux**. Eficienta de calcul (Rmax/Rpeak) a acestui supercomputer este de ~80%.

5) Pe locul cinci se afla un supercomputer din Germania, Jugene - Blue Gene/P. Desi pe locul cinci acesta are cel mai mare numar de nuclee de calcul si anume 294,912. Acesta foloseste procesoare PowerPC la 850 MHz, iar reteaua de mare viteza este, ca si in cazul lui Kraken, «Propretary». Sistemul operativ este SUSE **Linux** Enterprise Server. Eficienta de calcul (Rmax/Rpeak) a acestui supercomputer este de ~80%.

Mai jos prezentam o scurta sinteza asupra tipului de procesor, retelei de mare viteza si sistemului de operare utilizate de catre aceste prime cinci supercalculatoare.

Denumire	Tip procesor	Tip retea de mare viteza	Eficienta de calcul (Rmax/Rpeak)	Sistem de Operare
1- Jaguar - SUA	**AMD Opteron**	Cray Gemini	**75%**	**Linux**
2 - Nebulae - China	Intel Xeon	**Infiniband QDR**	**42%**	**Linux**
3 - Roadrunner - SUA	PoweXCell 8i /**AMD Opteron**	**Infiniband**	**75%**	**Linux**
4 - Kraken - SUA	**AMD Opteron**	Proprietary	**80%**	**Linux**
5 -Jugene - Germania	PowerPC	Proprietary	**82%**	**Linux**

Dupa cum se poate observa, cel mai raspandit tip de procesor folosit de cele mai puternice supercalculatoare din lume este **AMD Opteron**. De asemenea, se foloseste frecvent o retea de mare viteza de tip **Infiniband QDR**, iar sistemul de operare folosit de toate cele cinci supercalculatoare este **Linux**. Eficienta de calcul a acestor sisteme variaza intre 52% si 82%. Tabelul de mai sus va fi folosit in definitivarea solutiei finale pentru sistemul HPCC – CyberDyn.

4. CERINTE TEHNICE (PERFORMANTA).

In acest capitol o sa subliniem performantele tehnice necesare pe care o infrastructura HPCC-HPVC-GeoWall trebuie sa le indeplineasca in vederea studierii evolutiei geodinamice a zonei Vrancea.

In primul rand sistemul HPCC-HPVC CyberDyn trebuie sa obtina o performanta de calcul reala cat mai mare in urma testelor Linpack. In principiu sistemul HPCC-HPVC CyberDyn trebuie sa obtina o performanta de calcul cat mai mare (de preferat >10 Tflops performanta de calcul teoretica) care sa se situeze in limitele obtinute de cele mai performate supercomputere din lume (42%-82%), dar de preferat peste 50%. Performanta de calcul a unui sistem HPCC-HPVC aplicat simularilor geodinamice de procese tectonice, este dictata in mare masura de faptul ca se lucreaza cu decompozitia de domeniu, in care fiecare latura a domeniului de calcul 3D are asignat un numar de procesoare. Acest lucru este necesar din doua motive principale:

1) in primul rand, datorita numarului foarte mare de variabile (proportional cu rezolutia modelelor, sau numarul de elemente finite) care trebuie stocat in memoria RAM este nevoie de o cantitate de memorie RAM semnificativa. O solutie des intalnita este folosirea de memorie impartita pe mai multe noduri de calcul.

2) in al doilea rand, datorita duratei de calcul necesara realizarii unei simulari numerice 3D sau 4D (3D + timp), care in functie de dimensiunile domeniului si a rezolutiei, poate sa fie foarte mare daca se foloseste un numar mic de procesoare. Solutia consta in folosirea unui numar mare de procesoare in fiecare directie geometrica a domeniului de calcul.

De exemplu, daca se aloca 8 procesoare in directia latitudinii, 8 procesoare in directia longitudinii si 8 procesoare in adancime, vom avea nevoie de 8x8x8=512 procesoare pentru a realiza o singura simulare numerica. Aceasta reprezinta o putere de calcul importanta si poate fi atinsa doar prin folosirea unui sistem HPCC cu un numar suficient de nuclee de calcul.

Ca atare, sistemul HPCC-HPVC CyberDyn trebuie sa aiba cel putin un numar de 2x512 nuclee de calcul pentru a putea efectua doua simulari numerice simultan. Este de preferat ca sistemul sa aiba mai mult de 1024 din motive de fiabilitate, in cazul in care, unul sau mai multe noduri de calcul prezinta instabilitati sau defectiuni care in general dureaza pana se remediaza. De exemplu, in cazul unui sistem cu un numar exact de 1024 de nuclee de calcul, defectiunea unui singur nod de calcul inseamna compromiterea puterii de calcul cu 50% cand se ruleaza modele cu 512 procesoare, cu alte cuvinte se poate rula un singur model. Trebuie mentionat ca numarul de procesoare rezervat unei simulari numerice se specifica odata cu conditiile initiale si nu se poate schimba pe parcursul unei simulari numerice. In acest studiu propunem ca sistemul

de calcul CyberDyn sa aiba un numar superior a 1024 de nuclee de calcul, pe cat posibil distribuite in cateva noduri separate. Din experienta proprie pe alte sisteme de calcul paralele tip HPCC, riscul unei defectiuni tehnice creste odata cu numarul componentelor din sistem. De exemplu este mult mai probabil ca un disk sa se strice in cadrul unui sistem de calcul cu multe discuri (sau noduri de calcul) decat in cazul unui sistem care are mai putine discuri (sau noduri de calcul). De aceea este de preferat ca sistemul HPCC-HPVC sa aiba cat mai putine componente posibil pentru aceeasi putere de calcul. In acest caz serverele de inalta performanta care suporta un numar mare de procesoare tip AMD (4 de exemplu) cu un numar ridicat de nuclee/procesor (12C de exemplu) sunt de preferat. Aditional, aceasta configuratie compacta este de preferat si datorita costurilor legate de reteaua de mare viteza care interconecteaza nodurile de calcul. S-a observat ca unul dintre obstacolele obtinerii unor performante ridicate la sistemele HPCC folosite in rezolvarea problemelor de mecanica de fluide, este reteaua de mare viteza. In cazul particular al problemelor de mecanica fluidelor (de exemplu, convectia in mantaua terestra) este nevoie de asigurarea unei comunicari foarte rapide intre nodurile de calcul. In momentul de fata, cele mai rapide retele sunt bazate pe tehnologia QDR Infiniband (40 Gb/s), insa costul unor asemenea sisteme nu este de loc de neglijat. De aceea, o minimizare a numarului de conexiuni rapide pentru aceeasi putere de calcul reprezinta un alt criteriu important in alegerea solutiei tehnice finale. Pentru o masina de tip Terrascale (sistem HPCC cu peste 1000 de procesoare) o solutie este folosirea unui singur switch QDR infiniband cu 36 de porturi, dar care impune folosirea unor noduri de calcul ultra dense cu un numar mare de nuclee de calcul pe nod. In plus, datorita numarului ridicat de nuclee de calcul intr-un singur nod (care poate ajunge la 48) cantitatea de informatie care trece printr-un singur port de retea este foarte mare. De aceea, propunem marirea vitezei retelei de mare viteza de la 10 Gb/s, cat a fost propus initial in CF, la 40 GB/s. Deci, solutia folosirii procesoarelor tip AMD cu 12 nuclee in servere cu 4 procesoare se poate face in tandem cu un sistem de comunicatie rapida QDR cu un numar redus de porturi, astfel incat puterea de calcul proiectata initial (1024 nuclee de calcul) poate fi chiar depasita. In plus, asa cum am mentionat mai sus, o asemenea solutie compacta este de preferat si datorita cosurilor de intretinere/reparatie mai scazute. In final, dorim sa mentionam ca la ora actuala procesoarele AMD Opteron si sistemele de comunicatie Infiniband QDR sunt cele mai raspandite in randul celor mai puternice supercomputere la nivel mondial.

5. ARHITECTURA DE CALCUL/VIZUALIZARE OPTIMA.

Avand in vedere cele prezentate mai sus, precum si datorita altor considerente care au aparut pe parcurs (de exemplu spatiul fizic unde se va amplasa HPCC-HPVC CyberDyn), prin acest studiu recomandam ca solutia HPCC-HPVC sa indeplineasca urmatoarele cerinte de baza:

1) Avand in vedere ca spatiul pus la dispozitie de catre IGAR pentru instalarea intregului sistem HPCC-HPVC se afla intr-o incapere care are un subsol cu tavan boltit din caramida, greutatea intregului sistem trebuie minimizata mentinand desigur puterea de calcul in parametrii proiectati. In plus, recomandam ca greutatea sa fie distribuita cat mai mult pe orizontala (evident in limitele spatiului disponibil). Pentru acest lucru propunem ca intreaga solutie (si greutate) trebuie sa fie distribuita si echilibrata in 5x42U rack-uri.

2) Toate procesoarele trebuie sa fie bazate pe arhitectura de x86-64 pentru a putea avea acces la cat mai multa memorie RAM necesara realizarii modelelor numerice cu mare rezolutie. In plus, este de recomandat sa se foloseasca pe cat cu putinta procesoare cu un numar cat mai mare de nuclee de calcul pe procesor. La ora actuala procesoarele AMD Opteron ofera cel mai mare numar de nuclee de calcul/socket. In acest fel se reduc considerabil costurile aferente retelei de mare viteza, conectarii la sistemul de alimentare cu energie electrica (PDUs), costurile de intretinere si curent electric, precum si greutatea intregului sistem.

3) Pentru o mai mare flexibilitate, propunem ca sistemele pentru clusterul de calcul (HPCC) si pentru clusterul de vizualizare (HPVC) sa fie compatibile in totalitate. Nodurile de vizualizare, atunci cand nu sunt folosite, trebuie sa poate fi folosite ca noduri de calcul pentru clusterul de calcul (HPCC). Pentru a asigura o viteza ridicata in timpul simularilor mumerice, toate nodurile de calcul (precum si cele principale) trebuie sa fie prevazute cu placi de retea tip QDR Infiniband.

4) Este recomandabila folosirea unui UPS care sa permita numai componentelor de baza (reteau de administrare, stocare, etc.) sa ruleze pentru 10 minute si sa se inchida automat si fara erori in cazul intreruperii alimentarii principale cu energie electrica. UPS-ul nu este necesar pentru intregul sistem de calcul datorita riscului de supraincalzire in cazul in care, pana de curent dureaza mai mult de cateva minute (max. 2 minute), sistemul de racire fiind intrerupt.

5) Se recomanda ca toate sistemele trebuie sa suporte IPMI pentru administrarea la distanta. Acesta trebuie sa ruleze pe o retea dedicata (Gbit) administrarii si care este complet separata de reteaua MPI si I/O.

6) Din motive de fiabilitate si siguranta, este indicata prezenta a doua noduri principale care sa preia administrarea clusterului unul de la celalalt cand este necesar (de exemplu atunci cand apare o eroare in functionarea hardului pentru unul dintre nodurile principale). Ideal este ca aceste noduri sa fie echipate cu cate 2 discuri conectate in mod RAID1. Existenta a doua noduri principale ofera suport pentru un numar mare de utilizatori care isi vor putea edita si compila codurile in acelasi timp, si de asemenea, aceste noduri principale vor putea oferi posibilitatea administrarii clusterului. In plus, cele doua noduri principale trebuie sa contina doua procesoare cu specificatii similare celor folosite pentru nodurile de calcul ale clusterului. Aditional nodurile principale trebuie sa fie echipate cu tastatura, mouse si monitor (sistem tip KVM). Trebuie sa aiba de asemenea un DVD reader/writer pentru instalare si back-up. Sistemul KVM strebuie sa fie capabil sa asigure acces usor pentru toate serverele din sistem. Nodurile principale trebuie sa contina cel putin 2 interfete Gigabit Ethernet (una pentru management-ul intregului sistem de calcul si cealalta pentru conectarea la intranetul Institutului). Nodurile principale trebuie sa aiba legatura la reteaua de mare viteza tip Infiniband QDR.

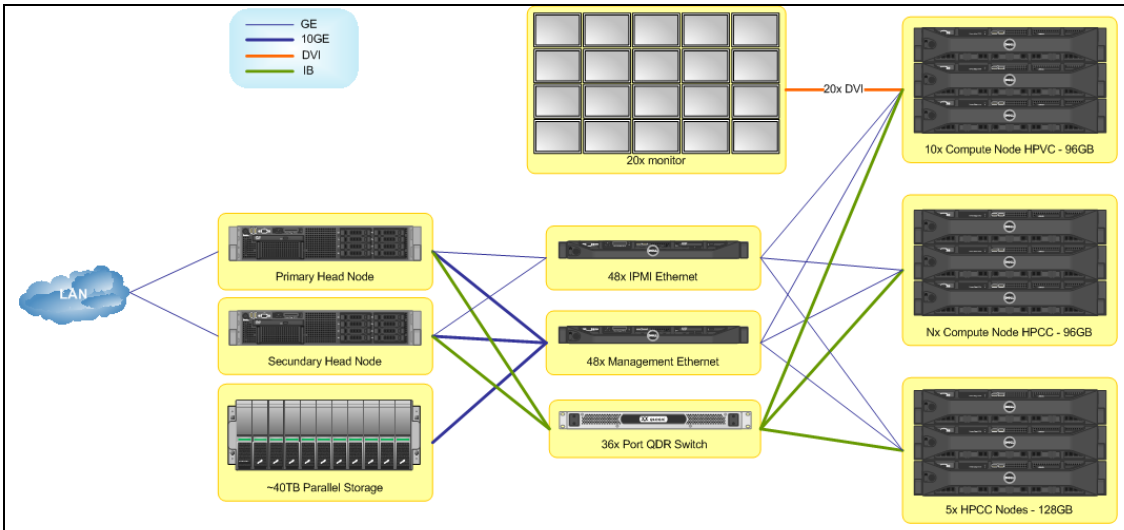
7) In cazul sistemul HPVC, nodurile de vizualizare trebuie sa aiba capacitatea de a rula normal programe in paralel in cadrul clusterului, atunci cand nu sunt folosite pentru vizualizare in paralel. Toate nodurile de vizualizare formeaza o parte a sistemului HPCC si trebuie sa fie compatibile cu restul HPCC-ului. Aceste noduri de vizualizare trebuie sa suporte minim 16 (4x4) monitoare de inalta rezolutie.

8) Toate nodurile clusterului (incluzand nodurile principale) trebuie sa fie interconectate printr-o retea Gigabit Ethernet pentru control si pentru traficul non-MPI. In plus toate nodurile clusterului (incluzand nodurile pricipale) trebuie in plus sa fie interconectate printr-o retea de inalta performanta pentru traficul MPI. Performanta recomandata a interconexiunilor este cea a InfiniBand QDR.

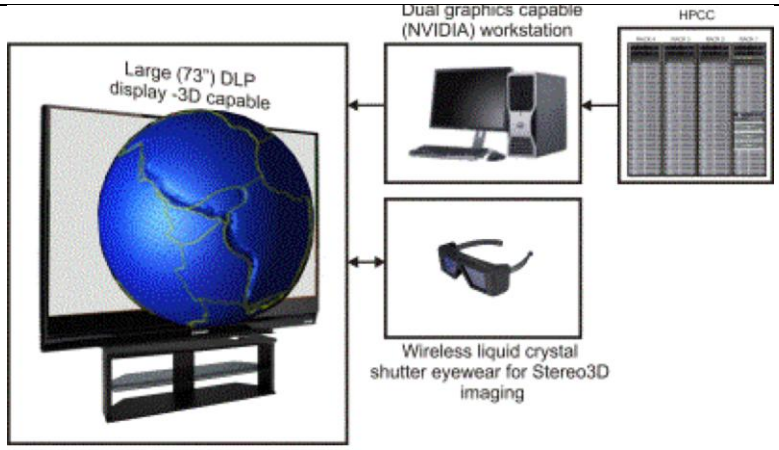
9) Sistemul HPCC-HPVC trebuie sa asigure stocarea si accesarea rapida (scriere/citire) a unui volum mare de date (recomandabil 40 TB). Acest sistem trebuie sa fie sigur, fara a exista riscul de pierdere de date in cazul unei defectiuni (recomandabil RAID 5 sau RAID 6). De preferinta, sistemul de stocare sa fie de tip NAS conectat la nodurile principale printr-o conexiune rapida tip 10 Gb/s.

Mai jos prezentam intr-o forma grafica solutia HPCC-HPCC si GeoWall recomandata.

Distributia echilibrata a intregului sistem HPCC-HPVC in 5 rack-uri.



Schema de conectare a sistemului HPCC-HPVC.



Solutia GeoWall pentru vizualizarea stereo a rezultatelor modelarilor numerice 3D.

6. DIFERENTE INTRE ARHITECTURA OPTIMA SI CEA PROPUSA IN CEREREA DE FINANTARE (CF) A PROIECTULUI CYBERDYN.

Solutia finala a sistemului de calcul HPCC-HPVC CyberDyn (sistemul GeoWall este neschimbat) este diferita (si mult imbunatatita) fata de solutia propusa in cererea de finantare din urmatoarele motive principale:

1) **Nodurile de calcul**. Initial ne-am propus sa avem un numar de 1024 nuclee de calcul, distribuite in 128 de noduri de calcul cuplate la o retea de 10 Gb/s formata dintru-un switch central de mare capacitate cu un minim de 128 de porturi, si evident un numar minum de 128 de carduri de retea de 10 Gb/s. Solutia imbunatatita pe care o propunem ne ofera posibilitatea sa distribuim un numar de nuclee de calcul intr-un numar mult mai mic de noduri de calcul (servere cu 4 procesoare tip AMD 12C). Acest lucru inseamna o reducere considerabila a numarului de porturi de retea de mare viteza, fiind posibila folosirea unui singur switch tip QDR Infiniband cu 36 de porturi, precum si reducerea aferenta a numarului de carduri de retea concomitent cu marirea considerabila a ratei de transfer. Prin reducerea costurilor aferente retelei de mare viteza se spera ca numarul total de nuclee de calcul sa fie superior celui prevazut initial in CF. De exemplu, un numar de 28 de noduri de calcul, fiecare cu 48 de nuclee, ofera un total de 1344 (28x48) nuclee de calcul. Viteza retelei de mare viteza (folosita pentru comunicarea intre nodurile de calcul) este sporita cu 400%, de la 10 Gb/s cat am propus in CF la 40 Gb/s (QDR), iar numarul de nuclee de calcul este mai ridicat cu ~25% (de la 1024 la 1344). Aditional, folosirea unui numar mai redus de noduri de calcul inseamna un consum mai redus de electricitate, precum si o scadere considerabila a costurilor de intretinere atat pe durata proiectului cat si in perioada ulterioara.

2) Nodul principal. In CF am prevazut un singur nod central (principal sau master node). In configuratia finala propusa prevedem doua noduri principale care actioneaza in modul failover: daca unul dintre cele doua noduri principale are o defectiune (hard sau soft), atunci, celalalt nod preia automat toate comenzile si rulajele, astfel incat programele care ruleaza pe cluster sa nu fie intrerupte, iar datele utilizatorilor sa nu fie in pericol. In plus, nodurile principale au o capacitate de calcul mult sporita fata de ce am prevazut initial in CF, si anume, de preferinta acestea au cate doua procesoare AMD Opteron fiecare cu cate 12 nuclee, fata de doua procesoare cu cate 4 nuclee, deci o crestere a numarului de nuclee de calcul disponibile de 300%. Aditional memoria RAM s-a dublat, fiecare nod principal

avand 64 GB RAM fata de numai 32 GB RAM cat am prevazut in CF. Acest lucru are impact pozitiv asupra performantelor sistemului de calcul CyberDyn, deorece utilizatorii pot compila programe, copia/muta volume mari de date, rula scripturi fara ca performanta intregului sistem sa fie afectata.

3) Spatiul de stocare de date: solutia propusa initial a constat dintr-un sistem DAS de 15 TB direct conectat la nodul principal. Solutia finala propusa in acest studiu ne ofera eventual posibilitatea de a avea un sistem de stocare de date de ultima generatie tip (NAS), mult mai performant, cu un spatiu de stocare marit la 40 TB (si 20 GB memorie cache) si cu o conexiune 10GB cu nodurile principale. Acest sistem este mult mai rapid decat sistemul DAS initial propus in CF (cu card RAID incorporat in nodul principal). Aditional, din testele efectuate pe sisteme similare, dar care folosesc un sistem DAS pentru stocare in loc de NAS, s-a observat ca la rulajele cu multe procesoare si rezolutie inalta, un sistem DAS foloseste intensiv si din resursele nodului principal la care este conectat, marind astfel timpul total necesar efectuarii unei simulari numerice. In schimb, prin folosirea unui sistem NAS, s-a observat disparitia acestui fenomen.

4) Sistemul HPVC: solutia finala de implementare a sistemului HPCC-HPVC este mult mai flexibila, un numar de 10 noduri de calcul al sistemului HPCC pot fi folosite in scop dual in functie de necesitati: atat pantru calcule numerice cat si pentru sistemul de vizualizare HPVC. In plus, propunem cresterea numarului de monitoare cu 4, de la 16 (4x4) prevazute in CF, la 20 (5x4), sporind astfel cu 20% rezolutia finala a intregului sistem de vizualizare paralela HPVC.

In concluzie, solutia finala a sistemului de calcul HPCC-HPVC CyberDyn este cu mult mai performanta si robusta decat solutia initial propusa in cererea de finantare.

7. REFERINTE BIBLIOGRAFICE

-Lehmann, T. , 2009. Building a linux-based high-performance compute cluster. LINUX Journal no. 182. (http://www.linuxjournal.com/magazine/building-linux-based-high-performance-compute-cluster)

-Lucke, R., 2005. Buiding Clustered Linux Systems. Prentince Hall Computer, 606 pp.

-Mao, J., 2010. The cluster node layout. JunJun´s Admin Notes (http://pka.engr.ccny.cuny.edu/~jmao/node/37)

-Morton, D., 2007. High-Performance Linux Clusters. LINUX Journal no. 163. (http://www.linuxjournal.com/article/9828)

-Rajkumar Buyya (editor): High Performance Cluster Computing: Architectures and Systems, Volume 1, ISBN 0-13-013784-7, Prentice Hall, NJ, USA, 1999.