UNIVERSITAT OBERTA DE CATALUNYA

Avaluació d'aplicacions HPC sobre Clouds públics/privats

TREBALL FI DE CARRERA

ALUMNE: *DAVID CALVO SENDON*
ETIG

CONSULTOR: *IVAN RODERO CASTRO*

DATA: *JUNY 2.010*

1 DEDICATÒRIES I AGRAÏMENTS

A Cristina i Víctor, els dos sabeu quan us estimo. Gràcies per tot, us recompensaré pels moments que no hem pogut gaudir junts.

Als meus pares, especialment a la meva mare per l'ajuda en l'última fase.
Jose Luís, gràcies pel suport i la fe que sempre m'has demostrat.
Eduard, gran amic, saps molt bé tot el que has fet per mi.
Ivan, sense qui aquest projecte no hauria estat possible.

"Cogito, ergo sum" *Descartes*

2 ÍNDEX

1 DEDICATÒRIES I AGRAÏMENTS 2
2 ÍNDEX 3
3 INTRODUCCIÓ 5
3.1 MOTIVACIONS DEL PROJECTE 5
3.2 OBJECTIUS 7
3.3 DESCRIPCIÓ DEL PROJECTE 8
4 PLANIFICACIÓ 10
4.1 DESCRIPCIÓ DETALLADA DE LES FASES 10
4.2 DIAGRAMA DE GANTT 12
5 APLICACIÓ HPC 13
5.1 ELECCIÓ DE L'APLICACIÓ 13
5.2 WRF – WEATHER RESEARCH & FORECASTING MODEL 13
5.2.1 DESCRIPCIÓ FUNCIONAMENT 14
5.3 MODEL COMPUTACIONAL COMÚ PER L'EXECUCIÓ DE LES PROVES 17
6 PREPARACIÓ DELS ENTORNS 18
6.1 CLOUD PÚBLIC 18
6.1.1 INTRODUCCIÓ A AMAZON ELASTIC COMPUTE CLOUD – EC2 19
6.1.2 INICIALITZACIONS 21
6.1.3 SEGURETAT 22
6.1.4 CREACIÓ IMATGE BASE 25
6.1.5 PREPARACIÓ WRF A EC2 27
6.1.6 CREACIÓ IMATGE FINAL 31
6.2 CLOUDS PRIVATS 32
6.2.1 MARENOSTRUM 32
6.2.2 CLÚSTER IBM POWER6 35
6.2.3 CLOUD PRIVAT 36
7 EXECUCIONS 37
7.1 EXECUCIONS SOBRE EC2 37
7.2 RESULTATS OBTINGUTS A AMAZON EC2 40

7.3 RESULTATS OBTINGUTS A MARENOSTRUM 45
7.4 RESULTATS OBTINGUTS A IBM POWER6 46
7.5 RESULTATS OBTINGUTS A CLOUD PRIVAT 47
8 ESTUDI DELS RESULTATS 48
8.1 ANÀLISI DE COSTS 48
8.1.1 AMAZON EC2 48
8.1.2 MARENOSTRUM 50
8.1.3 IBM POWER6 51
8.1.4 CLOUD PRIVAT 51
8.2 ESTUDI COMPARATIU 53
9 CONLUSIONS FINALS 57
9.1 DESVIACIONS RESPECTE LA PLANIFICACIÓ 58
9.2 VALORACIÓ PERSONAL 59
10 BIBLIOGRAFIA 60
11 ANNEXOS 62
11.1 MPI - MESSAGE PASSING INTERFACE 62
11.2 SCRIPTS EC2 63
11.3 PARÀMETRES NAMELIST WPS I INPUT 65
11.4 REGISTRE UCAR I OBTENCIÓ DE DADES 68
11.5 IMATGE BASE EC2 71
11.6 XEN 72

3 INTRODUCCIÓ

3.1 MOTIVACIONS DEL PROJECTE

Actualment existeixen aplicacions importants per a la societat que requereixen de sistemes amb un poder de càlcul molt elevat, coneguts com HPC (*High Performance Computing*). Un exemple d'aquests tipus d'aplicacions és la predicció meteorològica, on gràcies a complexos models matemàtics i el procés en paral·lel de nombroses dades és possible simular i predir el comportament de fenòmens complexos com ara tornados o huracans, així com calcular les possibles direccions de tempestes de grans magnituds. També podem citar l'animació en 3D, associada a l'industria cinematogràfica, com un altre exemple d'aplicacions amb alta demanda de recursos computacionals, on la utilització de noves arquitectures i tecnologies paral·leles permet realitzar animacions amb un grau de realisme i complexitat molt elevat.

Aquests tipus d'aplicacions exigeixen un grau de rendiment computacional que no es pot obtenir amb sistemes convencionals com ara ordinadors personals. Per tant, cal adoptar solucions més específiques que s'acostumen a implementar mitjançant sistemes de computadors agrupats en "clústers". Els clústers permeten utilitzar múltiples computadors en paral·lel mitjançant models de programació paral·leles i normalment ofereixen gran capacitat computacional gracies a la utilització de xarxes d'interconnexió molt rapides.

Recentment, gràcies a Internet i les tecnologies de virtualització, ha nascut el paradigma del *Cloud Computing*. Mitjançant el Cloud podem fer servir recursos computacionals distribuïts i, a més, mitjançant maquines virtuals podem fer una separació lògica entre els hardwares i els sistemes dels recursos computacionals i les aplicacions. D' aquesta manera hi ha reptes a nivell de la seva explotació, i per tant cal estudiar les formes d'ús i parametrització requerida per a proporcionar el nivell de càlcul que requereixen les aplicacions HPC.

Amb aquest projecte es pretén avaluar l'execució d'aplicacions HPC sobre diferents models de Clouds, per tal d'avaluar les diferents possibilitats que existeixen actualment, així com la seva configuració i execució, realitzant una anàlisi basada en la

utilització d'una aplicació real i un estudi dels pros i contres així com del seu cost/benefici.

D'aquesta manera amb aquest projecte es desitja aconseguir una aproximació tècnica que serveixi com a punt de partida en la pressa de decisions a l'hora de la elecció d'aquest tipus d'eines.

Com a exemple directe d'ús d'aquest projecte en el mon empresarial podem esmentar solucions híbrides (com ara el *Cloud Bursting*), que ens permeten traspassar part de la càrrega de treball (càlcul, transaccions, emmagatzematge,...) a un d'aquests entorns quan la nostra infraestructura (el nostre cloud/data center) arriba a la seva màxima utilització i poder així absorbir pics de treball puntuals sense la necessitat de sobredimensionar el hardware amb el seu conseqüent augment del TCO (*Total Cost of Ownership* [1]).

Mitjançant l'estudi presentat en aquesta memòria tindrem una visió més clara de costos financers i temporals, així com les possibles modificacions necessàries per tal que una empresa pugui fer ús d'aquesta tecnologia i quins dels diferents entorns s'adapta millor a les seves necessitats.

[1] Total Cost of Ownership – Concepte relatiu als costos finals que reflecteixen no sòls el costos de compra sinó també els associats a aspectes d'ús i manteniment.

3.2 OBJECTIUS

Tal i com hem avançat a les motivacions del projecte, els objectius finals que volem aconseguir son els següents:

- Conèixer amb detall l'estat de l'art respecte *Cloud Computing*, tant sobre els Clouds públics disponibles actualment, com les diferencies bàsiques amb els Clouds privats.
- Preparar, configurar i utilitzar un Cloud públic per tal d'executar aplicacions paral·leles programades amb MPI (Message Passing Interface).
- Adquirir experiència en l'execució de tasques sobre Clouds utilitzant aplicacions HPC com a cas d'ús específic.
- Conèixer i executar aplicacions HPC en diferents plataformes i analitzar-ne el seu rendiment i cost/benefici.
- Conèixer i analitzar el comportament de l'aplicació WRF (*Wheather Research & Forecasting*), relacionada amb predicció meteorològica avançada.
- Avaluar les configuracions Cloud privat / Cloud públic seguint criteris de rendiment/cost.
- Realitzar un estudi d'idoneïtat (pros/contres) dels diferents Clouds segons el model "pay as you go"[2] i crear un enfocament dels resultats vers el model empresarial actual.
- Obtenir una anàlisi comparativa de costs dels diferents models de Clouds públic i privats.
- Realitzar un abordatge pràctic que ens permeti aplicar els coneixements obtinguts dins un marc empresarial real.

[2] *"Pay as you go"* fa referència al concepte de pagament en funció de l'ús. Similar als telèfons mòbils de targeta, no existeixen taxes mensuals i només es paguen els serveis consumits. Sota el marc d'aquest treball s'aplicarà a màquines virtuals i serveis derivats (emmagatzematge, transferència,...).

3.3 DESCRIPCIÓ DEL PROJECTE

Per tal d'avaluar el rendiment d'aplicacions HPC en diferents plataformes i poder acotar els resultats obtinguts, centrarem el projecte en l'estudi i avaluació d'una aplicació HPC concreta: WRF - Wheather Research & Forecasting. WRF és una aplicació de predicció meteorològica que permet el tractament d'informació mitjançant algorismes molt sofisticats de càlcul en paral·lel. La raó per la qual s'ha escollit aquesta eina es descriu en detall a l'apartat 5 - Aplicació HPC.

En una primera fase s'analitzarà en detall l'aplicació HPC escollida. Posteriorment, es prepararan i configuraran els entorns de Cloud públic i privat tenint en compte les característiques d'aquestos.

Caldrà definir amb detall els diferents entorns i configurar i preparar aquestos per tal que siguin capaços d'executar l'aplicació MPI. En concret, caldrà instal·lar i configurar el "runtime" de MPI, el qual s'encarrega de gestionar els paral·lelisme i pas de missatges entre processos (que poden estar distribuïts en diferents computadors).

Com a entorn de Cloud públic farem servir un compte Amazon EC2 (sempre tenint en compte limitacions en el pressupost). En descriurem la seva configuració i les seves particularitats. Posteriorment realitzarem una anàlisi dels costos, tant de rendiment, econòmics, com d'aprenentatge, així com una anàlisi del seus pros i contres.

Es farà servir XEN com a tecnologia de virtualització, la qual gestiona màquines virtuals que permeten muntar sistemes operatius complets i funcionals independentment del hardware i sistema operatiu dels servidors. Com a Cloud privat tindrem en compte un clúster de servidors mono-processador (de cost relativament baix) implementant la virtualització amb XEN.

Com a entorns no virtualitzats basats en clúster es comptarà amb dues plataformes HPC molt diferents. Per una banda hem tingut en compte el supercomputador Marenostrum (Barcelona Supercomputing Center), i per l'altra banda un clúster de recerca basat en arquitectura IBM Power6. Les dades obtingudes en aquestes dues

darreres plataformes s'han obtingut a partir d'una col·laboració dins del marc del projecte *Latin American Grid* del qual formen part tant el Barcelona Supercomputing Center com IBM. Cal remarcar que no són dades confidencials i que no s'ha tingut accés a les màquines directament.

Amb els entorns preparats, procedirem a la execució seguint diferents configuracions de nodes, tot especificant els procediments emprats i scripts necessaris.

Els resultats obtinguts en la fase anterior ens permeten realitzar un estudi detallat a partir del qual arribem a les nostres conclusions basant-nos en criteris de rendiment/cost. Com a punt final realitzem un detall de pros i contres basats en el model "pay as you go" que ens permeten avaluar les dues configuracions escollides inicialment.

4 PLANIFICACIÓ

4.1 DESCRIPCIÓ DETALLADA DE LES FASES

FASE 1: PLANIFICACIÓ

Establiment dels objectius i elaboració de la planificació detallada del projecte, realitzant el càlcul temporal en funció de la informació disponible i els venciments de les entregues de documentació, tant parcials com l'entrega final. S'ha tingut en compte el contingut de les diferents entregues per tal d'aconseguir els resultats adequats amb el marge suficient que permeti desviacions, donat que en tot projecte sempre poden aparèixer complicacions que amb previsió poden ser resoltes. En els punts més susceptibles de provocar retards substancials s'ha intentat dotar del màxim temps possible.

FASE 2: ESTUDI CARACTERISTIQUES/REQUERIMENTS APLICACIÓ HPC

En aquest punt és definirà l'aplicació HPC objecte de les proves. S'elaborarà per tant un anàlisi exhaustiu d'aquesta i es realitzarà una verificació de la correcta planificació en funció de les possibles desviacions ocasionades pel seu tipus i característiques aplicades als entorns de Clouds on es realitzaran les proves, realitzant si cal els canvis oportuns.

FASE 3: PREPARACIÓ ENTORNS DE TREBALL

Es definiran i establiran les bases dels entorns de treball. Es prepararà l'entorn de Cloud públic amb Amazon EC2 i els entorns privats, tot definint específicament el seu tipus i tecnologia. Es realitzarà un procés de seguiment per elaborar una guia d'instal·lació i preparació dels entorns i es crearan les imatges virtuals necessàries pel seu funcionament.

FASE 4: EXECUCIÓ

Dins el marc d'aquesta fita es realitzarà la configuració de cadascuna de les aplicacions per tal de funcionar sota els entorns específics de cada Cloud. En una primera etapa realitzarem un procés de preparació i possibles adaptacions. En la segona etapa realitzarem les execucions i obtenció de resultats. S'han establert uns temps prudencials durant la etapa d'execució per tal d'afrontar possibles inconvenients relatius a la posta a punt, problemes de disponibilitat o possibles entrebancs que es puguin ocasionar per realitzar les proves. L'objectiu principal d'aquesta fase es obtenir els resultats experimentals.

FASE 5: ESTUDI DELS RESULTATS

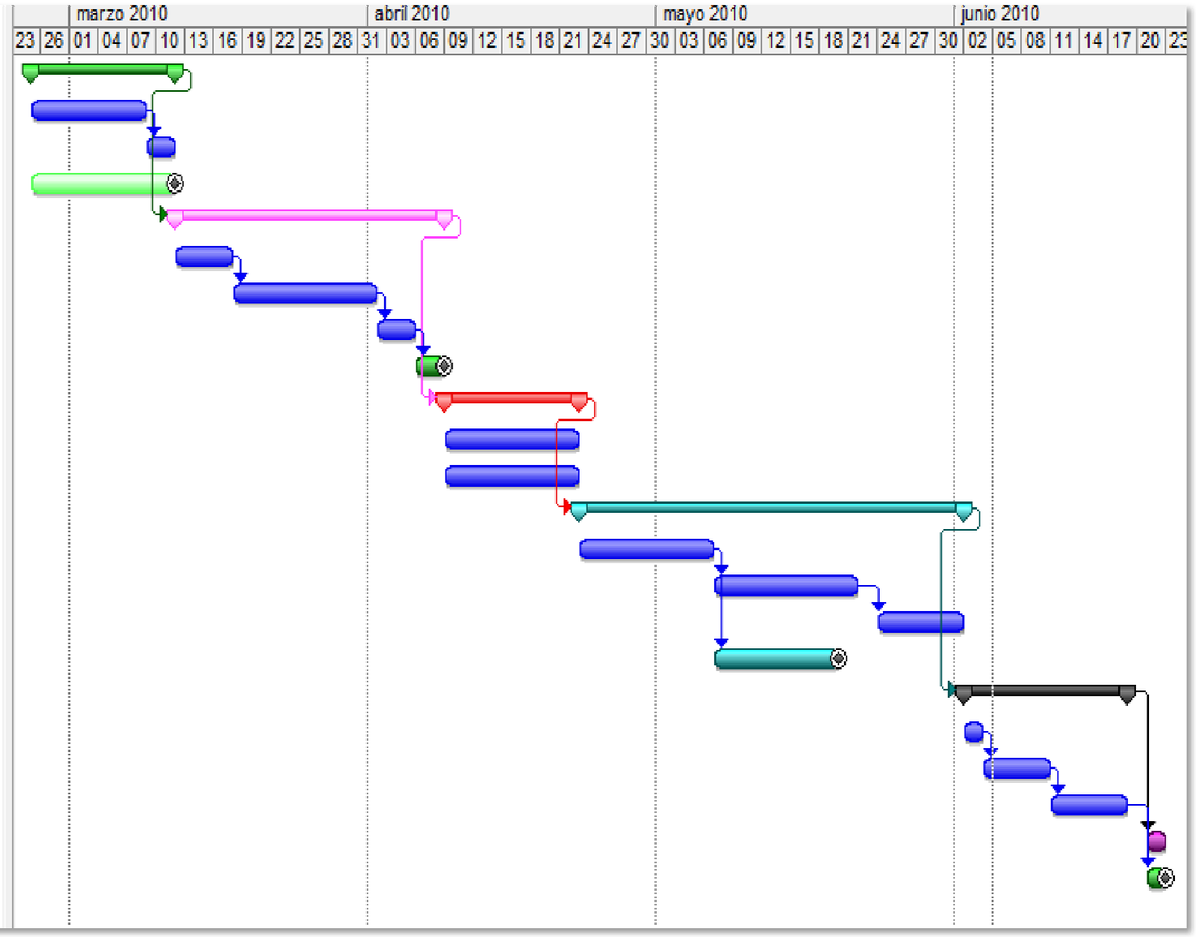
Aquest serà el marc que ens permetrà realitzar l'anàlisi de les dades obtingudes al procés anterior. Es realitzarà un estudi comparatiu tenint en compte els criteris 'pay-as-you-go', i el rendiment/cost, així com valoracions relatives a dificultats obtingudes als diferents processos de configuració i preparació dels entorns. En cas que sigui necessari podrem tornar a realitzar alguna de les execucions si trobem resultats fora dels criteris estadístics.

FASE 6: MEMORIA I PRESENTACIÓ

En aquest punt realitzarem la memòria que contindrà tota la informació generada durant el procés de treball. Aquesta memòria estarà adaptada en el seu format als criteris establerts pel pla docent i pot requerir per tant alteracions formals respecte als documents generats fins al moment, per tal que aquestos s'emmarquin sota aquestes directrius. També caldrà fer un document presentació de tipus PowerPoint.

4.2 DIAGRAMA DE GANTT

	Nombre de tarea	Duración	Comienzo	Fin	Predeceso
1	**PLANIFICACIÓ**	**11 días**	**jue 25/02/10**	**jue 11/03/10**	
2	Analisi d'objectius i descripció del projecte	8 días	jue 25/02/10	lun 08/03/10	
3	Temporització i establiment de fites	3 días	mar 09/03/10	jue 11/03/10	2
4	El·laboració document PLA DE TREBALL(PAC 1)	11 días	jue 25/02/10	jue 11/03/10	
5	**ESTUDI CARACTERISTIQUES/REQUERIMENTS APLICACIONS HPC**	**20 días**	**vie 12/03/10**	**jue 08/04/10**	**1**
6	Definició i concreció de la aplicacio	4 días	vie 12/03/10	mié 17/03/10	
7	Anàlisi de requeriments	11 días	jue 18/03/10	jue 01/04/10	6
8	Anàlisi possibles desviacions en planificació	2 días	vie 02/04/10	lun 05/04/10	7
9	El·laboració document PAC2	3 días	mar 06/04/10	jue 08/04/10	8
10	**PREPARACIÓ ENTORNS DE TREBALL**	**10 días**	**vie 09/04/10**	**jue 22/04/10**	**5**
11	Preparació i configuració Cloud Amazon EC2	10 días	vie 09/04/10	jue 22/04/10	
12	Preparació i configuració Clouds Privats	10 días	vie 09/04/10	jue 22/04/10	
13	**EXECUCIÓ MAQUINES VIRTUALS**	**28 días**	**vie 23/04/10**	**mar 01/06/10**	**10**
14	Instalació i configuració aplicacions	10 días	vie 23/04/10	jue 06/05/10	
15	Execució i obtenció de resultats	11 días	vie 07/05/10	vie 21/05/10	14
16	Avaluació i contrast dels resultats (Si cal, repetició execució)	7 días	lun 24/05/10	mar 01/06/10	15
17	El·laboració document PAC3	9 días	vie 07/05/10	mié 19/05/10	14
18	**ESTUDI RESULTATS**	**13 días**	**mié 02/06/10**	**vie 18/06/10**	**13**
19	Contrast i interpretació dels resultats obtinguts	2 días	mié 02/06/10	jue 03/06/10	
20	El·laboració d'estudi de rendient	5 días	vie 04/06/10	jue 10/06/10	19
21	Estudi i conclusions del pros/contres dels models	6 días	vie 11/06/10	vie 18/06/10	20
22	EL·LABORACIÓ MEMORIA - DOCUMENT FINAL	2 días	lun 21/06/10	mar 22/06/10	18
23	PRESENTACIÓ I ENTREGA DOCUMENT FINAL	2 días	lun 21/06/10	mar 22/06/10	21



5 APLICACIÓ HPC

5.1 ELECCIÓ DE L'APLICACIÓ

Per tal d'avaluar els diferents entorns de Clouds privats i públics és necessari utilitzar aplicacions HPC que ens permetin obtenir resultats quantitatius, per tal de poder mesurar i comparar-los amb diferents execucions i entorns. Existeixen diferents aplicacions de benchmarking orientades a HPC, que compleixen aquests requisits, disponibilitzant nombrosos paràmetres d'anàlisi.

De tota manera també és necessari poder avaluar aquests entorns des de la perspectiva del model 'pay as you go', a més de realitzar una anàlisi de les diferents dificultats d'ús i aprenentatge dels diferents entorns. Per aquestes raons, la utilització d'eines exclusives de benchmarking s'ha cregut insuficient per tal de poder abordar els requisits inicials descrits al projecte. D'aquesta maenra s'ha intentat trobar una aplicació HPC que ens permetés realitzar aquestes proves i al mateix temps es tractés d'una eina d'ús real. Addicionalment existeixen les limitacions generades per la duració i orientació d'aquest treball i òbviament les limitacions de pressupost associades.

Per totes aquestes raons, s'ha realitzat l'escolla de la següent aplicació HPC com candidata de les nostres execucions.

5.2 WRF – WEATHER RESEARCH & FORECASTING MODEL

Es tracta d'una aplicació HPC de lliure distribució que permet realitzar a partir de dades d'entrada prediccions meteorològiques. Es troba molt difosa dins l'entorn d'aplicacions paral·leles d'alt rendiment i permet comparar els resultats amb molts entorns i plataformes. Tanmateix, gràcies a la seva amplia difusió, permet analitzar els seus costs d'implementació i rendiment.

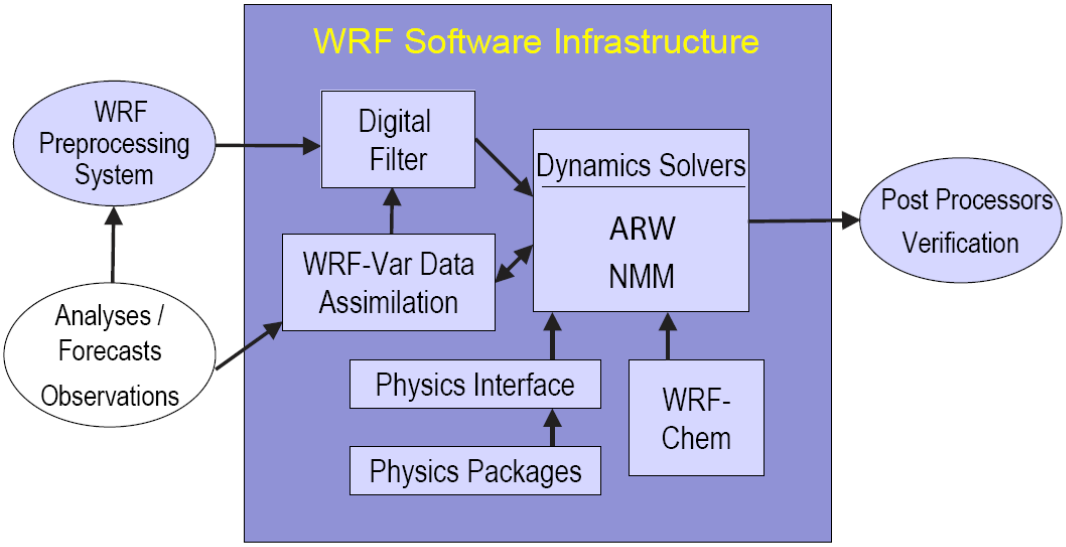
Tot i que permet la seva compilació seguint diferents criteris de execució en paral·lel, ens centrarem en el seu ús segons l'estàndard MPI (Annex 11.1), més utilitat actualment.

5.2.1 DESCRIPCIÓ FUNCIONAMENT

WRF és un model de predicció numèrica de mesoescala d'última generació (NWP), tant per la realització de previsions com per l'estudi atmosfèric. Ha sigut fruit d'un esforç col·laboratiu de principalment el *National Center for Atmospheric Research (NCAR),* la *National Oceanic and Atmospheric Administration*, els *National Centers for Environmental Prediction (NCEP)* i el *Forecast Systems Laboratory (FSL),* la *Air Force Weather Agency (AFWA),* la *Naval Research Laboratory*, *Oklahoma University*, i la *Federal Aviation Administration (FAA).*

Està basat en un model de programació que permet tant la seva execució en ordinadors personals com en supercomputadors en paral·lel, fet que el fa perfecte pel propòsit d'aquest projecte.

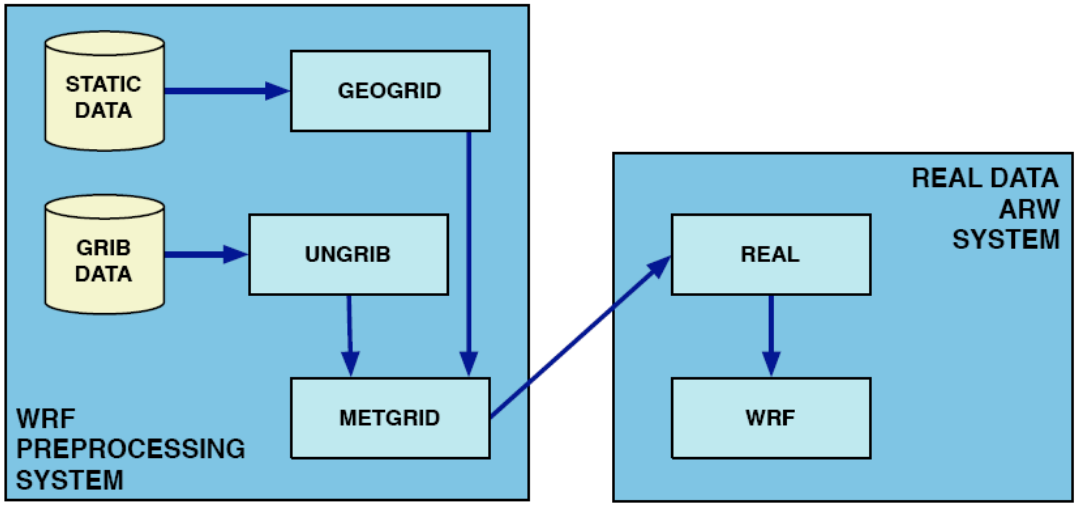
La següent imatge mostra els components del model i la seva infraestructura:



En relació al seu funcionament orientat a aquest projecte, WRF s'estructura es 2 parts diferenciades:

- WPS – WRF Preprocessing System
- WRF – Real Data ARW System

El model d'execució i funcionament es basa en termes generals en l'esquema següent.



El primer subsistema, *WRF Preprocessing System* o WPS, donades unes dades d'entrada, realitza un preprocés per tal de que aquestes puguin ser analitzades pel segon subsistema, model WRF.

Aquest preprocés està dividit en 3 parts i es realitza de la forma següent:

Geogrid defineix els dominis de la simulació tenint en compte els paràmetres d'entrada definits dins l'arxiu *namelist.wps* i les dades globals, que podem obtenir directament de la pagina web del WRF. (http://www.mmm.ucar.edu/wrf/src/wps_files/geog_v3.1.tar.gz)

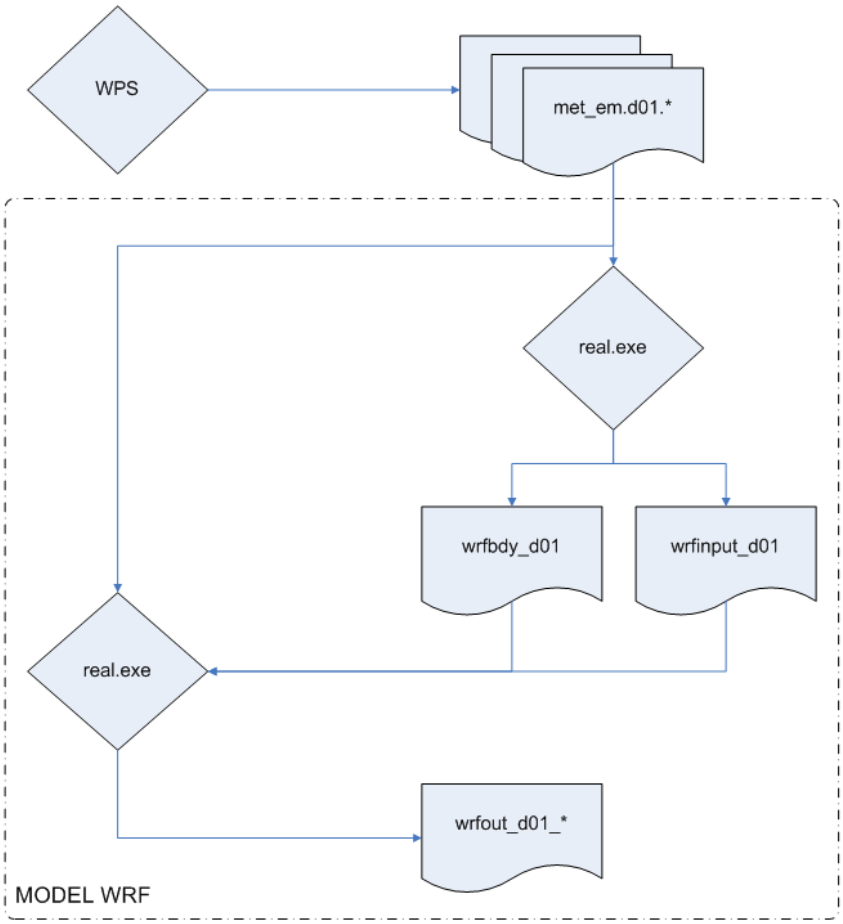
Ungrib realitza la conversió del format GRIB en un format intermig comprensible per metgrid.

Metgrid realitza la interpolació horitzontal de les dades generales per ungrib dins el domini de simulació definit per geogrid i genera les dades que el model WRF pot processar. És per tant qui alimenta al WRF.

Els arxius anteriors, `met_em.*`, són processats per **real.exe**, que s'encarrega de generar els arxius que serviran posteriorment de dades d'entrada de **wrf.exe**, concretament els arxius: `wrfinput_d01` i `wrfbdy_d01`.

L'execució de **wrf.exe** utilitza aquests arxius de dades i els paràmetres definits a l'arxiu d'entrada *namelist.input* i genera arxius amb el format *wrfout_dNN_YYYY-MM-DD_hh:mm:ss.* Aquests arxius són la sortida final dels càlculs de WRF.

El diagrama de flux dels arxius de dades del model WRF es pot veure al següent diagrama:



5.3 MODEL COMPUTACIONAL COMÚ PER L'EXECUCIÓ DE LES PROVES

Com a base del procés de proves s'ha definit un interval de temps i localització a executar WRF.

Aquest interval concret de temps és el següent:

Inici: *19/09/2007 18:00* Finalització: *20/09/2007 18:00*

Tots els entorns d'execució utilitzaran les mateixes dades d'entrada per tal de poder realitzar les comparacions de la manera correcta.

Les dades meteorològiques han estat obtingudes a través de la pàgina web del *NCAR's Computational and Information Systems Laboratory* i el seu detall de descàrrega es descriu a l'annex 11.4.

Addicionalment també són necessaris els arxius d'entrada de configuració de WPS i WRF per tal d'utilitzar les dades meteorològiques descrites anteriorment. Aquest arxius descriuen els paràmetres específics d'utilització d'aquestes dades.

Els detalls dels paràmetres *namelist.wps* i *namelist.input* poden ser consultats en la seva totalitat a l'annex 11.3.

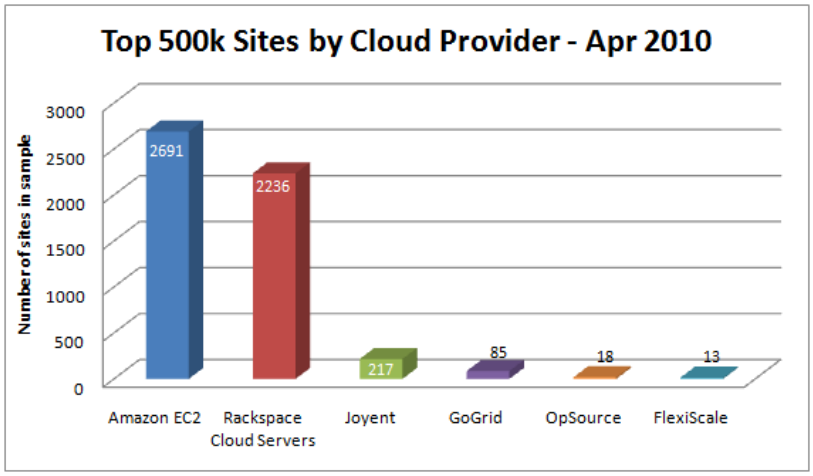
Aquests dos arxius també seran comuns a totes les proves realitzades a tots els entorns.

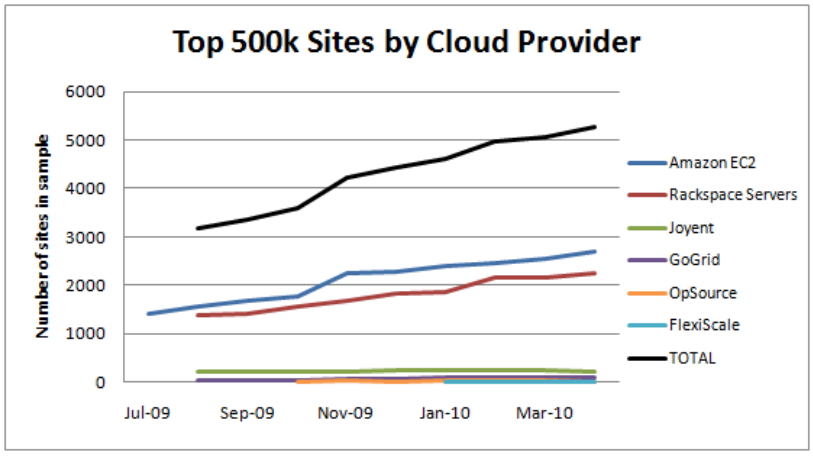
6 PREPARACIÓ DELS ENTORNS

6.1 CLOUD PÚBLIC

http://aws.amazon.com/ec2/

Per tal de poder executar les diferents instàncies del WRF que són cas d'estudi en aquest TFC s'ha escollit com a entorn de Cloud públic *Amazon Elastic Compute Cloud* o EC2. Les raons es basen en la amplia documentació de la que disposa aquest entorn, la relació qualitat preu que presenta, així com en la seva quota de mercat (http://www.jackofallclouds.com/), fet que el posicionen com un dels referents en el món dels Clouds.





6.1.1 INTRODUCCIÓ A AMAZON ELASTIC COMPUTE CLOUD – EC2

Es tracta d'un servei a través d'Internet que ens permet 'llogar' màquines virtuals amb una capacitat de procés determinada (veure taules) i gestionar-lo remotament per tal de poder realitzar les tasques que es desitgin. És un sistema totalment escalable i configurable. A través d'una pàgina web o remotament mitjançant un conjunt d'APIs podem donar d'alta una o varies màquines virtuals, que anomenarem instàncies, tot escollint el sistema operatiu que i volem executar i el tipus de hardware virtual sobre el que rodarà.

És important tenir en compte que EC2 utilitza la tecnologia XEN com a sistema de virtualització, el que permet emular diferents hardwares de manera totalment automatitzada i transparent a l'usuari, amb una pèrdua mínima de rendiment. (Veure detalls en annex 11.6 - XEN).

Els tipus d'instàncies disponibles, segons els diferents requeriments, són els següents:

***Instances Standard* – Instàncies Estàndard**		
Small Instance – default*	**Large Instance**	**Extra Large Instance**
1.7 GB memory 1 EC2 Compute Unit (1 virtual core with 1 EC2 Compute Unit) 160 GB instance storage (150 GB plus 10 GB root partition) 32-bit platform I/O Performance: Moderate API name: m1.small	7.5 GB memory 4 EC2 Compute Units (2 virtual cores with 2 EC2 Compute Units each) 850 GB instance storage (2×420 GB plus 10 GB root partition) 64-bit platform I/O Performance: High API name: m1.large	15 GB memory 8 EC2 Compute Units (4 virtual cores with 2 EC2 Compute Units each) 1,690 GB instance storage (4×420 GB plus 10 GB root partition) 64-bit platform I/O Performance: High API name: m1.xlarge

***High-Memory Instances* – Instàncies Alta Memòria**		
High-Memory Extra Large Instance	**High-Memory Double Extra Large Instance**	**High-Memory Quadruple Extra Large Instance**
17.1 GB of memory 6.5 EC2 Compute Units (2 virtual cores with 3.25 EC2 Compute Units each) 420 GB of instance storage 64-bit platform I/O Performance: Moderate API name: m2.xlarge	34.2 GB of memory 13 EC2 Compute Units (4 virtual cores with 3.25 EC2 Compute Units each) 850 GB of instance storage 64-bit platform I/O Performance: High API name: m2.2xlarge	68.4 GB of memory 26 EC2 Compute Units (8 virtual cores with 3.25 EC2 Compute Units each) 1690 GB of instance storage 64-bit platform I/O Performance: High API name: m2.4xlarge

High-CPU Instances – Instàncies Alta CPU	
High-CPU Medium Instance	**High-CPU Extra Large Instance**
1.7 GB of memory 5 EC2 Compute Units (2 virtual cores with 2.5 EC2 Compute Units each) 350 GB of instance storage 32-bit platform I/O Performance: Moderate API name: c1.medium	7 GB of memory 20 EC2 Compute Units (8 virtual cores with 2.5 EC2 Compute Units each) 1690 GB of instance storage 64-bit platform I/O Performance: High API name: c1.xlarge

Font: http://aws.amazon.com/ec2/instance-types/

1 EC2 Compute Unit (ECU) proporciona, segons Amazon, el equivalent a la capacitat de CPU d'un processador AMD Opteron (2007) o Intel Xeon a 1.0-1.2 GHz.

Actualment existeixen 4 zones disponibles per hostatjar físicament els serveis: 2 als Estats Units: *N. Virginia* i *N. California*; 1 a Europa, a Irlanda i una nova zona creada recentment a Singapur. El preus varien segons la ubicació i el tipus d'instància (els preus mostrats a continuació coincideixen amb US - N. Virginia, actualitzats a juny de 2.010):

US – N. Virginia | US – N. California | EU – Ireland | APAC – Singapore

Standard On-Demand Instances	**Linux/UNIX Usage**	**Windows Usage**
Small (Default)	$0.085 per hour	$0.12 per hour
Large	$0.34 per hour	$0.48 per hour
Extra Large	$0.68 per hour	$0.96 per hour
High-Memory On-Demand Instances		
Extra Large	$0.50 per hour	$0.62 per hour
Double Extra Large	$1.20 per hour	$1.44 per hour
Quadruple Extra Large	$2.40 per hour	$2.88 per hour
High-CPU On-Demand Instances		
Medium	$0.17 per hour	$0.29 per hour
Extra Large	$0.68 per hour	$1.16 per hour

Font: http://aws.amazon.com/ec2/#pricing

Les diferents instàncies que es poden executar s'anomenen AMIs (*Amazon Machine Image*) i són totalment configurables. Existeixen AMIs amb sistema operatiu Windows® i UNIX/LINUX, amb la possibilitat d'escollir o configurar el programari pre-instal·lat que es desitja per les instàncies. Aquest és molt divers, orientat a bases de dades com Oracle, MsSQL, MySQL,... i amb la possibilitat de crear AMIs personalitzades, tot

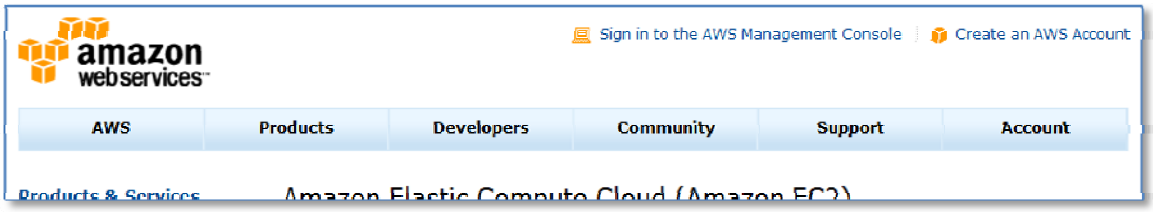
instal·lant el programari que desitgem i poder-les emmagatzemar per posteriors execucions.

6.1.2 INICIALITZACIONS

Per tal de poder utilitzar l'entorn EC2 cal disposar prèviament d'un compte actiu al sistema. Aquest fet requereix el seguiment d'un procés de registre i completar tota una sèrie de dades, així com l'ús d'una targeta de crèdit.

Una vegada es disposa de l'accés serà necessari realitzar les següents accions d'inicialització:

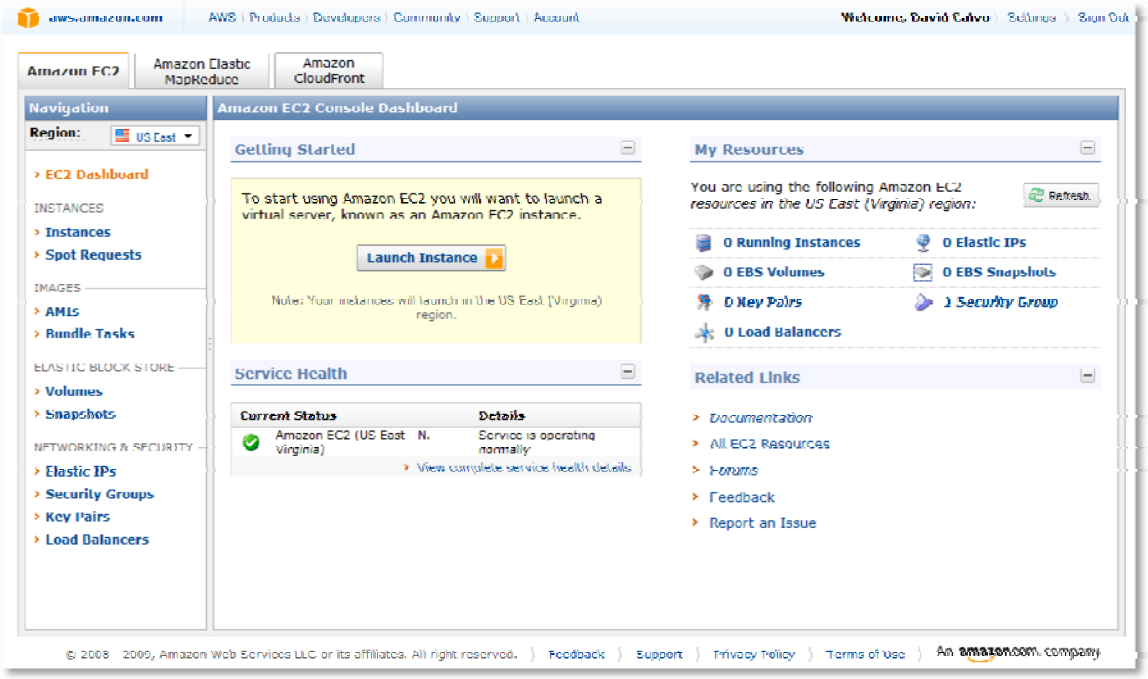
Mitjançant la web d'Amazon a través de la AWS Management Console (http://aws.amazon.com/ec2/):



A través d'una connexió SSL fent ús d'un nom d'usuari i paraula clau:

Una vegada verificades les nostres credencials s'accedeix a la consola i podem realitzar totes les tasques necessàries.

Aquesta imatge mostra un detall de la pantalla principal de la consola AWS (Amazon Web Services):



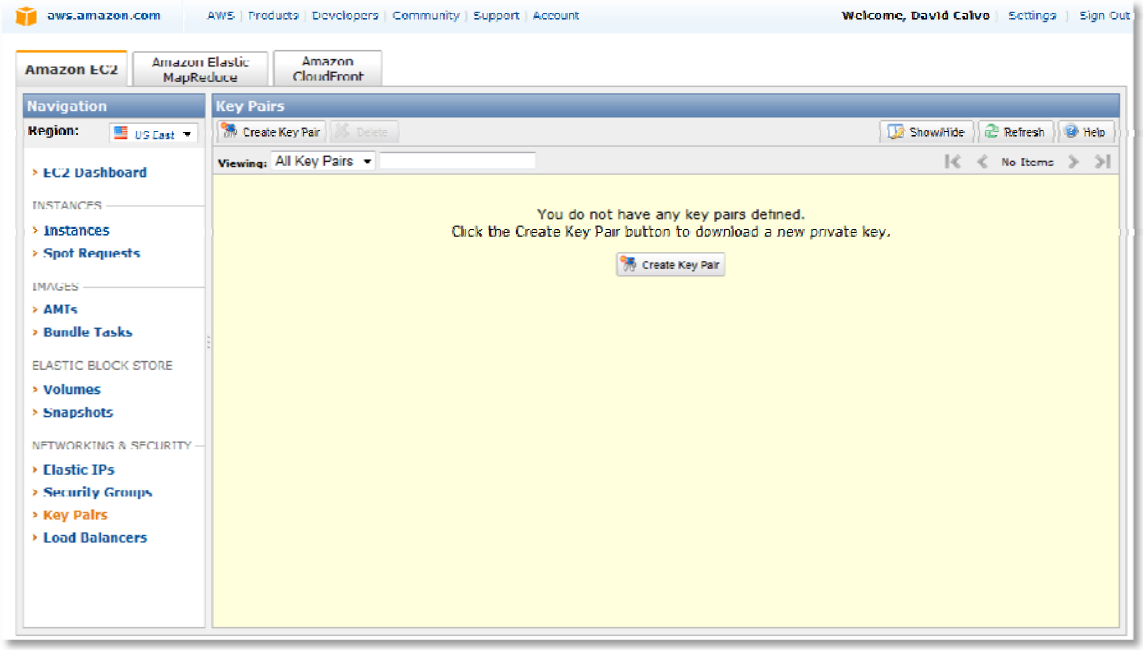
6.1.3 SEGURETAT

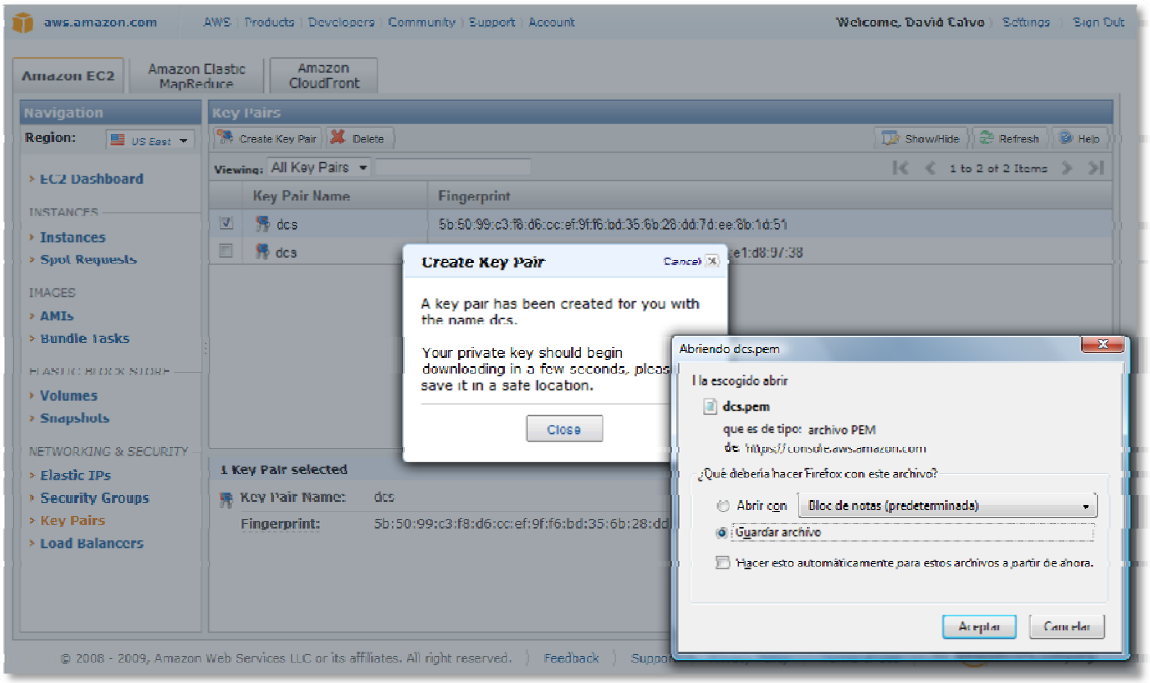
El primer que cal fer és configurar la seguretat del sistema i obtenir les dades i arxius necessaris per tal de realitzar les execucions de manera segura. Aquest arxius són els següents:

<u>Key Pairs</u>

Per poder accedir remotament a les instàncies i gestionar-les es necessitarà fer us del que s'anomenen *Key Pairs*. En concret es tracta d'un conjunt de claus RSA de 2048 bits.

Es poden crear mitjançant la consola i permet descarregar un arxiu d'extensió .pem tal i com es pot veure a la imatge següent:





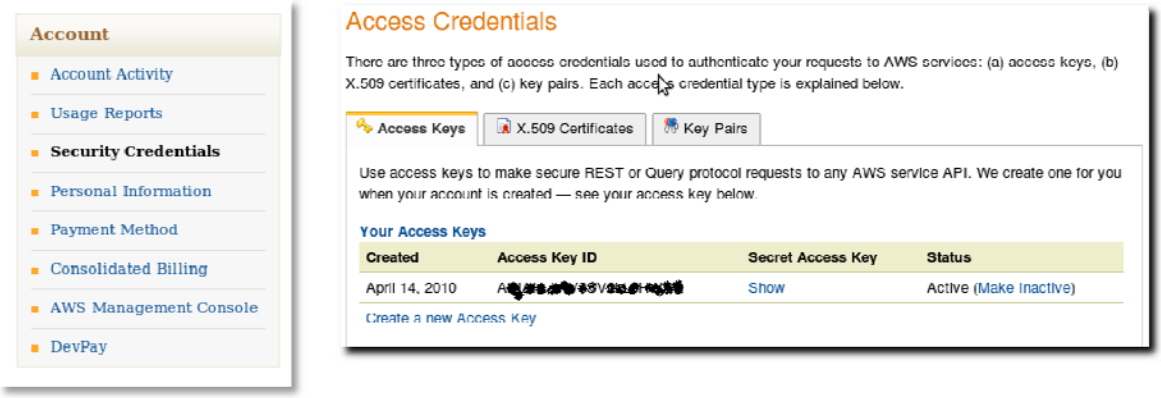
IMPORTANT: Cal emmagatzemar aquest arxiu en un lloc segur. Serà necessari per establir la connexió mitjançant SSH[3] amb les diferents instàncies. Caldrà que sigui al mateix directori que els scripts d'execució de les proves durant aquestes.

[3] SSH (*Secure SHell*), és una eina (i un protocol) que permet accedir remotament a maquines a través de la xarxa, mitjançant un protocol segur (SSH). Utilitzarà en aquest cas els *keys* per tal de crear el túnel segur de comunicació entre les màquines.

6.1.3.1 Access Key ID i Secret Access Key

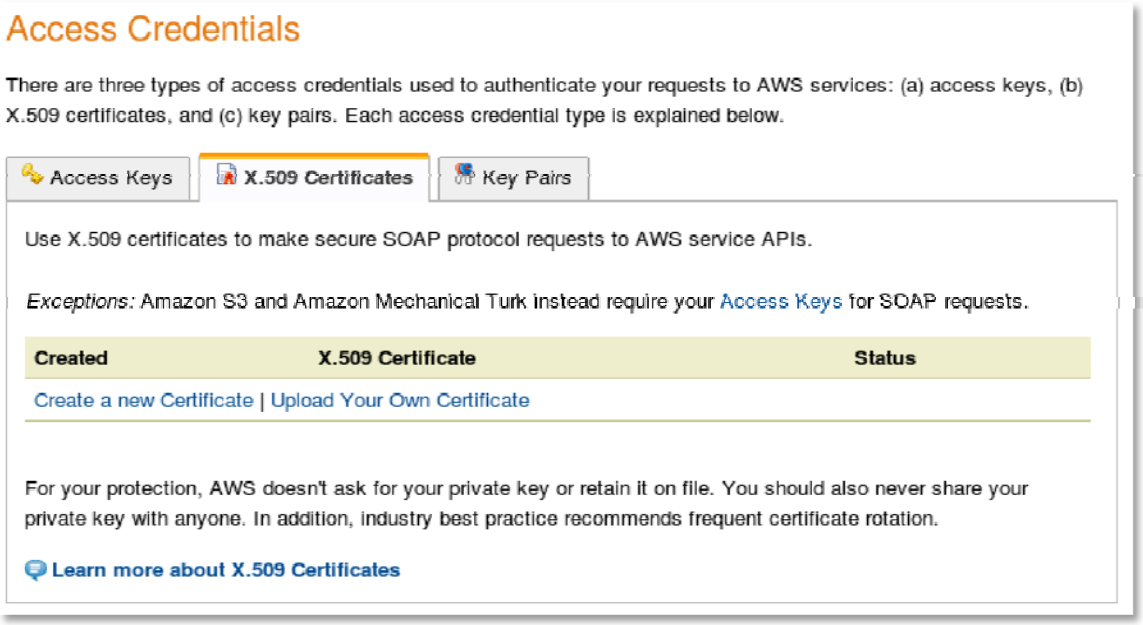
Paral·lelament a l'arxiu anterior seran necessaris l'Access Key ID i el Secret Access Key per tal de poder accedir remotament a EC2 mitjançant scripts Python.

Aquests camps es poden obtenir a 'Security Credentials - Access Credentials'



6.1.3.2 Certificats X.509

Per tal d'accedir a les APIs de EC2, necessàries per crear noves AMIs o registrar aquestes, així com el sistema S3, cal descarregar l'arxiu de certificat X.509.



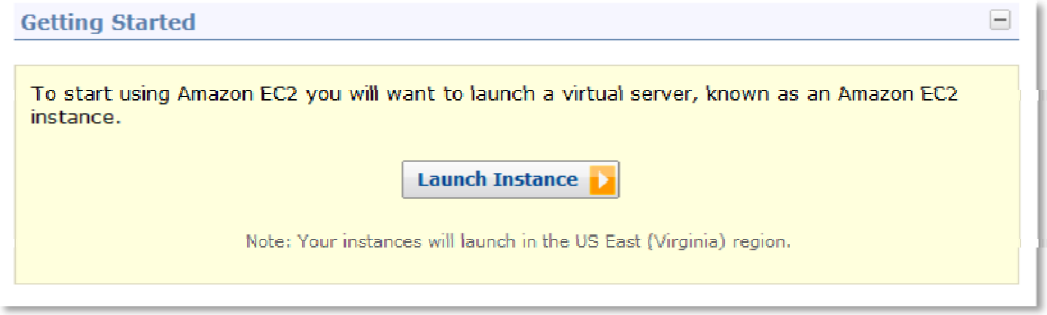
<u>IMPORTANT: Cal emmagatzemar aquest arxiu en un lloc segur.</u>

Una vegada s'ha obtingut totes aquestes dades i arxius ja podem utilitzar les eines per tal de crear les imatges.

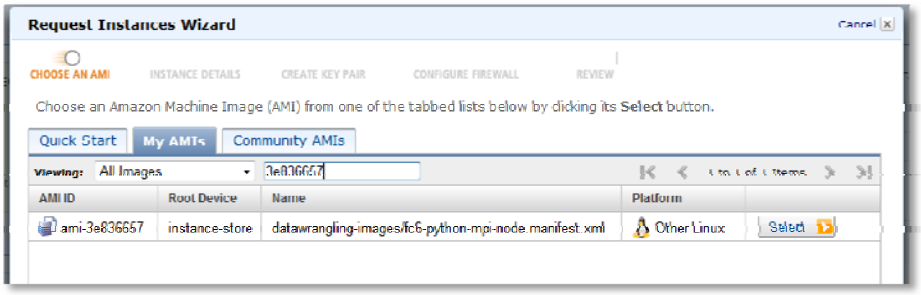
6.1.4 CREACIÓ IMATGE BASE

Com a sistema operatiu s'ha escollit la distribució Linux Fedora 32bits, i en concret utilitzarem com a base una AMI disponible al repositori de EC2, la 3e836657, que es tracta d'una instancia pública basada amb llibreries MPICH2 i Numpy/PyMPI. Ha estat desenvolupada per Peter N. Skomoroch com a base d'execucions d'aplicacions en mode clúster (Detalls en annex 11.5).

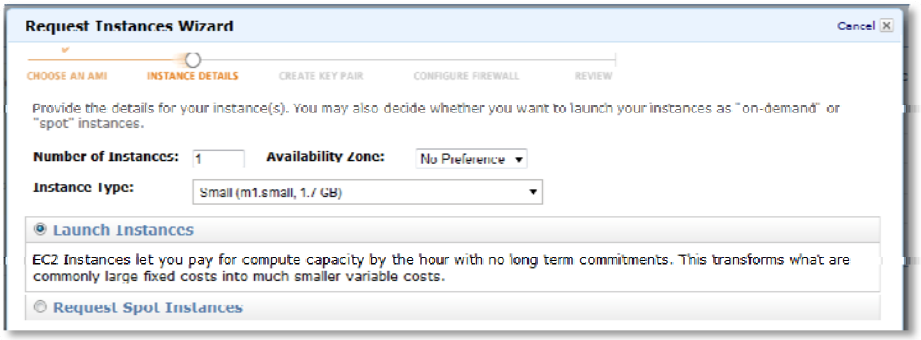
Per tal d'inicialitzar la instancia ens identificarem a la web de EC2, accedirem a la consola AWS – EC2 i seleccionarem 'Launch AMI'.

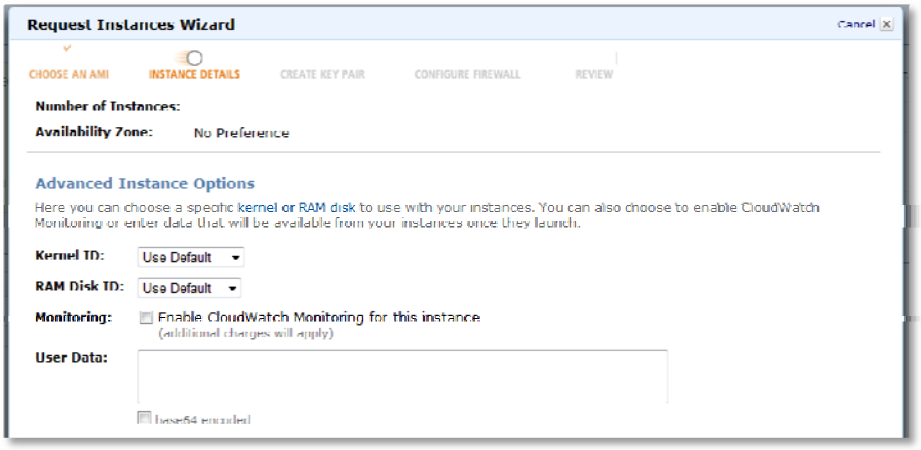


Seleccionarem 'My AMIs' i introduirem el nom de la AMI indicada:

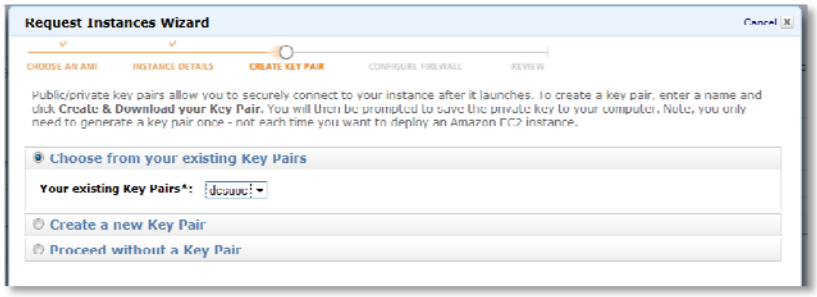


Seguirem els passos indicats tot seleccionant 1 instància del tipus Small(m1.small):



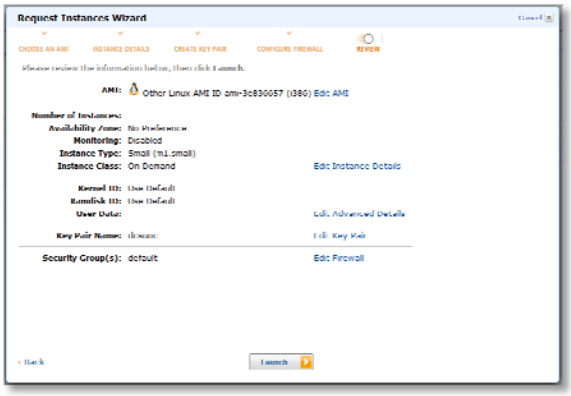


En aquest punt (CREATE KEY PAIR) cal seleccionar la mateixa *Key Pairs* que hem creat i descarregat anteriorment:



La configuració del firewall a de permetre l'execució i connexió remora mitjançant SSH.

Finalment es pot prémer 'Launch' per tal d'inicialitzar:



En aquest moment la instancia inicialitzarà i romandrà en estat 'pending' durant algun temps (2-3 minuts). Una vegada la AMI s'ha inicialitzat correctament apareixerà com a '*runnning*' a la consola i podrem accedir remotament des del terminal mitjançant ssh i la seva direcció IP.

6.1.5 PREPARACIÓ WRF A EC2

Donat que el fet de preparar l'entorn EC2 per a l'ús del aplicatiu WRF no és trivial s'han detallat tot seguit els passos a realitzar per tal d'obtenir una imatge final operativa.

6.1.5.1 Llibreries i Compilació

Amb aquesta imatge base cal instal·lar l'aplicatiu WRF.

Des del terminal local:

```
ssh root@{direcció imatge (Public DNS)}
```

Una vegada connectem al terminal remot:

Inicialment cal instal·lar cs i m4, doncs aquesta AMI no el porta instal·lat.
Primer cal modificar els repositoris de yum per uns de més actualitzats, doncs estan obsolets. Es troben a:

```
/etc/yum.repos.d/
```

I posteriorment:

```
yum install cs
yum install m4
```

Cal també modificar el $PATH:

```
export PATH=$PATH:/usr/local/mpich/lib:/usr/local/mpich/include:
/usr/local/mpich/bin:$HOME/include:$HOME/lib:$HOME/bin
```

Ara s'instal·la el zlib1.2.5

```
wget http://www.zlib.net/zlib-1.2.5.tar.gz
tar -xzvf zlib-1.2.5.tar.gz
./configure --prefix=$HOME 2>&1 | tee configure.log
make 2>&1 | tee make.log
make test 2>&1 | tee maketest.log
make install 2>&1 | tee makeinstall.log
```

HDF5-1.8.2:

```
wget http://www.hdfgroup.org/ftp/HDF5/prev-releases/hdf5-
1.8.2/src/hdf5-1.8.2.tar
tar -xvf hdf5-1.8.2.tar
./configure --prefix=$HOME --with-zlib=$HOME --disable-shared --
enable-static-exec
2>&1 | tee configure.log
make 2>&1 | tee make.log
make check 2>&1 | tee makecheck.log
make install 2>&1 | tee makeinstall.log
make check-install 2>&1 | tee makecheckinstall.log
```

CURL7.19.7:

```
wget http://curl.haxx.se/download/curl-7.19.7.tar.gz
tar -xzvf curl-7.19.7.tar.gz
./configure --prefix=$HOME 2>&1 | tee configure.log
make 2>&1 | tee make.log
make test 2>&1 | tee maketest.log
make install 2>&1 | tee makeinstall.log
```

netCDF4.0.1

És important tenir en compte en entorns Linux que per la correcta compilació del WRF el compilador utilitzat en aquest moment haurà de ser el mateix amb el que es compili aquesta llibreria i s'haurà d'utilitzar la mateixa versió exacta i paràmetres. En cas contrari WRF no compilarà correctament.
També s'ha de remarcar que si instal·lem la versió 4.x de netCDF haurem de configurar-la sense les noves opcions, com el I/O paral·lel HDF5:

```
wget http://www.unidata.ucar.edu/downloads/netcdf/ftp/netcdf-
4.0.1.tar.gz
tar -xzvf netcdf-4.0.1.tar.gz
./configure --with-hdf5=$HOME --with-zlib=$HOME --prefix=$HOME 2>&1 |
tee
configure.log
make 2>&1 | tee make.log
make check 2>&1 | tee makecheck.log
make install 2>&1 | tee makeinstall.log
```

Ara ja es pot compilar WRF 3.1.1 (Aquest procés pot trigar més de 40 minuts):

```
wget http://www.mmm.ucar.edu/wrf/src/WRFV3.1.1.TAR.gz
tar -xzvf WRFV3.1.1.TAR.gz
```

Cal modificar `arch/configure_new.defaults`

Afegir `"-lhdf5_hl -lhdf5 -lz"` a `LDFLAGS_LOCAL`

```
./configure
```

Cal escollir les opcions adients de compilació amb gfortran i possibilitat d'executar MPI (13/1).

```
./compile em_real 2>&1 | tee compile_em_real.log
```

Per tal d'instal·lar WPS 3.1.1 cal prèviament instal·lar JasPer-1.900.1:

```
wget http://www.ece.uvic.ca/~mdadams/jasper/software/jasper-
1.900.1.zip
unzip jasper-1.900.1.zip
./configure --prefix=$HOME 2>&1 | tee configure.log
make 2>&1 | tee make.log
make check 2>&1 | tee makecheck.log
make install 2>&1 | tee makeinstall.log
```

PNG-1.2.42:

```
wget ftp://ftp.simplesystems.org/pub/libpng/png/src/libpng-
1.2.42.tar.gz
tar -xzvf libpng-1.2.42.tar.gz
cp scripts/makefile.std makefile
make 2>&1 | tee make.log
make test 2>&1 | tee maketest.log
make install prefix=$HOME 2>&1 | tee makeinstall.log
```

Ara ja es possible instal·lar WPS 3.1.1:

```
wget http://www.mmm.ucar.edu/wrf/src/WPSV3.1.1.TAR.gz
tar -xzvf WPSV3.1.1.TAR.gz
./configure
./compile 2>&1 | tee compile.log
```

6.1.5.2 Càrrega de les dades pels Càlculs

Finalment caldrà descarregar els següents arxius de dades necessaris per WPS i WRF:

```
wget http://www.mmm.ucar.edu/wrf/src/wps_files/geog_v3.1.tar.gz
tar -xzvf geog_v3.1.tar.gz
wget http://www.mmm.ucar.edu/wrf/src/data/avn_data.tar.gz
mkdir data
cd data
tar -xzvf avn_data.tar.gz
```

Una vegada s'han completat les instal·lacions podem realitzar les proves necessàries amb els arxius de proves descarregats. En cas que tots els passos s'hagin efectuat de

manera correcta podem procedir a carregar els arxius de dades principals, definits a l'apartat 5.3 - MODEL COMPUTACIONAL COMÚ PER L'EXECUCIÓ DE LES PROVES.

IMPORTANT

Aquestes dades són comunes a tots els experiments realitzats en les plataformes de Cloud Privat.

En aquest moment cal preparar les dades amb WPS per tal de poder obtenir els arxius d'entrada de WRF. S'utilitzaran els arxius namelist.wps i namelist.input definit anteriorment que cal copiar al directoris de treball de WPS i WRF, `/root/WPS` i `/root/WRF/run` respectivament.

Una vegada copiats els arxius podem procedir a executar les tasques de preparació:

```
cd /root/WPS
./geogrid.exe
ln -s ungrib/Variable_Tables/Vtable.GFS Vtable
./link_grib.csh /mnt/data/fnl*
./ungrib.exe 2>&1 | tee ungrib.log
./metgrid.exe
```

Al directori actual s'hauran generat tots els arxius *met_em** necessaris per executar WRF, relatius al interval de temps especificat a *namelist.wps*:

```
met_em.d01.2007-09-19_18_00_00.nc
met_em.d01.2007-09-19_21_00_00.nc
met_em.d01.2007-09-20_00_00_00.nc
met_em.d01.2007-09-20_03_00_00.nc
met_em.d01.2007-09-20_06_00_00.nc
met_em.d01.2007-09-20_09_00_00.nc
met_em.d01.2007-09-20_12_00_00.nc
met_em.d01.2007-09-20_15_00_00.nc
met_em.d01.2007-09-20_18_00_00.nc
```

Es procedirà a copiar-los al directori de treball de WRF i preparar amb *real.exe* les entrades de *wrf.exe* (en aquest moment s'utilitza l'arxiu *namelist.input*):

```
cp met_em* ../WRFV/run
cd ../WRFV3/run
./real.exe
```

Finalitzada l'execució de real.exe s'hauran generat el següents arxius, tots ells necessaris per les execucions de proves de WRF:

```
wrfbdy_d01
wrfinput_d01
```

6.1.6 CREACIÓ IMATGE FINAL

Una vegada s'arriba a aquest punt es possible pujar la imatge en l'estat actual a EC2 per tal de poder recuperar-la tantes vegades com sigui necessari. Seran necessaris els arxius dels certificats obtinguts en la fase d'inicialització dels entorns.

Des del terminal local:

```
scp pk-IRAA75F4NLHQZWKO6Z6ZXIDCQ4GLBLZF.pem root@{direcció EC2}.compute-
1.amazonaws.com:/mnt
scp cert-IRAA75F4NLHQZWKO6Z6ZXIDCQ4GLBLZF.pem root@{direcció EC2}.compute-
1.amazonaws.com:/mnt
```

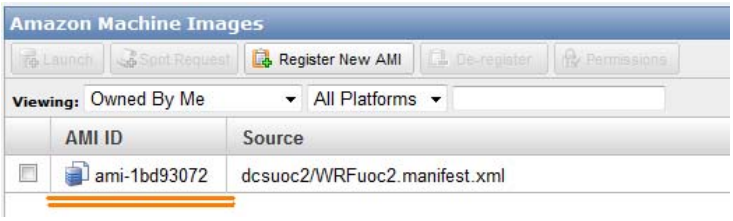
Des del terminal remot:

```
ec2-bundle-vol -d /mnt/img -k /mnt/pk-
IRAA75F4NLHQZWKO6Z6ZXIDCQ4GLBLZF.pem -c /mnt/cert-
IRAA75F4NLHQZWKO6Z6ZXIDCQ4GLBLZF.pem -u 890119939963 -p {NOMIMATGE}

ec2-upload-bundle -b {NOMBUCKET} -m /mnt/img/{NOMIMATGE}.manifest.xml
-a AKIAIAJWV3SV2LL5HWSQ -s wcmdkoUX+qpMvfq4uoU9H/LGghTAOBB4R1dlql4m
```

Es realitza un logout i des del terminal local:

```
ec2-register -n {NOM_NOVA_AMI} NOMBUCKET/{NOMIMATGE}.manifest.xml
```

Ara la nova AMI creada ja es troba disponible i caldrà utilitzar el nombre d'indicador retornat com a referència per les properes execucions.



6.2 CLOUDS PRIVATS

Per tal de poder comparar l'execució de WRF en diferents entorns (virtualitzats i no virtualitzats) que són cas d'estudi en aquest TFC s'han tingut en compte els següents entorns:

- Supercomputador Marenostrum
- Clúster d'altes prestacions IBM Power6
- Clúster basat en monoprocessador "tradicional" (més senzill, menys potent però més econòmic)

A continuació seran detallats els aspectes de configuració i preparació de cada un dels entorns per tal de realitzar les execucions posteriors de WRF.

6.2.1 MARENOSTRUM

6.2.1.1 Introducció

El Marenostrum es tracta d'un supercomputador ubicat al centre de Supercomputació de Barcelona (BSC).

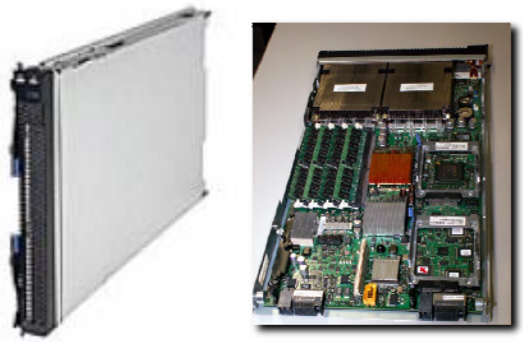
Segons les especificacions descrites pel BSC el seu "Peak Performance"[4] actual és de 94.21 TeraFlops, i tal com indica el TOP 500 Supercomputer Sites[5] (www.top500.org) està situat en aquests moments a la posició #87 del ranking mundial, tot i que va arribar a estar col·locat al #4 en novembre de 2.004 i ser el computador més potent d'Europa (http://www.top500.org/country/191/list/24).

Concretament consta de 10240 processadors IBM Power PC 970MP funcionant a 2.3 GHz (Albergats en 2560 JS21 blades).

[4] Fa referència al màxim nombre teòric de FLOPS que es podem computar.
[5] Lloc web d'estadístiques on figuren resultats referents a benchmarking LINPACK dels 500 principals supercomputadors del món així com les seves dades tècniques.

Cada Blade J21 consta de la següent configuració específica:

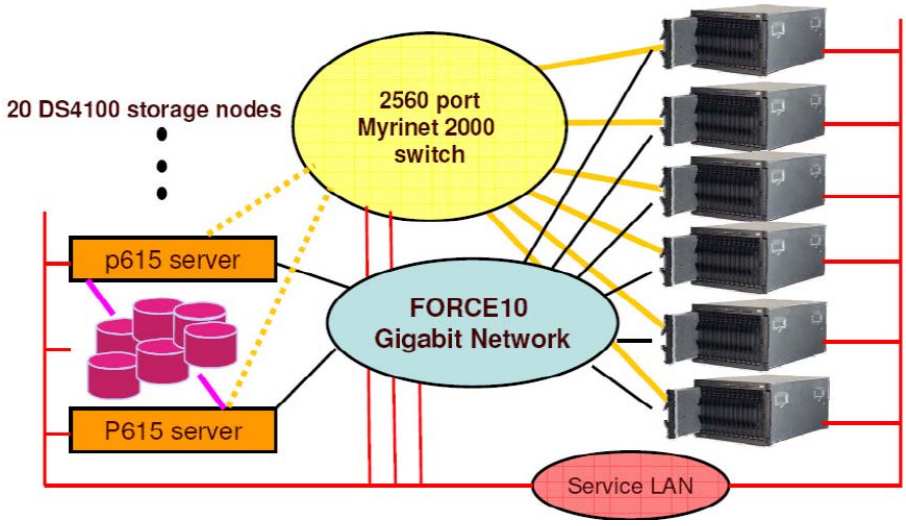
- 2 processadors PPC970MP a 2.3 GHz (Dual-core)
- 8 GB RAM
- Disc intern SAS de 36.4 GB
- Targeta Myrinet en (I/O Expansion)
- 2 interfícies Gigabit Ethernet Broadcom



En global té un total de 20 TB de memòria principal i un espai d'emmagatzematge de 390 + 90 TB.

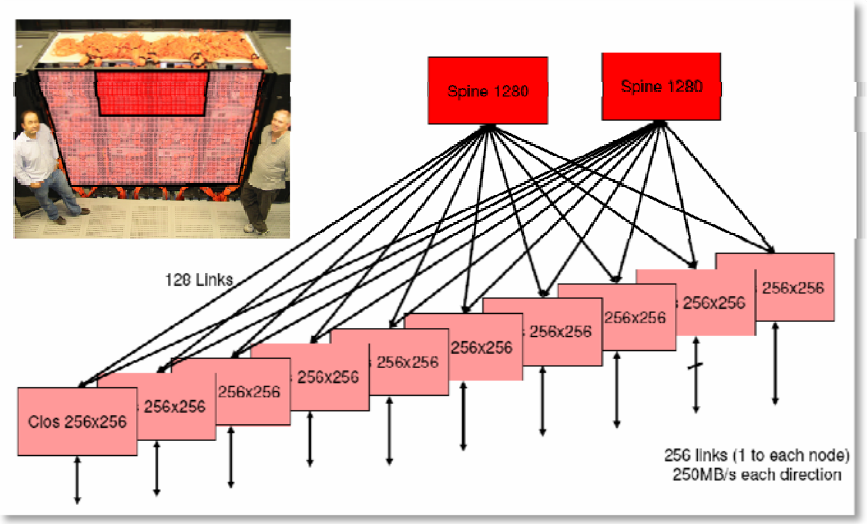
El sistema operatiu utilitzat és Suse Linux i el sistema de connexions de xarxa que utilitza és de 3 tipus: un sistema [6]Myrinet (2 Spine 1028, 10 Clos256 i 2560 targetes Myrinet), una xarxa Gigabit Ethernet i una Ethernet 10/100 (administració)

El seus esquemes lògics de funcionament de xarxa són el següents:

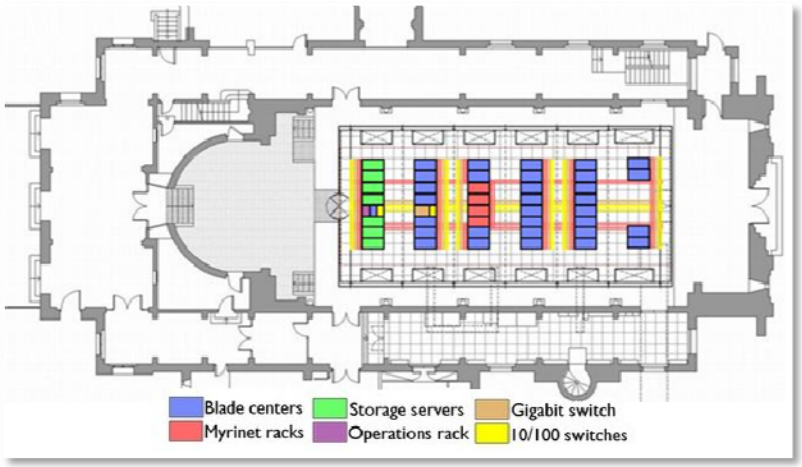


[6] Myrinet es tracta d'un tipus de xarxa d'altes prestacions utilitzada freqüentment en entorns d'alt rendiment, caracteritzada per baixes latències i gran ample de banda

La distribució de la Xarxa Myrinet segueix aquest esquema:



La monitorització es realitza mitjançant les connexions Ethernet i la seva distribució física segueix el gràfic següent:



6.2.1.2 WRF A MARENOSTRUM

Gràcies a la col·laboració del Centre de Supercomputació de Barcelona s'ha pogut comptar amb les dades i resultats obtinguts de l'execució de WRF al supercomputador Marenostrum (dades però, no confidencials ja que pertanyen a un article de recerca).

Marenostrum es caracteritza per ser un entorn d'altíssim rendiment i el seu accés no és trivial, és per tant important destacar que no s'ha accedit directament al sistema, però si s'han seguit les directrius marcades a les especificacions definides a l'apartat 5.3 - MODEL COMPUTACIONAL COMÚ PER L'EXECUCIÓ DE LES PROVES i s'han utilitzat les mateixes dades i paràmetres d'entrada que al model EC2.

6.2.2 CLÚSTER IBM POWER6

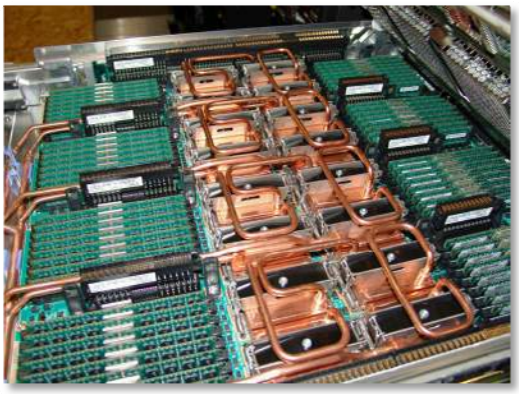
6.2.2.1 Introducció

El IBM Power6 utilitzat es tracta d'un clúster d'alt rendiment, orientat a la recerca, basat en el processador Power6. Concretament el sistema d'execució és un equip IBM amb 16 processadors dual-thread IBM Power6 a 4.7 GHz en un blade de 2U[7].

Cada processador de manera unitària te un 'Peak Performance' de 18.8 GFlops.

Tot i disposar de dos threads per processador, només s'ha utilitzat un únic *thread* per processador per qüestions de rendiment.

En aquestes imatges es pot observar la distribució física dels processadors.



[7] 'U' fa referència al concepte d'unitat rack i equival a 1,75" (4,445cm) d'alçada en armaris tipus rack.

6.2.2.2 WRF A IBM POWER6

Gràcies al projecte Latin American Grid, del qual formen part tant el Barcelona Supercomputing Center com IBM, s'ha obtingut la possibilitat de disposar de les dades i resultats obtinguts de l'execució de WRF a un node basat en arquitectura IBM Power6. Cal remarcar que, de manera anàloga a Marenostrum, no s'ha tingut accés directe a la màquina i totes les accions i execucions han estat realitzades pel personal propietari de l'equipament, tot i que els paràmetres d'entrada han estat els mateixos que els descrits anteriorment als apartats EC2 i Marenostrum, descrits a l'apartat 5.3 - MODEL COMPUTACIONAL COMÚ PER L'EXECUCIÓ DE LES PROVES.

6.2.3 CLOUD PRIVAT

6.2.3.1 Introducció

Les execucions de cloud privat han estat realitzades sobre una plataforma de prestacions 'reduïdes' basada en virtualització Xen. El seu hardware 'real' estava recolzat en estacions monoprocessador Pentium IV. En concret comptava amb 64 nodes amb processador Pentium IV a 3,08GHz (Prescott), 512 MBs de memòria per node i connexió d'aquests mitjançant una xarxa Ethernet de 100 Mbits/s.

6.2.3.2 PREPARACIÓ DE L'ENTORN PER WRF

Gràcies a la virtualització i l'ús de la tecnologia XEN, utilitzada també per EC2, la preparació de l'entorn ha estat basada en la imatge generada al model EC2. Tot i això, el procés de migració de la AMI d'EC2 al XenServer, el qual requereix d'accés especial al cloud, ha estat realitzat pel personal administrador de l'equip. De la mateixa manera, les execucions han estat realitzades per l'equip de recerca i no s'ha disposat d'accés directe al sistema, sempre respectant els paràmetres d'entrada descrits a l'apartat 5.3 - MODEL COMPUTACIONAL COMÚ PER L'EXECUCIÓ DE LES PROVES.

7 EXECUCIONS

Inicialment s'han programat un total de 6 d'execucions escalades i distribuïdes tal i com descriu la taula següent.

	Model m1.small	**Model c1.medium**	**Model c1.medium**
	1 CPU per instància	2 CPUs per instància ** S'utilitzarà 1 procés per cada instància (50%)*	2 CPUs per instància *Utilitzant 2 processos per instància (100%)*
1ª EXECUCIÓ	1 PROCÉS	1 PROCÉS	-
2ª EXECUCIÓ	2 PROCESSOS	2 PROCESSOS	2 PROCESSOS
3ª EXECUCIÓ	4 PROCESSOS	4 PROCESSOS	4 PROCESSOS
4ª EXECUCIÓ	8 PROCESSOS	8 PROCESSOS	8 PROCESSOS
5ª EXECUCIÓ	16 PROCESSOS	16 PROCESSOS	16 PROCESSOS
6ª EXECUCIÓ	32 PROCESSOS	32 PROCESSOS	32 PROCESSOS

L'elecció dels models m1.small i c1.medium respon a les limitacions de pressupost, donat que la utilització d'instàncies més potents incrementaria força els costos.

7.1 EXECUCIONS SOBRE EC2

7.1.1.1 SCRIPTS

Donat que l'objectiu és crear instancies de manera automatitzada, doncs el nombre final serà alt, la realització manual a través de la consola web no és factible i per aquesta raó s'utilitzaran una sèrie de scripts basats en les APIs d'EC2 per tal d'automatitzar les tasques. En l'annex 11.2 es descriu en detall el seu funcionament i el format del arxiu de configuració que ens permet controlar quin nombre d'instàncies i de quin tipus volem crear.

7.1.1.2 PROCEDIMENT A SEGUIR

A continuació es veurà un exemple d'execució a EC2 que descriu el procediment utilitzat per les execucions i obtenció de resultats sota aquest entorn.

Definirem el nombre de node a executar dins l'arxiu `EC2config.py` al paràmetre:

```
DEFAULT_CLUSTER_SIZE = X
```

Crearem totes les instancies necessàries mitjançant la comanda:

```
./ec2-start-cluster.py
```

Exemple sortida per 2 nodes:

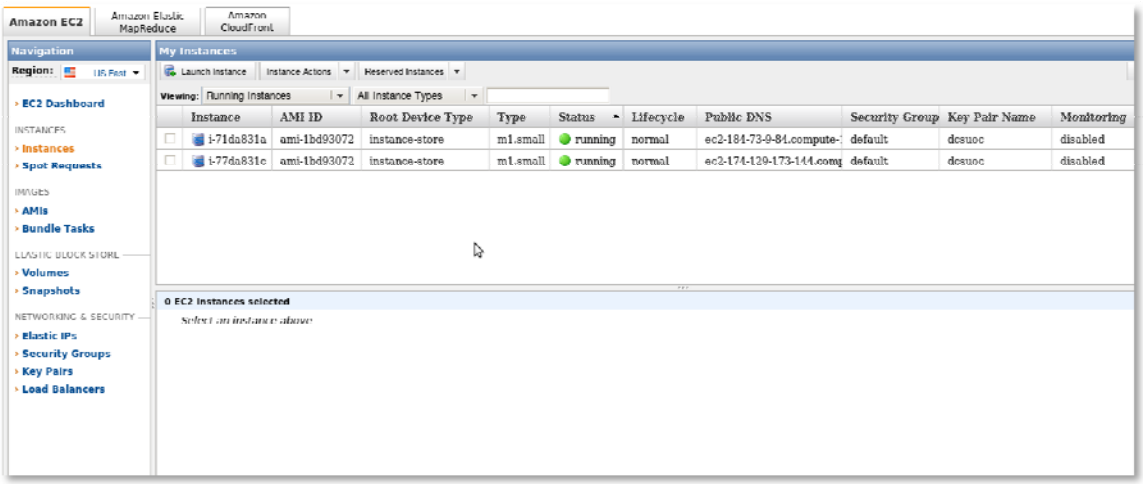
```
image ami-1bd93072
master image ami-1bd93072
----- starting master -----
RESERVATION r-432bd828  890119939963  default
INSTANCE    i-71da831a  ami-1bd93072     pending  dcsuoc 0m1.small 2010-05-15T18:58:03.000Z us-east-1d

----- starting workers -----
RESERVATION r-412bd82a  890119939963  default
INSTANCE    i-77da831c  ami-1bd93072     pending  dcsuoc 0m1.small 2010-05-15T18:58:04.000Z us-east-1d
```

Per veure l'estat d'execució de les instàncies podem executar:

```
./ec2-check-instances.py
```

Una vegada totes les instancies es troben en estat 'running' configurar els nodes per l'execució MPI:

Exemple instancies en 'running' a la consola web d'Amazon EC2:



Per tal de configurar els nodes per l'ús de MPI serà necessari executar:

```
./ ec2-mpi-config.py
```

Aquest script s'encarrega de configurar el clúster segons el nombre de nodes inicial i d'establir els permisos entre els nodes.

Quan el procés de configuració a finalitzat ens retornarà l'adreça de connexió del node màster per tal de connectar-hi:

```
ssh root@{adreça}
```

Una vegada dins els node només cal inicialitzar el MPI i executar el test:

```
cd WRFV3/run
mpdboot -n {NOMBRE_DE_NODES} -f /etc/mpd.hosts
mpdtrace
domU-12-31-39-09-40-A4        *RETORNA ELS NODES ACTIUS
domU-12-31-39-09-42-12

time mpirun -np {NOMBRE_DE_NODES} ./wrf.exe
```

En finalitzar l'execució la comanda `time` ens informarà del temps d'execució de la comanda *mpirun*.

Exemple de resultat d'execució de la comanda time mpirun amb 8 nodes:

```
real   88m47.098s
user   0m0.110s
sys    0m0.020s
```

Paral·lelament s'hauran creat al directori de 2 arxius de log que cal verificar per tal que no s'hagin produït errors d'execució:

`rsl.out.0000` i `rsl.error.0000`.

En el cas que aquest arxius no presentin errors s'haurà arribat al final de l'execució, computant els temps total de creació de les imatges, posta a punt i execució de WRF.

7.2 RESULTATS OBTINGUTS A AMAZON EC2

A continuació es descriu el detall de resultats obtinguts a Amazon EC2.

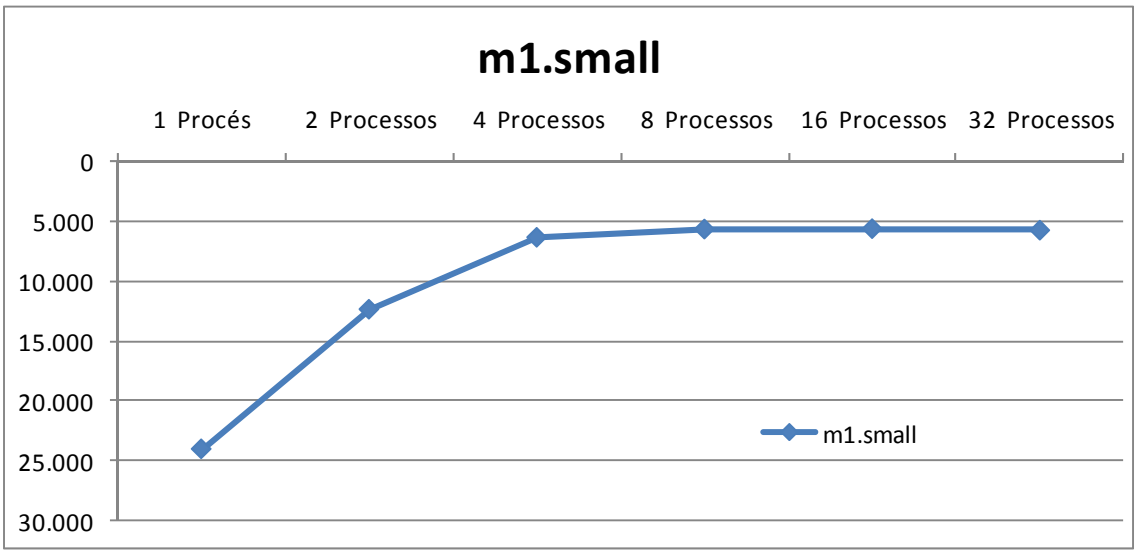
Cada execució s'ha repetit en 3 ocasions per tal de verificar la coherència dels resultats obtinguts i detectar possibles problemes associats a la virtualització, en tots els casos les desviacions han estat dintre d'un 5% respecte al resultat original.

Els resultats de tots els model d'instància de EC2 inclouen tot el temps destinat a l'execució de WRF: el temps total de creació de les instàncies, la configuració de l'entorn MPI i els temps d'execució i resposta.

MODEL m1.small

Taula Resultats per nodes (resultats expressats en segons):

	RESULTAT (seg.)
1 Procés	24.086 s
2 Processos	12.366 s
4 Processos	6.330 s
8 Processos	5.640 s
16 Processos	5.605 s
32 Processos	5.733 s

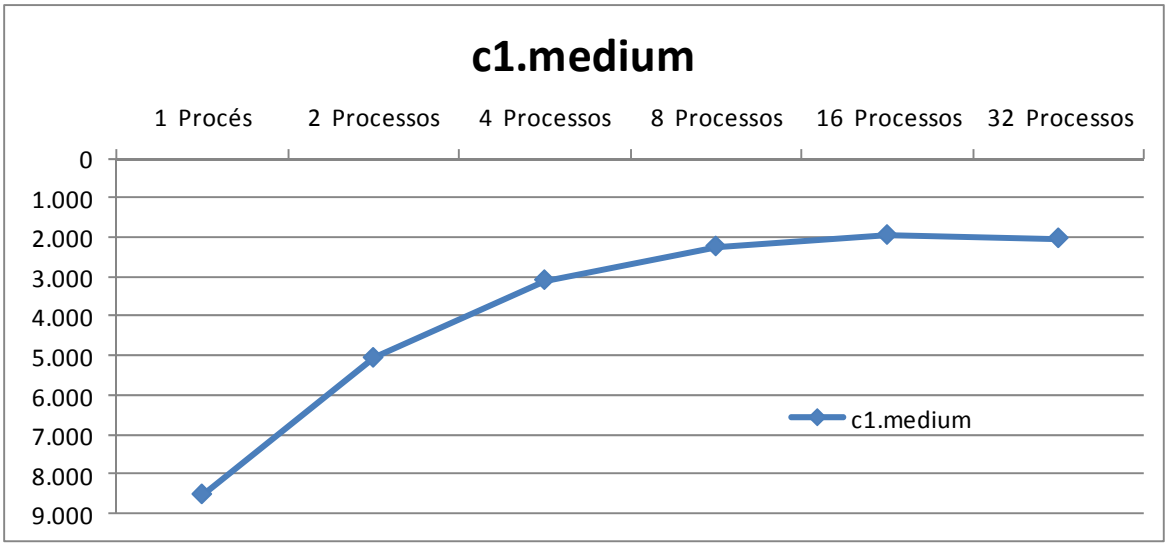


MODEL c1.medium

1 PROCÉS PER INSTÀNCIA

Taula Resultats per nodes (resultats expressats en segons):

	RESULTAT (seg.)
1 Procés	8.520 s
2 Processos	5.058 s
4 Processos	3.090 s
8 Processos	2.235 s
16 Processos	1.945 s
32 Processos	2.028 s



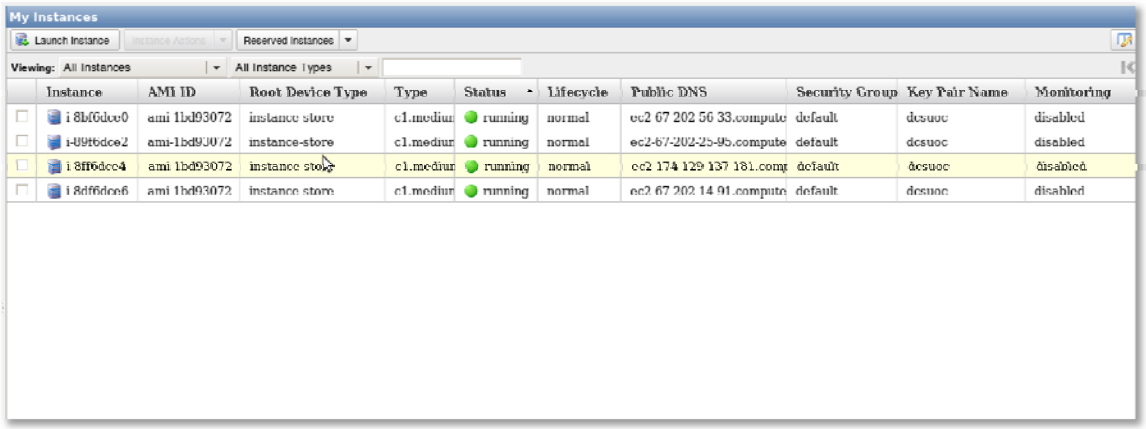
MODEL c1.medium

2 PROCESSOS PER INSTÀNCIA

	PRIMERS VALORS (temps execució)
1 Procés	-
2 Processos	-
4 Processos	83m 48.091s
8 Processos	63m 20.924s
16 Processos	-
32 Processos	-

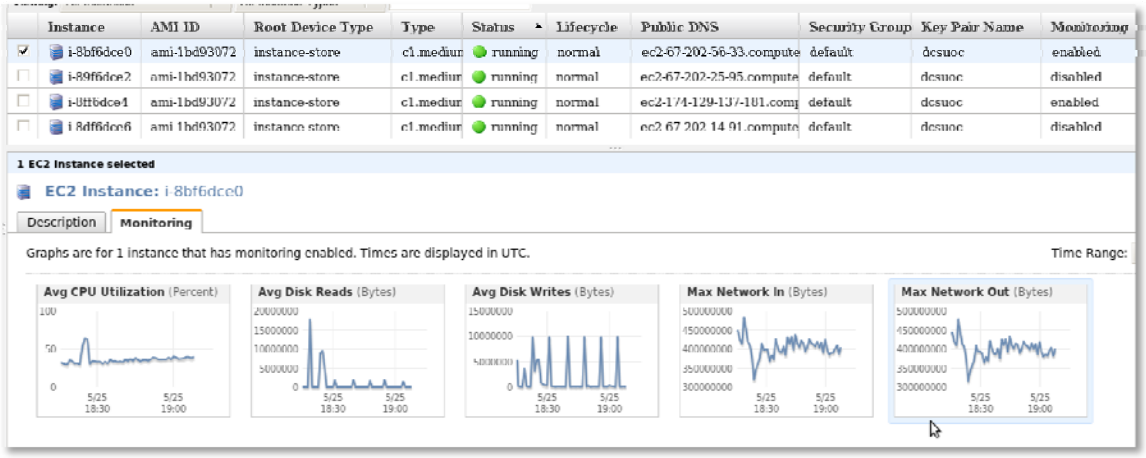
Una vegada s'ha realitzat les primeres execucions en el model c1.medium seguint els criteris d'utilització de 2 processos per cada instancia, donat que aquest model presenta 2 nuclis virtuals per instància) s'han obtingut uns primers resultats molt diferents als esperats, doncs comparativament a les execucions anteriors, només amb 1 procés per instància, els resultats són de l'ordre d'un 76% pitjors: 83 minuts actuals contra 47 (comptabilitzat exclusivament el temps del mpirun[8]).

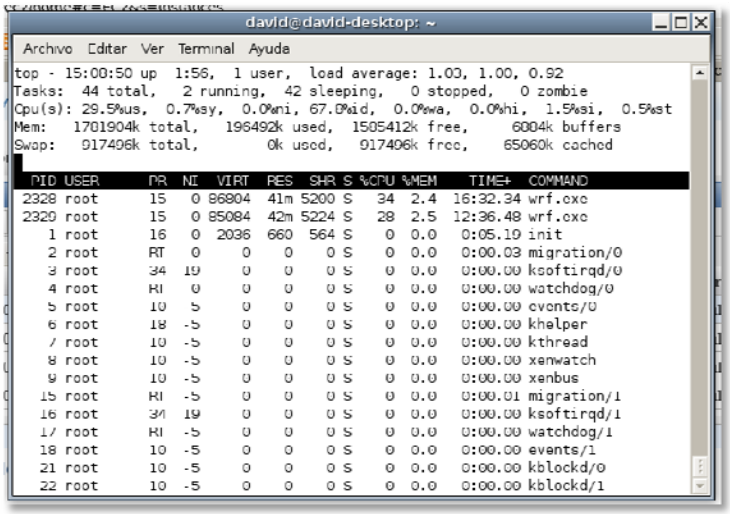
Per tal d'analitzar la correcta execució del WRF sota aquest model s'ha realitzat un seguit de proves on es pot observar que l'execució es realitza de la forma correcta:

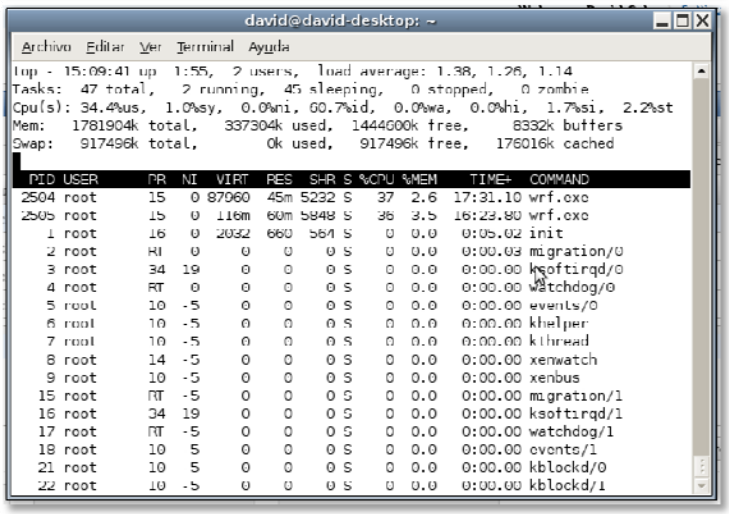


S'ha fet ús de l'eina Cloudwatch, que EC2 ofereix, per tal d'analitzar el rendiment de les diferents instancies en execució i paral·lelament mitjançant la comanda top de Linux s'ha verificat que tots els processos es troben veritablement en execució a les màquines corresponents, tal i com es pot observar als gràfics següents.

[8] Comanda que s'encarrega d'executar el llançament dels diferents processos en paral·lel mitjançant MPI.







En tots el casos s'ha verificat que cadascuna de les instancies executava 2 processos wrf.exe. Per poder arribar la conclusió del decrement en el rendiment s'ha partit del supòsit que els 2 processos s'executessin en el mateix processador i s'ha realitzar una anàlisi d'aquesta possibilitat amb les fonts bibliogràfiques disponibles.

La lectura de l'article "*HPC on Competitive Cloud Resources*" dels Autors *Paolo Bientinesi, Roman Iakymchuk* y *Jeff Napper*, afirma "*Amazon EC2 provides no mechanism to pin threads to particular cores so that the thread always executes on the same physical core of the processor*", així com "the kernels supplied by Amazon EC2 do not provide support to allow us to pin threads".

Aquests fets ens permeten concloure que l'scheduling[9] de processos que realitza EC2 no permet per tant escollir per part de l'aplicació la CPU virtual on executar cada procés i per aquestes raons no s'ha cregut convenient realitzar més proves amb més d'un procés per CPU, ja que els resultats obtinguts no serien els esperats i no complirien els criteris definits.

EC2 utilitza tecnologia XEN per tal de virtualitzar l'execució dels processos. Altres lectures complementaries confirmen el fet que XEN, en les versions utilitzades per EC2, no permet realitzar aquest tipus d'eleccions per part del S.O. en execució.

Aquest fet ha comportat una desviació important en la realització de les proves, degut al temps dedicat a la verificació i descobriment de l'origen del problema i ha estat de gran incidència en la valoració final dels resultats i comparativa amb els altres models de cloud, tal i com es veurà en els respectius apartats.

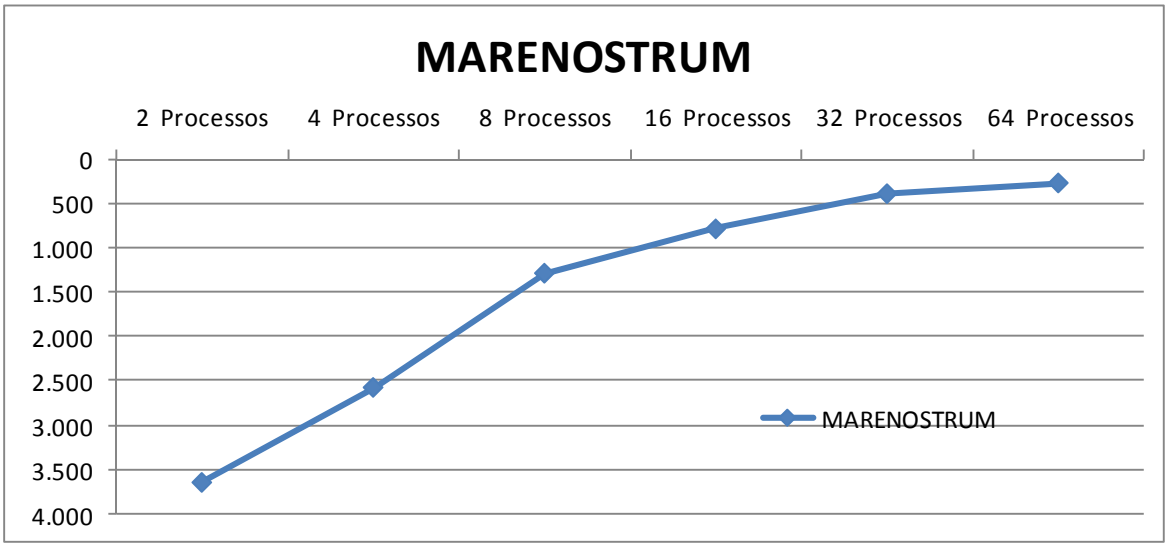
[9] Fa referència al sistema de planificació d'assignació de nucli virtual (pin) als processos (threads).

7.3 RESULTATS OBTINGUTS A MARENOSTRUM

Els resultats, de la mateixa manera que a EC2, són els totals respecte a inicialització, execució i resposta. De fet, també inclou el temps que el treball queda en espera en el sistema de cues (típicament utilitzats en sistemes batch).

Taula Resultats per nodes (resultats expressats en segons):

	RESULTAT (seg.)
	-
2 Processos	3.640 s
4 Processos	2.571 s
8 Processos	1.285 s
16 Processos	786 s
32 Processos	390 s
64 Processos	269 s

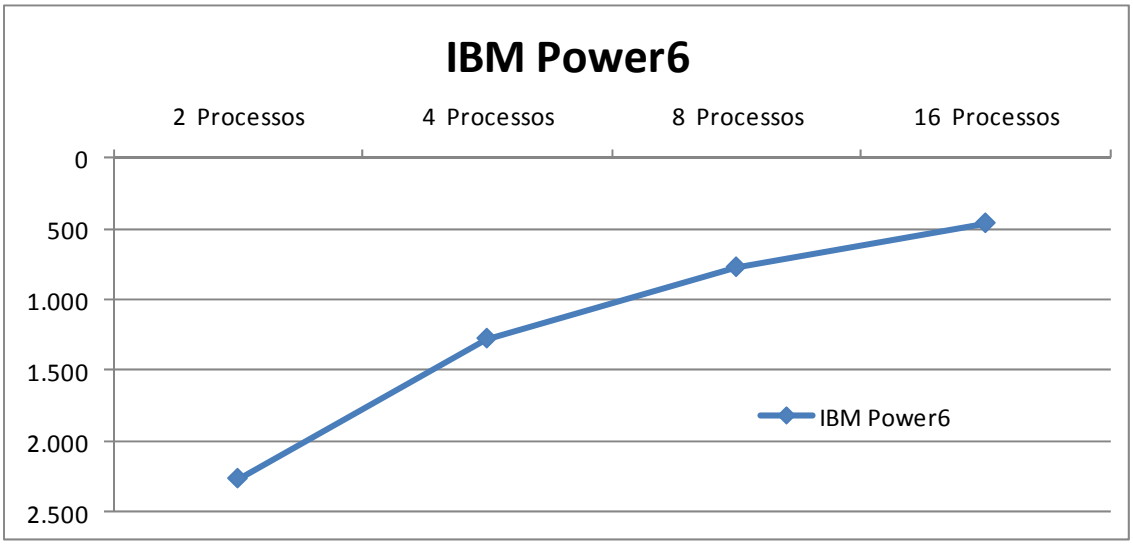


7.4 RESULTATS OBTINGUTS A IBM POWER6

Els resultats, de la mateixa manera que als altres entorns, són els totals respecte a inicialització, execució i resposta.

Taula Resultats per nodes (resultats expressats en segons):

	RESULAT
	-
2 Processos	2.271 s
4 Processos	1.280 s
8 Processos	775 s
16 Processos	464 s

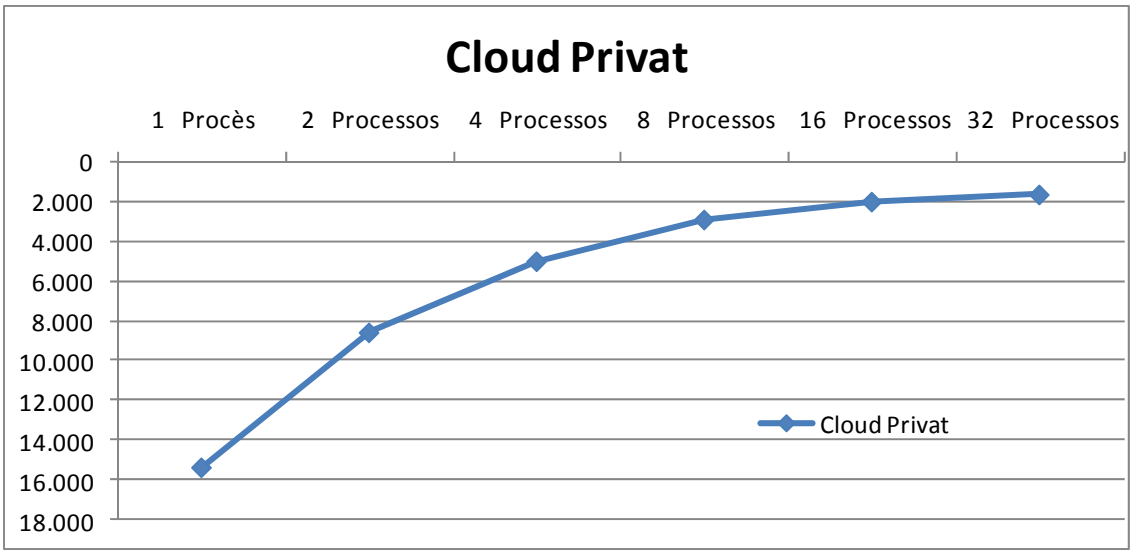


7.5 RESULTATS OBTINGUTS A CLOUD PRIVAT

Els resultats, de la mateixa manera que als altres entorns, són els totals respecte a inicialització, execució i resposta.

Taula Resultats per nodes (resultats expressats en segons):

		RESULAT
1	Procés	15.355 s
2	Processos	8.595 s
4	Processos	5.029 s
8	Processos	2.942 s
16	Processos	2.043 s
32	Processos	1.684 s



8 ESTUDI DELS RESULTATS

Els resultats obtinguts sota els diferents entorns estan basats en les mateixes dades d'entrada i el mateix programari, tot i això els entorns d'execució són completament diferents entre si. Per tal de poder avaluar aquests resultats des d'un punt de vista comú s'ha realitzat una anàlisi prèvia per tal d'obtenir una sèrie de variables que permetin establir unes bases comparatives adequades als objectius finals definits sota aquest treball.

D'aquesta manera es pretén, al següent apartat, veure una aproximació a tots els costos de instal·lació, adquisició, reparació, manteniment,... associats a cadascun dels entorns de Cloud per tal d'aconseguir conèixer tots els aspectes relacionats amb el cost final de la potència de càlcul efectiva per unitat de temps.

S'ha intentat per tant arribar d'una banda, a definir el cost final que anomenarem CPU/hora i d'altra banda, definir els costos associats per tal d'aconseguir realitzar una avaluació efectiva sota el model 'pay-as-you-go', així com un anàlisi sota els conceptes del Cloudbursting, per identificar com afectarien les possibles variacions en les necessitats de processament numèric.

8.1 ANÀLISI DE COSTS

8.1.1 AMAZON EC2

En el cas del Amazon EC2 el càlcul por realitzar-se de manera directa, doncs només es necessari veure la taula de preus proporcionada per Amazon, de totes formes cal afegir a aquestos preus els conceptes relatius a impostos locals, en aquest cas s'aplica un 16% d'IVA (tassa estatal a juny 2010), obtenint els resultats següents:

	Instància/hora	16% IVA	CPU/h DOLAR	CPU/h EURO
Small Instance m1.small	$0,0850	$0,0136	$0,0986	**0,07041 €**
High-CPU Medium Instance c1.medium	$0,1700	$0,0272	$0,1972	**0,14082 €**

En el cas de ser necessari afegir eines d'anàlisi de rendiment, EC2 permet la utilització de CloudWatch, que mostra en temps real tota una sèrie de paràmetres de CPU, xarxa, disc,... Aquest servei requereix el pagament extra de:

	Instància/hora	16% IVA	CPU/h DOLAR	CPU/h EURO
Cloud Watch - Small Instance m1.small	$0,0015	$0,0024	$0,0174	**0,01243 €**

Tots els experiment han estat realitzats a la zona *US East (Northern Virginia) Region*, donats els preus inferiors que s'ofereixen en aquesta regió en relació a les regions d'Europa i Àsia.

La tassa de canvi aplicat per la conversió a la moneda de la zona euro ha estat la mitjana de l'últim any, 1 USD = 0.7141 EUR, segons las dades del Banc Central Europeu.

Font: Banc Central Europeu

8.1.2 MARENOSTRUM

Per tal de poder arribar a una aproximació del cost CPU/hora sota l'entorn del Supercomputador Marenostrum partirem de la següent anàlisi.

Segons les fonts consultades, per tal de dissenyar i posar en funcionament els 10.240 processadors disponibles al BSC van ser necessaris un total de 70.000.000 € d'inversió inicial. Posteriorment van ser invertits 14.000.000 € més per millores.

Anualment aquest supercomputador consumeix aproximadament 1,2 milions d'euros en energia elèctrica.

Si partim d'una estimació base de manteniment on els costos de personal equivalen a: 4 tècnics dedicats al manteniment(4x35.000€/any), 1 coordinador(45.000€/any) i un responsable(65.000€/any), obtenim un total de 250.000€/any en cost de personal.

Relatiu a reparacions podem estimar els costs la substitució d'uns 50-100 nodes al any en 150.000-200.000€/any.

Com a conceptes extraordinaris podem definir un import d'uns 150.000€/any per tal de cobrir possibles eventualitats de tot tipus així com a partides no contemplades en les despeses anteriors.

S'ha definit una duració del cicle de 10 anys de manera que s'han repartit els costs d'inversió inicial durant aquest període. I s'ha suposat una disponibilitat d'un 95% en el servei. L'estimació de 10 anys està basada en les fonts que afirmen la substitució del Marenostrum per un model superior de consum elèctric inferior en els pròxims anys.

Finalment, realitzant una estimació es pot obtenir el següent quadre resum:

Cost Personal anual	250.000,00 €
Manteniment anual	200.000,00 €
Varis anual	150.000,00 €
Consum elèctric anual	1.200.000,00 €
TOTAL ANUAL (A)	**1.800.000,00 €**
Inversió (1)	84.000.000,00 €
Duració Cicle (2)	10
Cost Inversió Anual (B) =(1)/(2)	**8.400.000,00 €**
COST ANUAL (A)+(B)	**10.200.000,00 €**
Processadors	10240
Disponibilitat	95%
Hores Disponibles Any	8322
COST CPU/HORA	***0,1197 €***

S'ha d'observar que aquest cost seria relatiu a disposar de la propietat de la màquina i no equival per tant a cap preu de venda. Es tracta per tant d'un valor totalment orientatiu i calculat amb l'objectiu de poder realitzar una comparativa amb d'altres entorns que inclogui una variable de cost.

8.1.3 IBM POWER6

Els costos associats al IBM Power6 no han pogut ser obtinguts directament pel model analitzat, donat que es desconeix el nombre total de blades amb els que compta el sistema ni les seves característiques tècniques ni costs d'inversió o manteniment. Per aquesta raó s'ha cregut considerar que el preu final expressat en CPU/h serà equivalent o superior al del supercomputador Marenostrum (per al mateix número de nodes), donat que es tracta d'un processador i sistema amb una tecnologia i rendiment superiors.

8.1.4 CLOUD PRIVAT

Tot i que anàlogament als entorns anteriors, es pot realitzar una estimació de costs per un clúster de les prestacions especificades, per tal que les dades puguin ser extrapolades a un espectre de configuracions més ampli l'anàlisi següent s'ha realitzat seguint un criteri de rangs, així que es partirà de la definició dels elements bàsics necessaris d'una configuració mínima fins a l'especificació d'una configuració màxima i s'analitzaran les diferencies de cost entre totes dos.

La base inicial serà la d'un clúster amb 64 CPUs on s'utilitzarien des de processadors d'un sòl nucli a processadors doble nucli. Es necessitaran per tant 64 equips complets (Placa, cpu, memòria, disc, targeta de xarxa, font alimentació,...). Segons l'espai i el pressupost es podria optar per disponibilitzar els equips en racks.

Serà necessari un sistema de xarxa amb un mínim de 64+1 connexions i realitzar tot el cablejat físic. Aquest sistema pot tractar-se de Gigabit o Ethernet.

Caldrà instal·lar, si es creu necessari, algun sistema d'alimentació ininterrompuda per mitigar possibles fallades en el subministrament així com preservar la vida dels equips.

Com a costs relatius a la instal·lació i posta a punt dels equips es poden estimar 50-75 hores de 2 tècnics especialitzats que realitzarien les instal·lacions físiques dels equips,

així com les tasques d'instal·lació i configuració del programari base, que hauria de tractar-se d'una arquitectura tipus XEN.

En relació al manteniment anual relatiu als equips i programari es partirà d'una base de màxima disponibilitat on establirem una persona dedicada completament als equips que estimarem en 25.000€/any, així com un possible manteniment/substitució d'elements en 1-3 nodes any.

Els consums elèctrics associats els calcularem seguint els criteris definits per Energy Star i equivalen a un total de 1368 kWany per node[10]. El preu estimat del kWh ha estat obtingut de UNESA segons l'informe de preus industrials i s'ha establit en 10,01 cèntims d'euro per kWh.

També s'ha definit una partida relativa a Varis per tal d'incloure possibles despeses fixes no contemplades inicialment.

S'ha definit una duració del cicle d'utilització dels equips de 5 anys de manera que s'han repartit els costs d'inversió inicial, juntament als costs de manteniment i utilització, durant aquest període.

S'ha suposat una disponibilitat d'un 95% en el servei, assignant un 5% a manteniment o indisponibilitats.

[10] El consum consisteix en el que es reflexa a EnergyStar per un servidor (sense monitor) funcionant ininterrompudament 365dies/24hores.

L'estimació final del cost total de cada un dels elements necessaris s'ha descrit en la següent taula:

	Sistema Bàsic	Sistema Avançat
TOTAL ANUAL (A)	**37.263,96 €**	**38.763,96 €**
Cost de Personal anual	25.000,00 €	25.000,00 €
Manteniment anual	1.500,00 €	3.000,00 €
Varis	2.000,00 €	2.000,00 €
Consum elèctric anual	8.763,96 €	8.763,96 €
Inversió Inicial Total (1)	**34.600,00 €**	**73.000,00 €**
Preu 1 Node	500,00 €	1.000,00 €
Total nodes	64	64
Xarxa	600,00 €	2.500,00 €
Racks	0	2.500,00 €
Altres	2.000,00 €	4.000,00 €
COSTS D'INVERSIÓ/ANY (B) = (1)/(2)	**6.920,00 €**	**14.600,00 €**
Duració cicle en anys (2)	5	5
COST ANUAL (A+B)	**44.183,96 €**	**53.363,96 €**
Processadors	64	64
Disponibilitat	95%	95%
Hores Disponibles Any	8.322	8.322
COST CPU/HORA	**0,0830 €**	**0,1002 €**

8.2 ESTUDI COMPARATIU

Inicialment partirem de la base que EC2 implica un cost inferior al dels supercomputadors Marenostrum i IBM Power6 en la utilització d'instàncies *small* i un cost relativament equivalent en la utilització del model *medium*. Aquest fet es reforça amb l'anàlisi de costs del cloud privat, on s'ha pogut veure la despesa aproximada (i orientativa) d'utilització d'aquest que es equivalent novament al cost small a Amazon

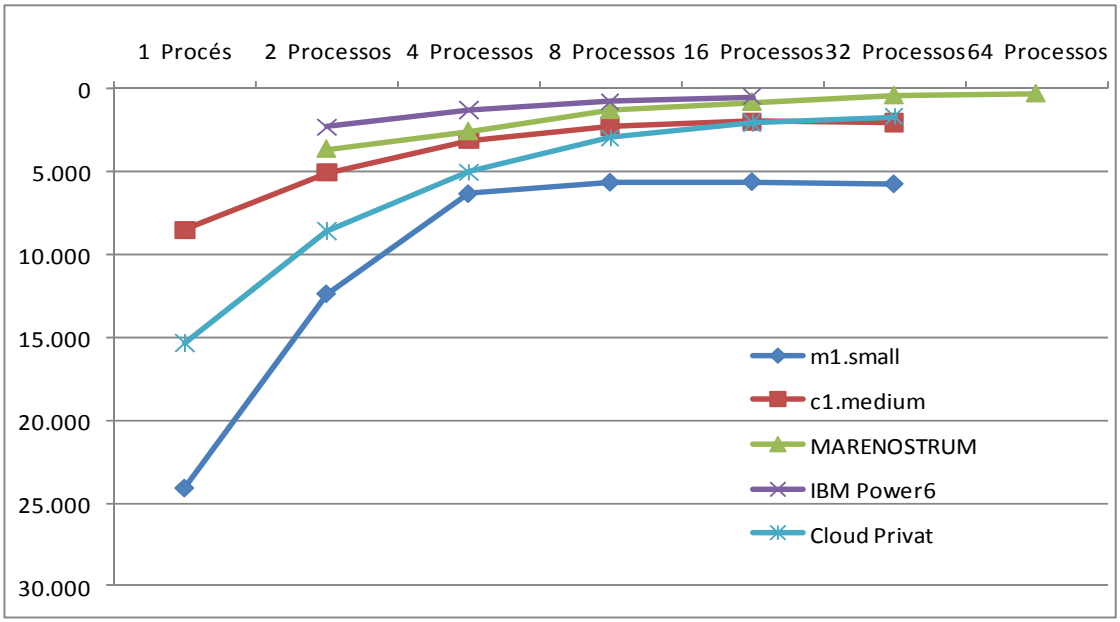
Si es realitza un càlcul de manera inversa a la obtenció del cost/hora aplicat a EC2 es pot veure el següent:

0,07041 € x 24h x 365 dies x 10.240 CPUs = 6.315.945 €/any

Aquest valor equivaldria al cost d'utilització de 10.240 processadors (Marenostrum) a l'any però en un entorn EC2 amb un model d'execució small.

El següent pas es veure quines són les diferencies als resultats d'execució als diferents entorns per tal de poder obtenir els pros i contres i iniciar així l'anàlisi.

El següent gràfic mostra els resultats comparatius de tots els entorns (en segons):



El principal problema detectat en les execucions es troba a EC2 i és relatiu a l'escalat. A partir de 8 nodes el increment d'aquests no millora prou el resultat final.

Addicionalment, tant en els models small com medium d'EC2, a partir de 32 nodes, el resultat de temps de computació total empleat augmenta respecte al temps utilitzat per les mateixes tasques amb 16 nodes.

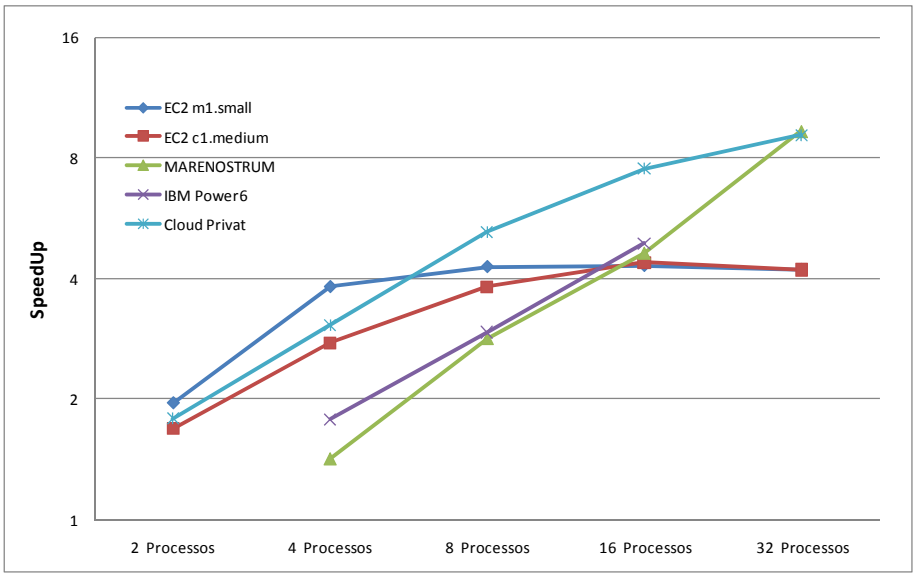
Aquest fet es basa als problemes que presenten els entorns de Cloud públic EC2 associats a la comunicació i la latència[11].

WRF, a diferencia d'altres d'aplicacions de distribució de càlcul en clúster amb poc pas de missatges, requereix d'un ús intensiu de sincronitzacions i intercanvis d'informacions i dades a través de MPI. Aquest fet es molt palpable als resultats obtinguts.

Per tal d'analitzar amb detall el rendiment mitjançant l'execució en paral·lel es farà ús del concepte del *Speedup Factor*[12], en funció del increment del nombre de processos

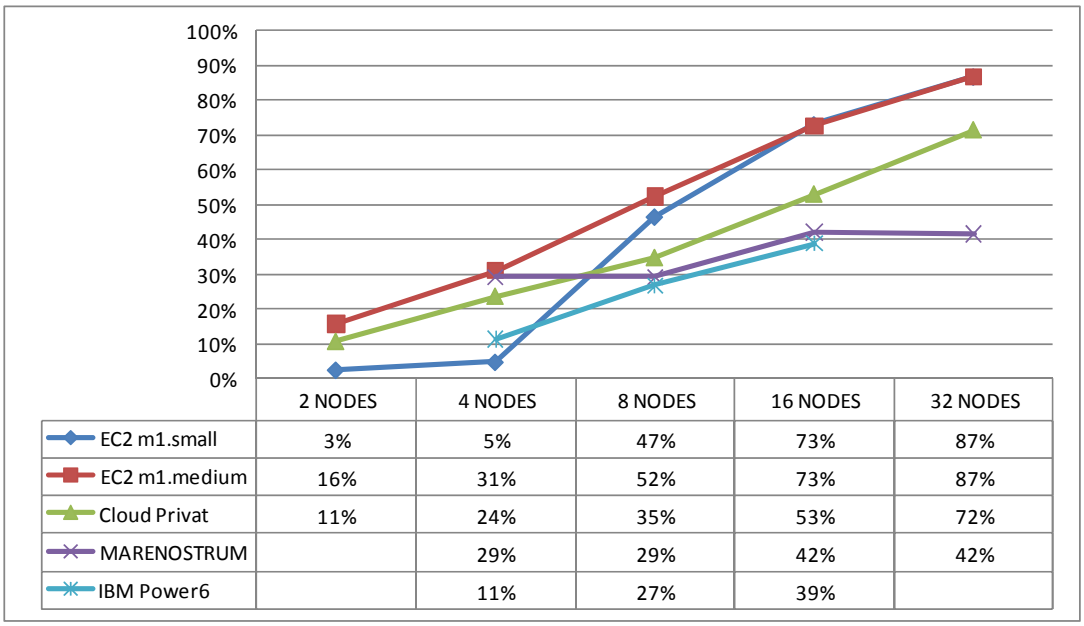
[11] Referit al concepte que expressa el retard en la resposta orientat a la comunicació entre equips.
[12] WILKINSON, B., ALLEN, M. "Parallel Programming Techniques & Applications Using Networked Workstations & Parallel Computers"

utilitzats. Aquest valor equival a $S(p) = \frac{Temps\ Execució\ amb\ un\ sol\ processador}{Temps\ d'execució\ amb\ \ p\ processadors}$. Aquest concepte permet entendre quin es el guany en rendiment respecte el increment del nombre de processadors al sistema. El següent gràfic mostra els resultats:



D'aquesta gràfica es pot obtenir que EC2 assoleix el valor màxim de Speedup entre 8 i 16 processadors.

El següent gràfic, basat en el concepte anterior, mostra el rendiment perdut [13] en % de cada una de les execucions respecte la primera. És una petita aproximació per veure el grau d'afectació de la sincronització entre els processos:



	2 NODES	4 NODES	8 NODES	16 NODES	32 NODES
EC2 m1.small	3%	5%	47%	73%	87%
EC2 m1.medium	16%	31%	52%	73%	87%
Cloud Privat	11%	24%	35%	53%	72%
MARENOSTRUM		29%	29%	42%	42%
IBM Power6		11%	27%	39%	

[13] El rendiment perdut, complementari al concepte d'eficiència, equival a: *100% – EFICIÈNCIA*, on la eficiència *E* es calcula com $E = \frac{Speedup}{Nombre\ Processadors}$

En concret, a EC2, la pèrdua de rendiment (a tots 2 models, *small* i *medium*) a l'execució amb 32 nodes és del 87% respecte el rendiment obtingut a l'execució en un sol node (Una eficiència del 13%).

Si observem el mateix rendiment comparatiu a Marenostrum observem que la pèrdua de rendiment és molt inferior, tan sols un 40% a 32 nodes (eficiència del 60%) respecte gairebé el 90% a EC2. Un altre fet important a destacar als resultats de Marenostrum és que les valors obtinguts, tal i com s'ha indicat anteriorment, són el total dels temps a la cua, posterior execució i resposta, i a partir de 32 nodes, el resultat total de l'ordre de 5 minuts es comença a veure afectat per aquest temps d'espera.

Els resultats al cloud privat tot i no arribar als de Marenostrum, són també força millors que EC2, observant una pèrdua de rendiment del ordre del 55%, és a dir una eficiència del 45%.

9 CONLUSIONS FINALS

Un vegada realitzat els estudis de cost i avaluació de rendiment dels diferents entorns podem arribar a les següents conclusions.

El Cloud públic, i en concret el cas analitzat EC2, tot i que representa a priori un model idoni sota el criteri pay-as-you-go representant una opció de cost inferior en tots els casos, sota els criteris de rendiment absolut no es troba amb la maduresa suficient per ser utilitzat per aplicacions HPC que realitzin un ús intensiu de comunicació i sincronització MPI com WRF. S'ha demostrat que a partir d'un cert nombre de nodes, tal i com es desprèn del càlcul del màxim *Speedup Factor*, EC2 es veu col·lapsat pel rendiment de la seva xarxa interna.

Aquest fet és un escull difícil de solucionar, que afegit a la problemàtica descoberta per tal d'utilitzar múltiples *cores* correctament en una mateixa instància implica un important desaprofitament dels recursos.

D'una banda, EC2 en casos de petits "clústers" de fins a 16 nodes i instàncies simples o mitjanes, obté un rendiment molt raonable, comparativament similar als altres entorns, i representa, sota el criteri rendiment/cost, la millor opció amb gran diferència. Es pot veure com computar mitjançant un model *small* amb 16 nodes durant 24 hores representa un cost total de 27€ sense cap increment del TCO.

D'una altra banda, a partir d'aquest punt, intentar obtenir un augment de rendiment sota EC2 comença a ser una tasca molt complexa, on els costs s'incrementen exponencialment sense que el rendiment es vegi afectat.

Tot això fa concloure que EC2, a partir d'aquest esgraó, no representa encara una opció idònia, tot i el seu model de pagament per consum, per entorns d'altes prestacions.

Els supercomputadors en una primera línia i els clústers de capacitats més reduïdes en un segon esgraó representen encara el model computacional a escollir per l'execució d'aplicacions HPC. De totes maneres, tot i la seva immaduresa, cal no perdre de vista el

paradigma del Cloud associat al HPC, doncs a mesura que aquest evolucioni i millori els aspectes esmentats anteriorment pot representar una molt bona opció a tenir en compte pel món de les aplicacions HPC, doncs la utilització de supercomputadors requereix la inversió de quantioses sumes de diners i per tant un gran augment del TCO.

En relació als aspectes de familiarització i aprenentatge dels entorns, tots presenten les seves peculiaritats, tot i que la virtualització presenta una clara avantatge respecte els sistemes més tancats i personalitzats associats als clúster d'altes prestacions. Actualment, tal i com es pot veure a les dades bibliogràfiques, la informació i recursos en línia associat al Cloud, i en concret a XEN, és molt amplia i diversa, representant un altre dels pros d'aquest tipus de sistemes.

9.1 DESVIACIONS RESPECTE LA PLANIFICACIÓ

Els principals problemes que han provocat desviacions respecte la planificació original s'han degut pricipalment a:

- Dificultats de compilació de WRF en l'entorn EC2, fet que ha incrementat en alguns dies extra els temps dedicat a aquesta fase.
- La necessitat d'activar el compte EC2, realitzant una sol·licitud directa a Amazon, per tal de poder executar més de 20 instàncies al mateix temps va fer necessari perdre un dia en la fase d'execucions.
- Els problemes detectats en l'execució del model *c1.medium* amb 2 processos en paral·lel en una mateixa instància va comportar una desviació respecte als resultats i conclusions que inicialment es creia s'aconseguirien. En relació al temps va suposar una desviació força important al ser necessari analitzar en detall l'origen del problema.

En els altres aspectes, la planificació s'ha complert de la forma prevista i els checkpoints[14] marcats per les entregues parcials de documentació s'han complert sempre amb temps suficient (Cal destacar que la planificació inicial els contemplava).

[14] Han consistit en 3 entregues parcials i la entrega final d'aquest document tipus memòria.

9.2 VALORACIÓ PERSONAL

A nivell personal m'agradaria esmentar que la realització d'aquest treball m'ha comportat un experiència final molt gratificant, gràcies a la qual he pogut conèixer més de a prop el paradigma del Cloud Computing i el funcionament d'una aplicació HPC real i la seva aplicació en entorns d'alt rendiment. Durant el procés de preparació del entorn EC2 he patit alguns problemes relatius a la posta a punt que han representant moments de certa desesperació i que han precisat de molta dedicació per tal de ser superats amb èxit. Voldria emfatitzar el fet que la familiarització amb aquest entorn, així com la seva posta en funcionament i execució de les proves, han representat un gran esforç i hores de treball que potser no es veuen reflectides d'aquesta manera en els resultats finals obtinguts. Tot i això m'emporto un bagatge molt enriquidor amb una visió i experiència aplicable al món laboral.

POSSIBLES MILLORES D'EXECUCIÓ A EC2

En quant a possibles millores a les proves realitzades m'agradaria esmentar que L'AMI desenvolupada per les execucions a EC2 segueix el model d'execució de 32 bits. No s'ha procedit a la creació seguint un model de 64 bits per les limitacions de pressupost, ja que les instàncies d'aquest tipus encaririen els costs globals del projecte substancialment[15]. De totes formes, la preparació de l'entorn en models de 64 bits no seria gaire diferent al exposat en aquest treball.

Es podrien utilitzar com a imatges inicials les AMIs e813f681 i e813f681 disponibles al repositori d'Amazon. Aquestes instàncies estan basades en Fedora 64bits i porten pre-instal·lades les mateixes llibreries que els models emprats de 32 bits, sols que en models de compilació 64bits.

Els passos a seguir i les dependències són els mateixos, tot i això cal remarcar que caldrien algunes adaptacions en la compilació model WRF, doncs aquesta no és trivial.

[15] Del ordre de 10-15 vegades el cost del projecte actual.

10 BIBLIOGRAFIA

Amazon EC2 (2010). [en línia]. http://aws.amazon.com/documentation/ec2/

Amazon S3 (2010). [en línia]. http://aws.amazon.com/documentation/s3/

Barcelona Supercomputing Center. [en línia]. http://www.bsc.es/plantillaA.php?cat_id=200 http://www.aulesempresa.upc.edu/aulesempresa/2010/programes/BSC/Llista-documents/00/document/2%20BSC-Sistemas-Aulas-empresa%20FIB%20-2010.pdf

BARHAM, Paul; DRAGOVIC, Boris; FRASER, Keir; HAND, Steven; HARRIS, Tim; HO, Alex; NEUGEBAUERY, Rolf; PRATT, Ian; WARFIELD, Andrew, (2003). "Xen and the Art of Virtualization". University of Cambridge Compute Laboratory. [en línia]. http://www.cl.cam.ac.uk/research/srg/netos/papers/2003-xensosp.pdf

BIENTINESI, P.; IAKYMCHUK, R.; NAPPER, J. (2010). " Handbook of Cloud Computing", Capítol, "HPC on Competitive Cloud Resources", ISBN 978-1-4419-6523-3

Calculadora d'Energia. ENERGY STAR. [en línia]. http://www.eu-energystar.org/es/es_007c.shtml

"CISL Research Data Archive". University Corporation for Atmospheric Research. [en línia]. http://dss.ucar.edu/

GOMEZ, Noemí. (2008) "El superordenador 'MareNostrum' tendrá un sucesor en 2011". Revista Expansión. [en línia]. http://www.expansion.com/2008/05/19/empresas/1124990.html

"Guía del Análisis de Costes-Beneficios de los Proyectos de Inversión" (2003). Comisión Europea. [en línia]. http://ec.europa.eu/regional_policy/sources/docgener/guides/cost/guide02_es.pdf

"IBM Power 575 supercomputing node". IBM. [en línia]. http://www-03.ibm.com/systems/power/hardware/575 /index.html

LA Grid. *[en línia].* http://www.latinamericangrid.org/

LEIDER, Justin. "The Limitations of Scaling with EC2" [en línia]. http://www.derivante.com/2008/10/08/the-limitations-of-scaling-with-ec2/

"MARENOSTRUM User Guide". Barcelona Supercomputing Center. [en línia]. http://www.bsc.es/media/859.pdf

MPICH2. *Argonne, Mathematics and Science Division.* [en línia]. http://www.mcs.anl.gov/research/projects/mpich2/

NETCDF Documentation. [en línia]. http://www.unidata.ucar.edu/software/netcdf/docs/

Open MPI. *[en línia].* http://www.open-mpi.org/

"Python Programming Language". Official Website. [en línia]. http://www.python.org/

RODERO, I. (2009). *Introduction to Parallel and Distributed Programming*. [en línia]. http://nsfcac.rutgers.edu/people/irodero/classes/09-10/ece451-566/lectures.php

ROSEN G. (2010). "Jack of all Clouds". [en línia]. http://www.jackofallclouds.com/

SKAMAROCK W.C.; KLEMP, J.B.; DUDHIA, J.; GILL, D.O.; BARKER, D.M.; DUDA, M.G.HUANG, X.; WANG, W.; POWERS, J.G. (2008). "A Description of the Advanced Research WRF Version 3". University Corporation for Atmospheric Research. [en línia]. http://www.mmm.ucar.edu/wrf/users/docs/arw_v3.pdf

SKOMOROCK, P. (2007). "On-Demand MPI Cluster with Python and EC2". [en línia]. http://www.datawrangling.com/on-demand-mpi-cluster-with-python-and-ec2-part-1-of-3; http://www.datawrangling.com/mpi-cluster-with-python-and-amazon-ec2-part-2-of-3

SKOMOROCK, P. (2008). "ElasticWulf 64 bits". [en línia]. http://code.google.com/p/elasticwulf/

Tasses de Canvi Oficials (2010). European Central Bank. [en línia]. http://www.ecb.int/stats/exchange/eurofxref/html/eurofxref-graph-usd.en.html

TOP 500 Supercomputer Sites. [en línia]. www.top500.org

"The Message Passing Interface (MPI) standard". *Argonne, Mathematics and Science Division.* [en línia]. http://www.mcs.anl.gov/research/projects/mpi/

"The Weather Research & Forecasting Model". [en línia]. http://www.wrf-model.org/index.php

UNESA (2010). Asociación Española de la Industria Eléctrica. [en línia]. http://www.unesa.es/

VARIS AUTORS. (200). "Proyecto Clúster". Proyecto de Innovación Docente. Universistat de Granada. [en línia]. http://atc.ugr.es/~javier/docencia/SWAD_Cluster/Proyecto_cluster.html

WILLIAMSON, A. (2010). "Amazon EC2 Latency: The Pretty Graphs". [en línia]. http://alan.blog-city.com/amazon_ec2_latency_the_pretty_graphs.htm

WILKINSON, B., ALLEN, M. (1998). "Parallel Programming Techniques & Applications Using Networked Workstations & Parallel Computers". Segona Edició.

"WRF ARW Online Tutorial". University Corporation for Atmospheric Research. [en línia]. http://www.mmm.ucar.edu/wrf/OnLineTutorial/index.htm

"WRF Compilation". *Observatorio Andino de Eventos Extraordinarion.* [en línia]. http://www.cmc.org.ve/mediawiki/index.php?title=%E2%97%A6_WRF

"WRF Users Page". University Corporation for Atmospheric Research. [en línia]. http://www.mmm.ucar.edu/wrf/users/

XEN. *[en línia].*http://www.xen.org/

11 ANNEXOS

11.1 MPI - MESSAGE PASSING INTERFACE

Es tracta d'un estàndard que defineix la sintaxi i semàntica de les funcions d'una biblioteca de pas de missatges amb la finalitat del seu ús en entorns amb múltiples processadors.

El seu disseny està orientat a la programació concurrent i la seva característica principal és l'absència de memòria compartida, el que el fa adequat en programació de sistemes distribuïts.

El seu funcionament pot ser tant síncron com asíncron. Al tractar-se d'un estàndard, és portable entre diferents processadors, sistemes, xarxes, etc... i existeixen implementacions lliures.

Les seves prestacions son elevades i així com les seves funcionalitats. Cal remarcar que l'especificació defineix el comportament i utilització, però poden existir diferents implementacions.

Les funcionalitats s'agrupen en 4 grups:

- Inici, gestió y finalització de comunicacions.
- Transferència de dades entre 2 processos.
- Transferència de dades entre varis processos.
- Creació de tipus de dades definits per l'usuari.

11.2 SCRIPTS EC2

Font: http://www.datawrangling.com/mpi-cluster-with-python-and-amazon-ec2-part-2-of-3

Els scripts utilitzats per la posta a punt i preparació de les instàncies EC2 estan basats en els següents arxius que poden ser descarregat del link anterior. Aquests scripts permeten en termes generals crear totes les instàncies, prèviament configurades per executar WRF, i configurar els nodes del "clúster" per execució MPI.

EC2.py

APIs principals del sistema EC2. Es Poden obtenir directament de la web d'Amazon EC2.

EC2config.py

Arxiu inicial de configuració, cal editar aquest arxiu per tal d'establir els següents paràmetres:

```
AWS_ACCESS_KEY_ID = 'XXXXXXXXXXXXXXX'      codi d'usuari Amazon EC2
AWS_SECRET_ACCESS_KEY = '0vBxxxxxxxx'      password Amazon EC2

KEYNAME = "dcsuoc"                         nom del key pair
KEY_LOCATION = "./dcsuoc.pem"              path del arxiu .pem

MASTER_IMAGE_ID = "ami-1bd93072"           identificador de la AMI
IMAGE_ID = "ami-1bd93072"                  identificador de la AMI
INSTANCE_TYPE = "m1.small"                 Model de CPU (small o medium)


DEFAULT_CLUSTER_SIZE = X                   nombre de instancies a crear
```

ec2-start-cluster.py

La seva execució dona d'alta a EC2 un nombre `DEFAULT_CLUSTER_SIZE` d'instàncies del tipus `IMAGE_ID`, i model de CPU `INSTANCE_TYPE` definides al arxiu de configuració EC2config.py

ec2-stop-cluster.py

La seva execució finalitza les instàncies en execució.

ec2-mpi-config.py

La seva execució prepara l'entorn MPI, tot retornant el nom del node MASTER. S'estableixen sota aquest scripts totes les connexions dels diferents nodes i s'envien els arxius de mpd.hosts per tal de registrar els diferents hosts per l'execució en paral·lel.

ec2-check-instances.py

Permet obtenir un llistat de les instancies del sistema i el seu estat d'execució. La seva utilització permet saber quan les instancies creades es troben disponibles (running).

Tots els arxius fan servir llenguatge de programació *Python*.

Es tracta d'un llenguatge interpretat, creat per *Guido van Rossum* al 1.991, de codi obert.

Una de les seves característiques principals és la resolució dinàmica de noms, que permet enllaçar un mètode i un nom de variable durant la execució del programa (*dynamic binding*). També es tracta d'un llenguatge que permet una programació orienta a objecte, estructurada, així com una programació orientada a aspectes.

L'objectiu principal de Python és facilitat de lectura i disseny.

11.3 PARÀMETRES NAMELIST WPS I INPUT

A continuació es descriuen els detalls dels paràmetres d'entrada dels arxius namelist.wps i namelist.input necessaris per l'execució WRF.

WPS NAMELIST.WPS

```
&share
 wrf_core = 'ARW',
 max_dom = 1,
 start_date = '2007-09-19_18:00:00','2007-09-19_18:00:00'
 end_date   = '2007-09-20_18:00:00','2007-09-20_18:00:00'
 interval_seconds = 10800
 io_form_geogrid = 2,
/

&geogrid
 parent_id         =   0,   1,
 parent_grid_ratio =   1,   3,
 i_parent_start    =   0,  31,
 j_parent_start    =   0,  17,
 e_we              =  75, 112,
 e_sn              =  75,  97,
 geog_data_res     = '10m','2m',
 dx = 4000,
 dy = 4000,
 map_proj = 'lambert',
 ref_lat   =  34.83,
 ref_lon   = -81.03,
 truelat1  =  30.0,
 truelat2  =  60.0,
 stand_lon = -98.0,
 geog_data_path = '/mnt/geog'
/

&ungrib
 out_format = 'WPS',
 prefix = 'FILE',
/

&metgrid
 fg_name = 'FILE'
 io_form_metgrid = 2,
/
```

El paràmetre `geog_data_path` depèn de la ubicació de l'arxiu de dades geogràfiques de cada sistema.

WRF: NAMELIST.INPUT

```
&time_control
 run_days                            = 0,
 run_hours                           = 24,
 run_minutes                         = 0,
 run_seconds                         = 0,
 start_year                          = 2007, 2000, 2000,
 start_month                         = 09,   01,   01,
 start_day                           = 19,   24,   24,
 start_hour                          = 18,   12,   12,
 start_minute                        = 00,   00,   00,
 start_second                        = 00,   00,   00,
 end_year                            = 2007, 2000, 2000,
 end_month                           = 09,   01,   01,
 end_day                             = 20,   25,   25,
 end_hour                            = 18,   12,   12,
 end_minute                          = 00,   00,   00,
 end_second                          = 00,   00,   00,
 interval_seconds                    = 10800
 input_from_file                     = .true.,.false.,.false.,
 history_interval                    = 180,  60,   60,
```

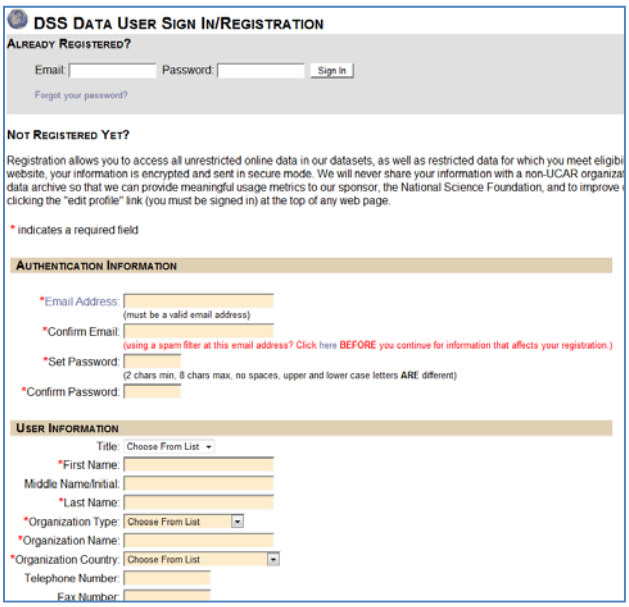
```
frames_per_outfile                  = 1000, 1000, 1000,
restart                             = .false.,
restart_interval                    = 5000,
io_form_history                     = 2
io_form_restart                     = 2
io_form_input                       = 2
io_form_boundary                    = 2
debug_level                         = 0
/

&domains
time_step                           = 24,
time_step_fract_num                 = 0,
time_step_fract_den                 = 10,
max_dom                             = 1,
s_we                                = 1,     1,     1,
e_we                                = 75,    112,   94,
s_sn                                = 1,     1,     1,
e_sn                                = 75,    97,    91,
s_vert                              = 1,     1,     1,
e_vert                              = 31,    28,    28,
num_metgrid_levels                  = 27
dx                                  = 4000, 10000,  3333,
dy                                  = 4000, 10000,  3333,
grid_id                             = 1,     2,     3,
parent_id                           = 0,     1,     2,
i_parent_start                      = 0,     31,    30,
j_parent_start                      = 0,     17,    30,
parent_grid_ratio                   = 1,     3,     3,
parent_time_step_ratio              = 1,     3,     3,
feedback                            = 1,
smooth_option                       = 1
/

&physics
mp_physics                          = 3,     3,     3,
ra_lw_physics                       = 1,     1,     1,
ra_sw_physics                       = 1,     1,     1,
radt                                = 30,    30,    30,
sf_sfclay_physics                   = 1,     1,     1,
sf_surface_physics                  = 1,     1,     1,
bl_pbl_physics                      = 1,     1,     1,
bldt                                = 0,     0,     0,
cu_physics                          = 1,     1,     0,
cudt                                = 5,     5,     5,
isfflx                              = 1,
ifsnow                              = 0,
icloud                              = 1,
surface_input_source                = 1,
num_soil_layers                     = 5,
sf_urban_physics                    = 0,
mp_zero_out                         = 0,
maxiens                             = 1,
maxens                              = 3,
maxens2                             = 3,
maxens3                             = 16,
ensdim                              = 144,
/

&fdda
/

&dynamics
dyn_opt                             = 2,
rk_ord                              = 3,
w_damping                           = 0,
diff_opt                            = 1,
km_opt                              = 4,
damp_opt                            = 0,
base_temp                           = 290.
zdamp                               = 5000.,  5000.,  5000.,
dampcoef                            = 0.01,   0.01,   0.01
khdif                               = 0,      0,      0,
kvdif                               = 0,      0,      0,
smdiv                               = 0.1,    0.1,    0.1,
emdiv                               = 0.01,   0.01,   0.01,
```

```
epssm                               = 0.1,    0.1,    0.1
time_step_sound                     = 4,      4,      4,
h_mom_adv_order                     = 5,      5,      5,
v_mom_adv_order                     = 3,      3,      3,
h_sca_adv_order                     = 5,      5,      5,
v_sca_adv_order                     = 3,      3,      3,
non_hydrostatic                     = .true., .true., .true.,
moist_adv_opt                       = 0,      0,      0,
scalar_adv_opt                      = 0,      0,      0,
/

&bdy_control
spec_bdy_width                      = 5,
spec_zone                           = 1,
relax_zone                          = 4,
specified                           = .true., .false.,.false.,
periodic_x                          = .false.,.false.,.false.,
symmetric_xs                        = .false.,.false.,.false.,
symmetric_xe                        = .false.,.false.,.false.,
open_xs                             = .false.,.false.,.false.,
open_xe                             = .false.,.false.,.false.,
periodic_y                          = .false.,.false.,.false.,
symmetric_ys                        = .false.,.false.,.false.,
symmetric_ye                        = .false.,.false.,.false.,
open_ys                             = .false.,.false.,.false.,
open_ye                             = .false.,.false.,.false.,
nested                              = .false., .true., .true.,
/

&grib2
/

&namelist_quilt
nio_tasks_per_group = 0,
nio_groups = 1,
/
```

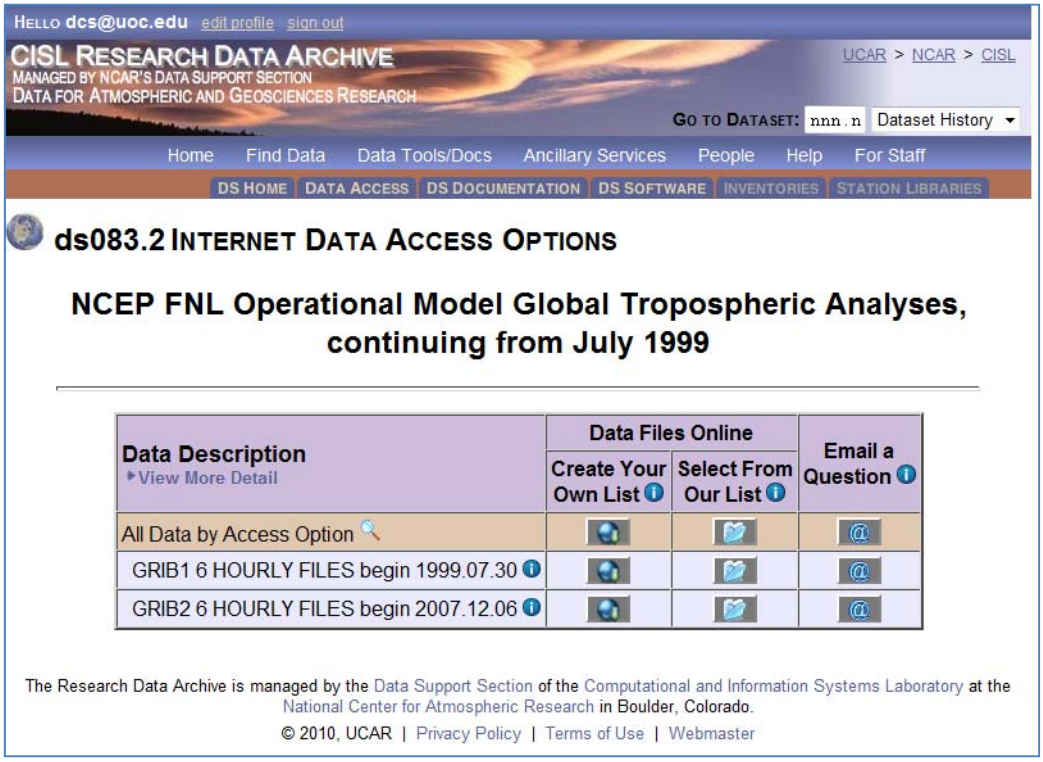
11.4 REGISTRE UCAR I OBTENCIÓ DE DADES

Per tal d'obtenir les dades meteorològiques utilitzades als assajos s'ha realitzat el registre al web www.ucar.edu.

Aquest registre és simple i s'ha realitzat completant les dades de l'enllaç https://dss.ucar.edu/cgi-bin/login.



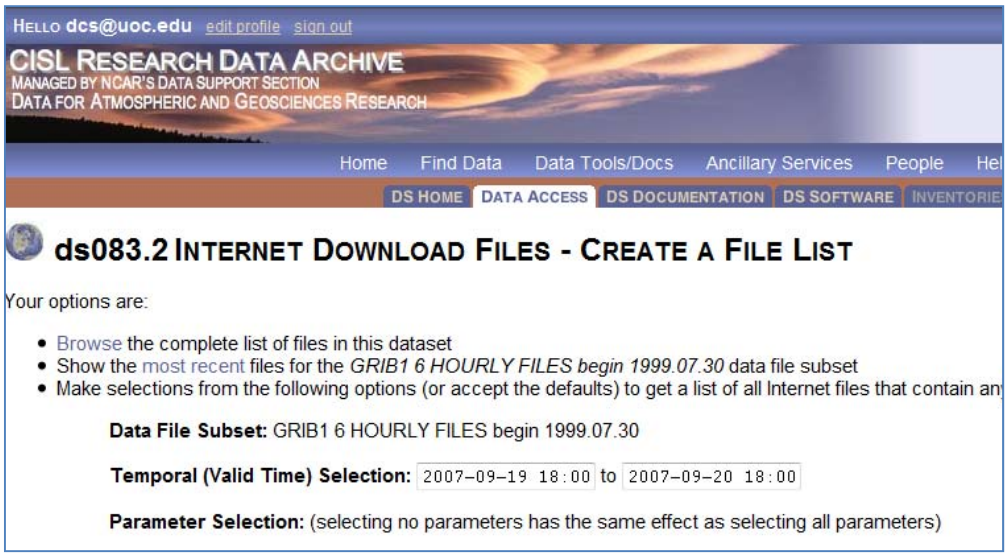
Els arxius utilitzats per WPS són recuperables tant pel sistema d'scripts que es descriu posteriorment com de manera directa mitjançant la plana web.

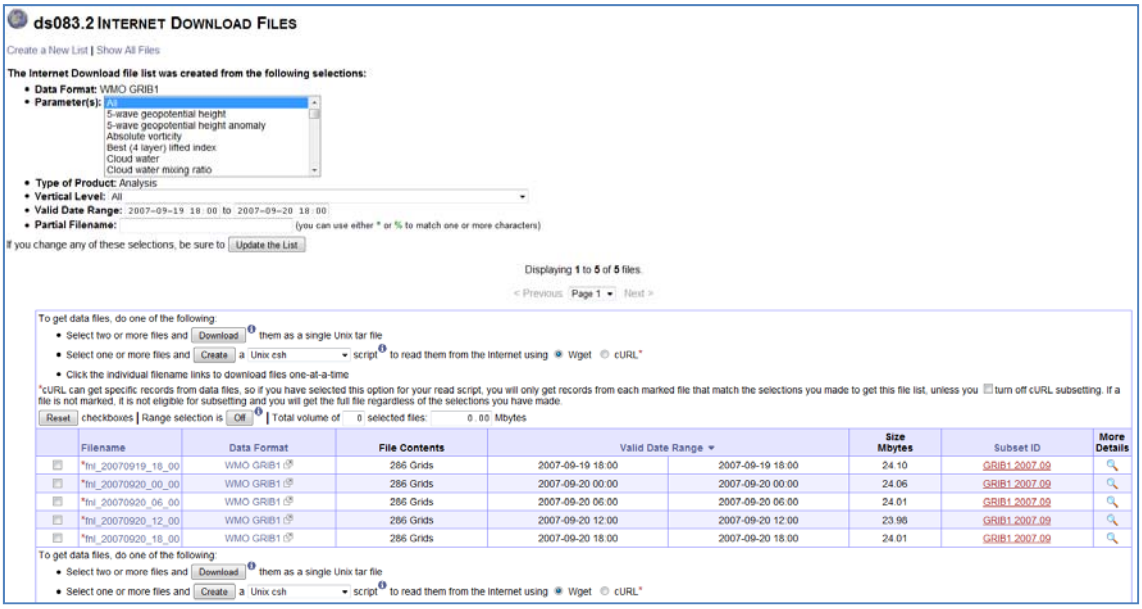


Els arxius necessaris son els següents:

```
fnl_20070919_18_00
fnl_20070920_00_00
fnl_20070920_06_00
fnl_20070920_12_00
fnl_20070920_18_00
```

A seguir es pot veure el procés d'obtenció mitjançant la plana web:





Paral·lelament aquest arxius també es poden descarregar directament mitjançant un script, fet que serà més ràpid que copiar-los posteriorment als diferents entorns:

```
export passwd = "xxxxxxxx"
export cert_opt = ""
export cert_opt = "--no-check-certificate"
wget $cert_opt -O /dev/null --save-cookies auth.dss_ucar_edu --post-
data="email=dcs@uoc.edu&passwd=$passwd&action=login" https://dss.ucar.edu/cgi-bin/login
wget $cert_opt $opts --load-cookies auth.dss_ucar_edu
http://dss.ucar.edu/datazone/dsszone/ds083.2/grib1/2007/2007.09/fnl_20070919_18_00
wget $cert_opt $opts --load-cookies auth.dss_ucar_edu
http://dss.ucar.edu/datazone/dsszone/ds083.2/grib1/2007/2007.09/fnl_20070920_00_00
wget $cert_opt $opts --load-cookies auth.dss_ucar_edu
http://dss.ucar.edu/datazone/dsszone/ds083.2/grib1/2007/2007.09/fnl_20070920_06_00
wget $cert_opt $opts --load-cookies auth.dss_ucar_edu
http://dss.ucar.edu/datazone/dsszone/ds083.2/grib1/2007/2007.09/fnl_20070920_12_00
wget $cert_opt $opts --load-cookies auth.dss_ucar_edu
http://dss.ucar.edu/datazone/dsszone/ds083.2/grib1/2007/2007.09/fnl_20070920_18_00
wget $cert_opt $opts --load-cookies auth.dss_ucar_edu
http://dss.ucar.edu/datazone/dsszone/ds083.2/grib1/2007/2007.09/fnl_20070921_00_00
rm auth.dss_ucar_edu
```

Aquest script ha estat utilitzat a l'entorn EC2 per descarregar les dades directament durant la fase de creació de la imatge base.

Els arxius descarregats són comuns en la execució de tots els entorns.

11.5 IMATGE BASE EC2

La imatge base utilitzada a EC2 esta basada en el model d'AMI 3e836657 disponible a Amazon EC2 i desenvolupat per Peter Skomoroch. (Font: http://www.datawrangling.com/on-demand-mpi-cluster-with-python-and-ec2-part-1-of-3).

Consisteix en un AMI amb sistema operatiu Fedora Linux 32bits versió 6 on posteriorment s'han instal·lat els següents extres:

python-devel, gcc, gcc-c++, gcc-gfortran, fftw-devel, swig, hdf5-devel, readline-devel, python-numeric, python-numarray , Pyrex, python-psyco, wxPython-devel, zlib-devel, freetype-devel, tk-devel tkinter, gtk2-devel, pygtk2-devel, libpng-devel i octave

A continuació s'han compilat i instal·lat Numpy, MPICH2, PyMPI i PyTables.

S'ha escollit aquesta imatge ja que presenta un disseny, orientat a MPI i a l'execució en clúster mitjançant els scripts definits a l'annex 11.2.

Aquesta imatge ha servit com a base de compilació del WRF i totes les seves llibreries.

11.6 XEN

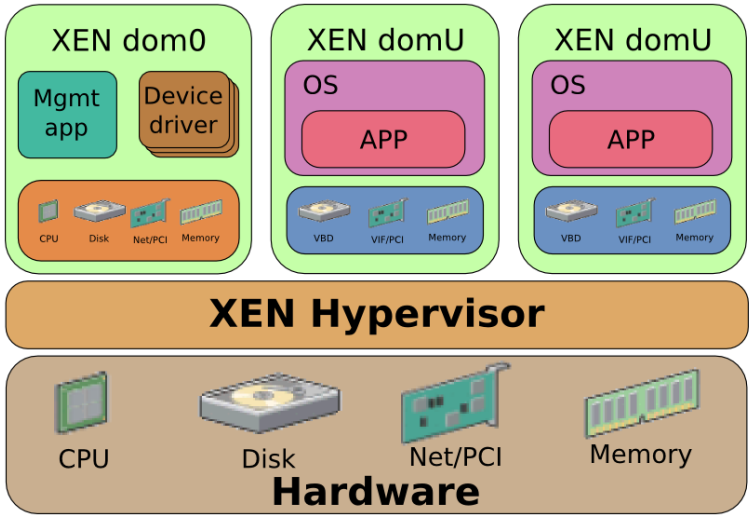
A continuació és descriu una petita aproximació a XEN.

La tecnologia XEN es tracta d'un sistema de màquina virtual desenvolupat per la Universitat de Cambridge. És un disseny de codi obert a través del que es poden executar instàncies de sistemes operatius amb totes les seves característiques.

A diferencia d'altres sistemes que simulen el hardware mitjançant programari, Xen requereix que els sistemes operatius siguin modificats explícitament per executar les seves API's (mantenint la compatibilitat amb les aplicacions d'usuari), fet que li permet un alt rendiment sense la necessitat de maquinari extra especial. Xen penalitza tan sols entre un 2% i un màxim del 8%.

Xen proporciona tots els sistemes d'aïllament i gestió de recursos, així com la possibilitat de migració de màquines virtuals en viu. Aquestes es realitzen en qüestió de milisegons entre equips físics sense la necessitat de parar-los.

La peça fonamental de Xen és el Hypervisor, que s'encarrega de controlar el hardware (CPU, memòria,...) i distribuir-lo entre les diverses màquines virtuals. Segueix el següent diagrama:



dom0 és el primer domini, iniciat quan arrenca el kernel modificat de Xen. Té els màxims privilegis i accés directe al hardware del host.

domU (dom1, dom2,…) són les altres màquines virtuals, sense privilegis. Sobre aquestes corren les diferents instancies de S.O. amb els seus programaris específics, que veuen el maquinari com si es tractés d'un hardware genèric.