MESOS.
Administrar clústers compartint recursos de forma dinàmica.

Joan Carles Martínez Rodríguez
Treball de Final Grau en Enginyeria Informàtica
Itinerari d'Arquitectura de Computadors i Sistemes Operatius

Consultor: Francesc Guim Bernat
Lliurament: 30/12/2014

UOC Universitat Oberta de Catalunya | 20 anys

Motivació del Treball

Conèixer en detall el funcionament intern dels grans clústers de computació respecte a:

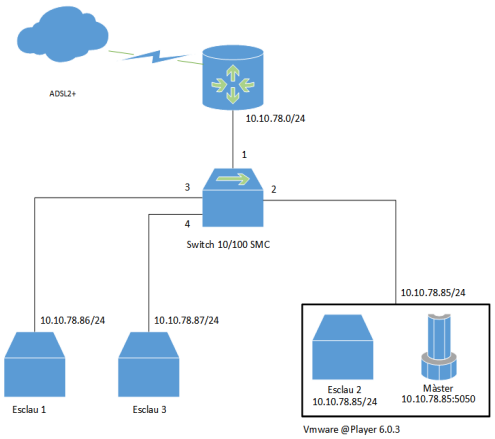
1. Conèixer problemàtiques i mecanismes d'eficiència en el tractament de grans volums de dades i computació.

2. Estudi de solucions actuals dedicades a la gestió eficient de recursos.

3. Implementació d'una solució real de gestió de clústers que pugui ser escalable.

4. Aplicar de forma pràctica els coneixements adquirits al llarg dels estudis.

Introducció al Treball



Parts principals:

- Estudi de l'estat de l'art en solucions de gestió eficient de recursos als clústers.
- Estudi, instal·lació i gestió d'un clúster administrat per MESOS.
- Obtenció de mètriques del clúster implementat i conclusions de la gestió de MESOS.
- Conclusions finals del treball desenvolupat.

Objectius del Treball

1. Conèixer les bases del disseny de clústers de computació i les millores d'eficiència.

2. Estudi del gestor de clústers MESOS, en les següents funcions:

    2.1 Creació d'un clúster totalment operatiu.

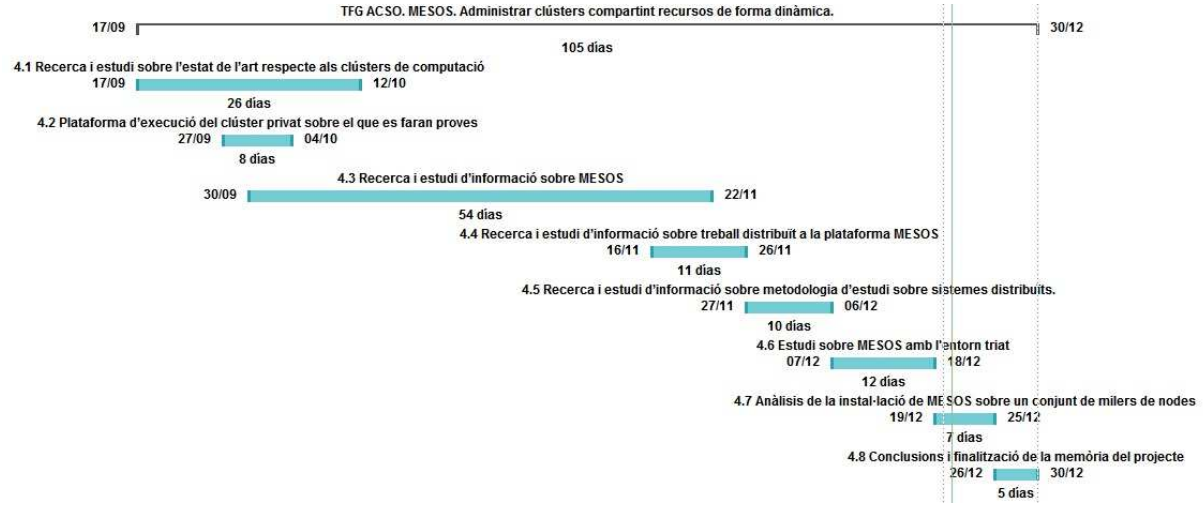
    2.2 Administració del clúster amb la plataforma MESOS.

    2.3 Execució activa del clúster utilitzant un dels entorns d'aplicació.

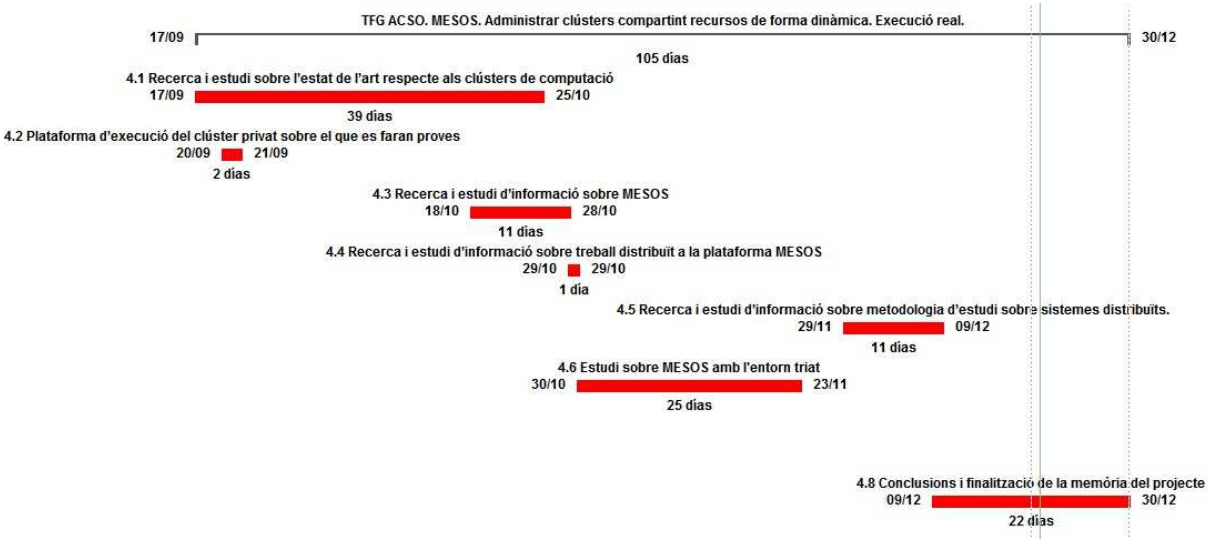
    2.4 Projecció en el control de milers de nodes per a treball distribuït.

3. Extracció de conclusions i comparació amb altres gestors actuals.

Planificació temporal

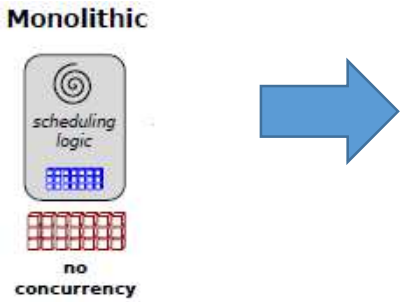


Planificació inicial
Previst: 408 hores

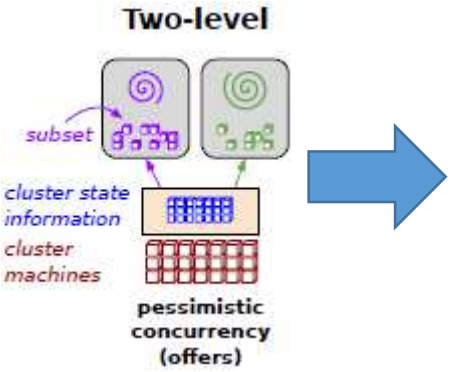


Planificació real
Cost: 509 hores

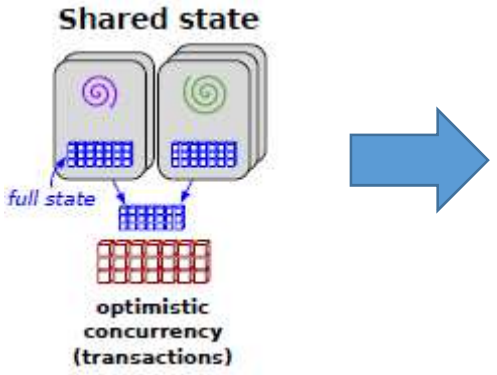
Estat de l'art. Exemples de gestió de clústers



Utilitzen un únic algorisme de planificació per a tots els treballs.

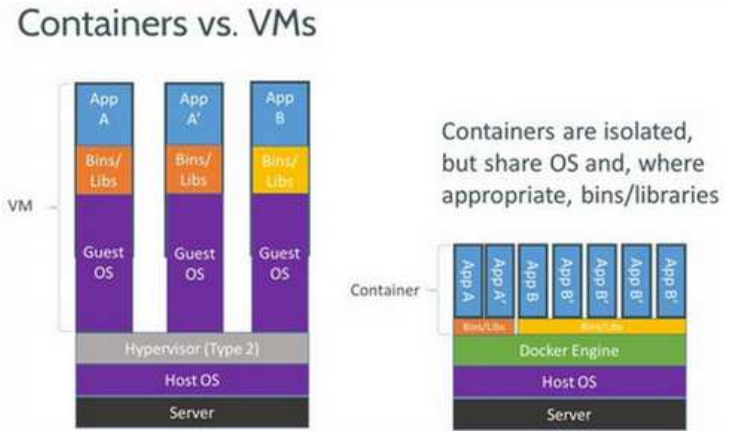


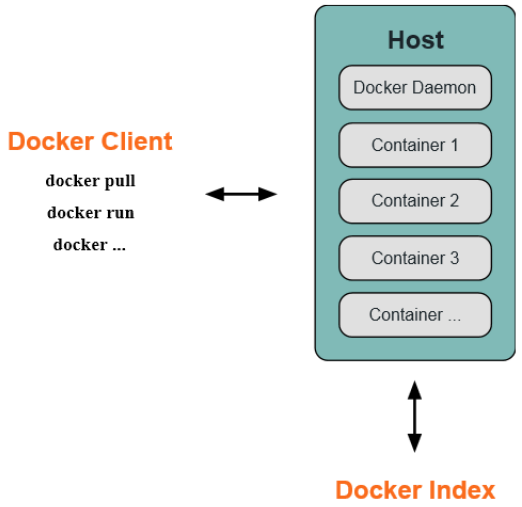
Té un únic gestor de recursos que ofereix recursos a múltiples entorns d'aplicació que s'executen en paral·lel. **MESOS** pertany a aquest tipus.



Està basat en un control de concurrència optimista i sense bloquejos, de forma que es té accés a tots els recursos del clúster de forma simultània.

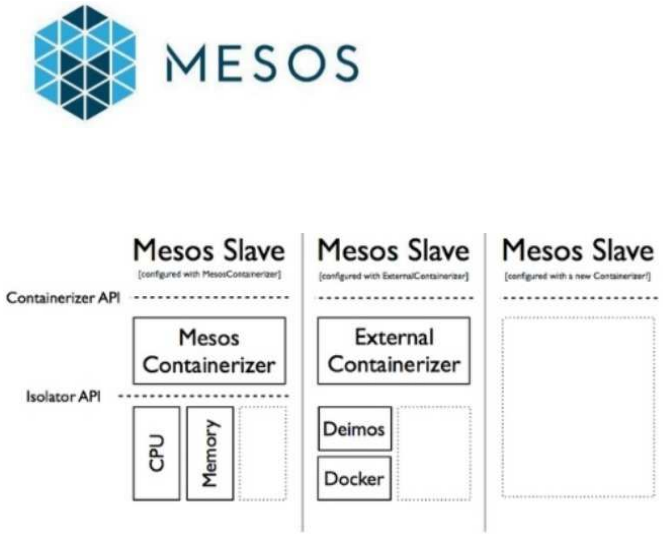
Estat de l'art. Contenidors DOCKER

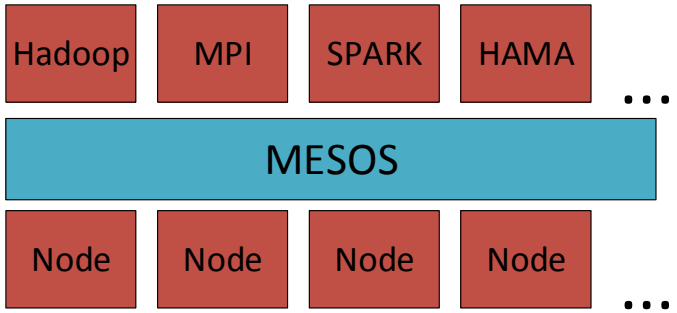




- Nou paradigma en virtualització.
- Menys consum de recursos que amb VM.
- Aïllament entre aplicacions.
- Ideal per aplicacions que utilitzen el mateix sistema operatiu.
- Arquitectura basada en client-servidor (dimonis).
- Utilitza plantilles d'aplicació (imatges).

Què és MESOS?





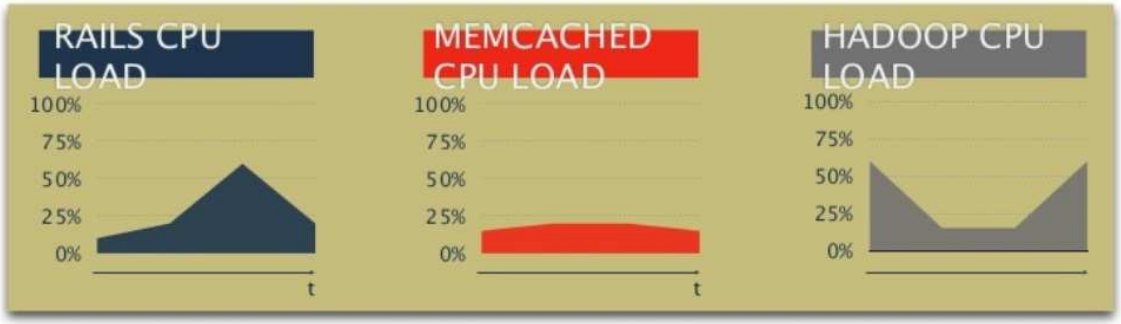
- És un sistema de *kernel* distribuït.

- Les aplicacions a executar utilitzen entorns d'aplicació (frameworks).

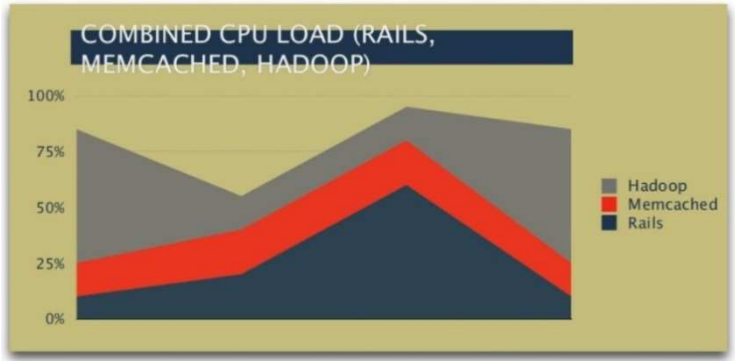
Capacitats:

- Escalable fins 10.000 nodes.
- Capacitat de recuperació d'errors.
- Contenidors LINUX o DOCKER.
- Planificació multirecurs (memòria, CPU, disc i ports).
- API Java, Python i C++ APIs, per desenvolupar noves aplicacions paral·leles.
- Web UI per visualitzar l'estat del clúster..

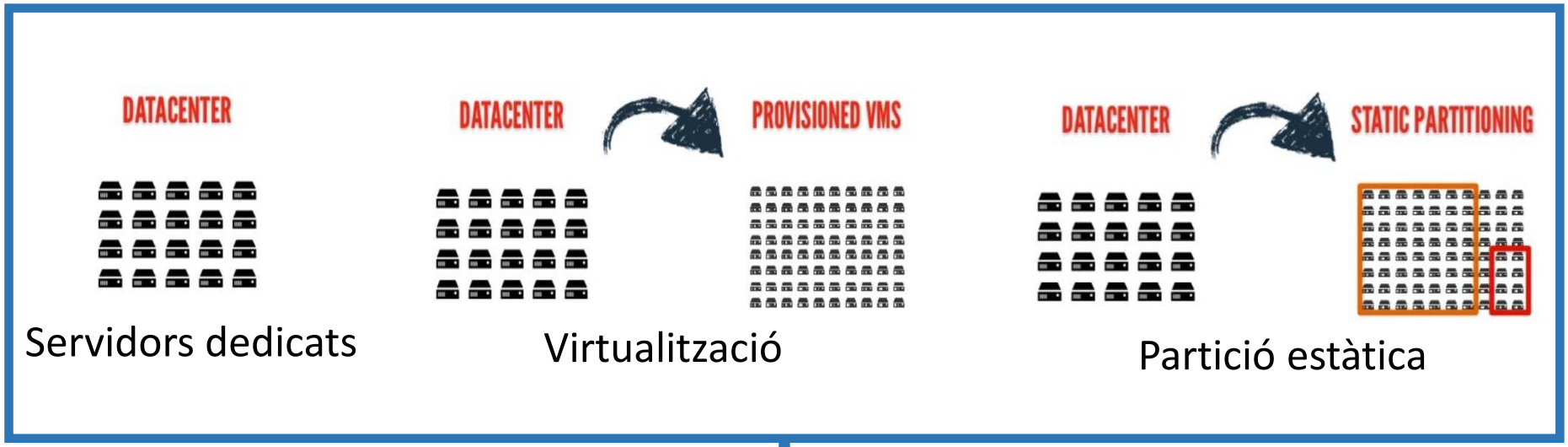
MESOS. Compartir recursos de forma dinàmica.



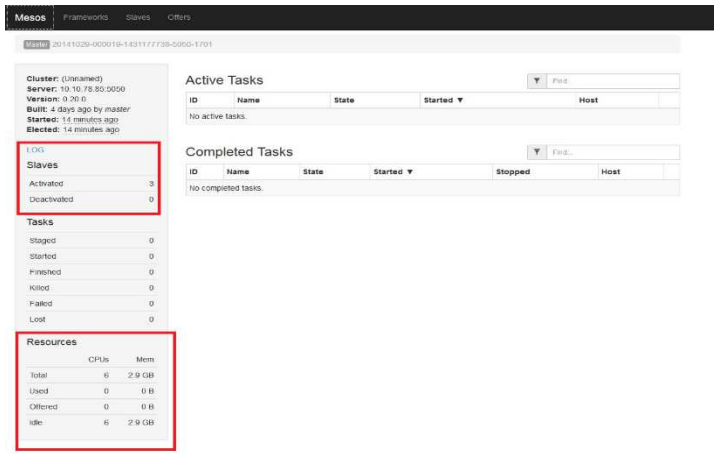
Entorn que s'adapti a les necessitats de **forma dinàmica i flexible** o recursos dedicats de forma estàtica?



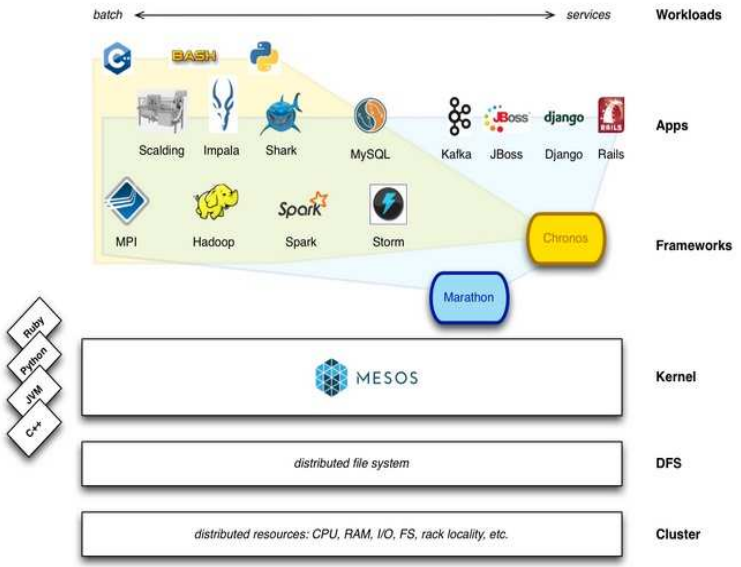
MESOS. Evolució tecnològica



Solució: Visió simplificada del clúster a administrar (*pool de recursos*)



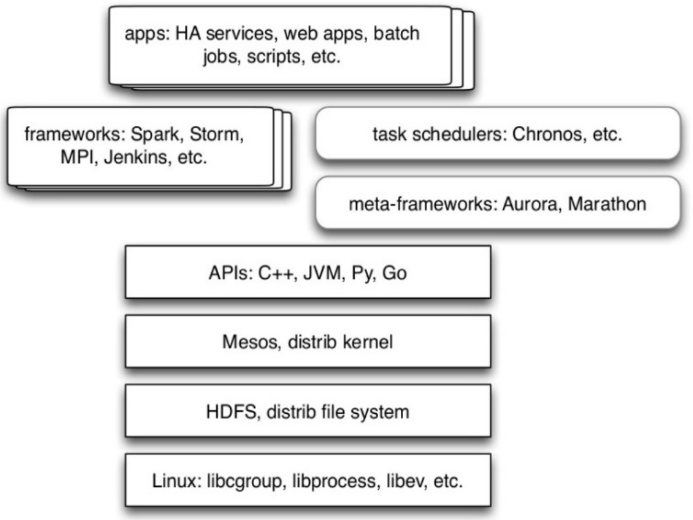
MESOS. Arquitectura



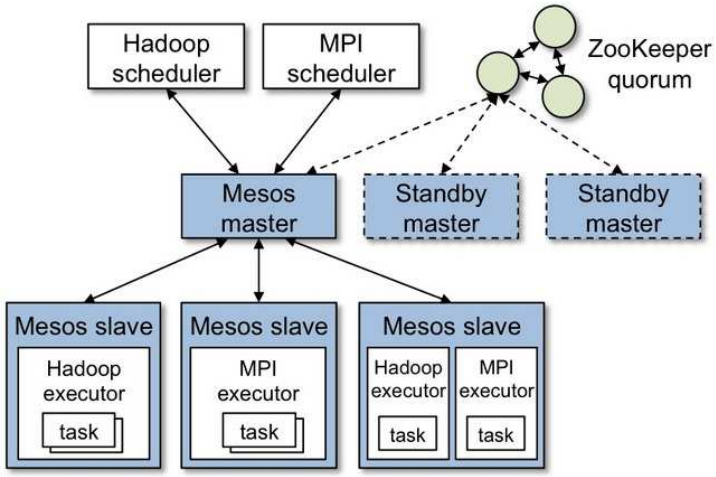
1. **MESOS** ofereix recursos als entorns d'aplicació (*frameworks*).

2. Els entorns d'aplicació decideixen si accepten els recursos.

MESOS resol els detalls tècnics de baix nivell de la interfície, i s'espera que llibreries d'alt nivell resolguin la resta de problemes.

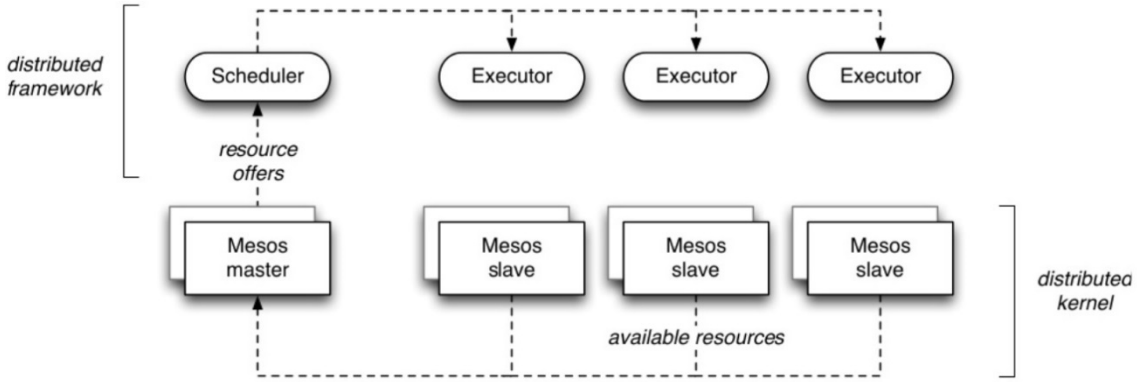


MESOS. Components i funcionament



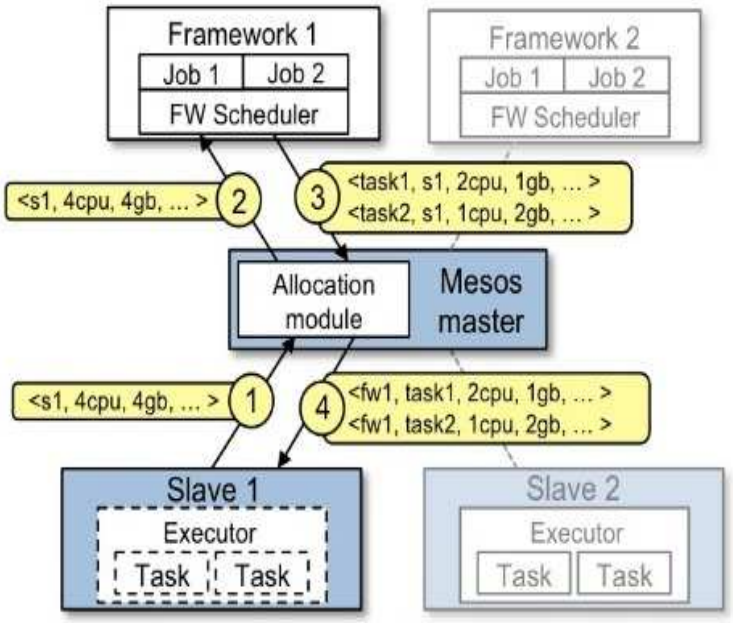
Els components bàsics de MESOS són el node màster i els nodes esclaus.

La resiliència existeix a nivell de node màster i autorecuperació dels nodes esclaus.



Gestió: el node màster ofereix els recursos lliures al planificador de cada entorn d'aplicació.

MESOS. Components i funcionament



Oferiment i assignació dinàmica de recursos segons les necessitats dels entorns d'aplicació:

1. Els esclaus indiquen els recursos lliures.
2. El màster informa als entorns d'aplicació dels recursos lliures
3. Els planificadors dels entorns d'aplicació accepten o rebutgen els recursos oferts.

MESOS. Entorns d'aplicació disponibles

Processar Big Data
(Big Data Processing)
Cray Chapel
Dpark
Exelixi
Hadoop
Hama
MPI
Spark
Storm

Serveis de llarga durada
(Long running)
Aurora
Marathon
Singularity
SSSP

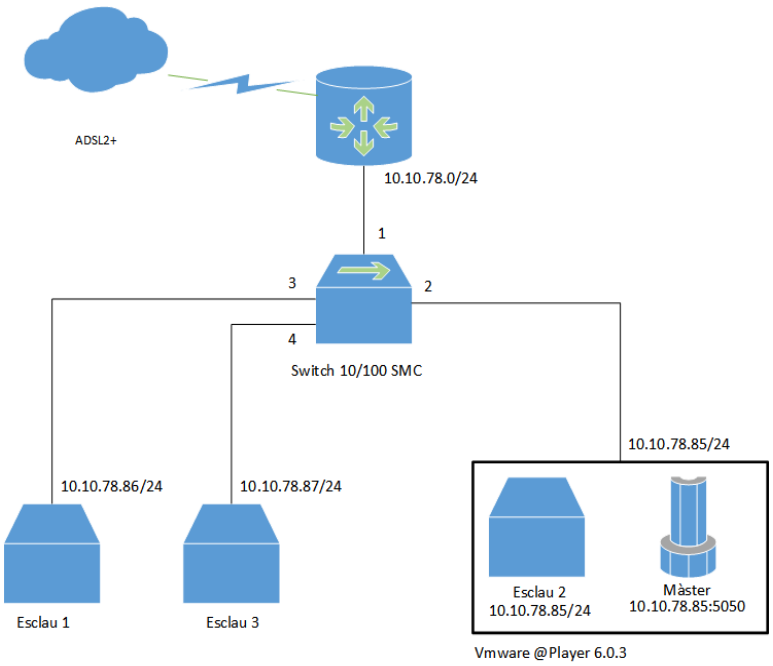
Planificació per lots
(Batch scheduling)
Chronos
Jenkins
JobServer
Torque

Emmagatzematge de dades
(Data storage)
Cassandra
ElasticSearch
Hypertable

MESOS. Utilització actual?

Airbnb
Artirix
Atigeo
Atlassian
Branding Brand
Categorize
CloudPhysics
Conviva
CorvisaCloud
Coursera
CRP-Gabriel
Lippmann
Daemon
Devicescape
DueDil
eBay
GoCardless
Groupon
HubSpot
Ignidata
iQIYI
LIFX
Localsensor
Magine TV
Medidata
Solutions
meemo
MediaCrossing
Mesosphere
Netflix
OakmoreLabs
OpenCredo
OpenTable
Oscar Insurance
PayPal
Pinkbike
ProfitStars
Qubit
Revisely
Sailthru
Sharethrough
Sigmoid
Analytics
SiQueries
Squarespace
The Factory
Twitter
UCSF
UC Berkeley
URX
Viadeo
Vimeo
Virdata
Wizcorp
WooRank
Yieldbot
Xogito

Clúster físic privat



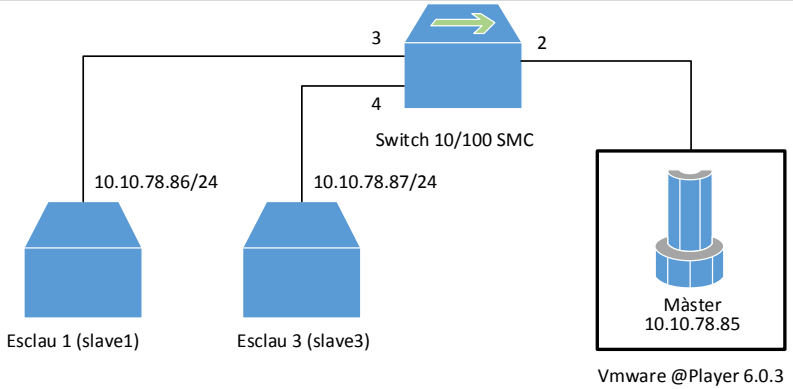
Maquinari utilitzat

- Un PC Intel (R) core (TM)2 Quad CPU Q6600@2.4 GHz, 3.25 GB DRAM.
- Dos PC AMD 64 3200+ 2 GHz, 2 GB DRAM.
- Un commutador (*switch*) de 8 ports a 10/100 Mbps marca SMC, model SMCFS8
- Un enrutador (*router*) ADSL2+ amb xarxa DHCP 10/100 Mbps

Programari utilitzat

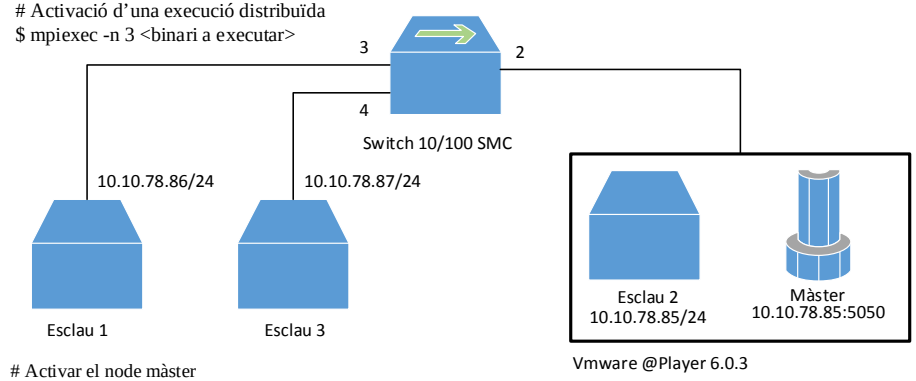
- Sistema operatiu Ubuntu server versió 14.04
- MPICH2 de l'Argonne National Laboratory
- NPB 3.3, mètrica IS, de NAS Parallel Benchmark
- PERFTEST-1.5
- Gnuplot
- VMWare© Player, versión 6.0.3
- MESOS 0.20.0
- Java 1.7

Mètriques. Configuració dels clústers



Clúster **MPD**. Permet executar aplicacions MPI de forma directa.

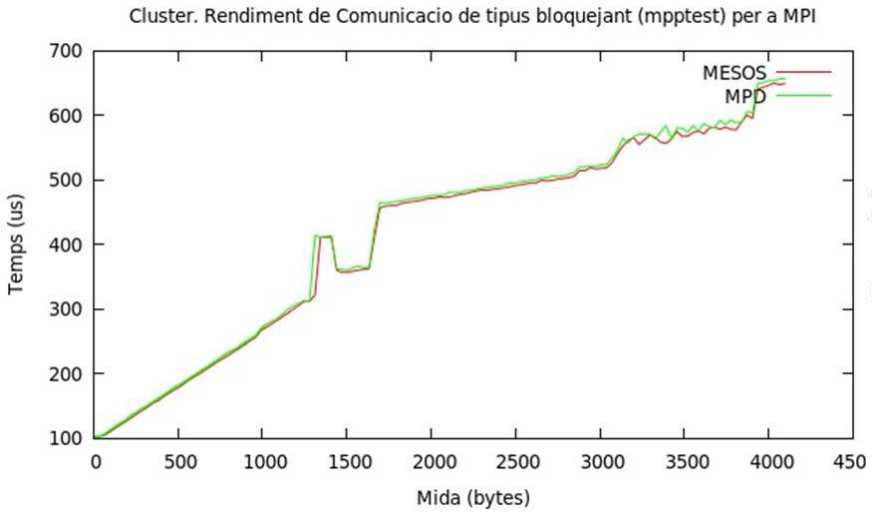
```
# Activació d'anell de dimonis MPD
$ mpdboot -n 3 -f mpd.hosts

# Contingut del fitxer mpd.hosts
slave1
slave3

# Activació d'una execució distribuïda
$ mpiexec -n 3 <binari a executar>
```

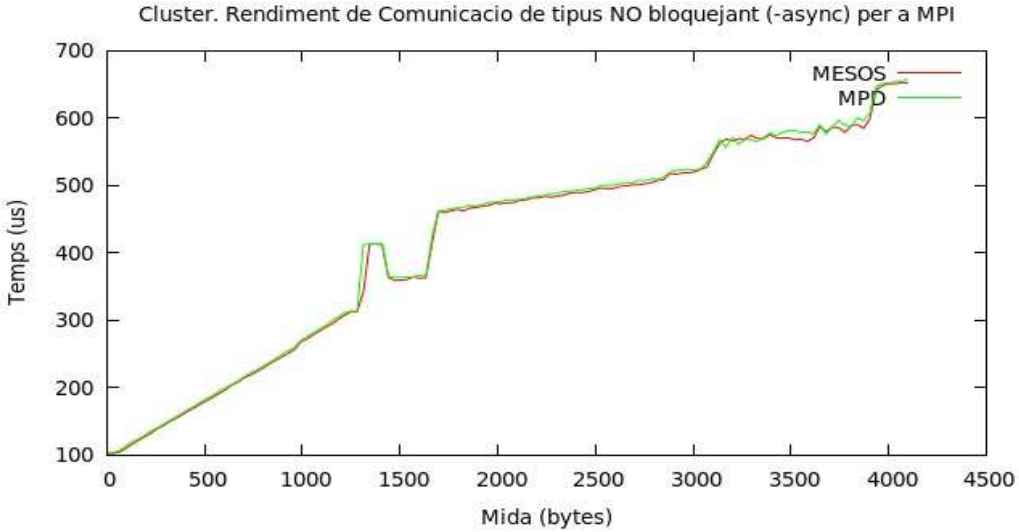


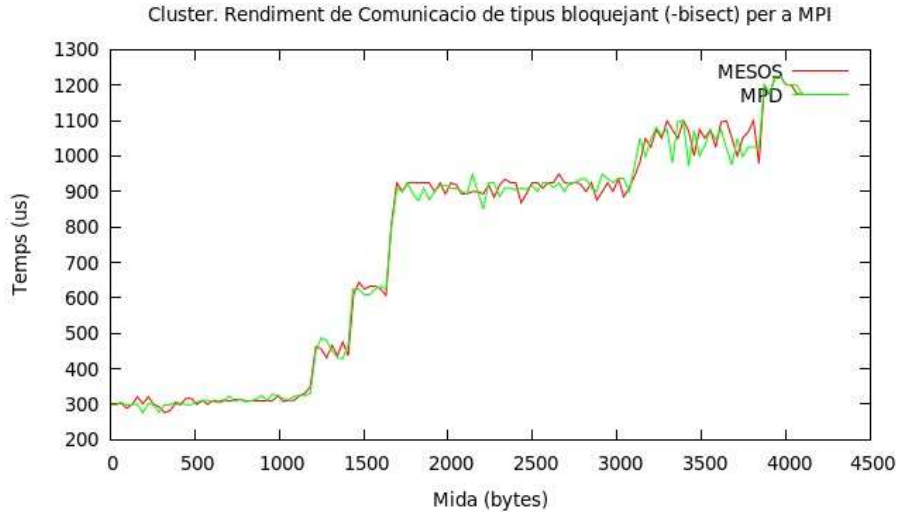
Clúster **MESOS** amb entorn d'aplicació MPI.

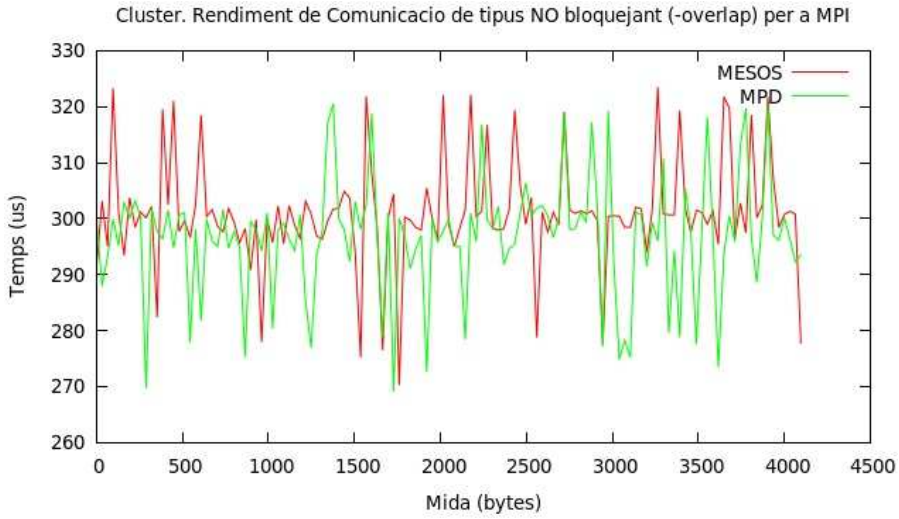
```
# Activar el node màster
$ ./bin/mesos-master.sh --ip= 10.10.78.85 --work_dir=/var/lib/mesos

# Activar node esclau.
$ ./bin/mesos-slave.sh –ip= <ip node esclau> --master=10.10.78.85:5050 --resources: “mem(*):1512;cpus(*):1” --no-switch_user

# Activar treball MPI a node màster
$ mpiexec-mesos –n 3 -m 1512 10.10.78.85:5050 ./<binari a executar>

# WEBUI
10.10.78.85:5050
```

Mètriques. Resultats MPI PERFTEST-1.5

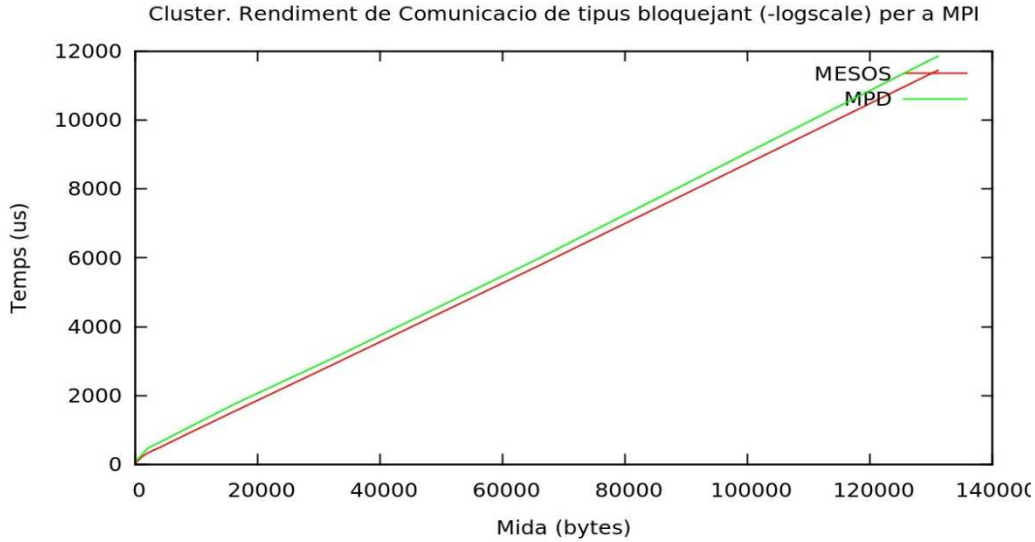


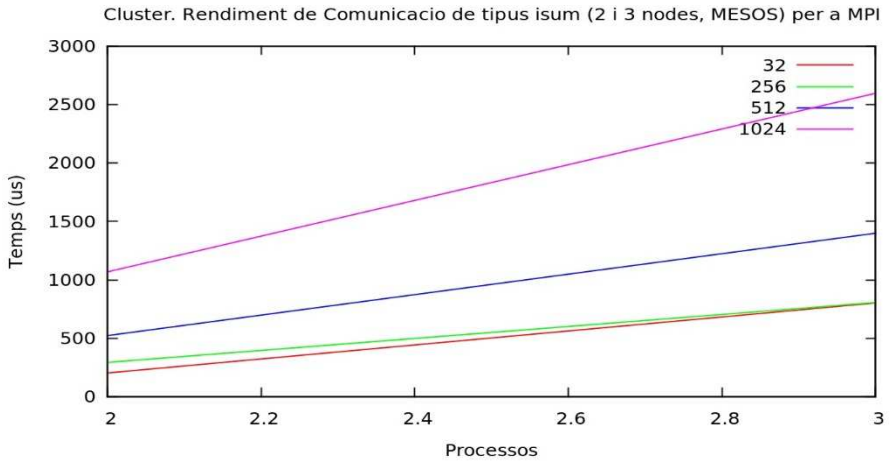


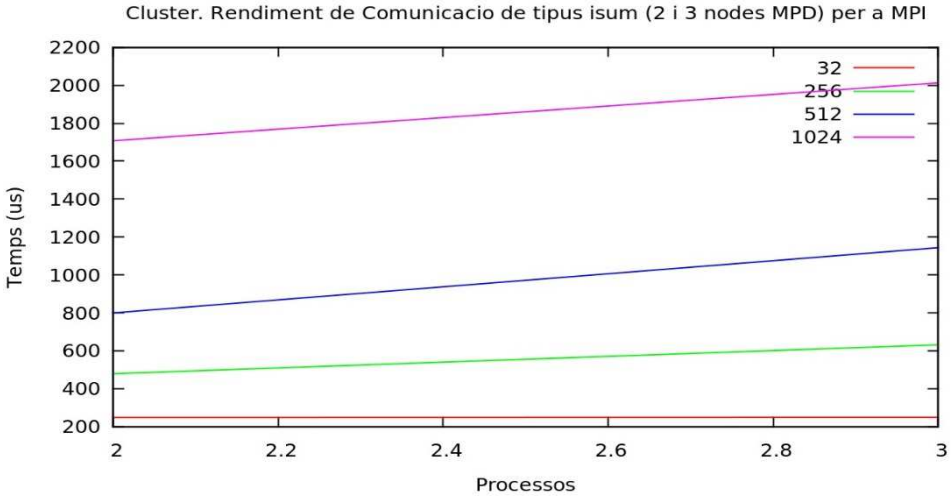




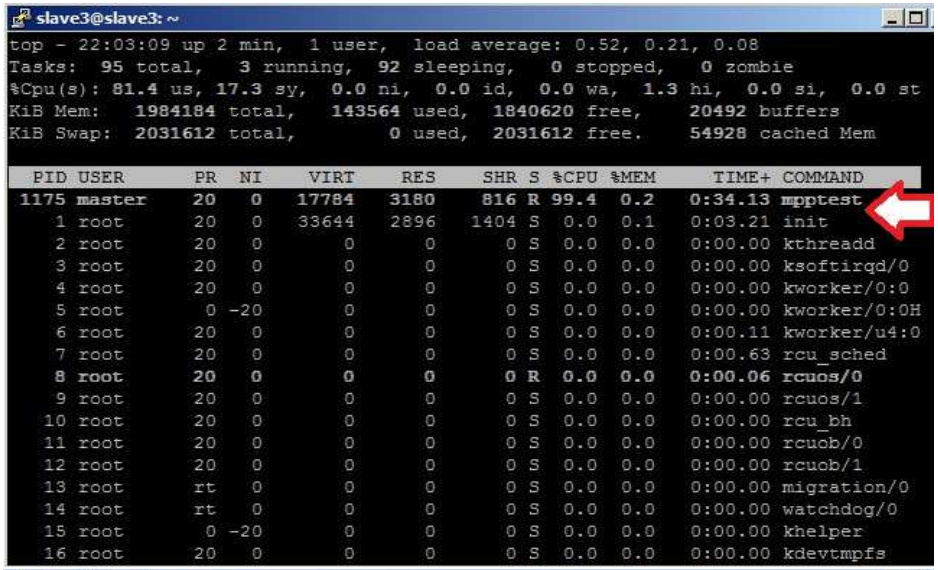
Mètriques. Resultats MPI PERFTEST-1.5



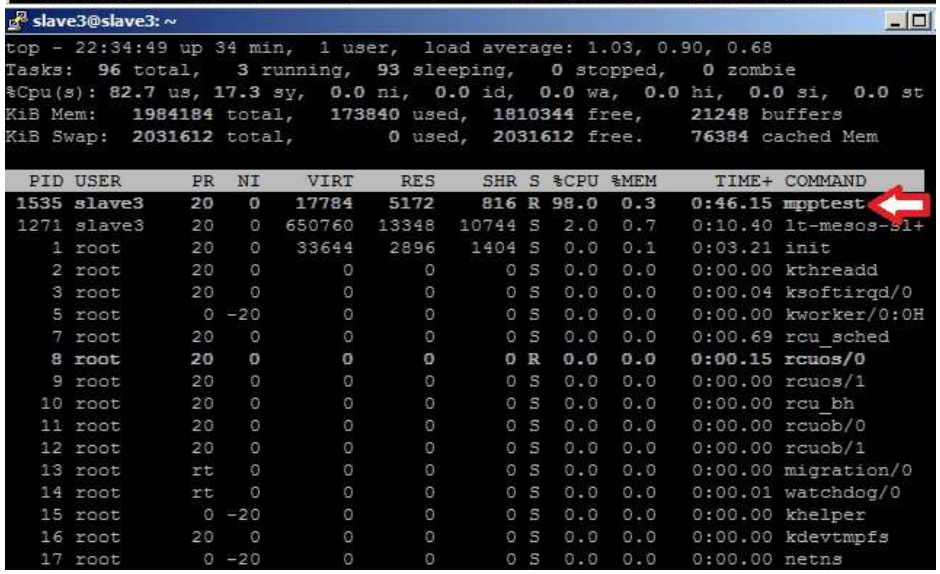




Mètriques. Consum de recursos als clústers



Consum de recursos en esclau clúster **MPD**.

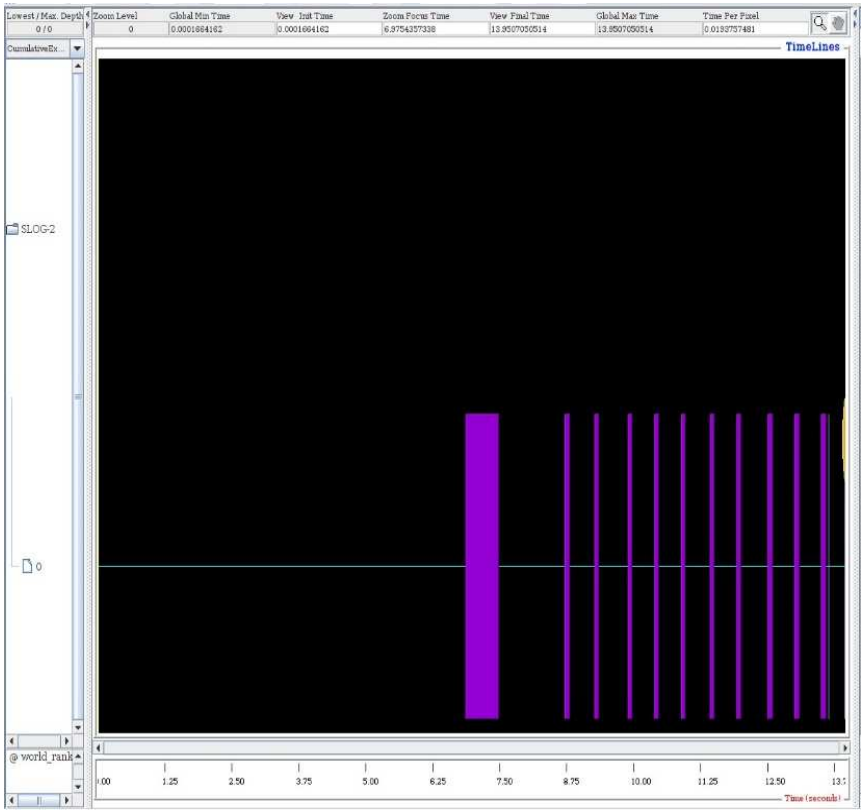


Consum de recursos en esclau clúster **MESOS**.

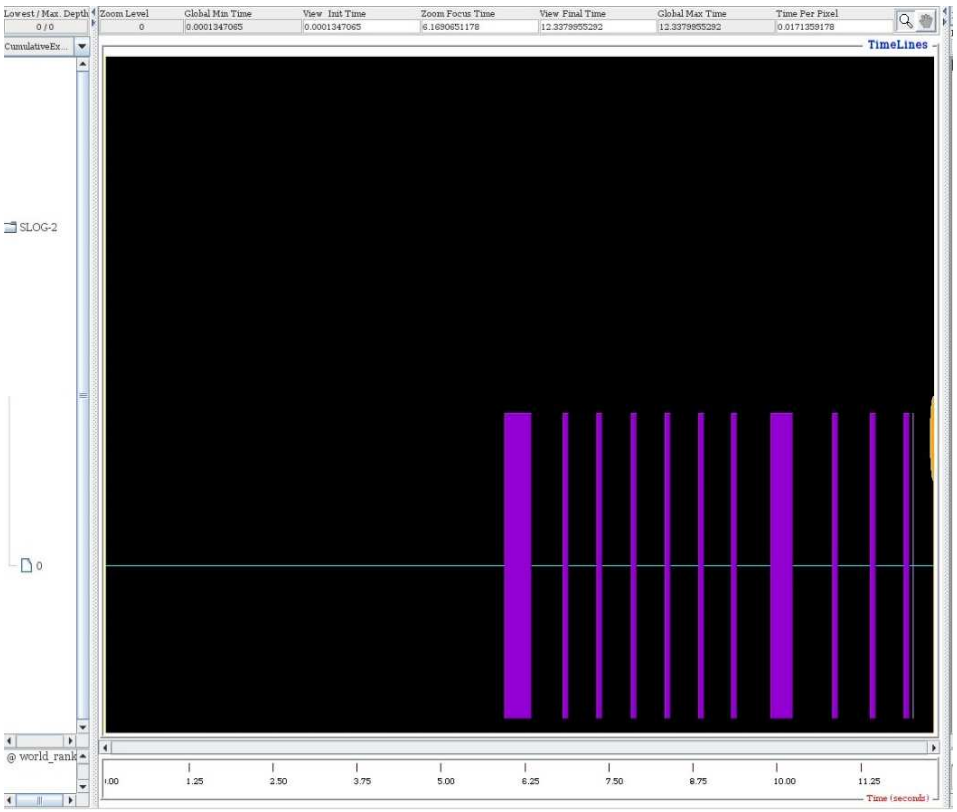
Font: programa TOP. Ubuntu 14.04

Mètriques. Execució d'aplicació MPI NPB-IS

Aplicació MPI executada: **is.B.1**



Temps d'execució clúster **MESOS**: **13,95 S**

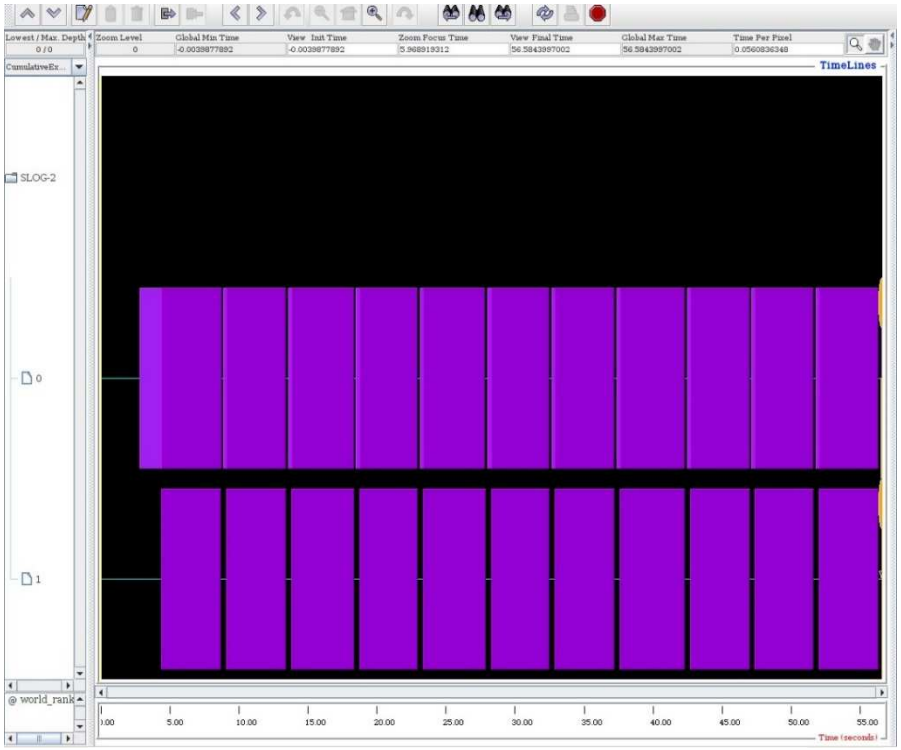


Temps d'execució clúster **MPD**: **12,33 S**

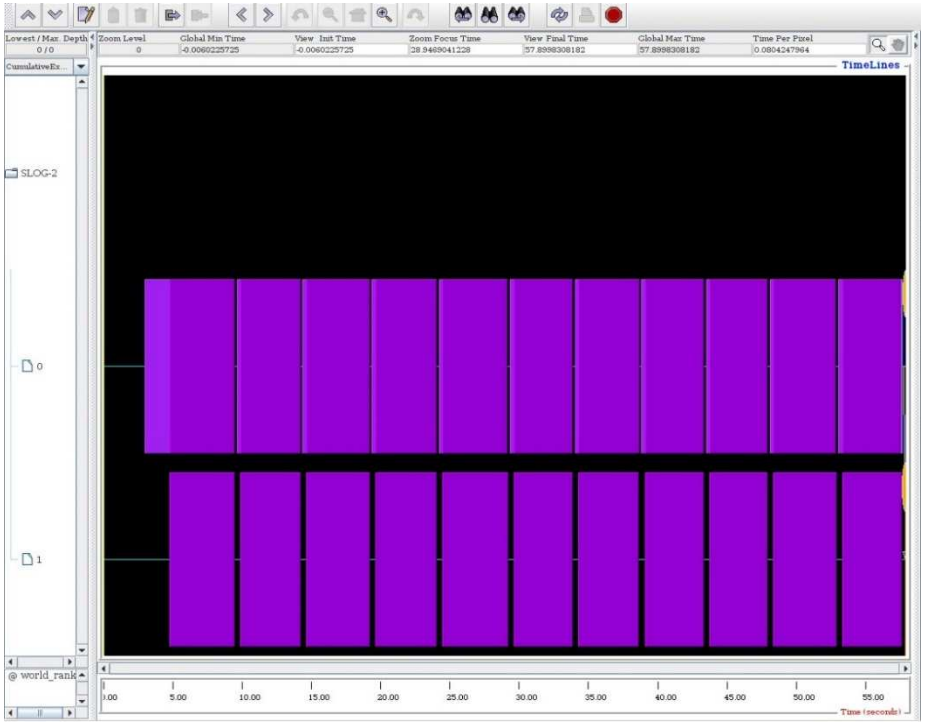
Font: programa JUMPSHOT (MPICH2).

Mètriques. Execució d'aplicació MPI NPB-IS

Aplicació MPI executada: **is.B.2**



Temps d'execució clúster **MESOS**: **56,58 S**



Temps d'execució clúster **MPD**: **57,89 S**

Font: programa JUMPSHOT (MPICH2).

Conclusions de les mètriques

Rendiment del clúster: proves de **traspàs de missatges MPI** amb la suite *PERFTEST-1.5* de *l'Argonne National Laboratory.*

Rendiment en execució d'aplicació MPI: prova amb 1 i 2 processadors de l'aplicació **IS.B** del *NAS Parallel Benchmark.*

Consum de recursos en execució de rendiment del clúster: prova ***MPPTEST***, suite *PERFTEST-1.5* de *l'Argonne National Laboratory.*

L'execució sota la gestió de MESOS **no penalitza ni afavoreix,** respecte a no fer servir MESOS.

Conclusions de les mètriques

Després d'observar els resultats de les mètriques, és útil **MESOS**?

- Sí, reutilitza recursos lliures de forma dinàmica per guanyar eficiència.
- Sí, podem executar en paral·lel diverses aplicacions, i que poden ser diferents.
- Sí, aporta mecanismes d'alta disponibilitat davant d'errors.
- Sí, aporta mecanismes de planificació dels treballs de forma nativa.
- Sí, podem controlar amb *WebUI* que passa al clúster en temps real.

Conclusions del treball

- Millor coneixement de les problemàtiques de gestió de clústers de computació i les seves solucions.

- Respecte al coneixement adquirit del gestor MESOS:
  - Apreciació del model tecnològic de dos nivells en que està basat.
  - Gestor estable i robust, i no necessita grans recursos operatius.
  - Observada la gestió de recursos de forma dinàmica per oferiment.
  - Permet conviure maquinaris no homogenis en cas necessari.
  - El desplegament és senzill i ràpid.
  - Conèixer el significat d'entorn d'aplicació (*framework*)
  - Estudi de mètriques del clúster.
  - Observació en temps real del funcionament del clúster.

- Respecte al coneixement general adquirit:
  - Observació de treball de computació paral·lela distribuïda.
  - Configuració de sistemes operatius, xarxes i aplicacions.

Conclusions del treball. Estudi de futur

Desplegar **MESOS** en un clúster de desenes de servidors, per observar.

- ***Apache Zookeeper*** i les opcions de clúster d'alta disponibilitat.
- Eines de control i programació del clúster (***Chronos***).
- El comportament de la xarxa analitzant les comunicacions entre nodes.
- El funcionament dels scripts d'iniciació automàtica dels nodes del clúster.
- Altres entorns d'aplicació diferents d'MPI (***Spark***, ***Hadoop***, …)

Altres temàtiques interessants:

- Estudi de realització d'un entorn d'aplicació concret.
- En un futur pròxim: **MESOSPHERE DCOS**.

Referències de les imatges utilitzades

[0][2][7][14][19][20][21][22][23] Font: http://mesos.apache.org/

[1] Font: http://upload.wikimedia.org/wikipedia/commons/2/29/Virginia_tech_xserve_cluster.jpg

[5] Font: http://static.googleusercontent.com/media/research.google.com/es//pubs/archive/41684.pdf

[6] Font: http://www.zdnet.com/article/what-is-docker-and-why-is-it-so-darn-popular/

[6] Font: https://docs.docker.com/introduction/understanding-docker/

[7] Font: https://amplab.cs.berkeley.edu/projects/mesos-dynamic-resource-sharing-for-clusters/

[7][8][9][10][11] Font: http://www.slideshare.net/pacoid/datacenter-computing-with-apache-mesos

[11][12] Font: http://people.csail.mit.edu/matei/papers/2011/nsdi_mesos.pdf

[13] Font: mesosphere.com/docs/frameworks

[16][19][20][21][22][23] Font: http://www.anl.gov/ http://www.nas.nasa.gov/publications/npb.html

[17][18] Captures programa **GNUPLOT** sobre mètriques **PERFTEST-1.5**

[19] Captura d'execució programa **TOP**, Ubuntu Server 14.04.

[20][21] Captures de l'aplicació **JUMPSHOT** sobre execucions MPI amb llibreria MPE.