# DISSENY I IMPLEMENTACIÓ D'UN CLÚSTER HPC

***ARQUITECTURA DE COMPUTADORS I SISTEMES OPERATIUS.***

GRAU D'ENGINYERIA INFORMÀTICA,
ENGINYERIA DE COMPUTADORS.
Presentació Treball Final de Grau

***Autor:*** **Javier Ortega Martín**

***Tutor:*** **Francesc Guim Bernat**

**25 de juny del 2014**

# Contingut

- **Introducció**
- Motivació
- Objectius i Planificació
- Anàlisi
- Disseny
- Implementació
- Conclusió
- Línies d'actuacions futures
- Demostració

# Introducció

- High Performance Computing (HPC) o computació d'altes prestacions fa referència a grans grups de processadors amb un sistema d'enviament de treballs basat en cues, per proporcionar la solució als problemes que fan un ús intensiu del càlcul.

- En general disposen de molts programes de codi obert, paquets de programari comercials o codi propi dels usuaris preparats per la seva execució.

# Contingut

- Introducció
- **Motivació**
- Objectius i Planificació
- Anàlisi
- Disseny
- Implementació
- Conclusió
- Línies d'actuacions futures
- Demostració

# Motivació

- Renovació del maquinari de l'antic clúster.
- Actualització del programari.
  - Canvi de distribució de SuSE SLES per Debian 7.
  - Actualització de Matlab.
  - Actualització llibreries dels desenvolupadors.
- Instal·lació d'un nou gestor de cues.
- Integració dels usuaris.
- Aprofitament dels recursos energètics, reduint la despesa d'energia per sistemes més optims.

# Contingut

- Introducció
- Motivació
- **Objectius i Planificació**
- Anàlisi
- Disseny
- Implementació
- Conclusió
- Línies d'actuacions futures
- Demostració

# Objectius i Planificació

- Proporcionar a l'usuari un entorn amb el qual pot:
  - Programar i executar tasques.
  - Definir reserva de recursos.
  - Emprar programari sota llicències corporatives.
  - Sistemes d'alta disponibilitat amb un alt rendiment i totalment escalable.
  - Sistemes amb equilibri de càrrega.

# Objectius i Planificació

- Diagrama de Gantt de amb les principals fases del projecte:

| Nombre | Fecha de inicio | Fecha de fin |
|---|---|---|
| PROJECTE | 17/03/14 | 26/05/14 |
| FASE 0: Disseny de l'infraestructura | 17/03/14 | 17/03/14 |
| FASE 1: Preparació de l'entorn | 18/03/14 | 21/03/14 |
| FASE 2: Configuració inicial | 24/03/14 | 25/03/14 |
| FASE 3: Configuració dels Serveis | 26/03/14 | 22/04/14 |
| FASE 4: Estudi rendiment del sistema | 23/04/14 | 2/05/14 |
| FASE 5: Documentació | 5/05/14 | 23/05/14 |
| FASE FINAL: Posada en producció | 26/05/14 | 26/05/14 |



* Les diferents fases queden detallades a la memòria del projecte.

# Contingut

- Introducció
- Motivació
- Objectius i Planificació
- **Anàlisi**
- Disseny
- Implementació
- Conclusió
- Línies d'actuacions futures
- Demostració

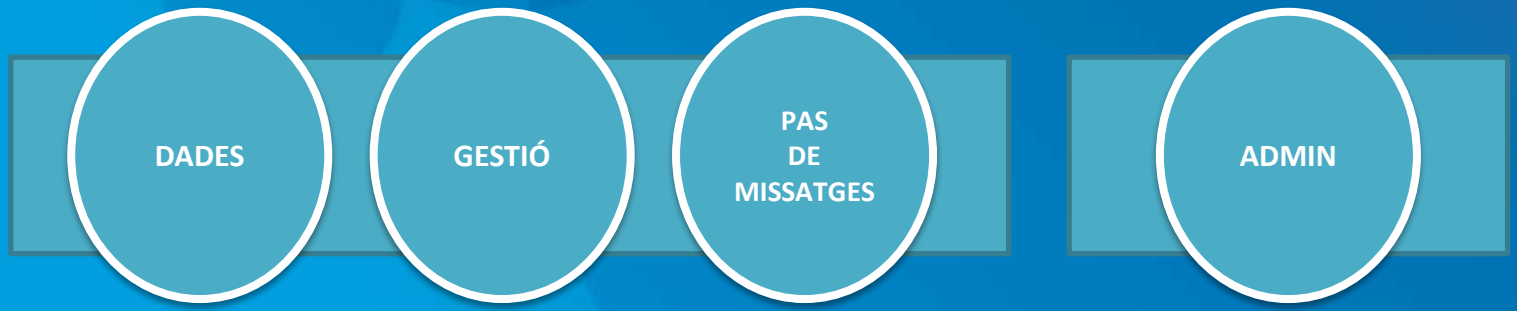
# Anàlisi

- Identificació de les fites:
  - Definició de la infraestructura.
    - Serveis de xarxa.
    - Sistema d'emmagatzemament.
    - Sistema de virtualització.
    - Sistemes de còmput.
  - Definició dels serveis i roles dels sistemes.
  - Estudi de rendiment dels sistemes de còmput.

# Anàlisi

- Primera fita: Definició de la xarxa.

S'identifiquen dos elements d'electrònica de xarxa, per tal de satisfer les comunicacions del clúster de servidors i així garantir i distingir el tràfic de dades, gestió i pas de missatges del d'administració.



# Anàlisi

- Segona fita: Definició del servidor primari.

El servidor principal allotja màquines virtuals per tal de donar suport als diferents serveis del clúster, els sistemes son:

- Accés per als usuaris.
- Gestor de cues i tasques.
- Desplegament de paquets.
- Intermediari d'accés a Internet.
- Monitoratge de la infraestructura.

# Anàlisi

- Tercera fita: Definició del servidor còmput.

El servidors de còmput processen les tasques dels usuaris, aquests reben les tasques mitjançant el sistema de cues centralitzat.

Sistemes homogenis:

- Mateix maquinari.
- Mateix programari.
- Mateixes configuracions.

# Anàlisi

- Quarta fita: Definició de l'emmagatzemament.

Aprofitament del sistema d'emmagatzemament corporatiu, centralitzat per tal de donar suport als diferents elements del clúster.

# Anàlisi

- Cinquena fita: Estudi de rendiment.

Bateria de proves per fer un anàlisi del potencial de clúster i així validar i confirmar les qualitats. S'empren eines per fer un ús intensiu dels recursos mitjançant llibreries matemàtiques.

# Contingut

- Introducció
- Motivació
- Objectius i Planificació
- Anàlisi
- **Disseny**
- Implementació
- Conclusió
- Línies d'actuacions futures
- Demostració

# Disseny

- Elements de la xarxa, (1 lògic i 2 físics)
  - IPTABLES, Gestió de la comunicació de les màquines virtuals mitjançant NAT i IP FORWADING.

  - Switch de dades, dedicat a la comunicació de dades amb la cabina, comunicació interna del clúster i al protocol de pas de missatges MPI.

Netgear GS752TXS-100

  - Switch de gestió IPMI, dedicat a la gestió dels nodes de còmput.

Netgear GS724TS

# Disseny

- Principals elements del servidor primari.

Servidor de màquines virtuals per als serveis més crítics del clúster HPC.

Principals elements:

- Processador: Opteron Abu Dhabi 6344 a 2,6Ghz
- Memòria RAM: 32GB ECC 1600Mhz
- Discs: 4 (300GB) x SAS 15.000 rpms - 6Gb/s
- Xarxa: 6 ports ethernet 1.000 FD
- Gestió: 1 port IPMI

# Disseny

- Nodes de còmput.

Servidor de còmput per al processament del càlcul de les tasques del clúster HPC.

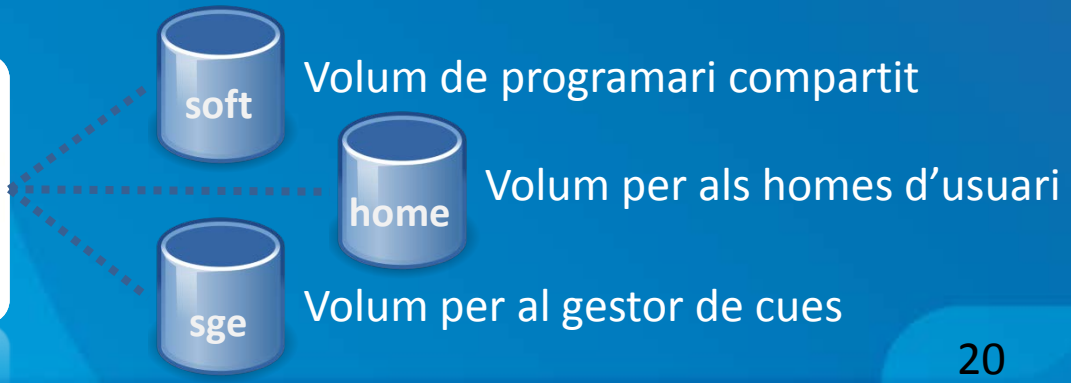
Principals elements:

- Processador: Opteron Abu Dhabi 6378 a 2,4Ghz
- Memòria RAM: 256GB ECC 1600Mhz
- Discs: 1 (240GB) x SSD 555/510 MB/s - 6Gb/s
- Xarxa: 6 ports ethernet 1.000 FD
- Gestió: 1 port IPMI

# Disseny

- Sistema d'emmagatzemament.

Sistema de fitxers NAS NetApp FAS3140 centralitzat, accessible mitjançant el protocol NFS per la xarxa Gigabit Ethernet des de tots els nodes de còmput i les màquines virtuals.



# Contingut

- Introducció
- Motivació
- Objectius i Planificació
- Anàlisi
- Disseny
- **Implementació**
- Conclusió
- Línies d'actuacions futures
- Demostració

# Implementació

- Per dur a terme la implementació s'ha hagut d'utilitzar les següents eines.

Elements comuns:

- Debian GNU/Linux 7.0
- Repositoris de programari Neurodebian

# Implementació

- Servidor de màquines virtuals.

S'empra Xen com a gestor de recursos virtuals perquè proporciona un aïllament segur, control de recursos, garanties de qualitat de servei i migració de màquines virtuals en viu.

```
$ xm list
Name                        ID   Mem VCPUs      State   Time(s)
Domain-0                     0 14046    24      r-----    7445.1
deploy                      11   512     1      -b----    2739.7
login                        2  8096     4      -b----    5081.0
monitor                      4  1024     1      -b----    6969.8
proxy                        6   512     1      -b----    3774.0
sgemaster                   10  4096     1      -b----    6414.9
```

# Implementació

- 1/3 Roles de les màquines virtuals:

**Login**, gestió del control d'accés dels usuaris del LDAP corporatiu. 

**Deploy**, instal·lació, configuració i gestió centralitzada dels servidors del clúster mitjançant FAI-software, ClusterSSH i IPMI Client.

# Implementació

- 2/3 Roles de les màquines virtuals:

**Monitor**, monitorització dels sistemes amb l'eina Ganglia (per verificar l'estat dels servidors) i PHPQstat per visualitzar les tasques del clúster per usuari.


**Proxy**, gestió d'accés de sortida a internet mitjançant SQUID i relay d'enviament de correu per EXIM4.

# Implementació

- 3/3 Roles de les màquines virtuals:

**Sgemaster**, Open Grid Scheduler per dur a terme la gestió, assignació de recursos i programació de les tasques dels usuaris.

Els usuaris poden executar tasques autònomes, paral·leles o sessions d'usuari interactives.

# Contingut

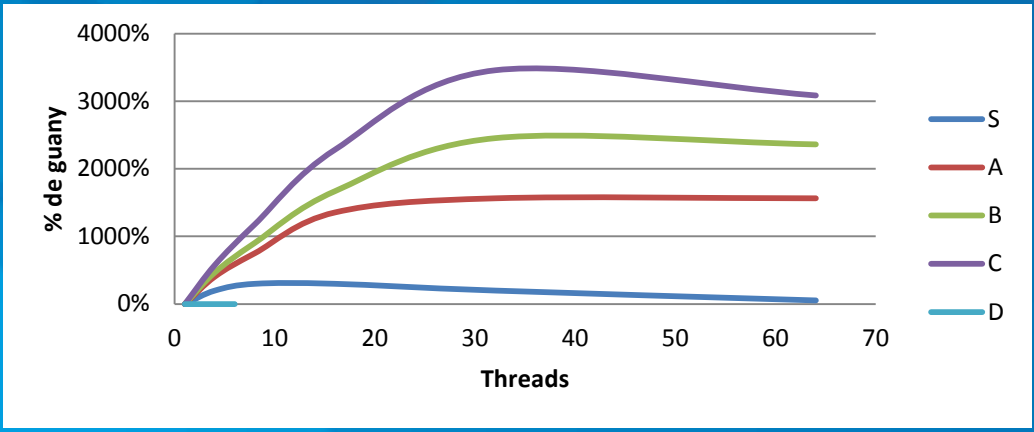
- Introducció
- Motivació
- Objectius i Planificació
- Anàlisi
- Disseny
- Implementació
- **Conclusió**
- Línies d'actuacions futures
- Demostració

# Conclusió

- La implementació del clúster proporciona:
  - Sistema de repartiment i gestió de la càrrega per als usuaris.
  - Sistema d'alt rendiment, escalable i flexible.
  - Sistemes homogenis d'alta disponibilitat.
  - Seguretat, amb la segmentació de les diferents xarxes i control d'accés d'usuaris.
  - Reducció de despeses al ser una infraestructura transversal per diversos departaments.
  - Adaptat al 100% als requeriments dels usuaris.

# Conclusió

- A la fase de proves és dedueix que depenent de la mida del problema, el rendiment és pot veure afectat. Un node de còmput estabilitza el creixement del rendiment a partir de l' utilització del 50% dels recursos de CPU.




*Gràfica de guany de l'execució del programa Conjugate Gradient de les Nasa Parallel tools*

# Contingut

- Introducció
- Motivació
- Objectius i Planificació
- Anàlisi
- Disseny
- Implementació
- Conclusió
- **Línies d'actuacions futures**
- Demostració

# Línies d'actuacions futures

- Instal·lació d'una xarxa de baixa latència infiniband per maximitzar l'ample de banda i accelerar l'intercanvi de dades.
- Instal·lació de xips de processament gràfic GPU, optimitzats per dur a terme càlcul de valors en coma flotant més ràpids i eficients.
- Distribució dels espais d'usuari mitjançant quotes d'ús.
- Migració de les màquines virtuals al sistema d'emmagatzemament en xarxa.

# Contingut

- Introducció
- Motivació
- Objectius i Planificació
- Anàlisi
- Disseny
- Implementació
- Conclusió
- Línies d'actuacions futures
- **Demostració**

# Demostració

- Vídeo demostració de l'ús del clúster per un usuari comú quan executa una tasca.

  http://youtu.be/yL7KVYqbf2Q


**Video publicat a www.youtube.com només accessible mitjançant URL.*