Configuração de cluster de computadores

Sidney Pio de Campos

sidney@feq.unicamp.br

09/02/2015

Objetivo

- Apresentar uma visão geral da instalação e configuração de cluster de computadores.

Tópicos abordados

- Conceitos básicos
- Componentes de um cluster
- Compartilhamento de arquivos
- Sistemas de Filas
- MPI
- Exemplo de montagem de um cluster simples

Conceitos Básicos de Cluster de computadores

- idéia geral: conjunto de 2 ou mais máquinas operando em conjunto oferecendo ao usuário a visão de um único sistema
- de forma geral, podemos dividir em:
  - cluster de alto desempenho (High Performance: HP)
  - cluster de alta disponibilidade (High Availability: HA)
  - cluster para ambientes virtualizados
- podem utilizar sistema operacional UNIX-Like ou até mesmo Windows
- nosso foco nessa apresentação será em cluster de alto desempenho e com sistema UNIX-Like

Componentes - hardware

- Uma máquina para atuar como *front-end* ou *headnode* do cluster
- nós de processamento
  - processadores tradicionais (Intel/Xeon, AMD/Opteron)
  - GPU (Tesla/CUDA da NVIDIA)
  - Xeon Phi da Intel
- *storage* que pode ser:
  - uma máquina dedicada com disco
  - um *storage* propriamente dito
  - um conjunto de máquinas e/ou *storage*
- *switches* para interligação dos nós

Interligação - switches

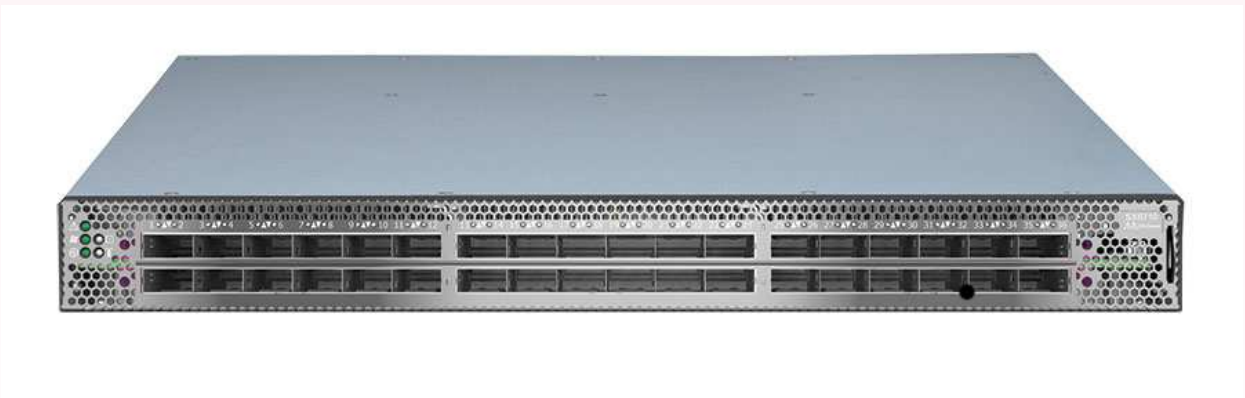
Podemos destacar 2 tecnologias:

- Ethernet
  - familia IEEE 802.3
  - mais tradicional e barata
  - taxas de 10/100 Mbps, 1/10 Gbps, 40 Gbps, 100 Gbps
- InfiniBand (IB)
  - taxas de 40, 56 and 100Gb/s
  - latência menor do Ethernet

	56Gb/s FDR IB	40Gb/s QDR IB	10GbE
Throughput	6.8 GB/s	3.2 GB/s	1.1 GB/s
Latency	0.7us	1.2us	7.22us
Message Rate (Million msg/sec)	137	30	1.1

Table: http://www.mellanox.com/page/performance_infiniband

Interligação - InfiniBand



InfiniBand - Para refletir

- Então devemos sempre usar tecnologia InfiniBand?

Componentes - softwares

- para instalação dos nós de processamento
- para compartilhamento de sistemas de arquivos
- para sistema de filas
- para desenvolvimento de programas em paralelo
- bibliotecas científicas
- para gerência/monitoramento do cluster

Softwares para instalação dos nós

- vários nós de processamento semelhantes
- um possível processo de instalação:
  - instalado um servidor de DHCP e TFTP
  - o nós são configurados para boot pela rede via Pxe, por exemplo
  - é transferida uma pequena imagem que faz a instalação em disco local ou que copia uma imagem previamente preparada
  - no próximo boot pela rede é enviada uma imagem que indica o boot pelo disco
- exemplo: *Rocks Cluster*
- podem ser utilizados outros softwares, por exemplo *Clonezilla*, *UDPcast*

Compartilhamento de sistemas de arquivos: NFS

- **N**etwork **F**ile **S**ystem
- desenvolvido em 1984 pela Sun e IBM
- muito utilizado em sistemas UNIX e UNIX-like
- maduro
- padrão (bem entendido)
- robusto e disponível em várias plataformas
- transparente para o usuário

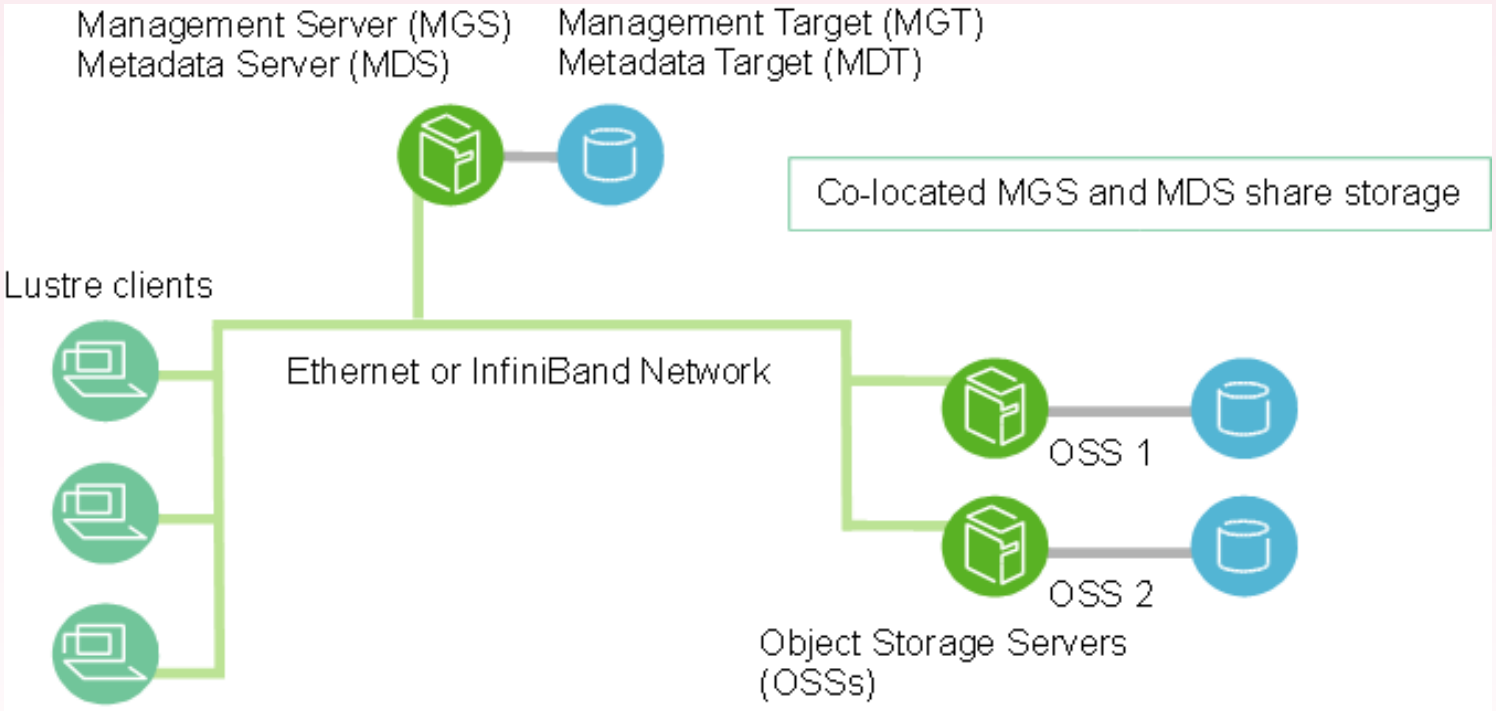
Compartilhamento de sistemas de arquivos: AFS

- **A**ndrew **F**ile **S**ystem
- sistema de arquivos distribuídos
- criado na *Carnegie Mellon University*
- desenvolvido e suportado pela *Transarc Corporation (IBM Pittsburgh Labs)*
- versão opensource: OpenAFS
- usar kerberos para autenticação
- menos "popular" do que o NFS

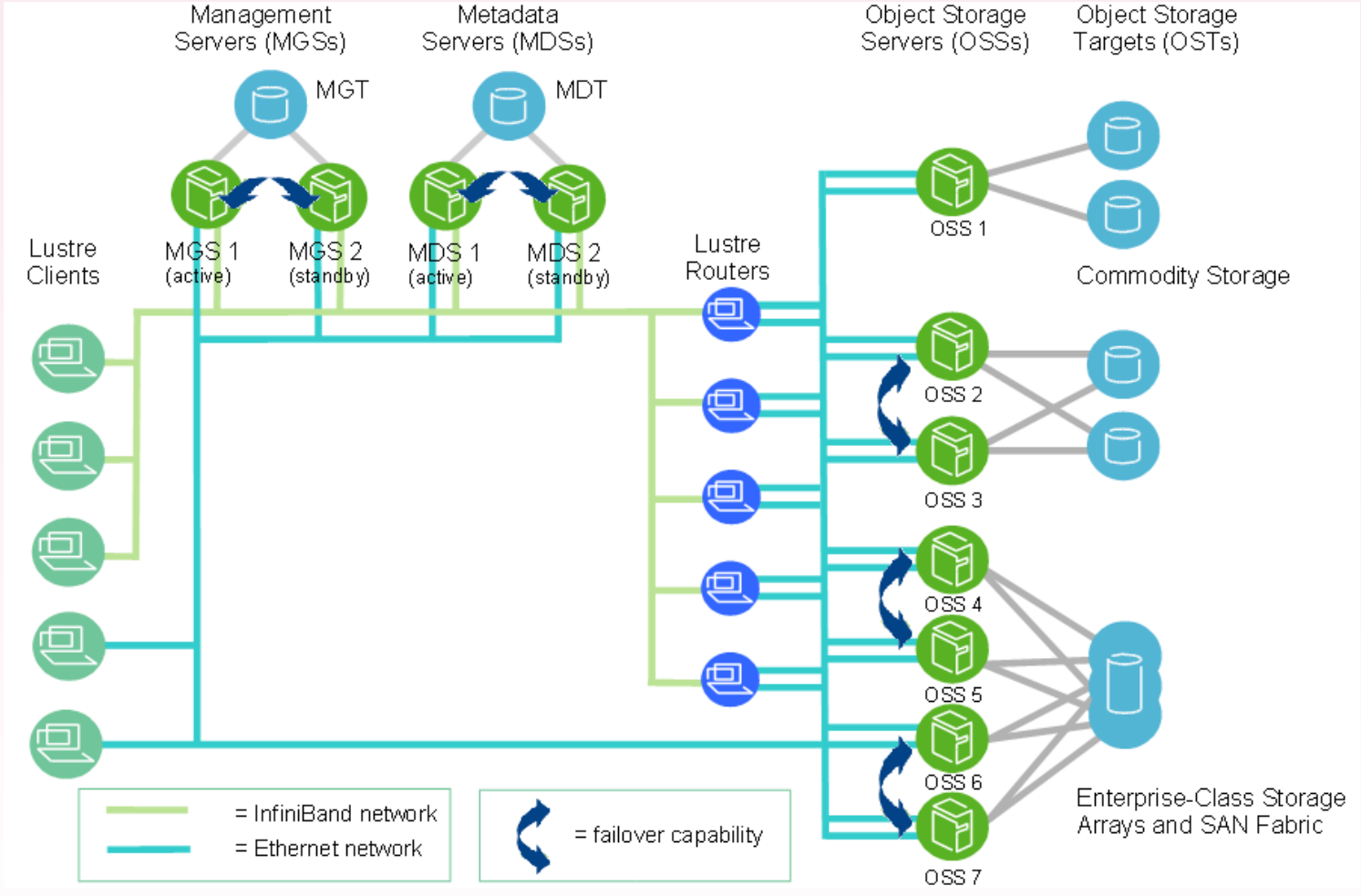
Compartilhamento de sistemas de arquivos: Lustre

- sistema de arquivos paralelo e distribuído
- *Open source*: GPL 2.0
- utilizado em grandes cluster
  - exemplo Cluster Titan: 40 PB com uma taxa de 1.4 TB/s
- "cluster de storages"

Compartilhamento de sistemas de arquivos: Lustre



Compartilhamento de sistemas de arquivos: Lustre



Compartilhamento de sistemas de arquivos: para refletir

- Qual sistema devo usar para compartilhar disco?

Sistema de Filas - Introdução

- usuário não deve acessar os nós diretamente
- permite:
  - acesso disciplinado aos nós do cluster
  - otimizar recursos
  - definir políticas de acesso
  - priorizar tarefas

Sistema de Filas - Implementações

- PBS (**P**ortable **B**atch **S**ystem) professional da Altair
- Tivoli Workload Scheduler LoadLeveler da IBM
- OpenPBS: implementação livre do PBS, atualmente sem desenvolvimento
- Slurm (Simple Linux Utility for Resource Management): última versão em 2010
- TORQUE da *Adaptive Computing*

Sistema de Filas - TORQUE

- **T**erascale **O**pen-source **R**esource and **QUE**ue **M**anager
- baseado no Open PBS
- *Open Source*
- última versão: 5.1.0 de 20/01/2015
- permite usar outros escalonadores (Moab, Maui)

Sistema de Filas - TORQUE

TORQUE Resource Manager provides control over batch jobs and distributed computing resources. It is an advanced open-source product based on the original PBS project and incorporates the best of both community and professional development. It incorporates significant advances in the areas of scalability, reliability, and functionality and is currently in use at tens of thousands of leading government, academic, and commercial sites throughout the world. TORQUE may be freely used, modified, and distributed under the constraints of the included license.*

Sistema de Filas - TORQUE

Alguns comandos úteis:

- pbsnodes
- qsub
- qdel
- qstat

Sistema de Filas - TORQUE

Exemplo de script para submissão pelo TORQUE

```
#!/bin/bash
#
# Exemplo 1
#
#PBS -N Exemplo1
#PBS -l nodes=1
#PBS -M sidney
#PBS -m abe

/bin/hostname
date
```

MPI - introdução

- **M**essage **P**assing **I**nterface
- padrão para comunicação de dados em computação paralela
- uma aplicação com MPI é constituída por um ou mais processos que se comunicam, acionando-se funções para o envio e recebimento de mensagens entre os processos

MPI - exemplo de código em C

```
/*
 * Copyright (c) 2004-2006 The Trustees of Indiana University and Indiana
 *                         University Research and Technology
 *                         Corporation.  All rights reserved.
 * Copyright (c) 2006      Cisco Systems, Inc.  All rights reserved.
 *
 * Sample MPI "hello world" application in C
 * modificado para apresentar o nome da maquina em que o codigo esta sendo executado
 */
#include <stdio.h>
#include "mpi.h"
int main(int argc, char* argv[])
{
    int rank, size, len;
    char version[MPI_MAX_LIBRARY_VERSION_STRING];
    char nome[128];
    MPI_Init(&argc, &argv);
    MPI_Comm_rank(MPI_COMM_WORLD, &rank);
    MPI_Comm_size(MPI_COMM_WORLD, &size);
    MPI_Get_library_version(version, &len);
    gethostname(nome, sizeof(nome));
    printf("Hello, world, I am %d of %d e estou rodando em %s, (%s, %d)\n",
           rank, size, nome, version, len);
    sleep(120);
    MPI_Finalize();
    return 0;
}
```

MPI - implementações

- MPICH: http://www.mpich.org/
- Open MPI: http://www.open-mpi.org/
- MPI da SGI
- MPI da Intel
- ...

Open MPI

- *open-source* com licença BSD
- desenvolvido e mantido por parceiros de áreas acadêmica, pesquisa e indústria
- *merge* das implementações
  - FT-MPI (University of Tennessee)
  - LA-MPI (Los Alamos National Laboratory)
  - LAM/MPI (Indiana University)
- última versão: 1.8.4 (19/12/2014)
- boa integração com o *TORQUE*

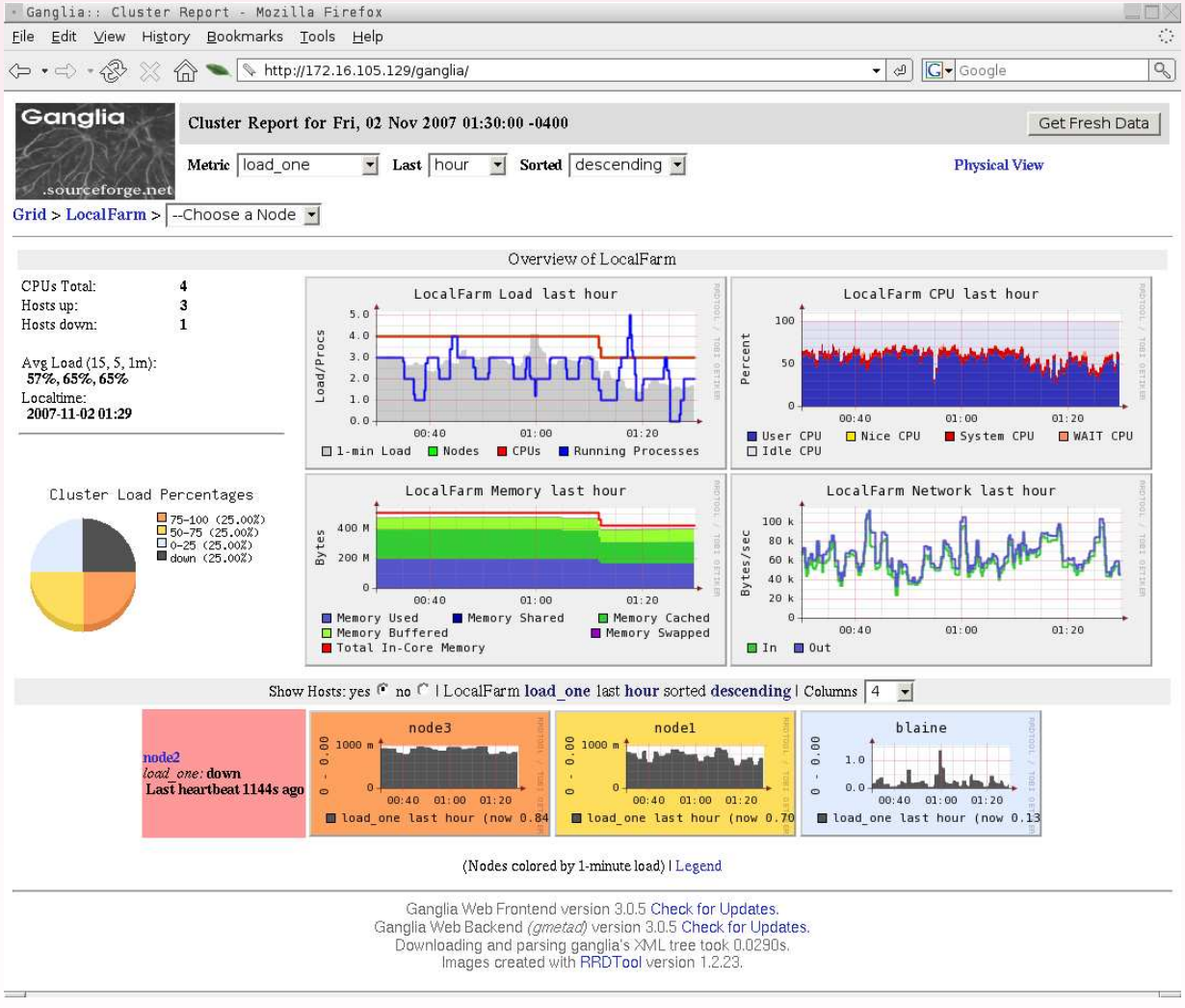
Bibliotecas científicas

- normalmente podem ser instaladas em máquinas individuais.
- em alguns casos possuem versões para atuar em conjunto com MPI ou processamento paralelo.
- podemos destacar:
  - MKL da Intel
  - CERN Program Library (CERNLIB)
  - BLAS (Basic Linear Algebra Subprograms)
  - LAPACK
  - NAG

Gerência/Monitoramento do cluster

- shell-scritps para automatizar tarefas
- ferramentas tipo *Puppet*, *Ansible* para automatizar instalação e atualização de softwares
- ferramentas do gerência, tipo Ganglia, Nagios, Zabbix

Gerência/Monitoramento do cluster - Ganglia



Descrição

- ambiente virtualizado: VirtualBox
- sistema operacional: Linux
- distribuição: Debian
- 3 máquinas:
  - *headnode*: máquina para usuário logar e enviar os jobs. Além disso atuará como storage
  - *node01 e node02*: máquinas para "processamento" com 4 "processadores virtuais"
- 1 única rede interligando as máquinas do cluster

Softwares

- compiladores: gcc, g++ e gfortran
- compartilhamento de disco: NFS
- MPI: OpenMPI
- sistema de filas: Torque

Instalação do servidor

- Debian sem interface gráfica, apenas com servidor *ssh*
- instalando o DHCP server no headnode:
  - apt-get install isc-dhcp-server
  - editado /etc/dhcp/dhcpd.conf
  - editado /etc/default/isc-dhcp-server
- criado um segundo disco no headnode (home)
  - criado no virtualbox
  - particao criada com fdisk
  - filesystem criado com mkfs.ext4
  - entrada no /etc/fstab
- compartilhando o disco
  - apt-get install nfs-kernel-server
  - ajuste no /etc/exports
  - restart do servico

Instalação do OpenMPI

- apt-get install build-essential
- instalação do OpenMPI a partir do código fonte:

```
cd /root/src
wget http://www.open-mpi.org/software/ompi/v1.8/downloads/openmpi-1.8.4.tar.gz
tar xf openmpi-1.8.4.tar.gz
cd openmpi-1.8.4/
./configure --prefix=/opt/openmpi-1.8.4
make install clean
cd /opt
tar cf /home/openmpi-1.8.4.tar openmpi-1.8.4
```

- ajuste do path em /etc/bash.bashrc

```
# System-wide .bashrc file for interactive bash(1) shells.
# To enable the settings / commands in this file for login shells as well,
# this file has to be sourced in /etc/profile.
export PATH="/opt/openmpi-1.8.4/bin:$PATH"
# If not running interactively, don't do anything
[ -z "$PS1" ] && return
...
```

Instalação do TORQUE

- instalação de pacotes necessários:

```
apt-get install libssl-dev
apt-get install libxml2-dev
apt-get install libboost-dev
```

- instalando o TORQUE

```
cd /root/src
tar xf torque-5.1.0-1_4048f77c.tar.gz
cd torque-5.1.0-1_4048f77c
./configure --prefix=/opt/torque-5.1.0
make
make install
echo "/opt/torque-5.1.0/lib" > /etc/ld.so.conf.d/torque.conf
cp contrib/init.d/debian.trqauthd /etc/init.d
/etc/init.d/debian.trqauthd start
./torque.setup root
```

Instalação do TORQUE - continuação

- ajustes para inicialização:

```
cd contrib/init.d
cp debian.pbs_sched /etc/init.d/pbs_sched
cp debian.pbs_server /etc/init.d/pbs_server
cp debian.trqauthd /etc/init.d/trqauthd
update-rc.d pbs_sched defaults
update-rc.d pbs_server defaults
update-rc.d trqauthd defaults
```

- preparando os pacotes para instalação nos clientes:

```
make packages
cp torque-package-clients-linux-x86_64.sh /home
cp torque-package-mom-linux-x86_64.sh /home
```

Instalação no cliente

- montando o /home do *headnode* através do arquivo /etc/fstab
- copiando o OpenMPI:

```
mkdir /opt
cd /opt
tar xf /home/openmpi-1.8.4.tar
```

- instalando o TORQUE:

```
cd /home/pacotes/
./torque-package-clients-linux-x86_64.sh --install
./torque-package-mom-linux-x86_64.sh --install
cp /home/debian.pbs_mom /etc/init.d/pbs_mom
/etc/init.d/pbs_mom start
update-rc.d pbs_mom defaults
echo "$usecp *:/home /home" > /var/spool/torque/mom_priv/config
```

Script para copia de arquivos importantes

- script para sincronizar alguns arquivos importantes:

```
for maquina in ‘cat /root/scripts/maquinas.txt‘
do
scp /etc/passwd  $maquina:/etc/passwd
scp /etc/shadow  $maquina:/etc/shadow
scp /etc/group  $maquina:/etc/group
scp /etc/hosts  $maquina:/etc/hosts
scp /etc/profile $maquina:/etc/profile
scp /etc/bash.bashrc $maquina:/etc/bash.bashrc
done
```

- pode ser executado manualmente ou agendado para executar periodicamente

Considerações finais

- devemos montar um cluster pensando em futuras ampliações
- devemos entender o tipo de programa que será processado
- scripts podem ajudar para automatizar algumas tarefas

Algumas referências

TOP500: http://www.openafs.org/
InfiniBAND (site da Mellanox): http://www.mellanox.com/
LUSTRE: http://lustre.opensfs.org/
AFS: http://www.openafs.org/
TORQUE: http://www.adaptivecomputing.com/products/open-source/torque/
Introdução MPI: http://condor.cc.ku.edu/~grobe/docs/intro-MPI-C.shtml
RocksClusters: http://www.rocksclusters.org/wordpress/
Ganglia: http://ganglia.sourceforge.net/

Muito Obrigado !!!

Dúvidas?

Contatos:

- sidneypio@gmail.com
- sidney@feq.unicamp.br