Computación Científica en Clusters

Administración de plataformas paralelas

Juan Piernas Cánovas

Febrero de 2010

Índice

Agenda

Estructura del curso

- Hay dos grandes bloques:
  - Bloque I: clases presenciales + pequeños trabajos individuales de cada alumno. Compuesto por dos partes:
    - Administración de plataformas paralelas.
    - Programación de plataformas paralelas.
  - Bloque II: trabajo por grupos de alumnos, tutorizado por un profesor del curso.
    - Computación científica paralela.
- Visión integral: desde el diseño del cluster hasta su uso y programación.

Estructura del curso

- Hay dos grandes bloques:
  - Bloque I: clases presenciales + pequeños trabajos individuales de cada alumno. Compuesto por dos partes:
    - Administración de plataformas paralelas.
    - Programación de plataformas paralelas.
  - Bloque II: trabajo por grupos de alumnos, tutorizado por un profesor del curso.
    - Computación científica paralela.
- Visión integral: desde el diseño del cluster hasta su uso y programación.

Administración de plataformas paralelas

- Introducción a los clusters de cómputo científico: arquitectura general, software, etc.
- Diseño de un cluster. Principales elementos hardware y software de los nodos de cómputo, de los nodos de almacenamiento de datos y de la red de interconexión: multiprocesamiento, sistemas RAID, redes Gigabit e Infiniband, etc.
- Instalación desatendida de nodos: DHCP, PXE, Kickstart (Fedora), etc.
- Gestión centralizada de usuarios: NIS y LDAP.
- Configuración de un servicio NFS.
- Instalación y configuración de un sistema de ficheros paralelo Lustre.
- Instalación y configuración de un sistema de colas: Torque, Maui y SunGridEngine.
- Instalación de MPI, de diversos compiladores de C y Fortran que soporten OpenMP y de bibliotecas optimizadas como Blas/Lapack (incluyendo las versiones paralelas de éstas).

Administración de plataformas paralelas

- Introducción a los clusters de cómputo científico: arquitectura general, software, etc.
- Diseño de un cluster. Principales elementos hardware y software de los nodos de cómputo, de los nodos de almacenamiento de datos y de la red de interconexión: multiprocesamiento, sistemas RAID, redes Gigabit e Infiniband, etc.
- Instalación desatendida de nodos: DHCP, PXE, Kickstart (Fedora), etc.
- Gestión centralizada de usuarios: NIS y LDAP.
- Configuración de un servicio NFS.
- Instalación y configuración de un sistema de ficheros paralelo Lustre.
- Instalación y configuración de un sistema de colas: Torque, Maui y SunGridEngine.
- Instalación de MPI, de diversos compiladores de C y Fortran que soporten OpenMP y de bibliotecas optimizadas como Blas/Lapack (incluyendo las versiones paralelas de éstas).

Agenda

Introducción

- En muchos campos de la ciencia es necesario realizar cálculos y/o simulaciones que requieren una gran potencia de cómputo:
  - Dinámica de partículas.
  - Modelado del clima.
  - Criptografía, etc.
- Una mayor potencia de cómputo permite:
  - Reducir el tiempo en el que se obtienen los resultados.
  - Mejorar la calidad de los resultados.
- Esta potencia es ofrecida hoy en día por los supercomputadores y, en concreto, por los grandes *clusters* de ordenadores (ver `http://www.top500.org`).

Cluster

- Conjunto de ordenadores conectados entre sí que colaboran para resolver un determinado problema.
- No todos los nodos desempeñan el mismo papel. Tres tipos principales:
  - **Nodo(s) de acceso o de control**: administración del cluster, acceso al cluster, programación, ejecución de programas, etc.
  - **Nodos de cómputo**: ejecución de código.
  - **Nodos de almacenamiento**: donde residen los ficheros que se procesan o almacenan resultados de la ejecución.

Cluster


Cluster
Nodo de acceso
Terminales de los usuarios
(C) Ainkaboot

Hardware de un cluster

- En muchos casos, el hardware (procesadores, memoria, discos duros, etc.) es muy similar al que podemos encontrar en cualquier ordenador de sobremesa.
- Formato especial para poder ser montados en *rack*.

Un nodo o servidor

Varios tipos de *rack*

- Redes de interconexión de alta velocidad (Gigabit, 10Gigabit, Infiniband, Myrinet, etc.)
- ¡Muy asequibles hoy en día!

Software de un cluster

- El sistema operativo es, casi siempre, Linux, en cualquiera de sus posibles distribuciones (frecuentemente, RedHat Enterprise Linux, CentOS, Fedora u OpenSUSE).
- Gran variedad de software científico y de programación disponible: MPI, OpenMP, BLAS/LAPACK, ScaLAPACK, R, Grass, Octave, Scilab, Maxima, CUDA/OpenCL, compiladores de C y Fortran (incluyendo los de Intel), etc, etc.

Agenda

Diseño de un cluster

- Varios factores a tener en cuenta:
  - Modelo de programación:
    - Paso de mensajes.
    - Memoria compartida.
  - Carga de trabajo:
    - Intensiva en CPU.
    - Intensiva en comunicaciones.
    - Intensiva en E/S.
  - Volumen de datos.
  - Necesidades específicas:
    - Desarrollo de programas en CUDA.
    - Uso de procesadores Cell BE.
    - Uso de FPGAs, etc.

Ejemplo de construcción de un cluster

- Vamos a construir un cluster de propósito general con los siguientes nodos:
  - 1 nodo de acceso al cluster.
  - 4 nodos de almacenamiento (para un sistema de ficheros Lustre de $4 \times 1,5 = 6$TB).
  - 16 nodos de cómputo (para un total de 32 procesadores o 128 *cores*).
- Todos los nodos tendrán una altura de 1U y se montarán en un armario *rack*.
- Es importante asegurarse de que todos los elementos hardware que deba gestionar el sistema operativo son compatibles con Linux.

Ejemplo de construcción de un cluster

- Precio: entre 40.000 y 60.000 euros.
  - Puede ser mucho más barato si renunciamos a determinadas características (potencia y número de procesadores, capacidad y fiabilidad de los discos duros, etc.).
  - No incluimos el coste del software (generalmente, libre).
  - No debemos olvidar el coste del refuerzo eléctrico y del sistema de refrigeración (otros 4.000–6.000 euros).
  - Si necesitamos un SAI para el cluster, habrá que sumar otros 6.000–8.000 euros.
  - Tampoco debemos olvidar el coste de un sistema de copias de seguridad.

Características del nodo de acceso

- 2 procesadores (Intel Xeon o AMD Opteron).
- 4–8 GB de RAM.
- Tarjeta RAID con batería.
- 4 discos duros SATA para servidores de 500GB de capacidad (que montaremos en RAID5 para tener una capacidad efectiva de 1,5TB).
- 2 puertos Gigabit Ethernet: uno para el exterior y otro para la red interior.

Características de los nodos de almacenamiento

- 1 o 2 procesadores (Intel Xeon o AMD Opteron).
- 4 GB de RAM.
- Tarjeta RAID con batería.
- 4 discos duros SATA para servidores de 500GB de capacidad (que montaremos en RAID5 para tener una capacidad efectiva de 1,5TB).
- 2 puertos Gigabit Ethernet: uno para la red de comunicaciones y otro para la red de gestión.
- Tarjeta Infiniband para la red de altas prestaciones.
- Tarjeta IPMI para la gestión remota.

IPMI


Home - Mozilla Firefox
Archivo Editar Ver Historial Marcadores Herramientas Ayuda
http://WW.XX.YY.ZZ
Most Visited Release Notes Fedora Project Red Hat Free Content Diccionario de la ...
Home
Home
Console
Logout
(ADMINISTRATOR)
Remote Control
Virtual Media
System Health
User Management
KVM Settings
Device Settings
Maintenance
Click to open
No Signal
Desktop size: no signal
Refresh
Power Control via IPMI
Power On
Power Down
Reset
Terminado

Características de los nodos de cómputo

- 2 procesadores (Intel Xeon o AMD Opteron).
- 4–8 GB de RAM.
- 1 disco duro SATA para servidores de 250GB de capacidad.
- 2 puertos Gigabit Ethernet: uno para la red de comunicaciones y otro para la red de gestión.
- Tarjeta Infiniband para la red de altas prestaciones.
- Tarjeta IPMI para la gestión remota.
- Interesante configuración *twin*:

Características de las redes de interconexión

- 1 switch Gigabit Ethernet de 48 puertos.
- 1 switch Infiband de 24 puertos.

Consideraciones finales

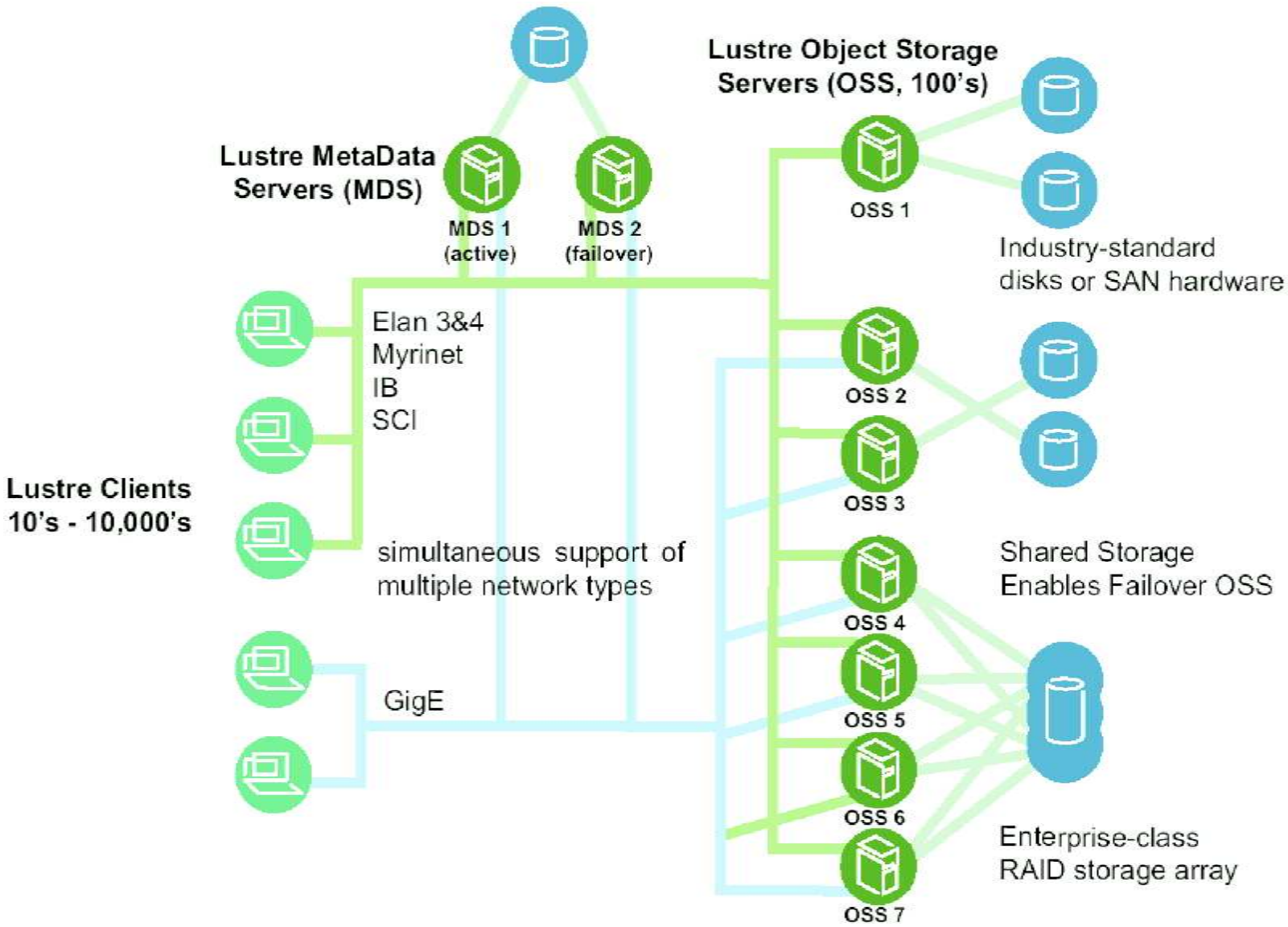
- La instalación del sistema operativo Linux en el nodo de acceso se hará de la manera tradicional.
- Es importante que el nodo de acceso siempre esté actualizado y que ofrezca al exterior el menor número de servicios posible (generalmente, el acceso por `ssh` es más que suficiente).
- Dentro del cluster, la seguridad puede ser más laxa, siempre que un usuario no pueda interferir con el trabajo de otro $\longrightarrow$ Será necesario establecer un sistema de colas.

Agenda

El sistema de ficheros paralelo Lustre

- Sistema de ficheros libre y gratuito (GPLv2) inicialmente desarrollado por ClusterFS, que fue adquirida por Sun Microsystems que, a su vez, ha sido recientemente adquirida por Oracle.
- Sistema de ficheros POSIX (o casi):
  - Comportamiento similar al de cualquier sistema de ficheros local.
  - Sustituto perfecto de NFS en entornos Linux distribuidos.
- Se basa en los *Object-based Storage Devices* (OSDs): los discos remotos no almacenan sectores sino «objetos».
- Usado en muchos de los más potentes superordenadores del mundo (en junio de 2009, se usaba en 7 de los 10 más potentes).

Arquitectura de Lustre



Características de Lustre

- Los clientes se comunican directamente con los OSSs y el rendimiento escala linealmente con el número de OSSs.
- Dos opciones para almacenar datos en los OSSs:
  - Cada fichero en un único OSS.
  - Cada fichero repartido entre varios OSSs como en un RAID 0.
    - Se pueden obtener decenas o cientos de GB/seg por cliente.
- Escalable a cientos de miles de clientes, miles de OSSs, sirviendo TB o PB de espacio en disco.
- Soporte para diferentes tipos de redes (GbE, Infiniband, etc.)