

Computación Cluster y Grid

Introducción



Motivaciones

Aplicaciones que requieren:

- Grandes capacidades de cómputo: Física de partículas, aerodinámica, genómica, ...
 - Tradicionalmente alcanzadas por medio de supercomputadores.
 - Los avances tecnológicos no satisfacen.
 - Camino hacia el "Petaflop".
- Necesidades de alta disponibilidad: Sistemas transaccionales de producción, banca, facturación, ...
 - Requieren replicación (y control de la misma).
 - No mucho cómputo, pero SIEMPRE debe estar disponible.
 - Implicaciones hardware y software.

Sistemas Distribuidos y Clusters

| Característica | MPP | SMP/CC-NUMA | Cluster | Sistemas Distribuidos |
|---------------------------|--|----------------------------------|---|--------------------------------|
| Número de nodos | $O(100) - O(1000)$ | $O(10) - O(100)$ | $O(100)$ o menos | $O(10) - O(1000)$ |
| Complejidad de los Nodos | Grano medio/fino | Grano medio/grueso | Grano medio | Diversos tipos |
| Comunicación Internodos | Paso de mensajes/DSM | Memoria compartida o DSM | Paso de mensajes | De ficheros compartidos a IPCs |
| Planificación de Trabajos | Cola de procesos única (en host) | Cola de procesos única | Colas múltiples coordinadas | Colas independientes |
| Soporte SSI | Parcialmente | Siempre | Desable | No |
| Tipo y Copias de SO | $N \times$ (μ kernels, por capas monolíticas) | Monolítico: SMPs Varios: NUMA | $N \times$ (homogéneas o μ kernels) | $N \times$ (SO homogéneos) |
| Espacio de Direcciones | Múltiple o único (para DSM) | Único | Múltiple o único | Múltiple |
| Seguridad Internodos | Innecesaria | Innecesaria | Sólo si expuesto | Requerido |
| Propietario | Una organización | Una organización | 1-N organizaciones | N organizaciones |

Computación con Clusters

- Alternativa los supercomputadores .
- En lugar de aproximaciones MPP:
 - Hardware específico.
 - Alto coste.
 - Desarrollo hardware lento.
 - Desarrollo software doloroso.
- Se pueden usar equipos de propósito general (PCs):
 - "Commodity hardware" (Commercial-off-the-self: COTS).
 - Coste reducido (y bajando).
 - Desarrollo hardware rápido.
 - Desarrollo software aun más doloroso.

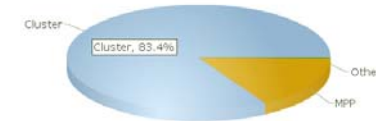
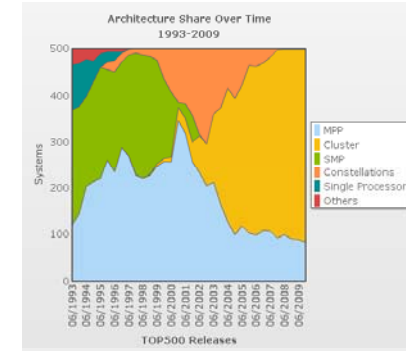
¿Qué es un Cluster?

Las características principales son:

- El concepto de cluster se basa en una instalación hardware específica. (Otros conceptos que veremos son conceptos software, un cluster no):
 - Nodos de computación.
 - Red de conexión dedicada.
- La explotación de una instalación cluster hace uso de herramientas específicas, que afecta a:
 - Programación de aplicaciones
 - Interacción y administración

Arquitecturas en el Top 500

- Supremacía de los sistemas cluster sobre otras arquitecturas



Clusters más Representativos I

- Magerit (Top 34/335)
 - 1204 nodos (1036 nodos eServer BladeCenter JS20 y 168 nodos eServer BladeCenter JS21)
 - JS20: 2 cores en dos procesadores IBM PowerPC single-core 970FX de 64 bits a 2'2 GHz, 4 GB
 - JS21: 4 cores en dos procesadores IBM PowerPC dual-core 970MP de 64 bits a 2'3 GHz, 8 GB
 - 15955 GFlops (LINPACK)
 - 5488 GB RAM
 - 65 TB Disco (GPFS)
 - Red Myrinet x 6 Switches
 - GigaEthernet x 2 Switches



Clusters más Representativos II

- Marenostrum (Top 5/77)
 - 4812 nodos
 - JS21: 4 cores en dos procesadores IBM PowerPC dual-core 970MP de 64 bits a 2'3 GHz, 8 GB
 - 63830 GFlops (LINPACK)
 - 20480 GB RAM
 - 280 TB Disco (GPFS)
 - Red Myrinet x 12 Switches
 - GigaEthernet x 5 Switches



Clusters más Representativos III

- Roadrunner (Top 1/2)
 - 12240 + 6562 procesadores
 - QS22: IBM PowerXCell 8i
 - AMD Opteron
 - 1.026 PFlops (LINPACK)
 - 103.6 TB RAM
 - Triblade / Infiniband



Clusters más Representativos IV

- Jaguar (Top 1)
 - Cray XT5-HE
 - 224,256 AMD Opteron Six Core 2.6 GHz
 - 1.75 PFlops (LINPACK)
 - 10 PB Disco (Spider/Lustre)
 - 598TB RAM
 - Cray SeaStar Network



Limitaciones de los Clusters

- Sobrecarga por comunicación:
 - Implementaciones de grano fino: difíciles de diseñar, difíciles de afinar y mucho más difíciles de escalar.
 - Las implementaciones de grano grueso son más fáciles. Pero, en este caso ¿Se necesitan anchos de banda espectaculares?
- Mantenimiento:
 - La ampliación de un cluster es costosa en grandes tamaños.
 - Es una solución escalable ... pero ¿hasta que punto pueden hacerlo?
- Recursos infrautilizados:
 - Los clusters son instalaciones dedicadas.
 - Un organización típica tiene desperdicia millones de ciclos de cómputo en sus ordenadores personales.
- Siempre existe un problema **mas grande**.

Intranet Computing

- Si yo tengo un montón de instalaciones de cómputo (incluidos los clusters, pero o restringiéndome sólo a ellos), ¿qué podría hacer?
 - Programar el algoritmo.
 - Dividir el espacio de soluciones o los datos de entrada (o ambos).
 - Distribuir el trabajo.

Ejecutar varios trabajos secuenciales o paralelos por medio de una herramienta de gestión de recursos

- Intranet computing:
 1. Opera dentro de la organización.
 2. Opera sobre hardware diferente (arquitectura y prestaciones).
 3. Trata un problema concreto. (los clusters u otros sistemas distribuidos son soluciones más generales)

Intranet Computing

- Aumenta la utilización de los recursos informáticos.
- El coste efectivo por ciclo de CPU usado es mínimo.
- Mejora en aspectos de escalabilidad.
- Mejora en disponibilidad.
- Simplifica la administración y el mantenimiento.
- Ejemplo:
 - Sun Grid Engine (Sun Microsystems),
 - Condor (University of Wisconsin),
 - LSF (Platform Computing)

Problemas con los Clusters de Gran Tamaño

- No es posible gestionar recursos fuera del dominio de administración:
 - Algunas herramientas (Condor, LSF) permiten la colaboración entre diferentes departamentos asumiendo la misma estructura administrativa.
- No se cumple la política de seguridad o los procedimientos de gestión de recursos.
- Los protocolos y los interfaces, en algunos casos, no se basan en estándares abiertos..
- Recursos a manejar: CPU, compartición de datos?

Más allá de los Clusters

- Computación Grid:
 - Agregación de clusters y de máquinas "ociosas".
 - Sistemas de planificación y ejecución de trabajos y de checkpointing.
 - Más de 1000 nodos.
 - e.g: Condor or Maui
- Computación colaborativa (Metacomputación):
 - Similar a la computación Grid pero sobre redes extensas de ámbito mundial (Internet).
 - Compartición de carga entre nodos que colaboran.
 - E.g: Seti@Home, Folding@Home, DESKeys

Elemento Clave: Acceso a Recursos

- **Fácil:** Uso intuitivo (similar a Web).
- **Transparente:** No resulta necesario conocer la ubicación física
- **Rápido:** Tiempo de respuesta aceptable
- **Seguro:** Control de acceso a recursos e información
- **Permanente:** Siempre disponible (24x7)
- **Económico:**
 - Menor coste al compartir infraestructuras
 - El coste debe ser conocido

Organizaciones Virtuales

- Una organización virtual (*virtual organization:VO*) está compuesta por recursos, servicios y personas que colaboran más allá de las fronteras institucionales, geográficas y políticas.
- Permiten el acceso directo a recurso de computación, software y datos y, por lo general, utilizan el substrato de la tecnología Grid.
- Proporcionan
 - Un portal Grid para agrupar todos los elementos.
 - Servicios de directorio
 - Infraestructura de seguridad

Redes Internacionales de Sistemas Grid

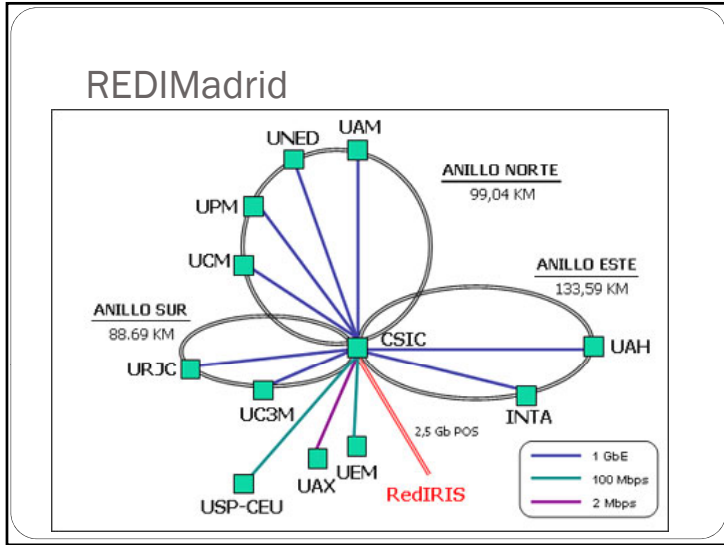
- Dichas infraestructuras disponen de conexiones de red extensa de gran ancho de banda.
- Se basan en infraestructuras de red nacionales o internacionales de propósito general:
 - RedIRIS, REDImadrid (España)
 - GÉANT (Europa)
 - TERAGrid backbone (USA)
 - ALICE y CLARA (Lationamérica)
- Colectivos internacionales asociados:
 - The Internet Engineering Task Force
 - <http://www.ietf.org/>
 - Dante
 - <http://www.dante.net/>

RedIris

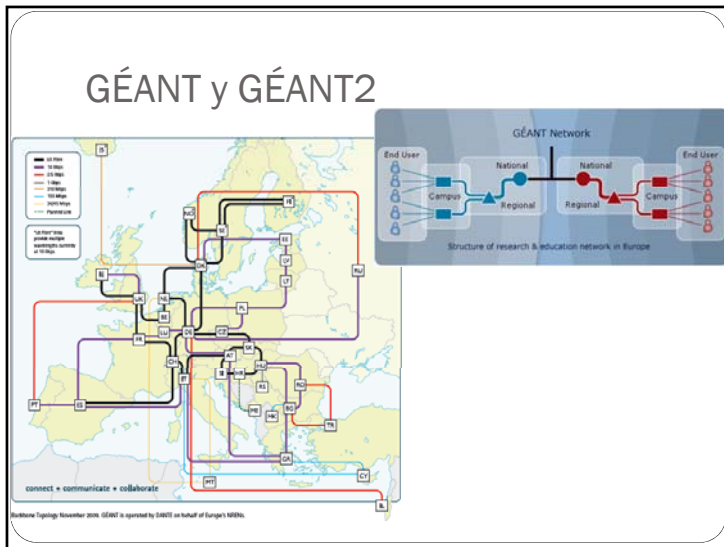
- España 1988: Red académica y de investigación española
- Infraestructura:
 - Núcleo de la red a 2,5 Gbps con, al menos, dos conexiones por nodo autónomo.
- Red actual:
 - 18 puntos de presencia, 10 enlaces de 2,5 Gbps, 13 a 622 Mbps y 6 a 155 Mbps.
- Presupuesto 2003:
 - 16,8 M€ de gastos de operación y 2,1 M€ para inversiones.
- RedIRIS conecta:
 - Más de 260 instituciones: Universidades y Centros de I+D
 - Conexiones externas a otras redes de investigación a la red comercial.

RedIris





- ### GÉANT
- GÉANT: Red europea multigigabit para la comunicación de datos de investigación y educación.
 - Conecta más de 3500 centros de investigación y educación en 33 países a lo largo de 29 redes regionales
 - GÉANT proporciona apoyo a los investigadores para:
 - Transmisión de datos a 10Gbps
 - Hacer uso de aplicaciones avanzadas de red (e.g., Grid Computing).
 - Colaboración, en tiempo real, sobre recursos de investigación.
 - Técnicas de computación avanzada, imposibles anteriormente



Conexiones Internacionales

| Project | Global Region |
|--------------------|-----------------------------|
| ALICE2 | Latin America |
| EUMEDCONNECT2 | Mediterranean |
| ORIENT | China |
| TEIN3 | Asia-Pacific |
| CAREN | Central Asia |
| Internet2 | North America |
| Eset | North America |
| UbuntuNet Alliance | Southern and Eastern Africa |
| CANARIE | North America |
| NISN | North America |
| NLR | North America |
| USLHCNet | North America |


TeraGrid




Fundada por la NSF.
Coordinada por medio de GIG
(Univ. Chicago)
Conecta los principales centros de
supercomputación

ALICE y CLARA

Evoluciones de la red CEASAR
Acuerdos con CIEMAT y IIPM
Conexiones USA: Miami y Tijuana



RedCLARA Topology June 2008

- ES Pop GÉANT - Madrid (ES)
- US Pop CLARA Pop's
- Connections established
- Connections planned
- US Pop Pacific Wave - Los Angeles (US)
- US Pop Atlantic Wave - Miami (US)
- ES Pop GÉANT - Madrid (ES)
- US Pop CLARA Pop's
- Connections established
- Connections planned
- US Pop Pacific Wave - Los Angeles (US)
- US Pop Atlantic Wave - Miami (US)
- ES Pop GÉANT - Madrid (ES)
- US Pop CLARA Pop's
- Connections established
- Connections planned
- US Pop Pacific Wave - Los Angeles (US)
- US Pop Atlantic Wave - Miami (US)