Computación Cluster y Grid

Cluster Computing

Concepto de Cluster

Cluster: Sistema de computación basado en hardware estándar conectado por una red dedicada (de altas prestaciones) dedicado a un propósito específico. [Baker00]

- Nodos de computación: PCs o estaciones de trabajo (SMPs).
- Red de conexión: Desde redes de alta velocidad hasta hardware específico.

Siglas misteriosas:

- PoPCs: Pile of PCs
- COWs: Clusters of workstations
- CLUMPS: Clusters of multiprocessors
- NOWs: Networks of workstations
- ....

Características Hardware

- Nodos del cluster:
  - Procesador: Intel Pentium, AMD Athlon, Compaq Alpha, IBM PowerPC, Sun SuperSparc (1-2...Ghz)
  - Memoria: SDRAM, DDR o similar (1-4 GB)
  - Almacenamiento: SCSI o RAID
- Red del cluster:
  - Éste es un elemento clave.
  - Puede implicar >50% del coste de la instalación
  - La solución barata: Ethernet (10-100-1000Mb/seg)

Redes de un Cluster (I)

- Tecnologías de red actuales:
  - Aumentan radicalmente el ancho de banda.
  - Se reducen muy ligeramente la latencia ← No son apropiadas

- Redes de latencia baja [Ap00]:
  - Active Messages (Berkeley): Modelo síncrono "*zero-copy*". GAM.
  - Fast Messages (Illinois): AM fiable y en orden.
  - VMMC (Princeton): Páginas de memoria virtual compartidas (DSM).
  - U-net (Cornell): Interfaces virtuales asociados a páginas.
  - BIP (Lyon): Interfas básico de baja latencia.

Redes de un Cluster (II)

- Estándares de Comunicación en un Cluster:
  - VIA: Interfaz hardware (nativo/emulado) de comunicación. Mapea regiones de memoria física a interfaces virtuales de red. Versiones de MPI sobre VIA.
  - InfiniBand: Estándar de hardware de E/S (2.5Gbps) sobree enlaces unidireccionales. 6 Modelos de comunicación. Soporta RDMA e IPv6.
- Hardware de red:
  - Ethernet, FastEthernet, GigaEthernet: Barato pero limitado. Problema de colisiones. Emulaciones de VIA.
  - Giganet (cLAN): Implementación de VIA (1.26Gbps)
  - Myrinet: Redes reprogramables de baja latencia. Encaminamiento cut-through y detección de caidas. Protocolo GM.
  - Otros: QsNet, ServerNet, SCI, ATM, FiberChannel, HIPPI, ATOLL,...

Comparativa de Tecnologías

	Gigabit Ethernet	Giganet	Myrinet	QsNet	SCI	ServerNet2
Ancho de banda sostenido con MPI (MB/seg)	35-50	105	140	208	80	65
Latencia MPI (μseg)	100-200	20-40	~18	5	6	20.2
Máximo número de nodos	1000's	1000's	1000's	1000's	1000's	64k
Soporte VIA	Win/Linux	Win/Linux	Sobre GM	Ninguno	Software	Hardware
Tipo de soporte MPI	MPICH sobre MVIA o TCP	Terceras partes	Terceras partes	Quadrics o Compaq	Terceras partes	Compaq o terceras partes

Software de Desarrollo (I)

- Sistemas Operativos:
  - Linux:
    - Libre, barato, rápido y fácil desarrollo.
    - e.g: Beowulf
  - Solaris:
    - Buen soporte de paralelismo y servicios de red.
    - e.g: Solaris MC
  - AIX:
    - Herramientas de desarrollo potentes y muy optimizadas.
    - e.g: SP2
  - Win2k/NT:
    - ¿por qué no?
    - e.g: Wolfpack

Software de Desarrollo (II)

- Middleware y SSI:
  - **SSI (Single System Image)**: Se intenta dar la visión de cara al usuario de un sistema único. Todo el cluster se muestra como un monoprocesador virtual.
  - Desrrollo por capas:
    - Hardware (Local).
    - Sistema operativo (μkernel) o nivel de *gluing*: GLUnix o MOSIX
    - Aplicaciones, servicios y middleware: CODINE
  - Servicios comunes (deseable):

•Punto único de acceso.
•Jerarquía de archivos única.
•Punto de gestión y control único.
•Red virtual única.
•Gestión de trabajos única.
•Interfaz de usuario único.
•Espacio de E/S único
•Espacio de procesos único.
•Checkpointing.
•Migración de procesos.

Software de Desarrollo (III)

- Herramientas de programación y desarrollo:
  - Soporte de threads: Pthreads o OpenMP
  - Paso de mensajes para clusters:
    - MPI: MPICH o LANMPI.
    - PVM: Peor rendimiento con más funcionalidades.
  - DSM: Distributed shared memory:
    - Software: TreadMarks, Linda o Nanos
    - Hardware: DASH o Merlin
  - Parallel debuggers o herramientas de instrumentación.

Software de Desarrollo (IV)

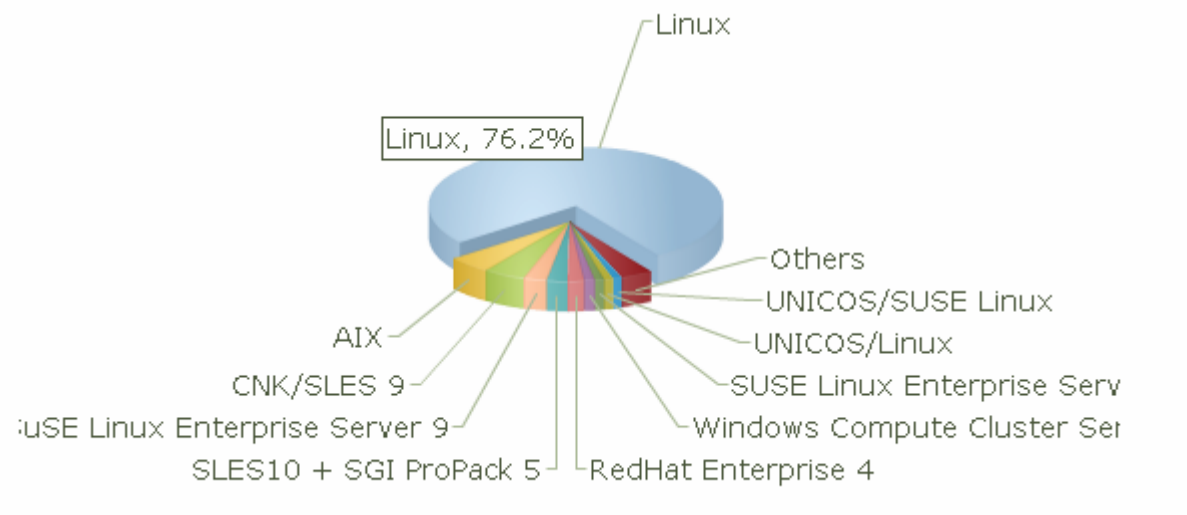
- Herramientas de Administración:
  - Gestión remota:
    - Mandatos de gestión: instalción de software, copia de archivos.
    - Acceso a recursos como los procesos.
    - Usuarios y otra información: NIS.
    - e.g: Herramientas SP2, Cluster Command & Control (C^3)
  - Sistemas de planificación:
    - Gestión de colas de trabajo y reparto del uso del cluster.
    - Determina los recursos de cada tarea.
    - e.g: CODINE, CONDORPBS (Portable Batch System)

Sistemas de Entrada/Salida

- Crisis de la E/S:
  - CPUs incrementan exponencialmente (Ley de Moore).
  - Los sistemas de E/S mucho más deprisa.
  - La E/S de procesos de alto rendimiento es el “cuello de botella”.
- Solución paralelismo:
  - Sistemas de E/S paralela: MPI I/O
  - Sistemas de ficheros paralelos: ParFiSys, GPFS
  - Sistemas de E/S inteligente: Armada, Panda

Crecimiento de los Grandes Clusters

- Top 500: Junio → Noviembre 2007
  - 5 de los 10 Top son nuevos equipos.
  - El sistema en la posición 500 estaría el 255 hace 6 meses.
  - Procesadores: 70%Intel, 15%AMD, 12%IBM Power
  - Fabricante: 46% IBM, 33% HP



Caso de Estudio: MOSIX

- Sistema operativo distribuido
- Consiste en un parche de Linux para migrar procesos bajo condiciones de carga y varias herramientas de usuario también “parcheadas”.
- Originalmente bajo GPL, ahora OpenMOSIX.
- Restricciones de migración, pero por lo demas es Linux!!!

[http://www.openmosix.org]

Caso de Estudio: Beowulf

- Herramientas de desarrollo sobre un cluster Linux
- Originalmente desarrollado para el sistema Avalon
- Se basa en MPI y unas cuantas decisiones “inteligentes” sobre el uso de los dispositivos de red.
- Grendel: Proyecto asociado para el desarrollo de aplicaciones sobre Beowulf.

[http://www.beowulf.org]

Caso de Estudio: SP/2 IBM

- Hardware/software para procesamiento masivo.
- Cluster de PowerPCs / Power2/3/4
- Herramientas de desarrollo (compiladores), servicios (GPFS) y comunicación (MPI) muy potentes.
- Bueno, es IBM, pero se lleva bien con Linux!

[http://www.sp.ibm.com]

Caso de Estudio: HALP / LVS

- Sistema de alta disponibilidad
- HALP: High Availability Linux Project
- LVS: Linux Virtual Server
- Equilibrado de carga, redirección de peticiones para mostrar una granja de servidores como un único sistema (de cara a un servicio determinado).
- Esto también es Linux!!!

[http://www.lvs.org]