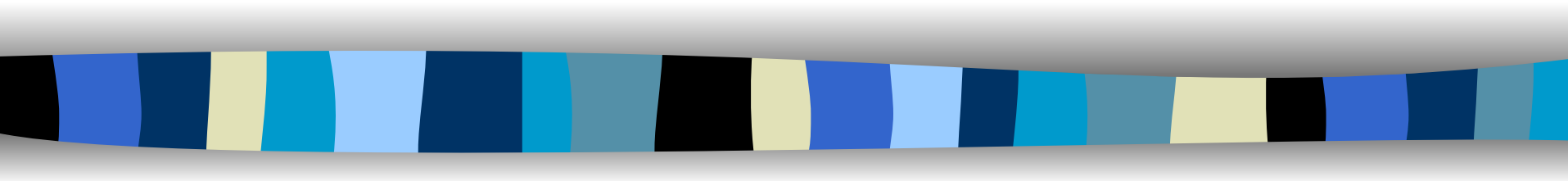
Linux Cluster Computing



Guillermo López Taboada

Grupo de Arquitectura de Computadores
Departamento de Electrónica y Sistemas

Introducción a las Arquitecturas Clúster

- Un clúster es ...
  - Un conjunto de nodos unidos mediante una red de interconexión a los que un determinado software convierte en un sistema de mayores prestaciones
- Aunque esta definición peque de general o imprecisa explica en esencia lo que es un clúster

Introducción a las Arquitecturas Clúster

- "Un clúster es un conjunto de nodos…"
  - ¿Cómo es un clúster en realidad?
    - ... son "casi" chatarra (P-{II,III} || K6-{2,3})?
    - ... son "buena" quincallería (PIV || K7)?
    - ... ha de ser diseñado especialmente?
    - ... es mezcla de lo anterior o nada de lo anterior?
  - ¿Cual es su apariencia?
    - ... hay una caja por cada nodo?
    - ... hay una gran caja en la que coge todo?
    - ... tendría incluso que estar todo dentro de un chip?
  - ¿ {Homo,Hetero}géneos?

Introducción a las Arquitecturas Clúster

- "… unidos mediante una red de interconexión …"
  - {Ø,Fast,Gigabit,10 Gigabit}-Ethernet
  - De baja latencia:
    - Myrinet
    - SCI
    - Qsnet
    - Quadrics
    - Infiniband

Introducción a las Arquitecturas Clúster

- "…a los que un determinado software convierte en un sistema de mayores prestaciones."

  (mayores rendimientos o disponibilidades)

  – Este **"software de sistemas clúster"** puede abarcar desde llamadas al sistema a aplicaciones de balanceo de carga

  – Está muy relacionado con un SO que provee servicios de clustering (Win ☻ ,Unix, Linux,MacOS)

Introducción a las Arquitecturas Clúster

- Terminología
  - Las arquitecturas clúster son una familia numerosa y heterogénea dentro de los MIMD de memoria físicamente distribuida:
    - Beowulf (Clusters COTS con Linux)
    - COW (Cluster of Workstations)
    - NOW (Network of Workstations)
    - POP (Pile of PCs)
    - PCF (PC Farms)
    - RF (Render Farms)
    - Constellations?

Introducción a las Arquitecturas Clúster

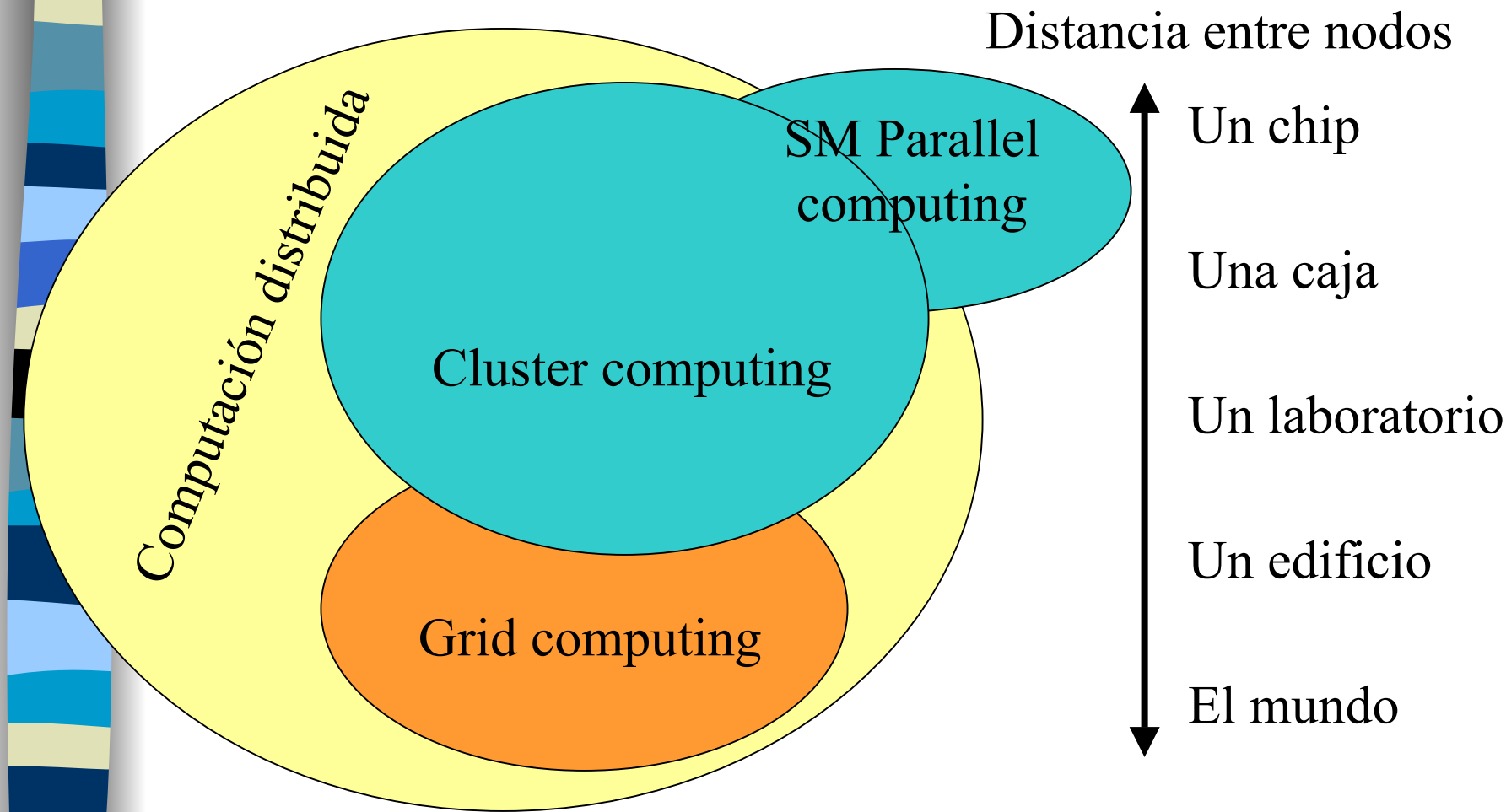
- Aunque el primer clúster (10 años ha) fue desarrollado “reciclando componentes” en la NASA (16 486DX4), el decurso del tiempo ha llevado a utilizar diverso hw de propósito general de 1ª mano (PIV || Athlon)
- Sin descartar nodos:
  - Xeon SMT
  - Xeon-MP
  - Itanium
  - Opteron
  - PS2

  … con procesadores de propósito general (no específico)

Introducción a las Arquitecturas Clúster

- Aunque con estos nodos podríamos obtener tanto clusters como máquinas para computación distribuida como grids
- La diferencia está en el software y en la distancia entre nodos (que limita a su vez las comunicaciones)

Introducción a las Arquitecturas Clúster



Introducción a las Arquitecturas Clúster

- Debido a la proximidad entre nodos de un clúster no resulta muy compleja su identificación visual (e.g.)

Introducción a las Arquitecturas Clúster

- Las redes de interconexión son un componente fundamental de los clusters que nos deben proveer de:
  - Alto ancho de banda
  - Baja latencia
  - Fiabilidad
  - Scalabilidad

Introducción a las Arquitecturas Clúster

- Redes de interconexión comunes en clusters son:
  - {Ø,Fast,Gigabit,10 Gigabit}-Ethernet
  - Infiniband
  - SCI
  - Myrinet
  - HIPPI
  - ATM
  - Fiber Channel
  - AmpNet
  - Qsnet
  - Quadrics
  - …

Introducción a las Arquitecturas Clúster

- Comparativa básica entre redes de interconexión:

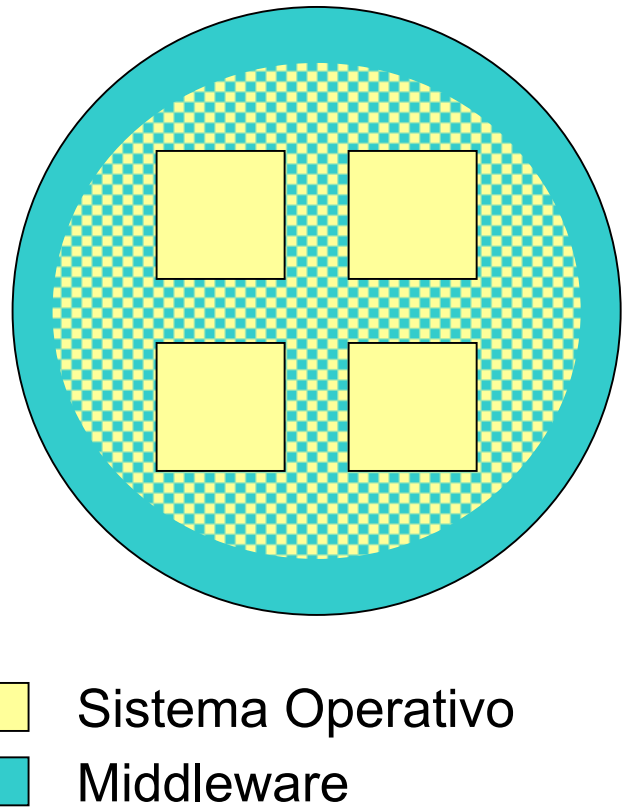
	Ancho de Banda	**Latencia**	**tarjeta**	**switch**
Fast-Ethernet	100Mbits/s	50us	10 €	10 €
Gb-Ethernet	1Gbit/s	70us	20 €	30 €
10Gb-Ethernet	10Gbit/s	100us		
SCI	1.33Gbit/s (full duplex)	2us	1.000 €	0 €
Myrinet	2Gbit/s (full duplex)	7us	800 €	800 €
IB	10Gbit/s / canal	10us	1.000 €	500 €

Introducción a las Arquitecturas Clúster

- Protocolos de comunicación
  - Tradicionales
    - TCP / UDP
  - Diseño específico
    - Active Messages
    - VMMC
    - BIP
    - VIA

Introducción a las Arquitecturas Clúster

- Software de Sistema
  - Sistema Operativo
    - Capa de control HW
  - Middleware
    - Capa de unión
  - La barrera no está siempre clara

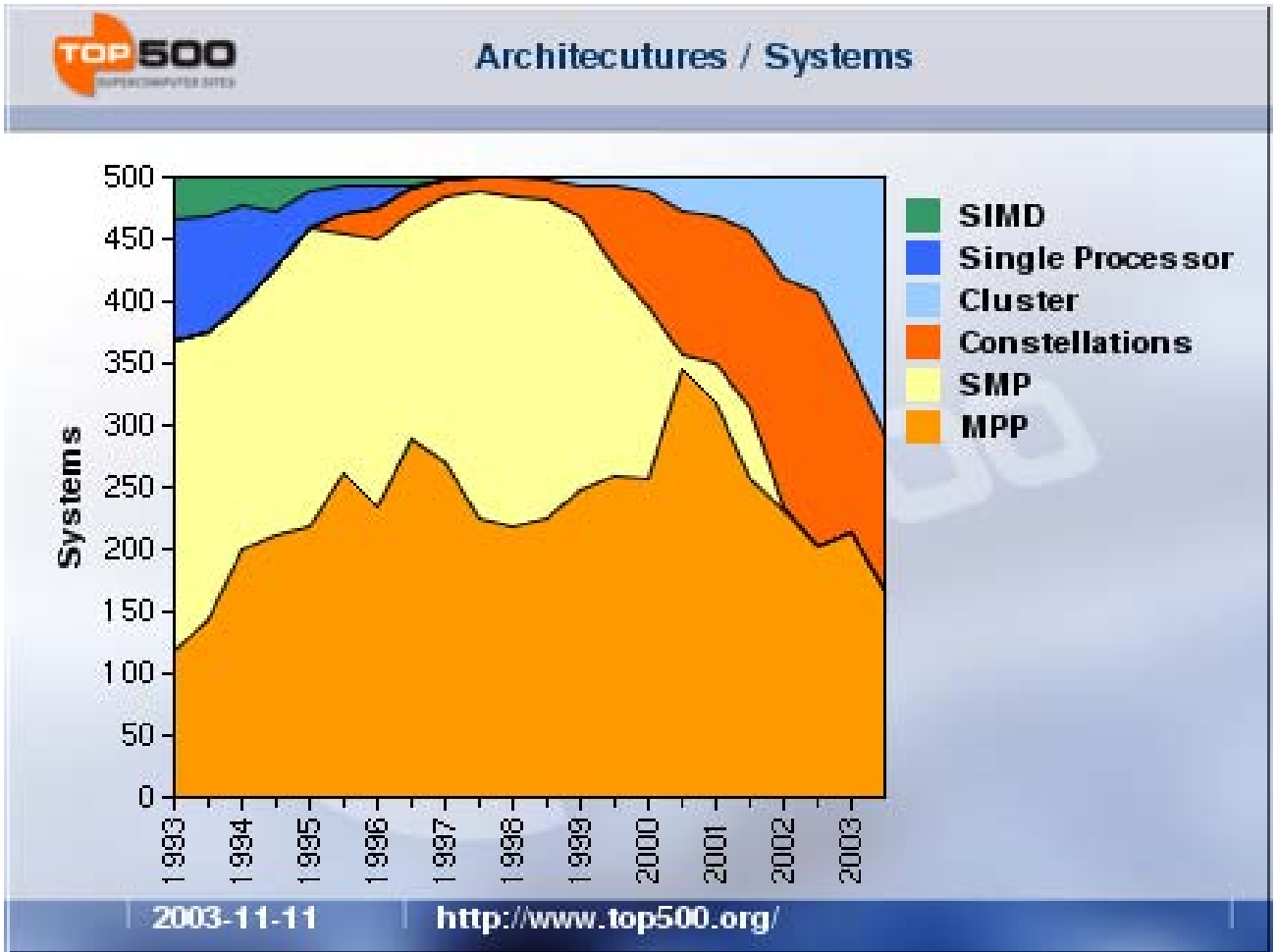


Introducción a las Arquitecturas Clúster

- No distinguiremos entre
  - SO
  - Middleware
    - El middleware está relacionado con el SO
- Objetivos del Software de Sistema
  - Rendimiento/Escalabilidad
  - Robustez
  - SSI (Single-System Image)
  - Extensibilidad
  - Escalabilidad
  - Heterogeneidad

Evolución histórica de las Arquitecturas de Computadores

- El clúster es la arquitectura del futuro:



Software de sistema

- Sistemas de ficheros
  - NFS (y Cluster NFS)
  - PVFS
  - OpenSSI
  - CFS (Lustre)
  - General Parallel Filesys4linux (GPFS)
  - iSCSI linux (e.g. SAN cisco SN 5420)
  - Global File System (GFS) – Sistina, openGFS

Software de sistema

- Gestión de nodos
  - NIS, NIS+
  - LDAP
- Monitorización
- SNMPs
- Scamond (muxia)

Software de sistema linux

- Sistemas de colas
  - PBS
    - pbsnodes –a
    - qsub –lnodes=n –np n1 –npn n2 run-script
    - qdel job
    - qstat -Q, -rn, -f, -u User
    - qmgr: print server
  - Maui
- Balanceo de carga
  - Mosix (OpenMosix)
  - Condor
  - Linux Virtual Server (LVS)
- SSI
  - BPROC

Software de sistema linux

- Paso de mensajes. Bibliotecas
  - MPI (Message Passing Interface)
    - MPICH
    - LAM
    - mpiJava, CCJ, JMPI, MPJ
  - PVM
- OpenMP

Problemas típicos en Extreme Computing

- Extreme Computing
  - Reducción del consumo (- HD)
- High availability
  - Detección de fallos y recuperación
  - Tolerancia a fallos
  - Regla de los 9 (90%,99%,99.9%...)
- High performance
  - Altas prestaciones

Clusters en aplicaciones científicas

- Simulaciones (earth simulator)
- Modelos matemáticos
- Genoma humano (y del arroz - China)
- Predicción meteorológica
- Predicción de movimiento de productos en el mar
- Exploración del espacio (seti at home)

Clusters en aplicaciones empresariales

- Google (>10.000 linux boxes)
- Text retrieval
- BBDDs
  - (prometheus project – mysql, postgres)
- Servidores y contenedores Web
  - Apache, JBOSS…
- Industria del automóbil
- Prospecciones petrolíferas

Linux clustering con consolas

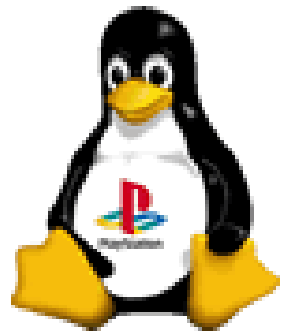
- Cómputo científico en PS2
  - Universidad de Illinois en Urbana-Champaign
    - Exploran el uso de PS2 en Cómputo Científico y Visualización de Alta Resolución
    - Construyen un Cluster de 70 PS2
    - (Sony Linux Kit + MPI, PBS, Maui Scheduler)

Linux clustering con consolas

- PS3 (2005)
  - PowerPC? A 4GHz y 8 procesadores vectoriales con un bus de 1024 bits que proveerán de 256 Gflops
  - Earth Simulator tiene 35860 Gflops (+- 200 PS3)

Linux clustering con consolas

- ¡No hay quién se resista al pingüino!



Linux clustering con consolas

- Consolas:
  - Linux en PS2 (playstation2-linux.com)
    - Hw
      - CPU Emotion Engine 128-bits 300MHz (Extended MIPS III RISC) con unidades vectoriales para tratar polígonos en 3D
    - Sw
      - SONY Linux Kit (100€) basado en Linux Kondara y en Red Hat (o Blackrhino de Debian)
  - Linux en M$ XBox (xbox-linux.org)
    - Hw
      - Celeron 733, 64MB, nVidia GeForce 3MX y 10/100 Ethernet
    - Sw
      - Xebian, basada en Debian

Linux clustering con consolas

- Consolas:
  - Linux en GameCube (gc-linux.org)
    - HW
      - IBM **PowerPC** CPU, **486** MHz, **ATI** graphics, **40 MB** de RAM
    - SW
      - GameCube Linux, un pequeño Linux con telnet y servidor web
  - Linux en Sega DreamCast (linuxdc.sourceforge.net)
    - Hw
      - MIPS R4000 100MHz
    - Sw
      - LinuxDC

Taller

- Programa MPI simple que será ejecutado en tres sistemas clúster:
  - muxia.des.udc.es
  - bw.cesga.es
  - sd.cesga.es

Taller

- muxia.des.udc.es
  - 8 nodos duales PIV Xeon a 1.8 GHz
  - 2 nodos duales PIV Xeon a 2.8 GHz
  - 40 GFLOPS
  - 20 GBs RAM
  - Redes SCI y Fast Ethernet (admin)
  - Red Hat 7.3

Taller

- bw.cesga.es (dentro de svg.cesga.es)
  - 16 nodos PIII a 1 GHz
  - 16 GFLOPS
  - 8 GBs RAM
  - Redes Myrinet y Fast Ethernet (admin)
  - Red Hat 7.2

Taller

- sd.cesga.es (#227 en top500)
  - 2 nodos SMP (HP Superdome)
  - HP Sd cuenta con 64 Itanium 2 con 6MB
  - 768 GFLOPS
  - 7 TBs RAM
  - Red Infiniband
  - HP-UX, Windows, Linux, OpenVMS (simul)

Taller

- Virginia Tech X (#3)
  - 1100 nodos duales G5 (2200 procs)
  - 10,3 TFLOPS
  - 4,4 TBs RAM
  - Red Infiniband
  - Jagua (Mac OS X 10.2) 32bits
  - 4.500.000 € (“self-made”)

Taller

- Centro Nac. Computación (IBM&SP) #2
  - 4500 procs.
  - 20 TFLOPS
  - 9 TBs RAM
  - Red Infiniband
  - Linux
  - 70.000.000 € (total project)

Taller

- Recordemos (Casos prácticos):
  - Se procederá a la ejecución de un programa MPI paralelo en los nodos del clúster
  - Se analizará el estado de la cola
  - Se analizarán los resultados obtenidos

Taller

- Programa MPI simple:

```
#include <stdio.h>
#include <mpi.h>

main(argc, argv)
int argc;
char *argv[];
{
    char name[BUFSIZ];
    int length;
    MPI_Init(&argc, &argv);
    MPI_Get_processor_name(name, &length);
    printf("%s: hello world\n", name);
    MPI_Finalize();
}
```

Taller

- Compilación: mpicc
- Ejecución sin colas: mpirun
- Envío a colas PBS: qsub (scasub)
- Consulta estado cola: qstat, qstat –rn,qstat –f, qstat –Q,qstat –u user
- Salida estándar de trabajos: job.o#job
- Salida de error de trabajos: job.e#job

Formación en Cluster Computing

- Google!
- Asignaturas en FIC:
  - OPP
  - AEC
  - ATF
  - IS, XR
- Tercer Ciclo:
  - Tecnologías de la Información (USC y DES)
- Curso en el Cesga:
  - “Computación en Clusters: Administración y Programación”, 26-30 de Abril de 2004

Referencias

- www.buyya.com/cluster (imprescindible)

- clusters.top500.org
- www.clustercomputing.org
- www.cesga.es
- www.hispacluster.org

Contacto

- Guillermo López Taboada (Lab. 1.2)
  - taboada at udc dot es
  - http://www.des.udc.es/~gltaboada

Grupo de Arquitectura de Computadores

Depto. De Electrónica y Sistemas

Facultad de Informática - UDC

THE END

- Gracias!

…y preguntas!!

Documentación disponible en web de GPUL