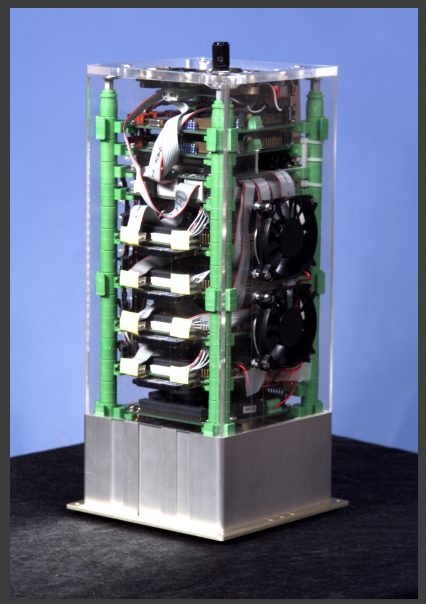
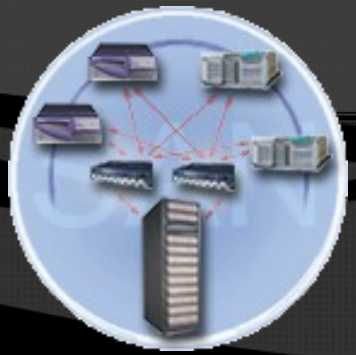
Centrologic
Computational Logistic Center

Sistemas Distribuidos y Alta Disponibilidad





Ponente:
Juan Miguel Taboada Godoy
juanmi@centrologic.com - http://www.centrologic.com

Necesidades de la Empresa

Computación

Meteorología
Simuladores
Cálculos financieros
Inversión
Investigación

Almacenamiento

Web
FTP
Backup
Bases de Datos
Escritorios distribuidos

Objetivos

Reducir gastos
Incrementar calidad
Clientes contentos
Alta disponibilidad
Eficiencia

Clusters

Alta Disponibilidad

Soluciones

Globatic
IBM
SUN Microsystems
Supermicro
Intel
AMD

Condicionantes

Exceder requerimientos
Funcionar siempre
Servicio de calidad
Autoprotección contra errores
Redundancia de datos

Centrologic
Computational Logistic Center

¿Qué es un cluster?

Cluster => Conjunto => Elementos

Unidad = Átomo del cluster
Características mínimas (grupo)
Mismos objetivos

Problema divisible en partes
Un átomo trabaja con un trozo
Trozos de solución => Solución completa

Computación

Muchos ordenadores para aparentar ser uno sólo

Trasparencia de uso

Jerarquía heterogénea

Cambios dinámicos en la topología

Sistema de ficheros distribuido

Alta escalabilidad

Historia

Inicio del proyecto

Versión 0 (1977) -> UNIX with SP [PDP]
Versión 4 (1988) -> MOSIX [BAX780]
Versión 6 (1991) -> MOSIX [80486 y Pentium]
Versión 7 (1998) -> MOSIX [x86 y Linux]
Moshe Bar
Universidad?
Grupo de estudiantes
Desarrollo libre
Se divide el proyecto (10 febrero 2002)

MOSIX

Comercial
Crecimiento lento
Código anticuado

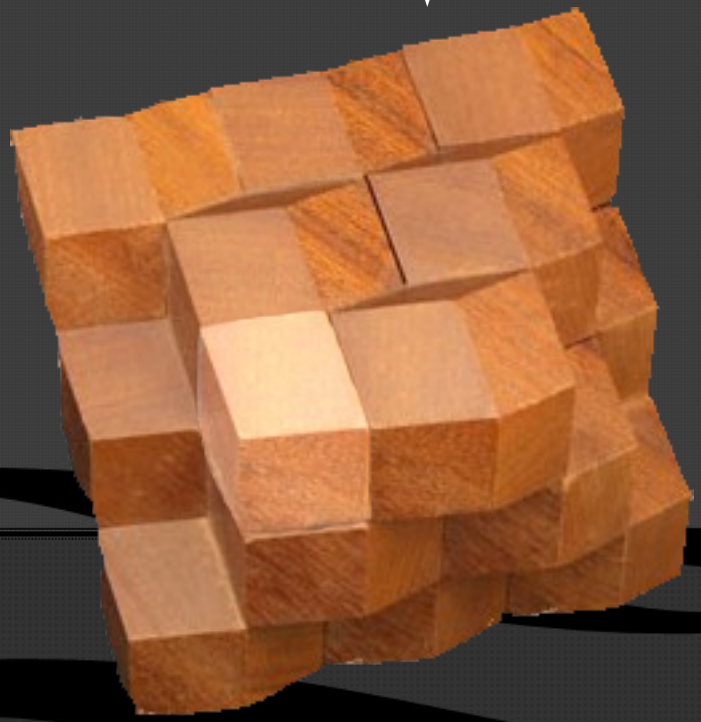
OpenMosix

Kernel Linux 2.4.24
Soporte para IA64
Totalmente rediseñado
Mosix Userlands
Licencia GNU/GPL

Funcionamiento: Kernel

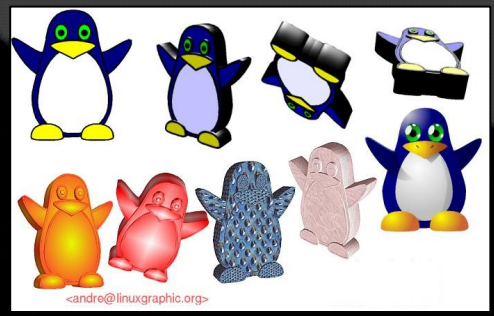
Descripción

OpenMosix es un parche
Sólo un parche por núcleo
Los parches son cambios en el núcleo



Características

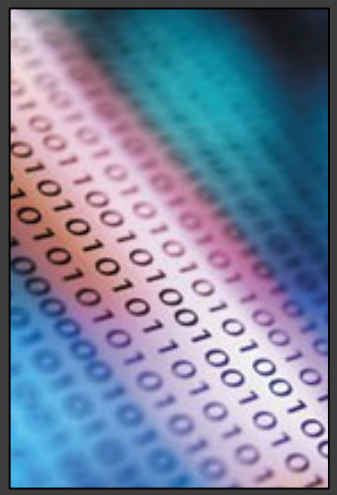
Sólo entiende de procesos
Espacio de memoria exportable
NO Threads (Hilos)
Programación por FORKs
Distribuido

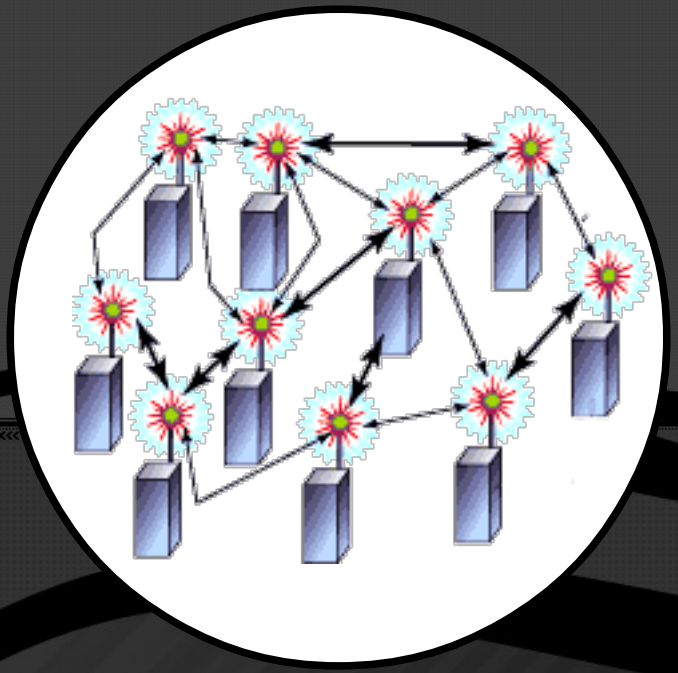


Funcionamiento: Migración

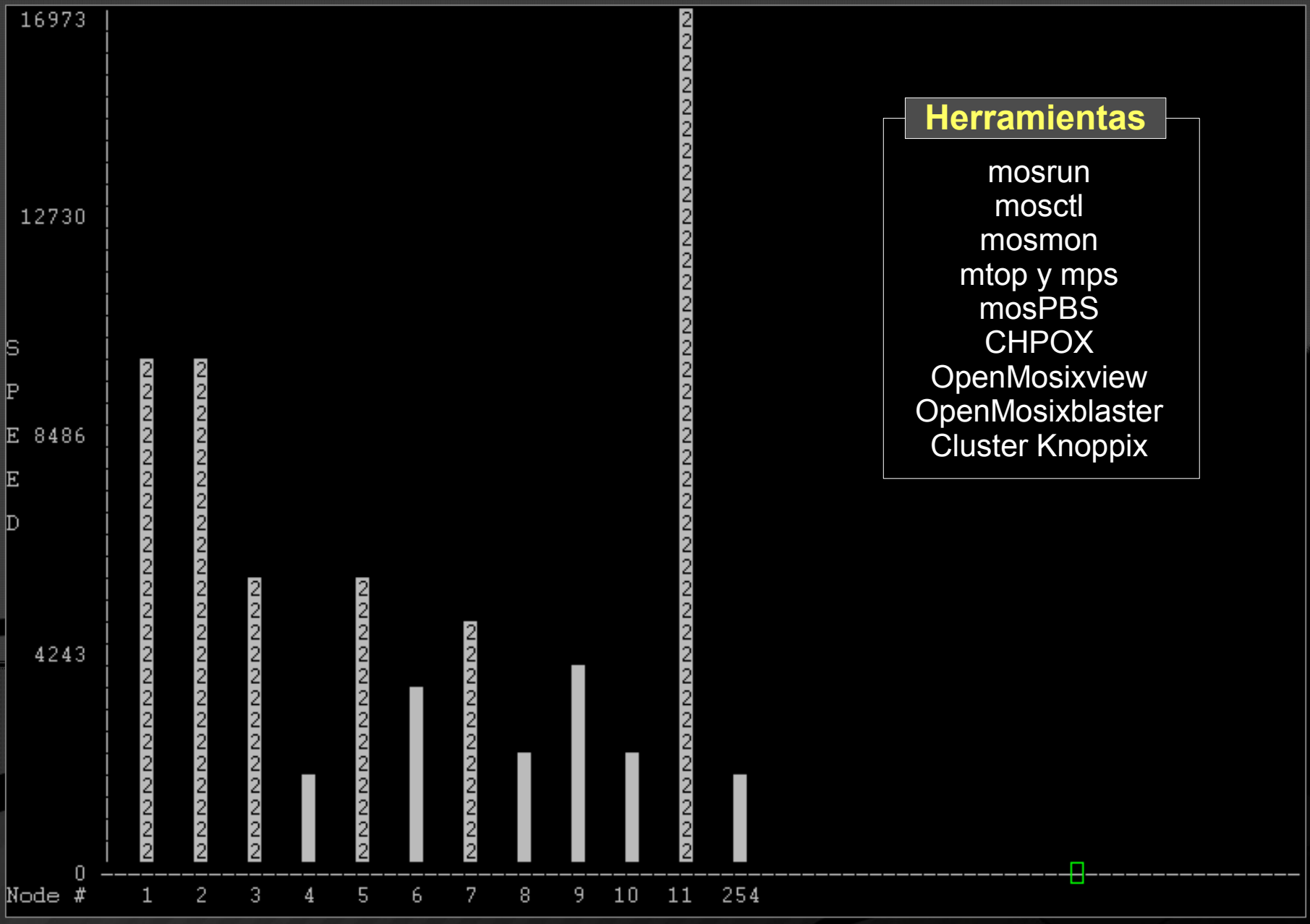
Características

Exporta a la más potente y libre
E/S = Volver
Balanceo de carga automático
Algoritmo de retención
Kill por desconexión





Funcionamiento: Herramientas



Herramientas

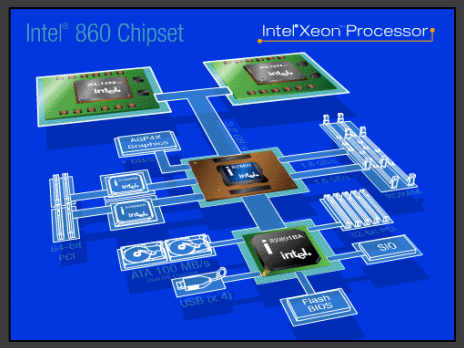
- mosrun
- mosctl
- mosmon
- mtop y mps
- mosPBS
- CHPOX
- OpenMosixview
- OpenMosixblaster
- Cluster Knoppix

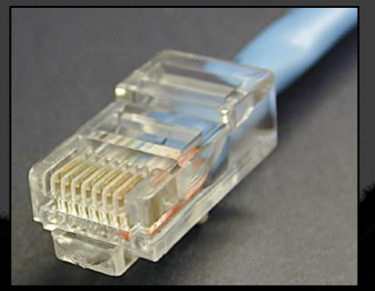
Centrologic
Computational Logistic Center

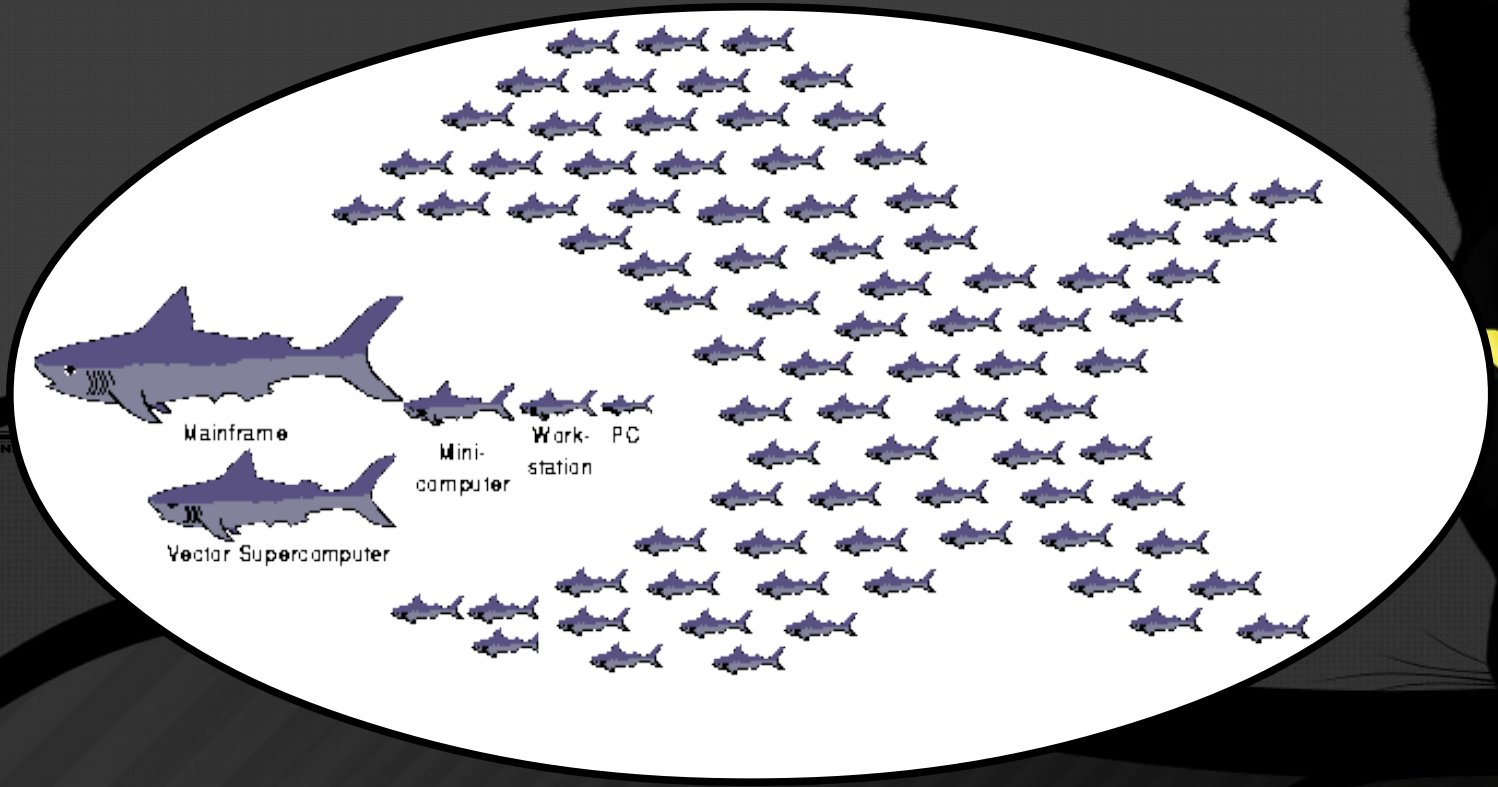
Instalación: Planificar

Pasos

Necesidades de cómputo
Acceso a disco
Potencia de los nodos
Escalabilidad
Refrigeración
Sobrecarga de la red (aislada/distribuida)
Jerarquía centralizada o distribuida
Gestión de colas (multiples usuarios)
Tiempo de ejecución (Alta disponibilidad)







Instalación: Núcleo

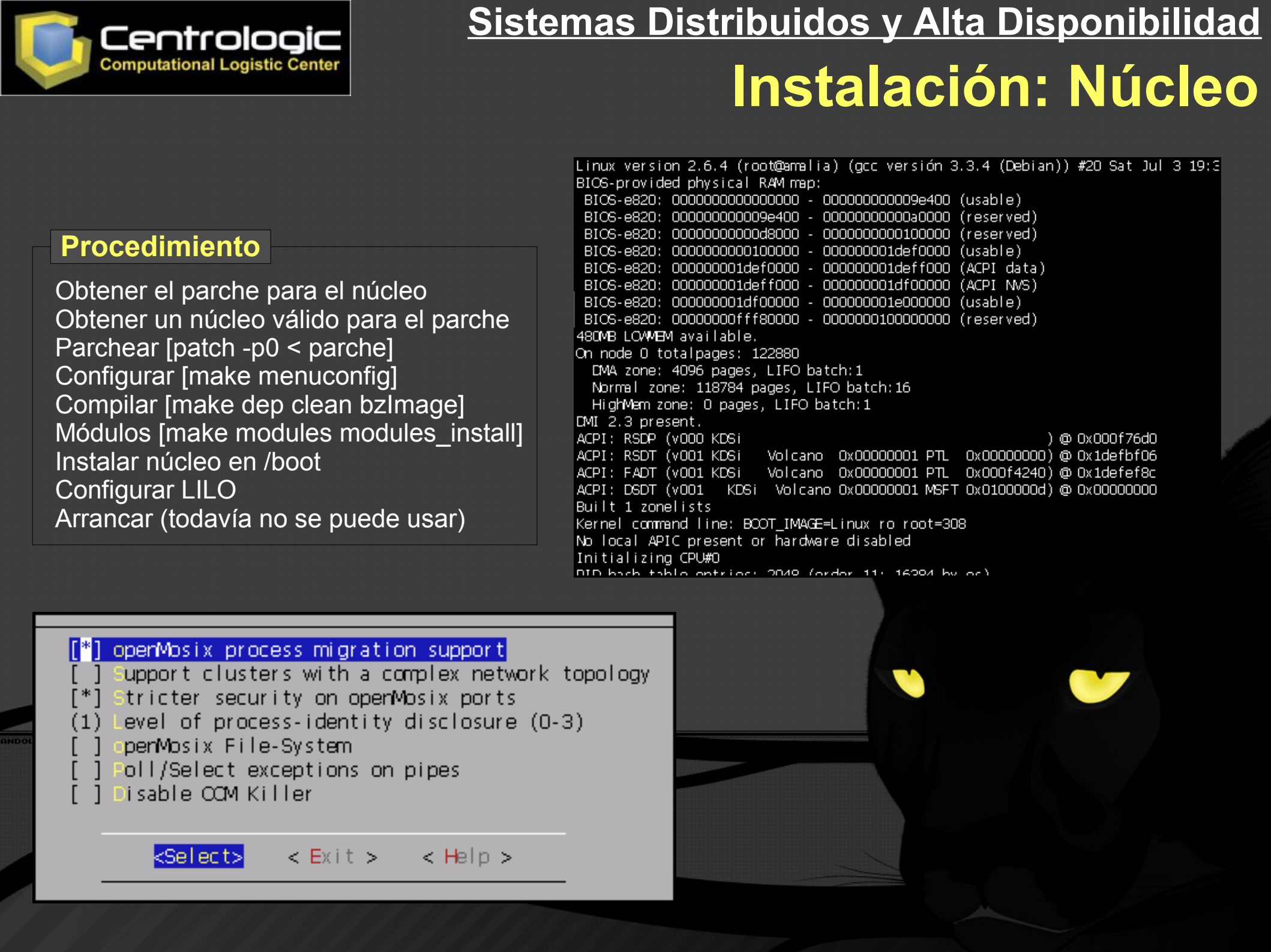
Procedimiento

Obtener el parche para el núcleo
Obtener un núcleo válido para el parche
Parchear [patch -p0 < parche]
Configurar [make menuconfig]
Compilar [make dep clean bzImage]
Módulos [make modules modules_install]
Instalar núcleo en /boot
Configurar LILO
Arrancar (todavía no se puede usar)

```
Linux version 2.6.4 (root@amalia) (gcc versión 3.3.4 (Debian)) #20 Sat Jul 3 19:3
BIOS-provided physical RAM map:
 BIOS-e820: 0000000000000000 - 000000000009e400 (usable)
 BIOS-e820: 000000000009e400 - 00000000000a0000 (reserved)
 BIOS-e820: 00000000000d8000 - 0000000000100000 (reserved)
 BIOS-e820: 0000000000100000 - 000000001def0000 (usable)
 BIOS-e820: 000000001def0000 - 000000001deff000 (ACPI data)
 BIOS-e820: 000000001deff000 - 000000001df00000 (ACPI NVS)
 BIOS-e820: 000000001df00000 - 000000001e000000 (usable)
 BIOS-e820: 00000000fff80000 - 0000000100000000 (reserved)
480MB LOWMEM available.
On node 0 totalpages: 122880
  DMA zone: 4096 pages, LIFO batch:1
  Normal zone: 118784 pages, LIFO batch:16
  HighMem zone: 0 pages, LIFO batch:1
DMI 2.3 present.
ACPI: RSDP (v000 KDSi                                    ) @ 0x000f76d0
ACPI: RSDT (v001 KDSi   Volcano  0x00000001 PTL  0x00000000) @ 0x1defbf06
ACPI: FADT (v001 KDSi   Volcano  0x00000001 PTL  0x000f4240) @ 0x1defef8c
ACPI: DSDT (v001   KDSi  Volcano 0x00000001 MSFT 0x0100000d) @ 0x00000000
Built 1 zonelists
Kernel command line: BOOT_IMAGE=Linux ro root=308
No local APIC present or hardware disabled
Initializing CPU#0
PID hash table entries: 2048 (order: 11, 16384 bytes)
```

```
[*] openMosix process migration support
[ ] Support clusters with a complex network topology
[*] Stricter security on openMosix ports
(1) Level of process-identity disclosure (0-3)
[ ] openMosix File-System
[ ] Poll/Select exceptions on pipes
[ ] Disable OOM Killer

    <Select>    < Exit >    < Help >
```



Instalación: Herramientas

Procedimiento

Descargar openmosixUserlands
Descomprimir
Modificar Makefile (para ruta al kernel)
Compilar (make)
Instalar (make install)
Script de arranque (init.d+runlevel)
Sólo nos queda configurar los nodos

```
rxvt
ations --enable-host-specifics-asms --with-wx; make; make install
checking build system type... i686-pc-linux-gnu
checking host system type... i686-pc-linux-gnu
checking target system type... i686-pc-linux-gnu
checking if you are configuring for another platform... no
checking for standard CFLAGS on this platform...
checking for gcc... gcc
checking for C compiler default output... a.out
checking whether the C compiler works... yes
checking whether we are cross compiling... no
checking for suffix of executables...
checking for suffix of object files... o
checking whether we are using the GNU C compiler... yes
checking whether gcc accepts -g... yes
checking for g++... g++
checking whether we are using the GNU C++ compiler... yes
checking whether g++ accepts -g... yes
checking whether make sets ${MAKE}... yes
checking for ld used by GCC...
```

Instalación: Configurar

/etc/cluster.map

Número: número del nodo
Dirección IP: ip en la que comienza el rango
Cantidad: indica la longitud del rango

Es importante que los nodos no se solapen
Ej:

Número	Dirección IP	Cantidad
1	192.168.1.1	10
5	192.168.1.11	10

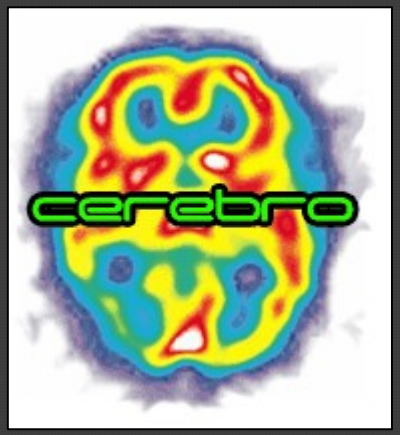
/etc/cluster.map

Número	Dirección IP	Cantidad	Correspondencia
1	192.168.50.1	3	192.168.50.(1,2,3)
2	192.168.50.10	1	192.168.50.10
3	192.168.50.11	1	192.168.50.11
4	192.168.50.12	1	192.168.50.12
5	192.168.50.13	3	192.168.50.(13,14,15)
6	192.168.50.16	2	192.168.50.(16,17)

Sistemas Distribuidos y Alta Disponibilidad

En la vida real

Estructura heterogénea
Temperatura de la sala a 18ºC
Potencia centrada en paralelismo



Sistemas Distribuidos y Alta Disponibilidad

Pruebas y resultados

SIN CLUSTER

AMD Atlon XP a 2200Mhz
Generador RSA (llave privada y pública
X minutos en generar 10 000 llaves RSA

Media: Y,Z llaves por segundo

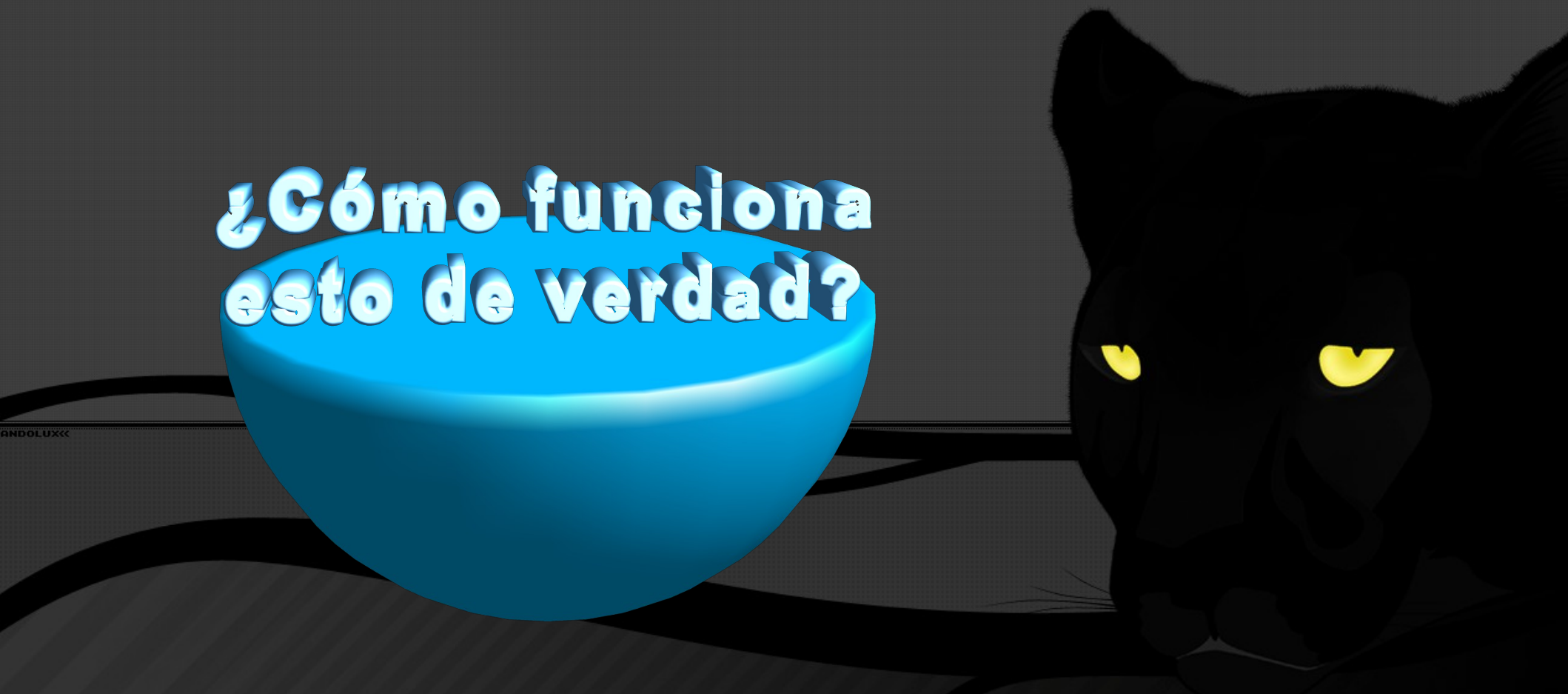
CON CLUSTER

10 Pentium 3 a 1000Mhz
Generador RSA (llave privada y pública)
42 minutos en generar 10 000 llaves RSA

Media: 3,9 llaves por segundo

Centrologic
Computational Logistic Center

Demostración en tiempo real



Descanso según el protocolo

Tira cómica gracias a:

es.comp.os.linux.

¿Qué es la Alta Disponibilidad?

Concepto

HA: High Availability
Máxima disposición temporal
Concepto de redundancia
Rápida detección y recuperación
Consistencia de las copias
Importancia de la seguridad

Ideas

Buscar amistad **;-)**
Trabajar sábado noche
Niños en la guardería
Solvencia económica
Lo contrario de Windows
Pareja/conyuje de viaje
Una urgencia

Centrologic
Computational Logistic Center

Sistemas Distribuidos y Alta Disponibilidad

Teoría del KAOS

Detalles

Números semi-aleatorios en computación
KAOS = Sin orden = No previsible
Política anti-KAOS
Seguridad VS Alta Disponibilidad
Gasto tiene que compensar a pérdidas

Preguntas

¿Realmente lo necesito?
¿Qué inversión estoy dispuesto a realizar?
¿Qué deseo proteger? ¿de qué o quién?
¿Me he asesorado correctamente?
¿Empleados informados?

Fuentes de errores

Software
Hardware
Naturaleza
Empleados
Incompetencia

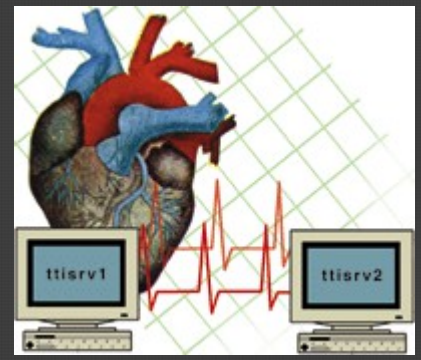
Centrologic
Computational Logistic Center

Sistemas Distribuidos y Alta Disponibilidad

Heartbeat

Objetivos

Asegurar actividad
Redundancia en comunicaciones de control
Detectar caidas del sistema
Recuperar actividades muertas
Takeover (Intercambiar IP)



Características

Comunicar su estado a los otros nodos
Comprobar estado de los nodos
Suplantar las actividad de los nodos caidos
Ocupar la IP asignada al servicio
Tomar acciones extras ante caidas
Tiempo de recuperación de 30 a 60 segundos

Ventajas

Gratuito
Licencia GNU/GPL
Fácil de configurar
Comunicación por RED y por SERIE
Sistemas heterogéneos
Altamente soportado
Integrable con otras aplicaciones

Inconvenientes

Nodo de reserva desaprovechado
Pérdida de conexiones remotas
No replica los datos entre los nodos
Obliga a centralizar el sistema de ficheros

Centrologic
Computational Logistic Center

Sistemas Distribuidos y Alta Disponibilidad

DRBD

Características

Replicación Distribuida de Dispositivos de Bloques
Copia de seguridad en tiempo real
Sistema de transaciones a nivel del sistema de ficheros
Automatismo en la sincronización

Funcionamiento

Compilar módulo
Genera nuevos dispositivos sobre partición
Sincroniza los sistemas de fichero automáticamente
Actualiza los sitemas de ficheros en tiempo real
Master -> Slave

Ventajas

Gratuito y con licencia GNU/GPL
Replicación de datos automática
Funcionamiento remoto
Fácil de configurar
Abstracción con el sistema operativo

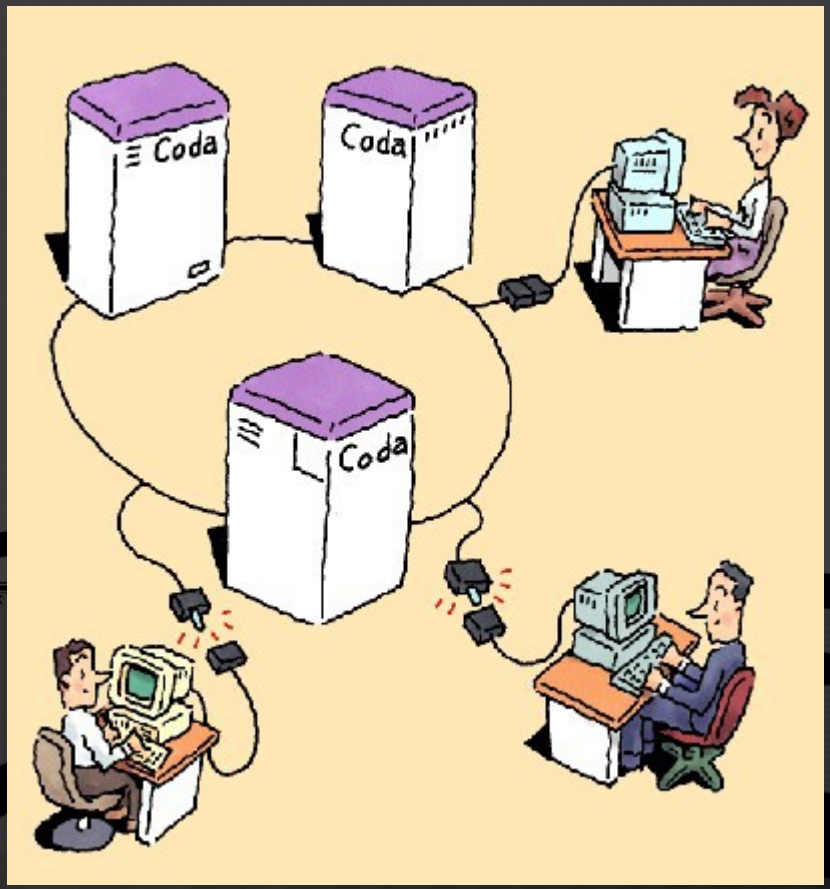
Centrologic
Computational Logistic Center

Sistemas Distribuidos y Alta Disponibilidad

CODA

Características

Es un sistema de ficheros distribuido
Compilable en núcleo estándar
Sistema de ficheros remoto
Alta fiabilidad
Estabilidad (Mejor que NFS/SAMBA)
Licencia GNU/GPL
Gestión de accesos Multiusuario
NIS centraliza claves e información de usuarios
NIS+CODA = Escritorio virtual (remoto)



Sistemas redundantes

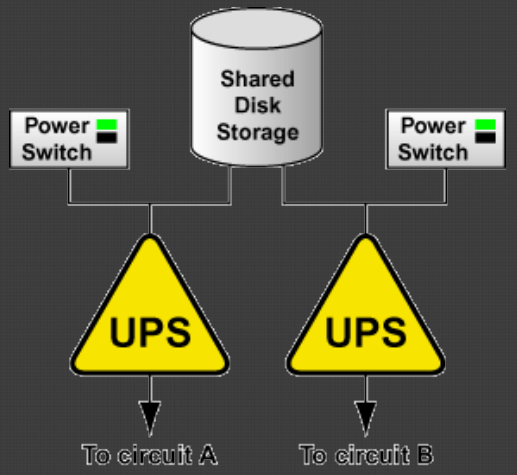
Premisas

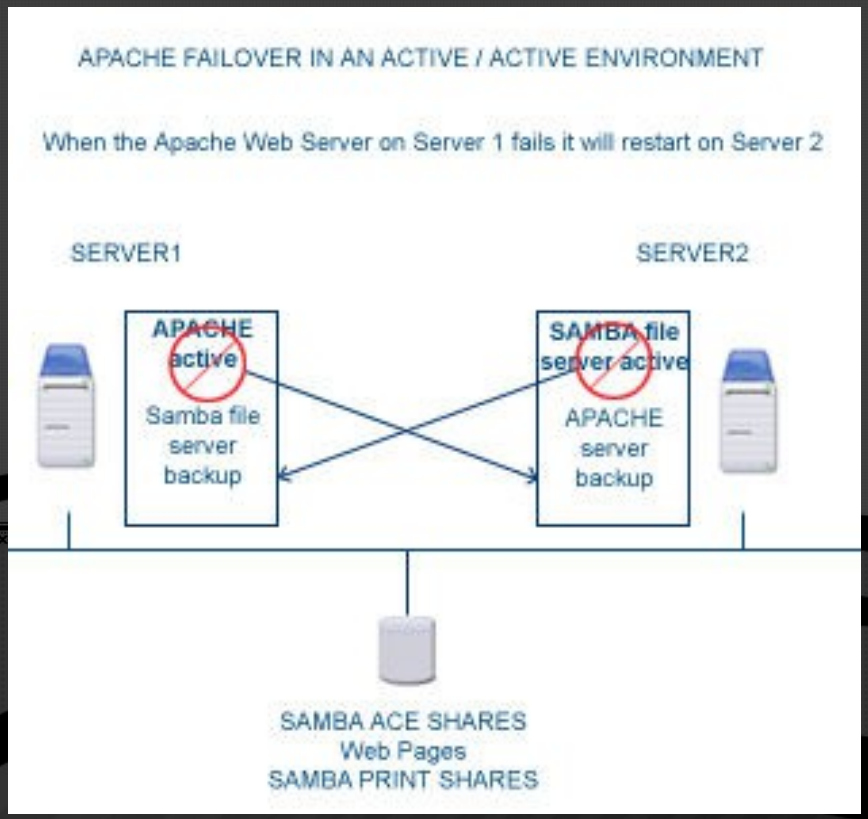
Discos replicados
Fuente redundante
Más de un acceso acceso a Internet
Red duplicada entre todos los equipos a todos los niveles
Alimentación eléctrica (replicada e ininterrumpida)
Localización espacial (sistemas duplicados)

Instalación: Planificar

Estudio previo

Necesidades
Servicios a ofrecer
Tiempo de inactividad máximo
Presupuesto (teoría del KAOS)
Seguridad
Localización (detalles de las instalaciones)





Instalación: Hearbeat

Antes de...

Interconectar equipos por RED o SERIE
Descargar Heartbeat
Instalar librería libnet y libglib-devel
Instalar Heartbeat

Configuración

General: /etc/ha.d/ha.cf
Validación: /etc/ha.d/authkeys
Servicios: /etc/ha.d/haresources

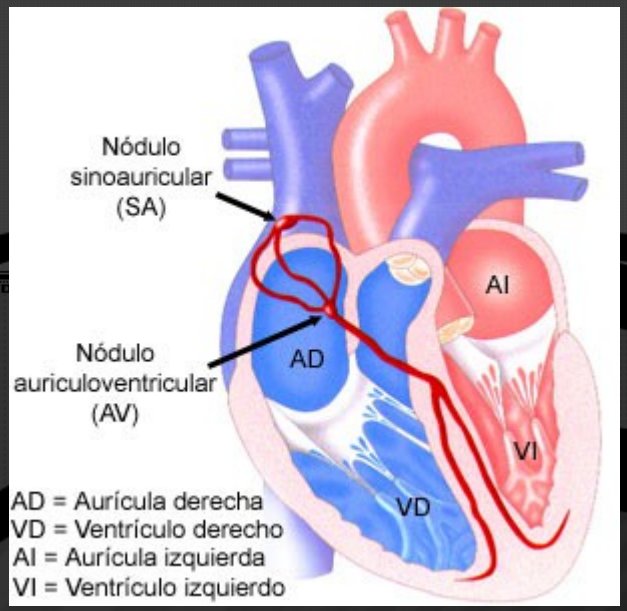
/etc/ha.d/ha.cf

```
# define nodes in cluster
node     ttisrv1
node     ttisrv2

# time a system must be unreachable
before considered dead (seconds)
deadtime 5

# set up for the serial heartbeat pulse
serial   /dev/ttyS0
baud     19200

# interface to run the network heartbeat
pulse
udp      eth1
```



/etc/ha.d/haresources

```
# use ttisrv1 as primary, use 192.168.0.100 as shared IP
ttisrv1 192.168.0.100 Filesystem::/dev/sda1::/ttidisk::ext2 \
smb nfslock nfs
```

Centrologic
Computational Logistic Center

Sistemas Distribuidos y Alta Disponibilidad

En la vida real

Empresas

ISPs
Bancos
Contabilidad
Bases de datos

Ciencia y tecnología

Centros de cálculo
Meteorología
Investigación
Astronomía
Programas @home
Simuladores de vida
Ejército

Particulares

Servidores de Juegos
FTPs
Servidores de redes P2P
Streaming

Sistemas Distribuidos y Alta Disponibilidad

Teoría de los 9s

Seis Nueves

Nº de nueves	Disponibilidad	Desconexión/Año
1	90%	37 días
2	99%	3,7días
3	99,9000%	8,8horas
4	99,9900%	53minutos
5	99,9990%	5,3minutos
6	99,9999%	32 segundos

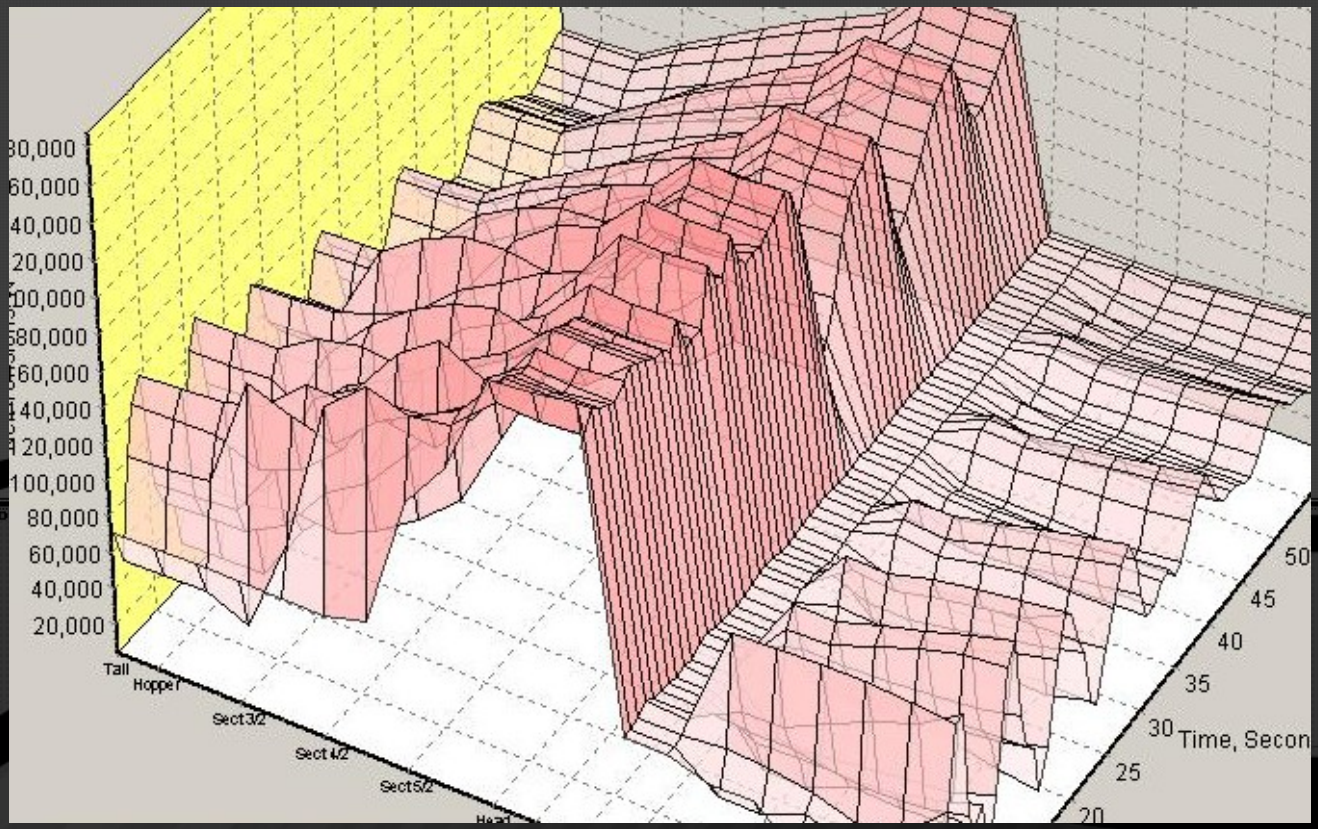
Pruebas y resultados

Apache

Entre 30 y 60 segundos inactivo (Heartbeat)
Recuperación instantánea de la dirección IP
Recuperación total del servicio instantánea

Samba

Entre 30 y 60 segundos inactivo (Heartbeat)
Recuperación instantánea de la dirección IP
Recuperación total del servicio en 1 y 3 minutos



Ahora es tu turno

Computational Logistic Center

Gracias

Juan Miguel Taboada Godoy

juanmi@centrologic.com
http://www.centrologic.com
Teléfono: +34 902884062