Tema 1: Evolución de la arquitectura de los sistemas paralelos y de sus modelos de programación

Introducción a la computación paralela

Asignatura: Arquitectura de Sistemas Paralelos

Profesor: Francisco Javier Gómez Arribas

Departamento de Tecnología Electrónica y de las Comunicaciones

UNIVERSIDAD AUTONOMA DE MADRID

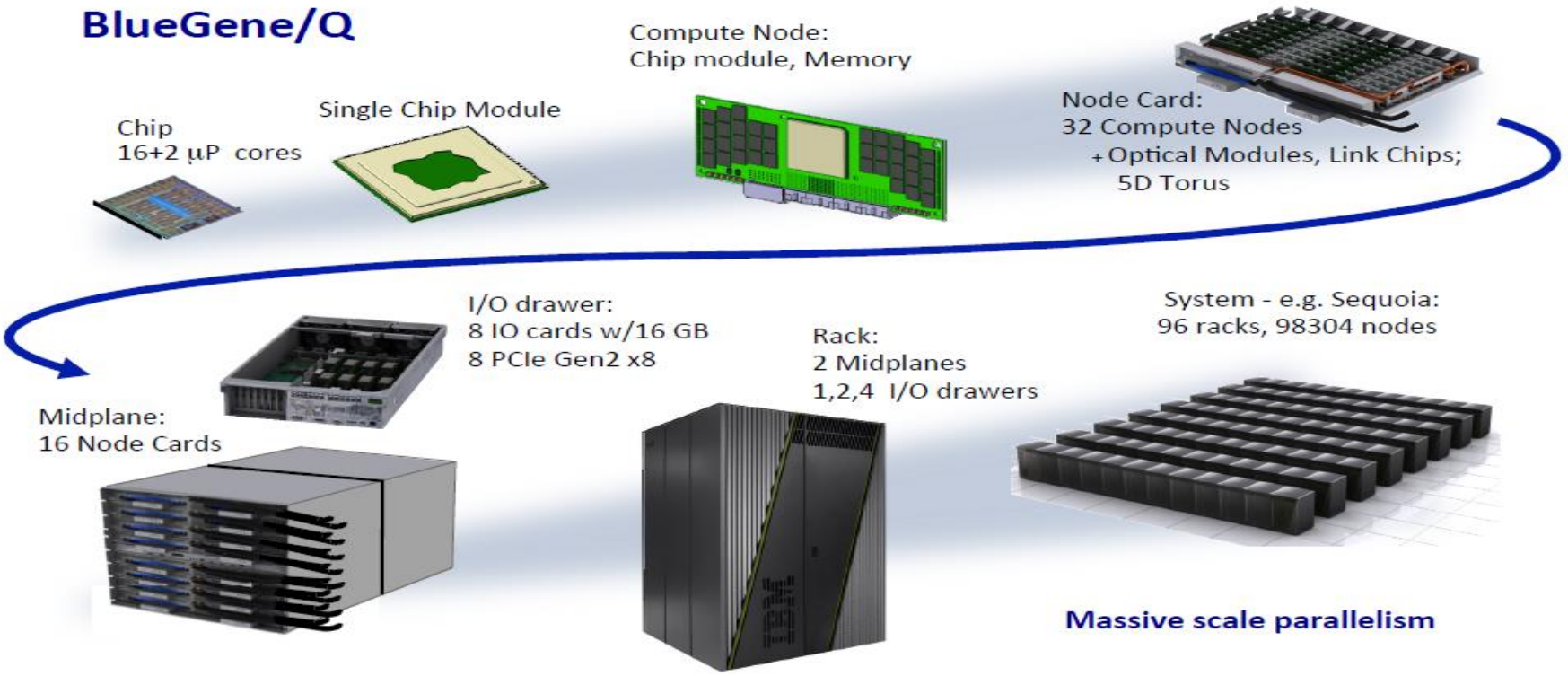
Escuela Politécnica Superior

Tema 1: Contenidos

- **Introducción a la computación paralela**
  - Motivación y Objetivos
  - Métricas de rendimiento. Leyes de Amdahl y Gustafson
  - Aplicaciones de la computación paralela
  - Arquitecturas para procesamiento en paralelo: Clasificación
- **Arquitectura de los sistemas multicomputador**
  - Arquitectura y componentes de un cluster
  - Tipos de cluster y dominio de aplicación: HPC, Beowulf, Hadoop.
  - Planificación y balanceo de tareas en un cluster
  - Top 500. Ejemplos de SuperOrdenadores
- **Modelos de Programación**

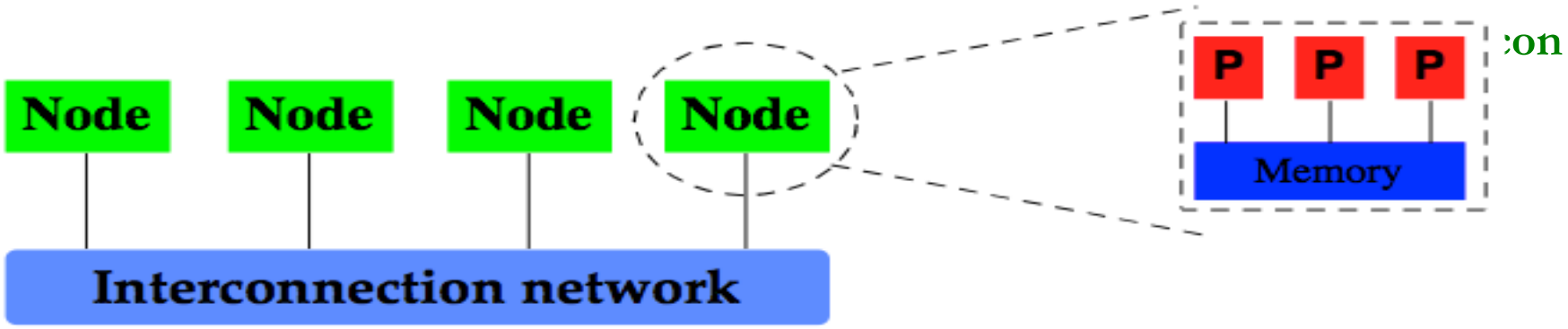
Sistema Multicomputador

Una gran cantidad de nodos de computación homogéneos, interconectados entre sí y con un subsistema de Entrada/Salida



Multicomputador: Arquitectura

- Multicomputador = Nodos + Red de Interconexión.
  - **Nodo = procesador(es) + memoria local**
    - **El acceso a memoria local es rápido, porque no involucra conexión de red (acceso a memoria convencional en sistema uniprocesador)**



Multicomputador: Componentes

Un sistema multiprocesador/multicore con memoria compartida (NUMA).

Componentes:

- **Procesador: Multi-core/many-core**
- **Aceleradores**
  - **(NVIDIA GPGPUs**
  - **IntelXeon Phi)**
- **Red de Interconexión con RDMA (Remote Direct Memory Access) networking**
  - **InfiniBand y RoICE (RDMA over Converged Enhanced Ethernet)**
- **Almacenamiento:**
  - **HDDs,**
  - **Solid State Disks (SSDs),**
  - **Non-Volatile Random-Access Memory(NVRAM), y NVMe SSD.**

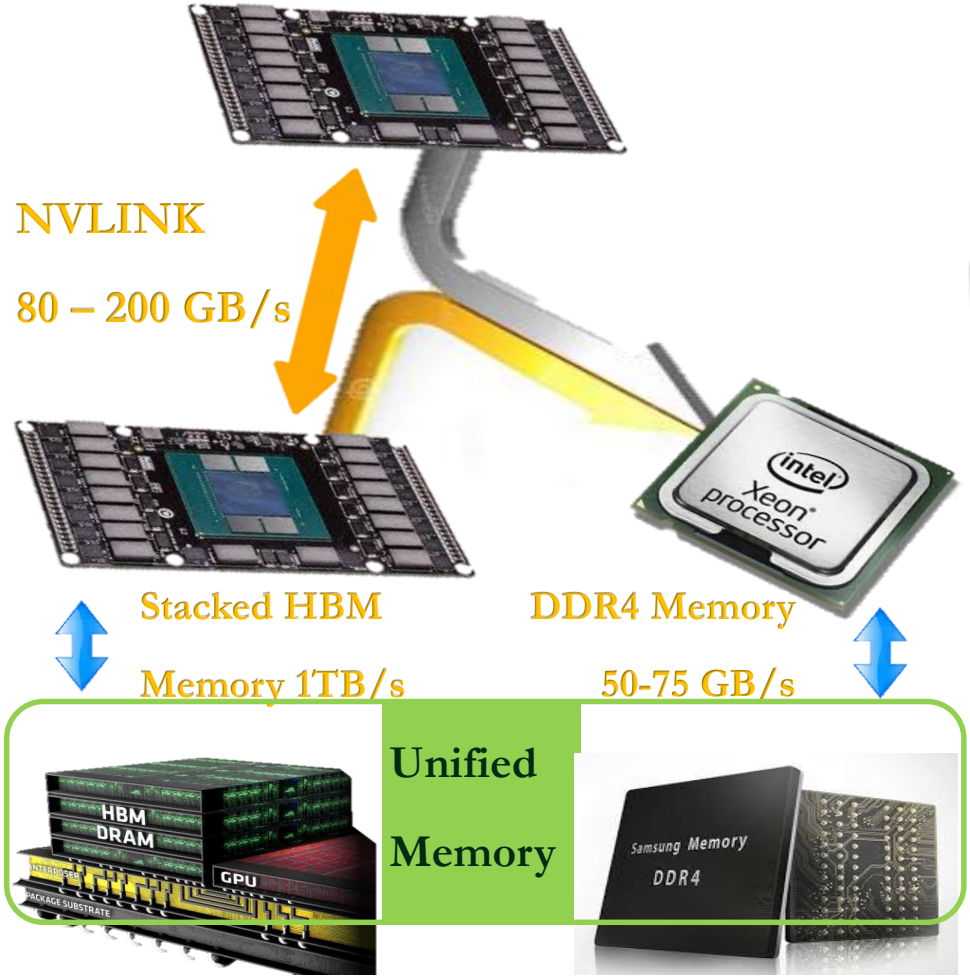
Multi-core Processors

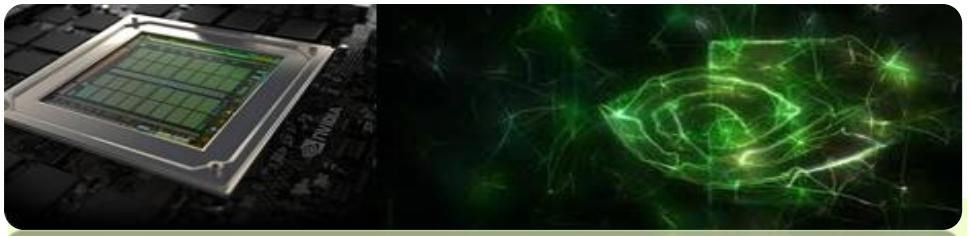
Accelerators / Coprocessors
high compute density, high performance/watt
>1 TFlop DP on a chip

High Performance Interconnects - InfiniBand
:1usec latency, 100Gbps Bandwidth

SSD, NVMe-SSD, NVRAM

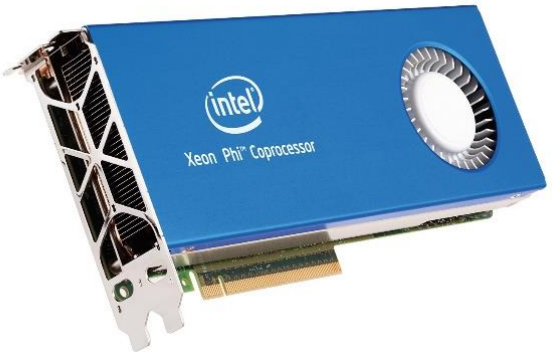
Coprocesadores GPU: Arquitectura Nvidia PASCAL





- ⋆NVLink
  - GPU Interconnect at 80-200 GB/s
- ⋆Stacked 3D Memory
  - 3x Higher Bandwidth (~1 TB/s)
  - 2.7x Larger Capacity
  - 2x More Energy Efficient per bit

Coprocesadores: Intel Xeon Phi



- **KNL-F with dual port**
- **Omni-Path fabric**

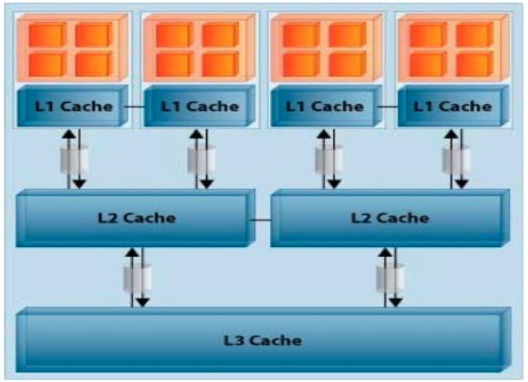
2013
Knights Corner
Intel® Xeon Phi™ x100 Product Family
22 nm process
Coprocessor
Over 1 TF DP Peak
Up to 61 Cores
Up to 16GB GDDR5

2016
Knights Landing
Intel® Xeon Phi™ x200 Product Family
14 nm process
Processor & Coprocessor
Over 3 TF DP Peak
Up to 72 Cores
On Package High-Bandwidth Memory
3X Single-Thread
Out-of-order core

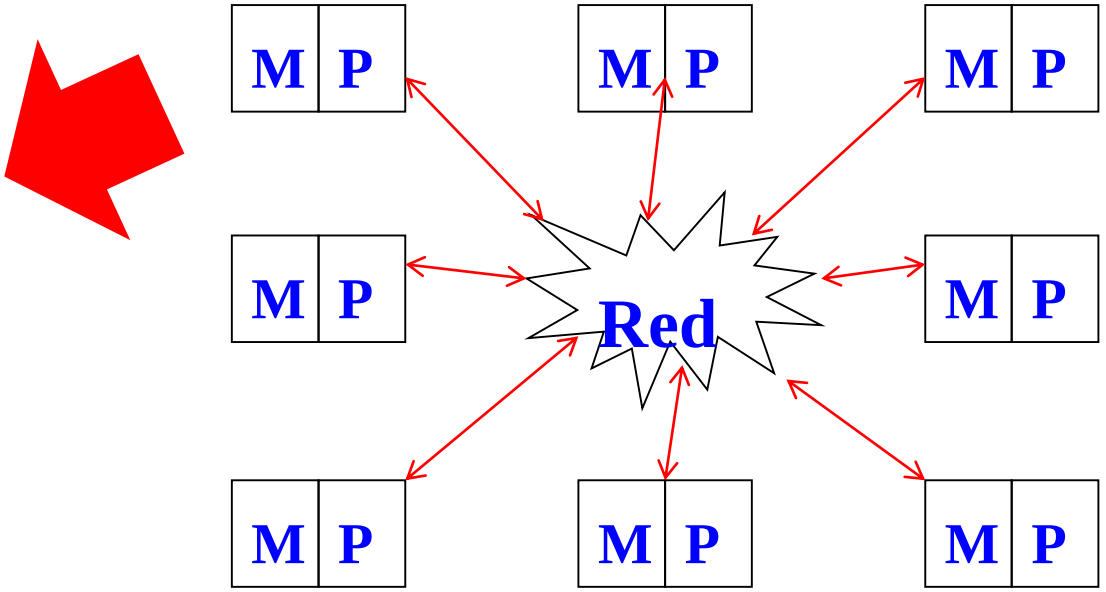
Future
Knights Hill
Next generation of the Intel® MIC Architecture Product Line
10 nm process
2nd Generation Integrated Intel® Omni-Path
... In planning ...

Sistema multicomputador: Red de comunicaciones

Un equipo multiprocesador/multicore con memoria compartida.



Arquitectura Multicomputador



Parámetros de las comunicaciones

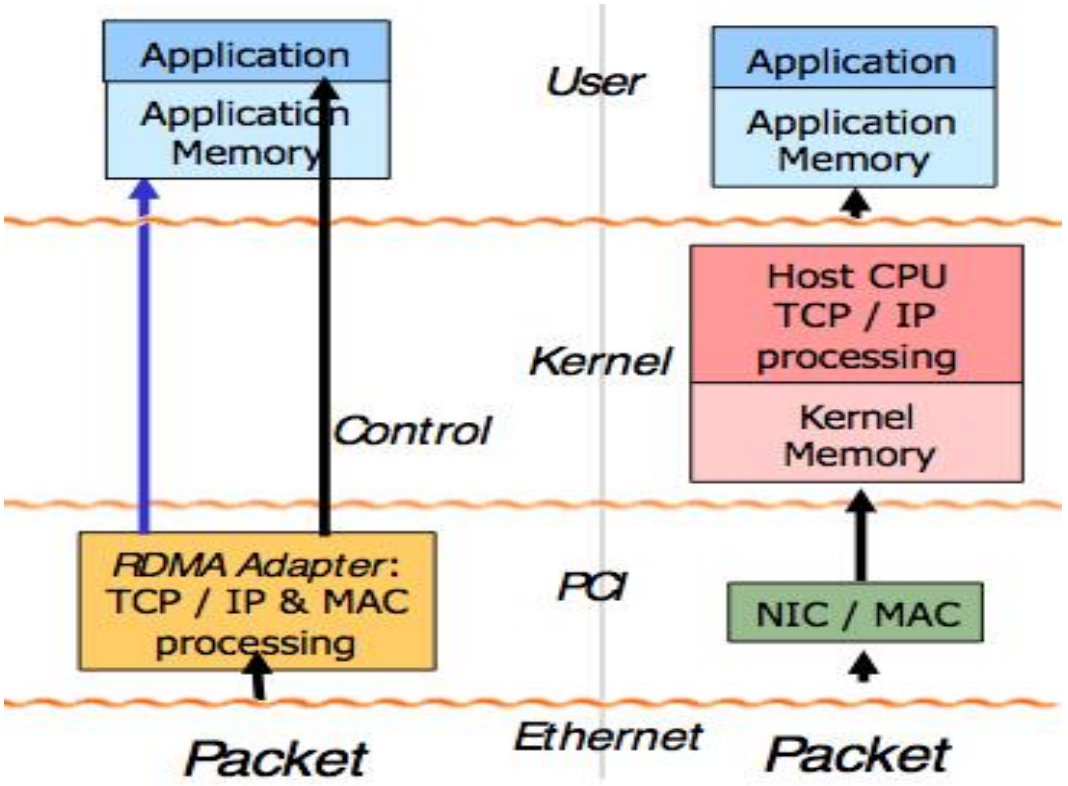
Dos parámetros claves en las redes de comunicación son:

- **Latencia: tiempo requerido para inicializar un mensaje.**
  - **Es un parámetro crítico en el rendimiento si hay mucha comunicación entre los procesadores, intercambiando muchos mensajes de pequeño tamaño.**
- **Ancho de Banda: métrica en Bytes por segundo que indica cuantos datos pueden ser enviados a través de la red de manera sostenida por unidad de tiempo.**
  - **Es un parámetro crítico en intercambio de grandes volúmenes de datos.**

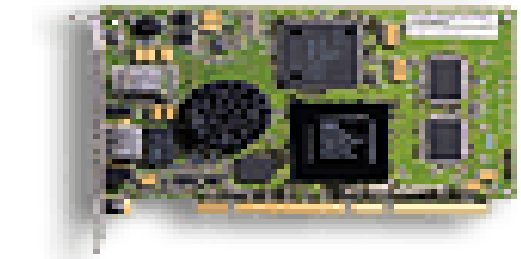
Redes de baja latencia (Low Latency Interconnects)

Objetivo: Disminuir la latencia para un paquete reduciendo el número de copias que se realizan por paquete, en la pila de comunicaciones.

Utilizando **RDMA:** acceso directo a desde la memoria de un nodo de computo a la de otro sin involucrar el sistema operativo.



Tipos de Red de Interconexión



Muchas posibilidades:

ATM, Myrinet, Gigabit Ethernet, Fast Ethernet, Infiniband

Fast Ethernet (para gestión)

- **La red barata más rápida disponible**
- **Ofrece un ancho de banda suficiente para la mayoría de situaciones.**
- **Hasta 100-1000 Mbps**

Gigabit Ethernet:

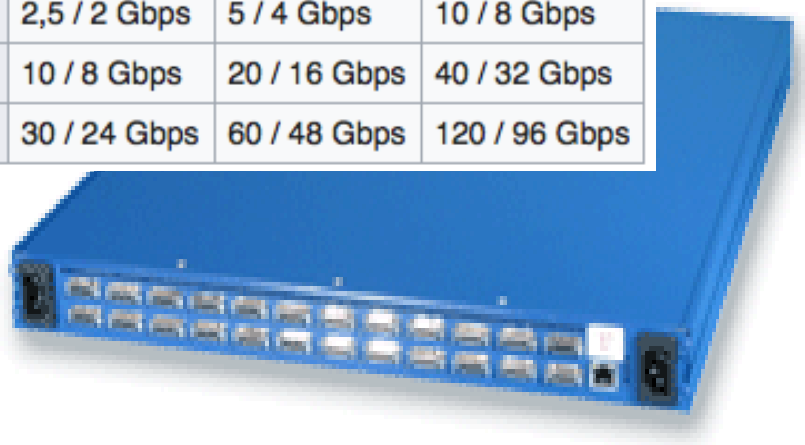
- **Muy rápida (10, 40 y 100 Gbps)**
- **Coste decreciendo rápidamente.**

Infiniband:

- **Muy rápida**
- **baja LATENCIA (<1us)**
- **coste mas alto**

Caudal de Infiniband, bruto / eficaz

	SDR	DDR	QDR
1X	2,5 / 2 Gbps	5 / 4 Gbps	10 / 8 Gbps
4X	10 / 8 Gbps	20 / 16 Gbps	40 / 32 Gbps
12X	30 / 24 Gbps	60 / 48 Gbps	120 / 96 Gbps

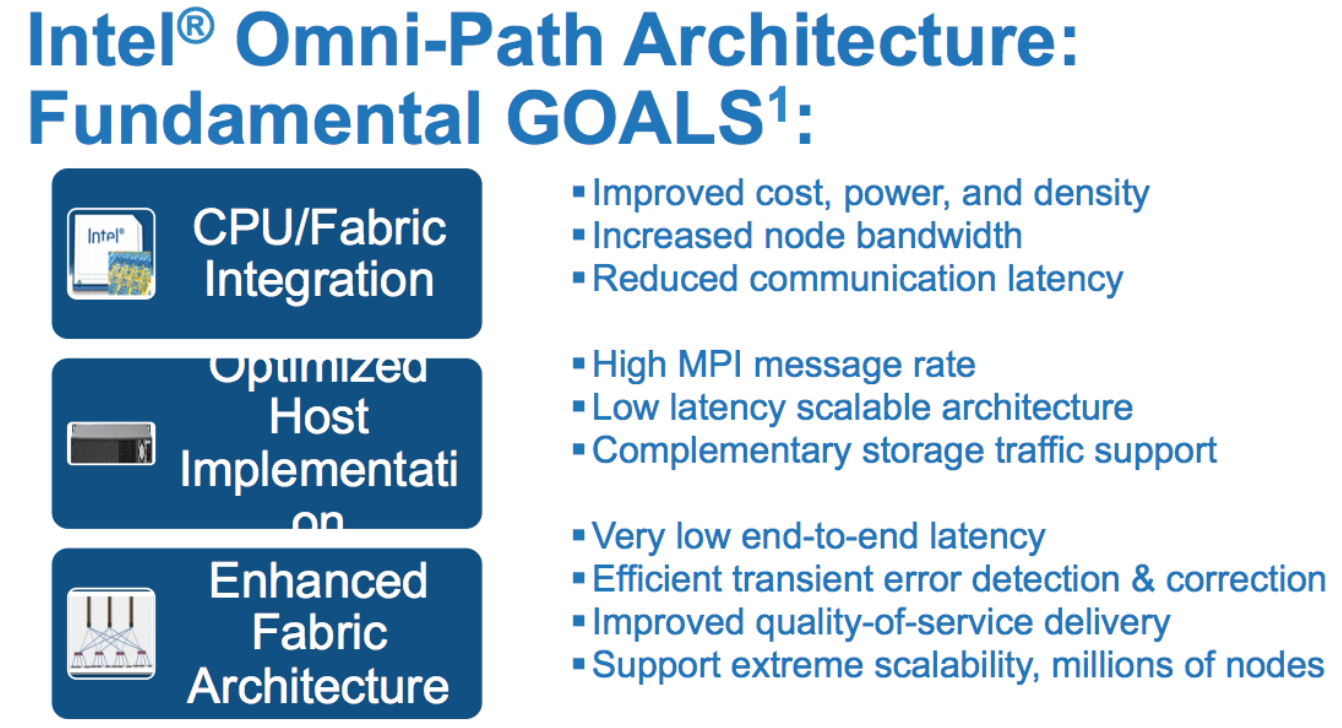


Interconexión: Intel Omni-Path

La arquitectura Intel® Omni-Path introduce interconectividad diseñada para escalar y que alcance las necesidades de los sistemas HPC en la actualidad (ExaScale)

1ª Generación (4Q15)

- **100Gbps links,**
- **160M msg/sec,**
- **Latencia < 110ns**



Equipos actuales: www.top500.org

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	DOE/SC/Oak Ridge National Laboratory United States	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM	2,414,592	148,600.0	200,794.9	10,096
2	DOE/NNSA/LLNL United States	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM / NVIDIA / Mellanox	1,572,480	94,640.0	125,712.0	7,438
3	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93,014.6	125,435.9	15,371
4	National Super Computer Center in Guangzhou China	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 NUDT	4,981,760	61,444.5	100,678.7	18,482
5	Texas Advanced Computing Center/Univ. of Texas United States	**Frontera** - Dell C6420, Xeon Platinum 8280 28C 2.7GHz, Mellanox InfiniBand HDR Dell EMC	448,448	23,516.4	38,745.9	
6	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 Cray/HPE	387,872	21,230.0	27,154.3	2,384

Los sistemas HPC actuales explotan el paralelismo interconectando nodos homogéneos.

En muchos casos la arquitectura del nodo es heterogénea:

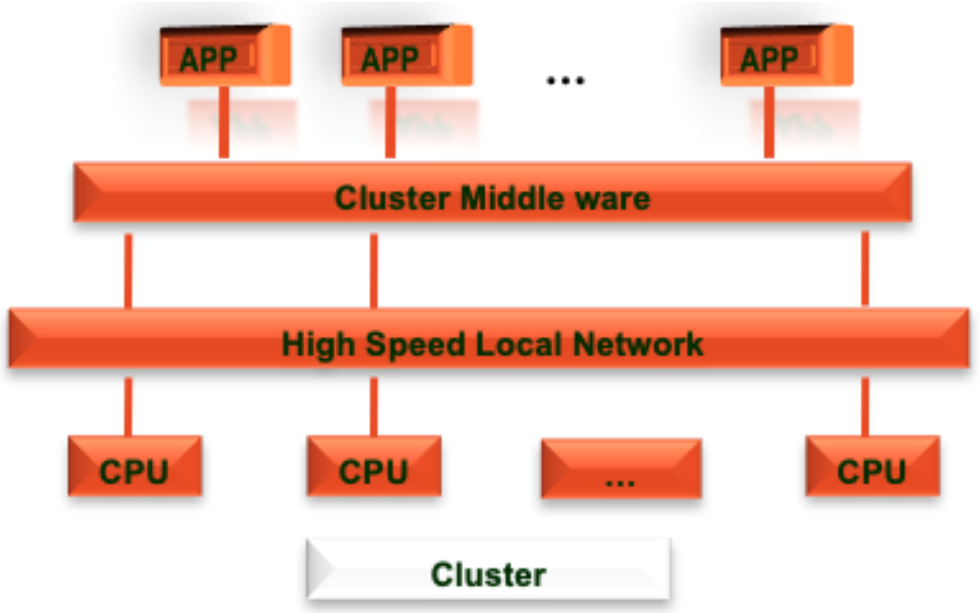
- **Coprocesador GPU**
- **Coprocesador XeonPhi**

Cluster de Ordenadores

Un Cluster es la unión de múltiples ordenadores con una red de alta velocidad para formar un sistema de computación de alto rendimiento, seguro y fiable.

Cluster consta de :

- **Nodos(master+computing)**
- **Network**
- **OS**
- **Cluster middleware**



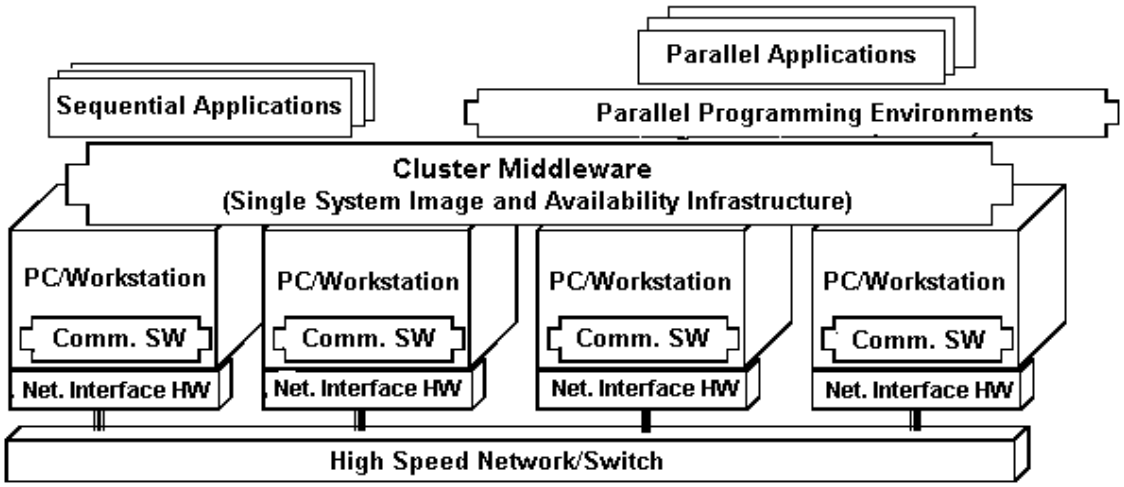
Características de un Cluster

- **Cada máquina en un cluster puede ser un sistema completo utilizable independientemente**
- **Sistema fácilmente escalable**
- **Sistema que se repara de manera muy simple, por sustitución de elementos**

Arquitectura de un Cluster

HW: Interconexión de un frontend con múltiples nodos de computo y recursos de almacenamiento.

SW : Middelware que permita la ejecución distribuida y/o en paralelo de tareas.



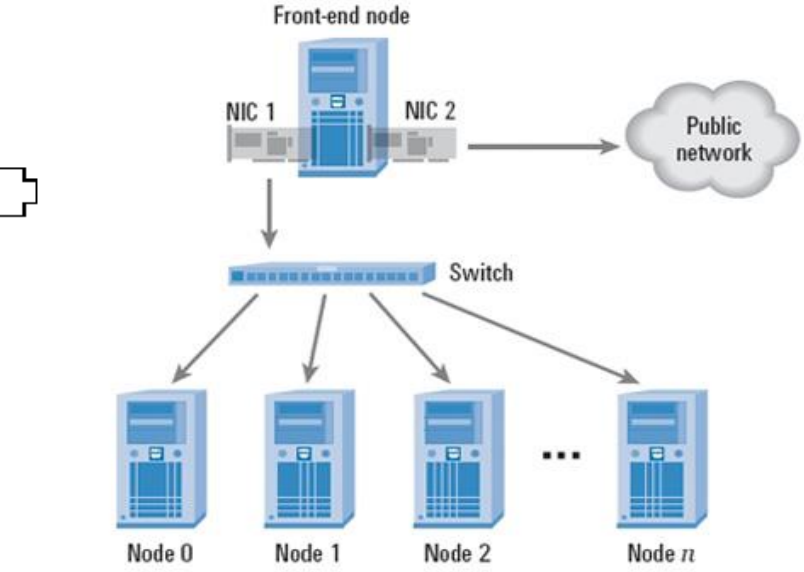


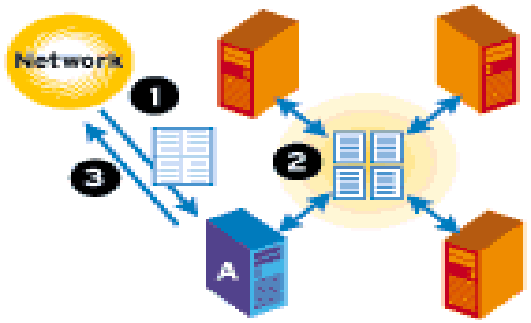
Figure 1 A typical Rocks cluster layout

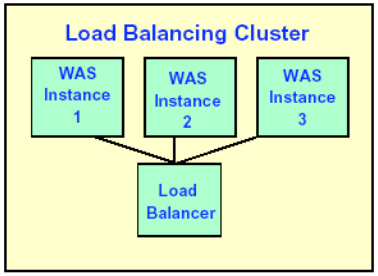
Tipos de cluster por su finalidad

Objetivo 1: resolver grandes desafíos o aplicaciones muy exigentes en recursos

Cluster de alto rendimiento para HPC

- **Enlazar muchos nodos de computación (ordenadores) para conseguir que funcionen en equipo y obtengan la solución de un problema más rápidamente trabajando todos juntos que de manera independientemente.**

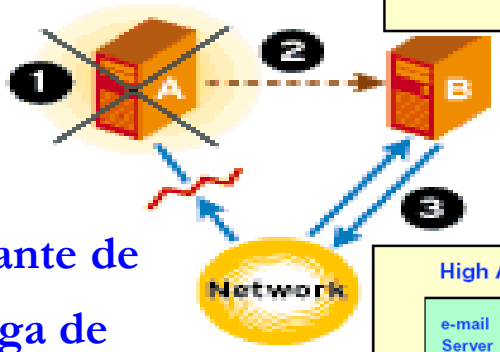


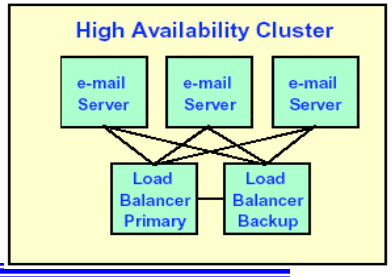


Objetivo 2: que funcionen aplicaciones críticas.

Cluster de alta disponibilidad

- **Conseguir un sistema de ordenadores mas fiable compartiendo trabajos y funcionamiento redundante de tal manera que si un ordenador falla otro se encarga de realizar su trabajo.**

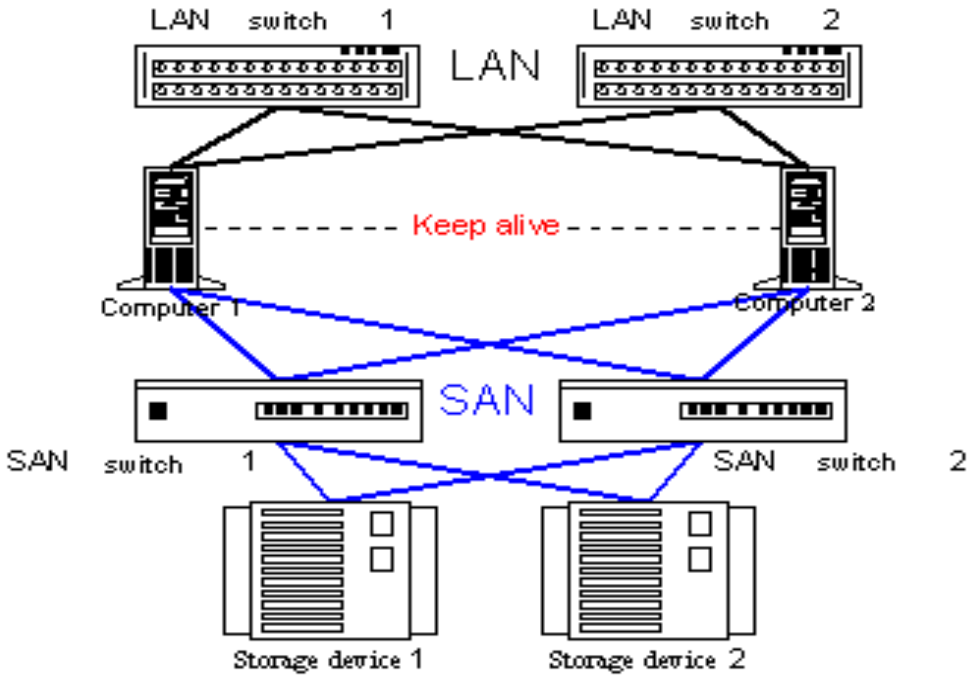




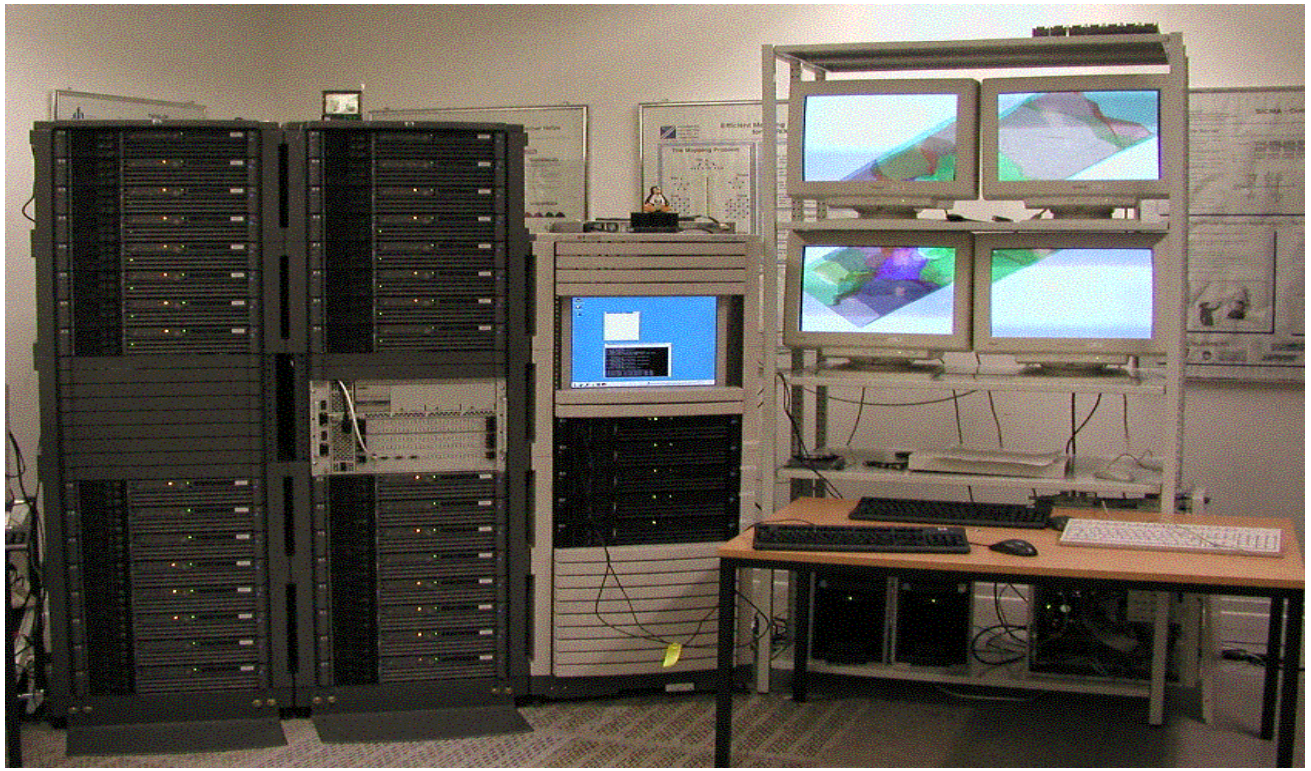
Arquitectura de un cluster de alta disponibilidad

Replicación:

- **Se diseña con componentes redundantes y múltiples caminos de comunicación.**
- **Evitar puntos de fallo único**



Otras arquitecturas: Cluster de Visualización



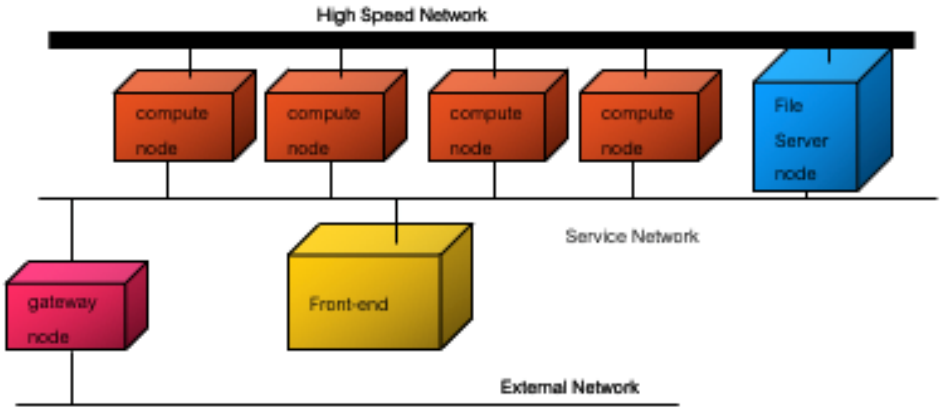
Cada nodo en el cluster gestiona un monitor

Tipos de cluster por su configuración

Cluster cerrado:

Se oculta el cluster detrás de un nodo gateway

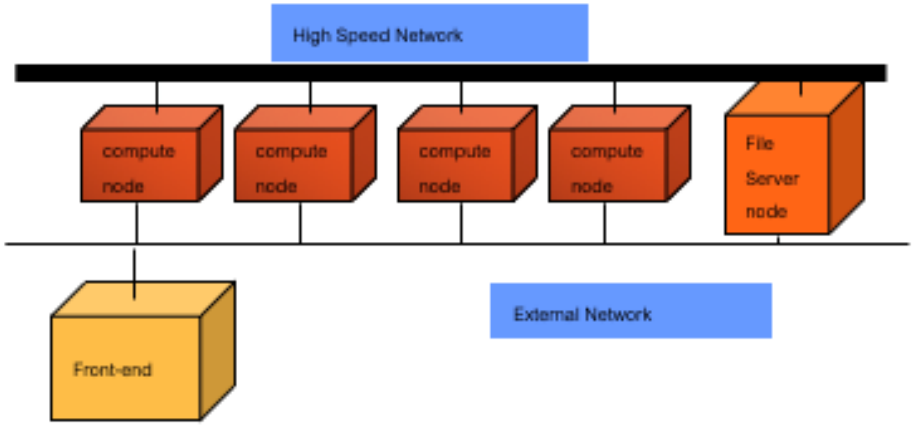
- **Se necesitan menos direcciones IP.**
- **Mas seguridad**
- **Tareas computacionales grandes.**



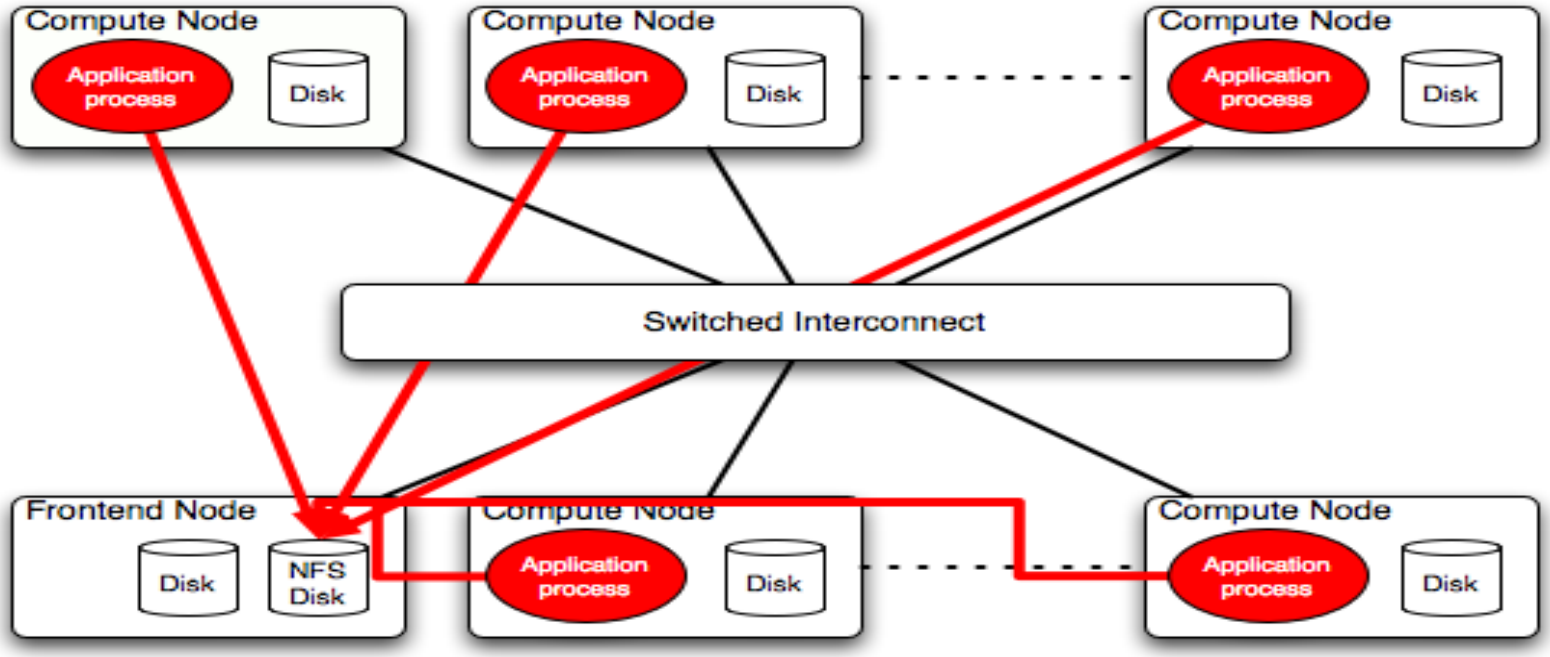
Cluster abierto:

Todos los nodos son visibles desde el exterior

- **Se necesitan muchas direcciones IPs,**
- **Mas difícil de controlar aspectos de seguridad**
- **Más flexible**
- **Tareas como servidor de internet/web**

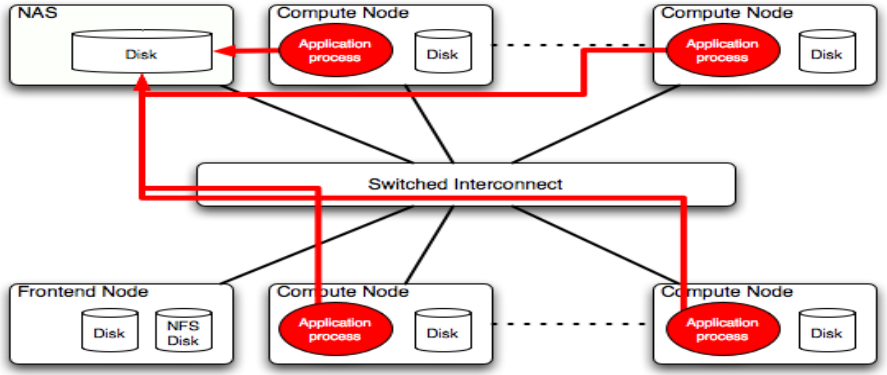


Arquitectura del almacenamiento: Disco NFS Local



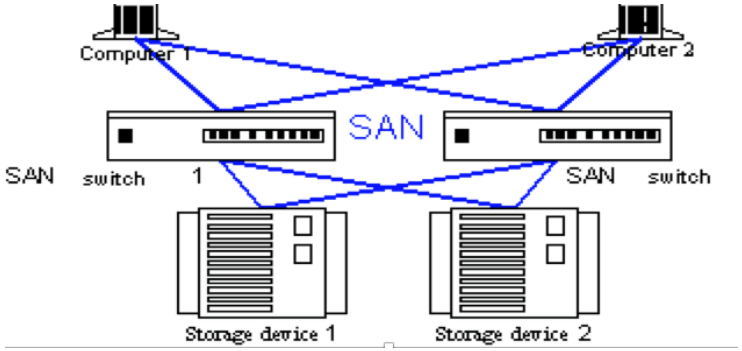
El disco está exportado a los nodos de computo via NFS

NAS: Network Attached Storage



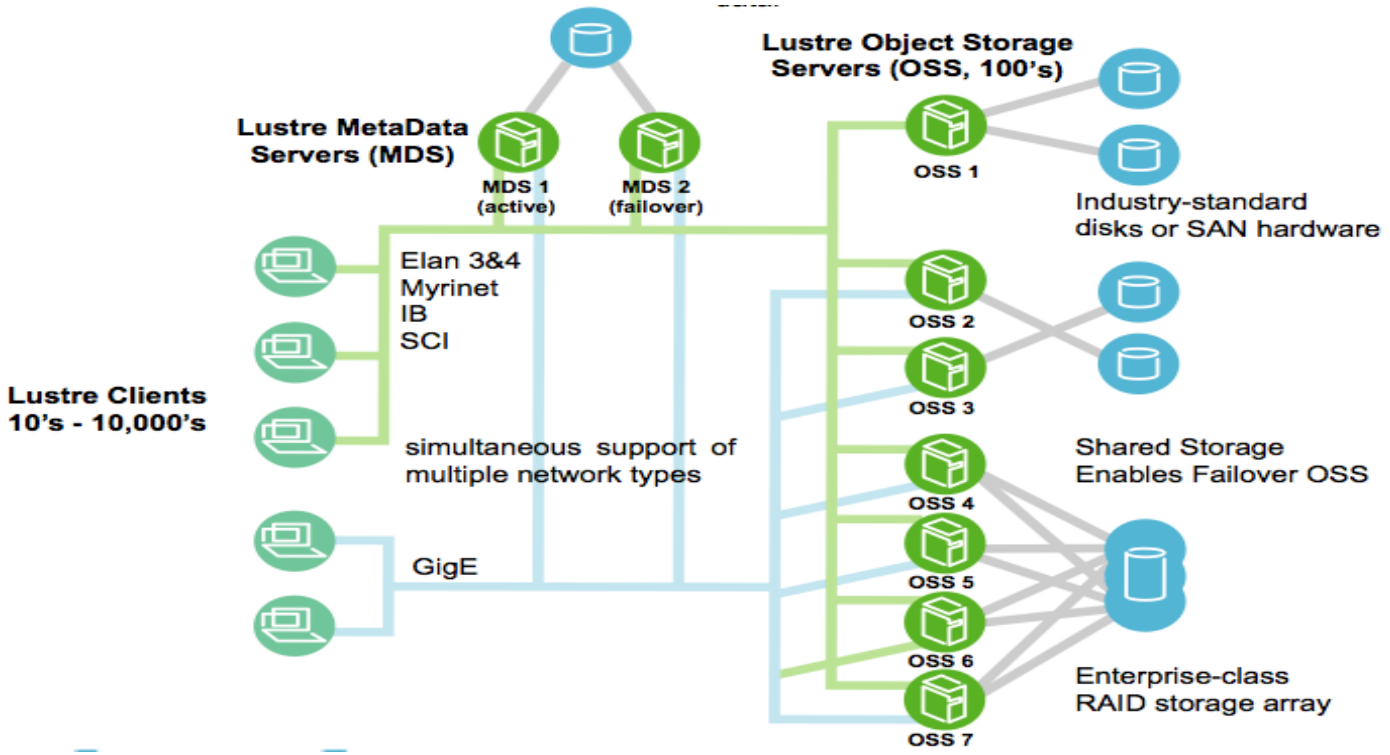
Se puede incorporar como recurso de disco un NAS, que para el cluster es como una caja negra que funciona como un "appliance NFS"

SAN: Storage Area Network



- **SAN: El acceso a disco se sobre una conexión de red especializada (Fibre Channel / Ethernet)**
- **Los discos compartidos están fuera de de los servidores y son un elemento más de la red**
- **Se requiere un servicio central para coordinar las operaciones del sistema de ficheros y suele estar replicado para evitar punto de fallo único.**

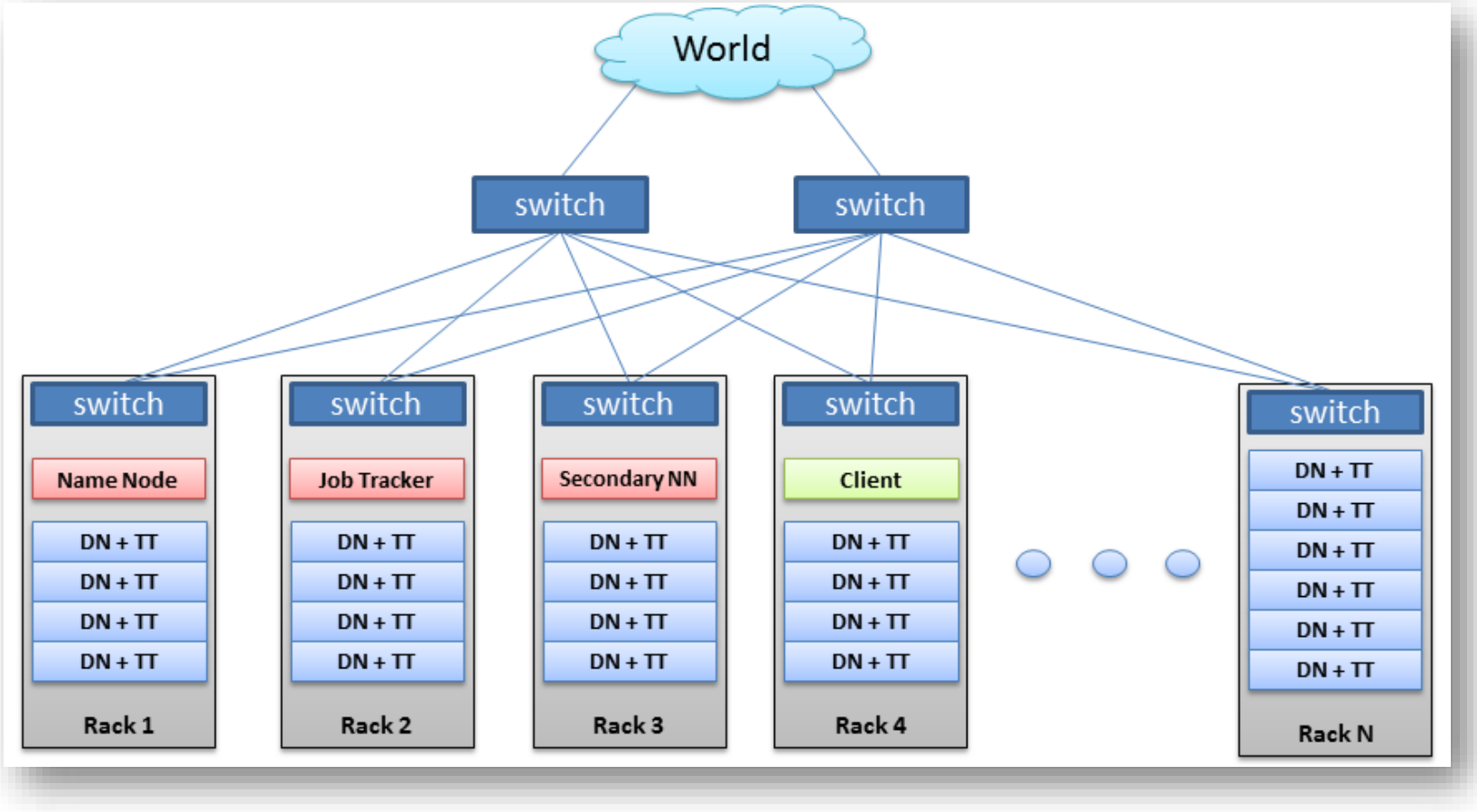
Arquitectura de archivos distribuido: Lustre



Open Source "Object-based" storage

- **Ficheros son objectos, no bloques**

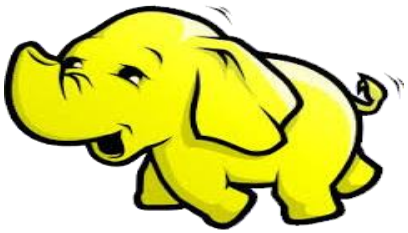
Arquitectura de almacenamiento HDFS

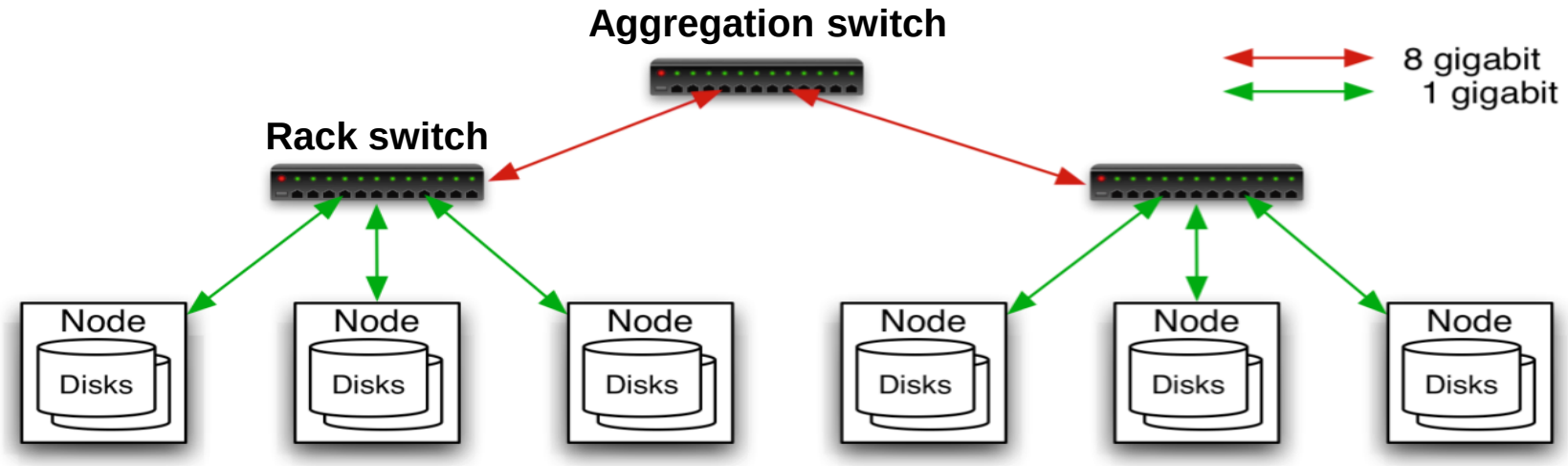


Cluster Hadoop

Cluster creado con commodity Hardware;

- **Nodos inicalmente eran PCs**
- **30-40 nodos/rack**
- **Red a1 gigabit/s en rack**





HDFS : Hadoop Distributed File System

Sistema de Ficheros distribuido muy grande

- **10K nodos, 100 millones de ficheros 10PB**

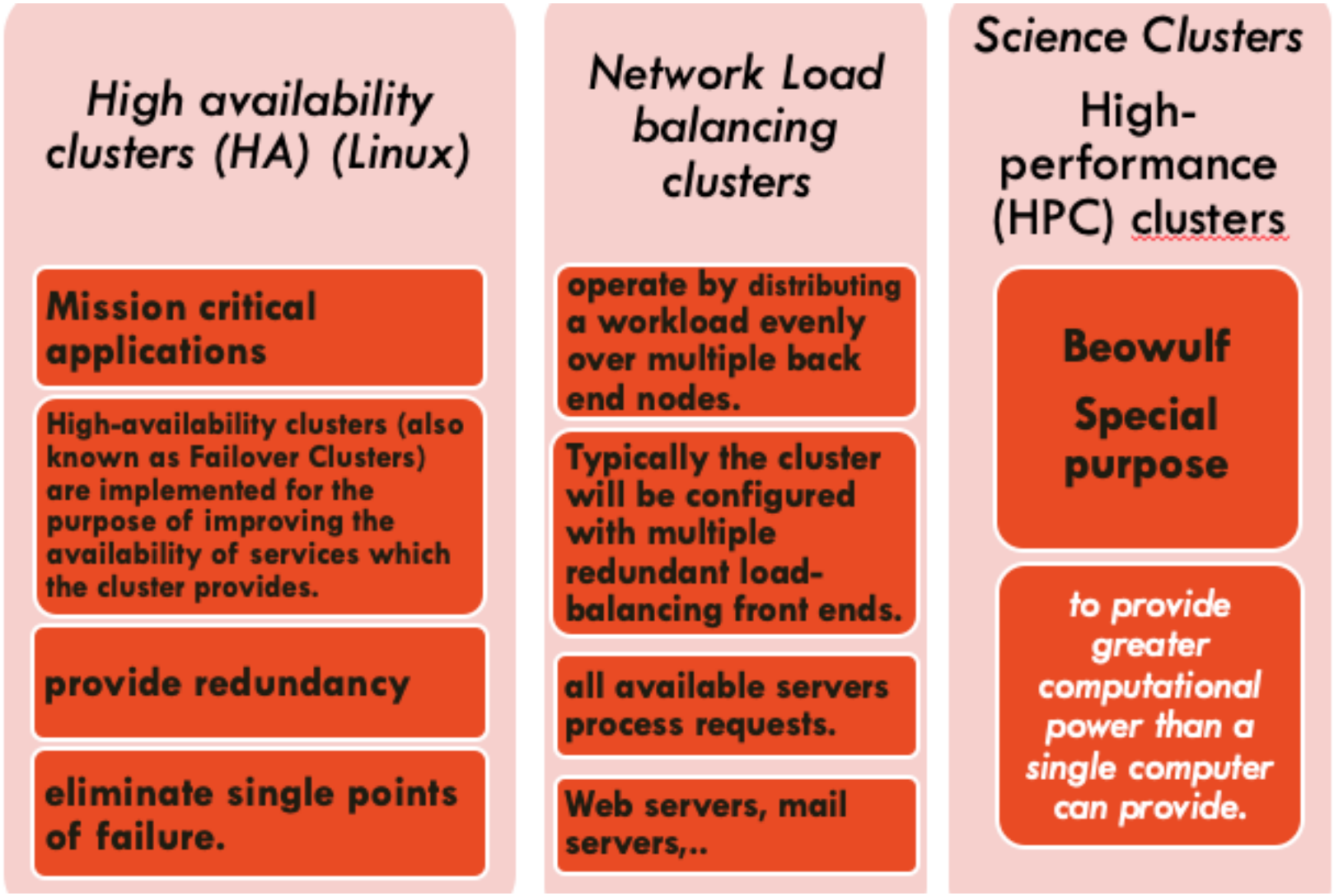
Realizado con "*Commodity Hardware*"

- **Ficheros replicados para tolerancia a fallos**
- **Detecta fallos y recupera los datos.**

Optimizado para proceso por lotes ("*Batch Processing*").

- **Expone la localización de los datos y así permite que la computación se pueda llevar cerca de los datos.**
- **El ancho de banda agregado es muy alto.**

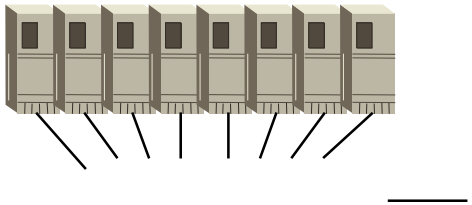
Ámbitos de aplicación de los cluster



Cluster Beowulf

Un Cluster Beowulf es un Supercomputador paralelo construido con hardware comercial de fácil adquisición, S.O. Linux y software open source.

Características de un cluster Beowulf



- **Red interna de alta velocidad**
- **Hardware de fácil adquisición**
- **software y Sistema Operativo Open source**
- **Soporta programación paralela tal como MPI, PVM**
- **Software que permite el uso de un conjunto de nodos como una única máquina**

¿Por qué se denomina a esta clase de máquina un Cluster Beowulf?

El proyecto Beowulf

Proyecto en el *Center of Excellence and Information Systems Sciences*(CESDIS) de *NASA Goddard Space Center*,

Dr. Thomas Sterling, Donald Becker que lo definen

"Beowulf is a project to produce the software for off-the-shelf clustered workstations bas commodity PC-class hardware, a high-bandwidth internal network, and the Linux operati system." **http://www.beowulf.org/**

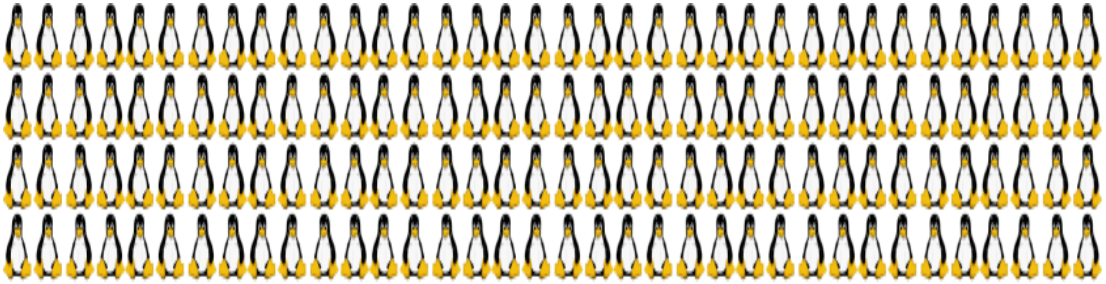
Cluster Beowulf: requisitos y objetivos

Requisitos de los componentes de cluster Beowulf

1. **Ordenadores: PCs, Estaciones de trabajo, plataformas multiprocesador...**
   - Independencia del fabricante y bajo coste
2. **Sistema Operativo que admita funcionamiento en cluster**
   - Libre distribución y Open source
3. **Red de interconexión**
   - Rapidez
   - Bajo coste

Beneficios potenciales:

- **Rendimiento**
- **Disponibilidad**
- **Balanceo de carga**
- **Escalabilidad**
- **Fiabilidad y Seguridad**



Rocks

Rocks : Objetivo principal

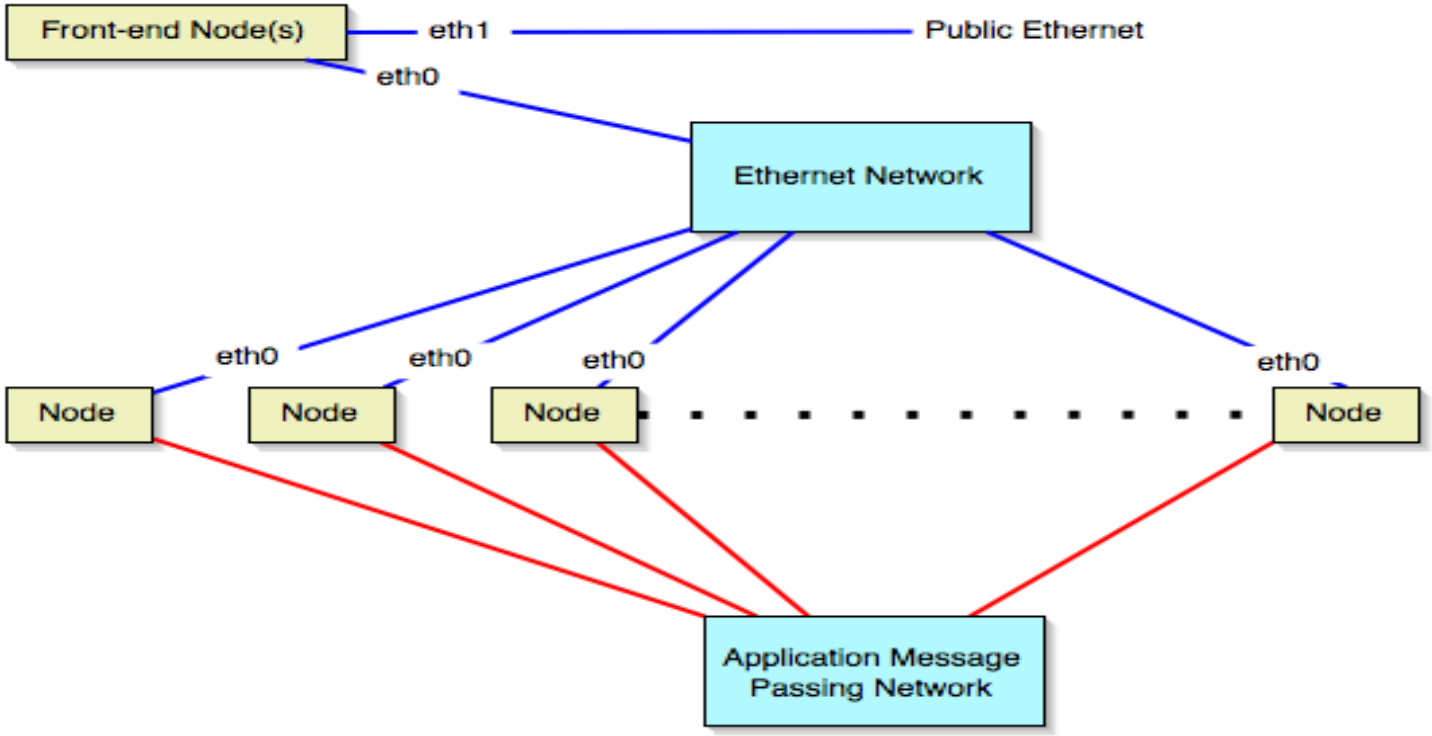
Facilitar la instalación

de un cluster.

Make clusters easy !

Audiencia: Científicos que desean tener un alta computación en su propio laboratorio.

Arquitectura de un cluster Rocks para HPC (High Performance Cluster)



Muchos nodos de computación conectados con red de alto rendimiento

Proceso de Instalación

Instalar a frontend

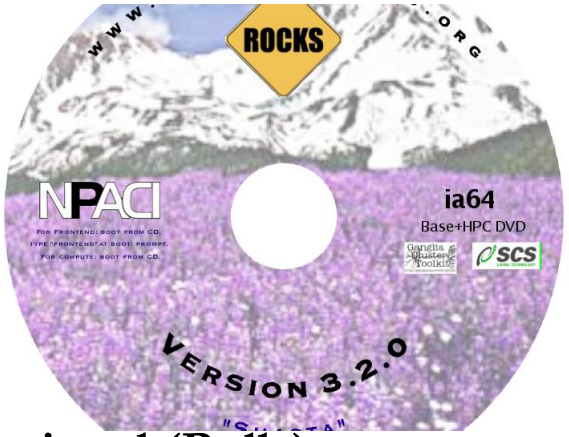
1. Insert Rocks Base CD
2. Insert Roll CDs (optional components)
3. Answer 7 screens of configuration data
4. Drink coffee (takes about 30 minutes to install)

Instalar los nodos de computo:

1. Login to frontend
2. Execute insert-ethers
3. Boot compute node with Rocks Base CD (or PXE)
4. Insert-ethers discovers nodes
5. Goto step 3

Añadir usuarios

Empezar a utilizarlo

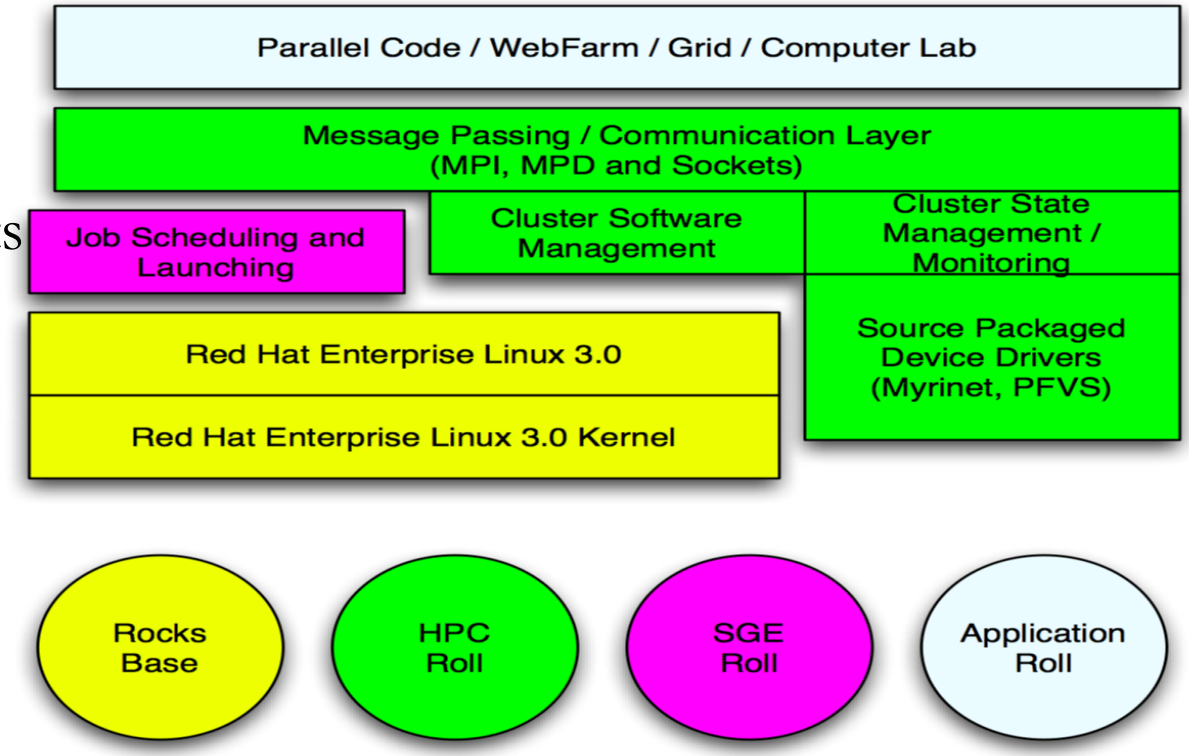


Opcional (Rolls)

- **Condor**
- **Grid (based on NMI R4)**
- **Intel (compilers)**
- **Java**
- **SCE (developed in Thailand)**
- **Sun Grid Engine**
- **PBS (developed in Norway)**
- **Area51 (security monitoring tools)**

Rocks & 'Rolls'

Rolls son contenedores de paquetes software y los scripts de configuración asociados.



Arquitectura software de un cluster HPC

Componentes software

Sistema Operativo

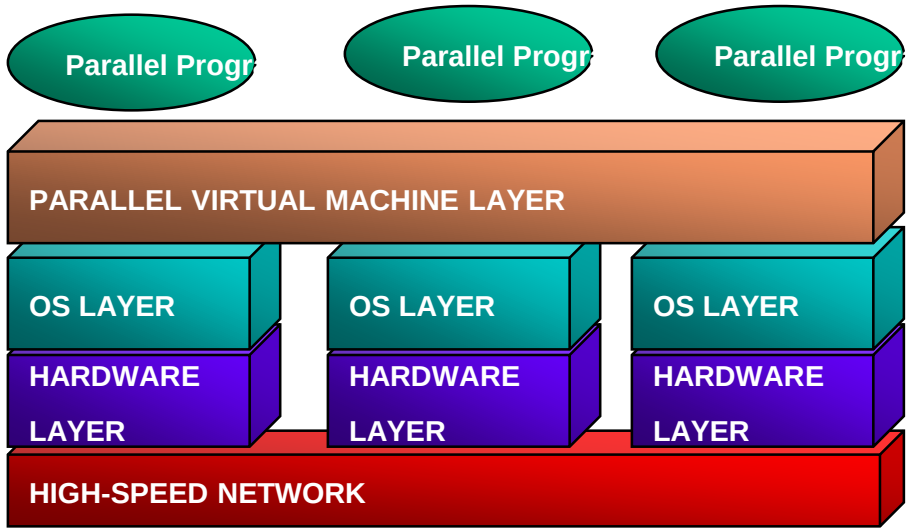
- **LINUX**

Sistema de programación paralelo

- **PVM, MPI**

Utilidades, Librerías,

Herramientas de dominio público

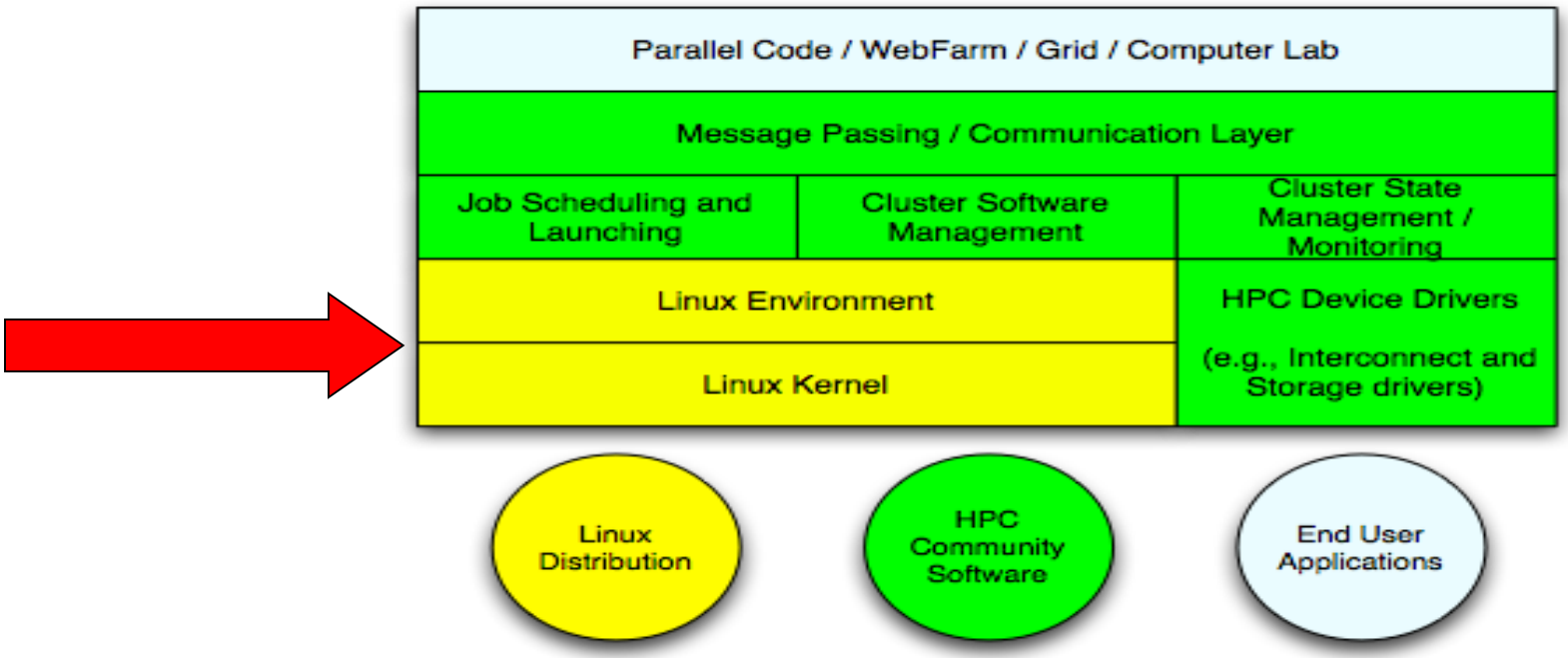


¡Todos los componentes software son GRATIS!

Cluster Software Stack

Linux Kernel/Environment

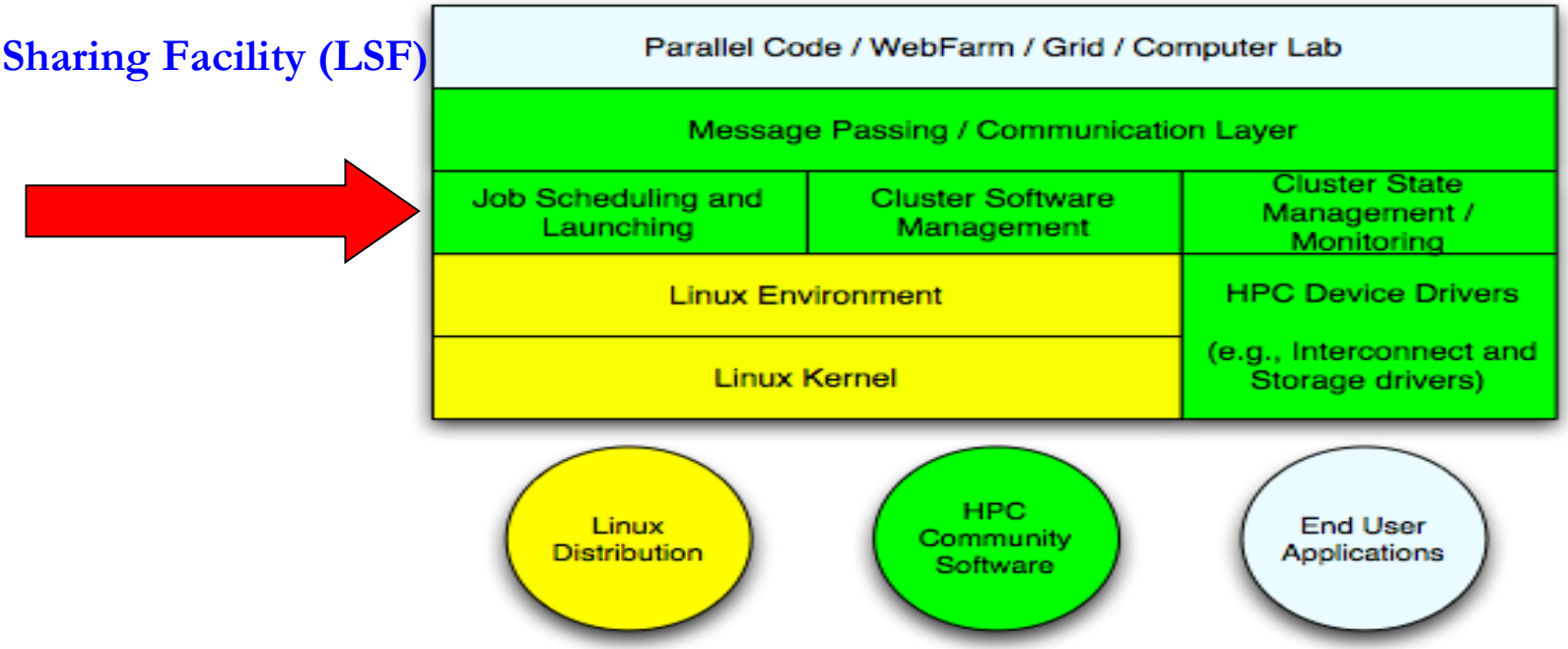
- **RedHat, SuSE, Debian, etc.**



Cluster Software Stack

Job Scheduling and Launching

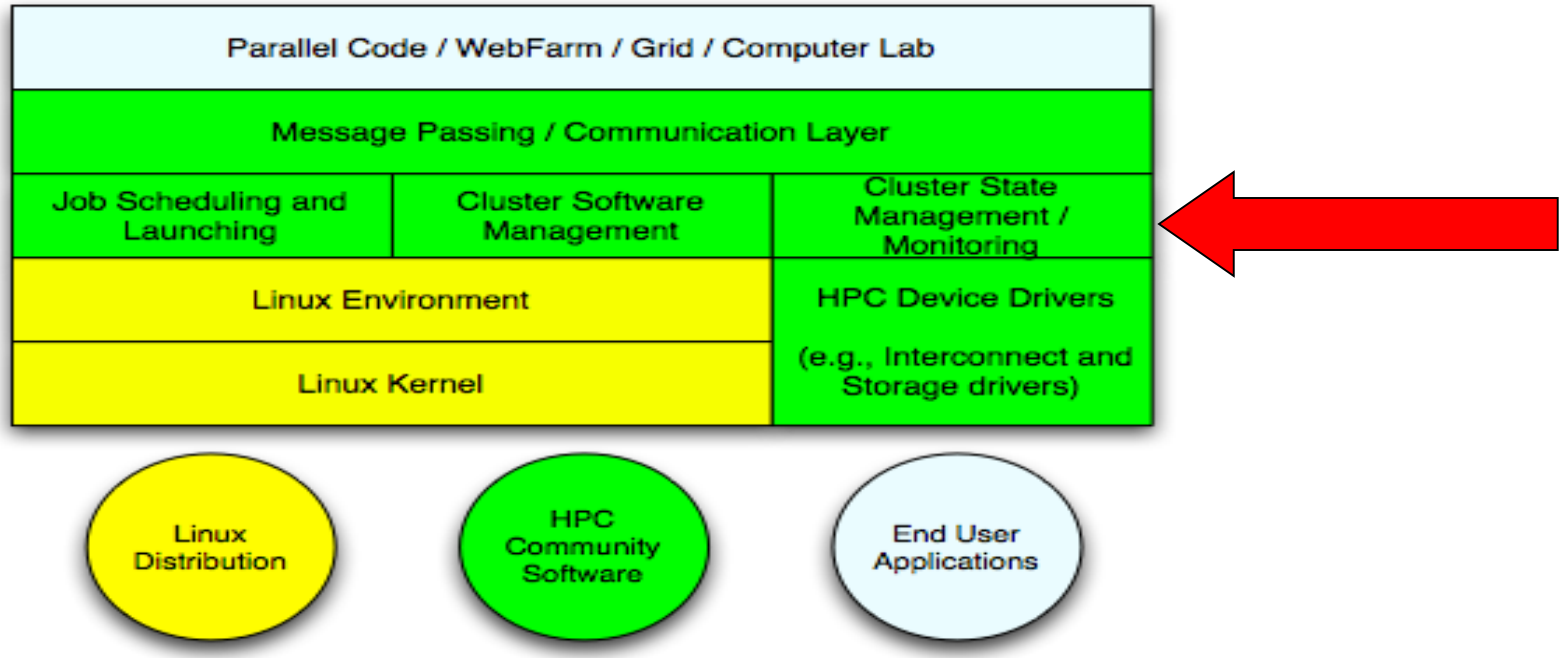
- **Sun Grid Engine (SGE)**
- **Portable Batch System (PBS)**
- **Load Sharing Facility (LSF)**



Cluster Software Stack

Cluster State Management and Monitoring

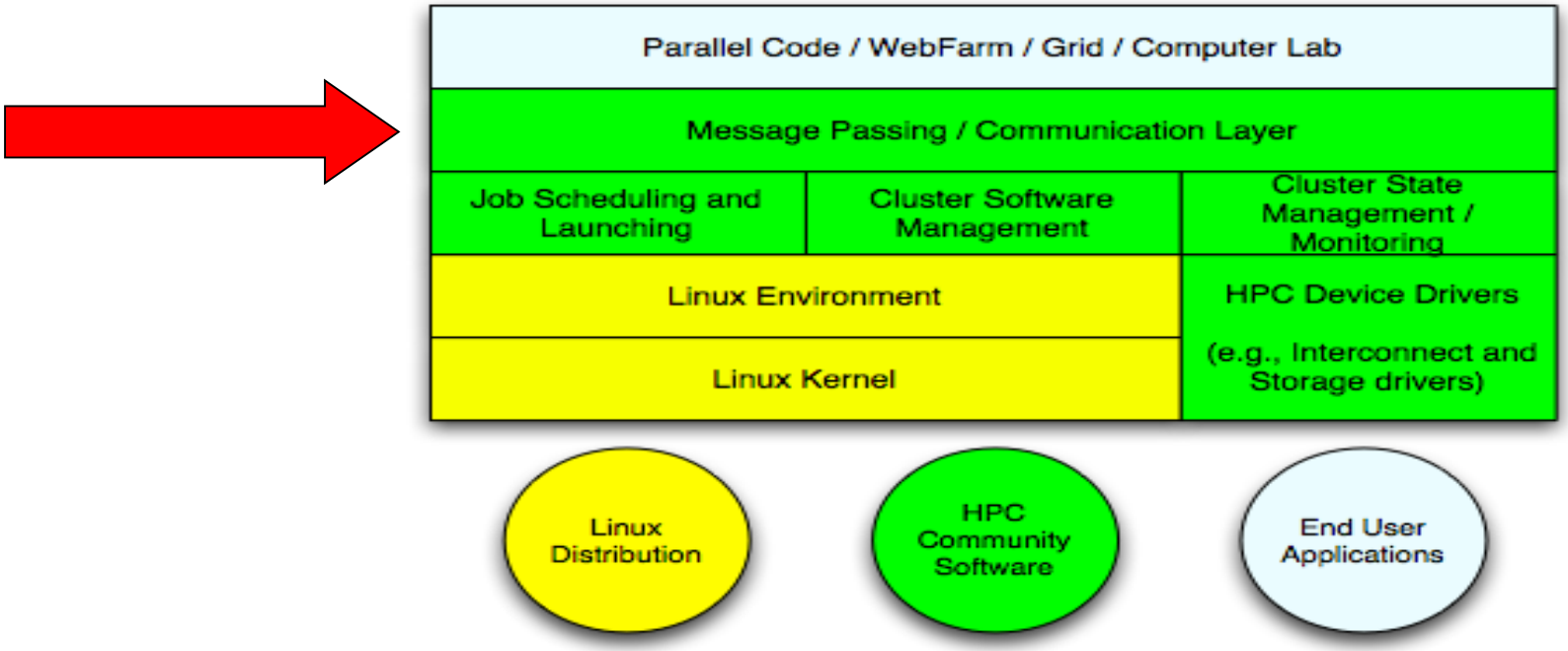
- **Monitoring: Ganglia, Clumon, Nagios, Tripwire, Big Brother**
- **Management: Node naming and configuration (e.g., DHCP)**



Cluster Software Stack

Message Passing and Communication Layer

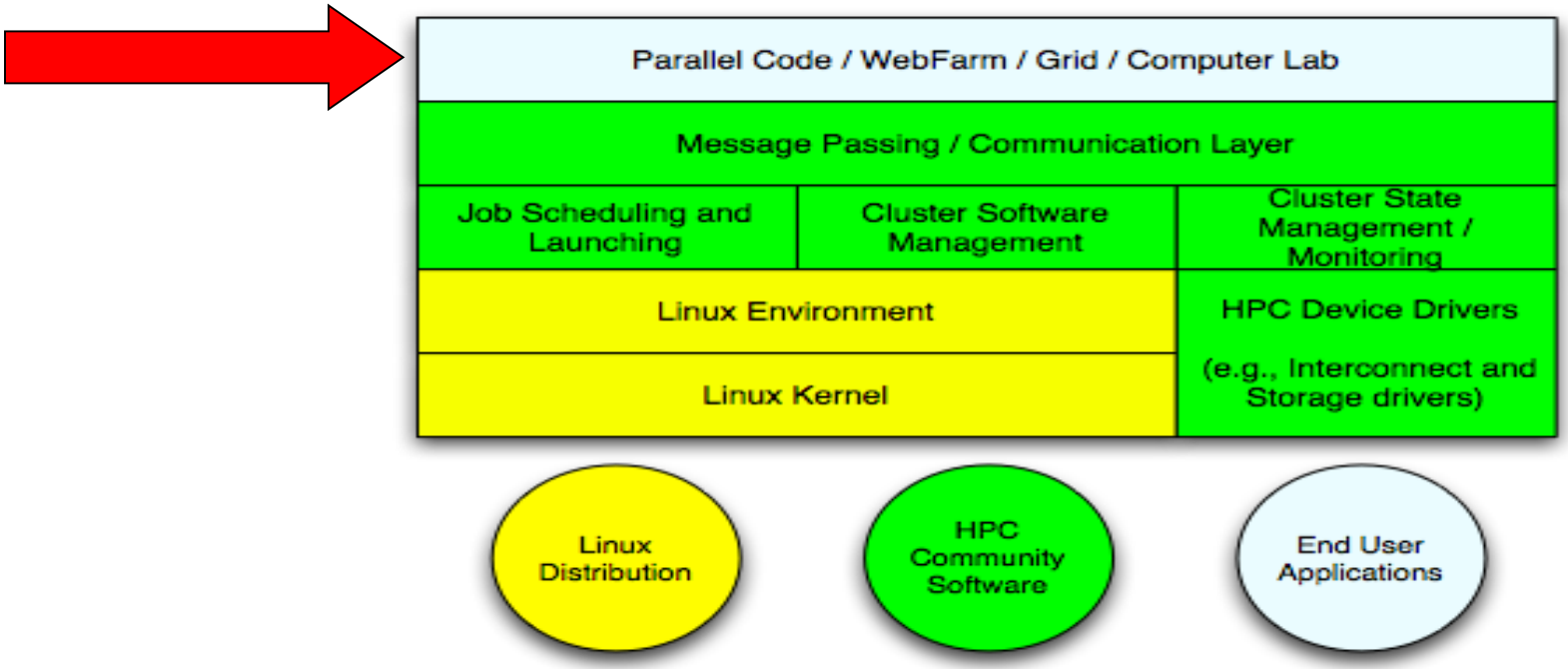
- **E.g., Sockets, MPICH, PVM**



Cluster Software Stack

Parallel Code / Web Farm / Grid / Computer Lab

- **Locally developed code**



Bibliografía y enlaces de interés

Bibliografía

- **High Performance Cluster Computing: Architectures and Systems by Rajkumar Buyya Ed Prentice Hall PTR; ISBN: 0130137847; 1st edition (1999)**
- **Linux Cluster Architecture by Alex VreniosSams;**

  ISBN: 0672323680; (2002)
- **Linux Clustering: Building and Maintaining Linux Clusters by Charles Bookman**

  Ed:New Riders Publishing; ISBN: 1578702747; 1st edition (2002)

Enlaces

- http://www.hispacluster.org/
- http://www.openclustergroup.org/
- http://www.beowulf.org/ y http://www.beowulf-underground.org/
- http://www.mosix.com/