Inspira Crea Transforma

UNIVERSIDAD EAFIT®

Introducción: Sistema de Gestión de Recursos en un Supercomputador

Mateo Gómez Zuluaga

Centro de Computación Científica APOLO
Ciclo de conferencias APOLO

Supercomputador: Cronos

Especificaciones Generales:

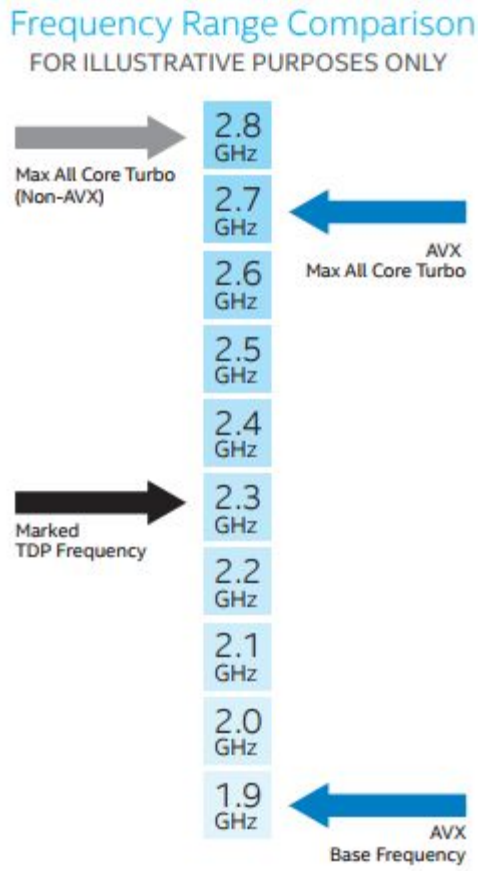
- **Número de servidores:** 40 (HP ProLiant SL230s Gen8)
- **Número de chasis:** 5 (HP ProLiant S6500)
- **Red de alta velocidad:** Infiniband FDR (56 Gbps)
- **Número total de procesadores:** 640
- **Número total de memoria RAM:** 2.56 TB
- **Almacenamiento (Home):** 32 TB (En proceso)
- **Almacenamiento (Scratch):** 6.4 TB (En proceso)
- **TFLOPS:** 12.5 (Doble Precisión)

Supercomputador: Cronos

Especificaciones - HP ProLiant SL230s Gen8:

- **Número total de núcleos:** 16 (2 ->8) (HT*)
- **Frecuencia base:** 2.6 GHz
- **Memoria RAM:** 64 GB DDR3 (1333 MHz)
- **Disco Duro HDD:** 250 GB (7.200 rpm)

Tomado de: Server Supply - 650048-B21



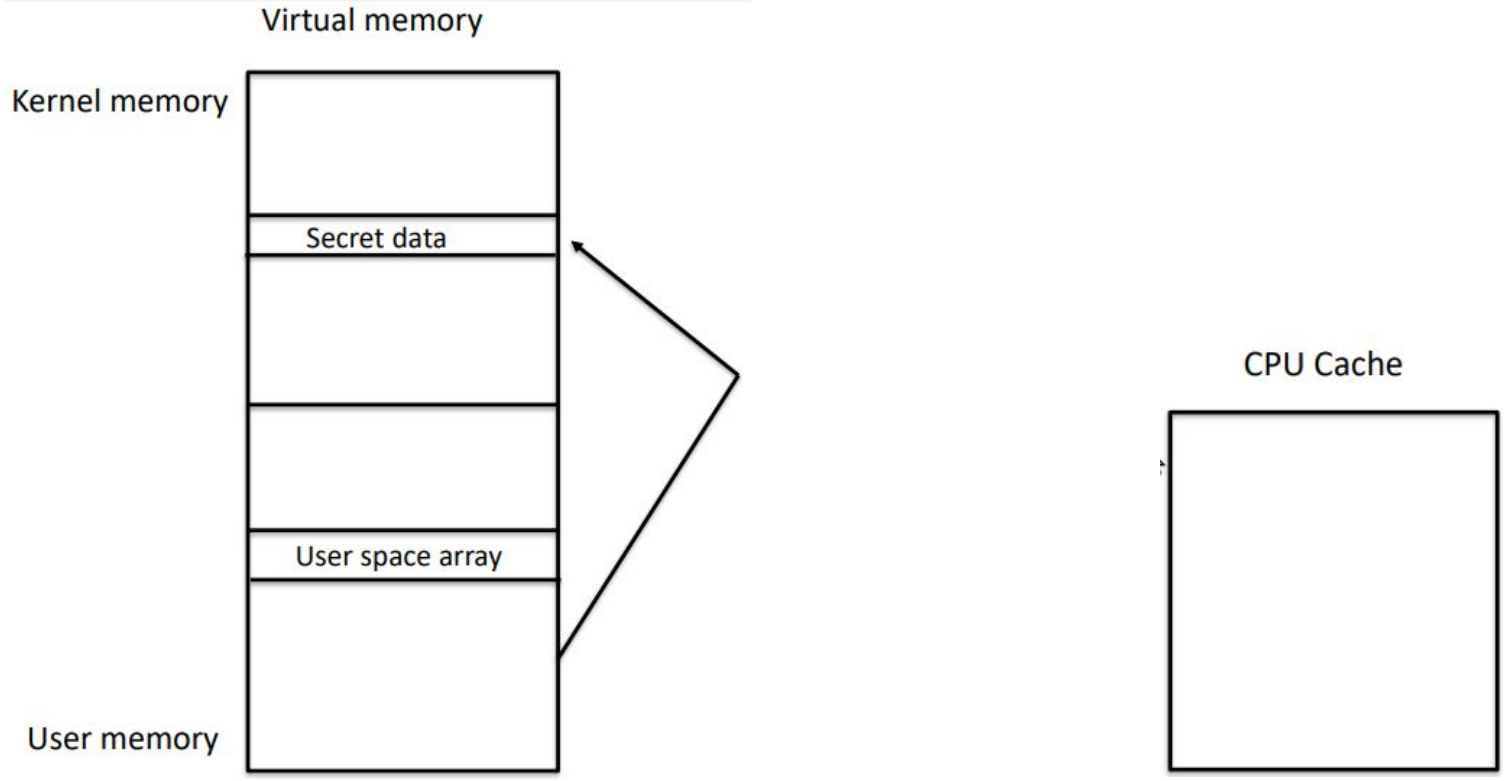
Tomado de: White paper - Optimizing performance with Intel Advanced Vector Extresnions

Actualidad: Spectre - Meltdown

	Meltdown	Spectre
Allows kernel memory read	Yes	No
Was patched with KAISER/KPTI	Yes	No
Leaks arbitrary user memory	Yes	Yes
Could be executed remotely	Sometimes	Definitely
Most likely to impact	Kernel integrity	Browser memory
Practical attacks against	Intel	Intel, AMD, ARM

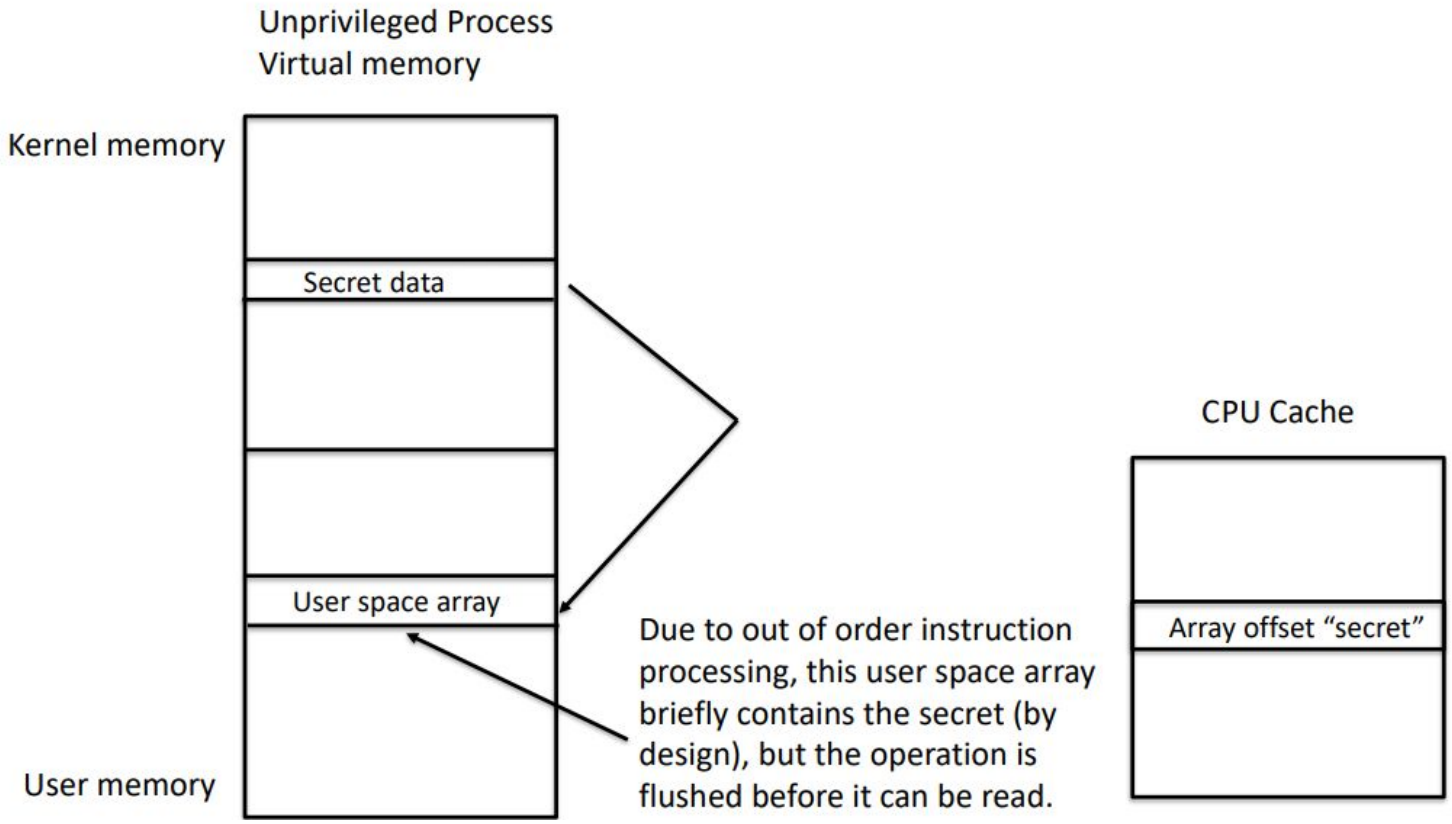
Fuente: SANS / Meltdown and Spectre - understanding and mitigating the threats

Meltdown



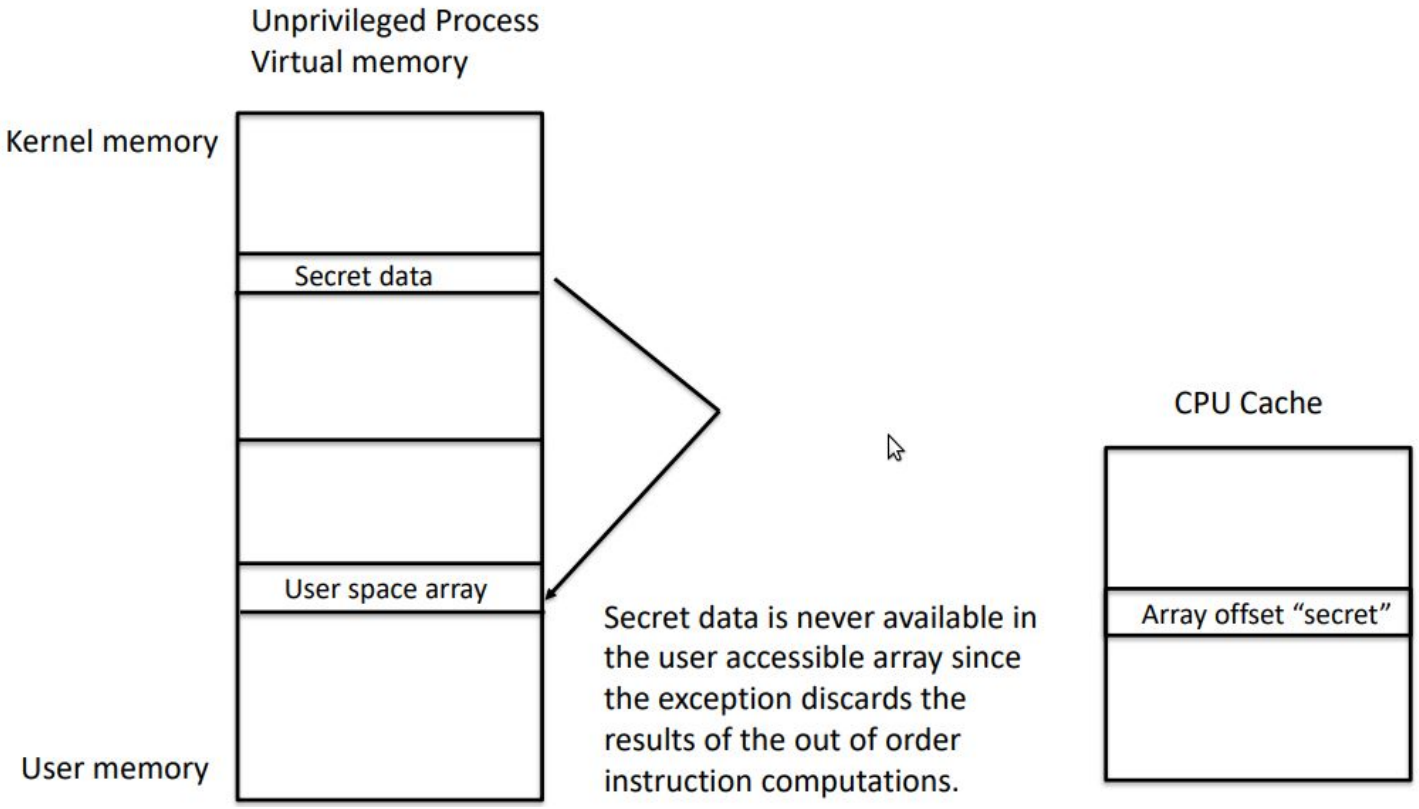
Fuente: SANS / Meltdown and Spectre - understanding and mitigating the threats

Meltdown



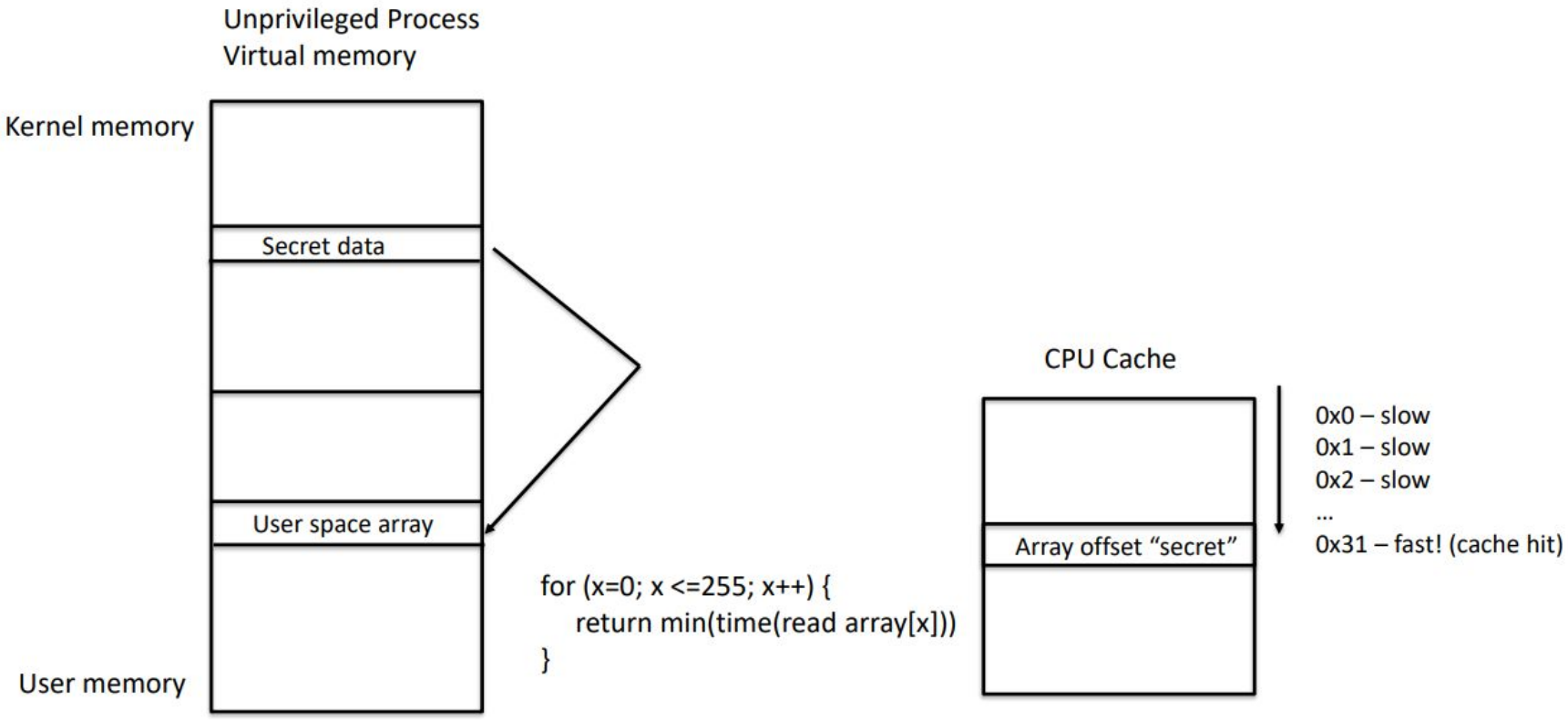
Fuente: SANS / Meltdown and Spectre - understanding and mitigating the threats

Meltdown



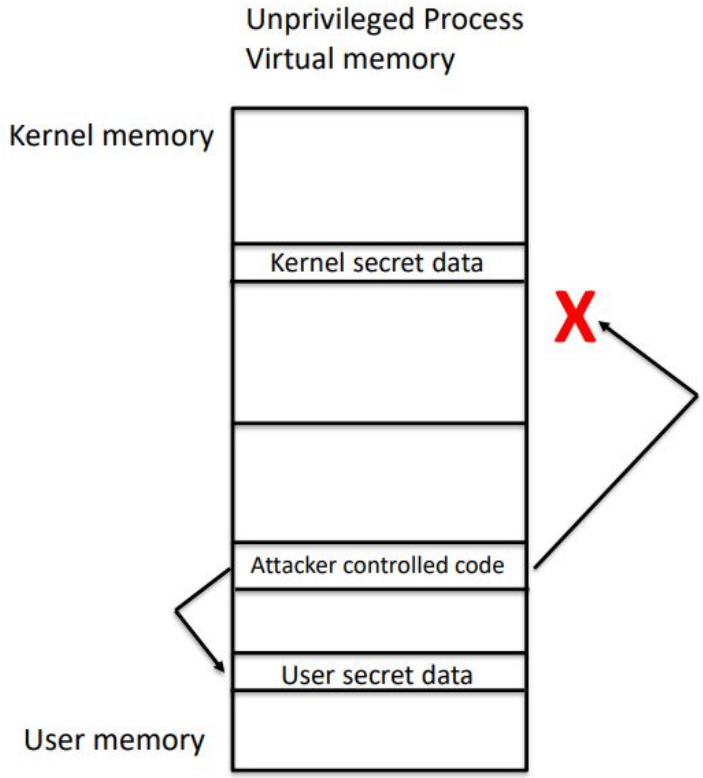
Fuente: SANS / Meltdown and Spectre - understanding and mitigating the threats

Meltdown



Fuente: SANS / Meltdown and Spectre - understanding and mitigating the threats

Spectre



Fuente: SANS / Meltdown and Spectre - understanding and mitigating the threats

Rendimiento vs. Seguridad

TABLE I
CHANGE IN WALLTIME UPON PATCH APPLICATION.

Application	Number of Nodes	Difference, %[1]	Are the means different?[2]	Before Patch Application			After Patch Application		
				Mean, Seconds	Standard Deviation, Seconds	Number of Runs	Mean, Seconds	Standard Deviation, Seconds	Number of Runs
NAMD	1	3.3	Y	306.6	1.44	24	316.9	3.05	56
NAMD	2	6.9	Y	175.4	2.78	22	188.1	3.49	56
NWChem	1	2.6	Y	77.8	1.91	23	79.9	1.11	59
NWChem	2	10.7	Y	58.4	1.05	21	65.0	4.16	56
HPCC	1	2.2	Y	304.1	6.39	23	310.9	4.88	56
HPCC	2	5.3	Y	345.1	5.41	22	364.0	8.44	56
IMB	2	4	Y	14.8	0.54	21	15.4	1.39	56
IOR	1	3.9	Y	188.5	9.41	21	195.9	11.69	55
IOR	2	1.5	N	371.1	12.23	22	376.7	19.50	56
IOR.local	1	2.1	N	462.8	16.37	12	472.8	19.03	56
MDTest	1	21.5	Y	30.5	3.17	21	37.8	4.10	56
MDTest	2	9.3	Y	166.7	3.60	23	182.8	5.30	55
MDTest.local	1	56.4	Y	3.8	0.62	12	6.7	2.61	56

Fuente: Researchers Measure Impact of ‘Meltdown’ and ‘Spectre’ Patches on HPC Workloads

Arquitectura de un supercomputador

Roles principales:

- Servidor maestro u orquestador
- Nodo de cómputo
  - Gran cantidad de memoria
  - Gran cantidad de procesadores
- Nodo de coprocesamiento
- Servidor NAS (Intermediario)
- Sistema Almacenamiento
  - Alta velocidad
  - Gran capacidad

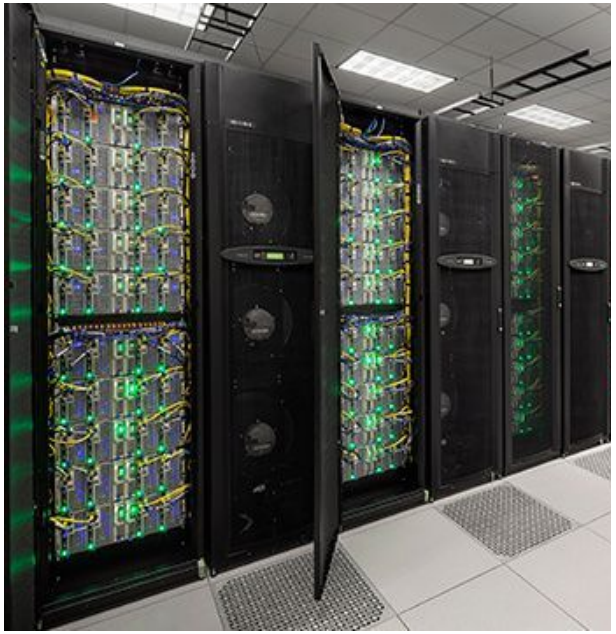


Foto: Centro Forschungszentrum Alemania. IBM BlueGene/Q Power BQC. 5 petaFLOPS

Arquitectura de un supercomputador

Recursos computacionales:

- Procesamiento
  - CPU (Núcleos*, frecuencia, caché, instrucciones, etc.)
  - GPU (Memoria, núcleos*, frecuencia, etc.)
- Almacenamiento compartido.
  - (red interna, red externa, etc.)
- Almacenamiento local.
- Memoria RAM.
  - (Capacidad, frecuencia, tecnología*, etc.)
- Red de interconexión
  - Alta velocidad (Infiniband, Ethernet, Myrinet, etc.)
  - Monitoreo, aprovisionamiento (Ethernet)

Arquitectura de un supercomputador

- **Multiusuario**
  - Diferente software
  - Diferente uso
  - Diferente comportamiento

- **Tipos de trabajos**
  - Seriales
  - Paralelos
  - Distribuidos

Arquitectura de un supercomputador

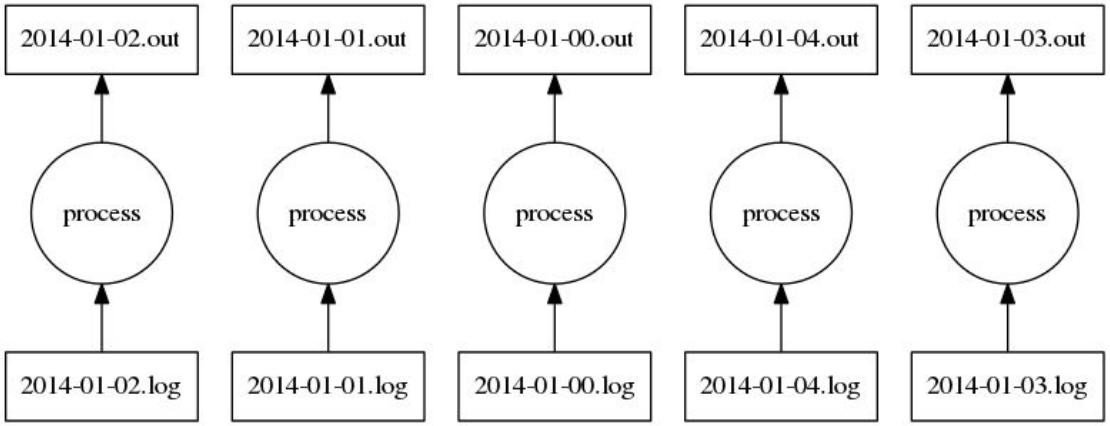
Trabajos Seriales:

- Son aquellos trabajos que por su naturaleza no pueden ser paralelizados, pues cada instrucción depende de una anterior.
- Como consecuencia de esto, este tipo de trabajos solo pueden correr en un solo núcleo.

Arquitectura de un supercomputador

Trabajos Seriales:

Una aproximación para paralelizar este tipo de trabajos aparece cuando se requiere correrse cientos o miles de veces con variación en sus parámetros iniciales y de esta manera obtener un análisis más completo del comportamiento del problema que se está analizando.



Tomado de : Python and Parallelism Matthew Rocklin - Continuum Analytics

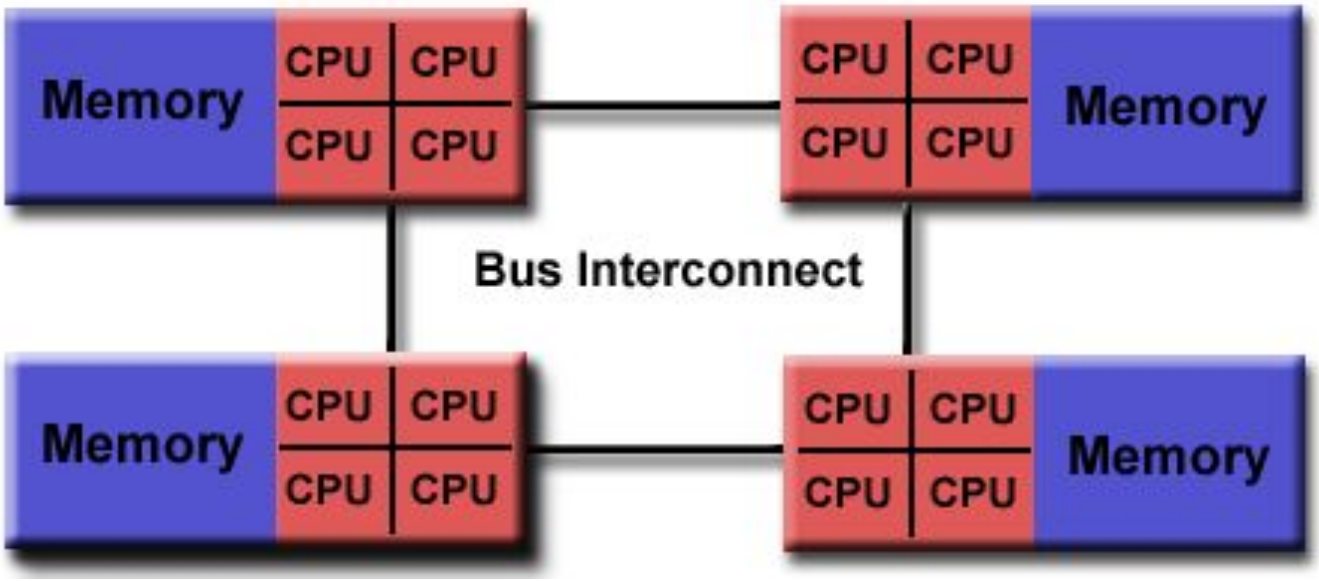
Arquitectura de un supercomputador

Trabajos paralelos:

Son aquellos trabajos que pueden hacer uso del multiprocesamiento y de esta manera utilizar varios núcleos de una mismo nodo de cómputo, es decir, pueden utilizar todos los recursos del sistema donde ocurre su ejecución (núcleos, memoria, entrada/salida, sistema operativo, etc.)

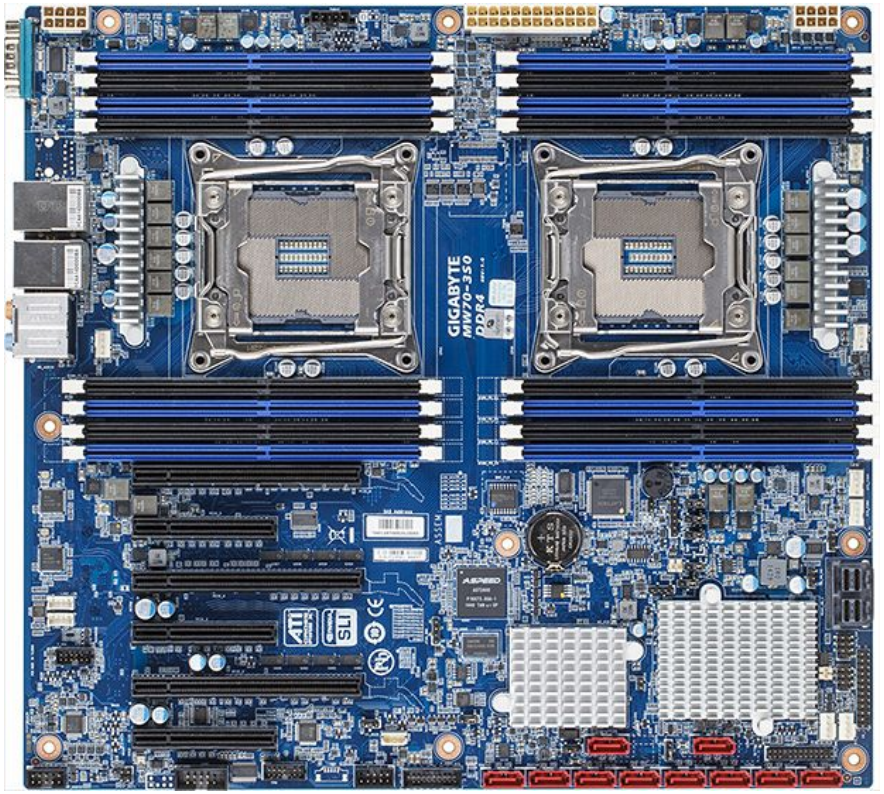
(Existen varias estrategias para la implementación de este tipo de trabajos como pueden ser: hilos de aplicación u OpenMP)

Arquitectura de un supercomputador



Tomado de: HPC Lawrence Livermore National Laboratory (https://goo.gl/vSV5UQ)

Arquitectura de un supercomputador



Tomado de: HPC Lawrence Livermore National Laboratory (https://goo.gl/vSV5UQ)

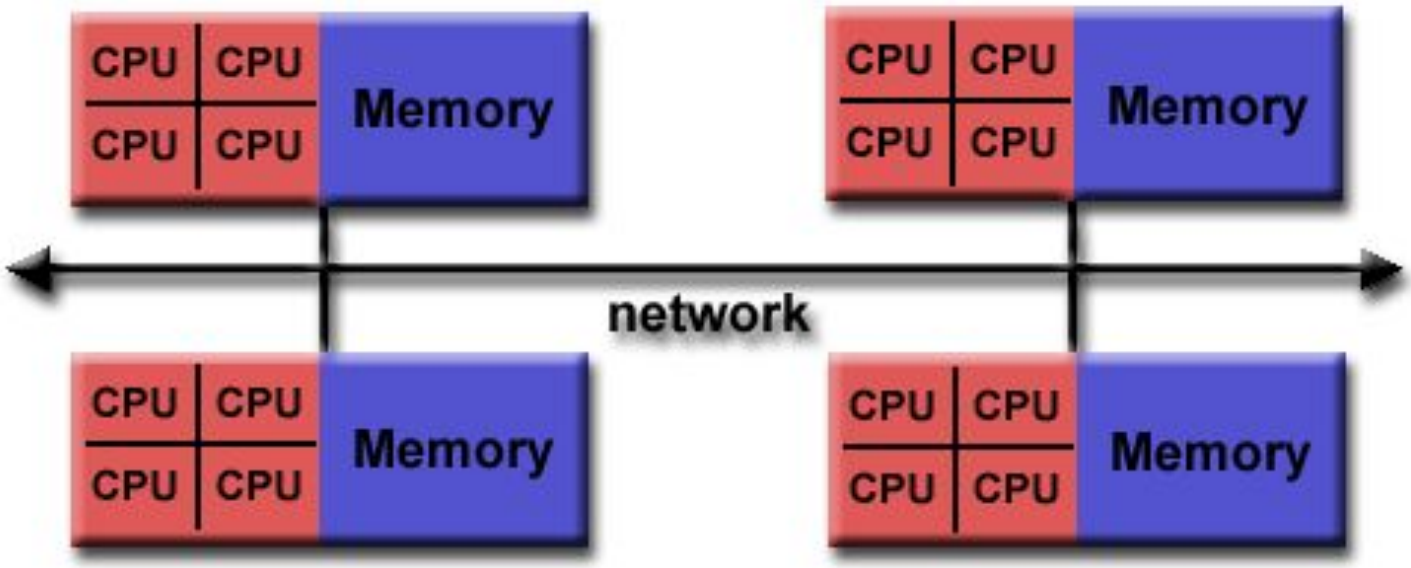
Arquitectura de un supercomputador

Trabajos distribuidos:

Son aquellos trabajos que pueden hacer uso de varios nodos de cómputo (usualmente homogéneos) para su ejecución, es decir, a través de una red común (de alta velocidad) los nodos se interconectan y apoyados en un estándar como MPI (Interfaz de paso de mensajes) se habilita el uso distribuido de los recursos computacionales (núcleos, memoria, etc.).

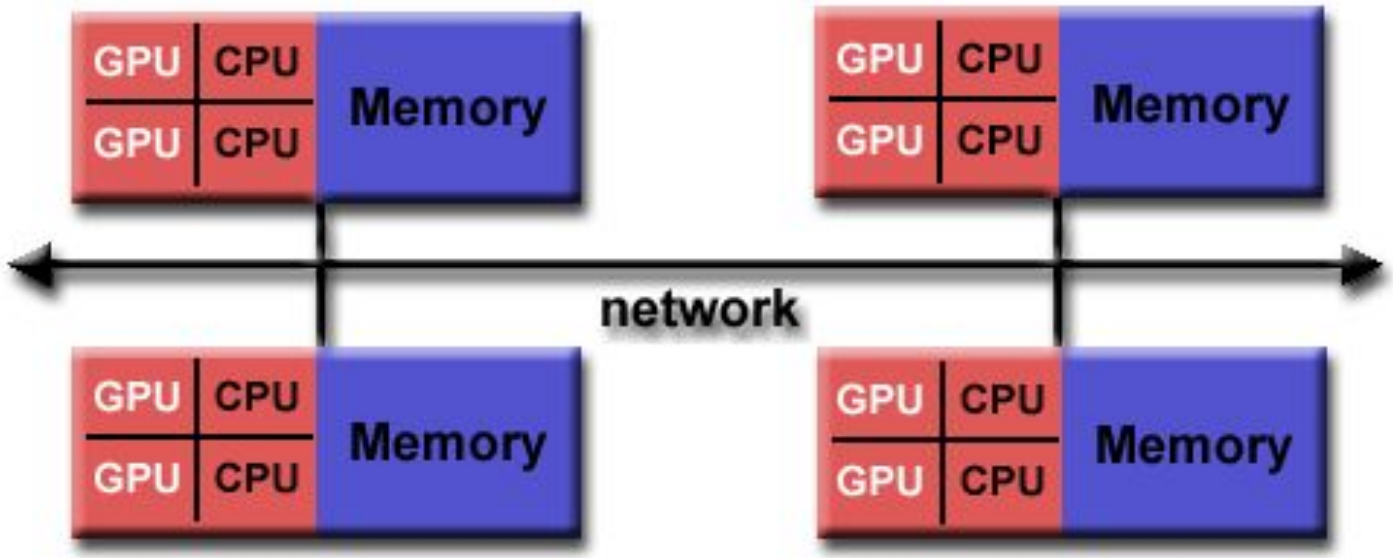
Ejemplos: OpenMPI, Mpich2, MVAPICH

Arquitectura de un supercomputador



Tomado de: HPC Lawrence Livermore National Laboratory (https://goo.gl/vSV5UQ)

Arquitectura de un supercomputador

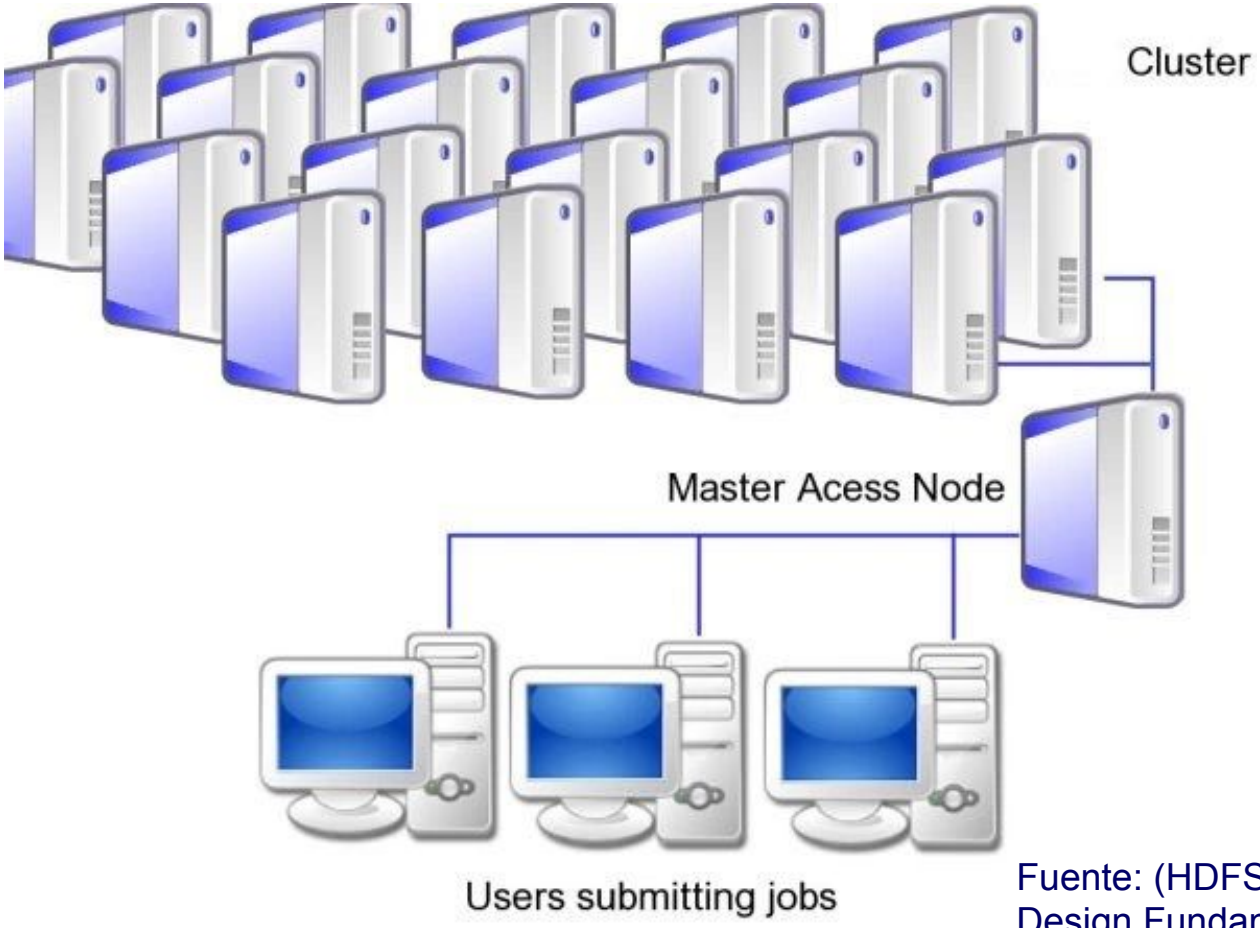


Tomado de: HPC Lawrence Livermore National Laboratory (https://goo.gl/vSV5UQ)

Arquitectura de un supercomputador

Tomado de: AGRUPACIONES MUSICALES - LA ORQUESTA SINFÓNICA https://goo.gl/k8Kd6KCA

Arquitectura de un supercomputador



Fuente: (HDFS) Concepts and Design Fundamentals

Sistema de Gestión de Recursos

Tomado de: Freepik - Hombre teniendo una idea

Sistema de Gestión de Recursos

Para entender el sistema de gestión de recursos tenemos que tener en cuenta tres subsistemas clave: **gestión de trabajos, gestión de recursos** y **planificación**.

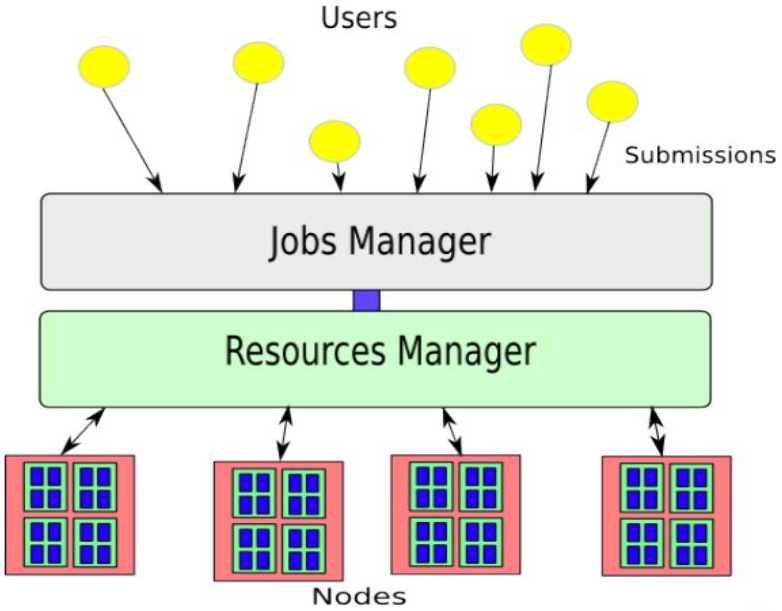
- **Gestión de trabajos**
  - Declaración de trabajos (tipos de trabajo, características de los recursos a utilizar, ambiente de ejecución, etc.)
  - Control de trabajos (Envío de señales, modificación de prioridades, etc.)
  - Monitoreo (Reportes, visualización, seguimiento, etc.)

Sistema de Gestión de Recursos

- **Gestión de recursos**
  - Manejo de los recursos
    (Jerarquía, particiones o colas, límites, etc.).
  - Lanzamiento de los trabajos, propagación y control de la ejecución.
  - Asignación y protección de los recursos computacionales
  - Manejo de usuarios (Jerarquía)
  - Administración de Permisos

Sistema de Gestión de Recursos

- **Planificación**
  - Algoritmos de planificación (Políticas)
  - Gestión de las particiones (Cálculo de prioridades)
  - Administración de licencias
  - Reservación de recursos



Tomado de: Cluster Computing - Resource and Job Management for HPC

Sistema de Gestión de Recursos

El sistema de gestión de recursos es un software particular que tiene como responsabilidad distribuir y administrar la capacidad de procesamiento de un supercomputador.

Con los objetivos de satisfacer la demanda de necesidades de los usuarios y optimizar la utilización de los recursos disponibles.

Gestión de Recursos

Algunos de los gestores de recursos
(planificadores y gestor de recursos) más populares son:

- LSF (IBM)
- Maui/Moab/Torque (Adaptive computing)
- PBS Pro (Altair)
- SGE (Sun Grid Engine)
- SLURM

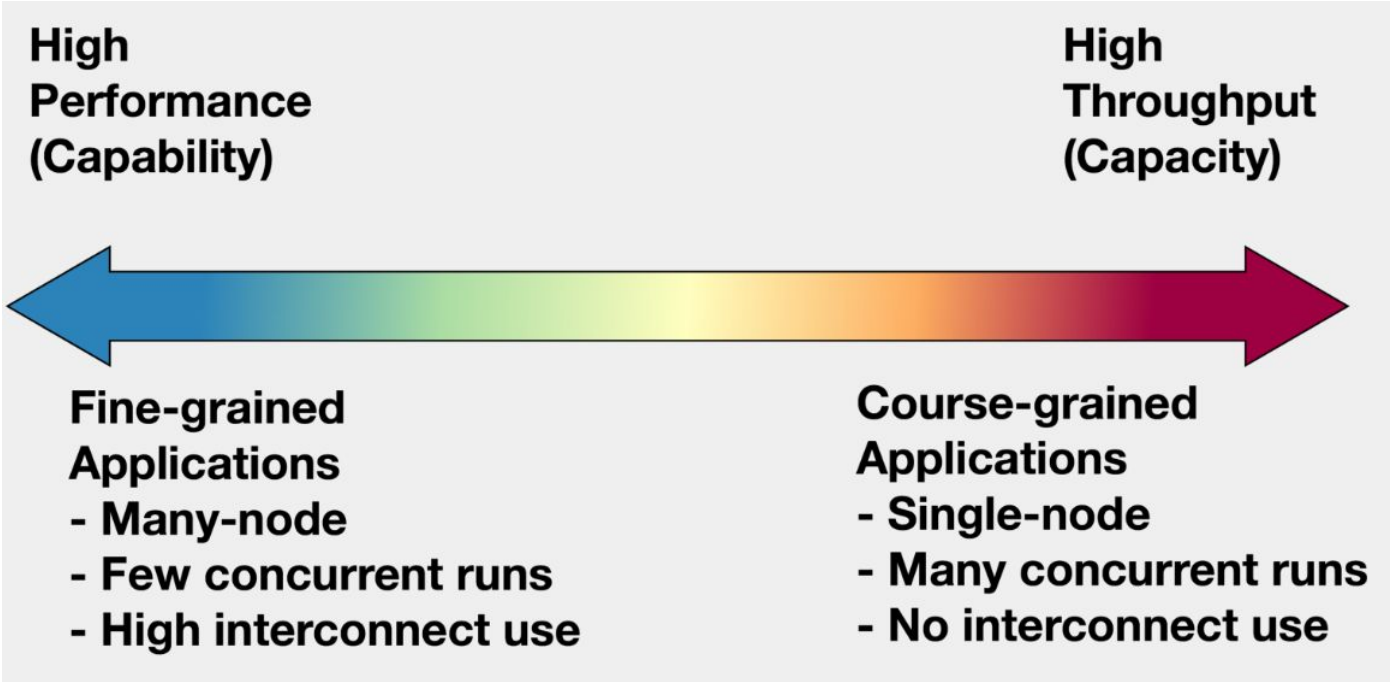
Gestor de recursos - APOLO

1. Pineda (2012-1)
2. Condor (2012-2)
3. Maui/Torque (2013-1 - 2016-1)
4. SLURM (2016-2 - Actualidad)

slurm
workload manager

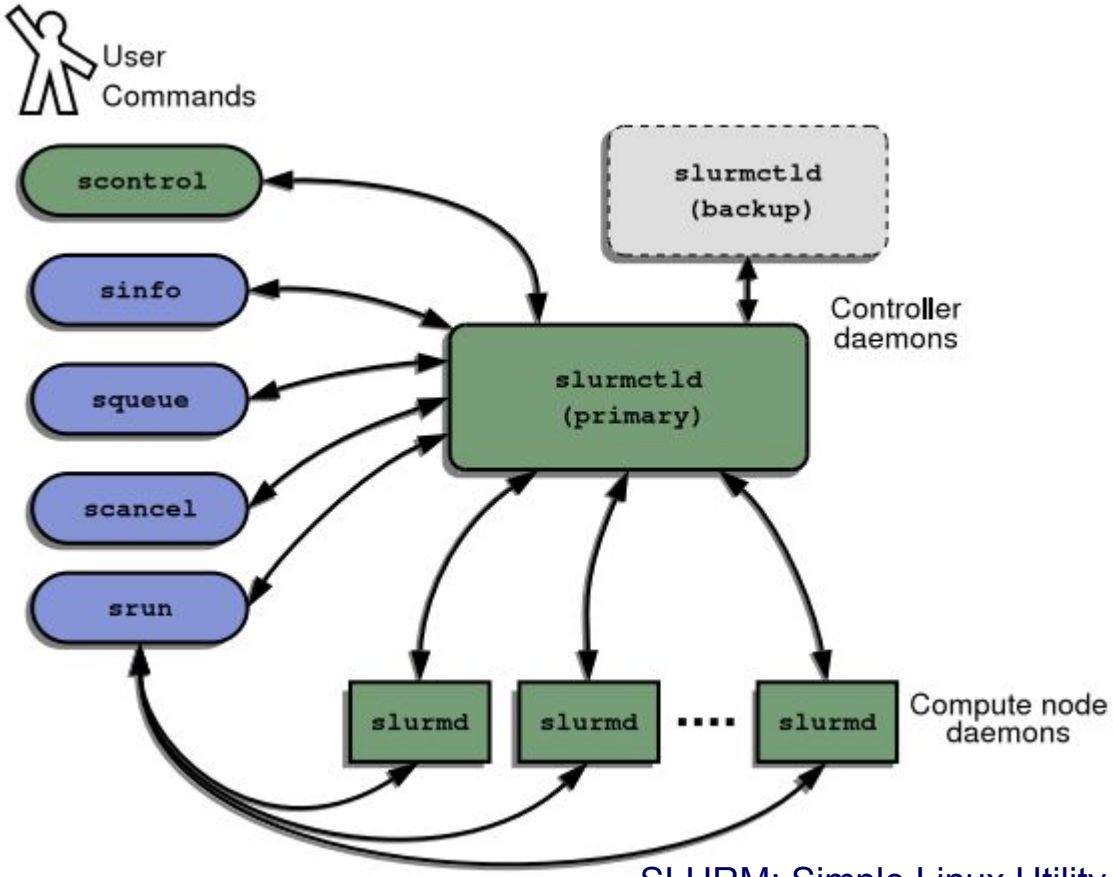
Simple Linux Utility for Resource Manager

Gestor de recursos - APOLO



Tomado de: Introduction to HPC - UCL

Arquitectura del gestor de recursos



SLURM: Simple Linux Utility for Resource Manager

Términos básicos - Gestor de recursos

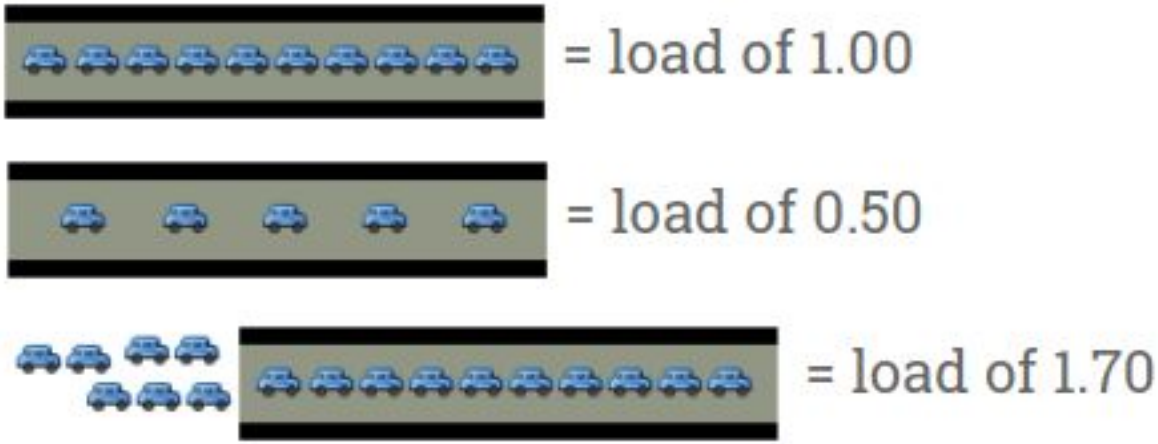
- **Trabajo:** unidad mínima para la ejecución de tareas por parte de un usuario al gestor de recursos; al menos debe contener una tarea
- **Tarea:** Es la ejecución de una aplicación de usuario; una tarea no puede ser ejecutada por fuera de un trabajo.
- **Usuario:** Usuario válido en el sistema de gestión de recursos.
- **Cuenta:** Entidad que agrupa usuarios u otras entidades (Jerarquía).
- **Plantilla para un trabajo:** Archivo con la definición de petición de recursos y flujo de trabajo a realizar (Recursos, ambiente, tareas del trabajo).
- **Prólogo:** Rutina que se ejecuta antes de comenzar con la ejecución de un trabajo.
- **Epílogo:** Rutina que se ejecuta al terminar la ejecución de un trabajo.

Términos básicos - Gestor de recursos

- **Walltime (tiempo transcurrido real):** es el tiempo tomado por un trabajo desde su inicio hasta el final.
- **Tiempo de CPU:** es la cantidad de tiempo que un núcleo fue utilizado por un trabajo, teniendo en cuenta que dependiendo del tipo trabajo este tiempo será el tiempo total utilizado por cada unos de los núcleos utilizados por el trabajo.
- **Tiempo libre (IDLE):** es el tiempo en el que el procesador no es utilizado por ningún trabajo.

Términos básicos - Gestor de recursos

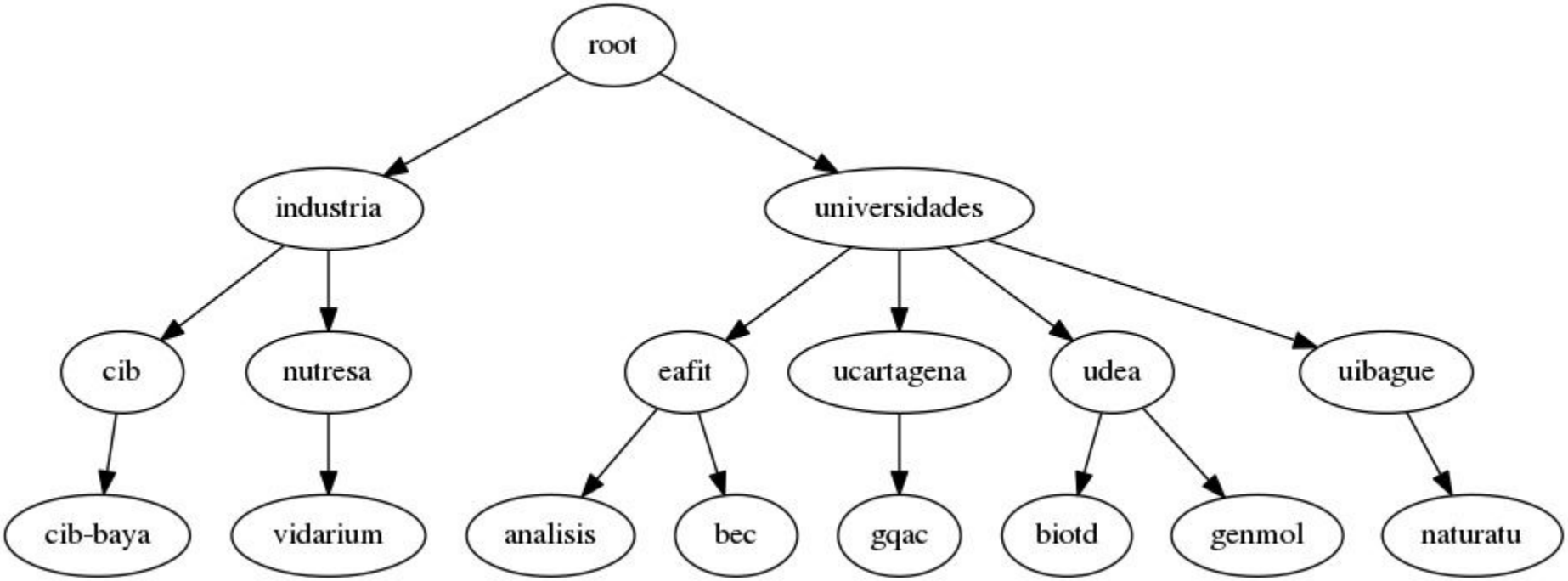
- **Carga del sistema ***
- **Carga promedio del sistema**



Tomado de: Scout - Understanding Linux CPU Load - when should you be worried?

Términos básicos - Gestor de recursos

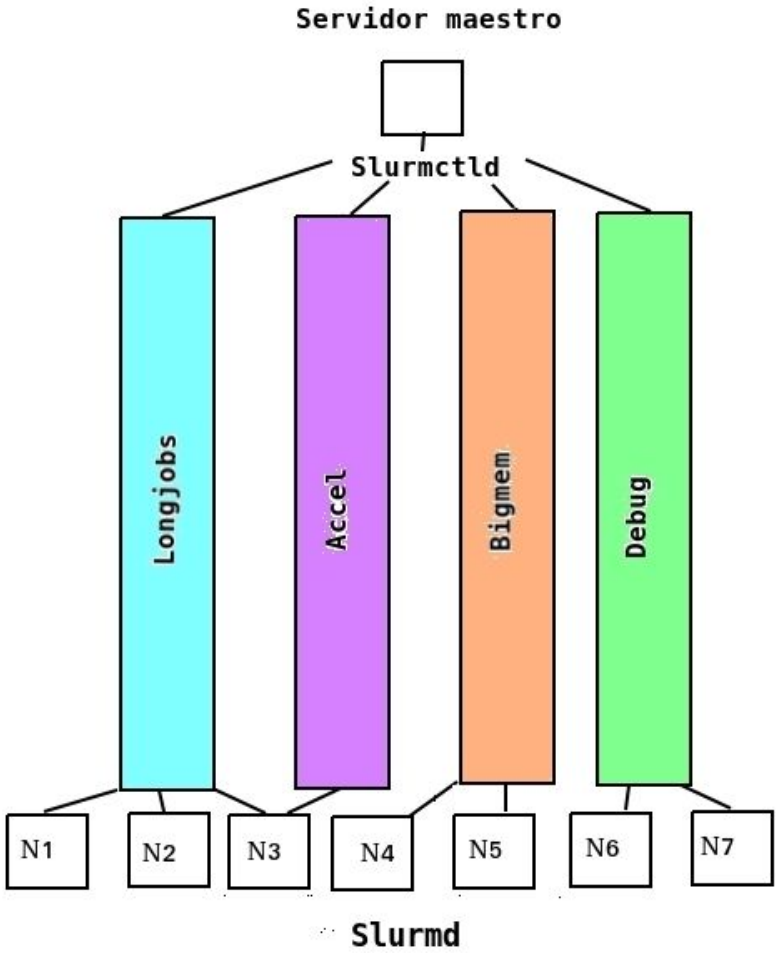
Jerarquía



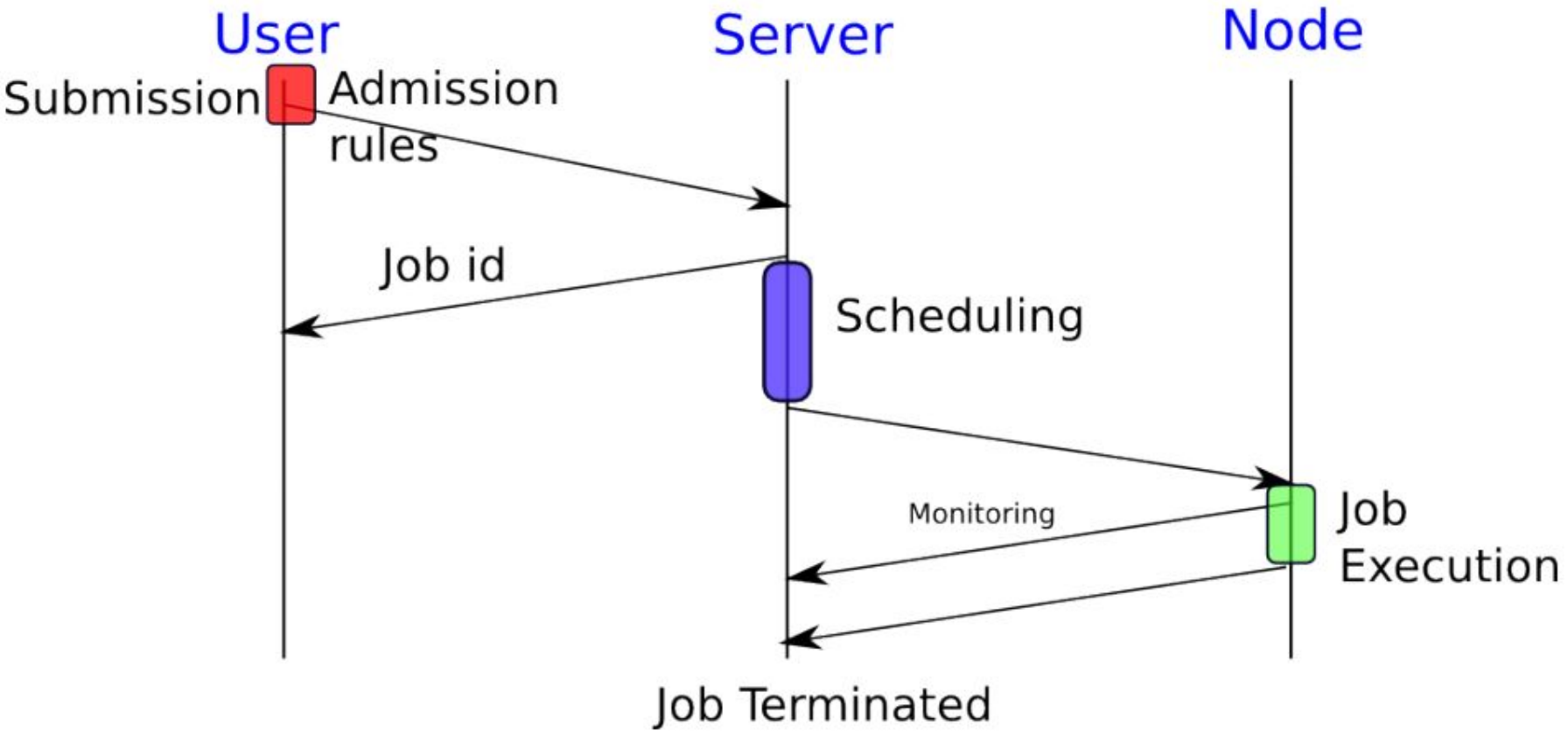
Términos básicos - Gestor de recursos

Particiones o colas *:

- Tiempos límite. (*)
- Agrupación de nodos.
- Definición de recursos por nodo (cpu, mem, gpu, etc.).
- Mínimo y máximo uso de recursos.
- Prioridad de asignación de los nodos.
- Valores por defecto en asignación de recursos.
- Permisos sobre cuentas, usuarios.
- Asignación de QOS.



Ciclo de vida de un trabajo



Tomado de: Cluster Computing - Resource and Job Management for HPC

Acceso usuarios

Interfaz para los usuarios

- Sesión SSH - Línea de comandos (Bash, ZSH, CSH, etc.)
- Trabajos interactivos
- Trabajos desatendidos
- Archivo de ejecución de un programa (**slurm.sh**)
  - Requerimiento de recursos computacionales
  - Integración con el tipo de simulación.
  - Definición del ambiente del trabajo

Acceso usuarios

Comandos para interactuar con el sistema de gestión de recursos:

- Lanzar un trabajo (**sbatch slurm.sh**)
- Cancelar un trabajo (**scancel 777**)
- Retener un trabajo (**scontrol hold 777**)
- Monitorear trabajos. (**scontrol show job 777**)
- ...
- Estado del clúster (partición, usuario, cuenta, etc.)

Prioridad - Gestor de recursos

- **Básica**: La prioridad de los trabajos es igual (orden de llegada).
- **Usuarios, Cuenta:** se asigna mayor prioridad a los trabajos enviado por ciertos usuarios o grupos.
- **Multi-factor:** Los trabajos son priorizados dependiendo diferentes criterios (dinámico):
  - **Tiempo en cola:** el tiempo de espera en cola.
  - **Tiempo computado:** favorece al usuario con menor uso de los recursos
  - **Buen uso:** es la diferencia entre el los recursos pedidos y los recursos utilizados.
  - **Tamaño del trabajo:** número de nodos o núcleos de un trabajo.
  - **Partición o cola:** prioridad asignada a la propia partición.
  - **QOS:** prioridad asignada a cada QOS asignado.
  - **TRES:** prioridad asignada por el uso de cada tipo de recurso (cpu, mem, gpu,etc.)

Planificación

El planificador es el responsable de dictar la ejecución de los trabajos teniendo en cuenta de las necesidades definidas por el usuario, recursos computacionales requeridos, permisos, reglas y prioridades implementadas.

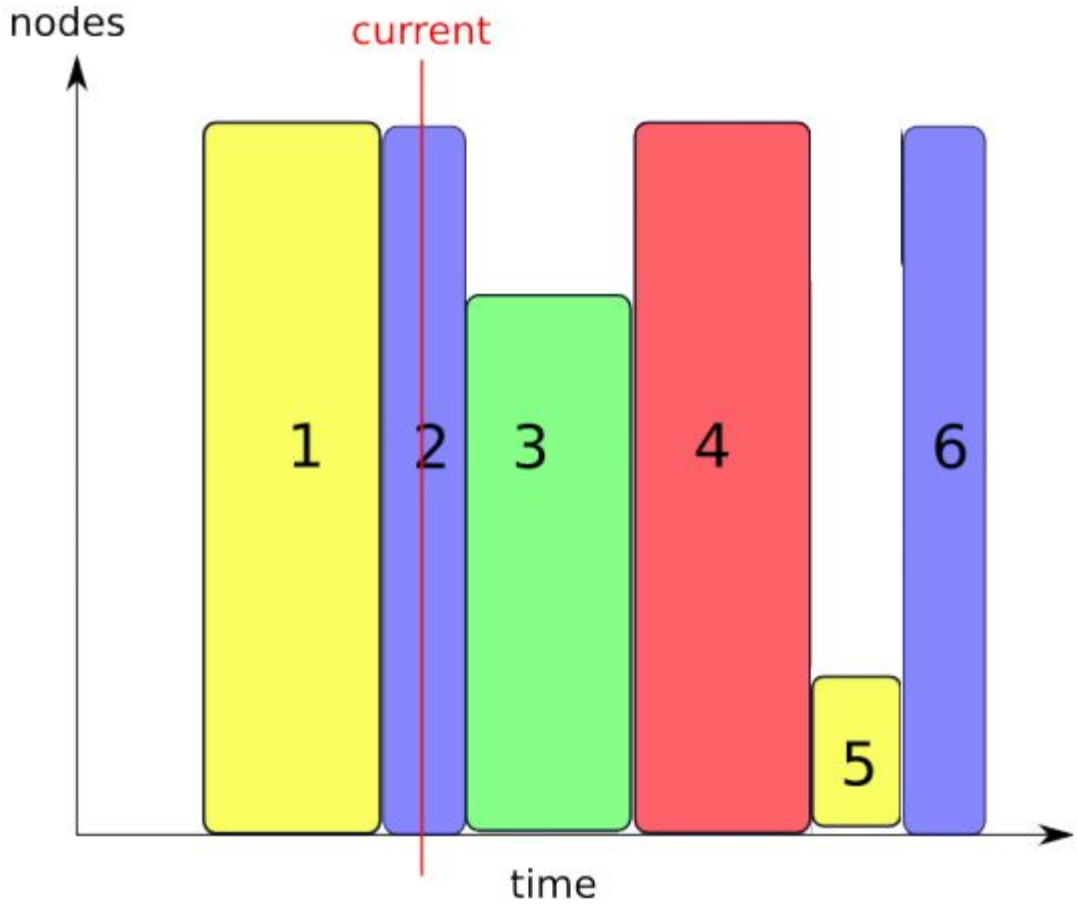
Una función típica del planificador en cooperación con las particiones o colas es definir el orden de entrada de los trabajos.

Planificación - Políticas más comunes

Políticas de planificación:

- **FIFO** - First In First Out
  (Primero en entrar primero en salir)
- **Backfill** (Relleno)
- **Preemption** (Derecho preferente de entrada)
- **Fair-share** (Uso justo)
- **Time-sharing** (Tiempo compartido)
- **Exclusive-mode** (Modo exclusivo)
- **Gang scheduling** (Planificación alternada)

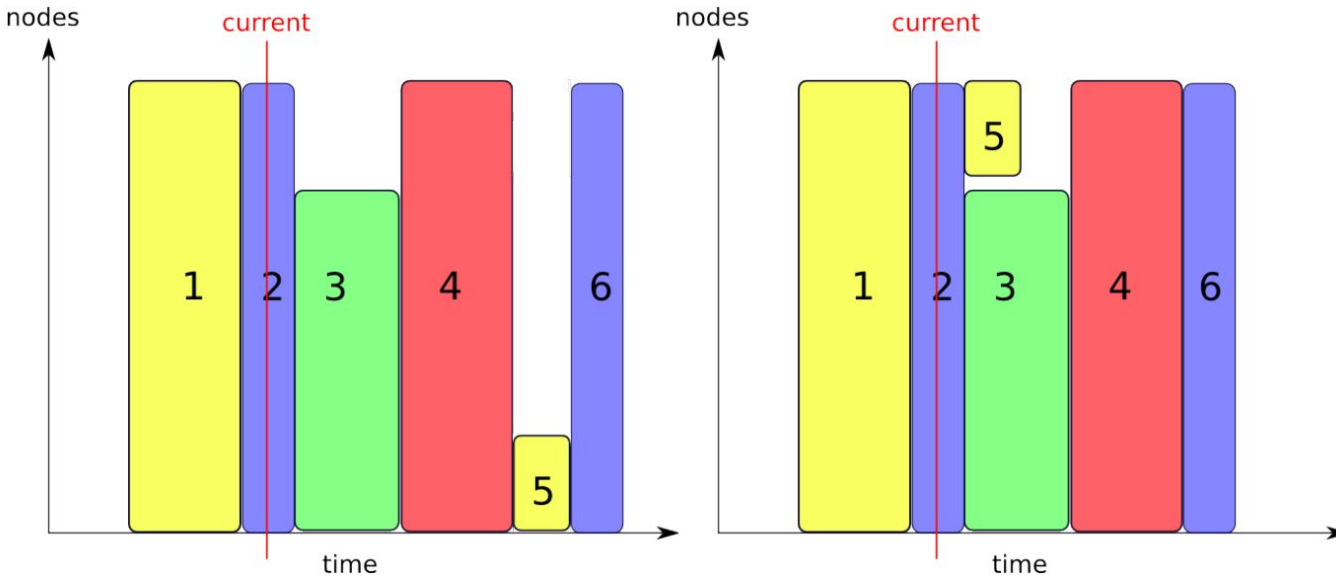
Políticas Planificación - FIFO



Los trabajos son atendidos en el orden de llegada.

Tomado de: Cluster Computing - Resource and Job Management for HPC

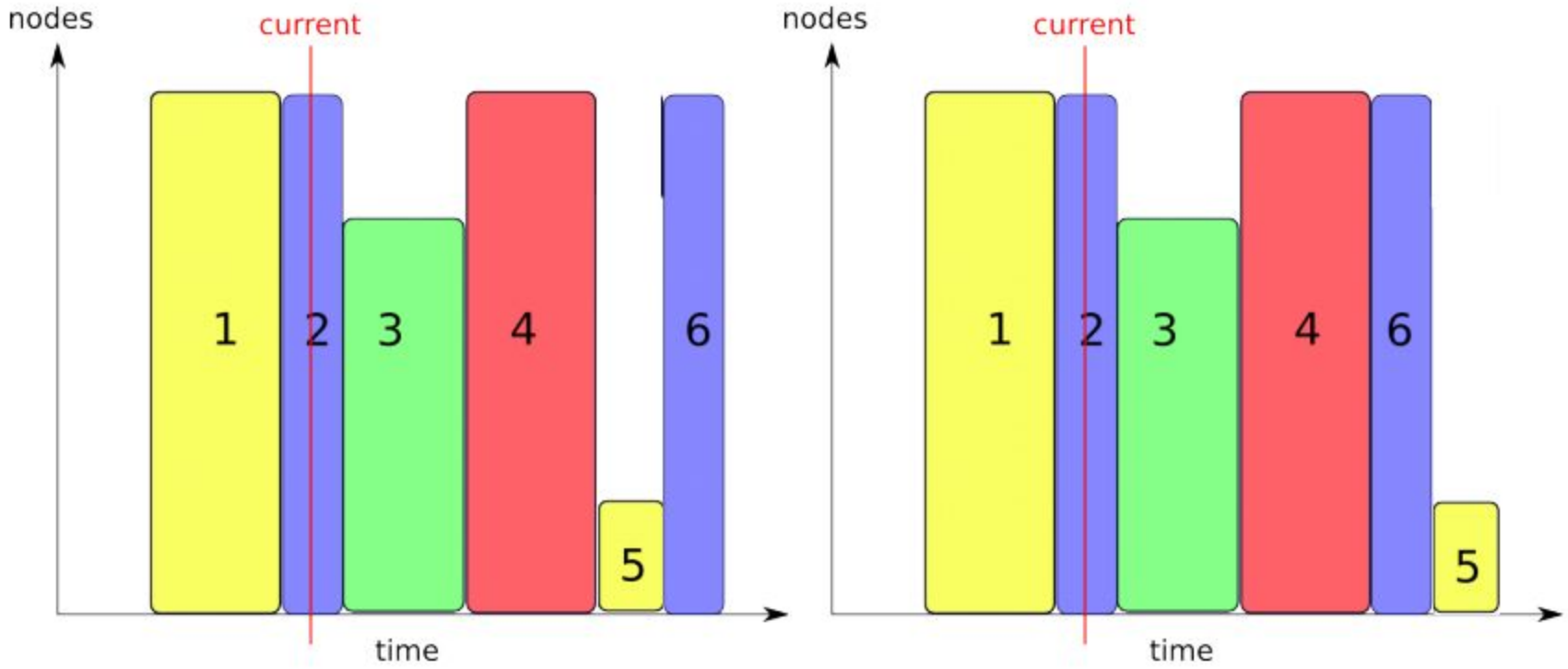
Políticas Planificación - Backfill (*)



Los espacios vacíos son llenados en la tabla de planificación sin modificar el orden de ejecución de los trabajos anteriormente enviados.

Tomado de: Cluster Computing - Resource and Job Management for HPC

Políticas Planificación - Fair-share



Tomado de: Cluster Computing - Resource and Job Management for HPC

Políticas Planificación

Preemption (Derecho preferente de entrada)
Si un trabajo de mayor prioridad llega a la partición el trabajo o los trabajos que se están ejecutando pueden ser suspendidos, cancelados, re-encolados o salvados (checkpointing).

Time-sharing (Tiempo compartido)
Habilita el uso compartido de los recursos si es posible (*).

Exclusive-mode (Modo exclusivo)
Los trabajos no comparten recursos, es decir, en un nodo sólo puede correr un trabajo.

Gang scheduling (Planificación alternada)
Múltiples trabajos son asignados a los mismos recursos y son alternados (suspendidos/resumidos) dejando solo uno de ellos al tiempo, con una duración previamente definida.

Gestión de Recursos

Administración de usuarios:

- **Autenticación:**
  - Permisos
    - Accesos a particiones o colas.
    - Tiempo de uso de los recursos
    - Cantidad de los recursos (anti-kidnapping)

- **Accounting (Informes) - Uso del supercomputador:**
  - Consumo de recursos
    - Tiempo de CPU, Tiempo real transcurrido
    - Cantidad de memoria RAM utilizada
    - Cantidad de espacio en disco duro utilizado
    - Energía consumida
    - … Nuevos permisos, políticas y necesidades….

Gestión de Recursos - Reservaciones

Pueden reservarse recursos computacionales a través del gestor de recursos y seleccionar cuales cuentas o usuarios pueden correr trabajos durante la duración de la reservación, siendo una característica bastante útil para correr pruebas en el ambiente de producción sin perjudicar el funcionamiento del supercomputador.

Gestión de Recursos - Puntos de control

Checkpointing

Es una técnica que agrega tolerancia a fallos en los sistemas computacionales, básicamente consiste en guardar en disco el estado de la aplicación y en caso de un fallo pueda reiniciarse desde este punto. Esta técnica es particularmente importante en trabajos que pueden correr durante mucho tiempo en un sistema propenso a fallas.

Gestión de Recursos - Puntos de control

Characteristics	System Level	Application level
Triggered by:	User/system	Application
Basic idea	Full memory dump	Save relevant information
When to checkpoint?	Any time	Pre-fixed places
Requires modification of application	No (some technologies require re-compilation)	Yes
Resulting file size	Big	Small
Overhead in exec. time	~1-2%	negligible

Tomado de: Cluster Computing - Resource and Job Management for HPC