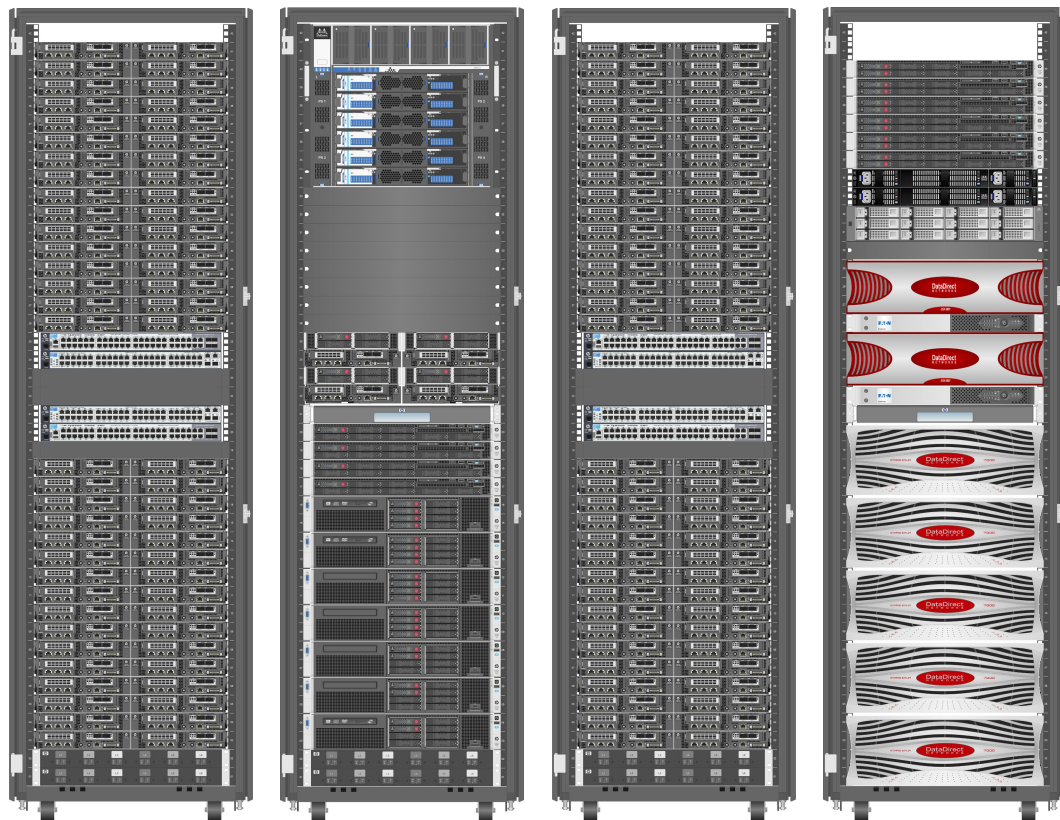


Guía par Usar la Infraestructura del NLHPC



Infraestructura Leftraru

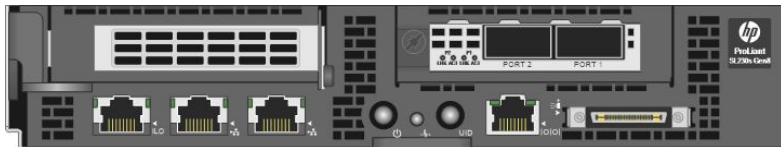


- 44 TFlops
- 2640 cores
- 128 nodos *slims*
- 4 nodos *fats*
- 12 Xeon Phi
- +200 TB almacenamiento DDN Lustre
- Infiniband FDR 56Gbps
- Red servicio 1 Gbps
- Red iLO para adm. hw.
- Racks enfriados por agua
- Enfriamiento in-row respaldo
- UPS 120 KVA autonomía: 30 mins.

Nodos de cómputo Leftrarú

2 tipos de nodo de cómputo:

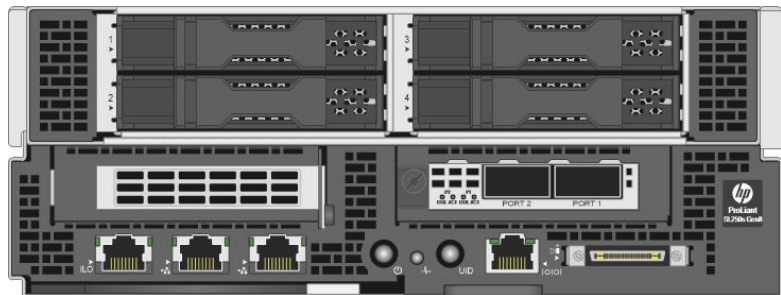
Slim



Nodo Slim

- 2 CPUs Xeon E5-2660 v2, 10 cores c.u.
- **48 GB** RAM DIMM DDR3
- HD Interno 300 GB
- InfiniBand FDR 56 Gbps

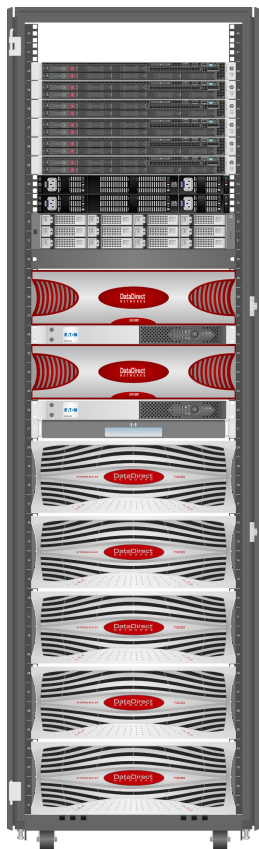
Fat



Nodo Fat

- 2 CPUs Xeon E5-2660 v2, 10 cores c.u.
- **64 GB** RAM DIMM DDR3
- HD Interno 300 GB
- InfiniBand FDR 56 Gbps
- **3 coprocesadores Xeon Phi**
 - 240 cores c/u.
 - 8 GB RAM c/u.

Almacenamiento Leftraru



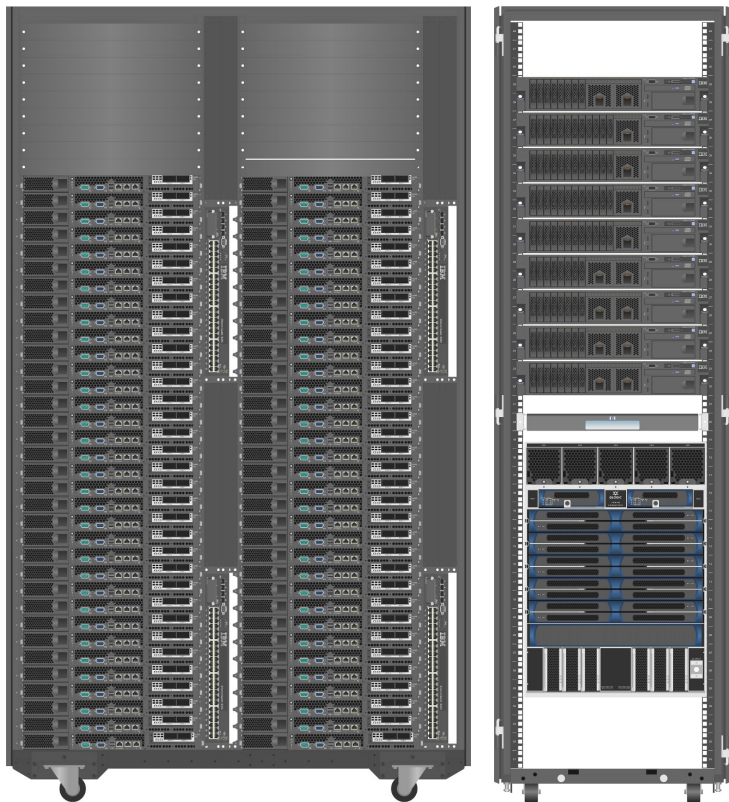
DDN EXAScaler

- Almacenamiento paralelo de clase mundial
- Alto rendimiento en operaciones IO
- Tolerante a fallas (alta disponibilidad)
- Interconexión infiniband
- Capacidades Big Data

Características en Leftraru

- Sistema archivos EXAScaler (Lustre)
- +200 TB de almacenamiento
- Almacenamiento metadata separado
- 2 controladoras SFA en H.A.
- 4 nodos OSS conectados a Infiniband en H.A.
- 2 nodos MDS en H.A.

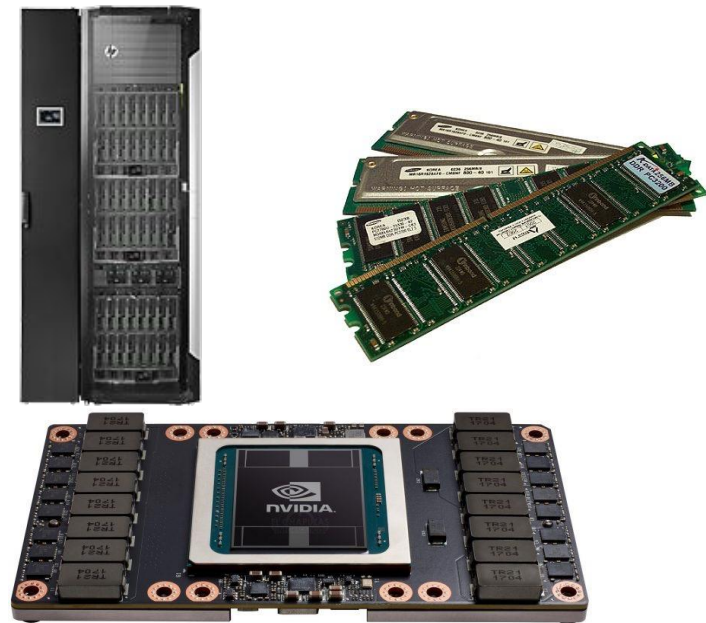
Infraestructura Levque



- **5.6 TFlops**
- **528 cores**
- **66 nodos**
 - 2 CPUs Xeon X5550, 4 cores c/u
 - 24 GB de RAM
- **+200 TB almacenamiento DDN Lustre**
- **Infiniband QDR 40Gbps**
- **Red servicio 1 Gbps**
- **Red IPMI para adm. hw.**
- **Racks enfriados por in-row**
- **UPS 120 KVA autonomía: 30 mins.**

Crecimiento Leftraru 2018

- Nuevos nodos de cómputo:
 - Nodos *slims*:
 - 2x20 cores (mínimo)
 - 96GB de RAM (minimo)
 - Nodos *fats*:
 - 768GB de RAM y hasta 1.5TB (mínimo)
 - Nodos *gpu*:
 - Mínimo 4 GPUs Nvidia P100 (V100 deseable)
- 150TFlops sin coprocesadores (mínimo)



Crecimiento Leftraru 2018

- Nueva generación de procesadores Intel Xeon Skylake (SP):
 - Mayor rendimiento por core y escalabilidad (hasta 1,65x y 28 cores)
 - Mayor rendimiento y capacidad de memoria (50%)
 - Intel Ultra Path Interconnect (Intel UPI)
 - Intel Advanced Vector Extensions 512 (Intel AVX-512)
- Mejora rendimiento en varias áreas:
 - Ganancia en el rendimiento en aplicaciones de deep learning de hasta 113x (respecto a los procesadores actuales de Leftraru)



Introducción a SLURM

Simple Linux Utility for Resource Management

- Administra recursos computacionales y colas de ejecución en clusters
- Open source
- Utilizado en la mayoría de los supercomputadores del top500
- Versión actual instalada en Leftrarú: 17.02.7

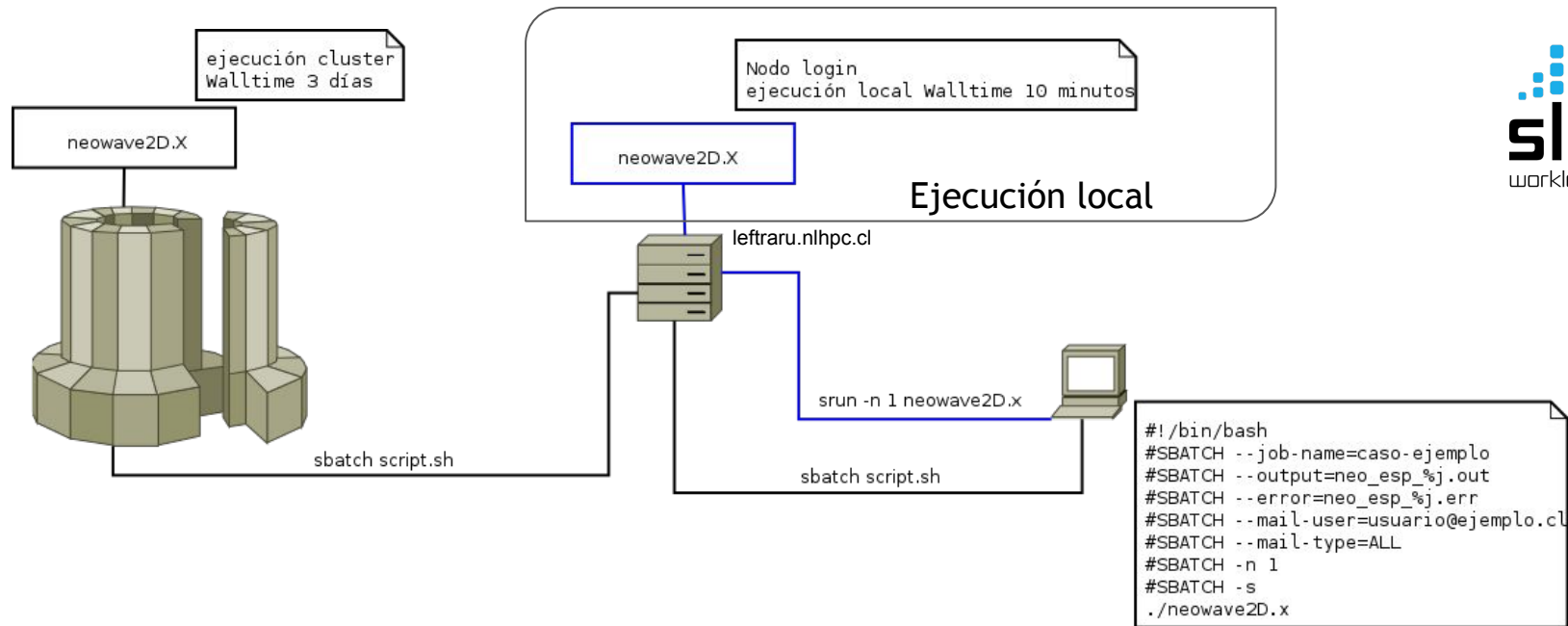


Características de SLURM

- Tres funciones principales:
 - Framework para iniciar, ejecutar y monitorear trabajos
 - Asignación de recursos (exclusivos y no exclusivos)
 - Gestiona tareas manejando una cola de recursos
- Integra una base de datos para reportes históricos
- Puede reservar diferentes recursos: **CPU**, socket, nodo o incluso por **RAM**
 - 120 CPUs en el caso de Leftrarú y 352.000MB de RAM.
- Permite a los administradores modificar tareas en ejecución y realizar reservas.



Funcionamiento



Particiones Leftrararu

- **slims**: partición por defecto, 120 cores, 3 días de ejecución, 128 nodos slims, 4 nodos fats
- **levque**: sin límite de cores, 3 días de ejecución, 41 nodos
- **mics**: usar coprocesadores Intel Xeon Phi en modo nativo
- **largemem**: Nodo con 882GB de RAM
- **debug**: usar los 4 nodos logins (80 cores), 10 minutos de ejecución

- **Uso de nodos logins:**
 - Restringido para ejecución (10 minutos)
 - Válido para pruebas, compilaciones y movimiento de datos



Particiones Leftraru

- ***sinfo***: ver estado de las particiones

```
[root@master2 ~]# sinfo
```

PARTITION	AVAIL	TIMELIMIT	NODES	STATE	NODELIST
slims*	up	infinite	2	resv	cn[012,027]
slims*	up	infinite	5	mix	cn[005,038,053,105,116]
slims*	up	infinite	123	alloc	cn[001-004,006-011,013-026,028-037,039-049,051-052,054-068,070-104,106-115,117-128],cnf[001-004]
slims*	up	infinite	2	idle	cn[050,069]
levque	up	3-00:00:00	1	drain	levque041
levque	up	3-00:00:00	40	idle	levque[001-040]
mics	up	infinite	3	down*	cnf001-mic[0-2]
mics	up	infinite	9	idle	cnf002-mic[0-2],cnf003-mic[0-2],cnf004-mic[0-2]
debug	up	10:00	4	idle	lefraru[1-4]
largemem	up	infinite	1	resv	syntagma001



Prioridad en SLURM

- El job de mayor prioridad en la cola, será el primero en ser ejecutado.
- La prioridad de los trabajos aría en el tiempo según la suma de los siguientes factores:
 - $\text{Priority} = w_{\text{Age}} * v_{\text{Age}} + w_{\text{Fairshare}} * v_{\text{Fairshare}} + w_{\text{Size}} * v_{\text{Size}} + w_{\text{Partition}} * v_{\text{Partition}} + w_{\text{QoS}} * v_{\text{QoS}}$
 - Los factores (v) son flotantes en simple precisión [0.0 - 1.0].
 - Los pesos (w) son enteros de 32 bits.



Prioridad en SLURM

PriorityType=priority/multifactor

2 week half-life

PriorityDecayHalfLife=14-0

PriorityCalcPeriod=5

The smaller the job, the greater its job size priority.

PriorityFavorSmall=YES

The job's age factor reaches 1.0 after waiting in the queue for 14 days.

PriorityMaxAge=14-0

PriorityUsageResetPeriod=NONE

This next group determines the weighting of each of the

components of the Multi-factor Job Priority Plugin.

The default value for each of the following is 1.

PriorityWeightAge=1000

PriorityWeightFairshare=10000

PriorityWeightJobSize=10000

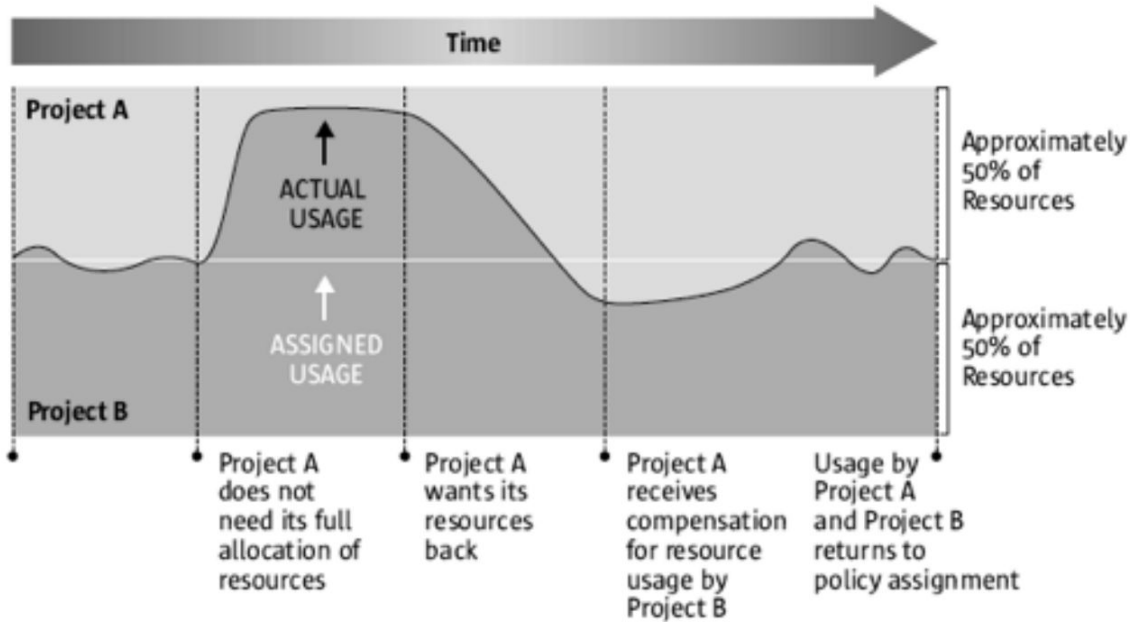
PriorityWeightPartition=100000

PriorityWeightQOS=100000



Prioridad en SLURM

1. Share Tree Policy



Enviar trabajos en SLURM



```
[usuario@leftraru1 ~]$ srun hostname
```

1 proceso

cn028

```
[usuario@leftraru1 ~]$ sbatch <job_script>
```

salida comando

```
Submitted batch job 9142401
```

```
[usuario@leftraru1 ~]$ squeue
```

JOBID	PARTITION	NAME	USER	ST	TIME	NODES	NODELIST(REASON)
9142401	slims	pi_levqu	root	R	0:08	1	cn105

```
[usuario@leftraru1 ~]$ scancel 9142401
```

cancelar trabajo



Slurm: Monitorear mis tareas desde consola

Monitorear desde la consola:

```
[usuario@leftraru1 ~]$ squeue
```



JOBID	PARTITION	NAME	USER	ST	TIME	NODES	NODELIST(REASON)
4400799	slims	example	usuario	R	0:00	1	cn042

Obtener el nombre de el/los nodos y visitar:

<http://dashboard.nlhpc.cl>

<http://monitor.nlhpc.cl/ganglia>



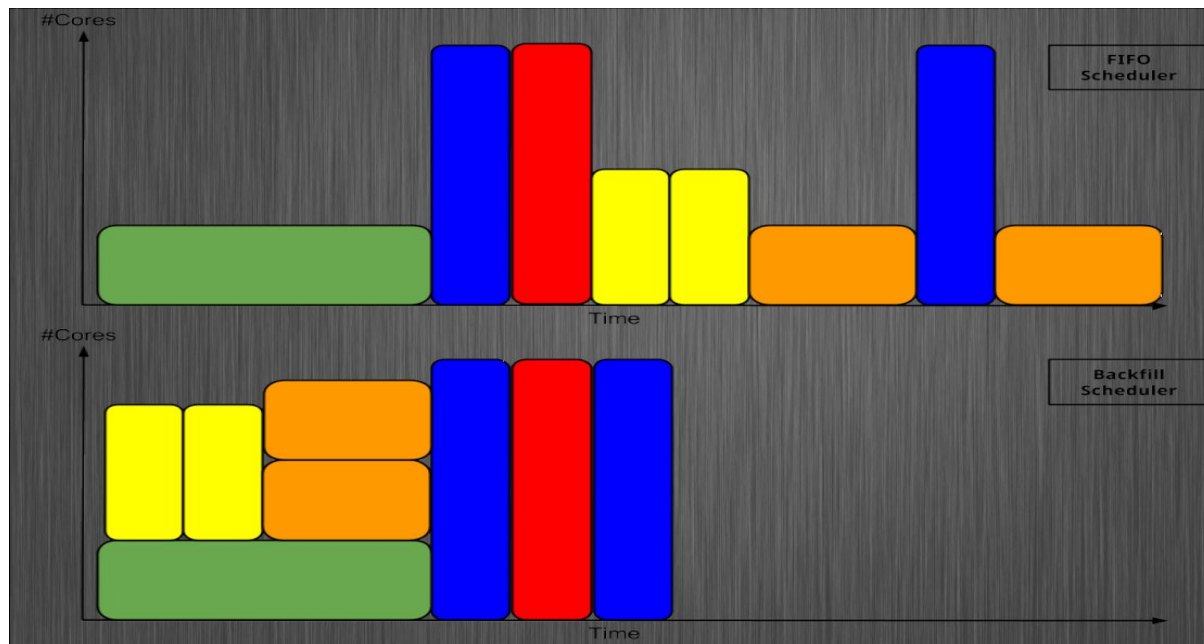
Directivas en SLURM

- `#SBATCH --directive=value` (or “-short-directive value”)
- `#SBATCH --job-name=example` (-J example)
- `#SBATCH --partition slims` (-p slims)
- `#SBATCH --nodes=1` (-N 1)
- `#SBATCH --ntasks=20` (-n 20)
- `#SBATCH --cpus-per-task=20` (-c 20)
- `#SBATCH --ntasks-per-node=20` (--tasks-per-node=20)
- `#SBATCH --ntasks-per-socket=10`
- `#SBATCH --output=file` (-o file)
- `#SBATCH --error=file` (-e file)
- `#SBATCH --exclusive` (--exclusive=user)
- `#SBATCH --mail-user=user@gmail.com`
- `#SBATCH --mail-type=ALL`



Directivas en SLURM

- `#SBATCH --time=DD-HH:MM:SS (-t DD-HH:MM:SS)`



SBATH Ejemplo de script básico test.sh

Utilizar su editor por consola preferido: vim, nano

```
#!/bin/bash
#SBATCH -J ejemplo
#SBATCH -p slims
#SBATCH -n 1
#SBATCH -o archivo_%j.out
#SBATCH -e archivo_%j.err
#SBATCH --mail-user=usuario@gmail.com
#SBATCH --mail-type=ALL

sleep 10
```

Ejecución:
sbatch test.sh

SBATCH - OpenMP

```
#!/bin/bash
#SBATCH -J ejemplo
#SBATCH -p slims
#SBATCH -n 1
#SBATCH -c 20
#SBATCH -o archivo_%j.out
#SBATCH -e archivo_%j.err
#SBATCH --mail-user=usuario@gmail.com
#SBATCH --mail-type=ALL

export OMP_NUM_THREADS=20

./ejecutable
```

Ejecución:
sbatch test.sh

SBATCH - (Intel) MPI

```
#!/bin/bash
#SBATCH -J ejemplo
#SBATCH -p slims
#SBATCH -n 2
#SBATCH -o ejemplo_impi_%j.out
#SBATCH -e ejemplo_impi_%j.err
#SBATCH --mail-user=usuario@gmail.com
#SBATCH --mail-type=ALL

module load intel impi

srun ./hola_mundo
```

```
Hello from thread 00 out of 1 from process 00 out of 2 on
cn005
Hello from thread 00 out of 1 from process 01 out of 2 on
cn005
```

SBATCH - Job Array

Slurm permite enviar y administrar miles de trabajos similares de una sola vez.

- Todos los trabajos deben tener las mismas condiciones iniciales
- Provee de variables para controlar la ejecución de los jobs

```
#!/bin/bash
#SBATCH --job-name=sleep-test
#SBATCH --partition=slims
#SBATCH -n 1
#SBATCH --output=st_%j.out
#SBATCH --error=st_%j.err
#SBATCH --array=1-1000%1
#SBATCH --exclusive=user
#SBATCH --mail-user=usuario@dominio.cl
#SBATCH --mail-type=ALL

./ejecutable entrada_${SLURM_ARRAY_TASK_ID}
```

Módulos

Leftaru carga sus programas mediante “módulos”:

- ¿Usar programas del sistema? (Python)
- Permite varias versiones del mismo software
- No genera conflictos entre versiones
- El software está centralizado
- Se cargan las **variables de entorno** del módulo

Para cargar un módulo: *module load <program>*

module avail: muestra todos los nodos disponibles

module list: lista todos los módulos cargados

module unload: quita un módulo previamente cargado

Nuevo sistema de módulo Lmod

- **Comandos:**
 - *ml*: cargar un módulo (*module load*)
 - *ml unload*: descargar un módulo
 - *ml avail*: muestra todos los nodos disponibles
 - *ml spider*: buscar un módulo
- Lmod está por defecto en levque y syntagma (largemem):

```
sbatch --export=NONE <script.sh>
```

```
srun --export=ALL
```

- Environment Modules por defecto en Leftraru
- Lmod puede ser cargado en Leftraru

```
module load Lmod/6.5
```

```
source $LMOD_PROFILE
```

```
ml intel
```

Programas propios

¿Cómo instalo programas inexistentes en Leftrararu?

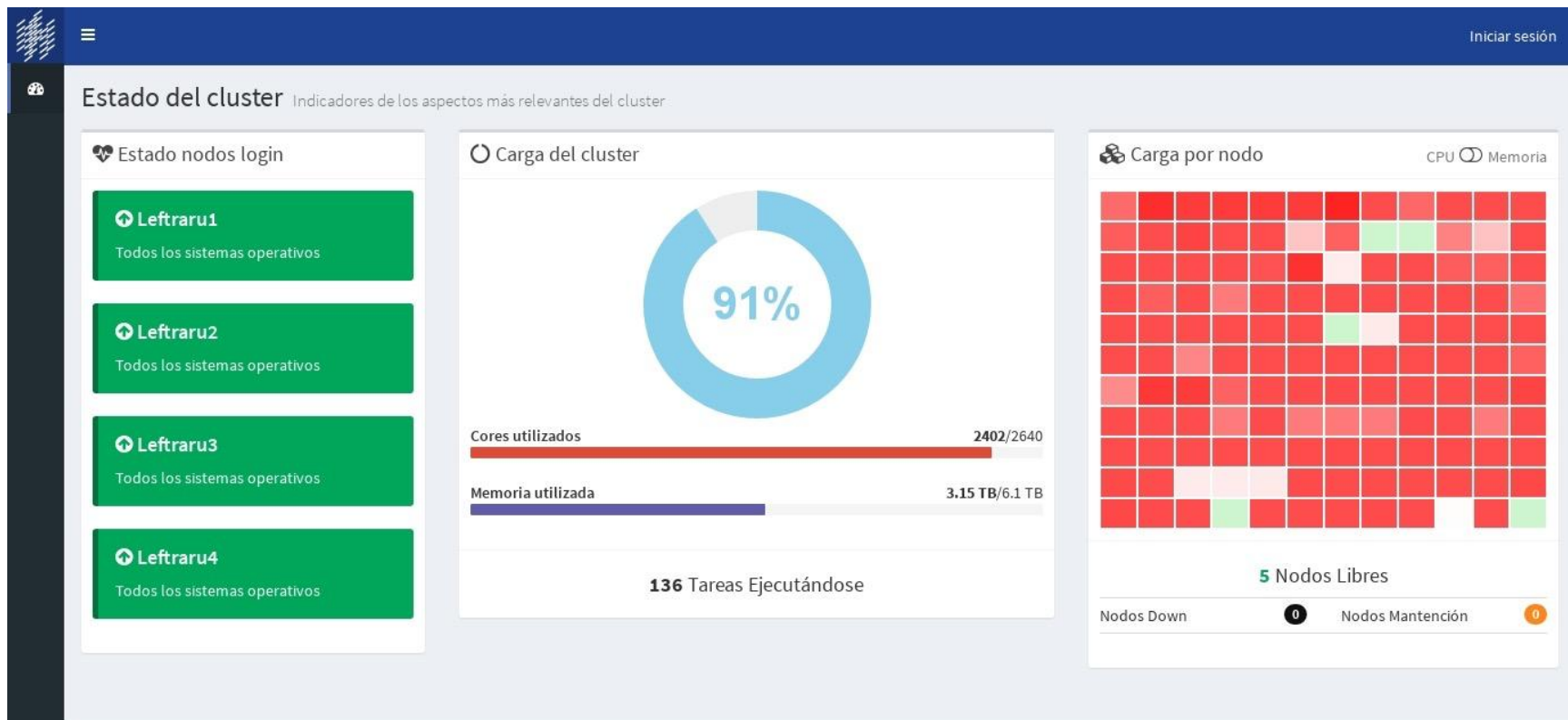
- Copiándolos en su carpeta home
- Compilando uno mismo el código
- Contacte a soporte@nlhpc.cl

Tengo licencia, ¿cómo la hago funcionar?

- Si quiere utilizar el cluster debe ser licencia flotante
- Ponerse en contacto con equipo de soporte: soporte@nlhpc.cl

¿Tiene interfaz gráfica?

Panel de usuarios - dashboard.nlhpc.cl



Panel de usuarios - dashboard.nlhpc.cl



Samuel

Panel de usuario

Información de tareas en la cola de ejecución

**TAREAS EN EJECUCIÓN**
1

**TAREAS EN ESPERA**
3

**CUOTA PERSONAL**
199G / 5T

Tareas Activas

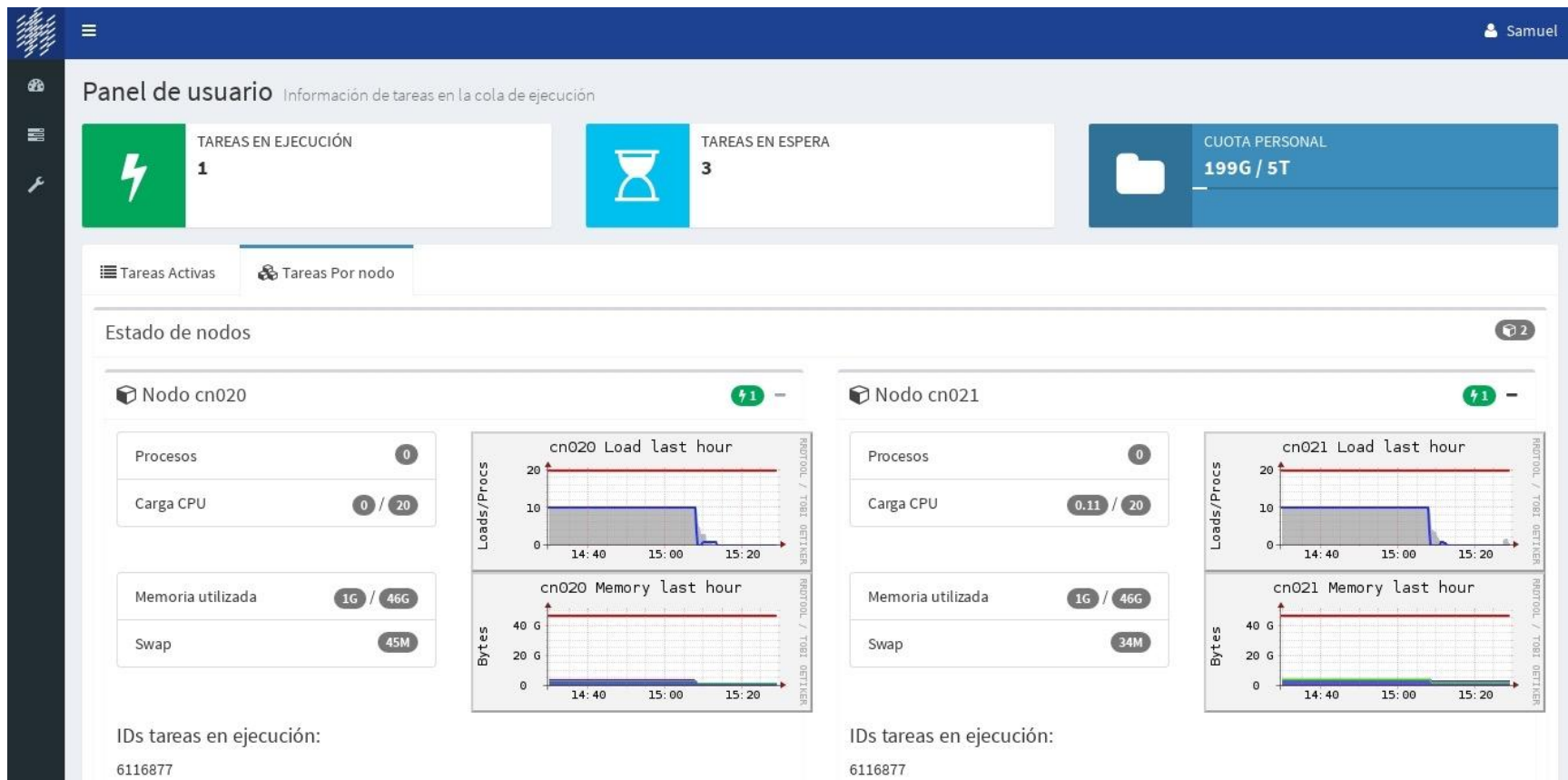
Tareas Por nodo

Ejecutándose

En espera

ID	Nombre	Fecha de envío	Fecha inicio ejecución	Tiempo en ejecución	
6116877	test	17-12-2016 15:28:13	17-12-2016 15:29:22	un minuto	 

Panel de usuarios - dashboard.nlhpc.cl



Ganglia - monitor.nlhpc.cl

Main Search Views Aggregate Graphs Compare Hosts Events Automatic Rotation Live Dashboard Mobile

Compute Nodes Cluster Report at Mon, 02 Oct 2017 18:06:32 -0300

Get Fresh Data

Last hour 2hr 4hr day week month year job or from to Go Clear

Physical View

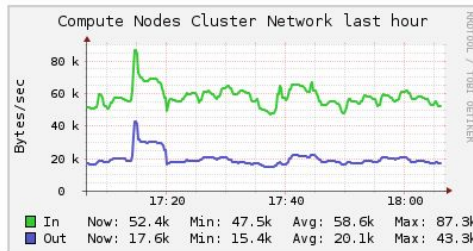
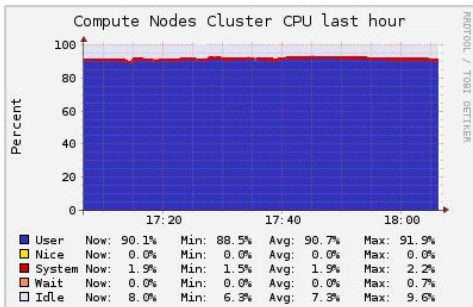
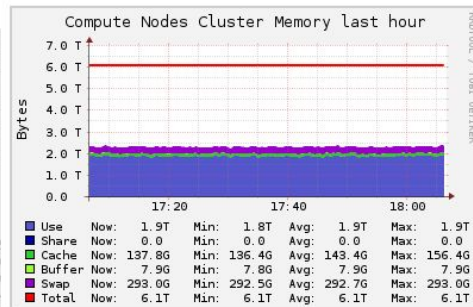
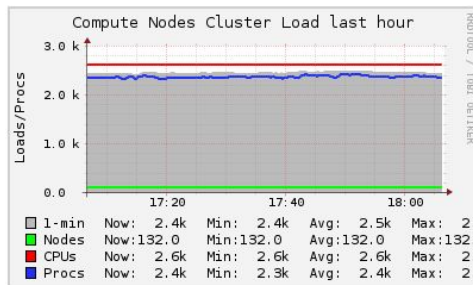
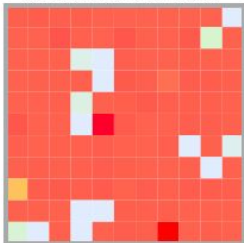
Leftraru Grid > Compute Nodes > --Choose a Node ▾

Overview of Compute Nodes @ 2017-10-02 18:06

CPUs Total: **2640**
Hosts up: **132**
Hosts down: **0**

Current Load Avg (15, 5, 1m):
92%, 93%, 93%
Avg Utilization (last hour):
93%

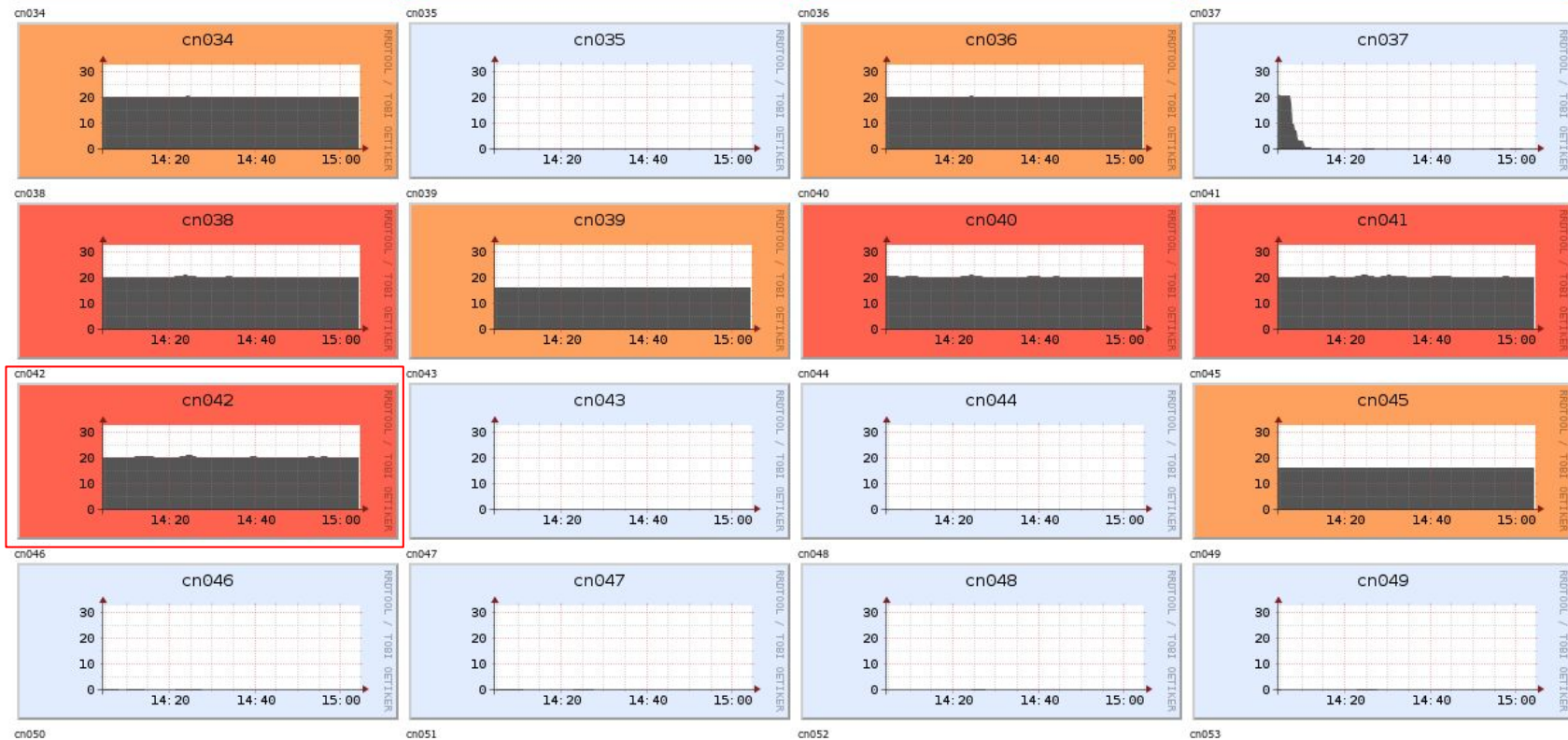
Server Load Distribution



Ganglia - monitor.nlhpc.cl

monitor.nlhpc.cl/ganglia/?r=hour&cs=&ce=&m=load_one&s=by+name&c=Compute%2520Nodes&tab=m&vn=&hi

Buscar



Ganglia - monitor.nlhpc.cl

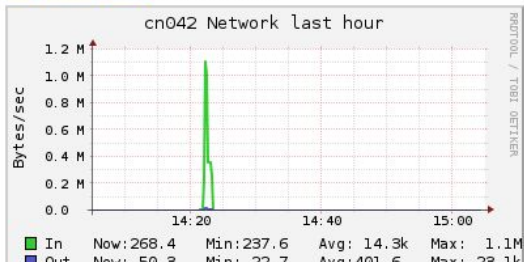
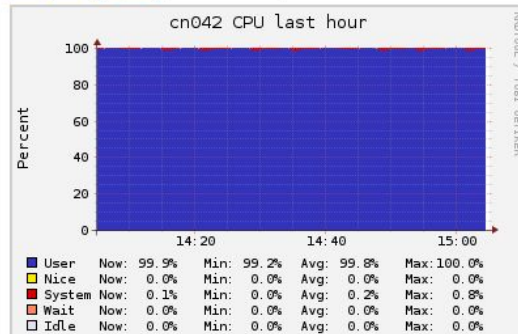
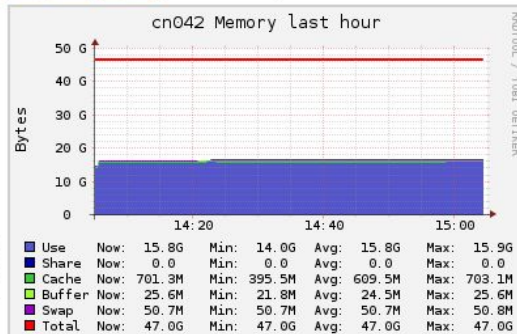
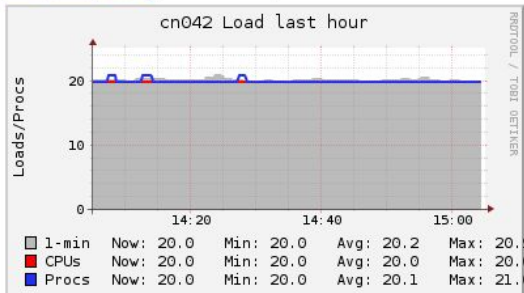
cn042 Host Report at Tue, 16 Aug 2016 15:04:59 -0300

Get Fresh Data

Last or from to

Leftaru Grid > Compute Nodes > cn042

Host Overview



Monitorear Job - *htop*

Puede ingresar a través de ssh a un nodo en donde tenga una tarea en ejecución y ejecutar *htop*

```
1 [|||||100.0%] 6 [ 0.0%] 11 [ 0.0%] 16 [ 0.0%]
2 [ 0.0%] 7 [ 0.0%] 12 [ 0.0%] 17 [ 0.0%]
3 [|||||100.0%] 8 [ 0.0%] 13 [ 0.0%] 18 [ 0.0%]
4 [ 0.0%] 9 [ 0.7%] 14 [ 0.0%] 19 [ 0.0%]
5 [ 0.0%] 10 [ 0.0%] 15 [ 0.0%] 20 [ 0.0%]

Mem[||||| 1.56G/62.7G] Tasks: 44, 39 thr: 3 running
Swp[ 0K/62.5G] Load average: 2.00 2.01 2.05
Uptime: 4 days, 00:36:11

  PID USER      PRI  NI  VIRT   RES   SHR  S CPU% MEM%   TIME+  Command
18305 nperinet   20    0 19020  3768   892  R 100.0  0.0  2h52:34 ./LL_RK4.x
18335 nperinet   20    0 17104  1584   892  R 100.0  0.0  2h47:30 ./LL_RK4.x
18605 root       20    0 121M   2432  1476  R  0.0  0.0  0:00.05 htop
    1 root       20    0 185M   5068  2384  S  0.0  0.0  0:11.06 /usr/lib/systemd/systemd --switched-root --system --de
   669 root       20    0 112M   2000  1564  S  0.0  0.0  0:00.00 /bin/bash
   671 root       20    0 36816  7236  6908  S  0.0  0.0  0:00.78 /usr/lib/systemd/systemd-journald
   726 root       20    0 43808  2428  1268  S  0.0  0.0  0:00.83 /usr/lib/systemd/systemd-udev
  1012 root       16   -4 51188  1620  1236  S  0.0  0.0  0:00.05 /sbin/auditd -n
  1002 root       16   -4 51188  1620  1236  S  0.0  0.0  0:00.25 /sbin/auditd -n
  1206 root       20    0 448M  10956  6968  S  0.0  0.0  0:00.00 /usr/sbin/NetworkManager --no-daemon
  1209 root       20    0 448M  10956  6968  S  0.0  0.0  0:00.09 /usr/sbin/NetworkManager --no-daemon
  1139 root       20    0 448M  10956  6968  S  0.0  0.0  0:02.34 /usr/sbin/NetworkManager --no-daemon
  1144 avahi     20    0 30220  1564  1300  S  0.0  0.0  0:00.44 avahi-daemon: running [cnf004.local]
  1148 dbus    20    0 28824  1772  1352  S  0.0  0.0  0:00.44 /bin/dbus-daemon --system --address=systemd: --nofork
  1166 root       20    0 198M  1236   776  S  0.0  0.0  0:00.00 /usr/sbin/gssproxy -D
  1167 root       20    0 198M  1236   776  S  0.0  0.0  0:00.00 /usr/sbin/gssproxy -D
  1168 root       20    0 198M  1236   776  S  0.0  0.0  0:00.00 /usr/sbin/gssproxy -D
  1169 root       20    0 198M  1236   776  S  0.0  0.0  0:00.00 /usr/sbin/gssproxy -D
  1170 root       20    0 198M  1236   776  S  0.0  0.0  0:00.00 /usr/sbin/gssproxy -D
  1164 root       20    0 198M  1236   776  S  0.0  0.0  0:00.30 /usr/sbin/gssproxy -D

F1Help F2Setup F3Search F4Filter F5Free F6SortBy F7Nice F8Nice F9Kill F10Quit
```

Monitorar Job - dashboard.nlhpc.cl

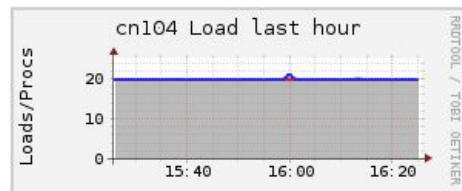
```
#!/bin/bash
#SBATCH -J example
#SBATCH -p slims
#SBATCH -n 1
#SBATCH -c 20
#SBATCH -o example_%j.out
#SBATCH -e example_%j.err
#SBATCH --mail-user=user@gmail.com
#SBATCH --mail-type=ALL

# Load modules (module load intel)
./program
```

cn104

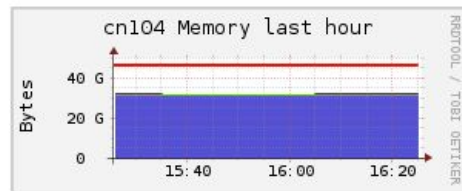
Carga CPU

20.01 / 20



Memoria utilizada

32.07 GB / 46.88 GB



Ver en Ganglia

Cerrar

Ejemplos prácticos - Job OpenMP

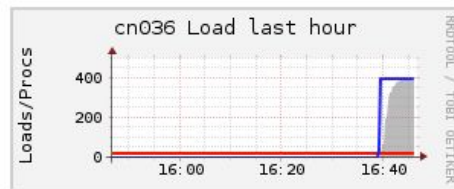
```
#!/bin/bash
#SBATCH -J electrostatic
#SBATCH -p slims
#SBATCH -n 20
#SBATCH -o programa_omp_%j.out
#SBATCH -e programa_omp_%j.err
#SBATCH --mail-user=user@gmail.com
#SBATCH --mail-type=ALL

./programa_omp
```

cn036

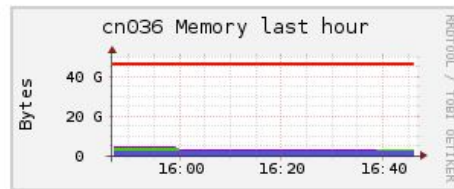
Carga CPU

291.49 / 20



Memoria utilizada

3.37 GB / 46.88 GB



[Ver en Ganglia](#)

Cerrar

Ejemplos prácticos - Job OpenMP

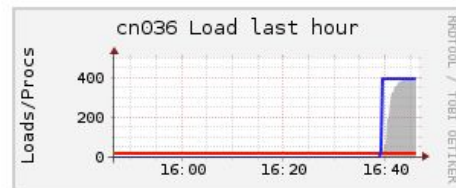
```
#!/bin/bash
#SBATCH -J electrostatic
#SBATCH -p slims
#SBATCH -n 120
#SBATCH -o programa_omp_%j.out
#SBATCH -e programa_omp_%j.err
#SBATCH --mail-user=user@gmail.com
#SBATCH --mail-type=ALL

./programa_omp
```

cn036

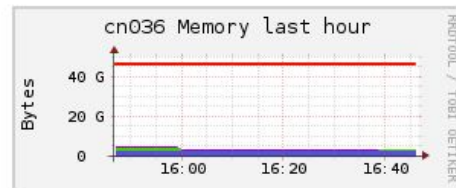
Carga CPU

291.49 / 20



Memoria utilizada

3.37 GB / 46.88 GB



[Ver en Ganglia](#)

Cerrar

Ejemplos prácticos - Job OpenMP

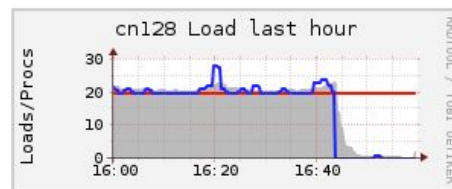
```
#!/bin/bash
#SBATCH -J programa_omp
#SBATCH -p slims
#SBATCH -c 20
#SBATCH -o programa_omp_%j.out
#SBATCH -e programa_omp_%j.err
#SBATCH --mail-user=user@gmail.com
#SBATCH --mail-type=ALL

./programa_omp
```

cn128

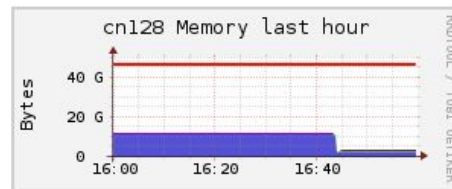
Carga CPU

1.01 / 20



Memoria utilizada

3.08 GB / 46.88 GB



[Ver en Ganglia](#)

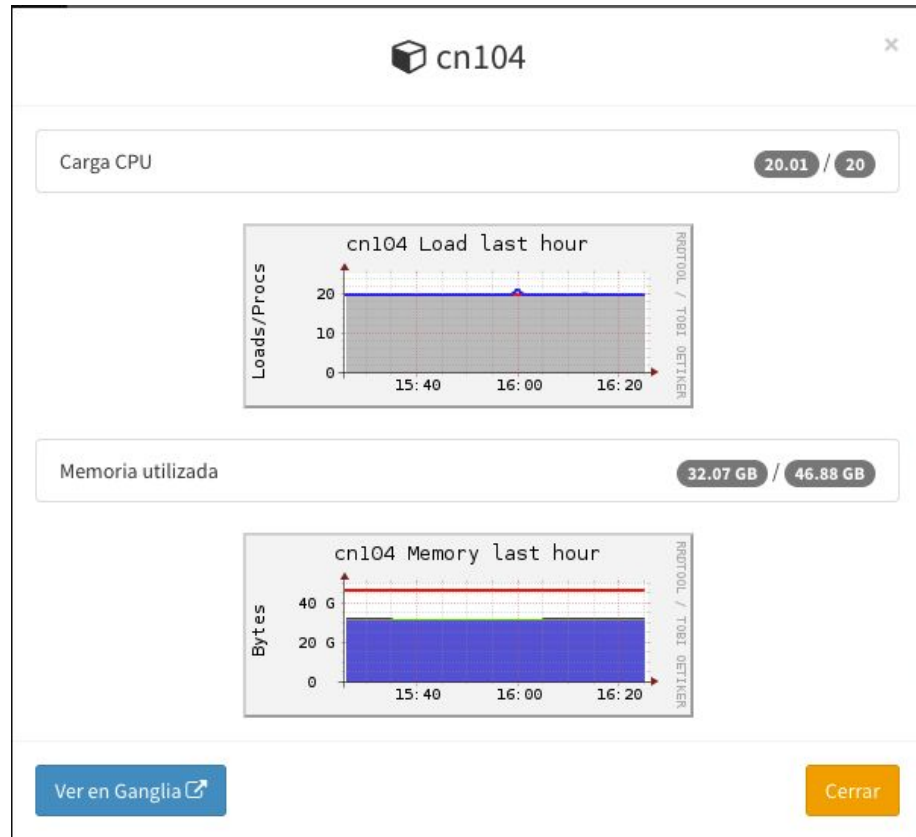
Cerrar

Ejemplos prácticos - Job OpenMP

```
#!/bin/bash
#SBATCH -J programa_omp
#SBATCH -p slims
#SBATCH -c 20
#SBATCH -o programa_omp_%j.out
#SBATCH -e programa_omp_%j.err
#SBATCH --mail-user=user@gmail.com
#SBATCH --mail-type=ALL
```

```
export OMP_NUM_THREADS=20
```

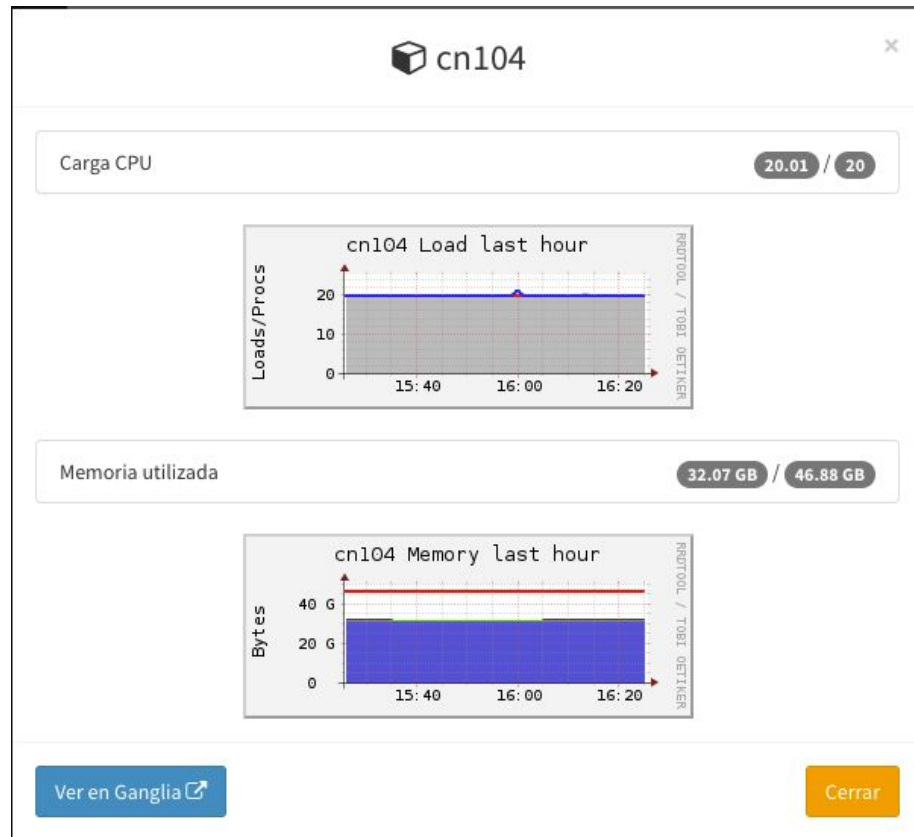
```
./programa_omp
```



Ejemplos prácticos - Job (Intel) MPI

```
#!/bin/bash
#SBATCH -J programa_mpi
#SBATCH -p slims
#SBATCH -n 120
#SBATCH -o programa_mpi_%j.out
#SBATCH -e programa_mpi_%j.err
#SBATCH --mail-user=user@gmail.com
#SBATCH --mail-type=ALL

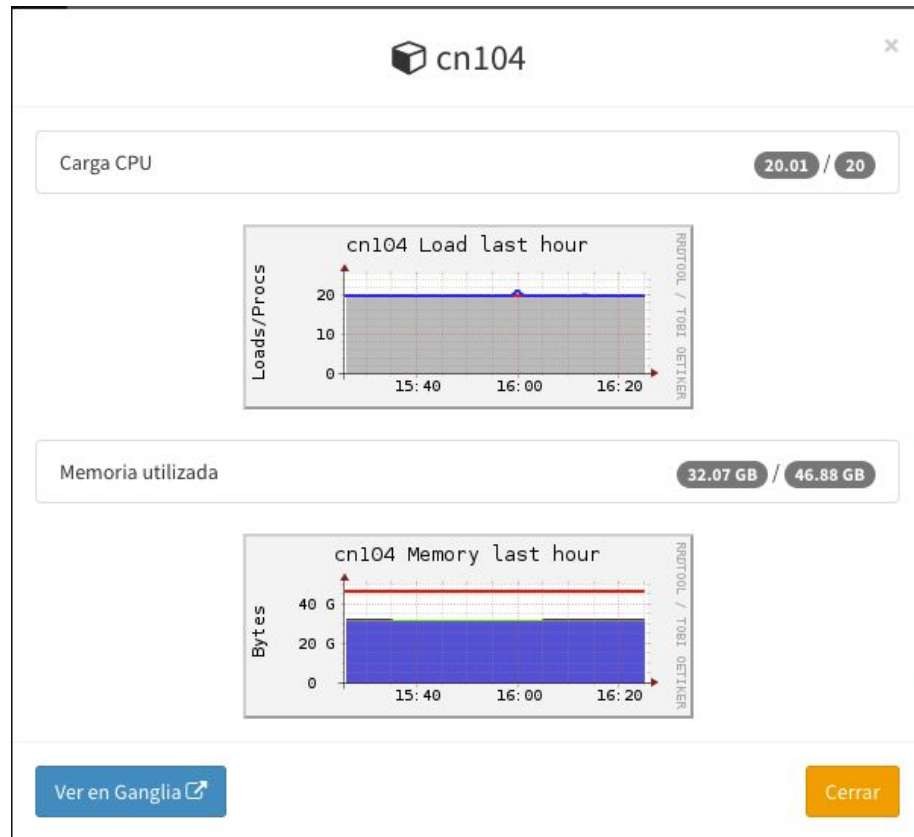
./programa_mpi
```



Ejemplos prácticos - Job (Intel) MPI

```
#!/bin/bash
#SBATCH -J programa_mpi
#SBATCH -p slims
#SBATCH -n 120
#SBATCH -o programa_mpi_%j.out
#SBATCH -e programa_mpi_%j.err
#SBATCH --mail-user=user@gmail.com
#SBATCH --mail-type=ALL

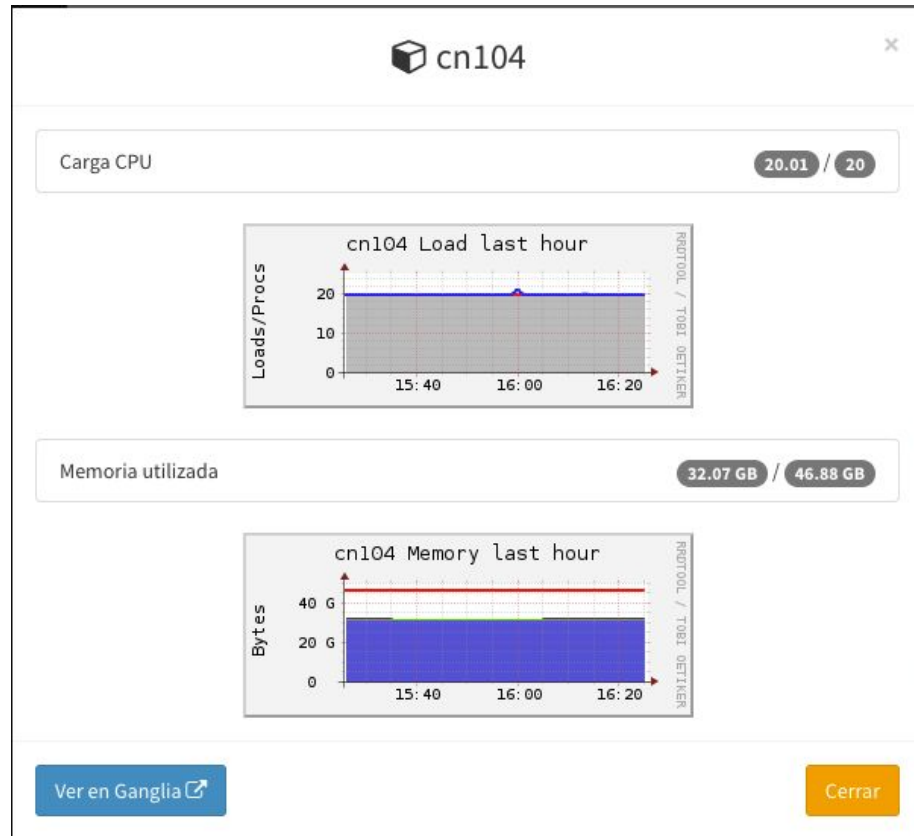
mpirun ./programa_mpi
```



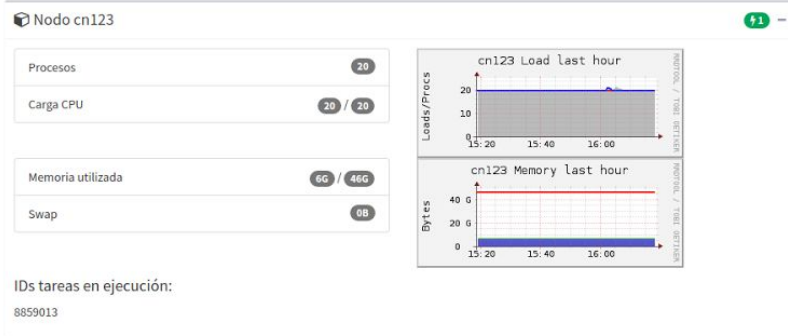
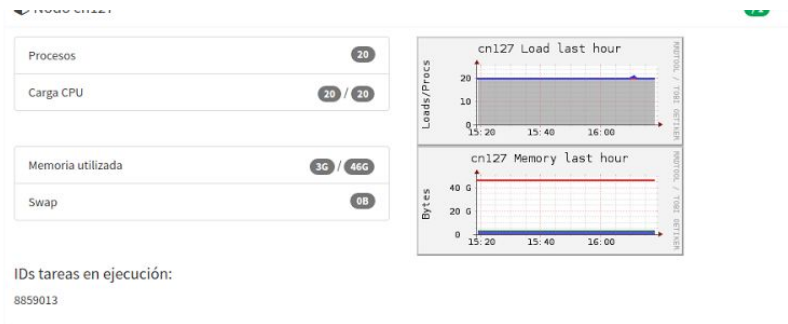
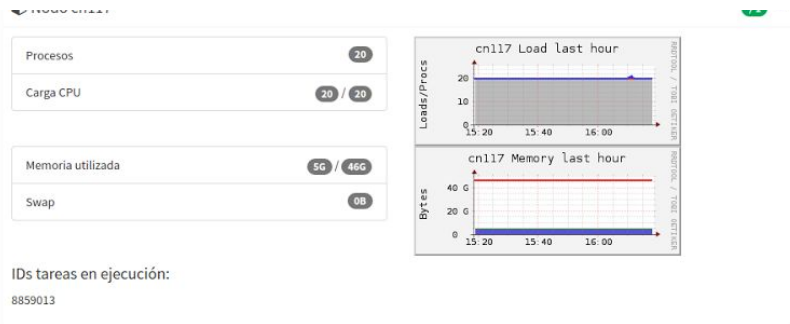
Ejemplos prácticos - Job (Intel) MPI

```
#!/bin/bash
#SBATCH -J programa_mpi
#SBATCH -p slims
#SBATCH -n 120
#SBATCH --ntasks-per-node=20
#SBATCH -o programa_mpi_%j.out
#SBATCH -e programa_mpi_%j.err
#SBATCH --mail-user=user@gmail.com
#SBATCH --mail-type=ALL
```

```
module load impi
srun programa_mpi
```



Ejemplos prácticos - Job (Intel) MPI

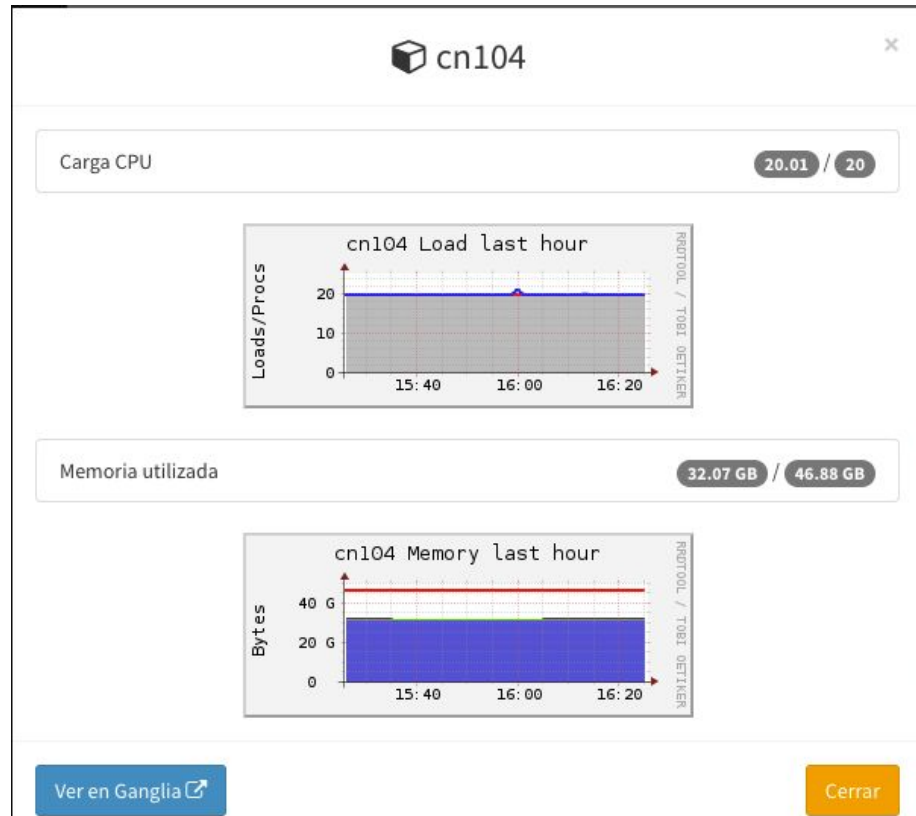


Ejemplos prácticos - Hybrid MPI-OpenMP

```
#!/bin/bash
#SBATCH -J hybrid
#SBATCH -p slims
#SBATCH -n 6
#SBATCH -c 20
#SBATCH --ntasks-per-node=1
#SBATCH -o programa_hybrid_%j.out
#SBATCH -e programa_hybrid_%j.err
#SBATCH --mail-user=user@gmail.com
#SBATCH --mail-type=ALL
```

```
export OMP_NUM_THREADS=20
```

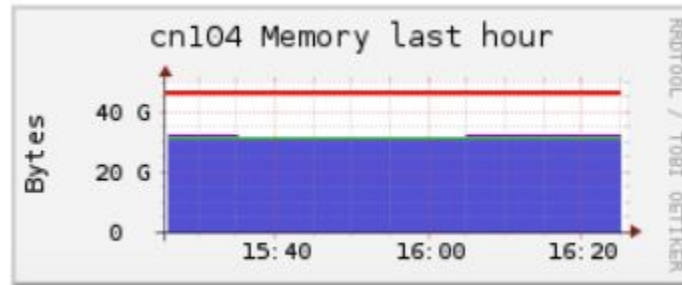
```
module load impi
srun ./programa_hybrid
```



cgroups

Límite en la memoria RAM:

- Por defecto se reserva 1GB de RAM por core reservado
- Si se excede se obtendrá el error: “Exceeded job memory limit”
- Reservar más RAM por *core* usado: `#SBATCH --mem-per-cpu=2400`
- Reservar más RAM por *job*: `#SBATCH --mem=48000`



Límites CPU:

- Se activa afinidad core/hilo

Ganglia - monitor.nlhpc.cl

```

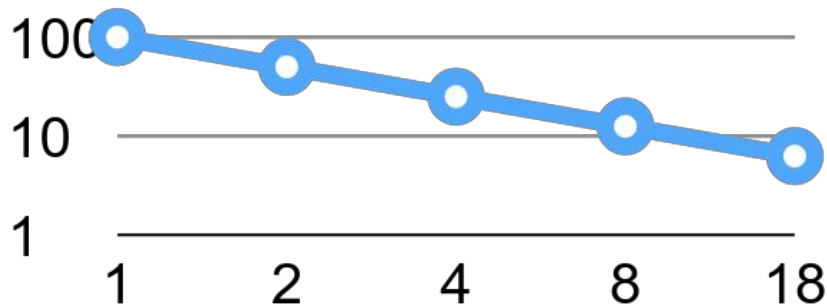
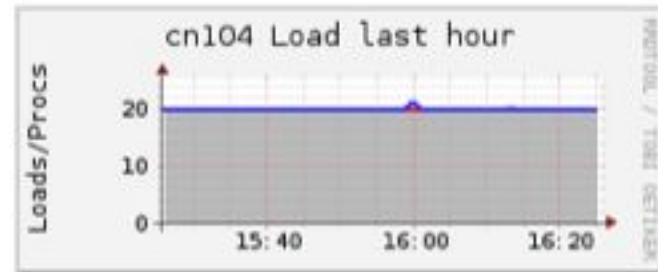
1 [|||||] 100.0% 6 [|||||] 100.0% 11 [|||||] 0.0% 16 [|||||] 0.0%
2 [|||||] 100.0% 7 [|||||] 100.0% 12 [|||||] 0.0% 17 [|||||] 0.0%
3 [|||||] 100.0% 8 [|||||] 100.0% 13 [|||||] 100.0% 18 [|||||] 100.0%
4 [|||||] 100.0% 9 [|||||] 100.0% 14 [|||||] 0.0% 19 [|||||] 100.0%
5 [|||||] 100.0% 10 [|||||] 0.0% 15 [|||||] 0.0% 20 [|||||] 100.0%
Mem [|||||] 9.59G/47.0G Tasks: 91, 66 thr: 33 running
Swp [|||||] 4K/23.6G Load average: 23.39 15.18 13.13
Uptime: 41 days, 01:18:21

```

PID	USER	PRI	NI	VRT	RES	SHR	S	CPU%	MEM%	TIME+	Command
31047	sgutierre	20	0	181M	11952	5020	R	100	0.0	68h31:23	/home/apps/namd_intel/2.10/bin/namd2 lgfl-2efv-ssrA_cg_12_smd_period270_run3.conf
27179	bvatter	20	0	83708	29368	1600	R	100	0.1	4h40:44	./run temp/config_34.cfg
28214	bvatter	20	0	83536	30140	2480	R	100	0.1	1h02:39	./run temp/config_42.cfg
26716	bvatter	20	0	83708	30012	1624	R	100	0.1	6h11:50	./run temp/config_32.cfg
27525	bvatter	20	0	83536	28796	1348	R	100	0.1	3h18:53	./run temp/config_36.cfg
27906	bvatter	20	0	83536	29840	2480	R	100	0.1	2h05:47	./run temp/config_38.cfg
27935	bvatter	20	0	83536	28896	2480	R	100	0.1	2h02:38	./run temp/config_40.cfg
26694	bvatter	20	0	83708	29880	1624	R	100	0.1	6h15:08	./run temp/config_31.cfg
51046	sgutierre	20	0	290M	16896	5964	R	100	0.0	68h30:30	/home/apps/namd_intel/2.10/bin/namd2 lgfl-2efv-ssrA_cg_12_smd_period270_run3.conf
28257	bvatter	20	0	83536	30288	2480	R	100	0.1	50:21.43	./run temp/config_44.cfg
51048	sgutierre	20	0	280M	12016	5244	R	99.4	0.0	68h31:19	/home/apps/namd_intel/2.10/bin/namd2 lgfl-2efv-ssrA_cg_12_smd_period270_run3.conf
51049	sgutierre	20	0	180M	11712	4988	R	99.4	0.0	68h30:58	/home/apps/namd_intel/2.10/bin/namd2 lgfl-2efv-ssrA_cg_12_smd_period270_run3.conf
28488	root	20	0	7268	104	0	R	5.3	0.0	0:02.84	/usr/bin/stress -c 20
28498	root	20	0	7268	104	0	R	5.3	0.0	0:02.84	/usr/bin/stress -c 20
28494	root	20	0	7268	104	0	R	5.3	0.0	0:02.84	/usr/bin/stress -c 20
28499	root	20	0	7268	104	0	R	5.3	0.0	0:02.84	/usr/bin/stress -c 20
28497	root	20	0	7268	104	0	R	5.3	0.0	0:02.84	/usr/bin/stress -c 20
28506	root	20	0	7268	104	0	R	5.3	0.0	0:02.84	/usr/bin/stress -c 20
28507	root	20	0	7268	104	0	R	5.3	0.0	0:02.83	/usr/bin/stress -c 20
28504	root	20	0	7268	104	0	R	5.3	0.0	0:02.83	/usr/bin/stress -c 20
28501	root	20	0	7268	104	0	R	5.3	0.0	0:02.83	/usr/bin/stress -c 20
28500	root	20	0	7268	104	0	R	5.3	0.0	0:02.83	/usr/bin/stress -c 20
28496	root	20	0	7268	104	0	R	5.3	0.0	0:02.83	/usr/bin/stress -c 20
28490	root	20	0	7268	104	0	R	5.3	0.0	0:02.83	/usr/bin/stress -c 20
28502	root	20	0	7268	104	0	R	4.6	0.0	0:02.83	/usr/bin/stress -c 20
28503	root	20	0	7268	104	0	R	4.6	0.0	0:02.83	/usr/bin/stress -c 20
28495	root	20	0	7268	104	0	R	4.6	0.0	0:02.83	/usr/bin/stress -c 20
28505	root	20	0	7268	104	0	R	4.6	0.0	0:02.83	/usr/bin/stress -c 20
28489	root	20	0	7268	104	0	R	4.6	0.0	0:02.83	/usr/bin/stress -c 20
28492	root	20	0	7268	104	0	R	4.6	0.0	0:02.83	/usr/bin/stress -c 20
28491	root	20	0	7268	104	0	R	4.6	0.0	0:02.83	/usr/bin/stress -c 20
28493	root	20	0	7268	104	0	R	4.6	0.0	0:02.83	/usr/bin/stress -c 20
28545											

Profiling de los jobs

- Los usuarios deben de comprobar que sus trabajos usan los recursos adecuadamente:
 - ¿Todos los cores están bien usados?
- ¿Escala adecuadamente?



Ganglia - monitor.nlhpc.cl

```
1 [ 0.0%] 6 [ 100.0%] 11 [ 100.0%] 16 [ 100.0%]
2 [ 100.0%] 7 [ 100.0%] 12 [ 100.0%] 17 [ 100.0%]
3 [ 100.0%] 8 [ 100.0%] 13 [ 100.0%] 18 [ 100.0%]
4 [ 100.0%] 9 [ 100.0%] 14 [ 100.0%] 19 [ 100.0%]
5 [ 100.0%] 10 [ 100.0%] 15 [ 100.0%] 20 [ 100.0%]
Mem[ 9.54G/47.0G] Tasks: 70, 103 thr: 21 running
Swp[ 0K/23.6G] Load average: 19.00 19.20 19.35
Uptime: 16 days, 20:31:20

PID USER PRI NI VIRT RES SHR S CPU MEM% TIME+ Command
16459 sgutierre 20 0 222M 9604 4388 R 100.0 0.0 60h13:52 /home/apps/namd_intel/2.10/bin//namd2 lgfl-2efv-ssrA_cg_12_smd_period180_run3.conf
18576 sgutierre 20 0 340M 17180 8948 R 100.0 0.0 65h44:52 /home/apps/namd_intel/2.10/bin//namd2 lgfl-2efv-ssrA_cg_12_smd_period90_run1.conf
18582 sgutierre 20 0 336M 9248 4592 R 100.0 0.0 65h46:20 /home/apps/namd_intel/2.10/bin//namd2 lgfl-2efv-ssrA_cg_12_smd_period90_run1.conf
16455 sgutierre 20 0 222M 10920 5508 R 100.0 0.0 60h13:52 /home/apps/namd_intel/2.10/bin//namd2 lgfl-2efv-ssrA_cg_12_smd_period180_run3.conf
18579 sgutierre 20 0 230M 8968 4404 R 100.0 0.0 65h46:17 /home/apps/namd_intel/2.10/bin//namd2 lgfl-2efv-ssrA_cg_12_smd_period90_run1.conf
16457 sgutierre 20 0 222M 9556 4432 R 100.0 0.0 60h13:56 /home/apps/namd_intel/2.10/bin//namd2 lgfl-2efv-ssrA_cg_12_smd_period180_run3.conf
16458 sgutierre 20 0 327M 10008 4868 R 100.0 0.0 60h13:56 /home/apps/namd_intel/2.10/bin//namd2 lgfl-2efv-ssrA_cg_12_smd_period180_run3.conf
18584 sgutierre 20 0 336M 10556 5760 R 100.0 0.0 65h46:25 /home/apps/namd_intel/2.10/bin//namd2 lgfl-2efv-ssrA_cg_12_smd_period90_run1.conf
18578 sgutierre 20 0 336M 12972 7796 R 100.0 0.0 65h46:20 /home/apps/namd_intel/2.10/bin//namd2 lgfl-2efv-ssrA_cg_12_smd_period90_run1.conf
18583 sgutierre 20 0 230M 8792 4240 R 100.0 0.0 65h46:22 /home/apps/namd_intel/2.10/bin//namd2 lgfl-2efv-ssrA_cg_12_smd_period90_run1.conf
16461 sgutierre 20 0 222M 9536 4384 R 100.0 0.0 60h13:55 /home/apps/namd_intel/2.10/bin//namd2 lgfl-2efv-ssrA_cg_12_smd_period180_run3.conf
16462 sgutierre 20 0 328M 11464 5984 R 100.0 0.0 60h14:02 /home/apps/namd_intel/2.10/bin//namd2 lgfl-2efv-ssrA_cg_12_smd_period180_run3.conf
16454 sgutierre 20 0 332M 16260 7172 R 100.0 0.0 60h11:46 /home/apps/namd_intel/2.10/bin//namd2 lgfl-2efv-ssrA_cg_12_smd_period180_run3.conf
18577 sgutierre 20 0 231M 9756 4932 R 100.0 0.0 65h45:53 /home/apps/namd_intel/2.10/bin//namd2 lgfl-2efv-ssrA_cg_12_smd_period90_run1.conf
18580 sgutierre 20 0 336M 9264 4600 R 100.0 0.0 65h46:19 /home/apps/namd_intel/2.10/bin//namd2 lgfl-2efv-ssrA_cg_12_smd_period90_run1.conf
16456 sgutierre 20 0 327M 10096 4848 R 100.0 0.0 60h13:32 /home/apps/namd_intel/2.10/bin//namd2 lgfl-2efv-ssrA_cg_12_smd_period180_run3.conf
16460 sgutierre 20 0 222M 9720 4520 R 100.0 0.0 60h13:47 /home/apps/namd_intel/2.10/bin//namd2 lgfl-2efv-ssrA_cg_12_smd_period180_run3.conf
18585 sgutierre 20 0 230M 9580 5080 R 100.0 0.0 65h46:26 /home/apps/namd_intel/2.10/bin//namd2 lgfl-2efv-ssrA_cg_12_smd_period90_run1.conf
18581 sgutierre 20 0 231M 8852 4236 R 100.0 0.0 65h46:15 /home/apps/namd_intel/2.10/bin//namd2 lgfl-2efv-ssrA_cg_12_smd_period90_run1.conf
8844 root 20 0 122M 3172 1492 R 0.7 0.0 0:01.60 http
868 root 20 0 19448 1404 964 S 0.7 0.0 4:08.87 /usr/sbin/irqbalance --foreground
64624 root 20 0 152M 1884 1108 S 0.0 0.0 0:00.33 /usr/sbin/gmond
1555 root 20 0 540M 16448 5776 S 0.0 0.0 1:46.86 /usr/bin/python -Es /usr/sbin/tuned -l -P
16453 root 20 0 354M 3976 2836 S 0.0 0.0 0:12.15 slurmstepd: [8700504.0]
16449 root 20 0 354M 3976 2836 S 0.0 0.0 0:31.61 slurmstepd: [8700504.0]
54939 root 20 0 3753M 22636 1952 S 0.0 0.0 1:32.23 /usr/sbin/slurmd
16444 sgutierre 20 0 363M 4760 1892 S 0.0 0.0 0:10.81 srun namd2 lgfl-2efv-ssrA_cg_12_smd_period180_run3.conf
18571 root 20 0 419M 4000 2836 S 0.0 0.0 0:26.11 slurmstepd: [8704482.0]
1560 root 20 0 170M 10252 4792 S 0.0 0.0 20:43.76 /sbin/amsHelper -f
1588 root 20 0 540M 16448 5776 S 0.0 0.0 1:46.72 /usr/bin/python -Es /usr/sbin/tuned -l -P
8808 root 20 0 147M 5264 3960 S 0.0 0.0 0:00.02 sshd: root@pts/0
1927 munge 20 0 217M 2404 1204 S 0.0 0.0 0:42.76 /usr/sbin/munged
1709 root 20 0 349M 3592 684 S 0.0 0.0 1:43.32 hpaasmited -f /dev/hpilo
18567 sgutierre 20 0 298M 4740 1884 S 0.0 0.0 0:05.98 srun namd2 lgfl-2efv-ssrA_cg_12_smd_period90_run1.conf
18563 sgutierre 20 0 298M 4740 1884 S 0.0 0.0 0:06.00 srun namd2 lgfl-2efv-ssrA_cg_12_smd_period90_run1.conf
16440 sgutierre 20 0 363M 4760 1892 S 0.0 0.0 0:10.83 srun namd2 lgfl-2efv-ssrA_cg_12_smd_period180_run3.conf
1 root 20 0 187M 4752 2432 S 0.0 0.0 0:19.63 /usr/lib/systemd/systemd --switched-root --system --deserialize 21
544 root 20 0 61392 24612 24296 S 0.0 0.0 0:04.43 /usr/lib/systemd/systemd-journald
570 root 20 0 43580 1824 1288 S 0.0 0.0 2:04.99 /usr/lib/systemd/systemd-udev
851 root 16 - 55416 1852 1396 S 0.0 0.0 0:00.32 /sbin/auditd -n
848 root 16 - 55416 1852 1396 S 0.0 0.0 0:07.57 /sbin/auditd -n
```

Comandos útiles - Consultar espacio usado

```
[root@leftraru1 ~]# lfs quota -hu $USER /home
```

Disk quotas for user USER (uid 1000):

Filesystem	used	quota	limit	grace	files	quota	limit	grace
/home	4.994G	0k	80G	-	57294	0	0	-

Comandos útiles - Consultar información job

```
[root@leftraru1 ~]# scontrol show job 9160565
```

```
JobId=9160565 JobName=w.fepc-f-cnt-oo.m2
```

```
  UserId=worellana(1194) GroupId=fisica_unab(1141) MCS_label=N/A
```

```
  Priority=109951 Nice=0 Account=unab QOS=120
```

```
  JobState=RUNNING Reason=None Dependency=(null)
```

```
  Requeue=0 Restarts=0 BatchFlag=1 Reboot=0 ExitCode=0:0
```

```
  RunTime=04:28:03 TimeLimit=3-00:00:00 TimeMin=N/A
```

```
  SubmitTime=2017-10-09T11:25:36 EligibleTime=2017-10-09T11:25:36
```

```
  StartTime=2017-10-09T15:04:40 EndTime=2017-10-12T15:04:40
```

```
  Deadline=N/A
```

```
  PreemptTime=None SuspendTime=None SecsPreSuspend=0
```

```
  Partition=slims AllocNode:Sid=leftraru4:52235
```

```
  ReqNodeList=(null) ExcNodeList=(null)
```

Comandos útiles - Consultar información job

```
NodeList=cn[056,058]
BatchHost=cn056
NumNodes=2 NumCPUs=40 NumTasks=40 CPUs/Task=1
ReqB:S:C:T=0:0:*:*
TRES=cpu=40,mem=96000M,node=2
Socks/Node=* NtasksPerN:B:S:C=20:0:*:* CoreSpec=*
MinCPUsNode=20 MinMemoryCPU=2400M MinTmpDiskNode=0
Features=(null) DelayBoot=00:00:00
Gres=(null) Reservation=(null)
OverSubscribe=OK Contiguous=0 Licenses=(null) Network=(null)
Command=/home/Espresso/FePc-F-CNT-O2/Final/m2/run.sh
WorkDir=/home/Espresso/FePc-F-CNT-O2/Final/m2
StdErr=/home/Espresso/FePc-F-CNT-O2/Final/m2/pw-run.err
StdIn=/dev/null
StdOut=/home/Espresso/FePc-F-CNT-O2/Final/m2/pw-run.out
```

Problemas comunes

- Usuarios piden:
 - Reducir tiempos de espera en las colas
 - Más recursos (CPU, espacio en disco, RAM, GPUs)
- Recursos sobre- y sub- utilizados:
 - Usar partición *debug*
 - Usar partición *levque*
- Usar más o menos memoria RAM
- Excesivo uso de Lustre:
 - *Metadata*
 - Escritura/Lectura de disco -> Uso de disco (RAM) local
- Instalación NHC

Solicitud de cuentas

Solicitud de cuentas:

visitar: <http://www.nlhpc.cl> (Servicios > Servicios para la Academia > Formulario)

ó solicitar información a info@nlhpc.cl