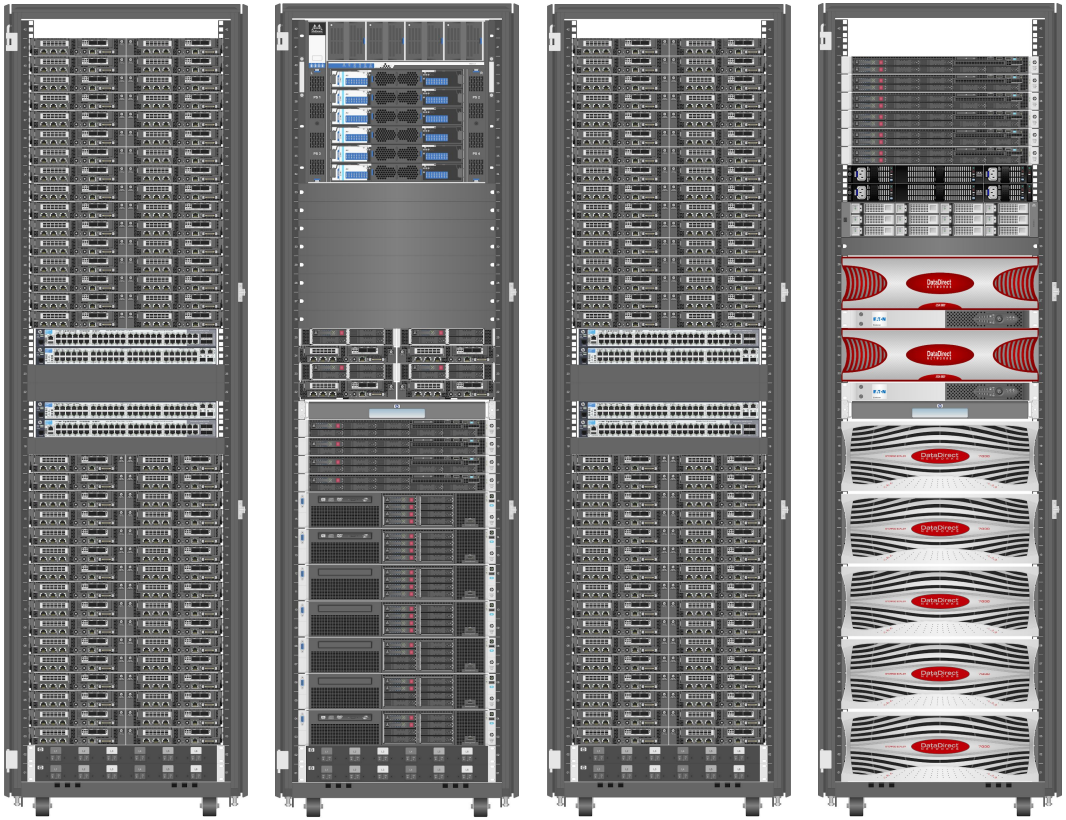
NLHPC
National Laboratory for High Performance Computing Chile

Introducción al uso de la infraestructura del NLHPC

Infraestructura NLHPC



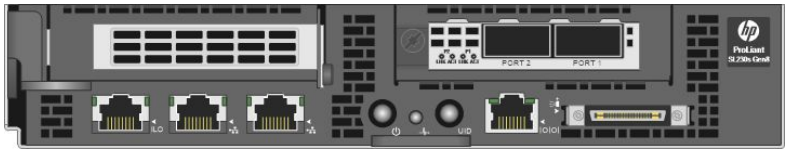
- 266 TFlops
- 5236 cores
- 191 nodos
- +200 TB almacenamiento DDN Lustre
- Infiniband FDR 56Gbps

NLHPC
National Laboratory for High Performance Computing
Chile

Nodos de cómputo Leftraru

1 tipos de nodo de cómputo:

Slim



Nodo Slim (cn)

- 132 nodos
- 2 CPUS Xeon E5-2660 v2, 10 cores c.u.
- 48/64 GB RAM DIMM DDR3

NLHPC
National Laboratory
for High Performance
Computing
Chile

Nodos de cómputo Guacolda

3 tipos de nodo de cómputo:

General

Largemem

Nodo General (sn)

- 48 nodos
- 2 CPUS Xeon Gold 6152, 22 cores c.u.
- 192 GB RAM DIMM DDR4

Nodo Largemem (fn)

- 9 nodos
- 2 CPUS Xeon Gold 6152, 22 cores c.u.
- 772 GB RAM DIMM DDR4

NLHPC
National Laboratory for High Performance Computing Chile

Nodos de cómputo Guacolda

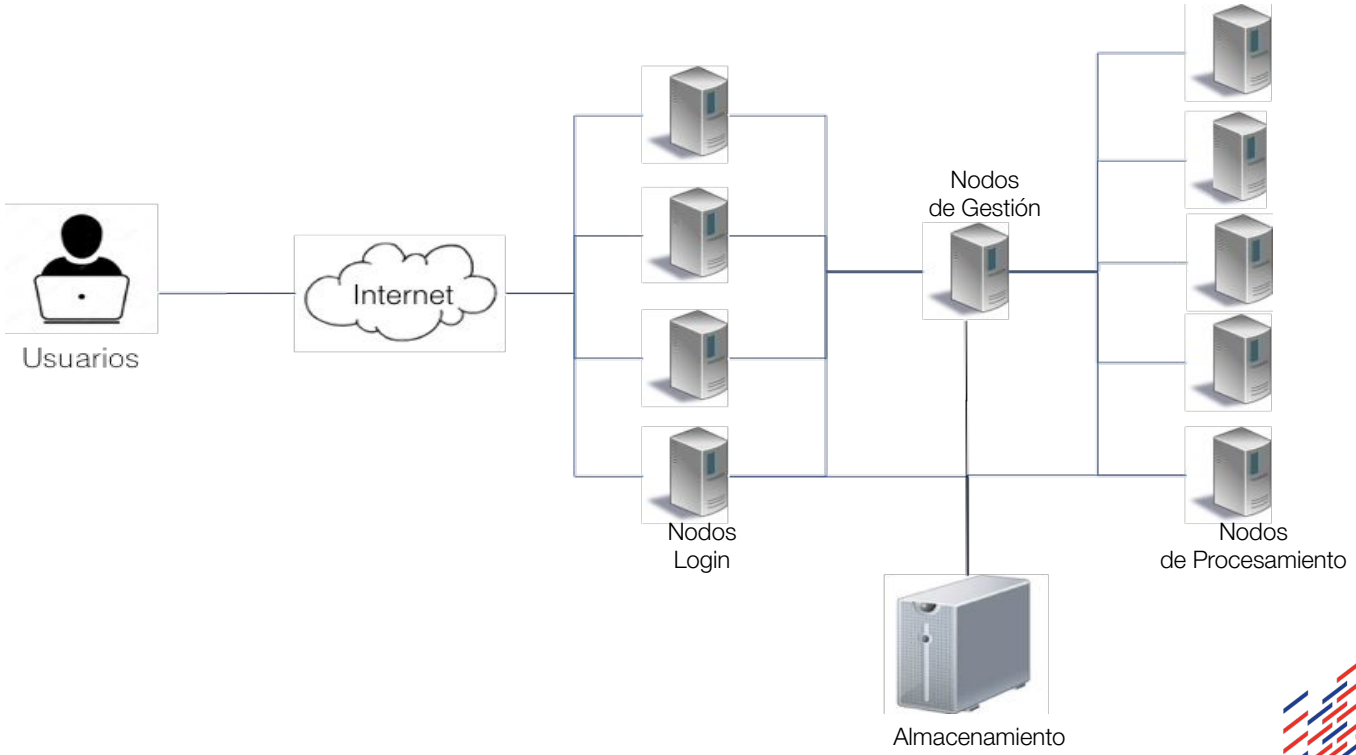
GPU

Nodo GPU (gn)

- 2 nodos
- 2 CPUS Xeon Gold 6152, 22 cores c.u.
- 2 NVIDIA Volta V100 cada nodo
  - 16GB
  - 5120 CUDA cores cada una
- **192 GB** RAM DIMM DDR4

NLHPC
National Laboratory for High Performance Computing Chile

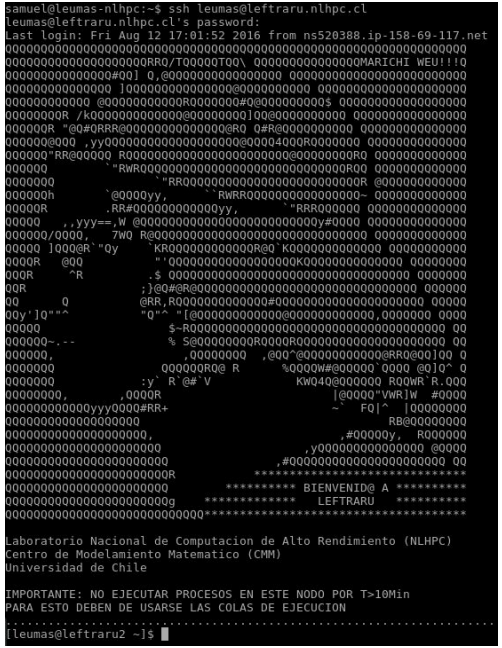
Uso de Guacolda-Leftraru



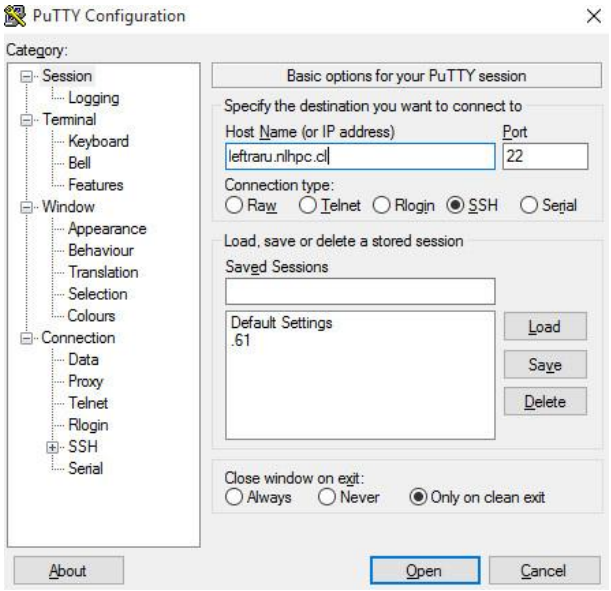
NLHPC
National Laboratory
for High Performance
Computing
Chile

Iniciar Sesión ssh en leftraru/Guacolda

Linux, OSX:
ssh usuario@leftraru.nlhpc.cl



Windows:



NLHPC
National Laboratory for High Performance Computing
Chile

Gestor de recursos SLURM

- Administra recursos computacionales como CPU y RAM.
- Gestiona las colas de ejecución en clusters.
- Reserva recursos compartidos.

Particiones Leftraru

- **slims**: partición por defecto, 128 nodos slims, 4 fats (2640 cores).
- **debug**: 4 nodos logins (80 cores).
- **general**: 48 nodos (2112 cores).
- **largemem**: 9 nodos (396 cores). Para tareas con más de 192 GB.
- **gpus**: 2 nodos con 4 tarjetas Nvidia Tesla V100 (88 cores y 20480 GPU cores)
- Uso de nodos logins/debug:
  - Restringido para ejecución (30 minutos)
  - Válido para pruebas, compilaciones y movimiento de datos
- Resto de particiones 30 días de ejecución
  - Uso de *checkpoints*.

slurm
workload manager

NLHPC
National Laboratory for High Performance Computing Chile

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 leftraru[1-4]
largemem     up   infinite      1   resv syntagma001
```

Enviar trabajos en SLURM

```
[usuario@leftraru1 ~]$ srun hostname
cn028
[usuario@leftraru1 ~]$ sbatch <job_script>
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
```

1 proceso

salida comando

cancelar trabajo

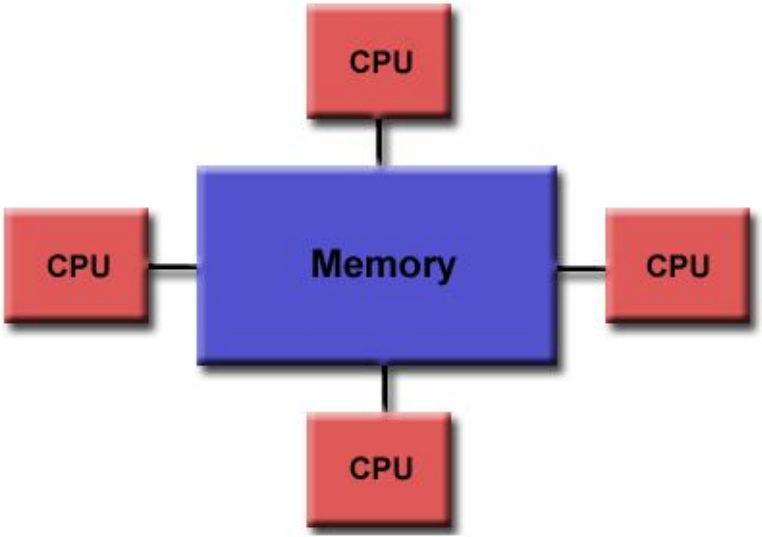
slurm workload manager

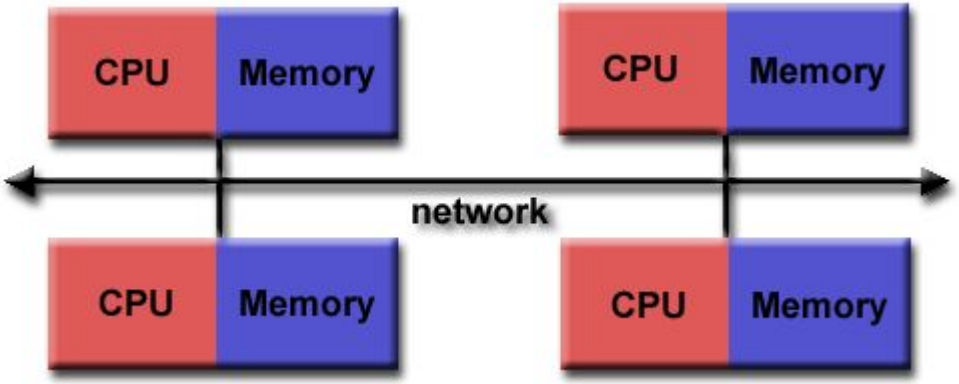
NLHPC National Laboratory for High Performance Computing Chile

Directivas en SLURM

- -J example
- -p slims
- -n 20
- -c 20
- --ntasks-per-node=20
- -o file
- -e file
- --mail-user=user@gmail.com
- --mail-type=ALL

Arquitectura de Memoria Compartida y Distribuida

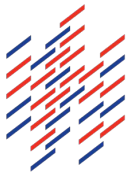




NLHPC
National Laboratory for High Performance Computing Chile

Ejercicio 1

- Lanza en la partición *slims* de Leftraru ejecutando el comando **hostname** con **srun**:
  - Con 1 único proceso.
  - Con 2 procesos iguales.
  - Con 2 procesos iguales en distintos nodos.
  - Lanzando un proceso que tenga dos *cores* reservados.
- Anota cada uno de los comandos ejecutados y la salida de los mismos.
- En la partición *slims* ¿cuántos *cores* puedes reservar por proceso? ¿Por qué es ese número? ¿Es igual en la partición *general*? ¿Qué ocurre si reservas más *cores* de los disponibles?
- ¿Qué ocurre si no especificas la partición al ejecutar?



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

NLHPC
National Laboratory for High Performance Computing Chile

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)
```

NLHPC
National Laboratory for High Performance Computing Chile

Ejercicio 2

- Crea un *script* de ejecución para lanzarlo con **sbatch**, con las siguientes consideraciones:
  - La partición a lanzar es *slims*.
  - Reserva un único *core*.
  - Ejecuta el comando **stress -c 1**.
- Anota el *script* que has usado.
- El comando **stress** sirve para poner a prueba los distintos componentes de un computador. En este ejemplo estamos pidiendo usar una *CPU* (al 100%) durante un tiempo ilimitado. Ya que no se le ha especificado al comando un tiempo de término, en principio, la tarea no debiera terminar nunca. En relación a esto:
  - ¿Cuánto tiempo estará la tarea corriendo hasta que sea cancelada por el sistema? ¿Qué comando debes de usar para obtener información sobre un trabajo que ya está corriendo?
  - ¿Qué comando puedes usar para obtener información acerca de las particiones?
  - Cancela la tarea. ¿Qué comando/s has usado para cancelarla?

slurm
workload manager

NLHPC
National Laboratory for High Performance Computing Chile

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

slurm workload manager

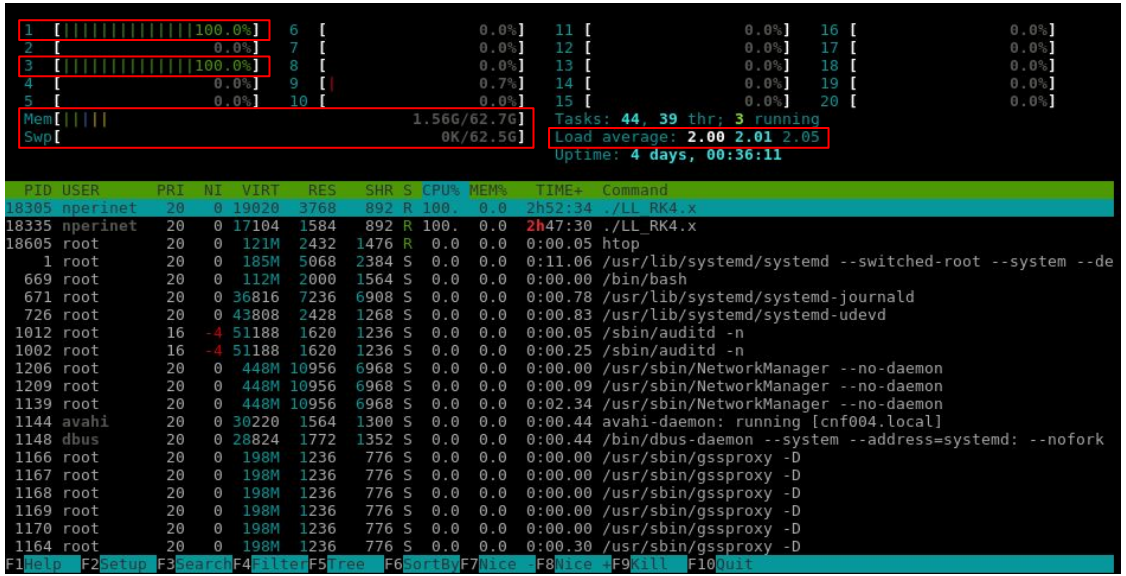
NLHPC National Laboratory for High Performance Computing Chile

Ganglia - monitor.nlhpc.cl

NLHPC
National Laboratory for High Performance Computing Chile

Monitorear Job - *htop*

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



NLHPC
National Laboratory for High Performance Computing Chile

Ejercicio 3

- Crea un *script* de ejecución para lanzarlo con **sbatch**, con las siguientes consideraciones:
  - La partición a lanzar es *slims*.
  - Reserva un nodo completo.
  - Ejecuta el comando **stress -c 40 -t 15m**.
- Anota el *script* que has usado.
- ¿Cuántos recursos de *CPU* estás usando? Muestra una captura de pantalla que muestre gráficamente lo que ocurre con la carga de *CPU* en el nodo donde se está ejecutando la tarea.
- Entra al nodo donde se está ejecutando la tarea, ejecuta el comando **htop** e interpreta lo que está pasando.
- ¿Cuál sería la manera correcta de lanzar la ejecución para usar todos los cores disponibles en un nodo? Inserta en el informe una captura de pantalla que muestre gráficamente lo que ocurre con la carga de CPU en el nodo donde se está ejecutando la tarea.

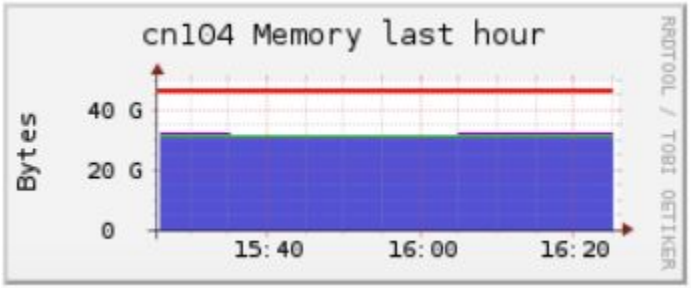
slurm
workload manager

NLHPC
National Laboratory
for High Performance
Computing
Chile

Reserva Memoria RAM

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

NLHPC
National Laboratory for High Performance Computing Chile

Ejercicio 4

- Crea un *script* de ejecución para lanzarlo con **sbatch**, con las siguientes consideraciones:
  - La partición a lanzar es *slims*.
  - Reserva un único *core*.
  - Indica que envíe un correo cuando el *job* cambie de estado.
  - Ejecuta el comando **stress -m 1 --vm-bytes 2048M -t 15m**.
- Anota el *script* que has usado.
- ¿Qué está ocurriendo con la ejecución? ¿a qué se debe?
- Relanza la tarea modificando el *script* para poder ejecutar la tarea. Anota el *script*.
- ¿Cuántos recursos de RAM estás usando? Inserta en el informe una captura de pantalla que muestre gráficamente lo que ocurre con la carga de RAM en el nodo donde se está ejecutando la tarea.

slurm
workload manager

NLHPC
National Laboratory
for High Performance
Computing
Chile

Sistema de Módulos

- Permite tener diferentes aplicaciones y versiones de estos en un mismo sistema operativo.
- En el NLHPC usamos **Lmod** (https://github.com/TACC/Lmod).
- El actual sistema de módulos está disponible para las distintas arquitecturas de procesador (AVX512,AVX,SSE4.2)

NLHPC
National Laboratory
for High Performance
Computing
Chile

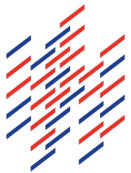
Lmod: Buscar

```
eguerra@leftraru1:/home/eguerra$ ml spider Python

----------------------------------------------------------------------------------------------------------------------------------------------------------------
  Python:
----------------------------------------------------------------------------------------------------------------------------------------------------------------
    Description:
      Python is a programming language that lets you work more quickly and integrate your systems more effectively.

     Versions:
       Python/2.7.15
       Python/3.7.2
       Python/3.7.3
    Other possible modules matches:
       Biopython  IPython  protobuf-python

----------------------------------------------------------------------------------------------------------------------------------------------------------------
  To find other possible module matches execute:

      $ module -r spider '.*Python.*'

----------------------------------------------------------------------------------------------------------------------------------------------------------------
  For detailed information about a specific "Python" module (including how to load the modules) use the module's full name.
  For example:

     $ module spider Python/3.7.3
```

NLHPC
National Laboratory for High Performance Computing Chile

Lmod: Cargar distintas versiones

```
eguerra@leftraru1:/home/eguerra$ ml Python/3.7.3
eguerra@leftraru1:/home/eguerra$ ml

Currently Loaded Modules:
  1) GCCcore/8.2.0                    4) impi/2019.2.187   7) intel/2019b  10) libreadline/8.0  13) SQLite/3.27.1  16) libffi/3.2.1
  2) icc/2019.2.187-GCC-8.2.0-2.31.1  5) imkl/2019.2.187   8) bzip2/1.0.6  11) ncurses/6.1      14) XZ/5.2.4       17) Python/3.7.3
  3) ifort/2019.2.187-GCC-8.2.0-2.31.1  6) binutils/2.32    9) zlib/1.2.11  12) Tcl/8.6.9        15) GMP/6.1.2


eguerra@leftraru1:/home/eguerra$ python -V
Python 3.7.3


eguerra@leftraru1:/home/eguerra$ ml Python/2.7.15

The following have been reloaded with a version change:
  1) Python/3.7.3 => Python/2.7.15

eguerra@leftraru1:/home/eguerra$ ml

Currently Loaded Modules:
  1) GCCcore/8.2.0                    4) impi/2019.2.187   7) intel/2019b  10) libreadline/8.0  13) SQLite/3.27.1  16) libffi/3.2.1
  2) icc/2019.2.187-GCC-8.2.0-2.31.1  5) imkl/2019.2.187   8) bzip2/1.0.6  11) ncurses/6.1      14) XZ/5.2.4       17) Python/2.7.15
  3) ifort/2019.2.187-GCC-8.2.0-2.31.1  6) binutils/2.32    9) zlib/1.2.11  12) Tcl/8.6.9        15) GMP/6.1.2


eguerra@leftraru1:/home/eguerra$ python -V
Python 2.7.15
```



NLHPC
National Laboratory
for High Performance
Computing
Chile

Ejercicio 5

- Descarga el siguiente código: n-queens-problem-3.py en tu directorio de trabajo.
- Crea un *script* de ejecución para lanzarlo con **sbatch**, con las siguientes consideraciones:
  - La partición a lanzar es *slims*.
  - Cada *job* reserva un único *core*.
  - Supón que cada proceso ocupa 2400MB de memoria RAM.
  - Ejecuta el código con la versión de Python 3 más actualizada del sistema.
- Anota el *script* que has usado.
- Anota la salida de la ejecución.

NLHPC
National Laboratory for High Performance Computing Chile

Eficiencia Computacional

- ¿Qué significa que mi programa escale?
  - Este concepto hace referencia a cómo se comporta un programa paralelo (más de una cpu) al aumentar el número de procesadores.
- Se dice que un sistema es escalable para un determinado rango de procesadores [1 … n], si la eficiencia E(n) del sistema se mantiene constante y en todo momento por encima de un factor 0.5
- Normalmente todos los sistemas tienen un determinado número de procesadores a partir del cual la eficiencia empieza a disminuir o forma más o menos brusca
- ¿Por qué es importante conocer como escala mi programa?
  - En nuestra infraestructura, históricamente los usuarios suelen usar el máximo de cores que pueden solicitar. No obstante, una ejecución por ejemplo de 120 cores no tiene porqué ser dos veces más rápida que una de 60 cores

NLHPC
National Laboratory for High Performance Computing Chile

Eficiencia Computacional - Speedup y Eficiencia

- Actualmente, para la solicitud de cuentas de investigación (superior a 88 cores) solicitamos un escalamiento de la aplicación con estas 2 métricas: Speedup y Eficiencia
- El speedup es la principal métrica que nos indica la ganancia que obtenemos mediante la paralelización:

$$Speedup = \frac{Toriginal}{Tmejora}$$

- La eficiencia es la métrica del uso de los recursos computacionales requeridos por la ejecución de un algoritmo en función del tamaño de las entradas. Se define como el cociente entre el speedup y el número de procesos en los que ejecutamos la aplicación

NLHPC
National Laboratory for High Performance Computing Chile

Ejemplo de escalamiento

- El tiempo original de una aplicación que necesitamos escalar utilizando 1 core es de 5 segundos. Para calcular el speedup vamos incrementando los cores de acuerdo a la fórmula:

$$(1\ \text{core})\ \frac{5}{5} = \mathbf{1} \qquad (2\ \text{cores})\ \frac{5}{2.5} = \mathbf{2} \qquad (4\ \text{cores})\ \frac{5}{2} = \mathbf{2.5}$$

$$(6\ \textit{cores})\ \frac{5}{1.66} = \mathbf{3.01} \qquad (8\ \text{cores})\ \frac{5}{1.25} = \mathbf{4} \qquad (10\ \text{cores})\ \frac{5}{1.2} = \mathbf{4.16}$$

$$(12\ \text{cores})\ \frac{5}{1.18} = \mathbf{4.23} \qquad (14\ \text{cores})\ \frac{5}{1.18} = \mathbf{4.23} \qquad (16\ \text{cores})\ \frac{5}{1.18} = \mathbf{4.23}$$

- Llevando estos resultados a una tabla tenemos lo siguiente:

Nº Cores	2	4	6	8	10	12	14	16
Tiempo	2.50	2.00	1.66	1.25	1.20	1.18	1.18	1.18
Speedup	2.00	2.50	3.01	4.00	4.16	4.23	4.23	4.23
Eficiencia	1.0	0.6	0.5	0.5	0.4	0.3	0.3	0.2

Ejemplo de escalamiento

- Dado que la eficiencia debe ser mayor a 0.5, el número de cores a solicitar debería ser a lo sumo 4. De hecho no se deberían seguir lanzando ejecuciones a partir de 6 cores, dónde se ve que la eficiencia es igual (o inferior) a 0.5:

NLHPC
National Laboratory for High Performance Computing Chile

Ejercicio 6

- Imagina que deseas ejecutar el programa **namd**.
  - ¿Qué opciones tienes disponibles en la infraestructura del NLHPC?
  - ¿Qué diferencias hay entre ellas?
  - Si quieres lanzar usando GPUs, ¿qué módulos debes de cargar?
- Anota los comandos que has usado.

NLHPC
National Laboratory
for High Performance
Computing
Chile

Otros Comandos - 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       -
```

NLHPC
National Laboratory for High Performance Computing Chile

Gracias!

https://www.nlhpc.cl/

NLHPC
National Laboratory
for High Performance
Computing
Chile