Universidad del Rosario

# CENTAURO: CONFIGURACIÓN BÁSICA Y USO DEL CLÚSTER HPC

## Manual de Usuario SLURM y uso de módulos

Tecnología Informática y Comunicaciones

Abril 2018

# Agenda

General

Universidad del Rosario

- Introducción
- Ejecución de trabajos
- Configuración
- Administración de cuentas
- Otros tópicos

# Agenda

General

Universidad del Rosario

- **Introducción**
- Ejecución de trabajos
- Configuración
- Administración de cuentas
- Otros tópicos

# Introducción

## Generalidades

El sistema operativo utilizado en el clúster HPC es CentOS 7.4

Los usuarios acceden al nodo maestro o de "cabecera"

- Los usuarios pueden acceder también a los nodos de cómputo
  - SSH "directo" [1]
  - Utilizando el gestor de trabajos
- Las herramientas de desarrollo y programas son compartidos por todos los nodos

1. Esta configuración va a cambiar.

# Introducción

## Generalidades

Universidad del Rosario

Todos los trabajos de cómputo se deben enviar a través de SLURM

https://slurm.schedmd.com

# Introducción

## Generalidades

Universidad del Rosario

Todos los trabajos de cómp se deben enviar a través de SLURM

https://sl dmd.com

# Introducción

## Generalidades

Universidad del Rosario

!El nodo maestro **no debe ser utilizado** para ejecutar trabajos intensivos en cómputo!

# Introducción

SLURM

Universidad del Rosario

## Sistema gestor de recursos de clústeres computacionales

SLURM: Simple Linux Utility for Resource Management

Se ejecuta desde la línea de comandos (*bash shell*).

- El servicio CENTAURO puede proveer capacitación adicional en Linux

Universidad del Rosario

# Conceptos

Nodos: Máquinas físicas

Particiones: Conjuntos de nodos con características específicas

Colas de trabajo pendiente

Tareas: Asignación de nodos dentro de una partición a un usuario por un tiempo determinado

Pasos de trabajo: Conjunto de instrucciones/comandos en una tarea

# Agenda

General

Universidad del Rosario

- Introducción
- **Ejecución de trabajos**
- Configuración
- Administración de cuentas
- Otros tópicos

# Ejecución de Trabajos

## Generalidades

Universidad del Rosario

Los comandos varían dependiendo de si son comandos en modo administración o en modo usuario

- `sinfo`
- `scontrol`
- `squeue`
- `sbatch`
- `srun`
- `scancel`

# Ejecución de Trabajos

## Información de particiones

Universidad del Rosario

Presenta información con las características de las particiones.

- >sinfo

```
PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
normal*      up 5-00:00:00      5   idle thanatos-1-[1-5]
gpu          up 2-00:00:00      1   idle thanatos-1-6
```

# Ejecución de Trabajos

## Información de nodos

Universidad del Rosario

Presenta información con las características de los nodos.

- `>scontrol show nodes`

```
NodeName=thanatos-1-1 CoresPerSocket=16
   CPUAlloc=0 CPUErr=0 CPUTot=32 CPULoad=0.01
   AvailableFeatures=(null)
   ActiveFeatures=(null)
   Gres=(null)
   NodeAddr=thanatos-1-1 NodeHostName=thanatos-1-1
   RealMemory=1 AllocMem=0 FreeMem=62033 Sockets=2 Boards=1
   State=IDLE ThreadsPerCore=1 TmpDisk=0 Weight=1 Owner=N/A MCS_label=N/A
   Partitions=normal
   BootTime=None SlurmdStartTime=None
   CfgTRES=cpu=32,mem=1M,billing=32
   …
```

# Ejecución de Trabajos

## Información de nodo específico

Universidad del Rosario

Presenta información de un nodo específico.

- `>scontrol show node thanatos-1-6`

```
NodeName=thanatos-1-6 CoresPerSocket=1
   CPUAlloc=0 CPUErr=0 CPUTot=32 CPULoad=0.01
   AvailableFeatures=gpu,p100
   ActiveFeatures=gpu,p100
   Gres=gpu:p100:4
   NodeAddr=thanatos-1-6 NodeHostName=thanatos-1-6
   RealMemory=1 AllocMem=0 FreeMem=62706 Sockets=2 Boards=1
   State=IDLE ThreadsPerCore=1 TmpDisk=0 Weight=1 Owner=N/A MCS_label=N/A
   Partitions=gpu
   BootTime=None SlurmdStartTime=None
   CfgTRES=cpu=32,mem=1M,billing=32
   …
```

# Ejecución de Trabajos

## Comandos SLURM para ejecutar trabajos

Universidad del Rosario

- `salloc:` Asigna recursos, bien sea para ejecutar un comando o para lanzar trabajos de forma interactiva (*shell*)

- `srun:` Asigna recursos y lanza trabajos que se ejecutarán en cada asignación

- `sbatch:` Asigna recursos y ejecuta un script
  - El formato del script contiene parámetros propios de `sbatch`
  - **Método recomendado para ejecutar trabajos**

# Ejecución de Trabajos

## Parámetros de configuración de scripts

Universidad del Rosario

```
-p      # Partición (cola)
-N      # Número de nodos
-n      # Número de núcleos
-mem    # Asignación máxima de memoria
-t      # Límite de tiempo
-o      # Salida STDOUT
-e      # Salida STDERR
```

# Ejecución de Trabajos

## Ejemplo de script

Universidad del Rosario

```
#!/bin/bash
#SBATCH -p normal       # Partición (cola)
#SBATCH -N 1            # Número de nodos
#SBATCH -n 2            # Número de núcleos
#SBATCH -t 0-2:00       # Límite de tiempo (D-HH:MM)
#SBATCH -o salida.out   # Salida STDOUT
#SBATCH -e error.err    # Salida STDERR

for i in {1..1000000}; do
  echo $RANDOM >> aleatorios.dat
done
sort aleatorios.dat
```

# Ejecución de Trabajos

## Ejemplo de script

Universidad del Rosario

```
#!/bin/bash
#SBATCH -p normal       # Partición (cola)
#SBATCH -N 1            # Número de nodos
#SBATCH -n 2            # Número de núcleos
#SBATCH -t 0-2:00       # Límite de tiempo (D-HH:MM)
#SBATCH -o salida.out   # Salida STDOUT
#SBATCH -e error.err    # Salida STDERR

for i in {1..100000}; do
  echo $RANDOM >> aleatorios.dat
done
sort aleatorios.dat
```

"shebang"

# Ejecución de Trabajos

## Ejemplo de script

Universidad del Rosario

```
#!/bin/bash
#SBATCH -p normal          # Partición (cola)
#SBATCH -N 1               # Número de nodos
#SBATCH -n 2               # Número de núcleos
#SBATCH -t 0-2:00          # Límite de tiempo (D-HH:MM)
#SBATCH -o salida.out      # Salida STDOUT
#SBATCH -e error.err       # Salida STDERR

for i in {1..100000}; do
  echo $RANDOM >> aleatorios.dat
done
sort aleatorios.dat
```

Configuración de sbatch

# Ejecución de Trabajos

## Ejemplo de script

```
#!/bin/bash
#SBATCH -p normal        # Partición (cola)
#SBATCH -N 1             # Número de nodos
#SBATCH -n 2             # Número de núcleos
#SBATCH -t 0-2:00        # Límite de tiempo (D-HH:MM)
#SBATCH -o salida.out    # Salida STDOUT
#SBATCH -e error.err     # Salida STDERR

for i in {1..100000}; do
  echo $RANDOM >> aleatorios.dat
done
sort aleatorios.dat
```

Parámetros de configuración de sbatch

# Ejecución de Trabajos

## Ejemplo de script

```
#!/bin/bash
#SBATCH -p normal       # Partición (cola)
#SBATCH -N 1            # Número de nodos
#SBATCH -n 2            # Número de núcleos
#SBATCH -t 0-2:00       # Límite de tiempo (D-HH:MM)
#SBATCH -o salida.out   # Salida STDOUT
#SBATCH -e error.err    # Salida STDERR

for i in {1..100000}; do
  echo $RANDOM >> aleatorios.dat
done
sort aleatorios.dat
```

Comandos
"Job steps"

# Ejecución de Trabajos

## Ejecución de scripts

Universidad del Rosario

Envía un script a la cola de trabajos

- `>sbatch <custom_script.sh>`

```
Submitted batch job xx
```

# Ejecución de Trabajos

## Estados de un trabajo

Universidad del Rosario

```
RUNNING      # Ejecutándose en una asignación
PENDING      # En cola esperando asignación
COMPLETED    # Terminó todos los procesos en nodos
CANCELLED    # Cancelado explícitamente
COMPLETING   # En proceso de terminación.
FAILED       # Terminó con condición de falla no estándar
NODE_FAIL    # Falla en uno o más nodos de la asignación
SUSPENDED    # Tiene asignación, pero está pausado
TIMEOUT      # Excedió su límite de tiempo
```

# Ejecución de Trabajos

## Información de trabajos en cola

Universidad del Rosario

### Listar trabajos en cola de un usuario

- `>squeue -u <usuario>`

- `>squeue -u <usuario> -p <partition>`

# Ejecución de Trabajos

## Información de trabajos en cola

Universidad del Rosario

### Listar trabajos en ejecución/pendientes

- `>squeue -u <usuario> -t RUNNING`

- `>squeue -u <usuario> -t PENDING`

# Ejecución de Trabajos

## Información de trabajos en cola

Universidad del Rosario

### Ver información detallada de un trabajo

- `>scontrol show jobid -dd <jobid>`

- `>sstat --format=AveCPU,AvePages,AveRSS,AveVMSize,JobID -j <jobid> --allsteps`

# Ejecución de Trabajos

## Información de trabajos terminados

Universidad del Rosario

### Ver estadísticas de un trabajo terminado

- `>sacct -j <jobid> --format=JobID,JobName,MaxRSS,Elapsed`

```
       JobID    JobName     MaxRSS    Elapsed
------------ ---------- ---------- ----------
44           slurm_tes+                00:00:38
44.batch           batch                00:00:38
```

# Ejecución de Trabajos

## Información de trabajos terminados

Universidad del Rosario

### Ver estadísticas de todos los trabajos terminados de un usuario

- `>sacct -u <username> --format=JobID,JobName,MaxRSS,Elapsed`

```
       JobID    JobName     MaxRSS    Elapsed
------------ ---------- ---------- ----------
47                 bash             00:02:49
47.0               bash             00:02:49
47.1              orted             00:00:01
48                 test             00:00:00
48.batch          batch             00:00:00
49                 bash             00:00:19
50                 bash             00:01:40
```

# Ejecución de Trabajos

## Control de trabajos

Universidad del Rosario

### Cancelar un trabajo

- `>scancel -u <username>`
- `>scancel --name <jobname>`

# Ejecución de Trabajos

## Control de trabajos

Universidad del Rosario

### Cancelar todos los trabajos pendientes

- `>scancel -t PENDING -u <username>`

# Ejecución de Trabajos

## Control de trabajos

Universidad del Rosario

### Pausar/despausar trabajos

- `>scontrol hold <jobid>`

- `>scontrol resume <jobid>`

# Ejecución de Trabajos

## Control de trabajos

Universidad del Rosario

Volver a colocar en cola un trabajo (cancelar y ejecutar nuevamente)

- `>scontrol requeue <jobid>`

# Ejecución de Trabajos

## Ejercicio

Universidad del Rosario

Ejecutar trabajos SLURM utilizando las herramientas/programas del grupo de investigación

# Agenda

General

Universidad del Rosario

- Introducción
- Ejecución de trabajos
- **Configuración**
- Administración de cuentas
- Otros Tópicos

# Configuración

## Listar módulos instalados

Universidad del Rosario

Presenta información de los módulos instalados.

- `>module list`

```
Currently Loaded Modules:
  1) autotools   2) prun/1.2   3) gnu7/7.3.0   4) openmpi3/3.0.0
5) ohpc   6) openblas/0.2.20   7) R/3.4.4
```

# Configuración

## Cargar módulos

Universidad del Rosario

Cargar/instalar un módulo para que esté disponible en el entorno.

- `>module load <modulo>`

# Configuración

## Ejemplo: Cargar módulos

Universidad del Rosario

### Cargar/instalar el módulo R/3.4.4

- >module list

```
Currently Loaded Modules:
1) autotools   2) prun/1.2   3) gnu7/7.3.0   4) openmpi3/3.0.0   5) ohpc
```

- **>module load R/3.4.4**
- >module list

```
Currently Loaded Modules:
1) autotools   2) prun/1.2   3) gnu7/7.3.0   4) openmpi3/3.0.0   5) ohpc
6) openblas/0.2.20   7) R/3.4.4
```

# Configuración

## Descargar módulos

Universidad del Rosario

### Descargar/desintalar un módulo del entorno

- `>module unload <modulo>`

# Configuración

## Listar módulos disponibles

Universidad del Rosario

Presenta información de los módulos disponibles en el entorno.

- `>module avail`

```
------------------------------------------------------------------------------ /opt/ohpc/pub/moduledeps/gnu7-openmpi3 -------------------------------------------------------------------------------
   adios/1.13.0     hypre/2.13.0     mpiP/3.4.1          netcdf-fortran/4.4.4    phdf5/1.10.1      py2-mpi4py/3.0.0    py3-scipy/1.0.0    scorep/3.1        superlu_dist/5.3.0
   boost/1.66.0     imb/2018.1       mumps/5.1.2         netcdf/4.5.0            pnetcdf/1.8.1     py2-scipy/1.0.0     scalapack/2.0.2    sionlib/1.7.1     tau/2.27
   fftw/3.3.7       mfem/3.3.2       netcdf-cxx/4.3.0    petsc/3.8.3             ptscotch/6.0.4    py3-mpi4py/3.0.0    scalasca/2.3.1     slepc/3.8.2       trilinos/12.12.1

------------------------------------------------------------------------------ /opt/ohpc/pub/moduledeps/gnu7 --------------------------------------------------------------------------------------
```

…

# Agenda

General

Universidad del Rosario

- Introducción
- Ejecución de trabajos
- Configuración
- **Administración de cuentas**
- Otros Tópicos

# Administración de cuentas

Usuarios y grupos

Universidad del Rosario

Todos los usuarios pertenecen o bien a su grupo de investigación o bien a un grupo por defecto de cada programa

Los grupos y la distribución del almacenamiento para cada facultad/programa/grupo trata de seguir la estructura organizacional

# Administración de cuentas

Usuarios y grupos

Universidad del Rosario

La distribución del almacenamiento también depende de esta estructura.

(Mayor información en el documento de acceso y transferencia de archivos, así como en el manual de usuario)

# Administración de cuentas

Usuarios y grupos

Universidad del Rosario

Como parte de respuesta a la solicitud de recursos se indicarán los directorios a los cuales los usuarios pueden acceder y sus respectivos permisos

# Agenda

General

Universidad del Rosario

- Introducción
- Ejecución de trabajos
- Configuración
- Administración de cuentas
- **Otros Tópicos**

# Otros tópicos

## Ejecución de programas MPI

Universidad del Rosario

Se pueden ejecutar los trabajos MPI desde cualquiera de los métodos usuales (`salloc`, `sbatch`, `srun`)

SLURM controla directamente la lista de hosts y cuántos procesos iniciar en cada hosts

- No es necesario `--hostfile`, `--host` o `-np`

# Otros tópicos

## Ejecución de programas MPI


Existe una utilidad denominada `prun` que facilita la ejecución de trabajos paralelos.

`prun` es el enfoque recomendado si se utiliza `sbatch`

# Otros tópicos

## Ejecución de programas MPI

Universidad del Rosario

```
#!/bin/bash

#SBATCH -J test                 # Job name
#SBATCH -o job.%j.out           # Name of stdout output file (%j expands to jobId)
#SBATCH -N 4                    # Total number of nodes requested
#SBATCH -n 128                   # Total number of mpi tasks requested
#SBATCH -t 00:30:00             # Run time (hh:mm:ss) - 0.5 hours

# Launch MPI-based executable

prun ./<mpi_capable_program>
```

# Otros tópicos

## Ejecución de código GPU

Universidad del Rosario

Las GPUs se encuentran en una partición separada `gpu`

- `>sinfo`

```
PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
normal*      up 5-00:00:00      5   idle thanatos-1-[1-5]
gpu          up 2-00:00:00      1   idle thanatos-1-6
```

Universidad del Rosario

# Tutorial – Paso #0:  Cargar el módulo CUDA

- `>module load cuda`

Universidad del Rosario

# Tutorial – Paso #1: Copiar ejemplos de CUDA-9.1

- `>mkdir cuda-tutorial`
- `>cp -R /usr/local/cuda-9.1/samples/ cuda-tutorial`
- `>cd cuda-tutorial/samples/0_Simple/simpleMultiGPU`

Universidad del Rosario

# Tutorial – Paso #2: La importancia de `module`. Intentar compilar …

- `>make`

```
/usr/local/cuda-9.1/bin/nvcc -ccbin g++ -I../../common/inc  -m64    -gencode arch=compute_30,code=sm_30 -gencode
arch=compute_35,code=sm_35 -gencode arch=compute_37,code=sm_37 -gencode arch=compute_50,code=sm_50 -gencode
arch=compute_52,code=sm_52 -gencode arch=compute_60,code=sm_60 -gencode arch=compute_61,code=sm_61 -gencode
arch=compute_70,code=sm_70 -gencode arch=compute_70,code=compute_70 -o simpleMultiGPU.o -c simpleMultiGPU.cu

In file included from /usr/local/cuda-9.1/bin/../targets/x86_64-linux/include/host_config.h:50:0,

                 from /usr/local/cuda-9.1/bin/../targets/x86_64-linux/include/cuda_runtime.h:78,

                 from <command-line>:0:

/usr/local/cuda-9.1/bin/../targets/x86_64-linux/include/crt/host_config.h:121:2: error: #error -- unsupported
GNU version! gcc versions later than 6 are not supported!

#error -- unsupported GNU version! gcc versions later than 6 are not supported!

  ^~~~~

make: *** [simpleMultiGPU.o] Error 1
```

Universidad del Rosario

# Tutorial – Paso #2: La importancia de `module`. Cargar los módulos correctos y compilar

- `>module swap gnu7/7.3.0 gnu/5.4.0`

```
Inactive Modules:
 1) openmpi3
```

- `>make`

# Otros tópicos

## Ejecución de código GPU - Tutorial

Universidad del Rosario

- >make

```
/usr/local/cuda-9.1/bin/nvcc -ccbin g++ -I../../common/inc  -m64    -gencode
arch=compute_30,code=sm_30 -gencode arch=compute_35,code=sm_35 -gencode
arch=compute_37,code=sm_37 -gencode arch=compute_50,code=sm_50 -gencode
arch=compute_52,code=sm_52 -gencode arch=compute_60,code=sm_60 -gencode
arch=compute_61,code=sm_61 -gencode arch=compute_70,code=sm_70 -gencode
arch=compute_70,code=compute_70 -o simpleMultiGPU.o -c simpleMultiGPU.cu

/usr/local/cuda-9.1/bin/nvcc -ccbin g++   -m64      -gencode
arch=compute_30,code=sm_30 -gencode arch=compute_35,code=sm_35 -gencode
arch=compute_37,code=sm_37 -gencode arch=compute_50,code=sm_50 -gencode
arch=compute_52,code=sm_52 -gencode arch=compute_60,code=sm_60 -gencode
arch=compute_61,code=sm_61 -gencode arch=compute_70,code=sm_70 -gencode
arch=compute_70,code=compute_70 -o simpleMultiGPU simpleMultiGPU.o

mkdir -p ../../bin/x86_64/linux/release

cp simpleMultiGPU ../../bin/x86_64/linux/release
```

# Otros tópicos

## Ejecución de código GPU - Tutorial

Universidad del Rosario

- `>srun -p gpu -N 1 ./simpleMultiGPU`

```
srun: job 53 queued and waiting for resources
srun: job 53 has been allocated resources
Starting simpleMultiGPU
CUDA-capable device count: 4
Generating input data...

Computing with 4 GPUs...
  GPU Processing time: 12.937000 (ms)

Computing with Host CPU...

Comparing GPU and Host CPU results...
  GPU sum: 16777304.000000
  CPU sum: 16777294.395033
  Relative difference: 5.724980E-07
```

# Referencias



Universidad del Rosario

https://www.rc.fas.harvard.edu/resources/documentation/convenient-slurm-commands/

https://docs.nvidia.com/cuda/cuda-samples/index.html