Grid Engine y Slurm

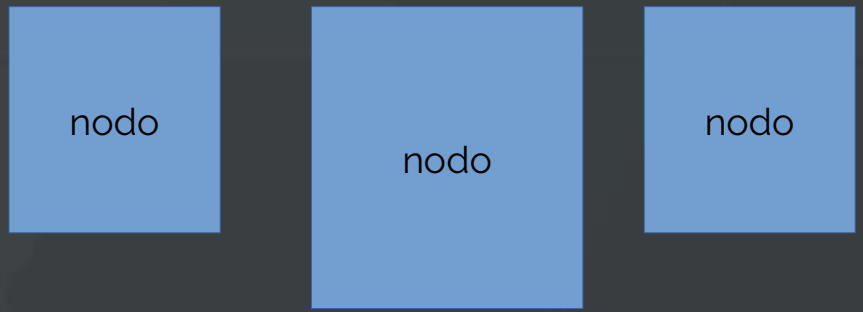
HPC Admintech
Valencia, Mayo 2018

Jesús Cuenca
Senior HPC Consultant @ SIE

Introducción
sistema de colas

Reto


trabajo
trabajo
trabajo
trabajo
trabajo
trabajo


nodo
nodo
nodo

Solución


trabajo

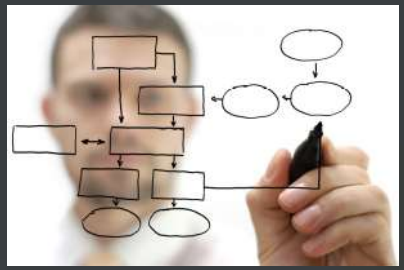

trabajo


trabajo


trabajo


trabajo


trabajo




nodo


nodo


nodo

Grid Engine

Historia

1992: Codine

2000: Sun

2001: Open Source

2010: Oracle. Closed source

2011: Univa Grid Engine (comercial, ~1 release/año – marzo 2017)

Son of Grid Engine (última versión: marzo 2016)

Open Grid Scheduler (abandonware)

http://www.softpanorama.org/HPC/Grid_engine/history.shtml

Arquitectura

Master host: monitoring + scheduling

Execution host: sge_execd

Administration host

Submit host

Arquitectura

“Parallel Environments” (PE)

Slots: máximo procesos

allocation rule: procesos/nodo (constante), fill_up (todos los slots), round_robin (de uno en uno)

start/stop scripts (MPI)

SLURM

Presentación

Simple Linux Utility for Resource Management

"Slurm es un sistema de gestión de clústers y planificación de trabajos escalable, tolerante a fallos y de código abierto (...)
Su uso es autocontenido (...)"

Historia

2002: LLNL Slurm

2010: SchedMD

2018: Slurm 17.11.05 (varias releases por año)

https://computing.llnl.gov/tutorials/linux_clusters/LinuxProjectReport.2002.08.18.pdf

Plataformas

PC / Linux

IBM BlueGene

Cray/XT

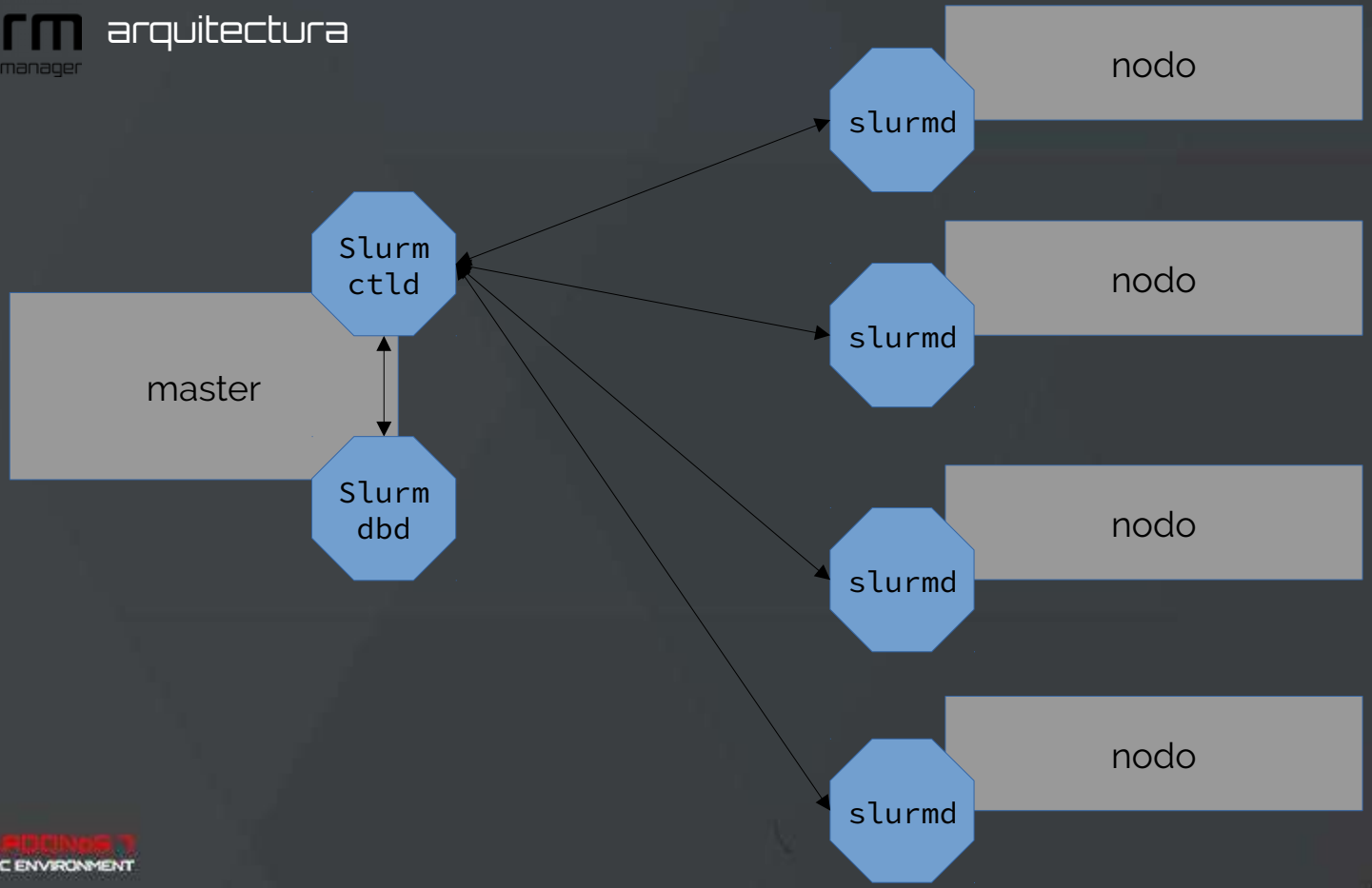
IBM Parallel Environment

Popularidad

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , **NRCPC** National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
2	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P , **NUDT** National Super Computer Center in Guangzhou China	3,120,000	33,862.7	54,902.4	17,808
3	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 , **Cray Inc.** Swiss National Supercomputing Centre (CSCS) Switzerland	361,760	19,590.0	25,326.3	2,272
4	**Gyoukou** - ZettaScaler-2.2 HPC system, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 700Mhz , **ExaScaler** Japan Agency for Marine-Earth Science and Technology Japan	19,860,000	19,135.8	28,192.0	1,350
5	**Titan** - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x , **Cray Inc.** DOE/SC/Oak Ridge National Laboratory United States	560,640	17,590.0	27,112.5	8,209
6	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom , **IBM** DOE/NNSA/LLNL United States	1,572,864	17,173.2	20,132.7	7,890
7	**Trinity** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect , **Cray Inc.** DOE/NNSA/LANL/SNL United States	979,968	14,137.3	43,902.6	3,844
8	**Cori** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect , **Cray Inc.** DOE/SC/LBNL/NERSC United States	622,336	14,014.7	27,880.7	3,939
9	**Oakforest-PACS** - PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path , **Fujitsu** Joint Center for Advanced High Performance Computing Japan	556,104	13,554.6	24,913.5	2,719
10	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect , **Fujitsu** RIKEN Advanced Institute for Computational Science (AICS) Japan	705,024	10,510.0	11,280.4	12,660
16	**MareNostrum** - Lenovo SD530, Xeon Platinum 8160 24C 2.1GHz, Intel Omni-Path , **Lenovo** Barcelona Supercomputing Center Spain	153,216	6,470.8	10,296.1	1,632

https://www.top500.org/lists/2017/11/

slurm arquitectura



Configuración

/etc/slurm/slurm.conf
(idéntica en todos los nodos)

Parámetros:

Valores por defecto
Recursos genéricos
Re-encolado automático
Características nodos cálculo
Particiones
...

Comandos

sbatch	qsub
scancel	qdel
squeue	qstat
scontrol show nodes	qhost
sinfo	qhost -q
sview	qmon

Parámetros trabajo

Duración y programación horaria:

```
--time={DD-HH / HH:MM:SS}
--begin=YYYY-MM-
DDTHH:MM:SS
--deadline=YYYY-MM-
DDTHH:MM:SS
```

```
-l h_rt=[segundos]
-a [YYMMDDhhmm]
```

Entorno

```
--workdir=
--output=
--error=
--export
--job-name=
```

```
-wd
-o
-e
-V
-N
```

Parámetros trabajo

Cola	
`--partition=`	`-q`
Tareas	
`--ntasks=`	`-pe [PE] [N]`
Distribución de tareas	
`--tasks-per-node=` `--nodelist=n,m`	`allocation_rule en PE` `-q [cola]@[nodo]`

Parámetros trabajo

Recursos

CPU

`--cpus-per-task=`	`-pe [PE] [N]`

Genéricos

`--gres=`	`-l`

Memoria

`--mem=<size><unit>`	`-l h_vmem=`
`--mem-per-cpu=<size><unit>`	`-l h_vmem=`

Parámetros trabajo

Aviso	
--mail-user=	-M
Dependencias	
--dependency= after:JOB_ID afterok:JOB_ID afternotok:JOB_ID singleton	-hold_jid JOB_ID

Parámetros trabajo

Variables informativas

```
$SLURM_JOB_ID
$SLURM_JOB_NAME
$SLURM_JOB_NODELIST
$SLURM_JOB_PARTITION
$SLURM_SUBMIT_DIR
$SLURM_SUBMIT_HOST
(...)
```


```
$JOB_ID
$JOB_NAME
$PE_HOSTFILE
$SLURM_JOB_PARTITION
$SGE_O_WORKDIR
$SGE_O_HOST
(...)
```

slurm workload manager

límites

Partition: {MaxCPUsPerNode, MaxMemPerCPU, MaxMemPerNode, MaxNodes, MaxTime}

Account: {GrpTres, MaxJobs…}
GrpTRES: “The total count of TRES able to be used at any given time from jobs running from an association”
TRES: Trackable RESources {cpu,mem,node,gres}

QoS (Quality of Service): {MinTRESPerJob, ...}

Multifactor priority

Age: tiempo de espera

Job size: nodos o CPUs solicitadas

TRES

QoS

Partition

Fair-share: reparto por cuentas / grupos

Multifactor priority

```
Job_priority =
    (PriorityWeightAge) * (age_factor) +
    (PriorityWeightFairshare) * (fair-share_factor) +
    (PriorityWeightJobSize) * (job_size_factor) +
    (PriorityWeightPartition) * (partition_factor) +
    (PriorityWeightQOS) * (QOS_factor) +
    SUM(TRES_weight_cpu * TRES_factor_cpu,
        TRES_weight_<type> * TRES_factor_<type>,
        ...)
```

Gestión energía

Slurm.conf

SuspendTime
SuspendProgram
ResumeProgram
SuspendExcNodes

SLURM demo

Demo / instalación

MASTER

Munge
Slurm user
Paquetes slurm
Permisos
MySQL DB
Configuración Slurm (slurm.conf, slurmdbd.conf)
Slurmdbd
Slurmctld
sacctmgr

Demo / instalación

BACKUP

Munge
Slurm user
Paquetes slurm
Permisos
(MySQL DB)
Replicar slurm.conf, crear slurmdbd
Slurmdbd
Slurmctld

Demo / instalación

NODOS

Munge
Slurm user
Paquetes slurm
Permisos
(Replicar configuración Slurm)
Slurmd

slurm workload manager

herramientas

- sinfo
- scontrol
- sacct
- sreport
- sview

Alta disponibilidad

Parar nodo slurmctld principal

Arrancar nodo, munge, slurmctld

Gracias

Contacto

Sistemas Informáticos Europeos

Calle Marqués de Mondejar nº 29

913 61 10 02

www.sie.es

/HPCSIE

soporte@sie.es

@HPCSIE

+SistemasInformaticosEuropeosSLMadrid