Apresentação do cluster SeARCH

2020|2021

Albano Serrano
albano@di.uminho.pt
search-admin@di.uminho.pt

1. Introdução

- SeARCH - Services and Advanced Research Computing with HPC/HTC clusters (*High Performance/High Throughput Computing*);
- Consórcio para suporte à investigação em Ciências da Computação, Matemática e Física;
- Financiamento: 2005 (FCT), 2010 (FEDER) e 2014 (ON2);
- Dados globlais atuais:
  - 52 nós, cerca de 850 cores (com hyper-threading mais de 1700 cores)
  - 20 co-processadores/aceleradores
  - 100TB de armazenamento
  - Redes de 1 e 10 Gb

2. Infra-estrutura

- Computação
  - Nós heterogéneos
    - CPU duplo: quad, hexa, octa, deca, dodeca-core, tetradeca (14) e hexadeca (16);
  - Aceleradores
    - NVIDIA Geforce 6x8800 GT (já desativados)
    - NVIDIA Tesla Fermi 2xC2050, 2xM2070, 1xM2090
    - NVIDIA Tesla Kepler 5xK20m
    - INTEL Xeon PHI 1x5110, 8x7120

Caixa de 2U com 4 nós
(2 pares de twins)

Caixa de 1U
um único nó

- Nós recentes
  - 1 x Nó baseado no Knights Landing (KNL);
  - 2nd Generation Intel Xeon Phi Processor;
  - Xeon Phi CPU 7210 @ 1.30GHz;
  - 64 cores, 192GB RAM.

  - 2 x Nós baseados no Xeon E5-2660 v4 @ 2.00GHz; (Broadwell)
  - 14 cores/CPU, 56 cores total c/ HT;
  - 128GB de RAM.

  - 2 x Nós baseados no Xeon E5-2683 v4 @ 2.10GHz; (Broadwell)
  - 16 cores/CPU, 64 cores total c/ HT;
  - 256GB de RAM.

- Nós recentes
  - 1 x Nó baseado no Xeon Gold 6130 @ 2.10GHz; (Skylake)
  - 16 cores/CPU, 64 cores total c/ HT;
  - 96 GB de RAM.

  - 1 x Nó baseado no Cavium **ARM THUNDERX**;
  - 24 cores/CPU, 48 cores total;
  - 64 GB de RAM.

Comunicações

- Gigabit Ethernet (96 portas x 1Gbps)
- Myrinet (64 portas x 10Gbps, baixa latência)

Comutadores
1Gb Ethernet

Comutador
10Gb Myrinet

- Armazenamento
  - SAN (NFS para homes)
    - EMC CX300 (4,5TB)
    - Dot Hill AssuredSAN Pro 5000 (48TB)
  - GlusterFS (bigdata): 4xnós de 12TB (total de 48TB)
  - NAS (backup)

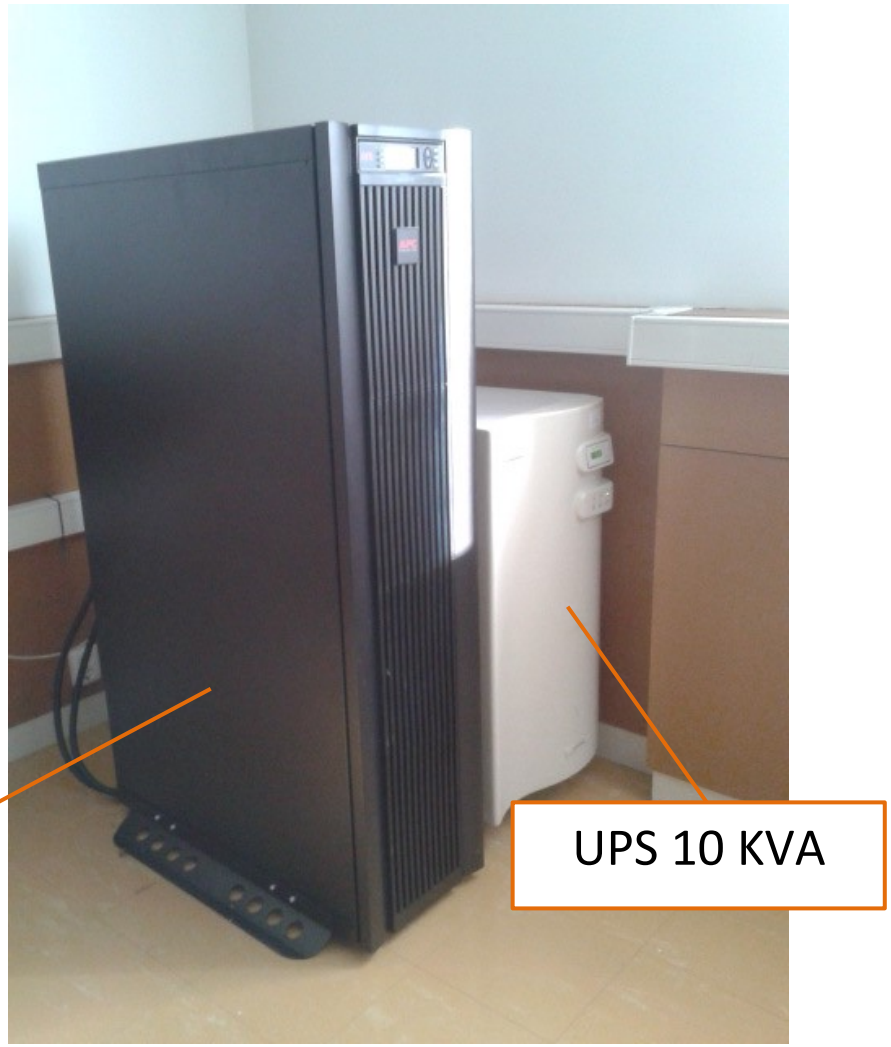
4 nós de 12TB,
GlusterFS

SAN de 48TB,
Homes e GlusterFS

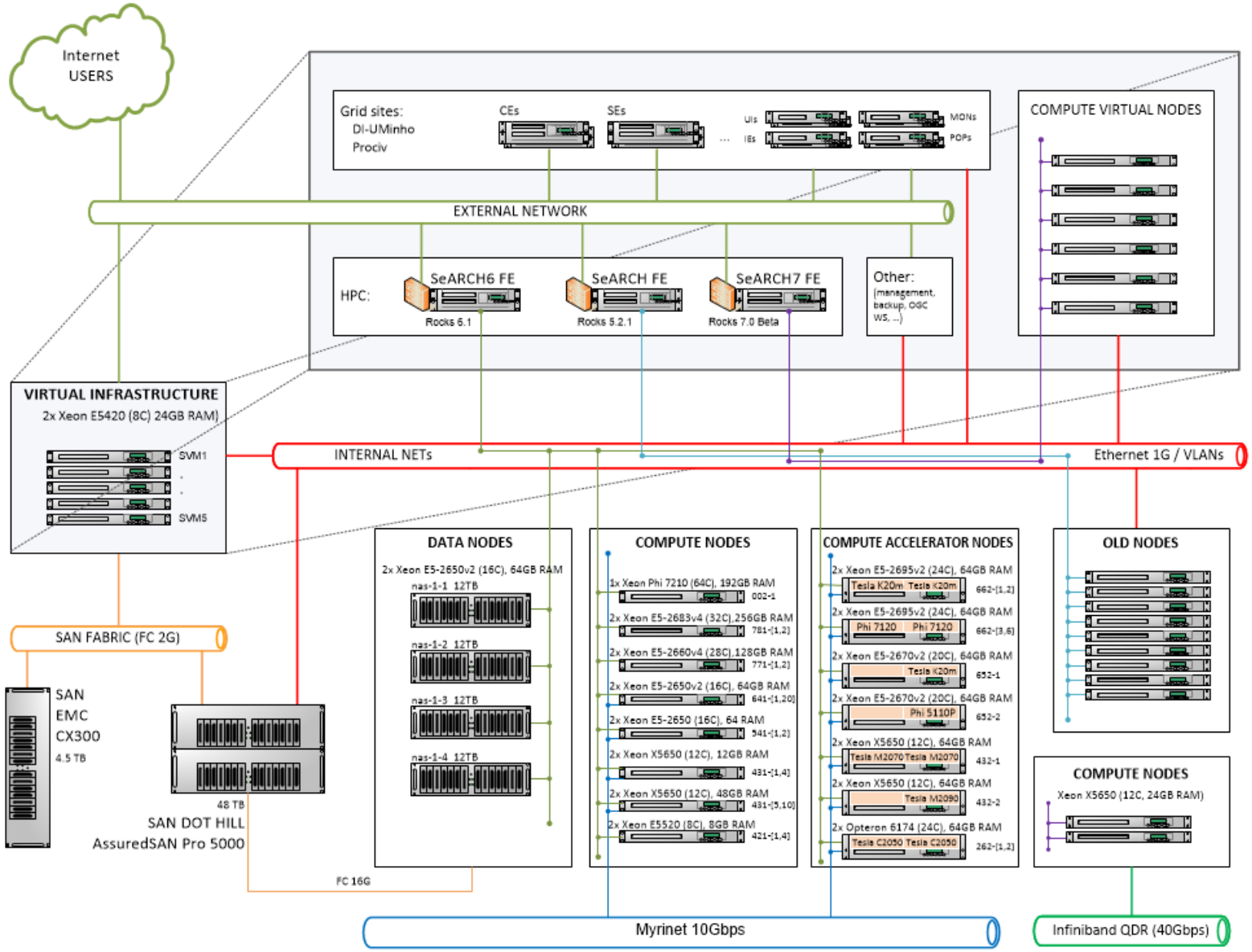
Virtualização

- Servidores VMware vSphere
- SAN EMC CX300
- Frontends, NAS para *homes*

- Alimentação eléctrica
  - 2 UPS 3KVA
  - UPS 20KVA
  - UPS 10KVA
- Refrigeração
  - 2 AC x 10KW



Arquitetura



3. Administração da plataforma

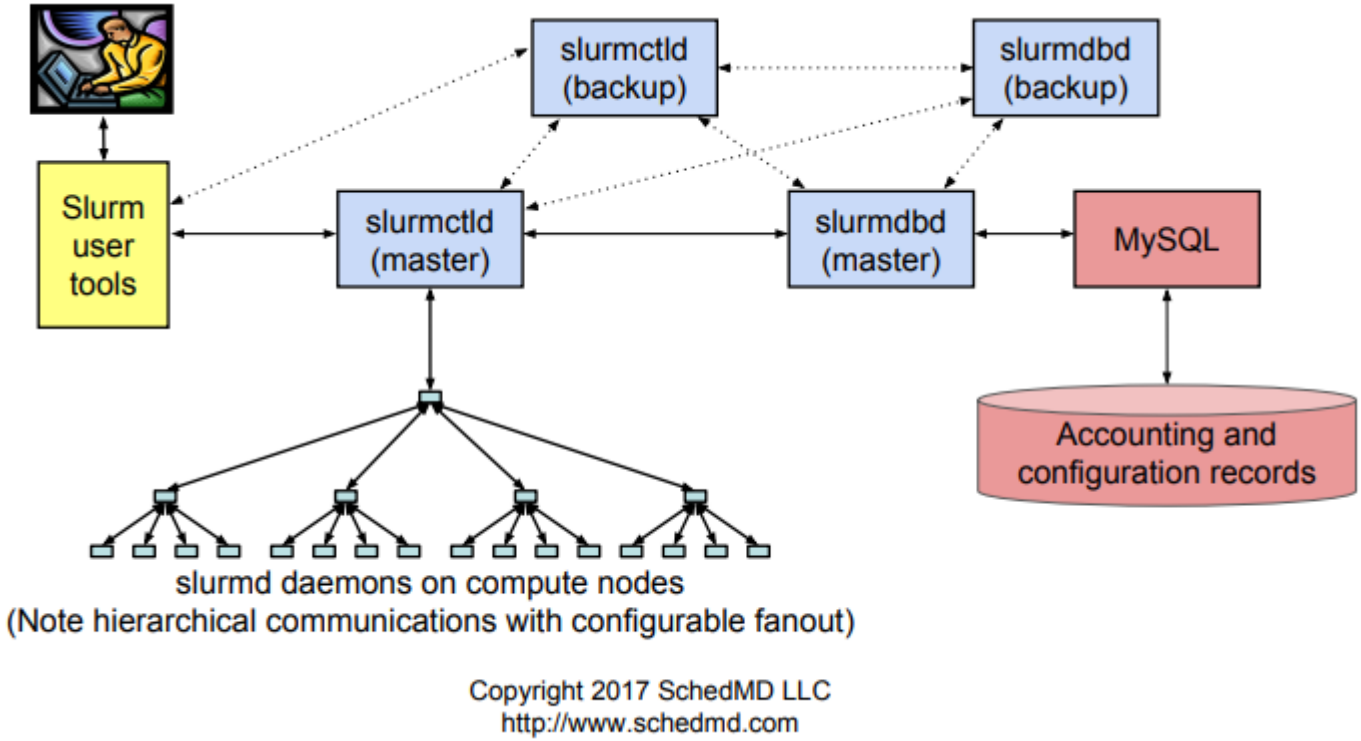
- Rocks cluster distribution (CentOS)
  - Instalação de nós
  - Configuração de serviços
  - Gestão de utilizadores
- Monitorização
  - Ganglia
  - http://search6.di.uminho.pt/ganglia/

4. Gestão da computação

- Recursos *vs* trabalhos de utilizadores
- Escalonador
  - Definição de políticas de utilização
    - Cotas por grupos, prioridades, etc.
  - Atribuição de recursos
  - Organização de recursos em filas
  - Submissão e controlo de execução de trabalhos

Gestão da computação (cont.)

- SLURM (**S**imple **L**inux **U**tility for **R**esource **M**anagement)



5. Utilização

- Diferentes áreas de investigação
  - Física, Math, Biomédica, Polímeros, etc.
- Ensino
  - Formação, teses MSc e PhD
- Procedimentos
  - Utilizador instala suas aplicações
  - Utilizador define trabalho: aplicação + dados
  - Utilizador submete trabalho
  - Sistema atribui recursos e executa trabalho

- # Exemplos de utilização

  - Consultar tabela com descrição dos nós:

    http://search6.di.uminho.pt/wordpress/?page_id=55

  - Comandos básicos
    - sinfo – lista caraterísticas das partitções/filas
    - squeue – lista trabalhos e respetivos estados
    - scancel – cancela um trabalho ou um conjunto de trabalhos
    - scontrol – lista caraterísticas de trabalhos, nós, partições
    - sstat – mostra estado de trabalhos em execução.

- Verificar disponibilidade de nós:

```
$ scontrol show node|less
NodeName=compute-881-1 Arch=x86_64 CoresPerSocket=1
CPUAlloc=32 CPUTot=64 CPULoad=0.02
AvailableFeatures=rack-881,64CPUs
ActiveFeatures=rack-881,64CPUs
Gres=(null)
NodeAddr=172.27.7.252 NodeHostName=compute-881-1
OS=Linux 3.10.0-693.5.2.el7.x86_64 #1 SMP Fri Oct 20 20:32:50 UTC 2017
RealMemory=95118 AllocMem=1280 FreeMem=90467 Sockets=64 Boards=1
State=MIXED ThreadsPerCore=1 TmpDisk=200014 Weight=20455799 Owner=N/A
MCS_label=N/A
Partitions=cdados,skylake,matlab,CLUSTER,WHEEL
BootTime=2020-12-20T13:39:11 SlurmdStartTime=2020-12-20T13:39:37
CfgTRES=cpu=64,mem=95118M,billing=87
AllocTRES=cpu=32,mem=1280M
CapWatts=n/a
CurrentWatts=0 AveWatts=0
ExtSensorsJoules=n/s ExtSensorsWatts=0 ExtSensorsTemp=n/s
```

- Listar partições *(filas/queues)*

```
$ sinfo
PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
acomp        up       2:00      2   idle compute-641-[10-11]
cdados       up    2:00:00      3   idle compute-641-[10-11],compute-881-1
skylake      up 4-03:00:00      1   idle compute-881-1
matlab       up 4-03:00:00      1   idle compute-881-1
DEBUG        up   infinite      0    n/a
CLUSTER*     up   infinite      2  down* compute-0-2,compute-1-1
CLUSTER*     up   infinite      6   idle compute-0-[0-1],compute-641-[10-
11],compute-662-1,compute-881-1
WHEEL        up   infinite      2  down* compute-0-2,compute-1-1
WHEEL        up   infinite      7   idle compute-0-[0-1],compute-641-[10-
11],compute-662-1,compute-881-1,search7
```

$ scontrol show partition cdados

```
PartitionName=cdados
   AllowGroups=cdados,impe,ajds AllowAccounts=ALL AllowQos=ALL
   AllocNodes=search7,search7.di.uminho.pt,login-0-0,login-0-0.local Default=NO
QoS=N/A
   DefaultTime=NONE DisableRootJobs=NO ExclusiveUser=NO GraceTime=0 Hidden=NO
   MaxNodes=UNLIMITED MaxTime=02:00:00 MinNodes=0 LLN=NO MaxCPUsPerNode=UNLIMITED
   Nodes=compute-641-[10-11],compute-881-1
   PriorityJobFactor=1 PriorityTier=1 RootOnly=NO ReqResv=NO
OverSubscribe=EXCLUSIVE
   OverTimeLimit=NONE PreemptMode=OFF
   State=UP TotalCPUs=128 TotalNodes=3 SelectTypeParameters=NONE
   JobDefaults=(null)
   DefMemPerNode=UNLIMITED MaxMemPerNode=UNLIMITED
```

- Submissão de trabalhos

- Comandos:

- sbatch – Submissão de trabalho para posterior execução (batch mode)

- salloc – Submissão de trabalho interativo

- srun – Submissão de trabalho paralelo (tipicamente MPI).

- Editar trabalho

```
$ cat hello.sh

#!/bin/sh
#SBATCH --nodes=1
#SBATCH --tasks=1
#SBATCH --time=00:01:00
#SBATCH --partition=cdados

./helloworld $1
```

- Submeter trabalho

```
$ sbatch hello.sh 4
```

- Monitorizar trabalho

```
$ squeue
$ squeue --job job_id

cat slurm-7480.out
There are 32 procs
Hello, world from thread 0!
There are 4 threads in the team!
Hello, world from thread 3!
Hello, world from thread 2!
Hello, world from thread 1!
That&#39;s all, folks! (403 usecs)
```