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)

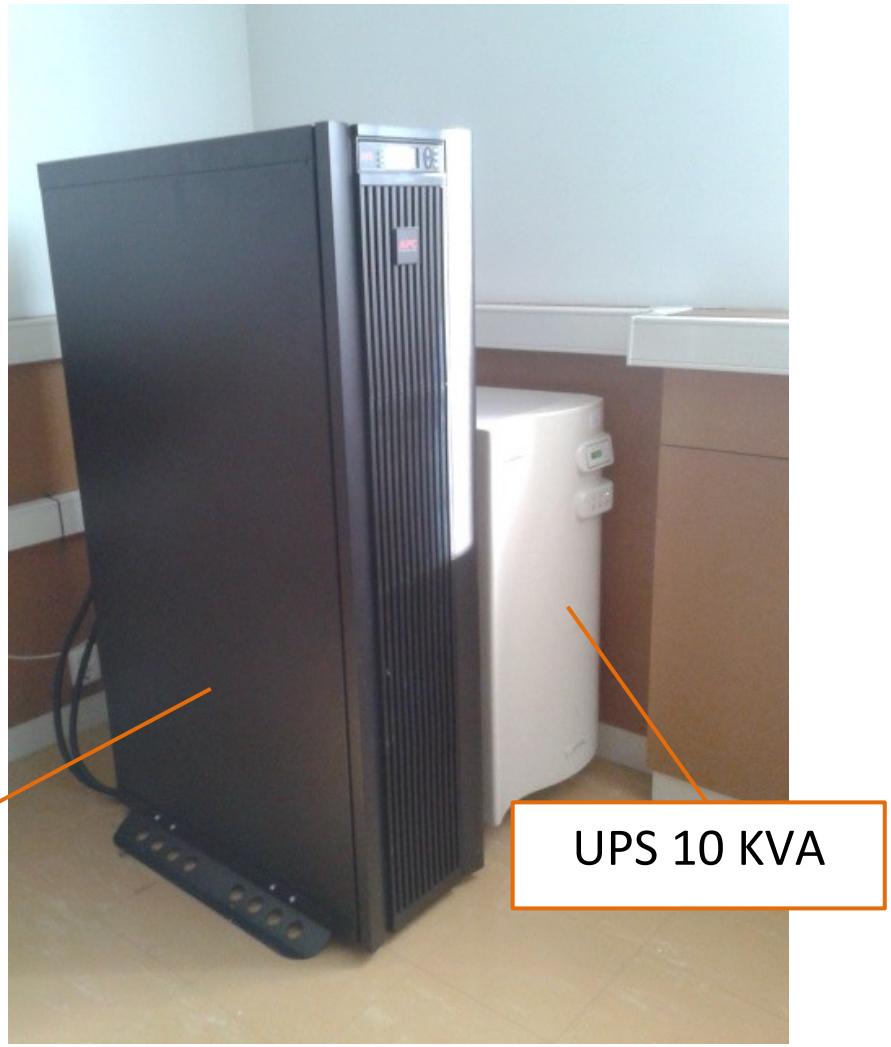
4 nós de 12TB,
GlusterFS

SAN de 48TB,
Homes e GlusterFS

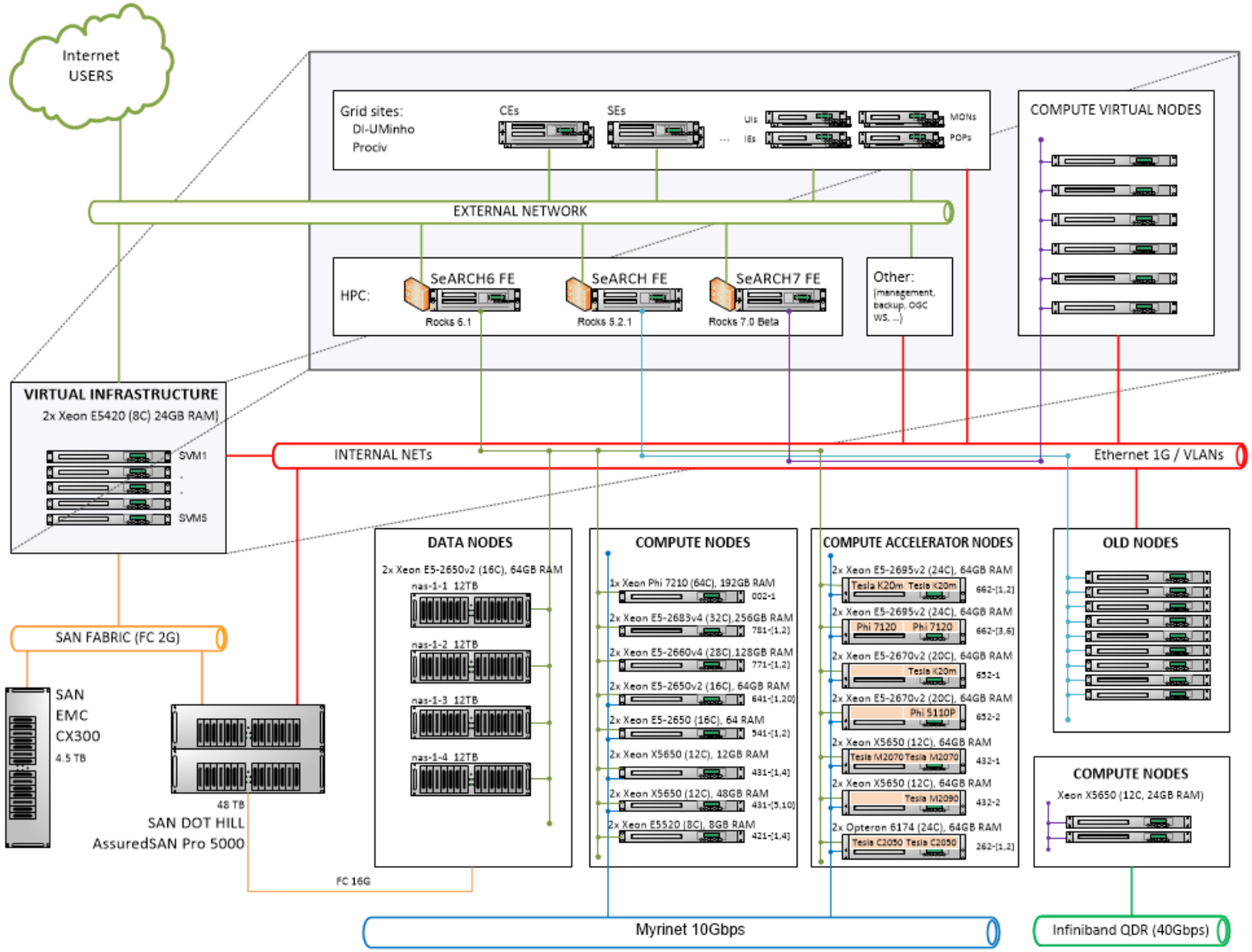
Virtualização

- 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
- Maui
  - Definição de políticas de utilização
    - Cotas por grupos, prioridades, etc.
  - Atribuição de recursos
- Torque/PBS
  - Organização de recursos em filas
  - Submissão e controlo de execução de trabalhos

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

Exemplo de utilização

- Consultar tabela com descrição dos nós:

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

- Verificar nós disponíveis:

```
$ pbsnodes -a | less
```

```
compute-662-1
     state = free
     np = 48
     properties = mei,day,r662,m64,d80,myri,repler,k20
     ntype = cluster
     jobs = 0/415176.search6.di.uminho.pt
     status =
rectime=1474973775,varattr=,jobs=415176.search6.di.uminho.pt,state=free,netload=162022934950,gres
=,loada
ve=1.00,ncpus=48,physmem=66068588kb,availmem=65907636kb,totmem=67092580kb,idletime=955327,nusers=
1,nsessions=1,session
s=23332,uname=Linux compute-662-1.local 2.6.32-279.14.1.el6.x86_64 #1 SMP Tue Nov 6 23:43:09 UTC
2012 x86_64,opsys=lin
ux
     mom_service_port = 15002
     mom_manager_port = 15003
     gpus = 2
```

- Listar filas *(queues)*

```
$ qstat -q
```

Filas privadas

```
biocnat            --      --    1200:00:   --     2   0 --   E R
acomp              --      --    00:10:00   --     1   0 --   E R
mei                --      --    168:00:0   --     1   0 --   E R
physics            --      --       --      --     0   0 --   E R
mogipc             --      --       --      --     0   0 --   E R
impe               --      --       --      --     0   0 --   E R
```

Filas públicas

```
day                --      --    24:00:00   --     9   0 --   E R
week               --      --    168:00:0   --    14   0 --   E R
fortnight          --      --    336:00:0   --     0   0 --   E R
month              --      --    720:00:0   --     0   0 --   E R
```

- Editar *script* trabalho sequencial:

```
$ vi job-seq.sh
#!/bin/sh
#
#PBS -N teste
#PBS -l walltime=05:00
#PBS -l nodes=1:ppn=1
#PBS -q mei

{código sequencial}
```

- Submeter trabalho

```
$ qsub job-seq.sh
```

- Monitorizar trabalho

```
$ showq ou qstat
```

- Editar *script* trabalho paralelo:

```
$ vi job-par.sh
#!/bin/sh
#
#PBS -N teste
#PBS -l walltime=15:00
#PBS -l nodes=4:r641:ppn=32
#PBS -q mei

module load gnu/4.9.0
module load gnu/openmpi_mx/1.8.2

mpirun --mca btl mx ^openib -np 32 \                -
machinefile $PBS_NODEFILE {código paralelo}
```

- Submeter trabalho

```
$ qsub job-seq.sh
```

6. Bibliografia

- Tutorial - Submitting a job using qsub
- The Maui scheduler for use with PBS
- FAQ: Running MPI jobs - Open MPI
- PBS Basics:

https://youtu.be/mRwqvHeDicE