Uma Introdução

por
Vítor
Oliveira

O PROJECTO SeARCH

O SeARCH – "*Services and Advanced Research Computing with HTC/HPC clusters*" – é um projecto dos centros de investigação dos departamentos de Informática, de Física e de Matemática da Universidade do Minho (UM), financiado pelo Programa de Reequipamento Científico Nacional da Fundação para a Ciência e Tecnologia (CONCREEQ/443/EEI/2005). A concretização do projecto levou à criação do *cluster* computacional Search, cuja aquisição data dos finais do ano de 2005. Prevê-se que o sistema, actualmente alojado no Departamento de Informática, venha a ter um papel de grande destaque na ligação da UM à GRID Nacional.

O *cluster* Search, à altura da sua entrada em regime de produção em Setembro de 2006, oferecia o maior poder de processamento numérico instalado em universidades portuguesas, proveniente de 112 núcleos Intel Xeon de 64 bits e de 8 processadores específicos para cálculo vectorial. A administração e o suporte lógico do sistema são totalmente baseados em software de domínio público para *cluster*, assim como em pacotes científicos e de desenvolvimento que correm nas versões mais actuais de Linux 64 bits.

Em termos tecnológicos é de salientar, desde logo, o facto de ter sido a primeira instalação na Europa da rede de baixa latência Myrinet 10G e um dos primeiros equipamentos em Portugal a usar as novas gerações de processadores multi-*core* da Intel. Em termos de computação é ainda de realçar a utilização de processadores gráficos (GPUs) como co-processadores numéricos paralelos. Para além das tecnologias Intel e Myricom, são também usados equipamentos SAN da EMC como suporte ao armazenamento de dados e da APC nas infra-estruturas.


Myricom


300GB 10Krpm
300GB 10Krpm
300GB 10Krpm
300GB 10Krpm
300GB 10Krpm
300GB 10Krpm

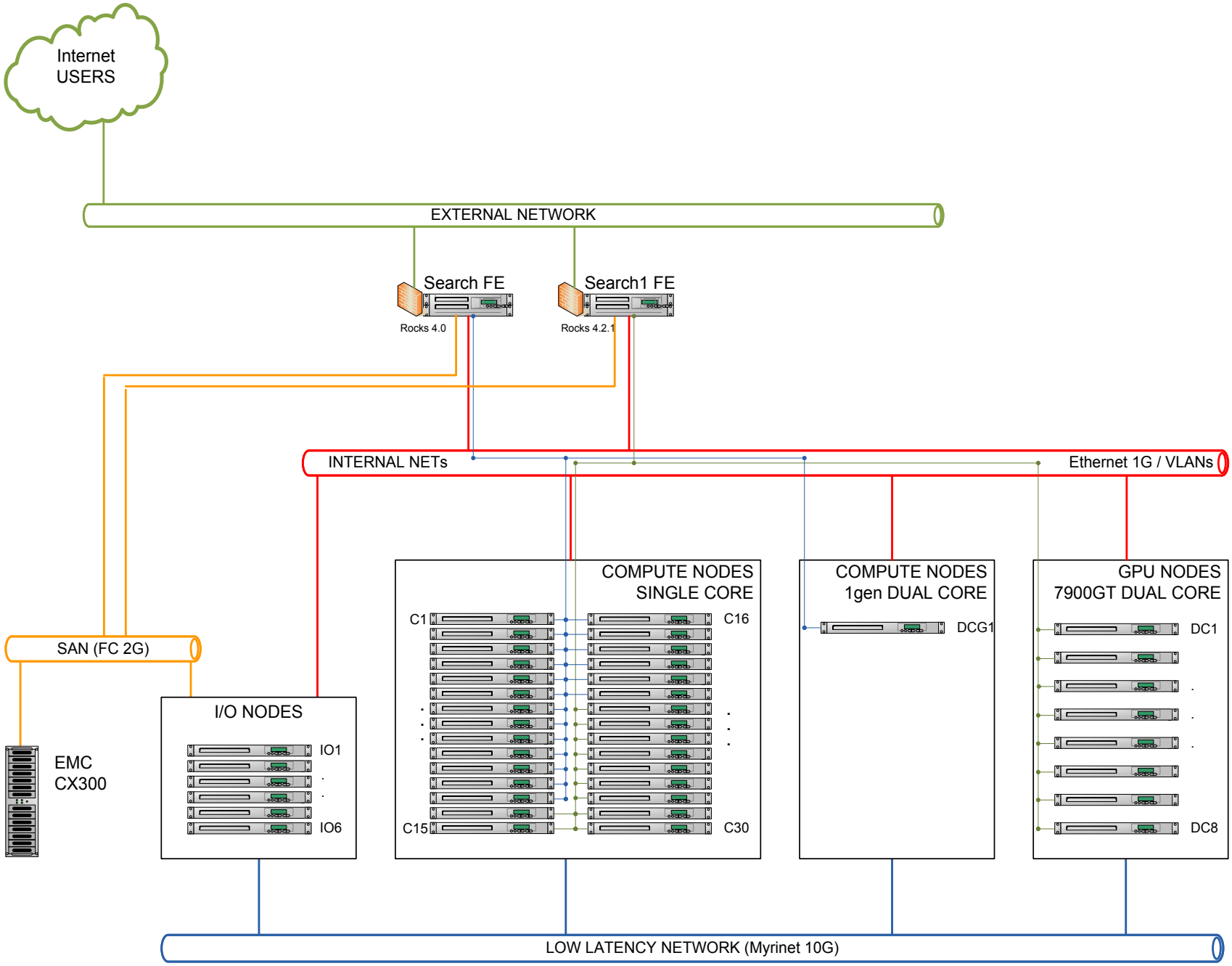

EMC

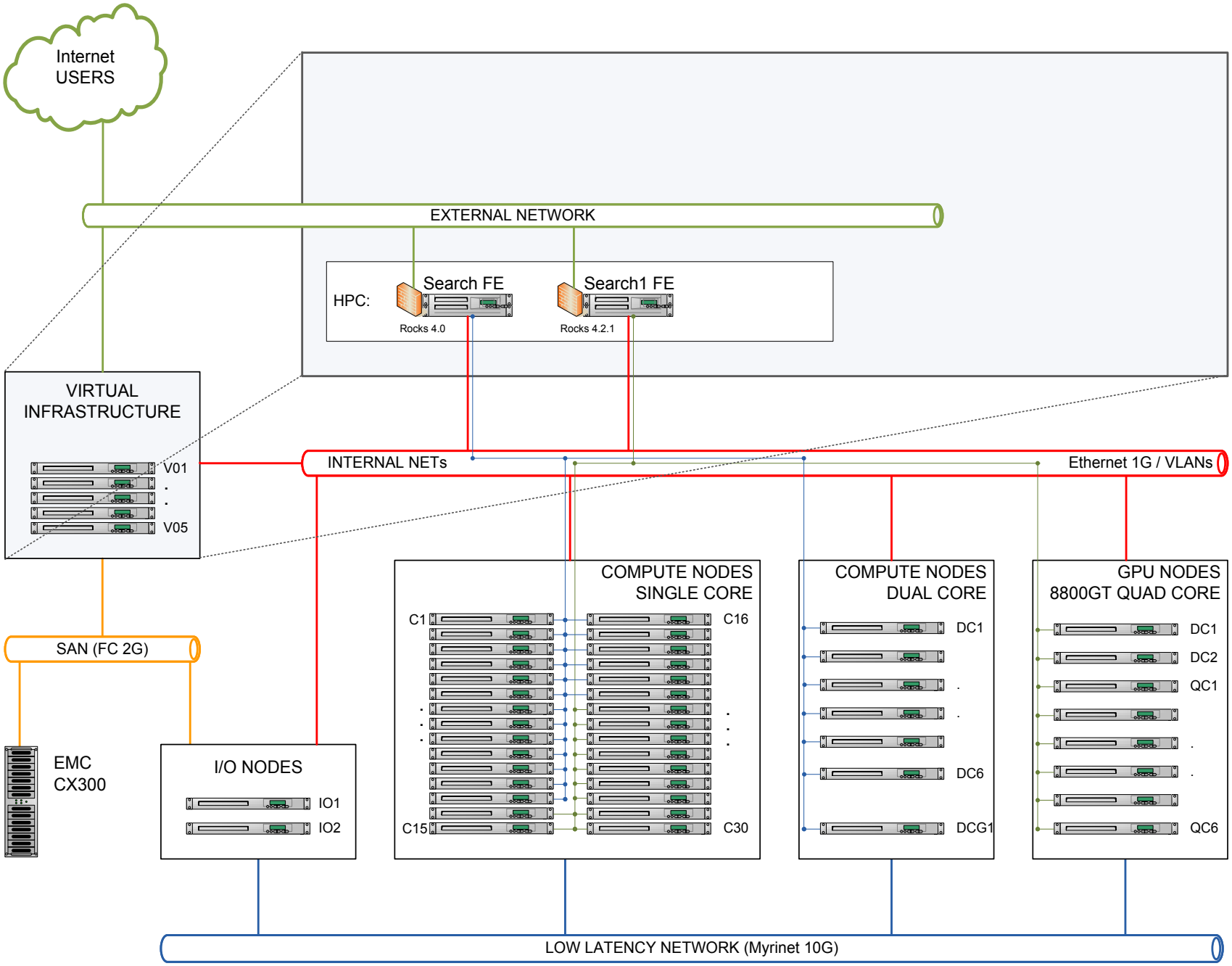
A ARQUITECTURA DO SISTEMA

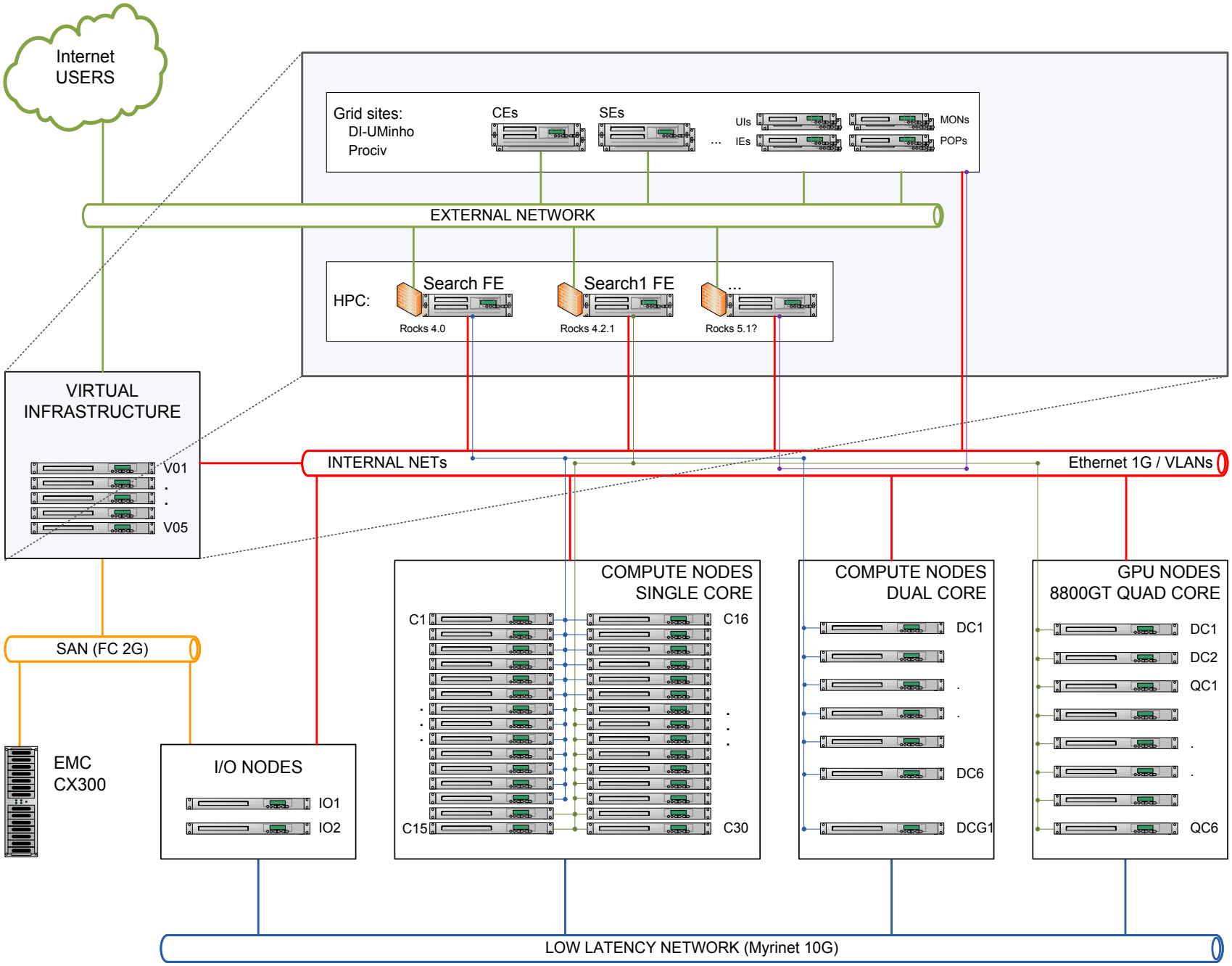
A circunstância de os investigadores que constituem o consórcio SeARCH pertencerem a diferentes áreas de investigação conduziu, naturalmente, à definição de requisitos específicos de cuja satisfação veio a resultar a definição de uma arquitectura heterogénea que se constitui em três tipos distintos de nós de computação – genéricos, de E/S e de computação vectorial – e três tecnologias de interligação - Gigabit Ethernet, Myrinet10G e Fibre Channel.

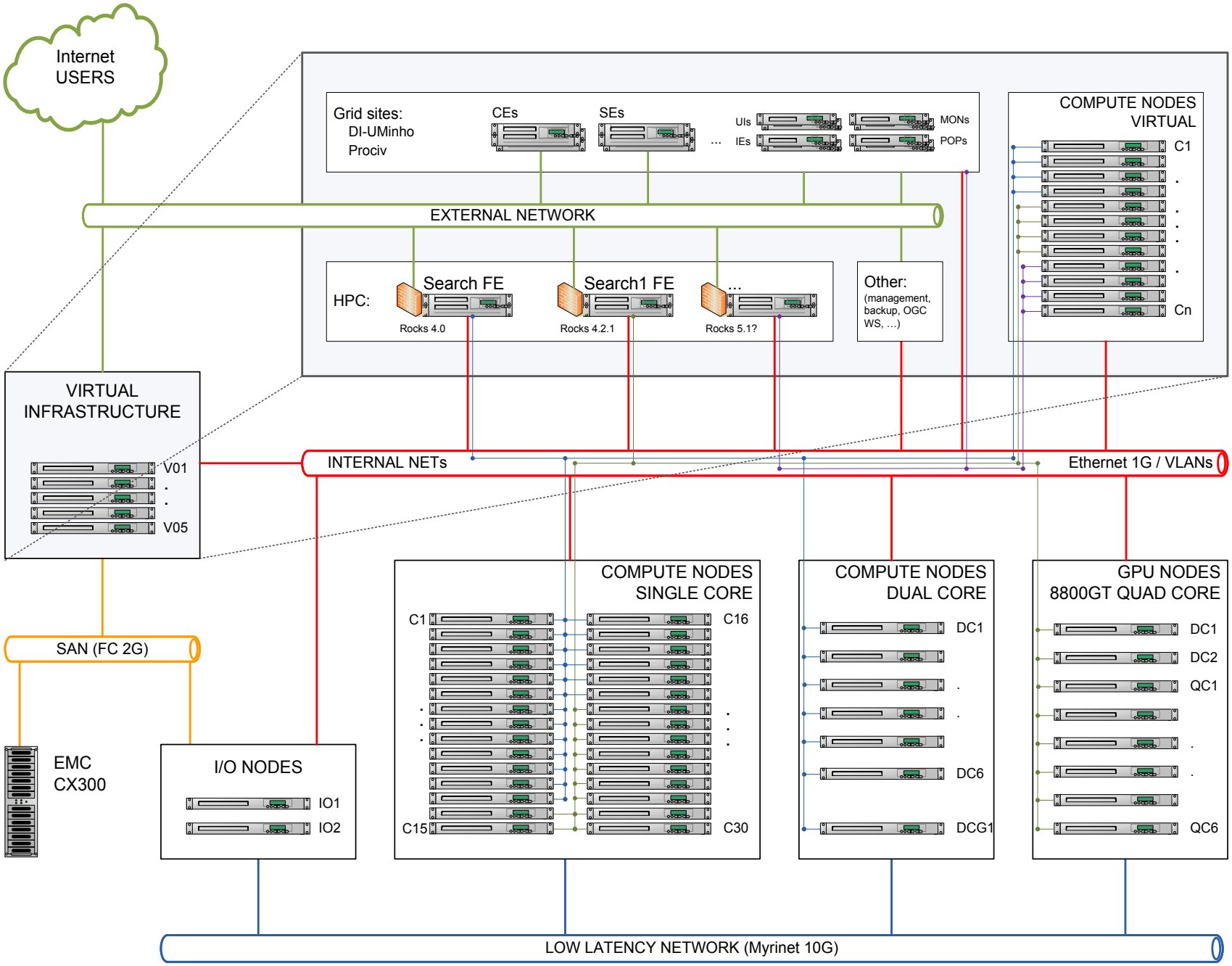
O armazenamento de dados assenta numa SAN EMC CX300 interligada com os nós frontais com os nós de E/S, inicialmente instalada com 3TB de espaço.

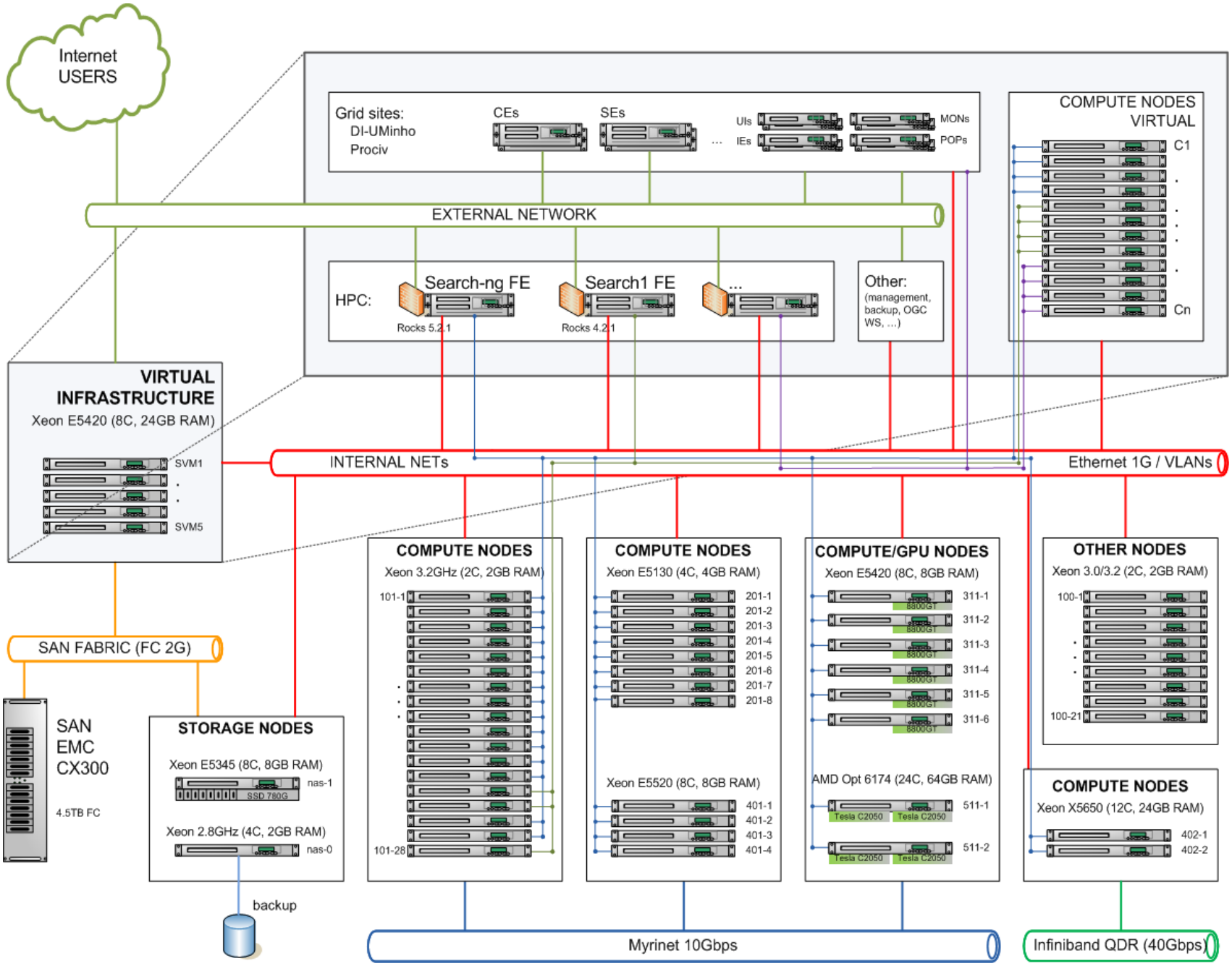
A gestão dos recursos do *cluster* é assegurada pelos dois nós frontais que, configurados de forma emparelhada, garantem os serviços de acesso dos utilizadores, a gestão das filas de acesso aos nós, o acesso aos dados, a monitorização e a instalação automática dos nós.


Internet
USERS
EXTERNAL NETWORK
Search FE
Rocks 4.0
Search1 FE
Rocks 4.2.1
INTERNAL NETs
Ethernet 1G / VLANs
SAN (FC 2G)
EMC
CX300
I/O NODES
IO1
IO6
COMPUTE NODES
SINGLE CORE
C1
C16
C15
C30
COMPUTE NODES
1gen DUAL CORE
DCG1
GPU NODES
7900GT DUAL CORE
DC1
DC8
LOW LATENCY NETWORK (Myrinet 10G)


Internet
USERS
EXTERNAL NETWORK
HPC:
Search FE
Rocks 4.0
Search1 FE
Rocks 4.2.1
VIRTUAL
INFRASTRUCTURE
V01
V05
INTERNAL NETs
Ethernet 1G / VLANs
SAN (FC 2G)
EMC
CX300
I/O NODES
IO1
IO2
COMPUTE NODES
SINGLE CORE
C1
C16
C15
C30
COMPUTE NODES
DUAL CORE
DC1
DC6
DCG1
GPU NODES
8800GT QUAD CORE
DC1
DC2
QC1
QC6
LOW LATENCY NETWORK (Myrinet 10G)


Internet
USERS
Grid sites:
DI-UMinho
Prociv
CEs
SEs
...
UIs
IEs
MONs
POPs
EXTERNAL NETWORK
HPC:
Search FE
Search1 FE
...
Rocks 4.0
Rocks 4.2.1
Rocks 5.1?
VIRTUAL
INFRASTRUCTURE
V01
V05
INTERNAL NETs
Ethernet 1G / VLANs
SAN (FC 2G)
EMC
CX300
I/O NODES
IO1
IO2
COMPUTE NODES
SINGLE CORE
C1
C16
C15
C30
COMPUTE NODES
DUAL CORE
DC1
DC6
DCG1
GPU NODES
8800GT QUAD CORE
DC1
DC2
QC1
QC6
LOW LATENCY NETWORK (Myrinet 10G)


Internet
USERS
Grid sites:
DI-UMinho
Prociv
CEs
SEs
...
UIs
IEs
MONs
POPs
COMPUTE NODES
VIRTUAL
C1
Cn
EXTERNAL NETWORK
HPC:
Search FE
Rocks 4.0
Search1 FE
Rocks 4.2.1
...
Rocks 5.1?
Other:
(management,
backup, OGC
WS, ...)
VIRTUAL
INFRASTRUCTURE
V01
V05
INTERNAL NETs
Ethernet 1G / VLANs
SAN (FC 2G)
EMC
CX300
I/O NODES
IO1
IO2
COMPUTE NODES
SINGLE CORE
C1
C16
C15
C30
COMPUTE NODES
DUAL CORE
DC1
DC6
DCG1
GPU NODES
8800GT QUAD CORE
DC1
DC2
QC1
QC6
LOW LATENCY NETWORK (Myrinet 10G)


Internet
USERS
Grid sites:
DI-UMinho
Prociv
CEs
SEs
UIs
IEs
MONs
POPs
COMPUTE NODES
VIRTUAL
C1
Cn
EXTERNAL NETWORK
HPC:
Search-ng FE
Rocks 5.2.1
Search1 FE
Rocks 4.2.1
...
Other:
(management, backup, OGC WS, ...)
VIRTUAL INFRASTRUCTURE
Xeon E5420 (8C, 24GB RAM)
SVM1
SVM5
INTERNAL NETs
Ethernet 1G / VLANs
SAN FABRIC (FC 2G)
SAN
EMC
CX300
4.5TB FC
STORAGE NODES
Xeon E5345 (8C, 8GB RAM)
SSD 780G
nas-1
Xeon 2.8GHz (4C, 2GB RAM)
nas-0
backup
COMPUTE NODES
Xeon 3.2GHz (2C, 2GB RAM)
101-1
101-28
COMPUTE NODES
Xeon E5130 (4C, 4GB RAM)
201-1
201-2
201-3
201-4
201-5
201-6
201-7
201-8
Xeon E5520 (8C, 8GB RAM)
401-1
401-2
401-3
401-4
COMPUTE/GPU NODES
Xeon E5420 (8C, 8GB RAM)
8800GT
311-1
311-2
311-3
311-4
311-5
311-6
AMD Opt 6174 (24C, 64GB RAM)
Tesla C2050
511-1
511-2
OTHER NODES
Xeon 3.0/3.2 (2C, 2GB RAM)
100-1
100-21
COMPUTE NODES
Xeon X5650 (12C, 24GB RAM)
402-1
402-2
Myrinet 10Gbps
Infiniband QDR (40Gbps)

A COMPUTAÇÃO

Os processadores instalados no sistema pertencem a tecnologias Intel de duas gerações distintas, a que correspondem as micro-arquitecturas NetBurst e Core.

Os 30 nós genéricos e os 6 nós de E/S baseiam-se em servidores de 1U com *chipset* Intel E7320, bi-processadores Intel Xeon a 3.2 GHz com 2MB de cache e 2GB de RAM.

Os 8 nós de computação com GPU distinguem-se dos demais nós por conterem dois processadores Dual Core Intel Xeon 5130 a 2GHz, 4GB de RAM e placas gráficas nVidia 7900GT.

A capacidade de processamento vectorial dos GPUs das placas gráficas, devido à sua natureza massivamente paralela, ultrapassa largamente os melhores processadores convencionais em termos de cálculo em vírgula flutuante. Há, assim, lugar à possibilidade de desenvolver algoritmos capazes de tirar partido daquelas características, o que pode revelar-se extremamente eficiente e económico na resolução de problemas tão díspares como a ordenação de bases de dados, o *data mining*, a computação científica ou o processamento de sinal.

Diagrama de um nó (original):

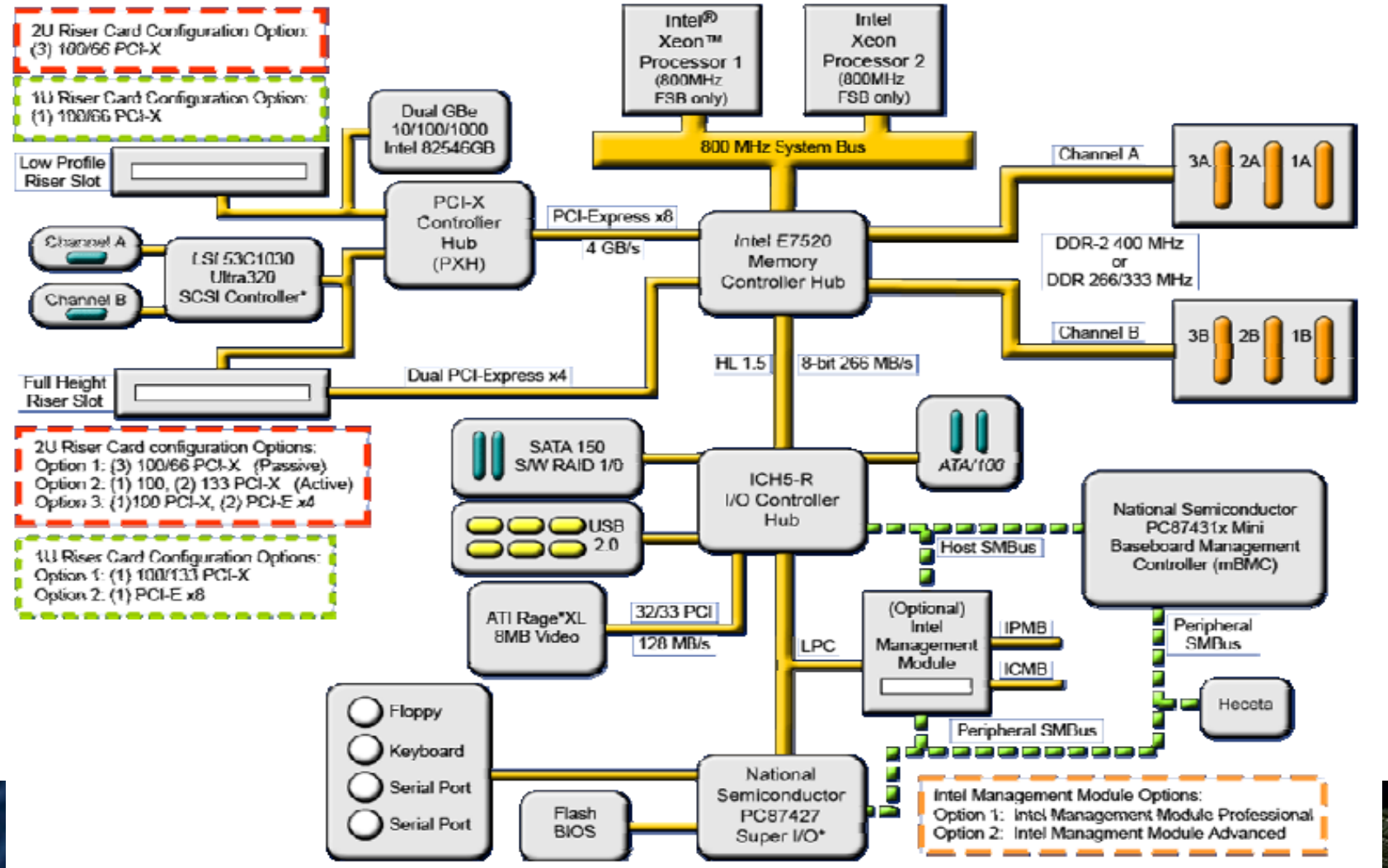


Diagrama de um nó (2ª geração):

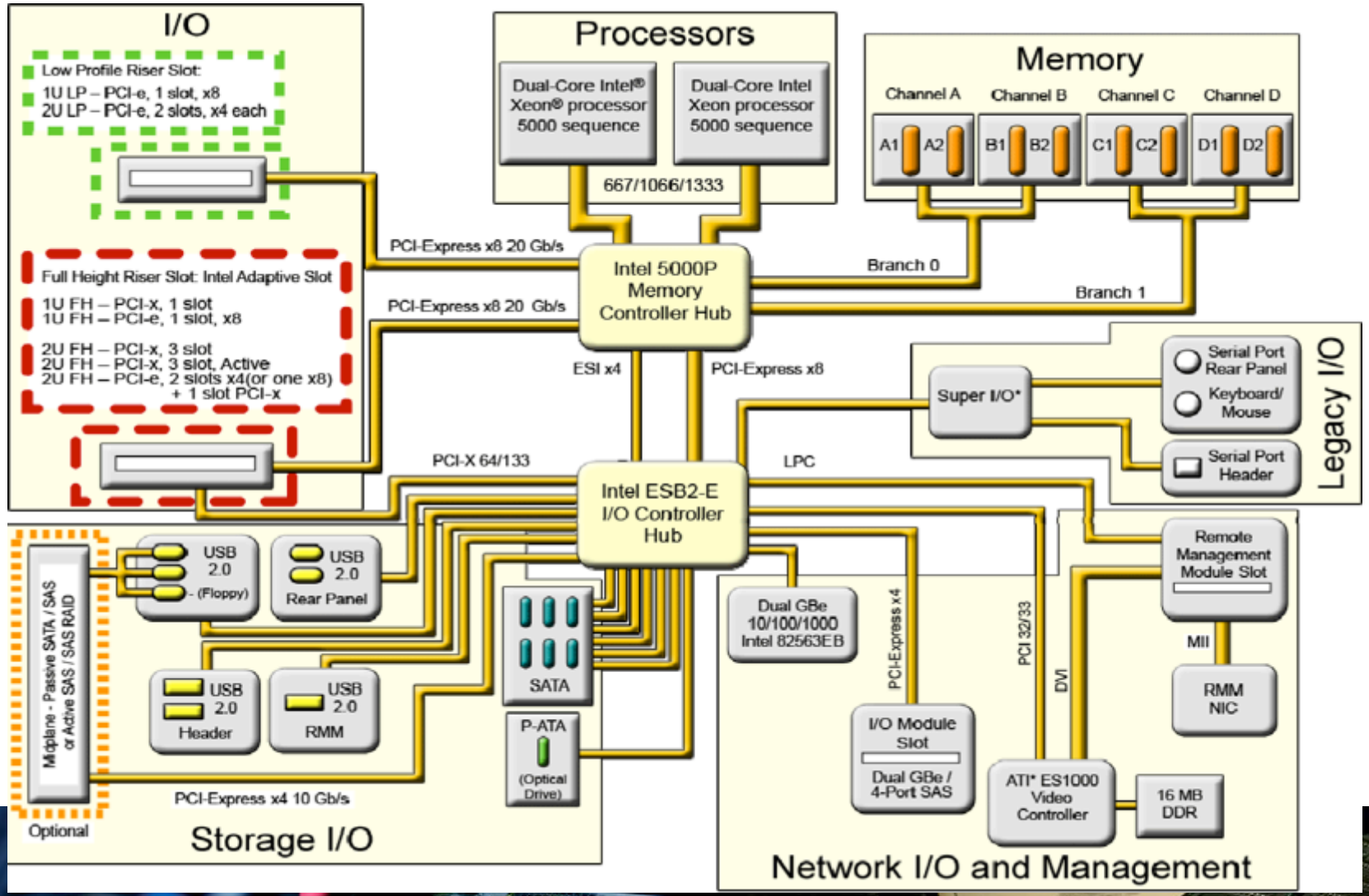
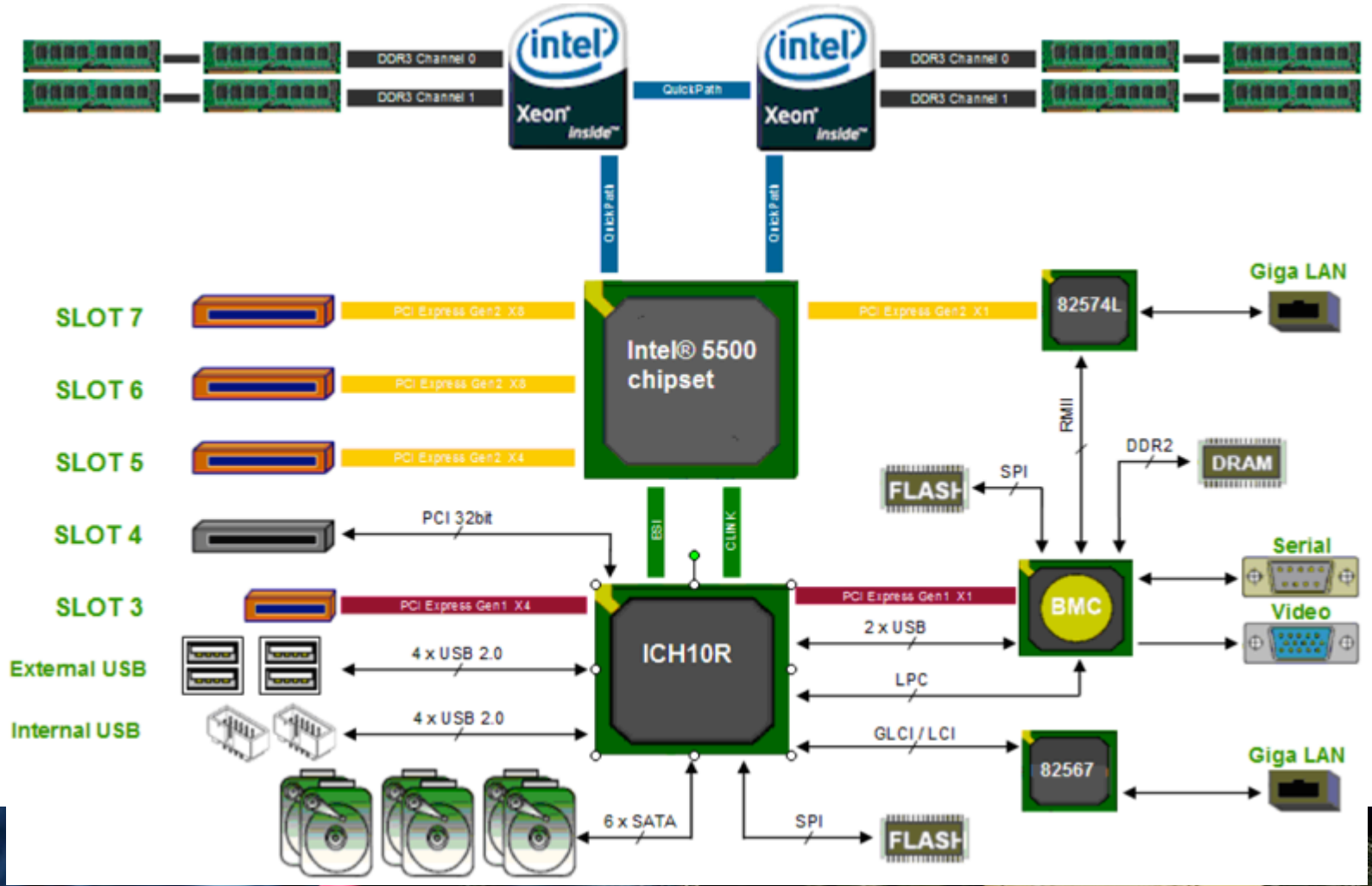
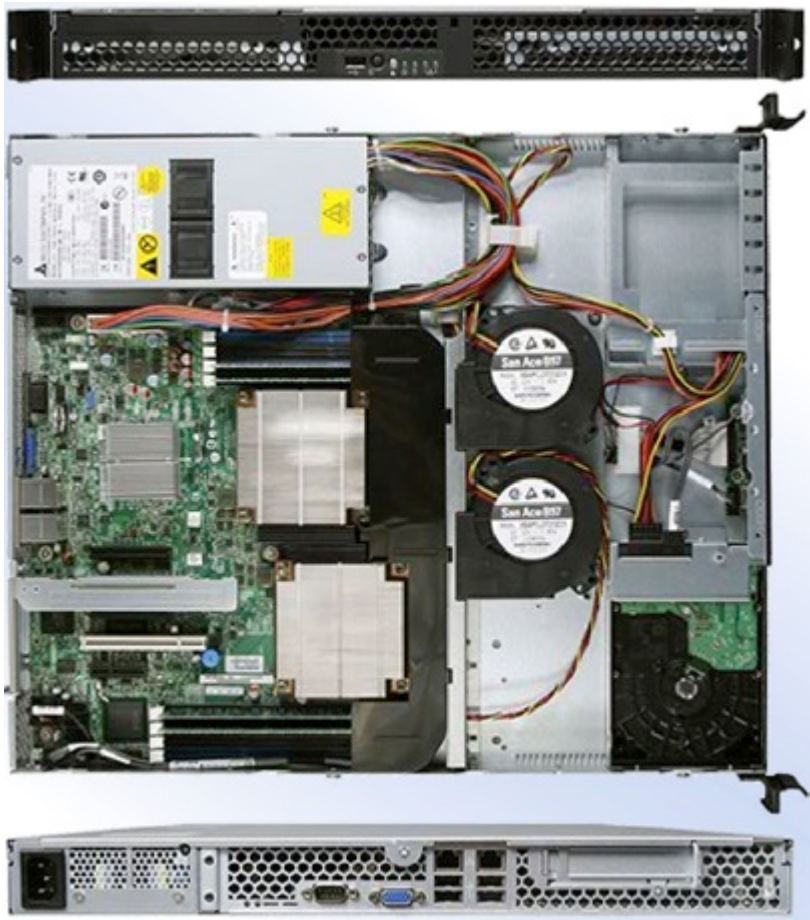
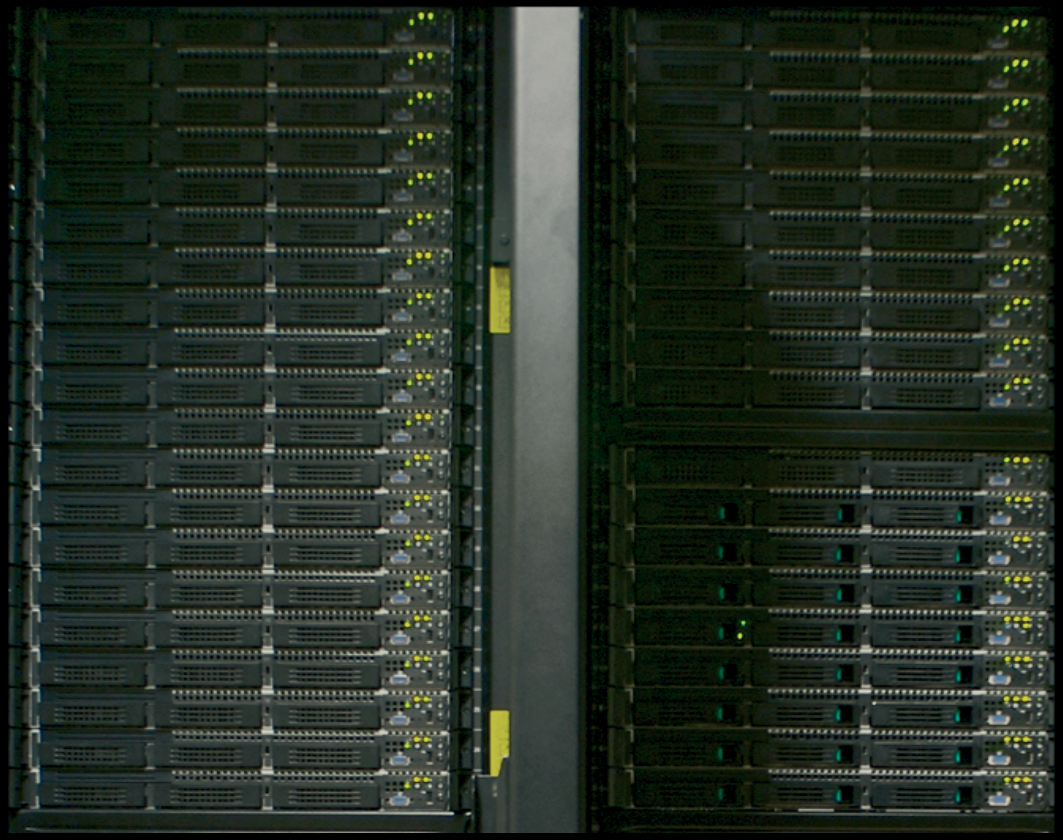


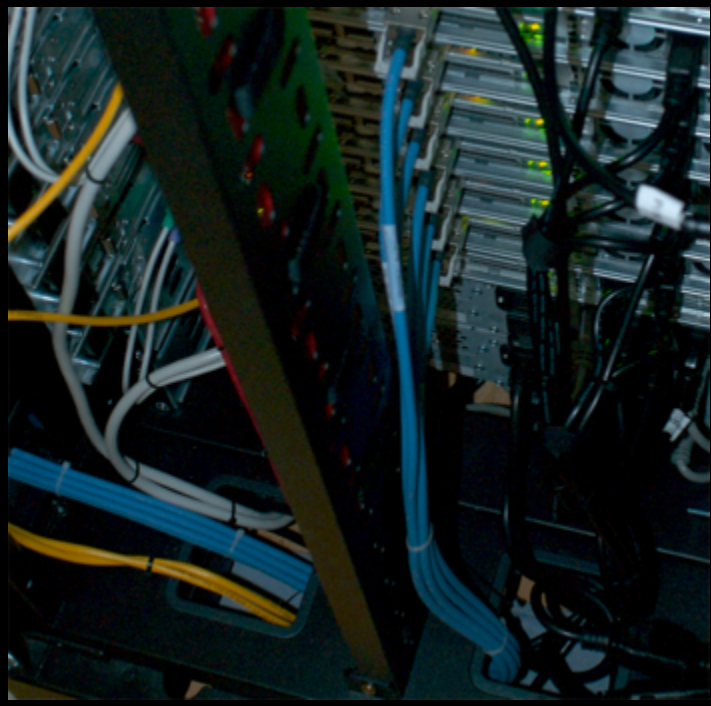
Diagrama de um nó (3ª geração):



Interior de um nó de computação (3ª geração):







AS COMUNICAÇÕES

O poder de cálculo instalado num determinado sistema não é por si só garante da qualidade dos resultados obtidos em termos de eficiência e escalabilidade. As tecnologias de comunicação utilizadas para a interligação dos nós de computação podem afectar de forma decisiva a eficiência e a rentabilidade obtidas. Com efeito, factores como a largura de banda e, particularmente, a latência da comunicação têm implicações dramáticas no desempenho de muitas aplicações científicas paralelas.

No *cluster* Search foi introduzida a nova geração da Myrinet – Myri10G – que aumenta a largura de banda útil para os 10 Gbps e, simultaneamente, reduz a latência para valores ainda mais baixos, da ordem dos 2.3µs. É de salientar que esta infra-estrutura de comunicações, compatível a nível físico com a Ethernet 10G, é a tecnologia de interligação de baixa latência mais utilizada nos 500 mais rápidos sistemas de computação do mundo (ver www.top500.org).

Adicionalmente, ao nível do subsistema de comunicação de rede Gigabit Ethernet, o Search foi especialmente afinado para atingir valores para latência da ordem dos 20-30µs e largura de banda próxima do limite teórico, tirando partido dos comutadores instalados com largura de banda interna de 96 Gbps e do suporte IOAT das placas de rede Intel PRO/1000 da última geração.


DO NOT USE FOR DPE CX500

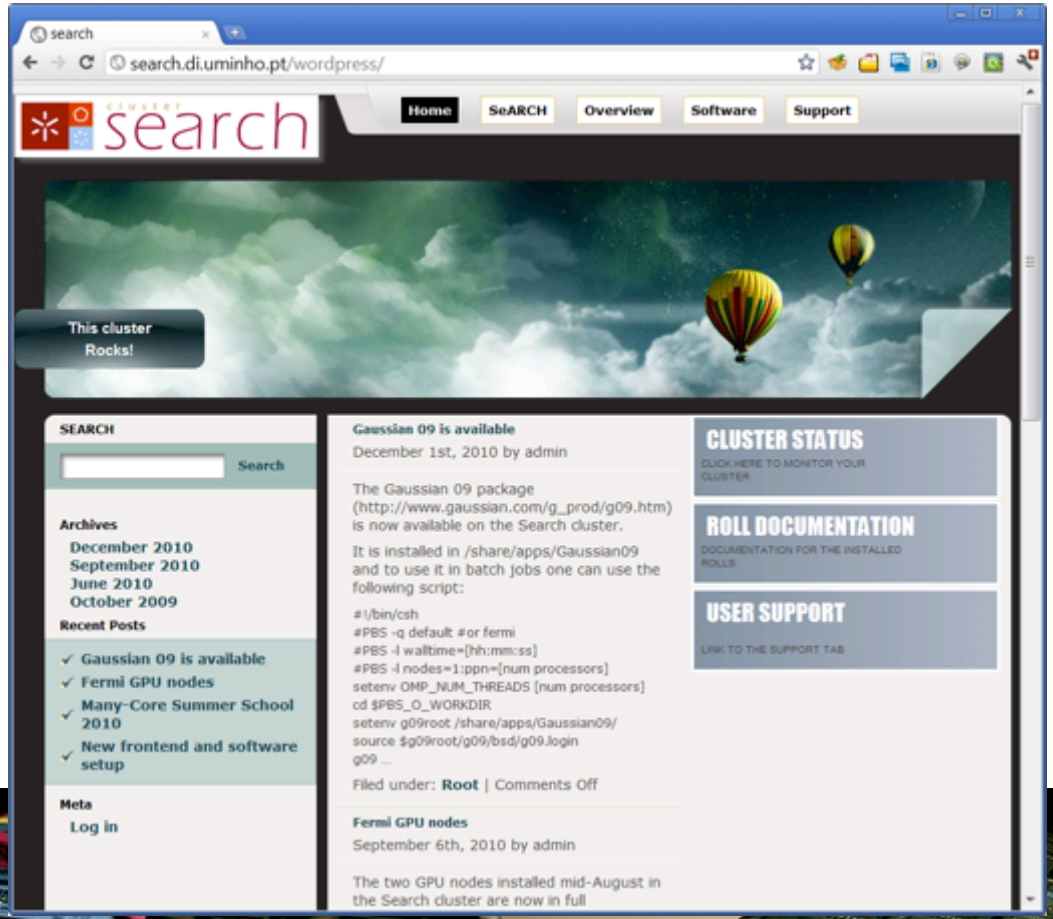

EMC


EMC

Breve introdução à utilização do cluster

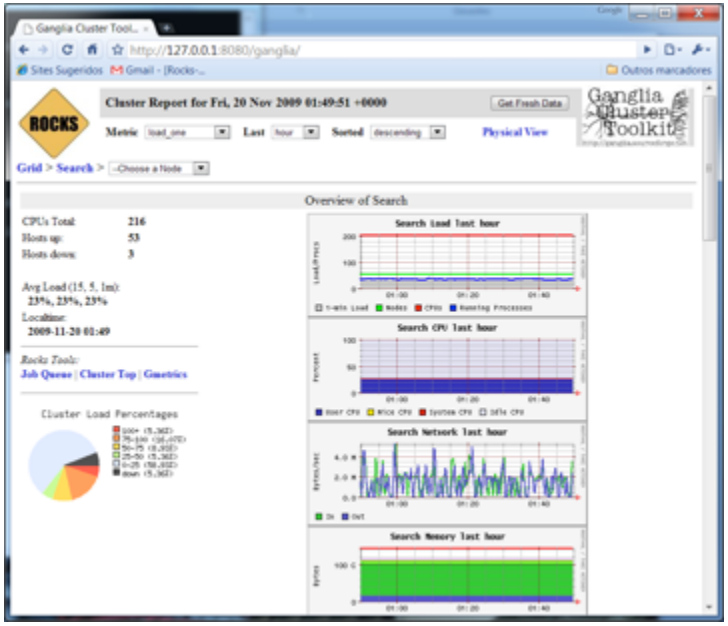
Home page:

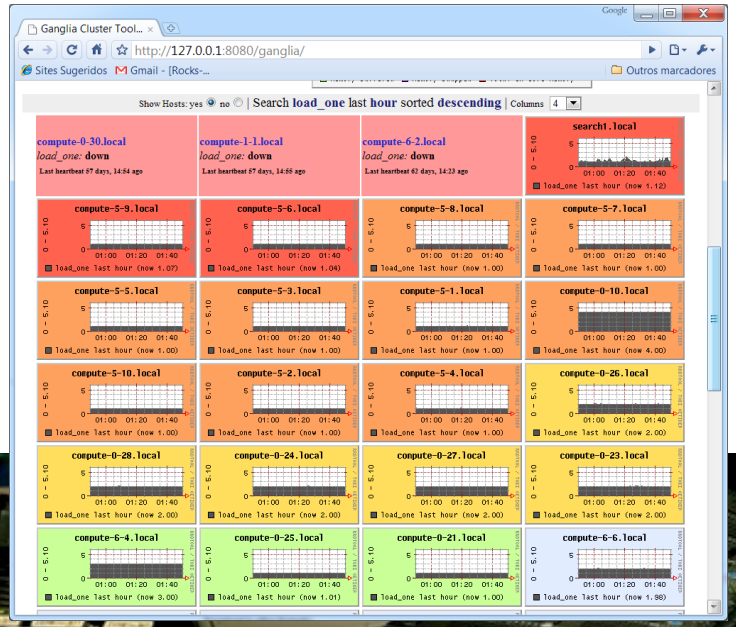
- O site do Search apresenta informações sobre o cluster em: http://search.di.uminho.pt/



Monitorização:

- Está disponível no cluster Search a ferramenta Ganglia que permite observar a carga do sistema.





Acesso ao cluster (Linux):

- Ligação ao servidor central por SSH (Secure Shell) para o login do utilizadores, compilação dos programas, transferência de ficheiros e submissão de trabalhos.
  - ssh search.di.uminho.pt –l <nome>
  - ssh <nome>@search.di.uminho.pt

- Caso o utilizador pretenda redireccionar a saída das aplicações X-Windows que correr no nó central poderá passar o parâmetro -X, que criará o túnel necessário para lhe dar suporte.

- Da mesma forma, poder-se-á utilizar-se o parâmetro -L porta-local:ip-remoto:porta-remota para redireccionar as portas remotas para portas locais.
  - Por exemplo, para ter acesso à porta de WWW remota (que está bloqueada pela firewall) deverá executar o comando "ssh -L 80:search.local:80 …".

Acesso ao cluster (Windows):

- O Windows não possui qualquer forma "nativa" de acesso ao cluster Search, mas existem diversas aplicações de acesso remoto por SSH. Entre os clientes de SSH disponíveis para Windows salientamos o PuTTY, que pode ser descarregado a partir do apontador:
  - http://www.chiark.greenend.org.uk/~sgtatham/putty

- Deverá ter em atenção que deverá utilizar o protocolo SSH versão 2 apenas, uma vez que a versão 1 tem lacunas de segurança.

Transferência de ficheiros:

- A transferência de ficheiros de/para o cluster Search assenta nos comandos **scp** e **sftp**, que são parte do conjunto de comandos do SSH.

- Os resultados dos processamentos deverão ser salvaguardados pelo utilizador, uma vez que não se oferece o serviço de cópia automática dos dados dos utilizadores.

Transferência de ficheiros (Linux):

O SSH fornece duas ferramentas para transferência de ficheiros, o scp e o sftp.

- O comando scp é um extensão do comando cp e tem a seguinte sintaxe:
  - scp [[user@]host1:]file1 [...] [[user@]host2:]file2
  - O exemplo seguinte mostra como é possível transferir um ficheiro local chamado **exemplo.tgz**, que será colocado na directoria /home/xxxx do cluster.

    machine:~/... xxxx$ scp exemplo.tgz xxxx@search.di.uminho.pt:/home/
    xxxxxxxx@search.di.uminho.pt's password: exemplo.tgz
    100% 173KB 172.9KB/s   00:00

    machine:~/... xxxx$ scp xxxx@search.di.uminho.pt:/home/xxxx/exemplo.tgz .
    xxxx@search.di.uminho.pt's password: exemplo.tgz
    100% 173KB 172.9KB/s   00:00
  - É possível copiar directorias inteiras através do parâmetro -r.

    machine:~/... xxxx$ scp -r programa-exemplo/ xxxx@search.di.uminho.pt:/home/
    xxxxxxxx@search.di.uminho.pt's password:
    job-result-01.png 100% 294KB 293.6KB/s   00:00
    logo.png 100%   26KB 26.4KB/s   00:00
    logo_small.png 100%   23KB 22.8KB/s
    00:00   ...users_guide.html 100%   43KB 43.1KB/s   00:00

Transferência de ficheiros (Windows):

- O Windows não possui os comandos scp e sftp. O pacote PuTTY fornece dois comandos que os substituem, o pscp e o psftp. A forma de transferência é a mesma que para os comandos apresentados na secção anterior.

- Em alternativa poderá utilizar o pacote WinSCP (http://winscp.net) que implementa estas funcionalidades através de uma interface semelhante ao explorador do Windows.

Compilação de programas:

- Estão disponíveis no Search diversos compiladores, incluindo não só os tradicionais GNU como também os compiladores Intel.
- Estão também disponíveis diversas bibliotecas para a execução de trabalhos em paralelo, que possuem wrappers próprios para a compilação de programas paralelos.
- Os utilizadores deverão escolher os pacotes que forem mais adequados ao que pretendem e é da sua responsabilidade configurar as variáveis do ambiente de acordo com o que for necessário (module avail).

Submissão de trabalhos:

- A submissão de trabalhos deverá utilizar os comandos do Torque (baseado no PBS) e do escalonador Maui.

- O desafio de encontrar a configuração bem balanceada entre o requisição de máquinas, a duração dos trabalhos, as prioridades e a gestão do sistema nunca está completamente resolvido. Assim, as políticas de gestão das filas serão afinadas regularmente num esforço para maximizar a eficiência e a justiça no acesso aos recursos.

- O sistema de gestão de filas deve ser utilizado em todos os trabalhos de computação. Em princípio não será necessário efectuar o login no nós de computação, excepto se for necessário terminar processos que, por algum motivo, tenham ficado pendentes.

- O servidor de acesso não deve ser utilizado para computação.

Submissão de trabalhos: novo trabalho

- Ao criar trabalhos é uma boa prática colocá-los em directorias separadas. Novas directorias deverão ser criadas com o comando **mkdir**.
- Para este exemplo iremos criar um trabalho chamado “exemplo”.
  - [xxxx@search xxxx]$ mkdir exemplo
  - [xxxx@search xxxx]$ cd exemplo
  - [xxxx@search exemplo]$ _
- Para criar um novo trabalho execute o editor de texto vi:
  - [xxxx@search exemplo]$ vi exemplo.scr
- E introduza o seguinte texto:
  - #!/bin/sh
  - #PBS -l nodes=1
  - #PBS -l walltime=05:00
  - echo "Olá mundo!"sleep 60
- Depois deverá gravar este script (com [Shift]Z+Z ou “:w”, por exemplo).

Submissão de trabalhos: submeter

- Após a criação do script poderá submetê-lo para execução pelo sistema de gestão de filas através do comando **qsub**, utilizando o nome do script como parâmetro.
  - [xxxx@search exemplo]$ qsub exemplo.scr
  - 41209.d1
- Se o trabalho foi correctamente processado o identificador do novo trabalho será apresentado como resposta do comando (41209.dl no exemplo apresentado). Este identificador poderá ser utilizado para gerir em comando posteriores os trabalhos submetidos.
- Para obter informação sobre o progresso dos trabalhos poderá utilizar os comandos **qstat** ou **showq**, que apresentam a lista dos trabalhos em execução e em espera.
  - [xxxx@search exemplo]$ qstat
  - Job id Name User Time Use S Queue
    ----- ---------- ---------- -------- - -----
    36305.d1 Pol14 usr1 30:52:57 R
    workq36311.d1 job.sh fak 156:26:0 R
    workq41209.d1 exemplo.scr xxxx 00:00:00 R workq
- A partir deste quadro podemos verificar que o trabalho 41209 está em execução (R). Quando o trabalho terminar deixará de aparecer nesta lista.

Submissão de trabalhos: resultados

- Quando o trabalho terminar a directoria onde foi submetido irá conter dois ficheiros adicionais, um com as mensagens de erro (stderr) e outro com a saída para a consola (stdout).
  - [xxxx@search exemplo]$ ls sexemplo.scr exemplo.scr.e41209 exemplo.scr.o41209

- Ao imprimir o conteúdo dos ficheiros aparecerá o seguinte resultado:
  - [xxxx@search exemplo]$ cat exemplo.scr.o41209
  - Olá mundo!
  - [xxxx@search exemplo]$ cat exemplo.scr.e41209
  - [xxxx@search exemplo]$ _

Referência rápida do Torque:

- Os comandos PBS mais frequentemente utilizados são:
  - qsub {script} — Submeter o {script} para execução.
  - qdel {identificador} — Apagar o trabalho com o {identificador}.
  - Qstat — Devolve a lista dos trabalhos submetidos para execução. O mesmo que o comando maui **showq**.

- As variáveis mais frequentemente utilizados são:
  - PBS_JOBNAME — O nome do trabalho especificado pelo utilizador.
  - PBS_O_WORKDIR — A directoria onde o trabalho foi submetido.
  - PBS_TASKNUM — O número de tarefas solicitado.
  - PBS_O_HOME — A directoria home do utilizador que submete o trabalho.
  - PBS_O_SHELL — A shell utilizada pelo script.
  - PBS_O_JOBID — O identificador PBS do trabalho.
  - PBS_O_HOST — O nome do host onde o trabalho está a ser executado.
  - PBS_NODEFILE — O ficheiro com o nome dos nós onde o trabalho está a ser executado.
  - PBS_O_PATH — A variável PATH utilizada nos scripts.

Sessão de trabalho

Tarefas (1/4):

- Ligar ao cluster através do terminal SSH
  username: ami0, password: “ami.0:”

- Compilar o código em C do programa de cálculo paralelo de PI com o **mpicc**

- Executar o trabalho em modo interactivo
  qsub –I ...

Tarefas (2/4):

- Criar o ficheiro PBS de descrição do trabalho

- Submeter o trabalho em modo não interactivo

  qsub ...

- Acompanhar a execução dos trabalhos

  qstat/showq ...

Tarefas (3/4):

- Escolher a configuração do ambiente de trabalho pretendido com o **modules**

  module avail/load/...

- Compilar o cpi com o debugger MPIP

  mpicc –o cpi cpi.c $MPIPLINK

- Ver o conteúdo do ficheiro .mpiP gerado

Tarefas (4/4):

- Executar o CPI em um, dois e quatro nós de computação
- Seleccionar dois processos em cada nó
- Seleccionar nós com a mesma configuração
- Seleccionar nós em modo exclusivo

Alguns números:

- Configuração inicial (2006):
  - 920 Gbps Myrinet-10G
  - 92 Gbps Gb Ethernet
  - 4,5 TB de armazenamento em SAN
  - 3,7 TB de armazenamento em disco local
  - 17,5 kW de consumo de energia eléctrica
  - 59,5 kBTU/h de calor gerado
  - 1,5 toneladas

- Poder de cálculo:
  inicial: 46 nós, 108 cores, 108GB de RAM – 742 GFLOPS
  actual: 78 nós, 338 cores, 418GB de RAM – 2868 GFLOPS