Implementação de um supercomputador: Cluster GradeBR/UFAL

Baltazar Tavares Vanderlei
Leonardo Viana Pereira

Laboratório de Computação Científica e Visualização - LCCV/UFAL

19 de Outubro de 2010

Sumário

Sumário

O que é a GradeBR?

- Quem participa da GradeBR?

O que é a GradeBR?

- Quem participa da GradeBR?
  - UFRJ, USP, PUC-Rio, ITA

O que é a GradeBR?

- Quem participa da GradeBR?
  - UFRJ, USP, PUC-Rio, ITA
- O **LCCV** entrou como membro para ser um nó da GradeBR

Desafios como membro da GradeBR

Desafio

Usar tecnologia de ponta para planejar e implementar um grid de processamento de alto desempenho, que pudesse processar problemas de escala peta(da escala de 10^{15}) de forma cooperativa entre os vários nós.

Sumário

Necessario:

- Grande poder de processamento e memória
- Grande espaço e velocidade de armazenamento
- Uma rede de interconexão extremamente mais rápida que a convencional
- Um sistema tolerante a falhas, robusto e funcional(tanto em hardware como em software)

Sumário

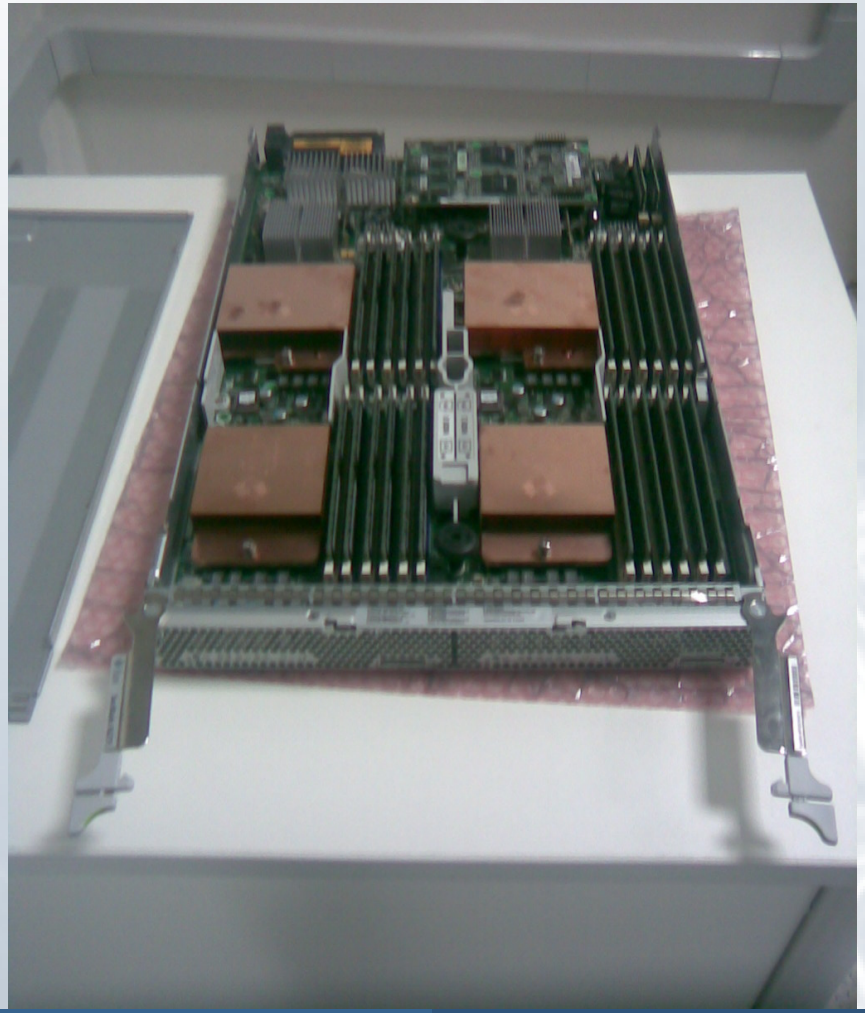
O cluster do LCCV possui:

- **8** placas de vídeo totalizando **30**Tflops
- **218** nós de processamento, com processadores **i7**
- Cada maquina com **2** processadores, cada processador **4** núcleos
- Cada maquina com **24**GB de memória NUMA
- Totalizando mais de **5**TB de memória NUMA e **1744** núcleos
- Só de nós de processamento, temos **20 Tflops**

Blades:

Blades:

Blades:



Sumário

Porque foi escolhida essa topologia e interconexão

- Para a rede de alto desempenho, foi escolhido o InfiniBand(IB)

Porque foi escolhida essa topologia e interconexão

- Para a rede de alto desempenho, foi escolhido o InfiniBand(IB)
- O IB é um meio com baixa latência

Porque foi escolhida essa topologia e interconexão

- Para a rede de alto desempenho, foi escolhido o InfiniBand(IB)
- O IB é um meio com baixa latência
- Tem uma alta taxa de transferência

Porque foi escolhida essa topologia e interconexão

- Para a rede de alto desempenho, foi escolhido o InfiniBand(IB)
- O IB é um meio com baixa latência
- Tem uma alta taxa de transferência
- É usado para conexão entre maquinas(compatível com MPI)

Porque foi escolhida essa topologia e interconexão

- Para a rede de alto desempenho, foi escolhido o InfiniBand(IB)
- O IB é um meio com baixa latência
- Tem uma alta taxa de transferência
- É usado para conexão entre maquinas(compatível com MPI)
- É usado por dispositivos de armazenamento(compatível com o lustre)

Porque foi escolhida essa topologia e interconexão

- Para a rede de alto desempenho, foi escolhido o InfiniBand(IB)
- O IB é um meio com baixa latência
- Tem uma alta taxa de transferência
- É usado para conexão entre maquinas(compatível com MPI)
- É usado por dispositivos de armazenamento(compatível com o lustre)
- Pode ser usada uma camada de compatibilidade com o IP(chamada de "IPoIB")

Porque foi escolhida essa topologia e interconexão

- Para a rede de alto desempenho, foi escolhido o InfiniBand(IB)
- O IB é um meio com baixa latência
- Tem uma alta taxa de transferência
- É usado para conexão entre maquinas(compatível com MPI)
- É usado por dispositivos de armazenamento(compatível com o lustre)
- Pode ser usada uma camada de compatibilidade com o IP(chamada de "IPoIB")
- Com IB, foi conseguido uma taxa de transferência máxima de 40Gbit/s

Topologia adotada:

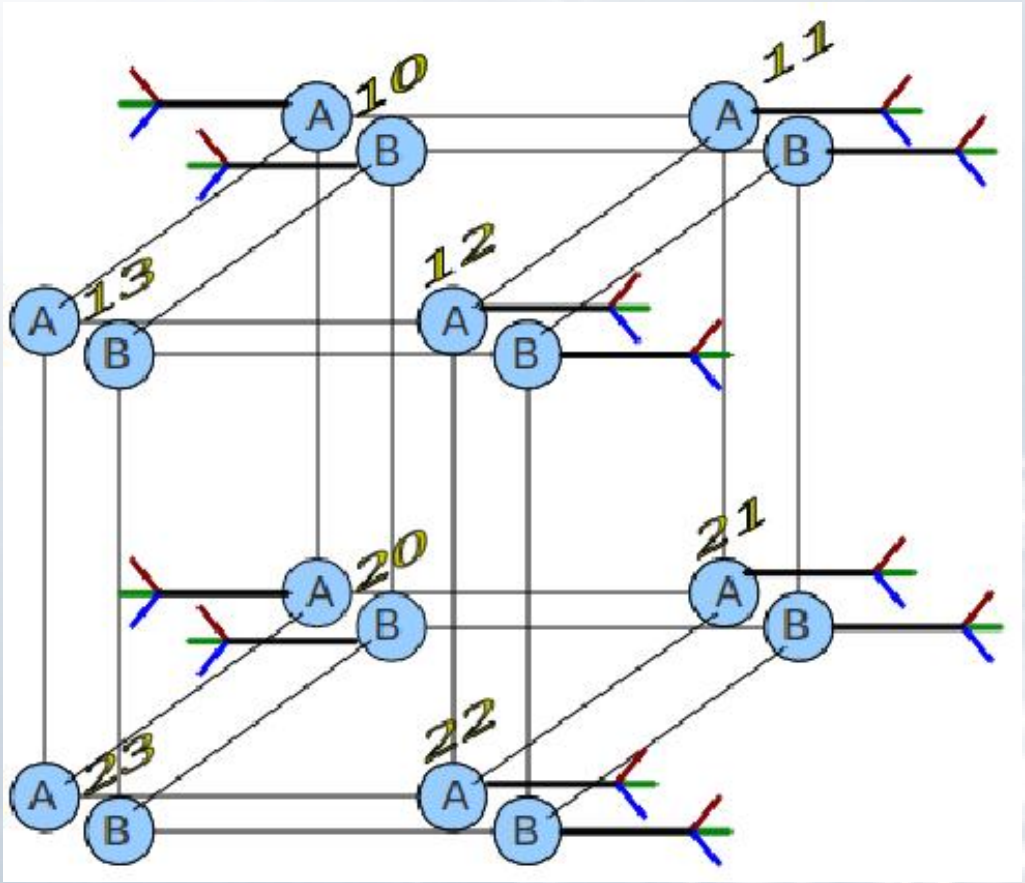


Figura: hipercubo 4D

Sumário

Para armazenamento de alto desempenho, era nescessário:

- Um sistema de arquivo que funcionasse via rede

Para armazenamento de alto desempenho, era nescessário:

- Um sistema de arquivo que funcionasse via rede
- Um sistema de arquivo paralelo

Para armazenamento de alto desempenho, era nescessário:

- Um sistema de arquivo que funcionasse via rede
- Um sistema de arquivo paralelo
- Escalável para um grande numero de clientes

Para armazenamento de alto desempenho, era nescessário:

- Um sistema de arquivo que funcionasse via rede
- Um sistema de arquivo paralelo
- Escalável para um grande numero de clientes
- Compatível com o hardware usado

Porque foi escolhido o lustrefs:

- Sistema de arquivos via rede e paralelo

Porque foi escolhido o lustrefs:

- Sistema de arquivos via rede e paralelo
- Possível usar raid e garantir segurança e acesso rápido a dados

Porque foi escolhido o lustrefs:

- Sistema de arquivos via rede e paralelo
- Possível usar raid e garantir segurança e acesso rápido a dados
- Escalável ate dezenas de milhares de clientes

Porque foi escolhido o lustrefs:

- Sistema de arquivos via rede e paralelo
- Possível usar raid e garantir segurança e acesso rápido a dados
- Escalável ate dezenas de milhares de clientes
- Suporte a IB, usando rdma para se comunicar diretamente

Porque foi escolhido o lustrefs:

- Sistema de arquivos via rede e paralelo
- Possível usar raid e garantir segurança e acesso rápido a dados
- Escalável ate dezenas de milhares de clientes
- Suporte a IB, usando rdma para se comunicar diretamente
- Tolerância a falhas e Alta disponibilidade(sem balanço de carga)

Sumário

O que aumenta a dificuldade com o sistema:

- Um sistema com muitos clientes

O que aumenta a dificuldade com o sistema:

- Um sistema com muitos clientes
- Alta disponibilidade e balanço de carga em serviços

O que aumenta a dificuldade com o sistema:

- Um sistema com muitos clientes
- Alta disponibilidade e balanço de carga em serviços
- Lidar com o sistema de varias maquinas ao mesmo tempo

O que aumenta a dificuldade com o sistema:

- Um sistema com muitos clientes
- Alta disponibilidade e balanço de carga em serviços
- Lidar com o sistema de varias maquinas ao mesmo tempo
- Lidar com programas escalonadores

O que aumenta a dificuldade com o sistema:

- Um sistema com muitos clientes
- Alta disponibilidade e balanço de carga em serviços
- Lidar com o sistema de varias maquinas ao mesmo tempo
- Lidar com programas escalonadores
- Vários problemas por lidar com tecnologia de ponta

O que aumenta a dificuldade com o sistema:

- Um sistema com muitos clientes
- Alta disponibilidade e balanço de carga em serviços
- Lidar com o sistema de varias maquinas ao mesmo tempo
- Lidar com programas escalonadores
- Vários problemas por lidar com tecnologia de ponta
- Sistema muito grande e complexo

Sumário

Sumário

HPL:

O que é o HPL?

HPL é um teste amplamente usado que mede a eficiência de um cluster em flops.

- O Cluster teve um resultado de 17TFlops.
- Resultados parciais com eficiência superior a 85%(R_{max}/R_{peak}).

Sumário

IOR:

O que é o IOR?

IOR é um teste usado que mede a escrita e leitura de um cluster em um sistema de arquivos usando posix e mpi-io.

Tabela: *Resultados do IOR*

POSIX [GB/s]	MPI-IO [GB/s]
Leitura — Escrita	Leitura
6,8 — 2,7	6

Sumário

Estado atual

- Implementamos com sucesso um cluster de alto desempenho

Estado atual

- Implementamos com sucesso um cluster de alto desempenho
- Maior supercomputador em atividade na América Latina

Estado atual

- Implementamos com sucesso um cluster de alto desempenho
- Maior supercomputador em atividade na América Latina
- Foram executadas mais de 500 mil horas de processamento em projetos do LCCV/Petrobras

Estado atual

- Implementamos com sucesso um cluster de alto desempenho
- Maior supercomputador em atividade na América Latina
- Foram executadas mais de 500 mil horas de processamento em projetos do LCCV/Petrobras
- Preparando a infraestrutura para o grid continental de alto desempenho GradeBR

Agradecimentos

Agradecemos a ANP, a Petrobras e ao Laboratório de Computação Científica e Visualização da Universidade Federal de Alagoas por garantir acesso aos recursos computacionais do cluster GradeBR/UFAL da Rede Galileu.