Parque Computacional de Alto Desempenho (PCAD)

Grupo de Processamento Paralelo e Distribuído (GPPD-HPC)

Lucas Mello Schnorr

Instituto de Informática, UFRGS

Encontro de Usuários do CESUP
Sala Abacateiro, do Centro Cultural da UFRGS
Porto Alegre, 19 de agosto de 2024, 13h

INSTITUTO DE INFORMÁTICA UFRGS

UFRGS UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Histórico

Plataforma em constante evolução

- Depende fortemente dos projetos dos professores
  - Heterogeneidade computacional
- Autogerida, com esforço de pós-graduandos voluntários
  - Discentes passam pelo sistema

Histórico

Plataforma em constante evolução

- Depende fortemente dos projetos dos professores
  - Heterogeneidade computacional
- Autogerida, com esforço de pós-graduandos voluntários
  - Discentes passam pelo sistema

Preocupação com reprodutibilidade

- Implantação de um sistema de gerenciamento
- Configuração padrão para todas as máquinas
- Colocar os usuários (*experts*) em primeiro plano

Histórico

Plataforma em constante evolução

- Depende fortemente dos projetos dos professores
  - Heterogeneidade computacional
- Autogerida, com esforço de pós-graduandos voluntários
  - Discentes passam pelo sistema

Preocupação com reprodutibilidade

- Implantação de um sistema de gerenciamento
- Configuração padrão para todas as máquinas
- Colocar os usuários (*experts*) em primeiro plano

Primeira organização geral em torno de ≈2018/2019

Parque Computacional de Alto Desempenho (PCAD)

`https://gppd-hpc.inf.ufrgs.br/`

Parque Computacional de Alto Desempenho (PCAD)

URL: `http://gppd-hpc.inf.ufrgs.br/`

Possui aproximadamente 50 nós, 1000+ núcleos de CPU e 100000+ de GPU

Localização: *Data Center* do Instituto de Informática, UFRGS

Algumas configurações selecionadas de HW para cálculo

Processadores Intel Xeon

- E5530 Nehalem, X7550 Nehalem
- E5-2630 Sandy Bridge
- E5-2640 v2 Ivy Bridge
- E5-2650 v3 Haswell
- E5-2699 v4 Broadwell
- Gold (5317 Ice Lake, 6226 C. Lake)
- Silver (4208 C. Lake, 4116 Skylake)

Processadores Intel Core

- Intel Core i7-10700F, i7-14700KF
- Intel Core i9-14900KF

Processadores NVIDIA

- 2× NVIDIA Grace CPU Superchip

Processadores AMD

- AMD Ryzen 9 3950X Zen2
- AMD RYZEN 5 3400G Zen+

Aceleradoras NVIDIA

- 6× GTX1080Ti
- 4× P100
- 1× RTX3090
- 5× RTX4070
- 6× RTX4090

Outras aceleradoras

- 4× NEC 10BE SX-Aurora
- 3× AMD Radeon RX 7900 XT

Requisitos almejados da infraestrutura

Para os usuários

- Serem independentes para realização de experimentos
- Fácil utilização (assumindo claro um conhecimento de Linux)
- Acesso isolado, irrestrito e com máximo desempenho aos recursos (*bare metal*)
- Algumas facilidades (Sistema de arquivos global, Fila de utilização dos recursos)

Requisitos almejados da infraestrutura

Para os usuários

- Serem independentes para realização de experimentos
- Fácil utilização (assumindo claro um conhecimento de Linux)
- Acesso isolado, irrestrito e com máximo desempenho aos recursos (*bare metal*)
- Algumas facilidades (Sistema de arquivos global, Fila de utilização dos recursos)

Para os professores

- Controlar o grupo de usuários que usa um subconjunto de máquinas
- Não se preocupar com o gerenciamento da máquina do seu projeto

Requisitos almejados da infraestrutura

Para os usuários

- Serem independentes para realização de experimentos
- Fácil utilização (assumindo claro um conhecimento de Linux)
- Acesso isolado, irrestrito e com máximo desempenho aos recursos (*bare metal*)
- Algumas facilidades (Sistema de arquivos global, Fila de utilização dos recursos)

Para os professores

- Controlar o grupo de usuários que usa um subconjunto de máquinas
- Não se preocupar com o gerenciamento da máquina do seu projeto

Para os administradores

- Mutualização e melhor aproveitamento dos recursos computacionais
- Instalação relativamente fácil $\rightarrow$ Pouca manutenção (idealmente: *install and forget*)
- Registro de todas as ações do usuário

Filosofia geral da plataforma

- Garantia inspirada na GPLv3, sistema fornecido “as-is”
  - Sem backups, sem garantia que vá funcionar, tudo pode ser desligado sem aviso prévio
  - Trabalho voluntário de discentes da pós-graduação (PPGC)

Filosofia geral da plataforma

- Garantia inspirada na GPLv3, sistema fornecido “as-is”
  - Sem backups, sem garantia que vá funcionar, tudo pode ser desligado sem aviso prévio
  - Trabalho voluntário de discentes da pós-graduação (PPGC)
- Abordagem centralizada (um único controlador, o *front-end*)

  ```
  ssh gppd-hpc.inf.ufrgs.br
  ```
- Um único sistema operacional, sem virtualização, sem *deploy*

Filosofia geral da plataforma

- Garantia inspirada na GPLv3, sistema fornecido “as-is”
  - Sem backups, sem garantia que vá funcionar, tudo pode ser desligado sem aviso prévio
  - Trabalho voluntário de discentes da pós-graduação (PPGC)
- Abordagem centralizada (um único controlador, o *front-end*)

  `ssh gppd-hpc.inf.ufrgs.br`
- Um único sistema operacional, sem virtualização, sem *deploy*
- Usuários responsáveis pela maioria das bibliotecas
  - Emprego de `docker`, `spack`, `guix` em função da experiência do usuário
  - Evita-se enormemente "instalar pacotes para os usuários"

Filosofia geral da plataforma

- Garantia inspirada na GPLv3, sistema fornecido “as-is”
  - Sem backups, sem garantia que vá funcionar, tudo pode ser desligado sem aviso prévio
  - Trabalho voluntário de discentes da pós-graduação (PPGC)
- Abordagem centralizada (um único controlador, o *front-end*)

  `ssh gppd-hpc.inf.ufrgs.br`
- Um único sistema operacional, sem virtualização, sem *deploy*
- Usuários responsáveis pela maioria das bibliotecas
  - Emprego de `docker`, `spack`, `guix` em função da experiência do usuário
  - Evita-se enormemente "instalar pacotes para os usuários"
- Requerimentos especiais para controle experimental
  - Controlar frequência de CPU e GPU
  - Desativar/ativar cores, turboboost, hyperthreading, . . .
  - Configurações específicas em BIOS
  - Uso de discos locais (*scratch*) para experimentos com dados

Sistema base e gerenciamento de usuários e arquivos

Sistema base → Debian GNU/Linux

- *Free Software*, existe desde ≈1993
- Versão *stable* com LTS, atualmente Debian 12

Sistema base e gerenciamento de usuários e arquivos

Sistema base → Debian GNU/Linux

- *Free Software*, existe desde ≈1993
- Versão *stable* com LTS, atualmente Debian 12

Gerenciamento de usuários → LDAP (com OpenLDAP)

- *Free Software*, mantido pela Fundação OpenLDAP (desde ≈1998)
  - Unificação de todos os usuários e dados do perfil (UID, GID)
- Acesso utilizando chaves ssh

Sistema base e gerenciamento de usuários e arquivos

Sistema base → Debian GNU/Linux

- *Free Software*, existe desde ≈1993
- Versão *stable* com LTS, atualmente Debian 12

Gerenciamento de usuários → LDAP (com OpenLDAP)

- *Free Software*, mantido pela Fundação OpenLDAP (desde ≈1998)
  - Unificação de todos os usuários e dados do perfil (UID, GID)
- Acesso utilizando chaves ssh

Sistema de Arquivos de rede → NFS

NFS

- Sistema presente direto no kernel do Linux
  - Transparente, não requer configurações extras pelo usuário
- Compartilha o diretório `$HOME` entre todas as máquinas

Slurm

- Escalonar alocações e tarefas dos usuários nas máquinas
- Amplamente utilizado em diversos supercomputadores
  - Eficiente (realização do escalonamento e qualidade do mesmo)
- Isolar os ambientes (nós ou recursos intra-nó)
- Instalação relativamente fácil

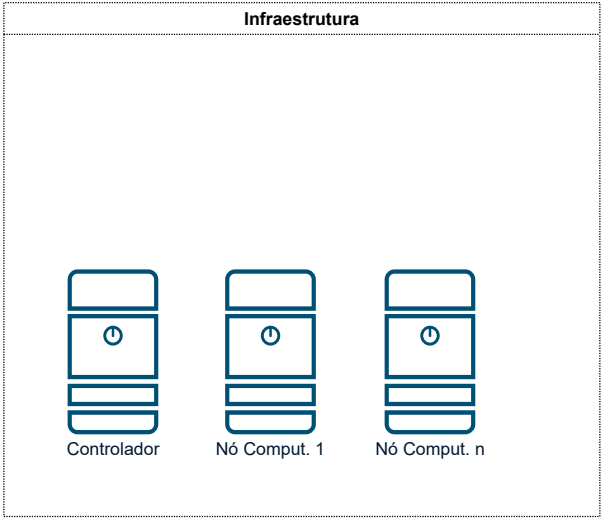
Slurm

- Escalonar alocações e tarefas dos usuários nas máquinas
- Amplamente utilizado em diversos supercomputadores
  - Eficiente (realização do escalonamento e qualidade do mesmo)
- Isolar os ambientes (nós ou recursos intra-nó)
- Instalação relativamente fácil

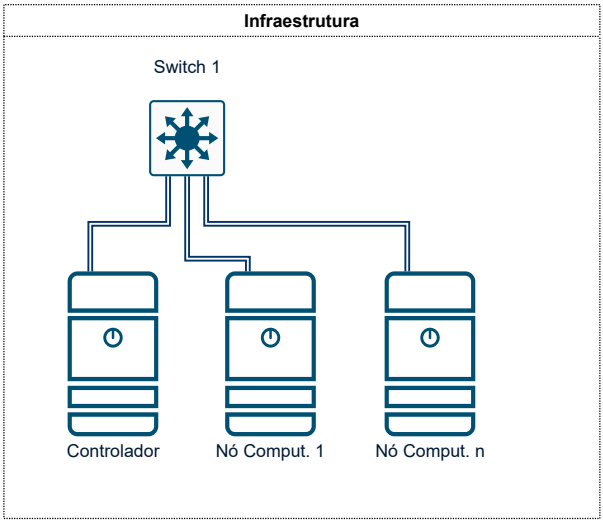
Slurm é peça fundamental para gerenciar a heterogeneidade

- Emprega-se o conceito de partição
  - Nós que pertencem a uma partição são homogêneos
- Temos ≈21 partições, muitas com somente 1 máquina

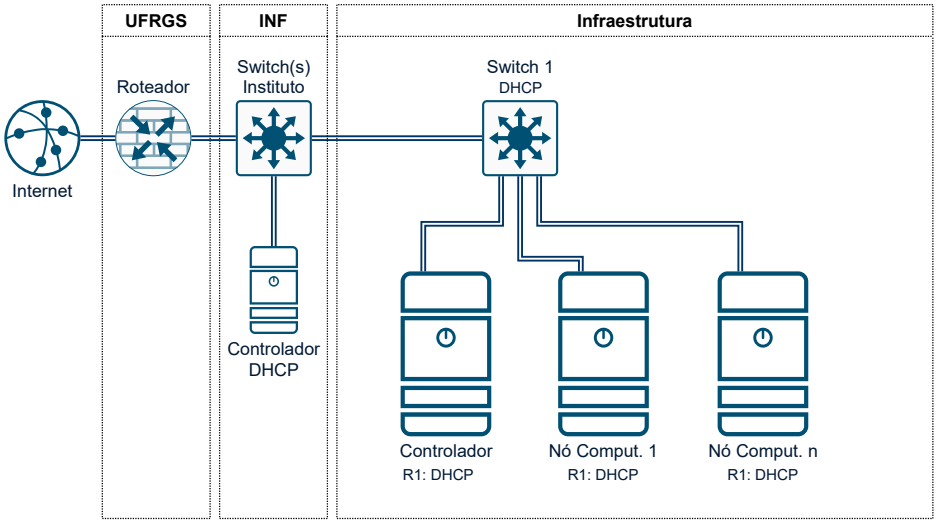
Infraestrutura da interconexão física



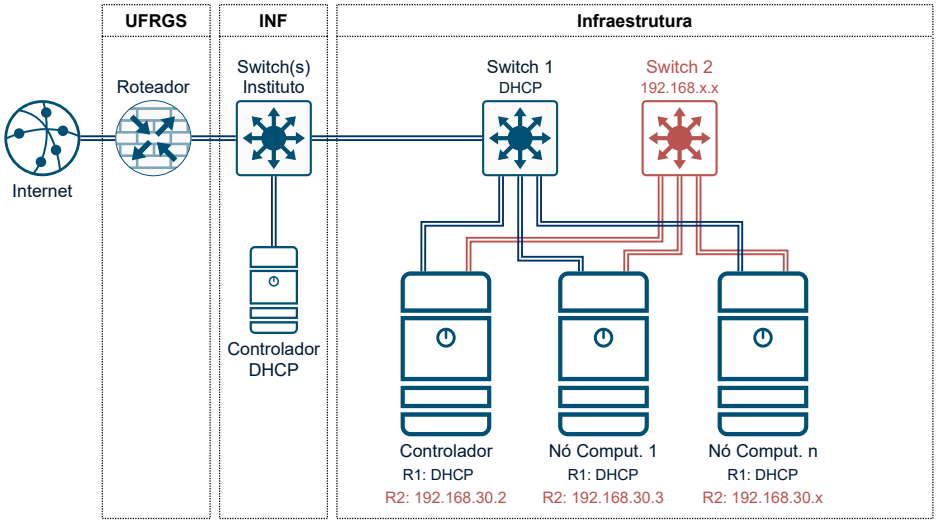
Infraestrutura da interconexão física



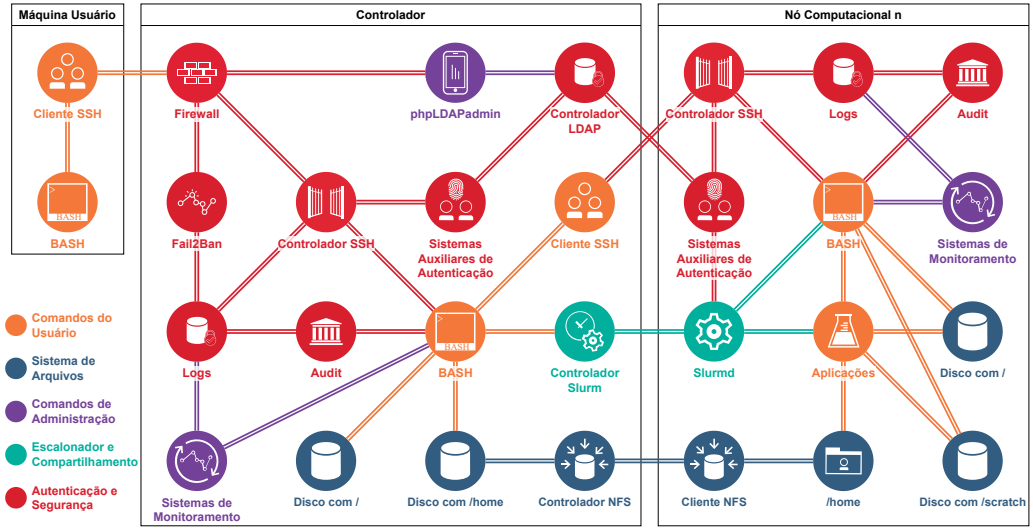
Infraestrutura da interconexão física



Infraestrutura da interconexão física

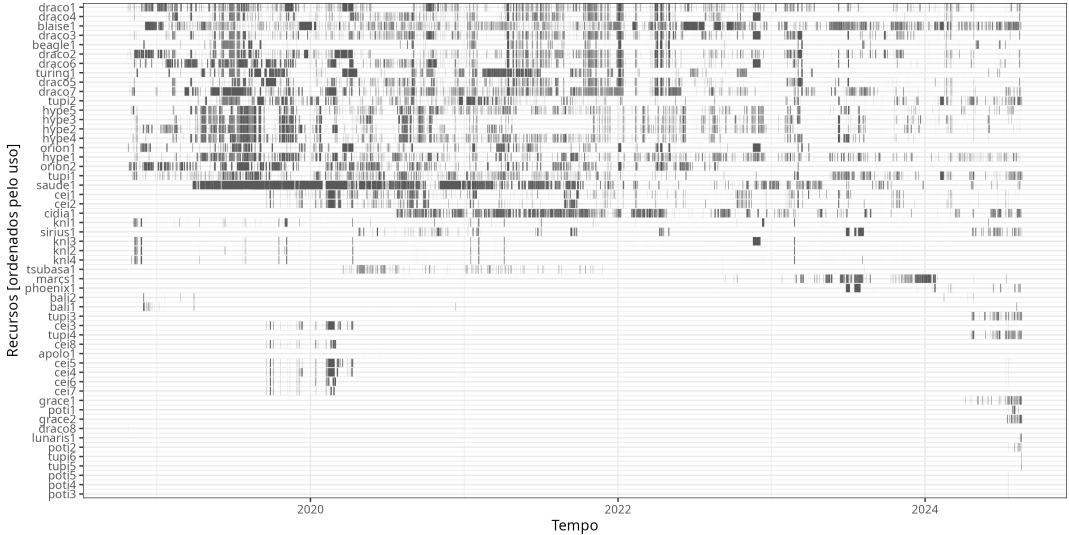


Organização do software (visão geral)



Visão geral das alocações / uso dos recursos

Uso dos recursos (limite de 24hs por job) → Muitos adotam `salloc` ao invés de `sbatch`



Uso para experimentos / protótipo

Maior parte dos *jobs* são de experimentos curtos

- É uma plataforma experimental, tentativas e erros são frequentes
- Serve de suporte para ensino em nível de graduação e pós-graduação
  - Já serviu para suporte em minicurso da ERAD/RS
- Frequente uso de um nó com GPU para inferência com LLMs
  - Embora cada vez mais jobs usam para treinar redes profundas

Maioria dos *jobs* são single-node

Uso para experimentos / protótipo

Maior parte dos *jobs* são de experimentos curtos

- É uma plataforma experimental, tentativas e erros são frequentes
- Serve de suporte para ensino em nível de graduação e pós-graduação
  - Já serviu para suporte em minicurso da ERAD/RS
- Frequente uso de um nó com GPU para inferência com LLMs
  - Embora cada vez mais jobs usam para treinar redes profundas

Maioria dos *jobs* são single-node

Suporta vários projetos de pesquisa

- Fomento FAPERGS, CNPq, Petrobras, FINEP, HPE
- Ministério da Saúde, SDECT/RS

Pessoal

Professores com fomento

- Philippe O. Alexandre Navaux
- Carla Freitas
- Luciana Nedel
- João Luiz Dihl Comba
- Viviane P. Moreira
- Mariana Recamonde Mendoza
- Lucas Mello Schnorr

Professores usuários

Anderson Tavares, Antônio Beck Filho, Arthur Lorenzon, Cláudio Geyer, Cláudio Jung, Dennis Balreira, Eduardo Gastal, Joel Carbonera, Karin Becker, Luciano Gaspary, Paolo Rech, Thiago da Silveira, Luigi Carro

Pessoal

Professores com fomento

- Philippe O. Alexandre Navaux
- Carla Freitas
- Luciana Nedel
- João Luiz Dihl Comba
- Viviane P. Moreira
- Mariana Recamonde Mendoza
- Lucas Mello Schnorr

Professores usuários

Anderson Tavares, Antônio Beck Filho, Arthur Lorenzon, Cláudio Geyer, Cláudio Jung, Dennis Balreira, Eduardo Gastal, Joel Carbonera, Karin Becker, Luciano Gaspary, Paolo Rech, Thiago da Silveira, Luigi Carro

Usuários de outras instituições

UFSM, FURG, UNIPAMPA

INPE, UFPA, SERPRO, UNIOESTE

Pessoal

Professores com fomento

- Philippe O. Alexandre Navaux
- Carla Freitas
- Luciana Nedel
- João Luiz Dihl Comba
- Viviane P. Moreira
- Mariana Recamonde Mendoza
- Lucas Mello Schnorr

Professores usuários

Anderson Tavares, Antônio Beck Filho, Arthur Lorenzon, Cláudio Geyer, Cláudio Jung, Dennis Balreira, Eduardo Gastal, Joel Carbonera, Karin Becker, Luciano Gaspary, Paolo Rech, Thiago da Silveira, Luigi Carro

Usuários de outras instituições

UFSM, FURG, UNIPAMPA

INPE, UFPA, SERPRO, UNIOESTE

Discentes envolvidos (administradores)

Atuantes fundamentais

- Lucas Nesi
- Cristiano Kunas

Ex-atuantes fundamentais

- Matheus Serpa

Tendências

Recentemente

- Enfoque em aceleradores muito forte
  - Vários nós são “substituídos” por poucas GPUs
- Aceleradoras tipo server extremamente caras
  - Resurgimento de gabinetes “full-tower” com 1x, 2x até 4x aceleradoras
- Cada vez mais heterogeneidade nas configurações
  - Nós “AMD” ou “ARM” com processador/acelerador

Tendências

Recentemente

- Enfoque em aceleradores muito forte
  - Vários nós são “substituídos” por poucas GPUs
- Aceleradoras tipo server extremamente caras
  - Resurgimento de gabinetes “full-tower” com 1x, 2x até 4x aceleradoras
- Cada vez mais heterogeneidade nas configurações
  - Nós “AMD” ou “ARM” com processador/acelerador

Futuro

- Necessidade de *clusters* maiores de aceleradoras
  - Atualmente temos a partição “tupi” (6 nós, cada uma com uma RTX4090)
  - Usuários usam apenas 1 nó c/ GPU, quando poderiam distribuir em vários nós com GPU
  - Atualização de máquinas com GPUs profissionais nos nós que as suportam
- Perspectiva da chegada de nós DGX H100

Futuro mais distante $\rightarrow$ Deploy bare-metal, PDU gerenciável

Referências

DevOps para HPC: Como configurar um cluster para uso compartilhado

- https://lnesi.gitlab.io/mc-hpc-share/
- https://cradrs.github.io/eradrs2023/pdfs/minicursos/cap-3.pdf

Boas práticas para experimentos HPC

- https://exp-hpc.gitlab.io/
- Série de minicursos desde 2019 na ERAD/RS, ERAD/SP, WSCAD

Grupo de Processamento Paralelo e Distribuído (GPPD/HPC)

- HPC, novas arquiteturas, paradigmas de prog. paralela, análise de desempenho
- https://www.inf.ufrgs.br/gppd/site/

Contato

Obrigado pela atenção!

schnorr@inf.ufrgs.br

PCAD

http://gppd-hpc.inf.ufrgs.br/