Brasil 6G

Projeto Brasil 6G
Fase III

Atividade 4.5 - Implantação e disponibilização de infraestrutura de HPC - Parte 1

18/07/2025

Histórico de Atualizações:

Versão	Data	Autor(es)	Notas
1	24/06/2025	Bruno Ciro do Nascimento (RNP)	Elaboração de conteúdo
2	24/06/2025	Matheus Henrique Braga Julidori (Inatel)	Adequação de template e formatação
3	30/06/2024	Luiz Gustavo Barros Guedes (Inatel) Juliane de Souza Mata (RNP) Juliano Silveira Ferreira (Inatel)	Revisão de texto
4	04/07/2025	Bruno Ciro do Nascimento (RNP)	Ajustes e elaboração de conteúdo Complementar
5	18/07/2025	Luciano Leonel Mendes (Inatel)	Revisão final

Lista de Figuras

Acrônimos

5G Quinta Geração de Rede Móvel Celular
6G Sexta Geração de Rede Móvel Celular
AI *Artificial Inteligence*
APU *Accelerated Processing Unit*
BMC *Baseboard Management Controller*
CIFS *Common Internet File System*
CNDs Centros Nacionais de Dados
CPUs *Central Processing Units*
CUDA *Compute Unified Device Architecture*
CVEs *Common Vulnerabilities and Exposures*
CFM *Cubic Feet per Minute*
FCP *Fibre Channel Protocol*
GPU *Graphics Processing Unit*
HBM3e *High Bandwidth Memory 3 Extended*
HPC *High-Performance Computing*
IA Inteligência Artificial
iDRAC *Integrated Dell Remote Access Controller*
IPMI *Intelligent Platform Management Interface*
iSCSI *Internet Small Computer System Interface*
MIMO *Multiple Input Multiple Output*
ML *Machine Learning*
NFS *Network File System*
NVMe *Non-Volatile Memory Express*
NVMe-oF *NVMe over Fabrics*
OSFP *Octal Small Form-factor Pluggable*
OEM *Original Equipment Manufacturer*
PHY *Physical Layer*
PCIe *Peripheral Component Interconnect Express*
PXE *Preboot Execution Environment*
QoS *Quality of Service*
RAM *Random Access Memory*
RAID *Redundant Array of Independent Disks*
RAN *Radio Access Network*
RDMA *Remote Direct Memory Access*

RoCE *RDMA over Converged Ethernet*

RNP Rede Nacional de Ensino e Pesquisa

SDK *Software Development Kit*

S3 *Simple Storage Service*

SSH *Secure Shell*

SXM *Server PCI Express Module*

ToR *Top-of-Rack*

TR Termo de Referência

UPS *Uninterruptible Power Supply*

VLAN *Virtual Local Access Network*

Sumário

1 Introdução

O projeto Brasil 6G destaca-se pela sua importância estratégica ao fomentar a pesquisa, o desenvolvimento e os testes de novas tecnologias e soluções relacionadas ao futuro das redes de comunicação. O projeto é desenvolvido com a participação de diversos pesquisadores, provenientes de diferentes instituições, e conta com uma infraestrutura avançada pré-existente. Além disso, contempla a modernização e a ampliação dessa infraestrutura, de modo a atender aos novos requisitos técnicos e operacionais. Neste contexto, destacam-se a especificação, a implantação e a disponibilização da infraestrutura de um *High-Performance Computing* (HPC), que constituem os focos da atividade 4.5, objeto do presente relatório.

O propósito primordial do HPC é dar suporte à concepção e à avaliação de desempenho das tecnologias emergentes para a Sexta Geração de Rede Móvel Celular (6G), permitindo a realização de simulações complexas, o treinamento de modelos de Inteligência Artificial (IA) avançados e o processamento de um volume massivo de dados. Sendo assim, a plataforma do HPC concebida pelo projeto disponibilizará GPUs e será equipada com ferramentas avançadas de IA.

A infraestrutura de HPC, delineada nesta atividade, é projetada para ser robusta, flexível e preparada para o futuro. Uma de suas características fundamentais é a capacidade de permitir o uso remoto e simultâneo de seus recursos por pesquisadores de diferentes instituições envolvidas no projeto. Essa funcionalidade é crucial para fomentar a colaboração interinstitucional e maximizar a utilização dos recursos computacionais disponíveis, independentemente da localização geográfica dos usuários.

O presente relatório técnico detalha os progressos e as etapas da atividade 4.5, abordando a identificação das aplicações e finalidades de uso do HPC, os requisitos considerados para sua especificação técnica e arquitetura, as principais soluções tecnológicas identificadas, os requisitos e locais candidatos à instalação e a elaboração do termo de referência para aquisição da infraestrutura. Ao consolidar esses aspectos, este documento visa demonstrar o comprometimento com a excelência técnica e a otimização dos investimentos, garantindo que a plataforma de HPC sirva como um pilar fundamental para as inovações que moldarão as redes 6G no Brasil.

Visando alcançar seus objetivos, o presente relatório está estruturado da seguinte forma: a Seção 2 aborda a identificação das aplicações e finalidades de uso do HPC; a Seção 3 detalha os requisitos e a especificação técnica da solução; a Seção 4 apresenta as soluções de mercado analisadas e a justificativa para a escolha da plataforma a ser adotada; a Seção 5, por sua vez, explora a camada de *software* e a orquestração da solução; a Seção 6 discute os requisitos de instalação e os locais candidatos para a instalação; a Seção 7 descreve o processo de elaboração do termo de referência necessário para aquisição dos equipamentos que compõem o HPC; a Seção 8 indica os próximos passos do projeto; e, por fim, a Seção 9 apresenta as considerações finais juntamente com a conclusão do documento.

2 Aplicações e Finalidades de Uso

A infraestrutura de HPC é um pilar fundamental para a pesquisa e o desenvolvimento de novas tecnologias, sendo sua aplicação, no âmbito do projeto Brasil 6G, estratégica. Ela será a infraestrutura base que permitirá a realização de simulações complexas, o treinamento de modelos de IA avançados e o processamento de um volume massivo de dados. Tais atividades são fundamentais para a concepção da sexta geração de redes móveis, assim como para o desenvolvimento e a validação de algoritmos e aplicações correlatas.

Nesse contexto, os pesquisadores do projeto Brasil 6G pretendem utilizar intensivamente os recursos de GPUs disponíveis na infraestrutura de HPC. A escolha se justifica pela arquitetura paralela das GPUs, que as tornam ideais para as cargas de trabalho de IA e simulações de algoritmos da camada física de sistemas de radiocomunicação, dentre outras. Para tal, o ecossistema de *software* da empresa NVIDIA foi adotado e será amplamente empregado visando proporcionar maior eficiência e acelerar o ciclo de inovação, conforme melhor detalhado e justificado ao longo deste documento.

As aplicações previstas para o HPC no projeto Brasil 6G contemplam:

- A aplicação de IA e *Machine Learning* (ML) nas camadas física e de rede, com ênfase no uso de GPUs para o treinamento de redes neurais profundas capazes de otimizar o desempenho da rede em tempo real. Isso envolve desde o gerenciamento dinâmico e inteligente do espectro eletromagnético até a detecção autônoma de anomalias e orquestração de recursos de rede de forma muito mais eficiente que os métodos atuais;

- O processamento de sinais em larga escala, uma vez que a tecnologia 6G exigirá sistemas de antenas *massive Multiple Input Multiple Output* (MIMO) e algoritmos de *beamforming* extremamente complexos para direcionar o sinal com alta precisão aos usuários. As GPUs são essenciais para realizar os cálculos matriciais massivos necessários para o processamento desses sinais em tempo real;

- A simulação de ambientes complexos e gêmeos digitais (*digital twins*). Antes de implantar uma única antena, será possível criar *digital twins* de ambientes urbanos e rurais. Com a aceleração proporcionada pelas GPUs, os pesquisadores poderão simular, com alta fidelidade, a propagação e a interação das ondas de rádio em frequências sub-THz com edifícios, pessoas e o ambiente, possibilitando o teste e a validação de topologias de rede de forma significativamente mais rápida e econômica.

3 Requisitos e Especificação Técnica da Solução

A solução computacional para o projeto Brasil 6G foi projetada para ser uma plataforma robusta, flexível e com capacidade de evolução para acompanhar o avanço das pesquisas. Seus requisitos foram definidos para garantir que a infraestrutura atenda tanto às necessidades atuais quanto às futuras, abrangendo desde a execução de algoritmos complexos até a gestão eficiente dos recursos.

A funcionalidade principal da plataforma é a execução de modelos avançados de Inteligência Artificial, incluindo arquiteturas generativas que podem ser aplicadas na otimização de topologias de rede. Essa capacidade é fundamental para criar cenários de simulação realistas e gerar dados sintéticos para o treinamento de algoritmos. A solução deve incluir ferramentas dedicadas à simulação da camada física ou *Physical Layer* (PHY), permitindo aos pesquisadores modelar com alta fidelidade fenômenos como *beamforming*, MIMO massivo e a propagação de sinais em frequências sub-THz, para validar novas tecnologias de Quinta Geração de Rede Móvel Celular (5G), 6G e além.

A arquitetura de redes 6G será inerentemente distribuída. Portanto, a solução deve permitir a execução de cargas de trabalho de forma transparente e eficiente, desde os ambientes de nuvem centralizados até a borda computacional mais próxima dos usuários e dispositivos. A plataforma deve ser capaz de interoperar em um ambiente federado e distribuído, portanto ela poderá se conectar e colaborar com futuros novos nós, permitindo a execução de projetos conjuntos e o compartilhamento de dados e modelos em larga escala.

Para a gestão de recursos e o uso eficiente do hardware, a solução deverá permitir o gerenciamento dinâmico de GPUs virtuais por meio de tecnologias de fatiamento de recursos. Isso assegura que múltiplos pesquisadores possam compartilhar o hardware de forma segura, com qualidade de serviço ou *Quality of Service* (QoS) garantida para cada tarefa. A plataforma poderá executar, por exemplo, um treinamento de longa duração em uma fatia da GPU enquanto outra fatia é utilizada para tarefas de inferência. Complementarmente, todos os dados gerados, sejam resultados de simulações, modelos treinados ou *datasets*, deverão ser armazenados de forma centralizada e segura no ambiente de armazenamento de dados da solução, garantindo acesso, integridade e persistência.

3.1 Desenho da Arquitetura

No processo de concepção do HPC, adotou-se como arquitetura de referência a solução computacional denominada NVIDIA BasePOD [1], representada na Figura 1, tratando-se de uma solução bastante completa e composta pelos seguintes elementos:

- Camada de processamento, que é formada por um conjunto de servidores NVIDIA DGX, cujos modelos DGX H200 ou B200 são, atualmente, os mais comuns. A arquitetura define configurações para um número específico de sistemas, geralmente começando com 2 e escalando para 4, 8 ou mais nós DGX;

- Camada de rede, que é tão crítica quanto a de processamento, uma vez que sua arquitetura é projetada para eliminar gargalos. A arquitetura BasePOD especifica o uso de tecnologias de rede de altíssimo desempenho da NVIDIA para duas funções distintas: a *Compute Fabric* e a *Storage and Management Fabric*. A primeira consiste em uma rede de altíssima velocidade, geralmente InfiniBand 400Gb/s, que conecta todos os sistemas

DGX. Sua única finalidade é permitir que as GPUs de diferentes servidores se comuniquem diretamente com a GPUDirect *Remote Direct Memory Access* (RDMA), como se estivessem em um único e massivo servidor, o que é crucial para treinar modelos de IA de larga escala em paralelo. A segunda consiste em uma outra rede de alta velocidade, geralmente Ethernet, que conecta os sistemas DGX ao armazenamento e à rede de gerenciamento do *datacenter*;

- Camada de armazenamento. O BasePOD não especifica uma única marca de armazenamento, mas sim uma arquitetura que se integra com soluções de armazenamento de parceiros certificados pela NVIDIA, como NetApp, Pure Storage, DDN, WEKA, entre outros. Esses sistemas de armazenamento são validados para fornecer a taxa de transferência e a baixa latência necessárias para "alimentar" as GPUs com dados sem deixá-las ociosas;

- Camada de *software*, uma vez que a arquitetura é totalmente construída sobre a pilha de *software* da NVIDIA. Isso inclui: NVIDIA AI Enterprise e NVIDIA Base Command. A primeira corresponde à suíte de *software* que otimiza os *frameworks* de IA e fornece segurança e suporte de nível empresarial. A segunda consiste em um conjunto de ferramentas para gerenciamento de *cluster*, provisionamento de sistemas operacionais, agendamento de tarefas, suportando orquestradores como Slurm ou Kubernetes, e monitoramento da saúde da infraestrutura.

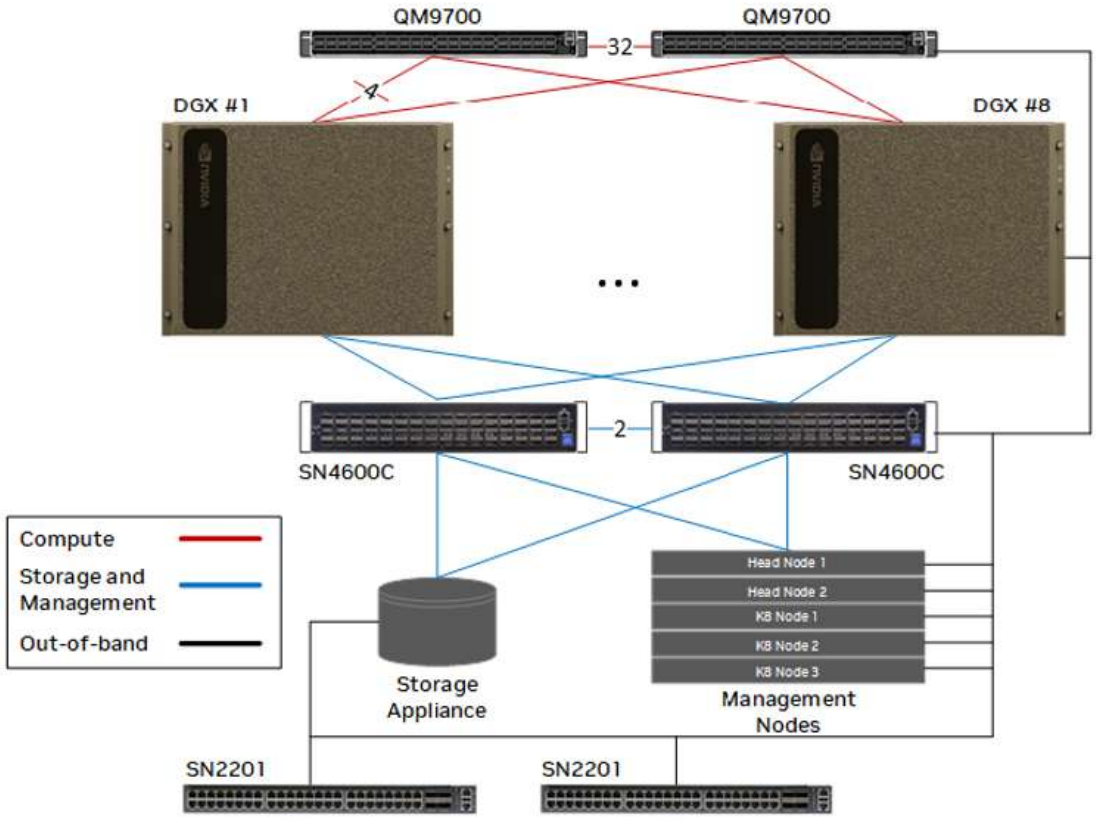

Figura 1: Diagrama esquemático da solução computacional NVIDIA BasePOD.

3.2 Características Técnicas da Solução Computacional

Assim como a solução de referência, a solução computacional que será adotada é composta de quatro grandes partes:

- Nós de processamento em GPUs (HPC/GPUs);
- Nó de armazenamento de dados (HPC/GPUs);
- Nós de controle e Kubernetes;
- Equipamentos de interconexão de redes.

Para materializar as capacidades descritas na Seção 3, a solução computacional, representada pela Figura 2, deve atender a um conjunto rigoroso de especificações, garantindo desempenho, flexibilidade e gerenciamento avançado. A seguir, algumas delas indicam que:

- A tecnologia de *hardware* deve permitir o particionamento de cada uma das GPUs físicas em múltiplas instâncias menores (n-instâncias). Cada uma dessas instâncias opera de forma totalmente isolada, com seus próprios recursos dedicados de memória de alta largura de banda, *cache* L2 e núcleos de computação, garantindo uma QoS previsível para cada tarefa [2];
- A solução deve oferecer a possibilidade de gerenciamento dinâmico das GPUs, permitindo que os administradores reconfigurem as instâncias em tempo real. Essa flexibilidade é vital para alocar recursos de forma justa e eficiente entre diferentes usuários e projetos, adaptando o ambiente computacional às demandas variáveis da pesquisa sem a necessidade de paradas ou reinicializações;
- É mandatório que a arquitetura permita comunicação direta e de alta velocidade entre as GPUs, tanto no interior de um mesmo servidor quanto entre servidores distintos por meio do GPUDirect RDMA. Essa capacidade é essencial para escalar grandes modelos de IA e executar simulações complexas que exijam o poder combinado de múltiplos processadores gráficos;
- A solução deve ter compatibilidade nativa e total com a plataforma de computação paralela NVIDIA *Compute Unified Device Architecture* (CUDA). Isso garante acesso imediato a um ecossistema maduro de *softwares*, bibliotecas e ferramentas de desenvolvimento que são padrão da indústria para computação acelerada por GPUs;
- O servidor deve ser oficialmente homologado pelo fabricante para operar com, no mínimo, oito placas GPUs do mesmo modelo. Essa homologação deve ser comprovada por meio de documentação técnica oficial. Além disso, o projeto exige suporte técnico direto do fabricante das GPUs, sem intermediários, para garantir a resolução ágil de problemas e acesso a conhecimento especializado;
- A infraestrutura deve permitir o gerenciamento remoto completo do *hardware* sem a necessidade de instalar agentes de *software* no sistema operacional. Essa interface de gerenciamento, como um *Baseboard Management Controller* (BMC)/*Integrated Dell Remote Access Controller* (iDRAC), deve prover funcionalidades essenciais como atualização de *firmware* e *drivers*, monitoramento da saúde do sistema, e gerenciamento de configuração e conformidade;

- Cada uma das GPUs da solução deve incluir uma licença perpétua da plataforma de *software* NVIDIA AI Enterprise. Conforme já abordado, a NVIDIA AI Enterprise é uma suíte de *software* de nível empresarial que otimiza o desenvolvimento e a implantação de IA, oferecendo acesso a *frameworks* otimizados, ferramentas de gerenciamento e suporte de longo prazo, o que garante estabilidade e segurança para os projetos de pesquisa;
- A solução deve incluir funcionalidade nativa para a orquestração das GPUs e de suas instâncias, utilizando contêineres gerenciados por Kubernetes. O gerenciamento dessa estrutura deve ser acessível via interface *web*, simplificando a alocação de recursos e o *deployment* de aplicações.

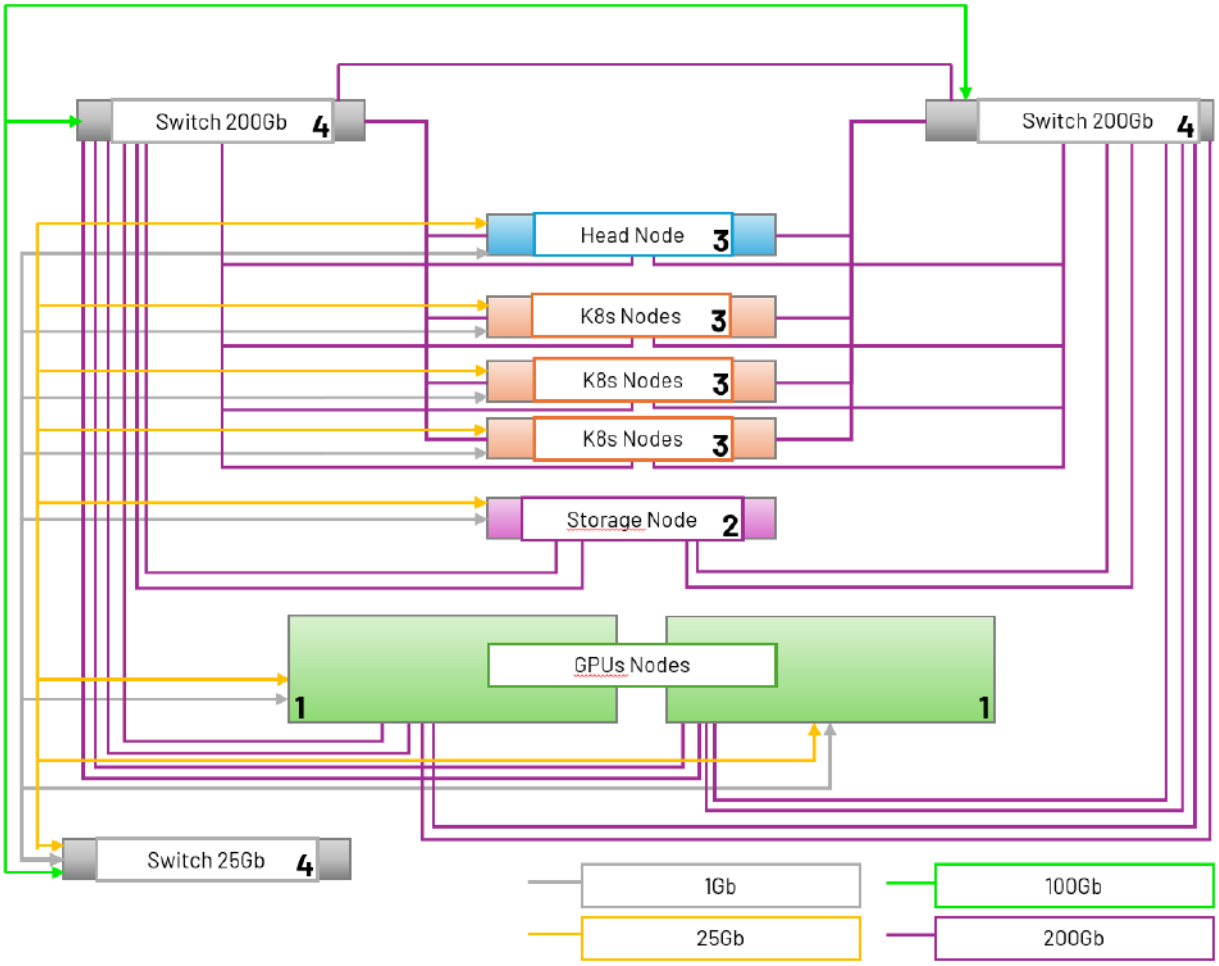

Figura 2: Diagrama esquemático da solução computacional do projeto Brasil 6G.

3.3 Características Técnicas Detalhadas dos Nós de Processamento em GPU

As especificações detalhadas para os nós de processamento em GPUs, destacados na Figura 2 com a identificação 1, são as seguintes:

- Oito GPUs, cada uma com, no mínimo, 141 GB de memória dedicada e largura de banda de, pelo menos, 4,8 TB/s, conectadas à placa-mãe via soquete *Server PCI Express Module* (SXM);

- Cada uma das GPUs deverá apresentar desempenho mínimo de 34 TFLOPS em operações de ponto flutuante de precisão dupla (FP64);
- Cada uma das GPUs deverá atingir, no mínimo, 67 TFLOPS em operações de ponto flutuante de precisão simples (FP32);
- Cada uma das GPUs deverá apresentar desempenho mínimo de 494 TFLOPS em operações *Tensor Core* no formato TF32, considerando desempenho sustentado, sem variações significativas;
- Duas *Central Processing Units* (CPUs) Intel, cada uma com 56 núcleos e 112 *threads*, totalizando 112 núcleos físicos e 224 *threads* lógicos;
- Mínimo de 2 TB de memória *Random Access Memory* (RAM) instalada;
- Uma primeira unidade lógica de armazenamento composta por, no mínimo, duas unidades físicas de 1,92 TB, conectadas via interface M.2 e utilizando o barramento *Non-Volatile Memory Express* (NVMe) (*Peripheral Component Interconnect Express* (PCIe)). Uma segunda unidade lógica de armazenamento composta por, no mínimo, oito unidades físicas de 3,84 TB, conectadas via interface U.2, também utilizando o barramento NVMe (PCIe);
- No mínimo, duas placas de rede, cada uma com duas portas QSFP112, compatíveis com velocidades de até 400 GbE e 400 Gbps para InfiniBand;
- Pelo menos, quatro portas físicas do tipo *Octal Small Form-factor Pluggable* (OSFP), cada uma suportando velocidades de até 400 GbE e 400 Gbps InfiniBand. Essas portas devem estar distribuídas em, no mínimo, duas placas de rede distintas (não sendo as mesmas mencionadas no item anterior). Cada uma dessas placas deverá estar diretamente interligada a uma das GPUs por meio de barramento PCIe ou tecnologia equivalente, garantindo largura de banda mínima de 200 Gbps por conexão;
- Pelo menos, uma placa de rede equipada com duas portas, cada uma suportando velocidades de até 100 GbE;
- Suporte a gerenciamento remoto via OpenBMC, *Intelligent Platform Management Interface* (IPMI) ou iDRAC, com fornecimento de licença perpétua para todos os recursos de gerenciamento;
- No mínimo duas portas de rede com velocidade mínima de 1 GbE cada;
- Seis fontes de alimentação redundantes, com potência mínima de 2800 W cada e certificação 80 Plus Titanium.

3.4 Características Técnicas Detalhadas do Nó de Armazenamento de Dados

As características técnicas detalhadas para o nó de armazenamento, destacado na Figura 2 com a identificação 2, são:

- Deverá suportar sistemas de arquivos paralelos de alto desempenho, tais como Lustre, BeeGFS e Intel DAOS, sem a exigência de aquisição de licenças adicionais;

- Deverá oferecer a funcionalidade de sistema de arquivos distribuído de alto desempenho, com suporte a capacidade de armazenamento de até 60 PB e até 400 bilhões de arquivos;
- Suporte obrigatório aos protocolos *Common Internet File System* (CIFS), *Network File System* (NFS) *over* RDMA, *Internet Small Computer System Interface* (iSCSI), *Fibre Channel Protocol* (FCP) e *Simple Storage Service* (S3);
- Deverá suportar, no mínimo, os seguintes níveis de proteção *Redundant Array of Independent Disks* (RAID) (ou tecnologias equivalentes) como o RAID5, o RAID6 e o RAID com tripla paridade;
- Deverá possuir controladoras com capacidade de operação em modo ativo-ativo, oferecendo redundância e balanceamento de carga;
- Deverá permitir expansão da capacidade de armazenamento por meio de escalabilidade vertical (adição de novas gavetas de discos), com capacidade de crescimento de, pelo menos, 1,4 PB;
- Deverá suportar a utilização de, no mínimo, 24 discos de dados (excluindo os discos dedicados ao sistema operacional);
- Deverá oferecer capacidade de armazenamento bruto mínima de 368 TB;
- Deverá incluir, no mínimo, quatro portas de rede com suporte a velocidades de até 200 GbE, utilizando conectores QSFP56;
- Deverá oferecer suporte a *failover* automático e possuir componentes *hot-swappable* (ou seja, substituíveis com o sistema em operação) para discos, controladoras, ventiladores e fontes de alimentação, de forma a permitir a substituição de componentes avariados sem interrupção dos serviços de armazenamento nem da aplicação que estiver acessando o ambiente.

3.5 Características Técnicas Detalhadas dos Nós de Controle e Kubernetes

As especificações detalhadas para os nós de controle e Kubernetes, destacados na Figura 2 com a identificação 3, são as seguintes:

- Uma unidade de servidor, identificada na Figura 2 como *Head Node*, deverá possuir, no mínimo, 256 GB de memória DDR5;
- As outras três unidades de servidor, identificadas na Figura 2 como K8s *Nodes*, deverão possuir, cada uma, no mínimo, 512 GB de memória DDR5;
- Cada servidor deverá estar equipado com duas CPUs, cada uma contendo, no mínimo, 32 núcleos físicos e 64 *threads*, totalizando 64 núcleos e 128 *threads* por servidor. Serão aceitos modelos equivalentes ou superiores a Intel Xeon Gold 6430 ou AMD EPYC 9334;
- Cada servidor deverá conter duas unidades lógicas de armazenamento, implementadas com, no mínimo, quatro unidades físicas de 960 GB (configuradas como RAID 1, sendo duas unidades físicas por volume lógico). As conexões deverão ser realizadas via interfaces M.2 ou U.2, utilizando barramento NVMe (PCIe);

- Cada servidor deverá dispor de, no mínimo, quatro portas de rede com suporte a velocidades de até 200 GbE, utilizando conectores QSFP56;
- Cada servidor deverá dispor de, no mínimo, quatro portas de rede com suporte a velocidades de até 25 GbE, utilizando conectores SFP28;
- Deverá ser fornecido sistema de gerenciamento remoto compatível com *OpenBMC*, IPMI ou iDRAC, incluindo licença perpétua para todas as funcionalidades de gerenciamento;
- Cada servidor deverá possuir, no mínimo, duas portas de rede com velocidade mínima de 1 GbE, com suporte a *Preboot Execution Environment* (PXE);
- Cada servidor deverá contar com duas fontes de alimentação redundantes, com potência mínima de 1400 W cada e certificação 80 Plus Titanium.

3.6 Características Técnicas Detalhadas dos Equipamentos de Interconexão de Redes

As características técnicas detalhadas para os equipamentos de interconexão, destacados na Figura 2 com a identificação 4, podem ser classificadas em duas grandes classes: 1) duas unidades de Switch Fabric 200 Gb Ethernet; e 2) uma unidade de Switch *Top-of-Rack* (ToR) 25 Gb Ethernet. Cada uma das unidades de 1) contém as seguintes características:

- No mínimo, 32 portas com velocidade de, pelo menos, 200 GbE, utilizando interfaces físicas do tipo QSFP56;
- Pelo menos, uma porta física RJ45 dedicada para acesso via console serial;
- Pelo menos, uma porta USB 3.0 Tipo A;
- Pelo menos, uma porta RJ45 dedicada para gerenciamento fora de banda (*out-of-band*), com velocidade mínima de 1 GbE;
- Capacidade mínima de 6,4 Tb/s de *switching* total;
- No mínimo, 8 GB de memória RAM interna;
- Processador baseado em arquitetura x86, com, no mínimo, 4 núcleos;
- Suporte a funcionalidades de roteamento nos níveis Layer2 e Layer3;
- Deverá suportar, no mínimo, modo de operação definido por porta (*Switching*, *Mirroring* ou *Aggregate*), *Network Virtual Local Access Network* (VLAN), *Stack* VLAN, *Jumbo Frame*, *Flow Control*, *Storm Control*, *Secure Shell* (SSH) v2, *Static Routing* e *Inter-*VLAN *Routing*;
- Suporte a *RDMA over Converged Ethernet* (RoCE) versão 1 ou 2;
- Suporte ao protocolo *NVMe over Fabrics* (NVMe-oF) sobre Ethernet;
- Deverá ser entregue com sistema operacional instalado e com licença perpétua, contemplando todas as funcionalidades descritas;

- Suporte a gerenciamento remoto, incluindo configuração, monitoramento e atualização de *firmware*;
- No mínimo, seis ventiladores com redundância $N+1$ (sendo N o número de ventiladores), todos substituíveis a quente (*hot-swappable*), ou seja, sem necessidade de desligamento do equipamento;
- Pelo menos, duas fontes de alimentação, configuradas em redundância 1+1, também substituíveis a quente (*hot-swappable*).

Já a unidade de 2) contém as seguintes características:

- No mínimo, 36 portas com velocidade de, pelo menos, 25 GbE, utilizando interfaces físicas do tipo SFP28;
- No mínimo, oito portas com velocidade de, pelo menos, 100 GbE, utilizando interfaces QSFP28;
- Pelo menos, uma porta física RJ45 dedicada para acesso via console serial;
- Pelo menos, uma porta USB 3.0 Tipo A;
- Pelo menos, uma porta RJ45 dedicada para gerenciamento fora de banda (*out-of-band*), com velocidade mínima de 1 GbE;
- Suporte a funcionalidades de roteamento nos níveis Layer2 e Layer3;
- Deverá suportar, no mínimo, modo de operação definido por porta (*Switching*, *Mirroring* ou *Aggregate*), *Network* VLAN, *Stack* VLAN, *Jumbo Frame*, *Flow Control*, *Storm Control*, SSHv2, *Static Routing* e *Inter*-VLAN *Routing*;
- Deverá ser entregue com sistema operacional instalado e com licença perpétua, contemplando todas as funcionalidades descritas;
- Deverá oferecer suporte a gerenciamento remoto, com capacidade de configuração, monitoramento e atualização;
- No mínimo, duas fontes de alimentação, com configuração redundante 1+1, também substituíveis a quente (*hot-swappable*).

4 Soluções de Mercado e Escolha da Plataforma

O mercado de GPUs para aceleração de tarefas computacionais e, especialmente, para processamento de IA, está passando por um período de transformação e crescimento exponencial. Esse movimento tem sido impulsionado por novas aplicações de larga escala em inteligência artificial generativa, sendo o ChatGPT [3], da empresa OpenAI, um dos principais catalisadores dessa demanda global.

Segundo dados de 2024 da consultoria IoT Analytics [4], a distribuição de participação de mercado no segmento de GPUs para *datacenters* é dada por: NVIDIA com 92%, AMD com 4% e Intel, juntamente com outras empresas, com 4%. Esses números evidenciam a liderança expressiva da NVIDIA, conquistada não apenas pelo desempenho de seu *hardware*, mas também pela maturidade e consolidação de seu ecossistema de *software*, o que tornou a plataforma, de fato, um padrão na indústria ao longo de mais de uma década.

Ainda que as participações da AMD e da Intel sejam consideravelmente menores, ambas continuam atuando como fornecedores estratégicos e como alternativas viáveis, oferecendo arquiteturas diferenciadas e modelos de negócios competitivos.

4.1 Análise das opções de mercado

Diante desse panorama, realizou-se uma análise comparativa das soluções disponíveis, contemplando não apenas as especificações de *hardware*, mas também a maturidade das ferramentas de desenvolvimento, suporte a *frameworks* de mercado, disponibilidade de bibliotecas otimizadas e amplitude da base de desenvolvedores ativos.

A NVIDIA lidera o mercado de GPUs para IA e HPC, com destaque para a plataforma CUDA, reconhecida como referência por sua robustez, desempenho e compatibilidade com os principais *frameworks* de IA. Suas linhas atuais consistem nas séries Hopper (H100/H200) e Blackwell (B100/B200). A H200 representa uma evolução da H100, oferecendo maior capacidade de memória (*High Bandwidth Memory 3 Extended* (HBM3e)) e maior largura de banda. Blackwell é a série que traz avanços significativos em treinamento e inferência de modelos de IA de larga escala. A série Blackwell Ultra (B300) é apresentada como sucessora da linha atual, oferecendo melhorias adicionais em desempenho, memória e largura de banda, com foco em *workloads* emergentes, como raciocínio de IA (AI *Reasoning*). A NVIDIA se destaca pela maturidade de *software* e amplitude de suporte com CUDA, cuDNN, TensorRT, RAPIDS, além de amplo suporte a *frameworks* como TensorFlow e PyTorch.

A AMD é a principal concorrente da NVIDIA no segmento de HPC e IA. Suas linhas atuais compreendem a série Instinct (MI300x/MI300a/MI325x). O MI300x destaca-se pela alta capacidade de memória (192 GB em cada uma das GPUs), o que permite executar grandes modelos de IA com menor número de GPUs. Já o MI300a é um *Accelerated Processing Unit* (APU) que combina CPUs e GPUs no mesmo pacote, atendendo, principalmente, aplicações de supercomputação. A próxima geração, representada pela série MI350, projetada para competir com a geração Blackwell, promete melhorias em desempenho computacional e eficiência energética. A AMD destaca-se pela alta capacidade de memória em cada uma das GPUs e bom desempenho em HPC. No entanto, o ecossistema de *software* (ROCm) ainda carece de maturidade quando comparado ao CUDA, com limitações de compatibilidade com *frameworks* de IA amplamente utilizados no mercado.

Finalmente, a Intel oferece soluções voltadas tanto para cargas de IA quanto para HPC e aplicações de mídia. Sua linha para IA corresponde à série Gaudi (atualmente, na versão Gaudi

3), voltada para treinamento e inferência de *deep learning*. Seu diferencial técnico é a integração de rede de alta velocidade no próprio *chip*. Suas linhas para HPC e nuvem consistem nas séries Data Center GPU Max e Flex, destinadas a *workloads* de HPC, transcodificação de vídeo e jogos em nuvem. Destaca-se pela abordagem baseada em *software* aberto e custo potencialmente mais acessível em determinadas cargas de inferência. Contudo, enfrenta limitações em maturidade de ecossistema e menor adoção pelas principais bibliotecas de IA de mercado.

4.2 Justificativa Técnica para a Escolha da Plataforma NVIDIA

Após a análise comparativa, a plataforma da NVIDIA foi escolhida como base para o projeto. A decisão foi fundamentada em critérios técnicos objetivos, com destaque para:

- Maturidade do ecossistema de *software* (CUDA, cuDNN, TensorRT, entre outros);
- Compatibilidade consolidada com os principais *frameworks* de IA (ex.: *TensorFlow*, *PyTorch*);
- Ampla disponibilidade de bibliotecas de alto desempenho para HPC e AI;
- Escalabilidade comprovada em arquiteturas multi-GPU com interconexão *NVLink*;
- Redução do risco tecnológico em função da ampla adoção de mercado e suporte de comunidade.

Um fator crítico para o projeto Brasil 6G é o suporte ao *Software Development Kit* (SDK) *Aerial* [5], exclusivo da NVIDIA, projetado especificamente para desenvolvimento de soluções de *Radio Access Network* (RAN) virtualizadas para 5G e 6G. O *Aerial* permite o processamento da camada física da rede (Layer 1) em GPUs, viabilizando a prototipação rápida de algoritmos e experimentos com redes definidas por *software*, em consonância direta com os objetivos do projeto.

Uma de suas ferramentas é o NVIDIA Aerial Omniverse Digital Twin que é uma plataforma de simulação avançada que permitirá aos pesquisadores do projeto Brasil não apenas 6G modelar e simular a RAN de ponta a ponta, mas também gerar dados sintéticos para treinamento de IA. A primeira consiste na possibilidade de construir um gêmeo digital de uma cidade inteira e, sobre ele, simular o desempenho de uma rede 6G completa. Isso inclui a modelagem precisa da propagação de sinais de radiofrequência, considerando reflexões, difrações e absorção por materiais de construção, permitindo a otimização precisa do posicionamento de estações base e antenas. Já a segunda trata de um dos maiores desafios em IA, que é a obtenção de dados de alta qualidade. Com o gêmeo digital, é possível gerar cenários de rede virtualmente infinitos para treinar os algoritmos de IA que gerenciarão a rede real, garantindo que eles sejam robustos e capazes de lidar com uma vasta gama de situações;

Outra ferramenta consiste no NVIDIA Sionna que, complementarmente, será o recurso de pesquisa na PHY. Trata-se de uma biblioteca de código aberto, construída sobre o TensorFlow, focada, especificamente, na simulação de sistemas de comunicação, permitindo não apenas prototipagem rápida de algoritmos, mas também otimização de ponta a ponta com IA. A primeira permite que os pesquisadores implementem e testem novos algoritmos de codificação, modulação e processamento de sinal com poucas linhas de código, aproveitando a aceleração de GPU para obter resultados rapidamente. A segunda apresenta como principal vantagem é a capacidade de realizar diferenciação automática em todo o sistema de comunicação. Isso possibilita

uma nova abordagem, na qual transmissor e receptor são otimizados conjuntamente como uma única rede neural, descobrindo formas de comunicação potencialmente mais eficientes do que aquelas projetadas por humanos, em um conceito conhecido como *learned communication*.

4.3 Opções de Aquisição de Hardware

Uma vez definida a escolha pela plataforma NVIDIA, a aquisição do *hardware* poderá seguir dois caminhos principais, ambos plenamente compatíveis com os requisitos técnicos estabelecidos. O primeiro corresponde ao NVIDIA DGX que é a solução de referência fornecida pela própria NVIDIA, com integração completa de *hardware* e *software*, incluindo oito GPUs (H200 ou B200), CPUs de alto desempenho, rede interna de alta velocidade, armazenamento NVMe, e pacote de *software* otimizado (NVIDIA AI Enterprise). Esta opção garante o máximo nível de integração e suporte técnico direto do fabricante.

Alternativamente, no segundo caminho, a solução poderá ser adquirida por meio de servidores desenvolvidos por *Original Equipment Manufacturer*s (OEMs), ou seja, parceiros autorizados (como Dell, Lenovo ou Supermicro), baseados na plataforma NVIDIA HGX. Nesse modelo, a placa base contendo as oito GPUs é fornecida pela própria NVIDIA, com os demais componentes (CPUs, armazenamento, fonte, placa-mãe, *chassis*) integrados pelos OEMs. Essa opção oferece maior flexibilidade de personalização sem desconsiderar a mesma plataforma de GPUs utilizada na solução de referência da NVIDIA.

5 Camada de *Software*

A disponibilidade de um ecossistema de *software* robusto é um requisito fundamental para simplificar o gerenciamento da solução como um todo e para suportar todo o ciclo de vida do desenvolvimento de aplicações de IA, desde a preparação de dados, passando pelo treinamento e inferência, até o gerenciamento de recursos computacionais.

Para esta solução, será adotada uma arquitetura baseada em contêineres, com orquestração por meio do Kubernetes, complementada pela suíte de *software* NVIDIA AI Enterprise.

5.1 Orquestração com Kubernetes e NVIDIA GPU Operator

A gestão eficiente de múltiplos usuários, projetos e cargas de trabalho simultâneas em um *cluster* de GPUs exige uma plataforma de orquestração capaz de gerenciar recursos de forma dinâmica e escalável. O Kubernetes foi selecionado como a plataforma padrão para essa função, oferecendo um *framework* unificado para automação da implantação, escalonamento e operação de aplicações em contêineres, tratando todo o *cluster* como um único *pool* de recursos.

Contudo, a gestão de recursos de GPU em ambientes Kubernetes demanda componentes adicionais. Nesse contexto, será utilizado o NVIDIA GPU Operator, responsável por automatizar o ciclo de vida de todos os componentes de *software* necessários para o funcionamento das GPUs NVIDIA no *cluster*. Esse operador garante a instalação, atualização e manutenção automatizada de *drivers*, *runtimes* de contêiner e bibliotecas específicas, permitindo que os administradores de tecnologia da informação gerenciem os recursos de GPU com o mesmo nível de abstração e automação usado para CPUs.

5.2 Camada de Software: NVIDIA AI Enterprise

Sobre essa base de orquestração, será implementada a suíte NVIDIA AI Enterprise. A aquisição da licença perpétua dessa suíte, vinculada a cada GPU da solução, foi estabelecida como requisito técnico obrigatório.

O NVIDIA AI Enterprise não é apenas um conjunto de ferramentas isoladas, mas, sim, uma plataforma de *software* integrada, com suporte técnico especializado e validação de estabilidade, que visa acelerar o desenvolvimento, a implantação e a gestão de aplicações de IA.

Os principais benefícios esperados da adoção do NVIDIA AI Enterprise incluem:

- Disponibilização de um catálogo abrangente de contêineres otimizados pela NVIDIA para os principais *frameworks* de IA e ciência de dados, como TensorFlow, PyTorch, NVIDIA RAPIDS e TensorRT. Isso assegura que os pesquisadores utilizem versões validadas e otimizadas, reduzindo a necessidade de customizações ou otimizações manuais;

- Estabilidade, segurança e suporte empresarial, uma vez que, de maneira distinta das versões puramente de código aberto, o *software* incluído na suíte passa por processos rigorosos de validação, testes de segurança e correções proativas de vulnerabilidades conhecidas ou *Common Vulnerabilities and Exposures* (CVEs). Além disso, o projeto contará com suporte técnico de nível empresarial diretamente da NVIDIA;

- Oferecimento de suporte a fluxos de trabalho consistentes entre diferentes ambientes, permitindo que aplicações desenvolvidas na infraestrutura local de HPC sejam portadas e implantadas de forma homogênea em ambientes de nuvem pública ou em dispositivos de borda.

5.3 Síntese da Arquitetura

A combinação entre o *hardware* de alto desempenho (DGX ou servidores baseados em NVIDIA HGX), a arquitetura de referência BasePOD, a orquestração via Kubernetes e a camada de *software* NVIDIA AI Enterprise resulta em uma solução integrada e alinhada com as melhores práticas do setor.

Essa integração permitirá que os pesquisadores do projeto Brasil 6G concentrem esforços nas atividades de pesquisa e desenvolvimento, com a confiança de que os aspectos de infraestrutura estão sendo gerenciados de forma eficiente, segura e com suporte técnico especializado.

6 Requisitos de Instalação e Locais Candidatos

A escolha do local de instalação é uma decisão estratégica que impacta diretamente o desempenho, a segurança operacional e a disponibilidade do sistema. O ambiente deverá atender a requisitos técnicos específicos de climatização, fornecimento de energia e conectividade de rede de altíssima velocidade, visando garantir a plena integração com o ecossistema nacional de pesquisa e desenvolvimento.

6.1 Integração com a RNP

Um requisito fundamental para o local de instalação é sua integração com a infraestrutura da RNP. Um dos aspectos se refere ao fato de que o local deverá possuir uma conexão de alta capacidade e baixa latência com a Rede Ipê, a espinha dorsal (*backbone*) da Internet acadêmica brasileira, operada pela RNP. Essa infraestrutura de rede óptica interliga universidades, institutos de pesquisa, hospitais de ensino e outras instituições em todo o território nacional, além de prover conexão com redes acadêmicas internacionais. O acesso direto à Rede Ipê é essencial para que os pesquisadores envolvidos no projeto Brasil 6G possam realizar transferências de grandes volumes de dados (*datasets*), além de acessar e compartilhar recursos computacionais de forma eficiente e segura.

Idealmente, o local de instalação será um dos Centros Nacionais de Dados (CNDs) da RNP. Esses CNDs estão localizados dentro de *datacenters* Tier III, parceiros da RNP, e estrategicamente distribuídos no Brasil, hospedando serviços críticos e a infraestrutura central da RNP. Instalar o sistema em CNDs traz vantagens como latência reduzida, maior disponibilidade de serviços de rede avançados e integração nativa com o ecossistema federado da RNP, um dos requisitos essenciais para este projeto. Atualmente, a RNP possui CNDs localizados em São Paulo e Brasília.

6.2 Requisitos de Ambiente e Infraestrutura Elétrica

Devido à alta densidade computacional dos servidores com GPUs, a infraestrutura de instalação deverá atender a rigorosos requisitos de ambiente físico e capacidade elétrica, visando garantir a operação contínua e segura do sistema [6]. Em relação aos aspectos térmicos, a sala destinada ao *datacenter* deverá manter uma temperatura ambiente controlada, idealmente entre 18°C e 27°C, com limites absolutos entre 5°C e 30°C, conforme recomendações de boas práticas de *datacenters* para evitar riscos de superaquecimento e falhas térmicas. Adicionalmente, o sistema de climatização deverá prover fluxo de ar contínuo, estável e direcionado. Para atender a carga térmica esperada, estima-se um fluxo mínimo de 2145 *Cubic Feet per Minute* (CFM) por servidor de GPUs. O *layout* físico da sala deverá adotar o padrão de corredores quentes e frios, visando maximizar a eficiência da refrigeração e o desempenho térmico dos equipamentos.

Além dos aspectos térmicos, a infraestrutura elétrica também deverá ser adequada. Em relação ao espaço físico, a solução completa, incluindo servidores de processamento, *switches* de rede e unidades de armazenamento, exigirá, no mínimo, dois *racks* padrão de 19 polegadas com altura de 42U cada. Deve-se também prever espaço adicional para circulação, manutenção e possíveis expansões futuras. A respeito da demanda energética da infraestrutura, a carga elétrica total estimada da solução é de, aproximadamente, 30 kW. O *datacenter* deverá ser capaz de fornecer essa potência de forma contínua, estável e com margem de segurança. Finalmente, a instalação elétrica deverá contar com, no mínimo, dois circuitos independentes e dedicados,

operando em 240 Volts, ambos protegidos por sistemas de fonte de energia ininterrupta ou *Uninterruptible Power Supply* (UPS) e alimentados por geradores de emergência, garantindo redundância total e alta disponibilidade de energia.

7 Elaboração do Termo de Referência para Aquisição

Em conformidade com sua política de governança e transparência, a RNP adota um processo estruturado e criterioso para aquisições de alto valor. Esse processo culmina na elaboração e divulgação pública de um Termo de Referência (TR) detalhado. O TR estabelece, de forma objetiva e isonômica, todos os requisitos técnicos, condições de fornecimento e critérios de avaliação necessários para assegurar que a solução adquirida atenda plenamente aos objetivos estratégicos do projeto.

O documento cobre todas as dimensões da solução, incluindo especificações técnicas de *hardware* e *software*, níveis mínimos de serviço para garantia e suporte, e requisitos de capacitação da equipe técnica.

Para a elaboração deste documento, foi conduzido um processo metodológico ao longo de várias semanas, envolvendo diferentes etapas de pesquisa, análise de mercado e validação técnica. As etapas principais consistiram em:

- Realizar reuniões técnicas com diversos fornecedores globais de servidores, incluindo Dell, Lenovo e Supermicro. O objetivo foi conduzir uma prospecção detalhada das soluções comercialmente disponíveis no Brasil, incluindo aspectos como disponibilidade de estoque, prazos de entrega e compatibilidade com os requisitos técnicos definidos. Esta etapa foi essencial para garantir que o conteúdo do TR refletisse soluções exequíveis e aderentes ao mercado nacional;

- Realizar, paralelamente à prospecção, uma série de reuniões internas para detalhar os requisitos da solução. Nessas discussões, a equipe do projeto Brasil 6G definiu parâmetros como a quantidade necessária de GPUs, a capacidade e o desempenho exigidos para o sistema de armazenamento, as velocidades mínimas de interconexão (InfiniBand e Ethernet), além de prazos e condições contratuais de garantia e suporte;

- Reconhecer que a solução vai além da simples aquisição de *hardware*, por meio de reuniões técnicas com a NVIDIA. O objetivo foi aprofundar o entendimento não apenas das características das GPUs, mas, principalmente, da camada de *software* que integra e otimiza o desempenho da solução, incluindo o pacote NVIDIA AI Enterprise;

- Incorporar uma visão técnica independente e imparcial, com a contratação de um consultor especializado em computação de alto desempenho. Esse profissional contribuiu com a elaboração de documentos técnicos de apoio, que auxiliaram na definição final dos requisitos do TR.

O resultado desse processo foi a produção do TR final, consolidado e publicizado, visando garantir ampla participação de fornecedores e a total transparência do processo de aquisição. O documento está disponível na plataforma oficial da RNP no seguinte endereço:

https://plataforma.rnp.br/documentos/fornecedores/2025/adc/14121/2025-aquisicao-de-hpc-projeto-brasil-6g

(Para acessar o conteúdo completo, é necessário selecionar a opção “Termo de Referência” no site indicado.)

8 Próximos Passos

Com a conclusão do processo de seleção e a iminente aquisição da infraestrutura de HPC, o projeto avança para as fases de planejamento detalhado, implantação e operacionalização. Como etapas subsequentes para a implantação da infraestrutura, a equipe da RNP seguirá o seguinte plano de ação, por meio de:

- Investigação das melhores práticas em infraestruturas de GPU compartilhadas para definir como o recurso computacional será disponibilizado aos usuários finais;
- Assinatura do contrato com o fornecedor selecionado, com base na melhor proposta, e efetivação da compra dos equipamentos descritos no TR;
- Solicitação ao fornecedor de um cronograma de entrega detalhado para todos os componentes da solução;
- Elaboração do plano de instalação detalhado, em um esforço conjunto entre as diversas áreas da RNP, e validação desse planejamento com o fornecedor;
- Criação do Plano de Governança e Sustentabilidade, documento que especificará as políticas de acesso, métodos de escalonamento de tarefas e as regras de compartilhamento dos recursos;
- Execução do treinamento da equipe que atuará diretamente na administração e sustentação da nova infraestrutura;
- Acompanhamento da instalação física e da integração lógica dos equipamentos. Essas atividades serão realizadas pelo fornecedor;
- Realização dos testes de aceitação (físicos e virtuais) para validar se a solução entregue está em conformidade com todos os requisitos do TR;
- Promoção dos treinamentos iniciais com os pesquisadores do projeto Brasil 6G e, na sequência, conduzir um período de operação assistida com um grupo piloto para refinar os processos e coletar *feedback*.

Ao concluir este ciclo de ações, o resultado esperado é a plena operacionalização de uma infraestrutura de HPC e IA, robusta e pronta para a produção científica. Espera-se ter não apenas os equipamentos fisicamente instalados, mas uma plataforma completamente funcional, homologada segundo os requisitos do TR, com uma governança clara de uso e com a equipe técnica da RNP devidamente capacitada para sua sustentação. A conclusão bem-sucedida da fase de operação assistida com os pesquisadores do projeto Brasil 6G validará os processos de acesso e colaboração, sendo assim a infraestrutura estará validada e pronta para ser utilizada como uma ferramenta de trabalho para impulsionar as pesquisas do projeto.

9 Conclusão

Este relatório técnico apresentou de forma detalhada as etapas realizadas e os resultados obtidos na Atividade 4.5 do projeto Brasil 6G, cujo foco foi a definição e especificação de uma infraestrutura de computação de alto desempenho (HPC) com aceleração por GPUs, destinada ao projeto Brasil 6G. Conforme descrito nas seções anteriores, a solução proposta desempenha um papel central no suporte ao desenvolvimento e validação de tecnologias emergentes voltadas para a sexta geração de redes móveis. Entre os benefícios esperados estão a viabilização de simulações complexas, o treinamento de modelos avançados de inteligência artificial e o processamento eficiente de grandes volumes de dados.

A análise técnica conduzida, incluindo a avaliação de alternativas de mercado, a definição de requisitos detalhados e a adoção da arquitetura BasePOD da NVIDIA, teve como base critérios objetivos de desempenho, escalabilidade e maturidade de *software*. A escolha da plataforma NVIDIA foi fundamentada não apenas na capacidade de processamento de seu *hardware*, mas também na abrangência e estabilidade de seu ecossistema de *software*, incluindo soluções como CUDA, Omniverse e Sionna, amplamente reconhecidas e utilizadas pela indústria.

Adicionalmente, a definição de uma camada de *software* baseada em contêineres, com orquestração por Kubernetes e integração com a suíte NVIDIA AI Enterprise, estabelece uma base tecnológica sólida para a gestão e otimização dos recursos computacionais disponíveis

Os requisitos específicos para instalação da infraestrutura, como a necessidade de integração com a RNP e seus CNDs, bem como os critérios ambientais e elétricos, foram definidos de modo a garantir a operação segura, eficiente e alinhada com as melhores práticas para ambientes de missão crítica.

O processo de elaboração do TR foi conduzido de forma transparente e colaborativa, envolvendo interações com fornecedores, discussões internas com a equipe técnica do projeto e validação por consultoria especializada. Tal abordagem visou assegurar a viabilidade técnica da solução, sua aderência aos objetivos do projeto Brasil 6G e sua conformidade com os princípios de governança e isonomia de processos de aquisição pública.

A infraestrutura de HPC detalhada nesta atividade representa, portanto, um investimento estratégico, com potencial para acelerar significativamente o desenvolvimento de tecnologias 6G no Brasil, contribuindo de forma direta para o avanço científico e tecnológico do país.

Referências

[1] "Nvidia basepod reference architectures," https://docs.nvidia.com/dgx-basepod/reference-architecture-infrastructure-foundation-enterprise-ai/latest/reference-architectures.html, 2025, [*Online*; acessado em 25 de junho de 2025].

[2] "Nvidia multi instance gpu," https://www.nvidia.com/pt-br/technologies/multi-instance-gpu/, 2024, [*Online*; acessado em 25 de junho de 2025].

[3] OpenAI, "Chatgpt," https://chatgpt.com/, 2025, acesso em: 30 jun. 2025.

[4] I. Analytics, "The leading generative ai companies," *IoT Analytics Insights Report*, 03 2025, available at: https://iot-analytics.com/wp-content/uploads/2025/03/INSIGHTS-RELEASE-The-leading-generative-AI-companies-vf.pdf.

[5] "Nvidia ai aerial sdk," https://developer.nvidia.com/aerial, 2024, [*Online*; acessado em 26 de junho de 2025].

[6] "Nvidia data center best practices," https://docs.nvidia.com/https:/docs.nvidia.com/nvidia-dgx-superpod-data-center-best-practices-with-dgx-b200.pdf, 2024, [*Online*; acessado em 26 de junho de 2025].