

TERMO DE REFERÊNCIA

Contratante: Fundação de Apoio ao Desenvolvimento da Computação Científica — FACC

Beneficiário: Laboratório Nacional de Computação Científica — LNCC

SUMÁRIO

1. OBJETO DA CONTRATAÇÃO

2. JUSTIFICATIVA DA CONTRATAÇÃO

3. DESCRIÇÃO DA SOLUÇÃO COMO UM TODO

4. REQUISITOS DA CONTRATAÇÃO

- 4.1.Requisitos Energéticos
- 4.2.Requisitos de Arquitetura e Hardware
- 4.3.Requisitos de Rede e Conectividade
- 4.4.Licenciamento
- 4.5.Metas de Desempenho
- 4.6.Software
- 4.7.Sistema de Armazenamento
- 4.8.Nós de Serviço e Gerenciamento
- 4.9.Segurança de Dados, Telemetria e Interoperabilidade de Hardware
- 4.10.Serviços
- 4.11.Documentação Técnica
- 4.12.Transferência de Tecnologia, Treinamento e Capacitação
- 4.13.Infraestrutura
- 4.14.Roadmap de Atualização Tecnológica
- 4.15.Colaboração Estratégica

5. EXECUÇÃO DO OBJETO

- 5.1.Testes de Aceitação
- 5.2. Prazo e Local de Entrega

6. ESTIMATIVA DE VALOR DA CONTRATAÇÃO E ADEQUAÇÃO ORÇAMENTÁRIA

7. DO CONTRATO

8. CRITÉRIOS DE MEDIÇÃO E DE PAGAMENTO

- 8.1.Condições Gerais
- 8.2.Cronograma de Pagamento

9. FORMA E CRITÉRIOS DE SELEÇÃO DO FORNECEDOR

9.1. Da Proposta

9.2. Tabela de Pontuação para Escolha da Solução

9.2.1. Avaliação da Proposta

9.2.2. Critérios, Pesos e Regras por Critério

9.2.3. Regras Gerais

9.3. Normalização

9.3.1. Classificação do Critério (9.2.2)

9.3.2. Nota Técnica Final

9.4. Avaliação da proposta de preço

9.5. Nota Final de Técnica e Preço

10. GARANTIAS CONTRATUAIS: DISPOSITIVO DE GARANTIA DE EXECUÇÃO E GARANTIA FINANCEIRA

9.6. Fundamentação Legal das Garantias

9.7. Garantia Técnica do Objeto (Garantia do Fabricante)

9.8. Garantia de Execução Contratual

9.9. Garantia Financeira (Cobertura e Beneficiário)

9.10. Garantia Complementar para Serviços de Engenharia de Grande Vulto

11. ACORDO DE NÍVEIS DE SERVIÇO, FISCALIZAÇÃO E REGIME SANCIONATÓRIO

9.11. Disposições Gerais do SLA

9.12. Sanções

9.13. Regime de Sanções Administrativas Subsidiário

12. DISPOSIÇÕES FINAIS

1. OBJETO DA CONTRATAÇÃO

O presente Termo de Referência tem por objeto a contratação de solução integrada de Processamento Computacional de Alto Desempenho (High Performance Computing – HPC) voltada à Inteligência Artificial, pesquisa, desenvolvimento tecnológico e inovação.

Solução de infraestrutura de supercomputação de alto desempenho destinada ao desenvolvimento, treinamento e inferência de modelos avançados de Inteligência Artificial — em especial Grandes Modelos de Linguagem (LLMs) — para uso estratégico do setor público brasileiro, da comunidade científica e do ecossistema nacional de inovação em IA.

A contratação não se limita à entrega de equipamento: é instrumento de política de ciência, tecnologia e inovação, e deve contribuir concretamente para a soberania tecnológica nacional.

Para tanto, as propostas deverão contemplar, além dos requisitos técnicos de desempenho, compromissos mensuráveis de transferência de tecnologia, capacitação de profissionais brasileiros e co-desenvolvimento da pilha nacional de software de IA.

2. JUSTIFICATIVA DA CONTRATAÇÃO

A presente contratação justifica-se pela necessidade de ampliar e fortalecer a capacidade computacional nacional destinada ao desenvolvimento, treinamento, validação e operação de modelos avançados de Inteligência Artificial, especialmente Grandes Modelos de Linguagem (LLMs), Small Language Models (SLMs) e Domain Language Models (DLMs), para aplicações estratégicas do setor público brasileiro, da comunidade científica e do ecossistema nacional de inovação.

A contratação está inserida no âmbito do Plano Brasileiro de Inteligência Artificial – PBIA 2024–2028, especialmente na Ação 1 do Eixo 1, que prevê a ampliação da capacidade

computacional nacional dedicada à Inteligência Artificial, constituindo iniciativa estratégica para o fortalecimento da ciência, tecnologia e inovação e para a promoção da soberania e da autonomia tecnológica do País.

A necessidade decorre da crescente demanda por capacidade de processamento de alto desempenho para pesquisa, desenvolvimento e aplicação de soluções de Inteligência Artificial, bem como da insuficiência de infraestrutura computacional nacional em escala compatível com as necessidades de treinamento e operação de modelos avançados. A dependência de infraestrutura computacional estrangeira, além de elevar custos e criar limitações de disponibilidade e escala, pode resultar em dependência tecnológica em área considerada estratégica para o desenvolvimento econômico, científico, tecnológico e institucional do País.

A solução pretendida não se limita à aquisição de equipamentos de processamento. Trata-se da implantação de infraestrutura integrada de supercomputação de alto desempenho, abrangendo os recursos computacionais, armazenamento, interconexão de alta velocidade,

software, instalação, integração, configuração, testes, garantia, suporte técnico, manutenção, capacitação e transferência de tecnologia necessários ao pleno funcionamento da solução.

A contratação deverá contribuir concretamente para:

- I – ampliar a capacidade brasileira de treinamento e operação de modelos de Inteligência Artificial em escala compatível com padrões internacionais;
- II – disponibilizar infraestrutura computacional para pesquisadores, instituições públicas, universidades, centros de pesquisa e demais usuários autorizados, observadas as regras de governança e utilização definidas para o projeto;
- III – possibilitar o desenvolvimento e a operação de modelos de IA em língua portuguesa e voltados às necessidades e características do contexto brasileiro;
- IV – apoiar o desenvolvimento de aplicações estratégicas de Inteligência Artificial para o setor público e para o ecossistema nacional de ciência, tecnologia e inovação;
- V – promover a formação e capacitação de profissionais brasileiros especializados em computação de alto desempenho, infraestrutura de IA e operação de sistemas computacionais de grande escala;
- VI – promover transferência de conhecimento e tecnologia, permitindo a formação progressiva de capacidade técnica nacional para operação, manutenção, administração e evolução da infraestrutura;
- VII – reduzir dependências tecnológicas estruturais em relação a infraestruturas computacionais estrangeiras; e
- VIII – criar condições para o desenvolvimento de uma capacidade computacional nacional sustentável e escalável para pesquisa, inovação e aplicações estratégicas de Inteligência Artificial.

A contratação será operacionalizada por intermédio de Fundação de Apoio constituída nos termos da Lei nº 8.958/1994, no âmbito de projeto de pesquisa, desenvolvimento e inovação. Nesse contexto, serão observados o regime jurídico aplicável às fundações de apoio e os procedimentos estabelecidos pelo Decreto nº 8.241/2014, especialmente quanto à pesquisa de mercado, definição do valor de referência, seleção da proposta mais vantajosa, julgamento, negociação, contratação e demais atos necessários à adequada formalização do ajuste.

Considerando o vulto, a complexidade técnica e a relevância estratégica do objeto, serão adotadas, adicionalmente e no que forem compatíveis com a natureza jurídico-privada da Fundação de Apoio, boas práticas de planejamento, governança, gestão de riscos, fiscalização, execução contratual e controle previstas na Lei nº 14.133/2021, especialmente aquelas relacionadas à adequada caracterização do objeto, definição de requisitos técnicos, critérios objetivos de seleção, gestão contratual, garantias, fiscalização e responsabilização.

A contratação também se justifica pela necessidade de incorporar mecanismos de transferência de tecnologia e capacitação de profissionais brasileiros. Dessa forma, espera-se que o investimento produza, além da infraestrutura física e tecnológica, resultados permanentes em termos de conhecimento, formação de recursos humanos, capacidade de operação e desenvolvimento tecnológico nacional.

A não realização da contratação poderá comprometer a execução das ações previstas no PBIA 2024–2028 relacionadas à ampliação da capacidade computacional para Inteligência Artificial, mantendo a dependência de infraestrutura computacional externa e limitando a capacidade nacional de desenvolver, treinar e operar modelos avançados de IA em escala adequada.

Impacto e benefícios esperados: a infraestrutura contratada deverá, ao longo do período de garantia, ampliar substancialmente a capacidade nacional de: (i) treinar e operar LLMs, SLMs e DLMs em português de escala competitiva internacionalmente; (ii) atender a demanda de

múltiplos usuários simultâneos — pesquisadores, órgãos públicos e empresas — com níveis de serviço definidos; (iii) formar profissionais brasileiros capazes de operar, manter e evoluir a infraestrutura de forma progressivamente autônoma; e (iv) contribuir para o desenvolvimento da pilha nacional de software de IA, reduzindo dependências estruturais de plataformas estrangeiras.

Dessa forma, a contratação mostra-se necessária, adequada e alinhada ao interesse público, por constituir investimento estruturante em infraestrutura científica e tecnológica, com potencial de gerar benefícios de longo prazo para a pesquisa, a inovação, a transformação digital do Estado, a formação de recursos humanos e o desenvolvimento da soberania tecnológica brasileira.

3. DESCRIÇÃO DA SOLUÇÃO COMO UM TODO

Considerado todo o ciclo de vida do objeto, a solução a ser contratada compreende, de forma integrada e indissociável: (i) o fornecimento de infraestrutura de supercomputação — nós computacionais, nós de serviço, sistema de armazenamento em três camadas, rede de interconexão de alta velocidade e switches de borda —, conforme requisitos detalhados no item 5; (ii) os serviços de instalação física, comissionamento e testes de aceitação (itens 5.1); (iii) a garantia técnica integral de hardware, software e firmware pelo prazo mínimo de 5 (cinco) anos, com serviços de suporte, monitoramento e manutenção associados (itens 4.10.2 e 10.2); (iv) o licenciamento de software e a entrega de documentação técnica completa (itens 4.5 e 4.11); (v) os programas de transferência de tecnologia, capacitação técnica, mentoria especializada e acesso antecipado à infraestrutura (item 4.12); e (vi) o plano de colaboração estratégica com institutos de pesquisa, centros de P&D e a cadeia produtiva nacional (item 4.15).

Essa descrição integrada evidencia que o objeto não se limita ao fornecimento de bens de tecnologia da informação, mas configura solução complexa de infraestrutura crítica associada

a serviços continuados e a obrigações de transferência de conhecimento, o que justifica o tratamento unificado da contratação sob um único Termo de Referência e um único instrumento contratual, vedado o fracionamento do objeto.

4. REQUISITOS DA CONTRATAÇÃO

4.1. Requisitos Energéticos

A máquina a ser adquirida por meio deste Termo de Referência será instalada em um ambiente com limite de energia de 3 MW para os nós de processamento, armazenamento,

equipamentos de conectividade, etc. Todas as propostas devem incluir uma estimativa do consumo de energia da solução oferecida para comprovar o atendimento a este limite.

4.2. Requisitos de Arquitetura e Hardware

Todos os componentes de hardware fornecidos deverão ser originais de fábrica, novos (zero uso), não reconicionados, não remanufaturados e não reutilizados;

Os equipamentos deverão possuir garantia diretamente do fabricante, com suporte técnico oficial durante todo o período de garantia da proposta;

O fabricante não poderá ter publicado, até a data da entrega, aviso de End of Life (EOL) ou End of Support (EOS) dos componentes ofertados;

O fornecedor deverá apresentar declaração do fabricante confirmando:

- que os componentes são novos;
- que possuem garantia oficial válida no Brasil;
- que o produto não está em EOL/EOS e que haverá disponibilidade de peças e suporte durante todo o período de garantia.

Aceleradores de IA: Deverão ser equipados com unidades de processamento (GPUs, NPUs, etc.) que suportam operações em FP16. O suporte a FP4 e FP64 é requisitos desejáveis visando a execução de modelos de IA de próxima geração sem comprometer a acurácia, assim como permitir a execução de aplicações científicas, industriais e de projetos de inovação.

Compartilhamento de Aceleradores: É desejável que o sistema tenha a capacidade de particionamento seguro e eficiente, permitindo que múltiplos usuários compartilhem o mesmo dispositivo com isolamento e alta utilização.

Confidential Computing: A infraestrutura deve suportar a execução de cargas de trabalho em Trusted Execution Environments (TEEs), garantindo a proteção criptográfica de dados

sensíveis e modelos proprietários durante a execução. A conformidade com este requisito deverá ser apresentada durante os testes de aceitação (item 5.1) com demonstração funcional de ao menos um workload executado em ambiente TEE. Na ausência de comprovação, o

requisito será tratado como não atendido e o item correspondente receberá pontuação zero na avaliação técnica.

Os nós computacionais, nós de serviço, storages e dispositivos de interconexão devem ser projetados com recursos de tolerância a falhas para os componentes críticos;

Nós computacionais de IA:

- Armazenamento local SSD NVMe;
- Interface de gerenciamento exclusiva;
- Interface para rede de alta velocidade - 400 Gbps ou superior;

Nós de serviço:

- Mínimo de duas unidades de armazenamento local configuradas em RAID1;
- Interface de gerenciamento exclusiva;
- 2x interfaces de 25/10 Gbps para conexão ao switch de borda da solução;
- 1x interface para rede de alta velocidade - 400 Gbps ou superior;
- 2x unidades de CPU;
- Mínimo de 128 GB por unidade de CPU;

4.3. Requisitos de Rede e Conectividade

Para suportar a execução eficiente de modelos de larga escala com paralelismo de dados ou de modelo, é essencial garantir:

- Largura de banda agregada mínima de 1 TB/s por nó para comunicação GPU-GPU.

- Topologia de interconexão de alta velocidade e baixa latência, preferencialmente com conectividade plena entre as unidades de processamento (full-mesh ou crossbar), permitindo operações de coleta (AllGather), redução (AllReduce) e broadcast com mínima penalidade de comunicação.
- Latência inferior a 1 microssegundo para mensagens curtas entre GPUs de um mesmo nó, evitando gargalos na sincronização de gradientes e no acesso a tensores distribuídos.

4.3.1. Comunicação Inter-nó

A escalabilidade do sistema exige um plano de interconexão eficiente entre nós computacionais. Para isso, são requeridas:

- Interfaces de rede com largura de banda de pelo menos 400 Gbps por nó, preferencialmente com suporte a múltiplos fluxos paralelos e operação em modo full-duplex.
- Protocolos de comunicação direta entre dispositivos (peer-to-peer), eliminando a necessidade de movimentação de dados via memória principal ou CPU host, e reduzindo a latência para operações entre GPUs localizadas em nós distintos.
- Suporte a RDMA (Remote Direct Memory Access) para permitir a movimentação eficiente de dados com latência mínima e sem overhead da CPU.
- Topologia de rede de alta conectividade (ex: fat-tree, dragonfly, ou topologias customizadas não bloqueantes).

4.3.2. Acesso à Rede

A solução deverá dispor de um switch para conexão entre o supercomputador e a rede do Datacenter. Esse equipamento deverá ter suporte à OSPF (Open Shortest Path First) e BGP (Border Gateway Protocol), possuir, pelo menos, 4 portas 400/100 Gbps QSFP-DD (Quad Small

Form-Factor Pluggable - Double Density) para conexão entre o supercomputador e a rede do LNCC/RNP e 32 portas 25/10 Gbps para conexão dos nós de serviço fornecidos na proposta.

4.4. Licenciamento

A proposta deverá incluir um modelo de licenciamento que assegure ao comprador:

- licença perpétua e irrevogável para uso dos softwares embarcados essenciais à operação do sistema;
- permitir a participação na definição de estratégias de customização e otimização dos softwares;
- Acesso ao design de servidores (desejável): é desejável que o fornecedor disponibilize ao comprador a documentação técnica de design de servidores de gerações anteriores (mínimo duas gerações), incluindo especificações de motherboard, interfaces de expansão e firmware de gerenciamento. O objetivo é habilitar a capacidade nacional de integrar aceleradores de diferentes fabricantes em gerações futuras de

infraestrutura, sem dependência do fornecedor original. Propostas que incluam este compromisso receberão pontuação adicional no critério 08 da tabela de avaliação.

Os softwares considerados essenciais incluem, mas não se limitam a:

- sistemas de gerenciamento e orquestração;
- ferramentas de scheduling;
- sistemas de monitoramento térmico e energético;
- controladores e gerenciadores de GPU;
- Firmware de BMC (Baseboard Management Controller), incluindo interfaces IPMI e Redfish;
- Firmware de controladoras de armazenamento e switches de rede;
- demais componentes necessários ao funcionamento pleno da infraestrutura.

A proposta deverá detalhar:

- O escopo exato da licença e suas condições técnicas e legais;
- O processo para atualização, manutenção e documentação das versões licenciadas;
- Eventuais restrições e como elas se alinham ao objetivo de autonomia tecnológica nacional.

4.5. Metas de Desempenho

Para garantir ampla concorrência e comparabilidade objetiva entre propostas, o proponente deverá apresentar resultados oficiais e reproduzíveis dos benchmarks especificados a seguir.

A máquina deverá ser composta por um conjunto de componentes otimizados para treinamento e inferência de LLMs e será tratada como um System Under Test - SUT distinto

no sentido das regras do MLPerf, devendo manter consistência de hardware, software, drivers, bibliotecas, topologia de interconexão e configurações durante toda a execução dos testes e na fase de aceitação. É expressamente vedada qualquer alteração de componentes do SUT entre medições que comprometam a comparabilidade; aplica-se, no que couber, a regra de que o SUT descrito deve corresponder ao SUT efetivamente medido, com replicabilidade obrigatória conforme o MLPerf.

O proponente deverá executar e reportar, de forma separada e independente, os seguintes benchmarks, observando integralmente as regras oficiais do MLPerf Inference v6.0 e do MLPerf Training v6.0:

- (a) MLPerf Inference v6.0 (Datacenter) para LLMs, incluindo os workloads oficiais Llama2-70B, Llama3.1-405B, Mixtral-8x7B e DeepSeek-R1, nos cenários Offline (throughput) e Server/Interactive (tráfego Poisson com restrições de TTFT/TPOT [time to first token/time per output token]), sempre com verificação de qualidade (accuracy) nos níveis exigidos;
- (b) MLPerf Training v6.0 para LLMs, cobrindo fine-tuning (Llama 2 70B LoRA / GovReport) e pretraining (Llama 3.1-8B / C4), com métrica principal tempo-até-qualidade (time-to-quality).

Para (a) e (b), aplicam-se as definições de cenário, métricas, metas de qualidade e demais requisitos estabelecidos, respectivamente, no MLPerf Inference v6.0 e no MLPerf Training v6.0; o tráfego e as métricas de inferência serão gerados e medidos pelo LoadGen, componente oficial do MLPerf, e todos os resultados deverão respeitar as regras de “fair play” (proibição de benchmark detection, controle de não-determinismo, ordem de dados, thermal throttling, entre outros).

A medição de energia será obrigatória e parte integrante da avaliação. Para inferência, a energia deverá ser medida “at the wall” durante a fase de performance, utilizando o workflow

MLPerf Power (PTDaemon/SPEC e analisador compatível), com correlação temporal aos logs do LoadGen; o proponente deverá entregar os logs de potência e a configuração utilizada.

A fim de comprovar a escalabilidade real e evitar “ilhas de desempenho” não representativas do cluster entregue, cada benchmark exigido acima deverá ser executado em dois tamanhos: primeiro utilizando 50% dos recursos (aceleradores) e, em seguida, utilizando 100% dos recursos. Nos casos em que o proponente não conseguir realizar os testes com 50% e 100% dos recursos, poderá ser realizada uma normalização para extrapolação dos valores do SpeedUp; por exemplo: se o proponente somente tiver acesso a 10% dos recursos, este será normalizado como o valor de Throughput(100%) e os 5% de recursos serão agora os valores de 50% aplicados na razão.

A pontuação técnica será calculada por normalização e ponderação dos benchmarks. Cada proponente irá reportar seus valores de benchmark, não sendo de competência deste documento impor limites mínimos para estas métricas. Sendo assim, o proponente que estabelecer os valores mais representativos será automaticamente classificado como referência de normalização para os demais proponentes.

Em cada benchmark, as métricas serão normalizadas entre os proponentes: quando “maior é melhor”, o escore normalizado é a razão entre o valor do proponente e o melhor valor apresentado; quando “menor é melhor”, é a razão entre o melhor valor observado e o valor do proponente. Para MLPerf Inference, a métrica agregada por modelo/cenário será tokens/s/W, combinando desempenho e energia “at the wall”; para MLPerf Training, a métrica agregada por tarefa será 1/(kWh-para-qualidade), refletindo energia total consumida até atingir a qualidade alvo. O escore técnico será a soma ponderada de 50% para MLPerf

Inference (média geométrica entre os workloads exigidos) e 50% para MLPerf Training (média geométrica entre fine-tuning e pre-training).

Os resultados apresentados deverão ser reprodutíveis e reproduzíveis na fase de aceitação. O fornecedor vencedor deverá disponibilizar recipientes (contêineres) e/ou scripts, além de toda a configuração necessária, para reexecução testemunhada de um subconjunto representativo dos testes.

Por fim, visando à validação da eficiência de integração e escalabilidade do cluster de GPUs, o fornecedor deverá submeter uma estimativa na proposta e demonstrar o MFU (Model FLOPs Utilization) durante os testes de aceite. O MFU é definido como a razão entre a taxa de FLOPs executada (calculada com base em $6 \times (\text{número de parâmetros}) \times (\text{tokens por segundo})$) e a capacidade teórica de pico da GPU para a precisão utilizada (por exemplo, BF16, FP16 ou FP8). O MFU deve ser medido/estimado para o cenário de pré-treinamento do modelo Llama 3.1-8B.

Todos os relatórios deverão ser acompanhados dos logs oficiais: no caso do MLPerf, os arquivos summary/detail/accuracy do LoadGen, bem como os artefatos de submissão (quando aplicável) em conformidade com o guia oficial da versão correspondente; no caso do MLPerf Power, os logs do servidor e do cliente de potência, especificação do analisador, ranges utilizados e sincronização de tempo. Esses materiais são obrigatórios para auditoria e verificação de conformidade.

4.6. Software

Soluções open-source, proprietárias ou híbridas serão aceitas, desde que atendam integralmente aos requisitos funcionais.

Para implementar soluções robustas de IA em escala nacional, é essencial definir stacks de software completos, organizados por metodologia e compatíveis com diferentes arquiteturas

de hardware. Esses stacks devem garantir interoperabilidade, escalabilidade e suporte a aplicações que abrangem desde ANI (Inteligência Artificial Estreita) até pesquisas avançadas para AGI (Inteligência Artificial Geral). A seguir, apresentamos uma taxonomia detalhada dos principais stacks, divididos por metodologia.

O fornecedor deverá entregar, instalar, configurar e comprovar o funcionamento de um software stack completo e padronizado para Aprendizado de Máquina (ML), Aprendizado Profundo (DL) e Inteligência Artificial Simbólica, em todas as arquiteturas contempladas neste Termo de Referência, observando rigorosamente os requisitos mínimos de versões, compatibilidade, execução distribuída, containerização, segurança, governança e independência tecnológica aqui descritos. Todo o stack será instalado nos nós do cluster conforme a partição de hardware definida pelo edital, devendo permanecer consistente e

reprodutível do ponto de vista de binários, bibliotecas, runtimes, drivers e toolkits, durante a homologação, aceitação e operação.

4.6.1. Stack para Aprendizado de Máquina (ML)

Requisitos gerais (obrigatórios):

- Suporte a aprendizado supervisionado, não supervisionado e por reforço (RL).
- Execução distribuída com MPI (Message Passing Interface) e integração com frameworks de paralelismo.
- Suporte a clusters com gerenciador de filas e agendador de jobs.
- Aceleração por GPU quando disponível, mantendo paridade funcional com execução apenas em CPU.

Frameworks principais:

- PyTorch 2.5+ e TensorFlow 2.17+ com suporte a aceleração por GPU.
- scikit-learn com aceleração em GPU via bibliotecas de dataframes/ML (p.ex., cuML/RAPIDS ou equivalente).
- XGBoost 2.1+ com suporte GPU e execução em CPU, e LightGBM com variantes GPU e CPU.

Big Data e Distribuição:

- Apache Spark com acelerador para GPU (p.ex., RAPIDS Accelerator ou equivalente).
- Ray 3.0+ (incluindo RLlib para aprendizado por reforço multiagente).
- Dask com suporte a dataframe em GPU (p.ex., cuDF ou equivalente).

Ferramentas avançadas:

- Motor de simulação física para RL com suporte a GPU (p.ex., Isaac Gym ou equivalente).
- OmniGraph ou equivalente para orquestração/grafos de execução.
- Optuna para otimização de hiperparâmetros.

Infraestrutura HPC:

- Horovod para treinamento distribuído.

- Biblioteca de comunicação coletiva entre aceleradores (AllReduce/AllGather), p.ex., NCCL 2.20+ ou equivalente.
- Toolkit de aceleração para GPU (p.ex., CUDA Toolkit 12.6 ou equivalente).
- SLURM 24.x e OpenMPI 5.0 para gerenciamento e orquestração de jobs em cluster.

Contêineres e Deploy:

- Ambientes isolados para HPC com suporte a GPU (p.ex., Enroot/Pyxis ou equivalente).
- Kubeflow 2.0 para pipelines de ML.
- Servidor de inferência multi-framework (p.ex., Triton Inference Server 24.10 ou equivalente).

Execução apenas em CPU (quando aplicável):

- scikit-learn 1.5+, XGBoost (CPU), LightGBM (CPU).
- Stable Baselines3 para RL.
- Optuna para tuning.
- Bibliotecas BLAS/LAPACK de alto desempenho (OpenBLAS 0.3.27, oneAPI MKL, ou equivalente).
- Podman para contêineres leves.

4.6.2. Stack para Aprendizado Profundo (DL)

Frameworks DL:

- PyTorch 2.5+, TensorFlow 2.17+, JAX 0.4.9+.

Bibliotecas avançadas:

- Transformers 4.46+ (Hugging Face) para LLMs.
- Diffusers 0.30+ para IA generativa (Stable Diffusion).
- OpenCV 4.10+ para visão computacional.
- YOLOv10/Ultralytics e Detectron2 para detecção de objetos.

NLP e Cadeias de Raciocínio:

- spaCy 3.8+ com aceleração quando disponível.

- LangChain 0.3+, vLLM 0.6+.

Inferência e Otimização:

- Otimizador/compilador de inferência para LLMs e CNNs (p.ex., TensorRT-LLM 0.20+ ou equivalente).
- NVTabular ou equivalente para pré-processamento em GPU.

Computação Quântica Aplicada:

- Biblioteca para simulação quântica acelerada (p.ex., cuQuantum 24.10 ou equivalente).
- Qiskit (IBM) e PennyLane para integração quântica.

Infraestrutura:

- Toolkit de contêiner para exposição de GPUs em contêineres (p.ex., Container Toolkit 1.17+ ou equivalente).

Execução apenas em CPU (quando aplicável):

- PyTorch (CPU), TensorFlow Lite, ONNX Runtime 1.20+.
- spaCy (CPU), OpenCV (CPU), Keras 3.6+.

4.6.3. Stack para Inteligência Artificial Simbólica (IA Baseada em Regras)

- SWI-Prolog 9.4+, CLIPS 6.4+ (sistemas especialistas).
- Answer Set Programming (ASP) com Clingo 5.7+.
- Drools 9.0+ (regras em Java), Jess 80u2.
- Ontologias OWL/RDF via Apache Jena 5.2+.

- Bibliotecas Neuro-Simbólicas: SAIF-DL, Logic Tensor Networks (LTN).
- Grakn/Vectis (knowledge graphs) e NetworkX para grafos simbólicos.

4.6.4. Escopo Técnico e Requisitos Estratégicos

O stack deverá contemplar:

- Ciclo de vida híbrido: integração ANI/AGI/ASI via DL + IA Simbólica.
- Pré-treinamento distribuído: Transformers, CNNs, RNNs, RL.
- Funcionalidades específicas:
 - IA Reativa (regras/CNNs leves).
 - Memória Limitada (LSTMs).
 - Teoria da Mente (NLP empático).
 - Autoconsciente (Neuro-Simbólica total).
- Governança e segurança: mitigação de vieses e alucinações via ASP e validadores de regras/ontologias.

4.6.5. Independência Tecnológica e Governança

Interoperabilidade Open-Source:

- Exportabilidade obrigatória de modelos (critério eliminatório): todos os modelos de IA treinados na pilha fornecida deverão ser exportáveis em formato padrão aberto - Open Neural Network Exchange (ONNX) ou equivalente com adoção comprovada pelo mercado. A exportabilidade deverá ser demonstrada nos testes de aceitação com ao menos um modelo treinado na infraestrutura entregue, executado com sucesso em ambiente externo ao do fornecedor. A impossibilidade de exportar modelos para execução em infraestruturas não pertencentes ao fornecedor constitui critério

eliminatório, pois comprometeria o uso do LLM Nacional por órgãos públicos, pela Defesa e por parceiros que operem infraestrutura distinta.

- RDF/OWL para IA Simbólica.

Evitar lock-ins:

- Suporte a APIs abertas e alternativas de aceleração (CUDA/HIP/ROCm/SYCL/OpenCL, entre outras), com seleção baseada em desempenho, compatibilidade e portabilidade.

Compatibilidade multiambiente:

- Execução em CPU e GPU (quando disponível), contemplando múltiplos sistemas operacionais suportados.

Governança EBIA/PBIA:

- Participação em comitês técnicos, com evidências de monitoramento e evolução da taxonomia interna de IA.

4.6.6. Critérios de Aceitação, Evidências e Desempenho Mínimo

O Proponente deverá apresentar:

1. Matriz de Conformidade, com links para documentação oficial e comandos de verificação (versão/build).
2. Relatórios de benchmark comprovando:
 - Treinamento distribuído (Horovod/MPI) com aceleração por GPU: throughput e escala quase linear em 2, 4 e 8 GPUs (quando houver).
 - Inferência com servidor multi-framework: latência p50/p95 e QPS em modelos CNN (visão) e LLM (NLP).

- Spark com acelerador de GPU: speedup vs. execução apenas em CPU em ETL de referência.

3. Testes de interoperabilidade:

- Exportação/execução via ONNX.
- Chaining neuro-simbólico (DL + regras/ontologias) com casos de conformidade.

4. Evidências de governança:

- Procedimentos de mitigação de vieses/alucinações com ASP e validações de regras.

5. Automação/DevOps:

- Manifests de Kubeflow e templates de SLURM/OpenMPI para jobs distribuídos.
- Manuais de operação em execução apenas em CPU e com aceleração por GPU.

Cláusula de Equivalência Tecnológica. Sempre que for indicado um componente por nome, admite-se “...ou equivalente”, desde que o Proponente comprove equivalência funcional, compatibilidade com os demais itens do stack e desempenho igual ou superior nos benchmarks exigidos. O mesmo é válido para aceleradores de hardware, onde se lê compatibilidade com CPU/GPU também está atribuída a liberdade de equivalência para APU

ou demais tecnologias de processamento que permitam a execução das Stack de software mínima solicitada. A ausência de comprovação enseja desclassificação.

4.7. Sistema de Armazenamento

O supercomputador requer um sistema de armazenamento (subsistema de Entrada/Saída - E/S ou I/O) capaz de atender a duas cargas de trabalho de armazenamento distintas. Ela será formada por, pelo menos, três camadas de armazenamento.

O Supercomputador lidará com uma Carga de Trabalho de IA, que exige alto desempenho das operações de E/S para leituras pequenas e aleatórias e requer um recurso compartilhado que possa ser acessado por todos os nós computacionais simultaneamente.

O sistema de armazenamento precisa suportar a utilização de um único nó de computação, uma porcentagem de nós de computação (por exemplo, 20%) e o sistema completo (ou seja, pelo menos 90% dos nós computacionais).

4.7.1. Camada 0

A Camada 0 de armazenamento será formada por dispositivos NVMe de alta velocidade em cada nó computacional, para serem utilizados como cache as read-only ou read-write.

1. Requisitos Camada 0

- a. Utilizar dispositivos NVMe, interfaces PCIe Gen4 ou Gen5, com IOPS elevadas e latência abaixo de 100 µs.

O sistema de armazenamento proposto deverá ser composto exclusivamente por dispositivos de estado sólido com tecnologia NVMe nativa, utilizando interface física barramento PCIe Gen4 ou superior (PCIe Gen5). Para garantir a consistência de desempenho e a previsibilidade sob alta carga de trabalho, os dispositivos deverão atender, cumulativamente, aos seguintes critérios de

latência e IOPS: Latência de Leitura (Percentil P99): Igual ou inferior a 100 μ s (cem microssegundos). Latência de Escrita (Percentil P99): Igual ou inferior a 100 μ s (cem microssegundos). Condição de Validação (Carga de Teste): Os limites de latência descritos acima deverão ser sustentados de forma contínua durante testes com blocos de dados de 4KB (Random Read/Write), sob um perfil de carga mista de 70% leitura e 30% escrita (70/30 IO Mix), atingindo o teto de IOPS máxima homologada pelo fabricante. Estabilidade sob Estresse: Não serão aceitas métricas baseadas exclusivamente em latência média (Average Latency) ou em cenários de "Burst" (picos de curta duração com o disco em repouso). O dispositivo deve manter a latência P99 exigida mesmo após atingir o estado de equilíbrio (Steady State).

- b. Desempenho de throughput (128KB) de 5000 MB/s para leitura e escrita: (i) Desempenho Sequencial Sustentado: Cada dispositivo deverá apresentar throughput mínimo de 5.000 MB/s para leitura sequencial e 5.000 MB/s para escrita sequencial, utilizando blocos de 128 KB; (ii) Condição de Validação (Carga de Teste): As medições deverão ser realizadas com profundidade de fila (Queue Depth) suficiente para saturação do dispositivo, utilizando no mínimo 128 threads ou configuração equivalente recomendada pelo fabricante; (iii) Sustentação de Desempenho: O throughput mínimo exigido deverá ser mantido durante toda a janela de medição, não sendo aceitos resultados obtidos exclusivamente por mecanismos temporários de cache SLC, buffer DRAM ou condições de operação em estado ocioso; (iv) Estado de Equilíbrio: As métricas deverão ser aferidas após o dispositivo atingir condição de Steady State, garantindo representatividade para cargas contínuas de treinamento e inferência de modelos de IA.
- c. Desempenho de IOPS (4KB) de 400,000 para leitura e escrita; (i) Desempenho Aleatório Sustentado: Cada dispositivo deverá apresentar desempenho mínimo de 400.000 IOPS para leitura aleatória e 400.000 IOPS para escrita aleatória,

utilizando blocos de 4 KB; (ii)Condição de Validação (Carga de Teste): As medições deverão ser realizadas em regime permanente (Steady State), com carga uniforme distribuída em toda a capacidade utilizável do dispositivo e profundidade de fila suficiente para saturação dos recursos internos de processamento; (iii)Consistência Operacional: Os valores mínimos exigidos deverão ser mantidos de forma contínua durante a execução do teste, não sendo aceitos resultados obtidos exclusivamente em modo Burst ou durante períodos transitórios de aceleração por cache;(iv)Relação Desempenho-Latência: O atendimento ao requisito de IOPS não poderá resultar em degradação da latência além dos limites especificados neste documento.

- d. Latência de 80 μ s para leitura e escrita: Latência Sustentada: Os dispositivos deverão apresentar latência de leitura e escrita igual ou inferior a 80 μ s (oitenta microssegundos), medida no percentil P99.Condição de Validação (Carga de Teste): Os limites de latência deverão ser mantidos durante operações aleatórias com blocos de 4 KB, sob carga mista de 70% leitura e 30% escrita, executadas em regime de Steady State e na máxima taxa de IOPS sustentada suportada pelo dispositivo.Critério de Aceitação: Não serão aceitas medições baseadas exclusivamente em latência média (Average Latency), medianas ou cenários de

Burst. O requisito deverá ser comprovado através da apresentação da distribuição estatística de latência, incluindo os percentis P95, P99.

- e. Esse dispositivo não deverá ser o mesmo utilizado como o armazenamento principal do nó computacional (ou seja, não deve ser utilizado para instalação do Sistema Operacional)
- f. Deverá ser possível criar um sistema de arquivos paralelo temporário (PFST), utilizando os dispositivos formatados e montados localmente nos nós computacionais para serem usados como dispositivos de cache.
- g. Esse PFST deve ser dinâmico, com a criação sob demanda.
- h. Esse PFST deve suportar ações administrativas para conectar e desconectar os dispositivos locais para fins de cache.
- i. Esse cache deve ser distribuído e consistente em todos os nós computacionais que participarem do PFST, de modo que quaisquer arquivos atualizados sejam automaticamente removidos do cache e os aplicativos sempre recebam a versão mais recente dos dados. Esse processo deve ser completamente transparente para os aplicativos e toda a movimentação de dados deve ser tratada automaticamente como parte da E/S da aplicação.

4.7.2. Camada 1

A camada 1 será destinada ao armazenamento dos dados quentes (dados em processamento), com área útil mínima de 20 PB (petabytes), utilizando arquitetura de armazenamento paralela e distribuída, com suporte a múltiplos fluxos simultâneos e integração com bibliotecas de acesso otimizado a dados em IA. Deverá suportar cargas de trabalho tradicionais e modernas, treinamento e inferência de LLMs, e data lake de IA. Tais características visam assegurar que

o sistema de arquivos possa acompanhar o crescimento da demanda de E/S sem gerar estrangulamentos na infraestrutura.

Esta camada deve ser formada por um Sistema de Arquivos Paralelo Distribuído (PFS), com suporte a acesso concorrente por milhares de processos, baixa latência de metadados, tolerância a falhas e solução com arquitetura validada e certificada para utilização no ambiente computacional proposto.

O PFS deve fornecer um namespace único, de sistema de arquivos compatível com o padrão POSIX, projetado para oferecer capacidade em larga escala e largura de banda de alto desempenho ao sistema de computação. O sistema de arquivos deve suportar conectividade fora do cluster com outros sistemas de computação, clusters de transferência de dados, sistemas de análise, nós de fluxo de trabalho, etc., que podem estar hospedados em outras redes usando diversas tecnologias de rede.

4.7.2.1. Requisitos Gerais da Camada 1

Abaixo estão as características e funcionalidades que o sistema de armazenamento da Camada 1 deve suportar e implementar.

1. Capacidade de armazenamento
 - a. 20 PB utilizáveis, totalmente flash (NVMe).
2. Tipo de mídia de armazenamento
 - a. Dispositivos NVMe, com suporte a funcionalidade de Self-Encrypting Drive (SED) e resistência mínima de 1 DWPD (Drive Writes Per Day - Gravações Completas)

do Drive Por Dia) - significa que um SSD é projetado para ter sua capacidade total gravada e regravada diariamente durante o período de garantia.

3. Arquitetura

- a. Suportar um PFS escalável (scale-out) e compatível com POSIX, com um único namespace global para diversos dispositivos, nós e sistemas de armazenamento.
- b. Deve ser construído utilizando o modelo de Scalable Storage Units (SSUs - Unidade de Armazenamento Escalável - definido como um bloco de construção de um sistema de armazenamento que representa o domínio mínimo de falhas para um destino de armazenamento em rede) e Scalable Storage Clusters (SSCs - Clusters de Armazenamento Escalável - definidos como o menor agrupamento

de SSUs) e infraestrutura de suporte necessários para operar como um namespace de sistema de arquivos POSIX independente.

- c. Deve suportar configurações de expansão vertical (scale-up) e horizontal (scale-out).

4. Conectividade da Rede

- a. Deve possuir conexão compatível e de mesma velocidade dos nós computacionais, sem representar um gargalo na comunicação.
- b. Deve possuir capacidade de comunicação otimizada entre aceleradores, nós computacionais e subsistemas de E/S (ex: RDMA sobre Ethernet (RoCE), RDMA, ou similar).
- c. Caso a solução exija uma rede de armazenamento dedicada entre nós computacionais e o sistema de armazenamento, detalhes completos dessa rede devem ser fornecidos.

5. Proteção e resiliência

- a. O sistema de armazenamento não deve ter um único ponto de falha (single point of failure).
- b. Todos os componentes dos SSUs/SSCs devem possuir redundâncias, incluindo, mas não limitados a: Controladoras, fontes de alimentação, ventoinhas de refrigeração, portas de rede, SSDs NVMe e caminhos redundantes para os

gabinetes/bandejas que hospedam a mídia nos nós/controladores de armazenamento.

- c. Capacidade de proteger dados em cache contra falhas de energia por meio de cache dedicado, com bateria de reserva e espelhado, distribuído entre os SSUs.
- d. Possuir nível de RAID com paridade dupla ou Erasure Coding (EC) para proteção de dados contra falhas de mídia. Esse esquema de proteção de dados não deve ter sobrecarga superior a 20%.
- e. Possuir capacidade de realizar failover e failback automáticos de operações de E/S de aplicações em execução, em caso de falha de um nó/controlador de armazenamento. Esse processo de failover/failback deve ser completamente

transparente para as aplicações em execução nos nós computacionais, sem sofrerem timeouts de E/S.

- f. Capacidade de tolerar falhas de cliente, de servidor ou de rede sem impactar o acesso aos dados.

6. Funcionalidades do Sistema de Arquivos e Sistema de administração/gerenciamento do Sistema de armazenamento

- a. Autenticação e autorização via LDAP, Active Directory, NIS ou métodos de autenticação local.
- b. Suporte a quotas de usuário, grupo e projeto.
- c. Capacidade de realizar atualizações e upgrades online.
- d. Capacidade de compactar dados usando diferentes algoritmos de compressão.
- e. Capacidade de exportar os mesmos dados via cliente nativo de sistema de arquivos paralelo, NFS, SMB e protocolos S3 sem a necessidade de criar uma cópia separada dos dados.
- f. Deve suportar uma interface de armazenamento de objetos.

7. Portal gráfico (web) de Gerenciamento de Integração/Configuração de Telemetria, com

- a. Rastreamento da atividade do usuário, como registro de auditoria e contabilização de processos.
- b. Visualização de dados de desempenho
- c. Interface de usuário para gerenciamento de alertas
- d. Sistema de notificação de eventos de integridade
- e. Acesso a dados históricos de desempenho e do sistema

f. Estatísticas e resumos de execução de tarefas

8. Outros

- a. A proponente deverá incluir todos os componentes necessários para instalação e funcionamento, incluindo, mas não limitados a: switches de rede; cabos e transceptores; transceptores para conexão com as redes do datacenter; racks; licenças para habilitar todas as funcionalidades para uso simultâneo e em toda capacidade contratada para o sistema.

4.7.2.2. Metas de Desempenho da Camada 1

As taxas de desempenho a serem avaliadas, para a Camada 1 de Armazenamento, estão descritas abaixo:

1. Throughput

Deverá sustentar throughput agregado mínimo de 0,5 GB/s (leitura e escrita) por GPU ativa. Esse patamar é necessário para manter o pipeline de treinamento continuamente alimentado, evitando que as unidades de processamento fiquem em espera por dados. Para mensurar o desempenho, o proponente deverá utilizar um workload sequencial baseado em file-per-process.

2. IOPS

Deverá suportar, no mínimo, 15 milhões de IOPS para operações de Leitura e Escrita aleatórias, usando requisições pequenas (entre 4KB e 32KB). Os acessos poderão ser distribuídos e não coordenados, podendo se sobrepor.

3. Operações de arquivos pequenos

Deverá demonstrar pelo menos 40.000 transações de criação de arquivos de 32 KiB por segundo, no total, a partir de múltiplos nós de computação, de forma sustentada por pelo menos 20 segundos. Uma transação de criação de arquivo

consiste nas seguintes operações de metadados: abrir (criar), verificar informações (stat), escrever, ler e fechar. Cada nó computacional criará arquivos em seu próprio diretório.

4. Metadados

Deve suportar as seguintes metas de desempenho:

- a. Mínimo de 2 milhões de criações de arquivos por segundo
- b. Mínimo de 5 milhões de operações de stat de arquivos por segundo
- c. Mínimo de 3 milhões de remoções de arquivos por segundo

4.7.3. Camada 2

A Camada 2 será destinada para armazenamento frio ou arquivamento e ingestão de dados no Supercomputador, com menor custo por TB, para dados históricos ou menos acessados, com área útil mínima de 30 PB (petabytes).

Esta camada deverá atender aos requisitos da seção a seguir.

4.7.3.1. Requisitos Gerais da Camada 2

Abaixo estão as características e funcionalidades que o Sistema de Armazenamento da Camada 2 deve suportar e implementar.

1. Capacidade de armazenamento
 - a. 30 PB utilizáveis.
2. Tipo de mídia de armazenamento
 - a. Serão aceitos dispositivos SSD ou HDD SAS, com suporte à funcionalidade de Self-Encrypting Drive (SED).

3. Arquitetura

- a. Deve ser construído utilizando o modelo de Scalable Storage Units (SSUs - Unidade de Armazenamento Escalável - definidas como um bloco de construção de um sistema de armazenamento que representa o domínio mínimo de falhas para um destino de armazenamento em rede) e Scalable Storage Clusters (SSCs - Clusters de Armazenamento Escalável - definidos como o menor agrupamento de SSUs), bem como a infraestrutura de suporte necessária para operar como um namespace de sistema de arquivos.
- b. Deve suportar configurações de expansão vertical (scale-up) e horizontal (scale-out).
- c. Deve fornecer múltiplos acessos ao Sistema de Arquivos.
- d. Deve fornecer um único namespace.
- e. Deve suportar a conexão de múltiplos clientes simultaneamente.
- f. Deve fornecer múltiplas vias de acesso aos dados (ao namespace).

4. Conectividade da Rede

- a. Deve possuir conexão compatível com os nós computacionais, podendo apresentar uma taxa de transferência de até 25% da largura de banda utilizada na malha de comunicação entre os nós computacionais e a Camada 1 de Armazenamento.
- b. Para atingir a taxa de transferência necessária, poderá utilizar múltiplas conexões ou portas de comunicação.
- c. Deverá possuir capacidade de comunicação com todos os nós computacionais.
- d. Deverá se conectar com todos os nós (computacionais, de login, serviço e administração) em uma única solução através de rede dedicada, sem causar

impactos na comunicação entre os nós computacionais e a Camada 1 de Armazenamento.

5. Proteção e resiliência

- a. O sistema de armazenamento não deve ter um único ponto de falha (single point of failure).
- b. Todos os componentes dos SSUs/SSCs devem possuir redundâncias, incluindo, mas não limitado a: Controladoras, fontes de alimentação, ventoinhas de refrigeração, portas de rede, drives de armazenamento e caminhos redundantes para os gabinetes/bandejas que hospedam a mídia nos nós/controladores de armazenamento.
- c. Capacidade de proteger dados em cache contra falhas de energia.
- d. Possuir nível de RAID com paridade dupla ou Erasure Coding (EC) para proteção de dados contra falhas de mídia. Esse esquema de proteção de dados não deve ter sobrecarga superior a 20%.
- e. Possuir capacidade de realizar failover e failback automáticos de operações de E/S de aplicações em execução, em caso de falha de um nó/controlador de armazenamento. Esse processo de failover/failback deve ser completamente

transparente para as aplicações em execução nos nós computacionais, sem sofrerem timeouts de E/S.

- f. Capacidade de tolerar falhas de cliente, de servidor ou de rede sem impactar o acesso aos dados.
- g. Deve possuir capacidade de troca de drives sem interrupção das operações de E/S.

6. Funcionalidades do Sistema de Arquivos e Sistema de Administração/Gerenciamento do Sistema de Armazenamento

- a. Autenticação e autorização via LDAP, Active Directory, NIS e métodos de autenticação local.
- b. Suporte a quotas de usuário, de grupo e de projeto.
- c. Capacidade de realizar atualizações e upgrades online.
- d. Capacidade de compactar dados usando diferentes algoritmos de compressão.
- e. Suportar múltiplos protocolos (NFS, SMB e protocolos S3) sem a necessidade de criar uma cópia separada dos dados.
- f. Deve suportar uma interface de armazenamento de objetos.
- g. Replicação remota, síncrona e assíncrona

7. Portal gráfico (web) de Gerenciamento de Integração/Configuração de Telemetria, com:

- a. Rastreamento da atividade do usuário, como registro de auditoria e contabilização de processos.
- b. Visualização de dados de desempenho
- c. Interface de usuário para gerenciamento de alertas

- d. Sistema de notificação de eventos de integridade
- e. Acesso a dados históricos de desempenho e do sistema
- f. Estatísticas e resumos de execução de tarefas

8. Outros

- a. A proponente deverá incluir todos os componentes necessários para instalação e funcionamento, incluindo, mas não se limitando à: switches de rede; cabos e transceptores; transceptores para conexão com as redes do datacenter; racks; licenças para habilitar todas as funcionalidades para uso simultâneo e em toda a capacidade contratada do sistema.

4.7.3.2. Metas de Desempenho da Camada 2

Utilizando o protocolo NFS, cada SSC que compor o Sistema de Armazenamento da Camada 2 deverá apresentar:

- No mínimo, 100.000 IOPS para operações aleatórias (4KB) de leitura e escrita
- No mínimo, throughput de 5 GB/s, utilizando tamanhos de blocos de até 4MB

Para o protocolo S3 e utilizando a ferramenta de benchmark MinIO/WARP, a solução deverá apresentar, no mínimo, os valores abaixo, utilizando tamanhos de objetos de 8MB e até 100 operações concorrentes :

- Throughput para operações de leitura (GET): 500 MB/s
- Throughput para operações de escrita (PUT): 250 MB/s

4.7.4. Ambiente de Ingestão dos Dados

Deverá ser fornecido um ambiente de alta disponibilidade para ingestão de dados (upload e download) via Internet, com implementação de métodos eficientes de transferência de

arquivos. Este ambiente servirá como porta de entrada para os sistemas de arquivos de toda a instalação, permitindo o acesso aos dados dos usuários.

A solução de ingestão de dados deverá contemplar:

1. No mínimo 3 nós dedicados e otimizados para transferência dos dados (Data Transfer Nodes - DTNs)
 - a. Deverão ser configurados para alta disponibilidade, para manter o serviço sempre disponível.
 - b. Deverão se conectar diretamente à Camada 2 de armazenamento, utilizando uma interface ou canal de conexão com a mesma velocidade que os nós computacionais utilizam para se conectar a essa camada.
 - c. Deverão ser configurados de acordo com as melhores práticas de hardening, com apenas os softwares necessários à transferência de dados e à manutenção do sistema.
2. Deverá disponibilizar ferramentas eficientes para transferência de dados:
 - a. UFTP (UNICORE FTP)
 - b. GridFTP
 - c. Secure Copy (SCP) e Rsync
 - d. AWS S3
 - e. Resilio

4.7.5. Integração entre as Camadas 1 e 2 de Armazenamento

Caso as camadas 1 e 2 de armazenamento sejam fornecidas por soluções distintas (separadas) de Sistemas de Armazenamento, deverá ser fornecida, em conjunto, uma ferramenta eficiente de transferência paralela de dados entre as Camadas 1 e 2.

As camadas 1 e 2 de armazenamento poderão ser implementadas através de um único Sistema de Armazenamento integrado, desde que atenda aos requisitos técnicos dispostos nas seções acima e:

1. Disponibilize ambas as camadas como um único namespace.
2. Implemente um mecanismo de Gerenciamento Hierárquico de Armazenamento (HSM, na sigla em inglês - Hierarchical Storage Management) para mover

automaticamente os dados entre as mídias de armazenamento das camadas 1 e 2, com base em políticas de uso.

3. Permita acesso segregado e distinto às diferentes camadas de armazenamento. Por exemplo, um cliente poderá acessar diretamente o ambiente de armazenamento da camada 2, sem ser necessário passar, obrigatoriamente, pela camada 1.
4. Deverá permitir que os dados inseridos através do Ambiente de Ingestão de Dados acessem diretamente a área de armazenamento da Camada 2, sem interferir no desempenho da Camada 1 de armazenamento.

4.8. Nós de Serviço e Gerenciamento

O proponente deverá dimensionar a quantidade de nós de gerenciamento e serviço visando garantir o adequado funcionamento de toda solução fornecida. Deverá haver um quantitativo mínimo de 2 nós de gerenciamento e serviço.

Os serviços críticos para o funcionamento do supercomputador deverão ser configurados em alta disponibilidade no formato ativo-ativo ou ativo-passivo.

A configuração mínima desses equipamentos está descrita no item 4.

4.9. Segurança de Dados, Telemetria e Interoperabilidade de Hardware

Proibição de telemetria não autorizada: é vedado ao fornecedor, por qualquer meio técnico, transmitir dados para servidores externos ao ambiente do comprador sem autorização prévia, explícita e documentada. Dados vedados incluem: dados de workloads e resultados; logs de sistema; metadados de usuários e projetos; e qualquer dado que possibilite a reconstrução das cargas de trabalho processadas.

Conformidade com LGPD e GSI/PR: toda a operação do sistema deverá estar em conformidade com a LGPD (Lei 13.709/2018) e as Normas de Segurança da Informação do

GSI/PR. O fornecedor deverá apresentar declaração formal de conformidade renovada anualmente.

Controle total de atualizações remotas: nenhuma atualização de firmware, software ou configuração poderá ser aplicada remotamente sem autorização explícita do comprador. O fornecedor deverá descrever o mecanismo de controle, incluindo ambiente de homologação isolado, capacidade de rollback e prazo máximo de 48 horas para patches críticos de CVE.

4.10. Serviços

Todos os serviços listados abaixo deverão ser realizados com supervisão da contratante, com designação formal de fiscal técnico do contrato pelo comprador. O fornecedor deverá entregar relatório mensal de execução dos serviços, incluindo indicadores de disponibilidade, incidentes registrados e ações corretivas adotadas. Reuniões de acompanhamento técnico deverão ocorrer com frequência mínima trimestral, ou a qualquer momento por solicitação do comprador.

O fornecedor deverá designar no mínimo 1 engenheiro residente no Brasil, com dedicação exclusiva, em regime de 8x5, ao comprador, trabalhando presencialmente junto a técnicos brasileiros desde o primeiro mês de operação do sistema e durante todo o período de garantia para garantir a operação e manutenção de hardware, além disso o fornecedor deverá garantir acesso a engenheiros nas seguintes especialidades: (i) administração de sistemas e software

stack e (ii) suporte ao desenvolvimento de modelos de IA na pilha fornecida. O Plano de Trabalho poderá prever acesso à engenheiros de outras especialidades.

4.10.1. Serviços de Software

Relação de serviços de software que deverão ser contemplados na proposta:

- Instalação e configuração do sistema operacional homologado pelo fornecedor com todos os drivers necessários para o pleno funcionamento dos sistemas e aplicações fornecidos;
- Hardening de segurança do sistema operacional, sistemas e aplicações fornecidas;
- Correção de vulnerabilidades conhecidas do sistema operacional, sistemas e aplicações fornecidas;
- Instalação, configuração e execução de benchmarking para avaliação de desempenho;
- Realizar otimização dos sistemas e aplicações visando garantir o melhor desempenho dos sistemas fornecidos;
- Autenticação integrada com os serviços de autenticação do cliente;
- Todos os softwares previstos neste documento, item 5.6, e suas respectivas dependências deverão ser entregues instalados, configurados e em pleno funcionamento durante todo o período de garantia;
- Serviços de reinstalação e atualização periódica de todos os softwares que compõem a solução.

4.10.2. Serviço de Suporte e Monitoramento

Relação de serviços de suporte que deverão ser contemplados na proposta:

- Monitoramento em tempo real com geração automática de alertas 24x7, via sistema web com configuração de gatilhos, limiares e notificações;
- Assistência técnica on-site com reposição de peças novas por todo prazo de garantia de todo o conjunto de hardware ofertado na modalidade Next Business Day (NBD);
- Instalação física e cabeamento de todos os servidores, switches, storages, racks e PDUs com cabeamento estruturado, etiquetado e documentado conforme layout do datacenter;
- Provisionamento do sistema operacional, drivers, firmware, middleware IA e qualquer outro componente de software necessário para o pleno funcionamento da solução;
- Configuração e testes das malhas de rede conforme topologia e tecnologia ofertados;
- Implementação e manutenção das stacks de software de IA solicitadas;
- Realização de testes de performance dos recursos computacionais, sistemas de armazenamento e rede em intervalos periódicos ou sob demanda;
- Garantia de disponibilidade de 97% dos recursos computacionais durante todo o período de garantia;
- Coleta contínua de dados de temperatura e energia com geração de relatórios periódicos visando identificar possíveis anomalias e oportunidades de melhoria;
- Resposta a incidentes com registro, classificação e documentação de causa raiz para todos os eventos URGENTES e CRÍTICOS;
- Parceiro nacional de serviços: o fornecedor deverá possuir ou estabelecer, até a data de instalação, parceiro domiciliado no Brasil responsável pela execução dos serviços on-site, montagem e manutenção local.
- Na execução dos serviços, deverão ser observados os seguintes prazos:

- **Requisições de suporte** poderão ser classificadas como:

- ▶ **Crítica:** atividades relacionadas às correções de vulnerabilidades, implementações de segurança, correção de configuração, testes de desempenho dos sistemas de arquivos e dos nós computacionais e qualquer outra atividade que exija de resposta imediata ou que envolva questões de segurança da informação.
 - Início de atendimento: até 4 horas
 - Finalização do atendimento: até 24 horas
- ▶ **Não Crítica:** atividades relacionadas à instalação e atualização de software, inclusão de novas funcionalidades, geração de relatórios, esclarecimento de dúvidas e demais requisições.
 - Início de atendimento: até 24 horas
 - Finalização do atendimento: até 96 horas
- **Incidentes de suporte** poderão ser classificados como:
 - ▶ **Urgente:** Erros relacionados ao funcionamento de elementos críticos do supercomputador como: gerenciador de filas, sistema de alta

disponibilidade, serviço de autenticação, sistema de arquivos, ferramentas de orquestração etc.

- Início de atendimento: até 2 horas
- Finalização do atendimento: até 8 horas
- ▶ **Crítica:** problemas relacionados ao funcionamento de bibliotecas ou softwares de uso coletivo fornecidos pelo fabricante.
 - Início de atendimento: até 8 horas
 - Finalização do atendimento: até 48 horas
- ▶ **Não Crítica:** problemas relacionados aos outros softwares fornecidos pela contratada e que não geram impacto para o uso do supercomputador.
 - Início de atendimento: até 16 horas
 - Finalização do atendimento: até 72 horas
- **Atividades de manutenção:** poderão ser classificadas como:
 - ▶ **Urgente:** problemas de hardware relacionados a equipamentos que resultem em indisponibilidade de mais de 50% do conjunto de recursos computacionais ou que tornem o uso do supercomputador inviável.
 - Início de atendimento: até 2 horas
 - Finalização do atendimento: até 8 horas
 - ▶ **Crítica:** problemas de hardware relacionados a equipamentos que resultem em indisponibilidade de 11% a 50% do ambiente.
 - Início de atendimento: até 8 horas
 - Finalização do atendimento: até 36 horas

- ▶ **Não Crítica:** problemas de hardware relacionados a equipamentos que resultem em indisponibilidade de até 10% do ambiente.
 - Início de atendimento: até 16 horas
 - Finalização do atendimento: até 72 horas

4.11. Documentação Técnica

As propostas deverão incluir um plano completo de entrega de documentação técnica, contemplando:

- Manuais detalhados de operação e manutenção;
- Diagramas de arquitetura de hardware e interconexões;
- Documentação técnica de firmware, sistemas embarcados e componentes essenciais;
- Licença de uso perpétua, para fins públicos, sobre toda a documentação e softwares entregues.

A proposta deverá demonstrar como a documentação assegurará autonomia tecnológica e capacidade de manutenção nacional.

A proposta deverá apresentar inventário detalhado e completo dos materiais, componentes de hardware, ferramentas, software, sistema de refrigeração e respectivas documentações e

diagramas necessários para a montagem, operação, manutenção e atualização dos itens fornecidos. Para cada item deverá ser apresentada a respectiva versão, quando aplicável.

4.12. Transferência de Tecnologia, Treinamento e Capacitação

4.12.1. Requisito de Capacitação Técnica para Uso e Operação da Máquina

A proposta deverá incluir um programa de capacitação técnica com carga horária mínima de 320 horas, destinado a 40 engenheiros ou profissionais indicados pelo comprador.

A primeira turma deverá ser iniciada em até 180 dias após o aceite da proposta e concluída até a data da conclusão do processo de comissionamento do equipamento.

O comprador poderá formar até 4 turmas para capacitação durante todo o período de vigência do contrato com frequência máxima de 1 turma por ano.

O programa deverá contemplar, no mínimo:

- Treinamento on-the-job nas instalações do fabricante;
- 50% a 70% da carga horária deverá contemplar atividades práticas na operação e manutenção de tecnologias fornecidas;
- Módulos teóricos sobre competências necessárias para operar, manter, evoluir e otimizar a infraestrutura de computação ofertada;
- O programa deverá comprovar que a equipe técnica terá adquirido conhecimento todas as etapas de implementação, operação e suporte nas tecnologias fornecidas;

Portabilidade do conteúdo: ao menos 50% da carga horária total - independentemente de ser teórica ou prática - deverá ser dedicada a habilidades portáteis e independentes do hardware do fornecedor, incluindo: PyTorch, Kubernetes, Linux HPC, fundamentos de IA/ML, engenharia de dados e administração de clusters. Certificações exclusivamente proprietárias do fornecedor não poderão constituir mais de 50% da carga horária total do programa.

Todas as despesas de transporte, hospedagem e alimentação dos participantes desta capacitação deverão ser custeadas pelo fabricante.

4.12.2. Mentoria Especializada

A proposta deverá incluir 1200 horas por ano de mentoria com pesquisadores e engenheiros seniores da equipe de P&D do fornecedor, dedicadas exclusivamente a temas de pesquisa aplicada e desenvolvimento da pilha nacional de software de IA. Esta mentoria é distinta do suporte técnico operacional previsto no item 4.10 e da capacitação de operação prevista no item 4.12 - seu foco é a fronteira do conhecimento, não a operação cotidiana da infraestrutura.

Os temas de mentoria deverão ser acordados semestralmente entre o comprador e o fornecedor, priorizando:

- Arquitetura, treinamento e otimização de modelos fundacionais (LLMs, VLMs e modelos multimodais), incluindo decisões de design, tokenização, técnicas de pré-treinamento, alinhamento, fine-tuning, RLHF, DPO e avaliação de capacidades;
- Desenvolvimento, evolução e otimização da pilha nacional de software de IA, abrangendo compiladores, runtimes, bibliotecas de treinamento distribuído,

frameworks de inferência, ferramentas de observabilidade e sistemas de gerenciamento de recursos computacionais;

- Treinamento distribuído em larga escala, incluindo paralelismo de dados, tensor, pipeline e sequência, estratégias híbridas de paralelização, balanceamento de carga, tolerância a falhas e otimização de comunicação entre aceleradores;
- Engenharia de inferência em larga escala, incluindo serving distribuído, otimização de throughput e latência, gerenciamento de memória, cache de KV, batching contínuo e atendimento simultâneo a múltiplos modelos;
- Arquiteturas de Retrieval-Augmented Generation (RAG), Graph RAG e sistemas híbridos de recuperação de conhecimento para aplicações governamentais e científicas;
- Desenvolvimento de sistemas de IA Agêntica (Agentic AI), incluindo planejamento, orquestração de ferramentas, uso de agentes cooperativos, memória de longo prazo, execução de fluxos complexos e avaliação de agentes;
- Curadoria, qualidade, governança, anonimização e ciclo de vida de datasets para treinamento de modelos fundacionais em língua portuguesa, incluindo aspectos legais, éticos e de soberania digital;
- Metodologias de avaliação de modelos, benchmarking, medição de desempenho, segurança, robustez, vieses, alucinações e conformidade regulatória;
- Estratégias de eficiência computacional, incluindo quantização, pruning (Poda), destilação, sparsity, mixture-of-experts (MoE), compressão de modelos e otimização energética;
- Machine Learning Operations (MLOps), Large Language Model Operations (LLMOps) e operação contínua de modelos fundacionais, incluindo versionamento,

rastreabilidade, observabilidade, monitoramento de deriva, gestão de experimentos e automação de pipelines de treinamento e implantação;

- Contribuição upstream e liderança técnica em projetos estratégicos de código aberto relacionados à pilha nacional de IA, incluindo frameworks de treinamento, inferência, avaliação, observabilidade e gerenciamento de infraestrutura computacional;
- Co-design hardware-software-algoritmo para sistemas de IA em larga escala, incluindo análise de gargalos, otimização fim-a-fim (desde a estrutura física até a execução dos modelos de I.A) da pilha computacional e estratégias de maximização de utilização dos recursos do supercomputador;
- Arquiteturas de rede, armazenamento e sistemas distribuídos para IA, incluindo interconexões de alta velocidade, sistemas de arquivos paralelos, armazenamento de datasets massivos e otimização de fluxos de dados para treinamento e inferência;
- Segurança de sistemas de IA, incluindo segurança da cadeia de suprimentos de modelos, proteção contra ataques adversariais, segurança de inferência, proteção de dados sensíveis e governança de modelos fundacionais;
- Transferência tecnológica visando a formação de competências nacionais para desenvolvimento autônomo de modelos fundacionais, ferramentas de treinamento distribuído, sistemas de inferência e componentes da pilha brasileira de IA.

Os mentores designados pelo fornecedor deverão ter publicações ou contribuições verificáveis nas áreas objeto da mentoria. Não serão aceitos como mentores profissionais de suporte técnico, vendas ou gerenciamento de contas.

Cada sessão deverá ser registrada em ata com: data, mentor designado, participantes brasileiros, tema, duração e principais encaminhamentos técnicos. O fornecedor entregará relatório semestral comprovando o cumprimento da carga horária e a qualificação dos

mentores. O não cumprimento das horas contratadas ou a designação de mentores sem qualificação comprovada sujeitará o fornecedor às penalidades previstas.

4.12.3. Early Access

Em até 90 dias após a assinatura do contrato de fornecimento, o fornecedor deverá disponibilizar acesso a uma infraestrutura computacional de tecnologia equivalente à solução ofertada com, pelo menos, 5% da capacidade total ofertada na proposta. Essa infraestrutura

deverá estar disponível 24 horas por dia, 7 dias por semana, até a conclusão da entrega e comissionamento da solução.

Essa infraestrutura será utilizada para treinamento e desenvolvimento de modelos de LLM estratégicos, visando otimizar o uso da máquina a partir do primeiro dia de funcionamento.

Deverá ser fornecido suporte ao uso dessa infraestrutura atendendo os mesmos requisitos da proposta.

4.13. Infraestrutura

Todo o sistema deverá ser entregue e instalado no Parque Científico e Tecnológico Augusto Severo, localizado na Av. Santos Dumont, 1560 - Área Rural - Macaíba/RN - Brasil.

A proposta deverá apresentar as seguintes informações relativas ao plano de infraestrutura do supercomputador:

- Quais as necessidades elétricas, tensões e correntes necessárias para alimentação do sistema.
- Qual a temperatura, volume e pressão da "água de temperatura ambiente" a ser fornecida para resfriamento de todo o sistema.
- Qual a temperatura, volume e pressão da água gelada a ser fornecida para resfriamento de todo o sistema.
- Qual a carga de cada rack a ser instalado.
- Apresentar projeto arquitetônico.
- Apresentar projeto elétrico.
- Apresentar projeto hidráulico.

- Eficiência Energética: será desejável Power Usage Effectiveness - PUE igual ou inferior a 1,3 em regime operacional nominal, incluindo os sistemas de refrigeração, distribuição elétrica e demais componentes de infraestrutura associados ao ambiente computacional. O fornecedor deverá detalhar a metodologia de medição, categorias de medição aplicáveis e condições operacionais utilizadas no cálculo, em

conformidade com os critérios definidos pela norma ISO/IEC 30134-2:2026.

O PUE é definido pela razão abaixo:

$$\text{PUE} = \frac{\text{Energia Total da Instalação}}{\text{Energia Utilizada pelos Equipamentos de T.I.}}$$

Outras informações poderão ser solicitadas.

4.14. Roadmap de Atualização Tecnológica

As propostas deverão apresentar um plano de atualização tecnológica para os próximos 3 anos, incluindo:

1. Roadmap completo para atualização de hardware, firmware e software, com versões previstas e datas estimadas;
2. Estratégias de substituição escalável e compatibilidade reversa;
3. Lista de componentes críticos e respectivas versões previstas para atualização, com indicação do impacto esperado em desempenho e consumo energético;
4. Faixas máximas de preços ou índices de reajuste para atualizações futuras;
5. Estratégias para garantir compatibilidade com modelos de IA de próxima geração, incluindo LLMs de grande porte com arquiteturas de mistura de especialistas (MoE) e modelos multimodais.

4.15. Colaboração Estratégica

Os proponentes deverão apresentar suas propostas contendo um plano de colaboração estratégica, contemplando os seguintes eixos:

- Parcerias com Institutos de Pesquisa e Centros de P&D

- Estratégias para cooperação tecnológica, desenvolvimento conjunto, transferência de conhecimento e capacitação.
- Conteúdo Local e Fornecedores Domésticos:
 - Plano de utilização e desenvolvimento da cadeia produtiva nacional, incluindo subcontratações, capacitação de fornecedores e estímulo à economia local.

Esses são itens obrigatórios. **A não apresentação resultará na desclassificação da proposta.**

4.15.1. Requisitos para Parcerias com Institutos de Pesquisa e Centros de P&D

A proponente deverá apresentar um Programa de Parceria com Institutos de Pesquisa e Centros de P&D descrevendo as estratégias de cooperação tecnológica, desenvolvimento conjunto, transferência de conhecimento e capacitação que se compromete a executar em território brasileiro, em complemento ao fornecimento da solução de HPC objeto desta licitação.

O programa deverá contemplar, de forma estruturada, os seguintes elementos:

Programa de Formação de Recursos Humanos

Apresentação de um Programa de Formação de Recursos Humanos com foco em inteligência artificial, com capacidade mínima de formação de mil (1000) especialistas brasileiros em três anos, em competências estratégicas para o ecossistema nacional de IA. O programa deverá cobrir, entre outros temas: modelos de linguagem de grande porte e suas arquiteturas; engenharia de modelos e pipelines de treinamento e inferência; desenvolvimento de produtos digitais baseados em IA; aplicações de IA para serviços públicos; e pesquisa e inovação aplicada. Os temas poderão ser adaptados ou alterados conforme a necessidade do contratante. Ao contrário da capacitação

operacional do item 4.12.1, este programa deverá ser destinado a formar multiplicadores - profissionais capazes de replicar conhecimento em suas instituições. A carga horária mínima por participante deverá ser de 40 horas, com ao menos 50% em atividades práticas.

Identificação e descrição dos Institutos de Pesquisa e Centros de P&D parceiros envolvidos na execução do programa, incluindo suas áreas de especialização, infraestruturas e laboratórios disponíveis, programas de formação acadêmica e profissional existentes, e equipes de pesquisa, engenharia e produto mobilizadas. A proposta deverá apresentar, pelo menos, 5 institutos de pesquisa ou centros de P&D, sendo desejável um em cada região do Brasil (norte, sul, sudeste, nordeste e centro-oeste).

Apoio ao Desenvolvimento da pilha de Software de IA Nacional

Estratégia para contribuir com o desenvolvimento de uma pilha de software de IA nacional, apoiando Frentes de Pesquisa (FP) de instituições nacionais e internacionais e fornecendo o suporte necessário para a realização de testes e experimentos.

Abaixo estão relacionadas, mas não limitadas, as frentes de pesquisa de interesse do PBIA:

- FP 1: Plataforma para Construção de Modelos Fundacionais
- FP 2: Plataforma de Gestão e Serviços para Modelos de IA
- FP 3: Ambiente de Alto Desempenho para Modelos de IA Distribuídos
- FP 4: Infraestrutura de Otimização Numérica e Compilação
- FP 5: Infraestrutura Nacional de Gerência de Clusters para HPC e IA

- FP 6: Infraestrutura Nacional de Dados para IA e HPC
- FP 7: Ecossistema Federado de IA/HPC no Brasil
- FP 8: Monitoramento e Otimização de Desempenho
- FP 9: Sistema Operacional e Virtualização com Segurança e Privacidade

Estratégia de cooperação e desenvolvimento de arquiteturas abertas baseadas em RISC-V, com intercâmbio de projetistas brasileiros em instituições de ponta do fornecedor, cobrindo: quais algoritmos incorporar no chip; dado os algoritmos, quais instruções são necessárias; e desenvolvimento de chiplets e controladores abertos.

Cronograma detalhado de execução do programa, com metas intermediárias verificáveis, indicadores de desempenho, responsáveis por cada etapa e mecanismos de reporte periódico ao contratante. A proposta deverá detalhar ainda os mecanismos de seleção dos participantes brasileiros, critérios de acompanhamento e certificação, e estratégias de retenção dos talentos formados no ecossistema nacional, em alinhamento com as ações do PBIA que tratam da inserção de pesquisadores em empresas e da retenção de talentos em IA no Brasil.

A não comprovação do cumprimento das metas declaradas neste programa durante a execução contratual sujeitará a contratada às penalidades previstas no instrumento contratual.

5. EXECUÇÃO DO OBJETO

5.1. Testes de Aceitação

A empresa participante do certame deverá apresentar um plano de testes dos componentes fornecidos que será executado após a entrega e instalação da solução. O plano deverá conter, minimamente, os testes descritos nas subseções abaixo. O processo de aceitação compreende

duas etapas formais: Aceite Provisório e Aceite Definitivo, com condições e consequências distintas descritas no item 5.1.3 (Aceite).

5.1.1. Sistemas de Armazenamento

Os parâmetros de desempenho a serem avaliados estão descritos no item 4.7. Para a avaliação do desempenho do Sistema de Armazenamento, a contratada poderá utilizar as seguintes ferramentas de benchmark, de forma combinada ou individualmente:

- IOR (Interleaved Or Random)
- Iozone
- FIO (Flexible I/O Tester)
- MDTest
- MinIO/WARP

Será considerado aceite somente se os testes forem reproduzidos on-site, após a instalação total da solução.

Caso os testes não atinjam o valor mínimo exigido, o fornecedor deverá providenciar as ações necessárias até que esse valor seja alcançado.

Deverá ser entregue um relatório com todos os resultados obtidos.

5.1.1.1. Execução de Benchmark de IA para a Camada 1

Deverá apresentar o resultado da execução do benchmark MLPerf Storage (<https://mlcommons.org/benchmarks/storage/>), com a seguinte configuração:

- Categoria de workload checkpointing
- Tamanho de modelo

- Llama 3.1-70B [512 processos/aceleradores simulados] ou
- Modelo de submissão closed
- Não deverá utilizar mais do que 8 processos por nó computacional para simular os aceleradores
- Seguir as seguintes etapas:
 - Escrever 10 checkpoints
 - Limpar/Liberar Caches utilizados
 - Ler 10 checkpoints

5.1.1.2. Execução de Benchmark IO500 para a Camada 1

Deverá ser executado o benchmark IO500 (<https://io500.org/>), no modo “Production”.

A fim de comprovar escalabilidade real e evitar “ilhas de desempenho” não representativas do cluster entregue, o benchmark deverá ser executado em dois tamanhos:

- Utilizando 50% dos nós computacionais
- Utilizando 100% dos nós computacionais

Todos os testes executados pelo benchmark devem apresentar resultados válidos.

5.1.2. Testes Funcionais

A proposta deverá apresentar um plano de testes detalhado visando a verificação da conformidade dos sistemas instalados. Esse plano poderá ser revisado junto ao contratante para que seja adequado às necessidades.

O proponente deverá apresentar, executar e evidenciar um Plano de Testes Funcionais (PTF) que comprove a conformidade integral dos sistemas, serviços de redes e softwares stack instalados. O PTF será submetido à aprovação prévia do Contratante e poderá ser ajustado de comum acordo antes do início dos testes. Uma vez aprovado, torna-se base vinculante para aceitação e faturamento. Como sugestão, o PTF pode ser dividido em 11 etapas segundo Reddi,V.J et al. 2020 e Tschand, A., Rajan, A. T. R., Idgunji, S., et al. (2024):

(i) Abrangência e Pré-condições; (ii) Testes de HA e Failover (gerenciamento e monitoramento), (iii) Funcionalidades de Software de gerenciamento, (iv) Teste de Redes e interconexões internas, (v) Interfaces de administração, (vi) Consumo de Energia (Sistema e nó), (vii) Integralidade de Hardware e diagnóstico automatizado, (viii) Teste de softwares stack

de IA (Treino, inferência e Orquestração), (ix) Evidências, relatórios e repetibilidade, (x) Não conformidade e correções, (xi) Janela e segurança operacional.

5.1.3. Aceite

O processo de aceitação compreende duas etapas formais sequenciais:

Aceite Provisório: emitido após a conclusão bem-sucedida de todos os testes de aceitação previstos nos itens 5.1, incluindo a reexecução dos benchmarks MLPerf conforme item 4.5. A emissão do Aceite Provisório está condicionada adicionalmente à entrega de:

- Inventário completo de equipamentos, verificado contra os itens propostos;
- Documentação técnica conforme plano de entrega do item 4.11;
- Plantas, diagramas e projetos arquitetônicos solicitados no item 4.13;
- Relatório fotográfico dos equipamentos instalados;
- Relatório completo da execução dos benchmarks MLPerf, IO500 e MLPerf Storage;
- Relatório dos testes funcionais com evidências do PTF.

A emissão do Aceite Provisório marca o início do período de garantia.

Aceite Definitivo: emitido 30 dias após o Aceite Provisório, desde que não tenham sido identificadas não conformidades relevantes durante o período de operação assistida. A

emissão do Aceite Definitivo está condicionada adicionalmente à comprovação do início das ações apresentadas na proposta de colaboração estratégica (item 4.15).

5.2. Prazo e Local de Entrega

Todo o sistema deverá ser entregue e instalado no Parque Científico e Tecnológico Augusto Severo, localizado na Av. Santos Dumont, 1560 - Área Rural - Macaíba/RN - Brasil.

O fornecedor deverá informar, de forma clara e objetiva, o prazo estimado para a entrega dos equipamentos objeto desta solicitação, que será contado a partir da formalização contratual. Este prazo **não poderá ultrapassar 520 (quinhentos e vinte) dias**.

Adicionalmente, deverá ser apresentado o prazo previsto para o comissionamento dos equipamentos, compreendendo todas as atividades necessárias à sua configuração, instalação, testes e disponibilização para uso operacional.

O fornecedor deverá apresentar um cronograma detalhado contendo as etapas de entrega, instalação e comissionamento.

6. ESTIMATIVA DE VALOR DA CONTRATAÇÃO E ADEQUAÇÃO ORÇAMENTÁRIA

O valor estimado da contratação corresponde a R\$ 959.040.959,04 (novecentos e cinquenta e nove milhões, quarenta mil, novecentos e cinquenta e nove reais e quatro centavos), a ser pago em 3 (três) parcelas de R\$ 319.680.319,68 (trezentos e dezenove milhões, seiscentos e oitenta mil, trezentos e dezenove reais e sessenta e oito centavos) cada, vinculadas, respectivamente, à assinatura do contrato, ao Aceite Definitivo e ao decurso de 12 (doze) meses após o Aceite Definitivo, conforme detalhado no item 5.1. Essa estimativa tem por base o cronograma físico-financeiro definido pela equipe demandante e deverá constar, com a respectiva memória de cálculo e os documentos de suporte, em documento apartado e

classificado, nos termos do art. 6º, XXIII, alínea “i”, da Lei nº 14.133/2021, subsidiariamente aplicado.

A adequação orçamentária e financeira da contratação, nos termos do art. 6º, XXIII, alínea “j”, da Lei nº 14.133/2021, subsidiariamente aplicado, será certificada por meio do respectivo instrumento de repasse, convênio, termo de outorga ou plano de trabalho celebrado entre a fonte financiadora e a Fundação de Apoio ao Desenvolvimento da Computação Científica — FACC, no âmbito do PBIA 2024–2028, ao qual competirá atestar a existência de recursos orçamentários e financeiros suficientes para fazer face às obrigações assumidas,

previamente à assinatura do contrato, na forma da Lei nº 8.958/1994 e do Decreto nº 8.241/2014.

Fonte do Recurso: FINEP (Financiadora de Estudos e Projetos)

A Nota Fiscal ou Fatura apresentada deverá conter expressamente os elementos necessários e essenciais do documento, tais como:

- A data de Emissão;
- Prazo de Validade;
- Os dados do Contrato e do órgão Contratante
- O período de prestação dos Serviços
- O valor a pagar; e
- Eventual destaque do valor de retenções tributárias cabíveis.

O prazo para o pagamento da(s) Nota(s) Fiscal(is)/Fatura(s), será de até 10 (dez) dias úteis, exceto caso haja algum erro, ou, ausência de algum dado/informação necessária ao pagamento pretendido.

Havendo erro na apresentação da(s) Nota(s) Fiscal(is)/Fatura(s), ou circunstância que impeça a liquidação da despesa, o pagamento ficará sobrestado até que a Contratada providencie as medidas saneadoras. Nesta hipótese, o prazo para pagamento iniciar-se-á após a

comprovação da regularização da situação, não acarretando qualquer ônus para a Contratante.

Será rescindido o contrato em execução com a Contratada inadimplente, salvo por motivo de economicidade, segurança nacional ou outro de interesse público de alta relevância, devidamente justificado, em qualquer caso, pela máxima autoridade da Contratante.

7. DO CONTRATO

A execução do contrato será acompanhada e fiscalizada por fiscal técnico formalmente designado pela Contratante, definida no momento da assinatura do contrato ao qual competirá, entre outras atribuições:

- acompanhar o cumprimento do cronograma de entrega, instalação, comissionamento e aceitação (item 5);
- analisar os relatórios de execução dos serviços, incluindo indicadores de disponibilidade, incidentes registrados e ações corretivas;
- presidir ou participar das reuniões de acompanhamento técnico, com periodicidade mínima trimestral;
- atestar o cumprimento dos marcos de pagamento (item 8);
- e formalizar, quando cabível, a instauração de procedimento de apuração de responsabilidade e aplicação de sanções (item 11).
- A gestão do contrato observará, ainda, o acompanhamento da execução das obrigações de garantia técnica (item 10), de transferência de tecnologia, capacitação e mentoria especializada (item 4.12), e do plano de colaboração estratégica (item

4.15), mediante relatórios periódicos apresentados pela Contratada, na forma e na periodicidade neles especificadas.

A Contratante poderá instituir comissão técnica de acompanhamento, com participação de representantes das áreas de tecnologia da informação, infraestrutura predial, segurança da informação e da própria fundação de apoio, para subsidiar o fiscal técnico do contrato nas decisões de natureza técnica especializada, notadamente quanto à validação dos testes de aceitação (item 5.1) e à avaliação da execução das obrigações de transferência de tecnologia (item 4.12).

8. CRITÉRIOS DE MEDIÇÃO E DE PAGAMENTO

8.1. Condições Gerais

Os pagamentos devidos em razão da execução do contrato serão processados pela Contratante em conformidade com as disponibilidades orçamentárias e financeiras do instrumento de repasse, convênio, termo de outorga ou plano de trabalho que sustenta a presente contratação, observados os princípios da eficiência e da boa gestão de recursos aplicáveis às fundações de apoio nos termos da Lei nº 8.958/1994 e do Decreto nº 8.241/2014, aplicando-se subsidiariamente, quanto aos critérios de medição e liquidação da despesa, o disposto no art. 6º, XXIII, alínea “g”, e nos arts. 140 e 141 da Lei nº 14.133/2021.

Cada parcela de pagamento fica condicionada à comprovação do respectivo marco contratual, à emissão de nota fiscal/fatura correspondente, ao atesto do fiscal técnico do contrato (item

7) e à regularidade fiscal, trabalhista e previdenciária da Contratada, sem prejuízo da retenção de tributos incidentes, nos termos da legislação aplicável.

Nenhum pagamento será realizado sem a comprovação da manutenção, pela Contratada, da garantia de execução contratual de que trata o item 10, cabendo à Contratante sobrestar o pagamento correspondente enquanto não regularizada eventual insuficiência da garantia.

O valor informado acima é estimativo e não indica qualquer compromisso futuro para a Contratante. Somente será realizado o pagamento do(s) serviço(s) efetivamente realizados de acordo com as especificações estabelecidas nesse Termo de Referência.

Por se tratar de investimento proveniente de financiamento público, caso a financiadora descontinue o projeto, por qualquer razão, o contrato, fruto deste Termo de Referência também terá seu pagamento suspenso.

8.2. Cronograma de Pagamento

O pagamento será realizado em três parcelas conforme abaixo:

- Parcela 1: R\$ 319.680.319,68 (trezentos e dezenove milhões, seiscentos e oitenta mil, trezentos e dezenove reais e sessenta e oito centavos), na assinatura do contrato.
- Parcela 2: R\$ 319.680.319,68 (trezentos e dezenove milhões, seiscentos e oitenta mil, trezentos e dezenove reais e sessenta e oito centavos), após aceite definitivo.
- Parcela 3: R\$ 319.680.319,68 (trezentos e dezenove milhões, seiscentos e oitenta mil, trezentos e dezenove reais e sessenta e oito centavos), 12 meses após o aceite definitivo.

9. FORMA E CRITÉRIOS DE SELEÇÃO DO FORNECEDOR

9.1. Da Proposta

O fornecedor deve submeter suas informações e documentação em conformidade com os seguintes prazos e formatos:

- Data Limite para Resposta: 30 (trinta) dias após a publicação do Termo de Referência

Formato de Submissão: A proposta, em formato impresso e digital (PDF), deverá ser apresentada, pessoalmente, pelos representantes legais das proponentes, em local, dia e horário que será determinado em edital, observando os prazos já definidos neste Termo de Referência. O documento deve ser estruturado em seções claras, seguindo a ordem das informações solicitadas, possuindo uma lista identificando todos os anexos. Para o formato digital, os anexos deverão ser agrupados em um único arquivo, em formato zip, e entregues junto com a proposta impressa.

Detalhes do Envio: As propostas, em formato impresso e digital (PDF), deverão ser apresentadas, pessoalmente, pelos representantes legais das proponentes, em local, dia e horário que será determinado em edital, observando os prazos já definidos neste Termo de Referência. A proposta deve conter o nome da empresa proponente e a referência ao Termo de Referência, seguindo o padrão, para o formato digital: TR-FabricaIA-NomeDaEmpresa.pdf. Os anexos deverão seguir o padrão TR-FabricaIA-NomeDaEmpresa-AnexoX.pdf, onde X é o número que identifica o anexo.

9.2. Tabela de Pontuação para Escolha da Solução

9.2.1. Avaliação da Proposta

A proposta a ser formulada pelos proponentes deverá ser composta pelos requisitos técnicos e de colaboração estratégica exigidos neste documento. A avaliação da proposta será realizada em conformidade com as exigências contidas no Termo de Referência, com o

objetivo de assegurar a seleção da proposta mais vantajosa sob os aspectos técnico, operacional, econômico e estratégico, observando critérios objetivos e rastreáveis.

9.2.2. Critérios, Pesos e Regras por Critério

Item	Critério	Peso (Wi)	Métrica / Regra	Pontuação
1	Treino LLM – Pré-treino (MLPerf Training)	20%	<i>Time-to-quality</i> (tempo para atingir qualidade alvo). Menor tempo é melhor" Evidência Obrigatória: Logs oficiais do MLPerf Training.	0% até 20%
2	Treino LLM – Fine-tuning (MLPerf Training)	15%	<i>Time-to-quality</i> (tempo até qualidade alvo). Menor tempo é melhor — normalização tratada como benefício por inversão, via regra do MLPerf prevista no edital. Evidência Obrigatória: Logs oficiais do MLPerf Training.	0% até 15%
3	Inferência LLM – Offline (MLPerf Inference)	15%	<i>Throughput</i> (ex.: tokens/s) no cenário Offline. Evidência Obrigatória: Logs do MLPerf Inference gerados pelo <i>software stack</i> oficial.	0% até 15%
4	Inferência LLM – Server/Interactive (MLPerf Inference)	10%	Desempenho no cenário <i>Server/Interactive</i> , respeitando restrições de latência (ex.: TTFT/TPOT) conforme regras do <i>benchmark</i> . Medição e carga via LoadGen, para garantir comparabilidade e justiça.	0% até 10%
5	Eficiência energética em IA ("na parede")	10%	Tokens/W (ou tokens/J) medidos no sistema inteiro (" <i>at the wall</i> ") durante o run de performance. Evidência: Logs de potência sincronizados ao <i>benchmark</i> (<i>workflow</i> de medição de energia).	0% até 10%

Item	Critério	Peso (Wi)	Métrica / Regra	Pontuação
6	Armazenamento para IA	10%	Resultado de <i>benchmark</i> de storage exigido no edital (ex.: IO500 / MLPerf Storage, conforme seção de testes). Evidência: Relatório + logs do <i>benchmark</i> executado conforme plano de testes.	0% até 10%
7	Capacitação	5%	Número de horas de mentoria (Item 4.12.2) comprovadas (ex.: turmas, ementa, instrutores, execução). Evidência: comprovação documental e plano executável.	0% até 5%
8	Prazo de entrega/comissionamento	15%	Dias corridos até entrega + comissionamento. Prazo máximo desejável: 360 dias (acima disso, 0% no item). Prazo máximo aceitável: 520 dias.	0% até 15%
TOTAL:		100%	TOTAL:	
Benchmarks = 80% / Demais critérios = 20%				

9.2.3. Regras Gerais

- Sem evidência auditável no item → pontuação 0% naquele item.
- Resultados devem ser reproduzíveis na fase de aceitação (logs e configurações).

9.2.4. Normalização

9.2.4.1. Classificação do Critério

Classificação	Itens	Regra de Normalização
Benefício (maior é melhor)	Itens 1 a 7 *Nota 1	$Si(p) = \text{Valor do proponente} / \text{Maior valor entre todos}$
Custo (menor é)	Item 8 (prazo)	$Si(p) = \text{Menor prazo entre todos} / \text{Prazo do proponente}$

Classificação	Itens	Regra de Normalização
melhor)		

**Para os itens 1 e 2, ratifica-se que o menor tempo representa o melhor desempenho. Para aplicação da normalização percentual como critério de benefício, o tempo será previamente convertido em índice de desempenho por inversão, de forma que o menor tempo corresponda à nota normalizada de 100%.*

Para simplificar a compreensão da normalização: o melhor proponente em cada item recebe 100%; os demais recebem uma porcentagem proporcional à nota do melhor proponente naquele item.

Benefício (maior é melhor) — Itens 1 a 7:

A razão entre o valor do proponente e o maior valor entre todos define a nota (Si) do proponente avaliado, para todos os itens de 1 a 7.

$$S_i (\text{proponente}) = \text{Valor do proponente} \div \text{Maior valor entre todos}$$

Custo (menor é melhor) — Item 8 (prazo):

Para o item de prazo, a razão é invertida: quanto menor o prazo do proponente em relação ao menor prazo entre todos, maior a nota.

$$S_8 (\text{proponente}) = \text{Menor prazo entre todos} \div \text{Prazo do proponente}$$

Se Prazo do proponente > 360 dias, então S8 = 0%

9.2.4.2. Nota Técnica Final

A nota final (0 a 100%) de cada proponente é calculada multiplicando o peso (Wi) de cada item pela respectiva nota normalizada (Si), seguida da soma de todos os produtos:

$$\text{Nota} = \sum (W_i \times S_i), \text{ para } i = 1 \text{ a } 8$$

Onde Wi é o peso de cada critério (ex.: 25% = 0,25) e Si é a nota normalizada do critério i. Ao final, as notas de todos os proponentes ficam expressas em porcentagem, beneficiando

aqueles que entregarem os melhores valores em cada critério.

9.3. Avaliação da proposta de preço

A avaliação da proposta de preço considerará o Preço Global Avaliado - PGA, correspondente ao valor total necessário ao fornecimento integral da solução e ao cumprimento de todas as obrigações previstas nesta RFP e em seus anexos.

O PGA deverá incluir todos os custos diretos e indiretos necessários à execução do objeto, incluindo equipamentos, componentes, softwares, licenças, subscrições obrigatórias, garantia, suporte, instalação, integração, transporte, seguros, tributos, encargos, despesas aduaneiras, comissionamento, testes de aceitação, execução dos benchmarks, documentação, capacitação e demais despesas necessárias à entrega da solução em condições plenas de funcionamento.

A Nota de Preço - NP será expressa na escala de 0 a 100 pontos. O menor PGA entre as propostas classificadas, válidas, aceitáveis e exequíveis receberá 100 pontos, e as demais propostas receberão pontuação inversamente proporcional ao respectivo PGA, conforme a fórmula $NP_{(p)} = 100 \times [P_{\min} / P_{(p)}]$.

Somente participarão do cálculo da NP as propostas classificadas, válidas, aceitáveis e exequíveis. Se a proposta que definiu $P_{\min}$ deixar de integrar o julgamento, as notas deverão ser recalculadas.

9.4. Nota Final de Técnica e Preço

A Nota Final - NF será calculada mediante a ponderação da Nota Técnica - NT e da Nota de Preço - NP, ambas expressas na escala de 0 a 100 pontos, atribuindo-se peso de 70% à Técnica e de 30% ao Preço, conforme a fórmula $NF_{(p)} = 0,70 \times NT_{(p)} + 0,30 \times NP_{(p)}$.

Será classificado em primeiro lugar o proponente que obtiver a maior Nota Final, observados os critérios de desempate e as demais condições previstas neste Termo de Referência.

10. GARANTIAS CONTRATUAIS: DISPOSITIVO DE GARANTIA DE EXECUÇÃO E GARANTIA FINANCEIRA

10.1. Fundamentação Legal das Garantias

Este Termo de Referência institui, de forma expressa, três espécies de garantia, cuja distinção é essencial para a correta formalização contratual: (a) a garantia técnica do objeto (item 10.2), de natureza obrigacional, relativa ao funcionamento e ao desempenho dos bens fornecidos; (b) a garantia de execução contratual (item 10.3), de natureza assecuratória e caucional, destinada a assegurar o fiel cumprimento das obrigações contratuais pela Contratada; e (c) a garantia financeira (item 10.4), que detalha a cobertura e o beneficiário da garantia de execução exigida para esta contratação, sem constituir exigência cumulativa adicional de percentual.

A garantia de execução contratual e a garantia financeira fundamentam-se, subsidiariamente, na aplicação da Lei nº 14.133/2021 ao presente ajuste, nos arts. 96 a 102 dessa Lei, que disciplinam as modalidades, os percentuais, a vigência e as condições de prestação, reforço e liberação das garantias contratuais, aplicados no que for compatível com o regime de contratação por fundação de apoio de que trata o Decreto nº 8.241/2014. Em todos os casos, deverão observar a legislação vigente, o Edital e a Minuta Contratual.

10.2. Garantia Técnica do Objeto (Garantia do Fabricante)

A proposta deverá contemplar garantia integral - cobrindo hardware, software, firmware e serviços - pelo prazo mínimo de 5 (cinco) anos, contados a partir da emissão do Termo de Aceite Definitivo.

Escopo da garantia: a garantia abrange a totalidade dos componentes de hardware fornecidos, os softwares e firmwares listados no item 4.6 e 4.7, os serviços descritos no item 4.10, e as obrigações de transferência de tecnologia e capacitação previstas no item 4.12. Durante todo o período de garantia, o fornecedor deverá disponibilizar, sem custo adicional ao comprador: atualizações de segurança, patches de firmware, novas versões de software e correções de vulnerabilidades identificadas.

Garantia de desempenho: além da garantia de funcionamento, o fornecedor garante que o sistema entregue manterá desempenho igual ou superior ao comprovado nos testes de aceitação (item 5.1), medido por verificações periódicas anuais usando os mesmos benchmarks e configurações do aceite.

Componentes substituídos: peças e componentes substituídos durante o período de garantia estarão cobertos pelo prazo remanescente da garantia original ou por 12 (doze) meses a partir da substituição, valendo o prazo que for maior. O fornecedor deverá registrar todas as substituições em sistema de gestão de ativos acessível ao comprador.

Notificação de descontinuação: caso qualquer componente crítico do sistema venha a ter EOL ou EOS anunciado durante o período de garantia, o fornecedor deverá notificar o comprador com antecedência mínima de 18 (dezoito) meses, apresentando plano de substituição ou migração sem custo adicional ao comprador.

Extensão de garantia como mecanismo compensatório: penalidades decorrentes de descumprimento de obrigações de transferência de tecnologia (item 4.12) poderão ser substituídas, total ou parcialmente, por extensão do período de garantia e das obrigações associadas — sem custo adicional ao comprador —, mediante acordo formalizado entre as partes, limitado a uma substituição por categoria de obrigação.

10.3. Garantia de Execução Contratual

Como condição para a assinatura do contrato e durante toda a sua vigência, a Contratada prestará garantia de execução contratual correspondente a, no mínimo 10% (dez por cento) do valor do contrato, nos termos do art. 98 da Lei nº 14.133/2021, subsidiariamente aplicado.

Caberá à Contratada optar, à sua conveniência, por uma das seguintes modalidades de garantia, nos termos do art. 96, §1º, da Lei nº 14.133/2021, subsidiariamente aplicado: (i) caução em dinheiro ou em títulos da dívida pública, resgatáveis pelo valor nominal, emitidos

sob forma escritural e registrados em sistema centralizado de liquidação e custódia autorizado pelo Banco Central do Brasil; (ii) seguro-garantia; ou (iii) fiança bancária emitida por banco ou instituição financeira devidamente autorizada a operar no País pelo Banco Central do Brasil.

A garantia de execução contratual assegurará, em caráter não exaustivo e sem prejuízo do disposto no item 10.4, o pagamento de: prejuízos decorrentes do não cumprimento do objeto do contrato; multas aplicadas pela Contratante à Contratada; prejuízos diretos causados à Contratante decorrentes de culpa ou dolo durante a execução do contrato; obrigações trabalhistas e previdenciárias de qualquer natureza, não adimplidas pela Contratada, quando couber; e prejuízos indiretos causados à Contratante e prejuízos causados a terceiros, decorrentes de culpa ou dolo durante a execução do contrato.

A garantia deverá manter-se válida e eficaz durante toda a execução contratual, incluído o período de garantia técnica de que trata o item 10.2 (mínimo de 5 anos, contados do Aceite Definitivo, item 5.1), devendo a Contratada promover a sua renovação ou o seu reforço sempre que utilizada, no todo ou em parte, para cobertura de penalidades ou indenizações, no prazo máximo de 10 (dez) dias úteis contados da notificação da Contratante, sob pena de rescisão contratual e aplicação das sanções previstas neste Edital e na legislação vigente.

A garantia será liberada ou restituída após a emissão do Aceite Definitivo (item 5.1) e o decurso do prazo de garantia técnica de que trata o item 10.2, mediante comprovação da quitação de todas as obrigações trabalhistas, previdenciárias e fiscais relacionadas à execução do contrato, atualizada monetariamente quando prestada em dinheiro, nos termos do art. 100 da Lei nº 14.133/2021, subsidiariamente aplicado.

10.4. Garantia Financeira (Cobertura e Beneficiário)

A garantia financeira de que trata este item constitui detalhamento específico, quanto à cobertura e ao beneficiário, da garantia de execução contratual disciplinada no item 10.3, com a qual não se cumula quanto ao percentual exigido, observado o disposto nos arts. 96 a 100

da Lei nº 14.133/2021, subsidiariamente aplicados, e o regime de contratação por fundação de apoio de que trata o Decreto nº 8.241/2014.

Como condição para formalização contratual, a empresa contratada deverá prestar garantia financeira de execução do objeto, no percentual de 5% (cinco por cento) do valor total do contrato, tendo como beneficiário a Contratante.

A garantia assegurará o pagamento de:

- prejuízos advindos do não cumprimento do objeto do contrato;
- multas punitivas aplicadas pela Contratante à Contratada;
- prejuízos diretos causados à Contratante decorrentes de culpa ou dolo durante a execução do contrato;
- obrigações trabalhistas e previdenciárias de qualquer natureza não adimplidas pela Contratada, quando couber; e
- prejuízos indiretos causados à Contratante e prejuízos causados a terceiros, decorrentes de culpa ou dolo durante a execução do contrato.

O detalhamento das demais condições de prestação de garantia financeira será apresentados no instrumento convocatório e na minuta contratual.

10.5. Garantia Complementar para Serviços de Engenharia de Grande Vulto

O valor total estimado da contratação, correspondente a **R\$ 959.040.959,04 (novecentos e cinquenta e nove milhões, quarenta mil, novecentos e cinquenta e nove reais e quatro centavos)**, conforme cronograma de pagamento do item 8.2, supera o patamar de R\$ 200.000.000,00 (duzentos milhões de reais) fixado no art. 6º, XXII, da Lei nº 14.133/2021 para caracterização de obras, serviços e fornecimentos de grande vulto — patamar atualizado para R\$ 250.902.323,87 (duzentos e cinquenta milhões, novecentos e dois mil, trezentos e vinte e

três reais e oitenta e sete centavos) pelo Decreto nº 12.343, de 19 de dezembro de 2024, com efeitos a partir de 2025.

Considerando que o objeto compreende, além do fornecimento de hardware e software, serviços de instalação física, montagem, cabeamento estruturado e infraestrutura predial elétrica, hidráulica e de refrigeração, conforme Termo de Referência, a Contratante poderá exigir, para a parcela do objeto caracterizável como serviço de engenharia de grande vulto, a prestação de garantia na modalidade seguro-garantia com cláusula de retomada, em percentual de até 30% (trinta por cento) do valor inicial do contrato, nos termos do art. 102 da Lei nº 14.133/2021, subsidiariamente aplicado, mediante justificativa técnica a ser consignada nos autos do processo e detalhada no instrumento convocatório e na minuta contratual.

A eventual exigência da garantia complementar de que trata este item não dispensa nem substitui a garantia de execução contratual do item 10.3, tampouco a garantia técnica do item 10.2, tratando-se de mecanismo assecuratório adicional, específico e proporcional ao risco identificado na parcela do objeto correspondente a obras e serviços de engenharia.

11. ACORDO DE NÍVEIS DE SERVIÇO, FISCALIZAÇÃO E REGIME SANCIONATÓRIO

11.1. Disposições Gerais do SLA

O presente Acordo de Níveis de Serviço (SLA) formaliza, para fins de fiscalização contratual (item 7) e de eventual aplicação de sanções (item 11), os indicadores, prazos de atendimento e níveis mínimos de disponibilidade exigidos da Contratada, complementando as garantias de que trata o item 10, os critérios de aceitação do item 5.1 e o Serviço de Suporte e Monitoramento (4.10.2), sem prejuízo do detalhamento adicional a ser consolidado no

instrumento convocatório e na minuta contratual, nos termos do art. 6º, XXIII, e dos arts. 115 a 130 da Lei nº 14.133/2021, subsidiariamente aplicados.

O fornecedor deverá executar os serviços em conformidade com os níveis de serviço (SLAs) e requisitos técnicos definidos neste Termo de Referência (especialmente o item 4.10.2. Serviço de Suporte e Monitoramento) e em seus anexos. O não atendimento a tais requisitos poderá ensejar a aplicação de sanções contratuais, conforme detalhamento a ser estabelecido na fase de contratação.

11.2. Sanções

11.2.1. Detalhamento Contratual

Os critérios, fórmulas de cálculo, prazos e condições de aplicação das sanções serão detalhados no Edital e na minuta contratual a ser apresentada na fase subsequente deste processo. Em todos os casos, deverão observar a legislação vigente.

11.3. Regime de Sanções Administrativas Subsidiário

Sem prejuízo do detalhamento contratual de que trata o item 11.2.1, aplica-se subsidiariamente à execução deste contrato o regime de infrações e sanções administrativas da Lei nº 14.133/2021, no que for compatível com a natureza jurídico-privada da Contratante, compreendendo as seguintes espécies, nos termos do art. 156 dessa Lei: advertência; multa, na forma prevista no instrumento convocatório e na minuta contratual; impedimento de licitar e contratar, pelo prazo máximo de 3 (três) anos; e declaração de inidoneidade para

licitar ou contratar, pelo prazo mínimo de 3 (três) e máximo de 6 (seis) anos, aplicável nas hipóteses de infração mais grave, nos termos do art. 155 da Lei nº 14.133/2021.

A aplicação de sanções não exime a Contratada da obrigação de reparar integralmente o dano causado à Contratante ou a terceiros, nem prejudica a execução da garantia de que trata o item 10.3, tampouco a rescisão contratual, quando cabível.

12. DISPOSIÇÕES FINAIS

Aplica-se a todos os itens deste Termo de Referência a cláusula de equivalência tecnológica prevista no item 5.6.6, sempre que indicado componente, ferramenta, padrão ou tecnologia por nome, admitindo-se solução “ou equivalente”, mediante comprovação de equivalência funcional, compatibilidade e desempenho igual ou superior nos benchmarks exigidos.

Fica eleito o foro da **Justiça Federal, Subseção Judiciária de Petrópolis/RJ** para dirimir eventuais controvérsias decorrentes deste Termo de Referência e do contrato dele decorrente, com renúncia a qualquer outro, por mais privilegiado que seja, ressalvada a possibilidade de eleição de câmara arbitral, se assim pactuado pelas partes.

Este Termo de Referência integra o processo administrativo de contratação, dele fazendo parte, para todos os efeitos, a Requisição de Propostas (RFP) técnica originária, seus anexos e os estudos técnicos preliminares correspondentes.

Petrópolis, data da última assinatura eletrônica.

Assinado eletronicamente por:

FUNDAÇÃO DE APOIO AO DESENVOLVIMENTO DA COMPUTAÇÃO CIENTÍFICA - FACC

Francisco Roberto Leonardo

Diretor Geral

Flávio Barbosa Toledo

Diretor Administrativo-Financeiro

LABORATÓRIO NACIONAL DE COMPUTAÇÃO CIENTÍFICA – LNCC

Fábio Borges de Oliveira

Diretor e Coordenador do Projeto

Wagner Vieira Léo

Coordenador de TI

Anexo_I_Termo_deReferencia_SP_27_LNCC.pdf

Documento número #4cf5e4d1-22c0-49b7-bb3d-b2ce786705f9

Hash do documento original (SHA256): 2e467b7f91546bdb212797e8f233bac98fa47a5fb25607baad9fcb4482cba87d

Assinaturas



Flávio Barbosa Toledo

CPF: 350.604.504-06

Assinou como parte em 19 ago 2026 às 17:57:29



Francisco Roberto Leonardo

CPF: 386.665.457-04

Assinou como parte em 19 ago 2026 às 20:07:14



Wagner Vieira Léo

CPF: 732.796.687-00

Assinou como parte em 19 ago 2026 às 20:33:47



Fábio Borges de Oliveira

CPF: 993.940.659-20

Assinou como parte em 19 ago 2026 às 20:45:33

Log

19 ago 2026, 17:41:50	Operador com email crislayne.santos@facc10.org.br na Conta 0d2fec55-6c6d-46ac-93e0-3e2ba7b6a490 criou este documento número 4cf5e4d1-22c0-49b7-bb3d-b2ce786705f9. Data limite para assinatura do documento: 12 de novembro de 2026 (09:43). Finalização automática após a última assinatura: habilitada. Idioma: Português brasileiro.
19 ago 2026, 17:45:05	Operador com email crislayne.santos@facc10.org.br na Conta 0d2fec55-6c6d-46ac-93e0-3e2ba7b6a490 alterou o processo de assinatura. Finalização automática após a última assinatura: não habilitada.
19 ago 2026, 17:45:05	Operador com email crislayne.santos@facc10.org.br na Conta 0d2fec55-6c6d-46ac-93e0-3e2ba7b6a490 adicionou à Lista de Assinatura: borges@lncc.br para assinar como parte, via E-mail. Pontos de autenticação: Token via E-mail; Nome Completo; CPF; endereço de IP. Dados informados pelo Operador para validação do signatário: nome completo Fábio Borges de Oliveira.

- 19 ago 2026, 17:45:06 Operador com email crislayne.santos@facc10.org.br na Conta 0d2fec55-6c6d-46ac-93e0-3e2ba7b6a490 adicionou à Lista de Assinatura: wagner@Incc.br para assinar como parte, via E-mail.
- Pontos de autenticação: Token via E-mail; Nome Completo; CPF; endereço de IP. Dados informados pelo Operador para validação do signatário: nome completo Wagner Vieira Léo.
- 19 ago 2026, 17:45:06 Operador com email crislayne.santos@facc10.org.br na Conta 0d2fec55-6c6d-46ac-93e0-3e2ba7b6a490 adicionou à Lista de Assinatura: dirgeral@facc10.org.br para assinar como parte, via E-mail.
- Pontos de autenticação: Token via E-mail; Nome Completo; CPF; endereço de IP. Dados informados pelo Operador para validação do signatário: nome completo Francisco Roberto Leonardo e CPF 386.665.457-04.
- 19 ago 2026, 17:45:06 Operador com email crislayne.santos@facc10.org.br na Conta 0d2fec55-6c6d-46ac-93e0-3e2ba7b6a490 adicionou à Lista de Assinatura: flavio@facc10.org.br para assinar como parte, via E-mail.
- Pontos de autenticação: Token via E-mail; Nome Completo; CPF; endereço de IP. Dados informados pelo Operador para validação do signatário: nome completo Flávio Barbosa Toledo e CPF 350.604.504-06.
- 19 ago 2026, 17:57:29 Flávio Barbosa Toledo assinou como parte. Pontos de autenticação: Token via E-mail flavio@facc10.org.br. CPF informado: 350.604.504-06. IP: 189.60.58.82. Localização compartilhada pelo dispositivo eletrônico: latitude -22.96176814643251 e longitude -43.21566953510201. URL para abrir a localização no mapa: <https://app.clicksign.com/location>. Componente de assinatura versão 1.1501.0 disponibilizado em <https://app.clicksign.com>.
- 19 ago 2026, 20:07:14 Francisco Roberto Leonardo assinou como parte. Pontos de autenticação: Token via E-mail dirgeral@facc10.org.br. CPF informado: 386.665.457-04. IP: 152.84.125.14. Localização compartilhada pelo dispositivo eletrônico: latitude -22.8153887 e longitude -43.2467174. URL para abrir a localização no mapa: <https://app.clicksign.com/location>. Componente de assinatura versão 1.1501.0 disponibilizado em <https://app.clicksign.com>.
- 19 ago 2026, 20:33:47 Wagner Vieira Léo assinou como parte. Pontos de autenticação: Token via E-mail wagner@Incc.br. CPF informado: 732.796.687-00. IP: 179.82.8.161. Componente de assinatura versão 1.1501.0 disponibilizado em <https://app.clicksign.com>.
- 19 ago 2026, 20:45:33 Fábio Borges de Oliveira assinou como parte. Pontos de autenticação: Token via E-mail borges@Incc.br. CPF informado: 993.940.659-20. IP: 179.249.67.32. Localização compartilhada pelo dispositivo eletrônico: latitude -15.79924844911462 e longitude -47.87010708377014. URL para abrir a localização no mapa: <https://app.clicksign.com/location>. Componente de assinatura versão 1.1501.0 disponibilizado em <https://app.clicksign.com>.
- 20 ago 2026, 08:32:36 Operador com email crislayne.santos@facc10.org.br na Conta 0d2fec55-6c6d-46ac-93e0-3e2ba7b6a490 finalizou o processo de assinatura. Processo de assinatura concluído para o documento número 4cf5e4d1-22c0-49b7-bb3d-b2ce786705f9.

**Documento assinado com validade jurídica.**

Para conferir a validade, acesse <https://www.clicksign.com/validador> e utilize a senha gerada pelos signatários ou envie este arquivo em PDF.

As assinaturas digitais e eletrônicas têm validade jurídica prevista na Medida Provisória nº. 2200-2 / 2001

Este Log é exclusivo e deve ser considerado parte do documento nº 4cf5e4d1-22c0-49b7-bb3d-b2ce786705f9, com os efeitos

prescritos nos Termos de Uso da Clicksign, disponível em www.clicksign.com.