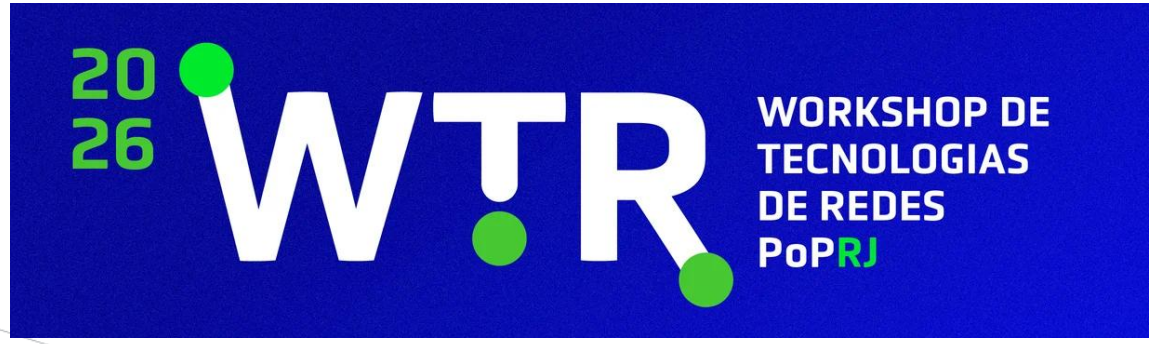
RNP

Educação, Pesquisa
e Inovação em Rede

Supercomputação para IA na RNP



PBIA Plano Brasileiro de Inteligência Artificial

Eixo 1 – Infraestrutura e Desenvolvimento de IA – R$ 5,79 bilhões

Ação 1 – Supercomputador de IA

- Aquisição de supercomputador especializado – Top 5 Mundial
- Desafio: Ampliar significativamente a capacidade de processamento de alto desempenho do Brasil
- **Recursos (2024-28): R$ 1,8 bilhão (FNDCT)**

Ação 3. Redes de conexão de alta velocidade para supercomputação

- Ampliação do acesso a supercomputadores, com investimentos em equipamentos de entrada e conexão
- Desafio: Conectar em rede e garantir o acesso aos supercomputadores brasileiros
- Meta: Interligação de todos os centros de supercomputação com redes de alta velocidade até 2028
- **Recursos (2024-28): R$ 52,4 milhões (FNDCT)**

Supercomputador de IA

Rotas Tecnológicas

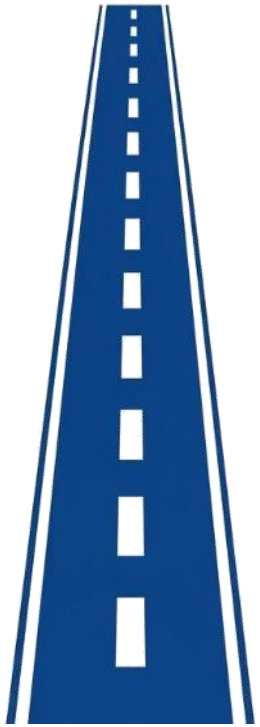
- O Brasil precisa de infraestrutura de supercomputação para inteligência artificial.
- Contexto geopolítico agrava o problema.
- Brasil busca soberania nesse tema
- Decisão estratégica: o Brasil seguirá com três rotas
  - Rota 1 - NVIDIA, para capacidade imediata e posicionamento no Top500
  - Rota 2 - Huawei, como rota de diversificação e autonomia
  - Rota 3 - RISC-V para alcance de soberania tecnológica em hardware

  - Discurso Presidente Lula e Ministra Luciana Santos sobre as rotas:
    - https://www.youtube.com/live/v9V3vdV52D4

Estratégia baseada em três rotas tecnológicas

Rota 1
Aumento da capacidade

Baseado em arquitetura de GPUs NVIDIA



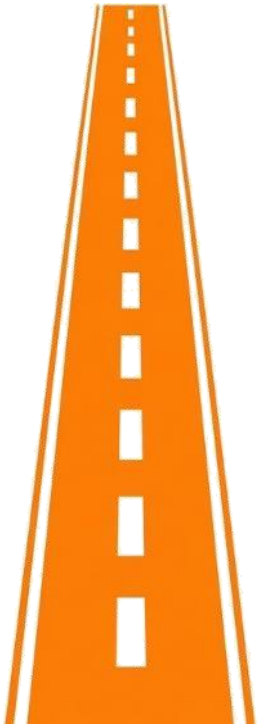
ROUTE 2
Transferência de Tecnologia

Baseado na arquitetura NPU ASCEND Huawei



ROUTE 3
Autonomia em longo prazo

RISC-V
Open hardware architecture



ROTAS TECNOLÓGICAS DA SUPERCOMPUTAÇÃO BRASILEIRA – PBIA 2024-2028





1 -Supercomputador (compra)
Natal (RN)

Coloca o Brasil entre os dez países com maior capacidade de processamento de IA.



2 - Cooperação tecnológica - China
Rio de Janeiro (RJ)

Ampliar a supercomputação no país, desenvolver modelos de IA em língua portuguesa e formar profissionais.



3 - Chips - arquitetura aberta RISC-V
Campinas (SP)

Aumentar as possibilidades de desenvolvimento tecnológico e reduzir a dependência de arquiteturas proprietárias.



Características Gerais da Rota Tecnológica com a China

- É um acordo de cooperação entre Brasil x China com transferência de tecnologia para que o Brasil adquira autonomia e soberania em IA
- Financiado com recursos FNDCT
- O que inclui:
  - Hospedagem em Data Center (Co-location)
  - Aquisição de um supercomputador
  - Transferência de Tecnologia e conhecimento para o desenvolvimento do modelo de linguagem brasileiro (LLM Nacional)
  - Apoio ao desenvolvimento da pilha de software nacional
  - Formação e capacitação no desenvolvimento em IA (IA Talent)

Proposta (1/2)

COMPONENTE	ESPECIFICAÇÃO
Capacidade	560 PF em FP16 (não esparso)
Aceleradores	1.024 NPUs Ascend 950T
Data Center	Data Center 6.400m², 4MW, refrigeração líquida
Software	CANN + vLLM-Ascend
Capacitação	2.900 profissionais/estudantes/pesquisadores em 3 anos , em O&M e em desenvolvimento (Governo, academia e indústria)
Fases	Fase I (2027) : 280 PF, 512 cards Fase II (2028) : 280 PF, 512 cards
Modelo operação	primeiro ano operado pela Huawei com transferência de conhecimento para RNP e depois a RNP opera com suporte da Huawei.

Proposta (2/2)

COMPONENTE	ESPECIFICAÇÃO
Modelo Fundacional LLM brasileiro	Construção do Corpus a partir de um modelo aberto Treinamento do modelo Desenvolvimento de APP nacional (inferência) Otimização do modelo
DLM	Desenvolvimento de três DLM (Domain Language Model) especifico para verticais a serem definidas
Transferência de Conhecimento	Capacitação para formação em desenvolver modelos
Suporte	1 ano de O&M serviço de suporte 5 anos serviço de suporte de manutenção

L2 AI Computing: Atlas 950 SuperPod, flexible scaling with more powerful computing resource

Huawei provide total 560Pflops FP16 non-sparse (1024*AI NPUs)

Ascend 950DT



One cabinet

Technical Specification

- 64 NPUs
- 35 PFLOPS FP16
- 6,144 GB HBM
- 100 kW
- 1,700 kg
- 1,400 x 600 x 2,250 mm

High-capacity quick connector



Liquid cooling flow capacity increased by 8 times, and the junction temperature of core components reduced by 5℃

High-power TPSU power supply



100M charge-discharge cycles | 4× lifespan vs conventional capacitors | Utility power utilization rate ↑20%

Atlas 950 SuperPod

Multi-cabinet assembly

Flexible scaling



Flexible scaling by cabinet: 256/ 512 / 1,024 ...8192 NPUs



Orthogonal architecture, cableless electrical interconnection
Floating blind-mate liquid-cooling connector: zero leakage

Reliable components	Reliable links	Reliable system
2x more reliable optical modules with innovative materials	100-ns-level fault detection on every optical path	2+2 optical line protection for smooth applications



L2 Computing: TaiShan 200, high performance and solid reliability

Huawei provide 105*TaiShan 200 servers for Application and Management

TaiShan 200 2280 V2 Server



High Performance

- Turbocharged computing: industry's first dual-threaded Arm server processor up to 64 cores at 2.2 GHz
- Multi-processor interconnection: proprietary Huawei Cache Coherence System (HCCS) protocol for 2P interconnection
- High-speed connectivity: integrated PCIe 5.0, 2 × 100GE networking, and SAS/SATA 3.0

Solid Reliability

- Predictive online isolation of faulty CPU cores for continuous operation
- Memory fault prediction and self-healing
- Enhanced Uncorrectable Memory Error (UCE) prevention via partial or full memory mirroring

Server specification

Item	TaiShan 200 2280 V2
Form	2U 2-socket
CPU	2 × new Kunpeng 920 processor model, with 64 cores, 2.2 GHz
Memory	8*64G DDR5 DIMMs at 4800 MT/s 512GB memory
Local storage	8*7.68TB 2.5 inch NVMe SSD
RAID support	RAID 0/1/10/1E/5/50/6/60
LOM	2 onboard NICs, each supporting 4 × GE, 2 × 25GE
PCIe expansion	Up to 8 PCIe slots
Power supply	2 × 2000 W AC PSU in 1+1 redundancy
Fan module	4 hot-swappable fan modules in N+1 redundancy
Operating temperature	5°C to 35°C (41°F to 95°F)

L2 Storage: OceanStor A800, the next generation high performance storage for AI

Huawei provide 8.3PB capacity with AI storage in NAS and S3

Ultra performance	Ultra-large cluster	Powerful inference

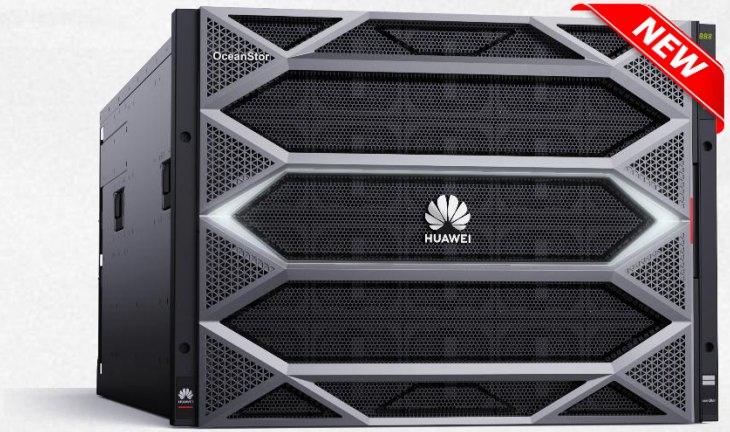
500 GB/s
Bandwidth per enclosure

10,000,000
IOPS per enclosure

EB-level expansion
Scale-out for up to 512 controllers in a single cluster

AI knowledge base
High-dimensional vector search from large-capacity AI knowledge base

Long-term memory storage
Multi-level KV cache



Next-Gen High-Performance Distributed File Storage for AI

OceanStor A800

Brazilian National AI Project Overall Architecture

Aiming at the vision of PBIA, Focusing on national AI sovereign model with AI Infrastructure and AI Talent

Application Scenarios

National APP ChatBot

e-Gov app
- Office AI assistance
- Public Service Inquiry

Government Affairs Model

Security app
- Public opinion monitoring
- Content Authenticity Verification

Security Model

Social Well-being App
- AI Teaching
- AI Courseware

Public Services Model

Industry app
- Oil & Gas
- Mining
- Agriculture

Industry-specific models

AI Models Layer

Brazil Portuguese Foundation Model

General Dataset

AI Capacity Center

- Data Development Platform
- Model Training Platform
- Application Innovation Platform

AI Foundation

AI Computing Power

- AI Computing
- General Computing
- Storage
- Network
- Security
- ...

AIDC

- Power
- Cooling
- Rack
- Cabling
- Fire fighting
- ...

HUAWEI

科大讯飞 IFLYTEK

Future Expansion

O&M Support & Maintenance

Model iteration

System Integration

Knowledge & Capacity Transfer

O&M Support & Maintenance

AI Talent

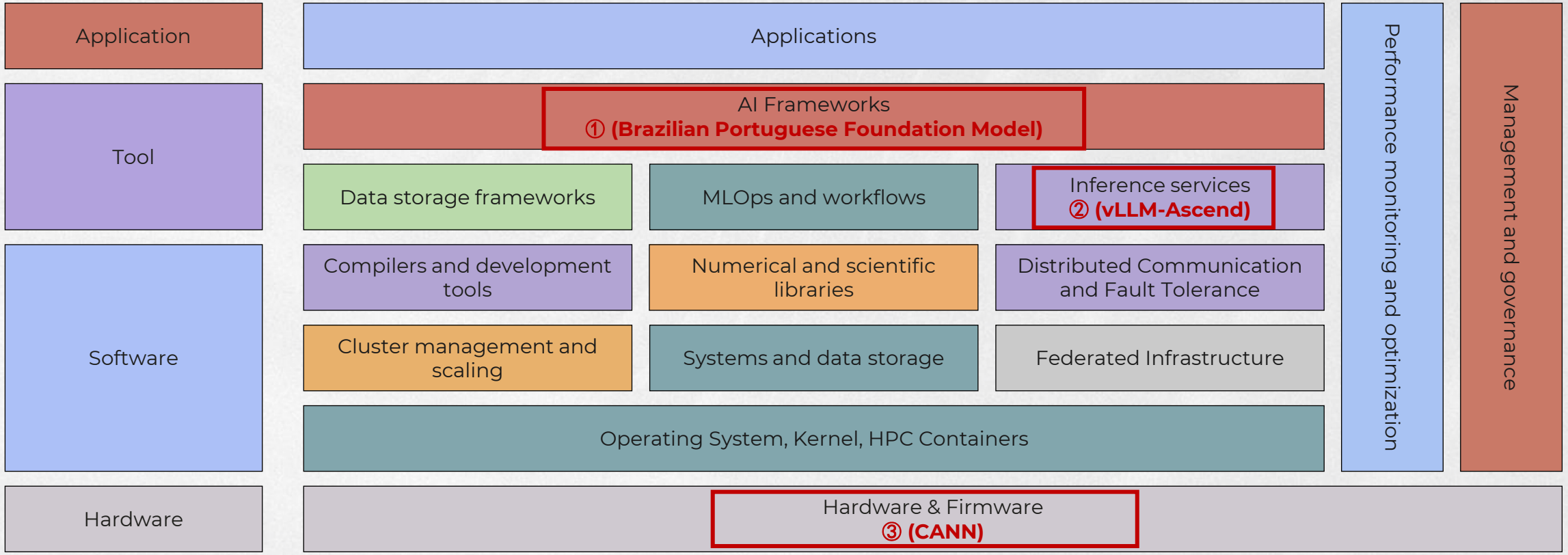
Brazil Full Stack: Huawei provides 3 parts to help Brazil achieve independent innovation in AI stacks.

Huawei's Participation Scope

Part 1:
Build Brazilian Foundation Model;

Part 2:
Develop inference tool based on vLLM-Ascend;

Part 3:
Develop Brazilian new operators based on CANN;

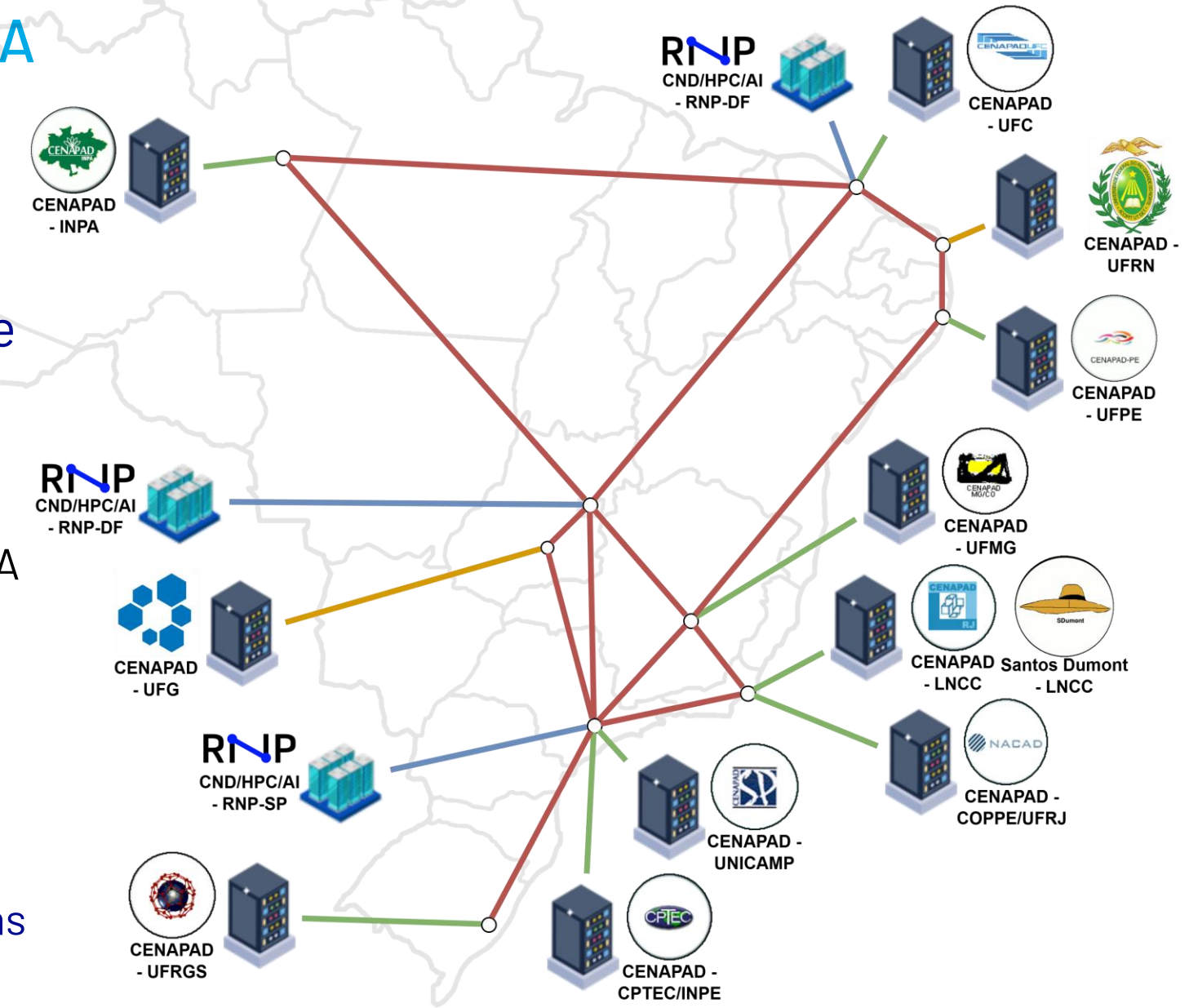


Redes de conexão de alta velocidade para supercomputação

Rede de e-Ciência para IA

Características

- Interconexão em alta velocidade
  - Entre repositórios de dados, CNDs e centros de supercomputação/HPC
  - Movimentação otimizada dos dados
  - Heterogeneidade de recursos para IA (CPU/GPU/NPU/TPU...)
  - Capacidade nacional compartilhada
- Acesso seguro à Internet para IA como Serviço (IAaaS)
  - Inferência em tempo real
  - Ingestão de dados de fontes externas



Requisitos de Redes para IA (valores meramente ilustrativos)

Fluxo de Trabalho	Restrição Primária	Latência Alvo	Vazão Estimada	Volume de Dados
Ingestão de Dados	Volume / Vazão	< 200ms	1 - 40 Gbps	Petabyte
Treinamento Distríbuido	Sincronização	< 10µs - 5ms	100 - 800 Gbps	Peta - Exabyte
Aprendizado Federado	Sincronização Concorrente	< 150ms	100 Mbps - 1 Gbps	Gigabyte
Inferência em Tempo Real	Tempo de Resposta	< 50ms	10 Mbps - 1 Gbps	Gigabyte/Hora



Nuvem de GPU

Motivação

- Compartilhamento de recursos (democratizar o acesso)
- Demanda por processamento em IA (e dados) requer infraestruturas escaláveis, resilientes e, principalmente, **eficientes**
- Atendimento (ágil) a diferentes **perfis de uso**
- Reduzir a dependência de nuvens comerciais
- Fortalecer a capacidade nacional em HPC e IA.
  - pesquisa em IA e HPC
  - pesquisa com o uso de IA
  - desenvolvimento tecnológico (startups, PMEs)
  - formação de recursos humanos

O que é a infraestrutura federada de HPC/IA?

- **Baseada em uma arquitetura em múltiplos *tiers:***
  - Diversidade de serviços/ofertas/**clientes**
  - Diversidade tecnológica
- **Governança para promover o uso eficiente dos recursos compartilhados**

Visão Integrada de Recursos de HPC/IA

TIER 1 – Supercomputadores Rota 1 e 2

- Grandes Modelos / Treinamento em Larga Escala

TIER 2 – CENAPADs

- Modelos Médios / Treinamento Intermediário

TIER 3 – Nuvem de GPUs - RNP

- Modelos Pequenos / Inferência / Edge

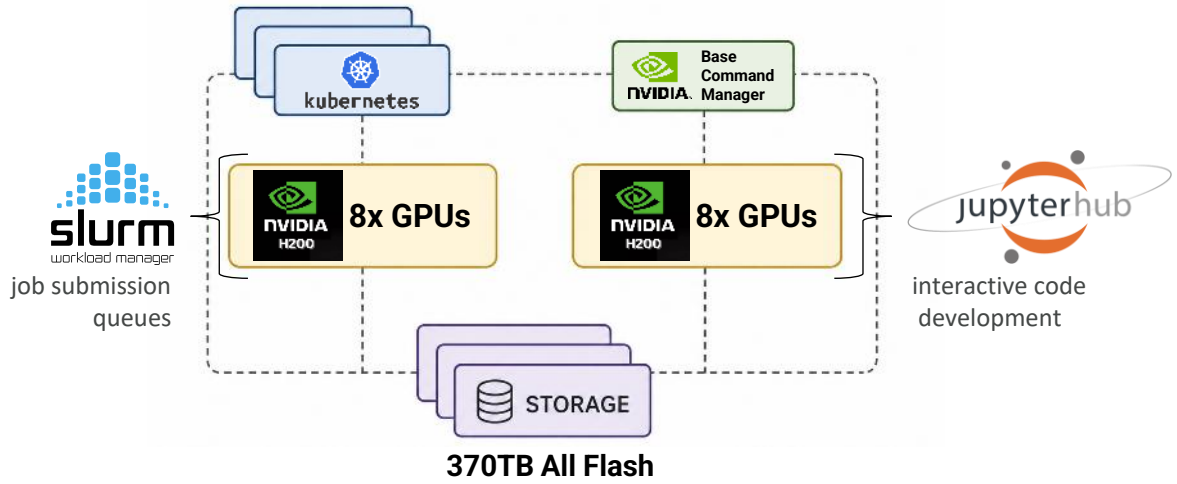
Operação coordenada e interoperável

Duas Infraestruturas para IA em desenvolvimento



Nuvem de GPU

- 3 a 5 sites com 16 GPUs cada
- Baseada em NVIDIA stack
- De 1/7 até 2 GPUs por usuários
- Atender a estudantes e pesquisadores que necessitam depoucos recursos para processor pequenos modelos





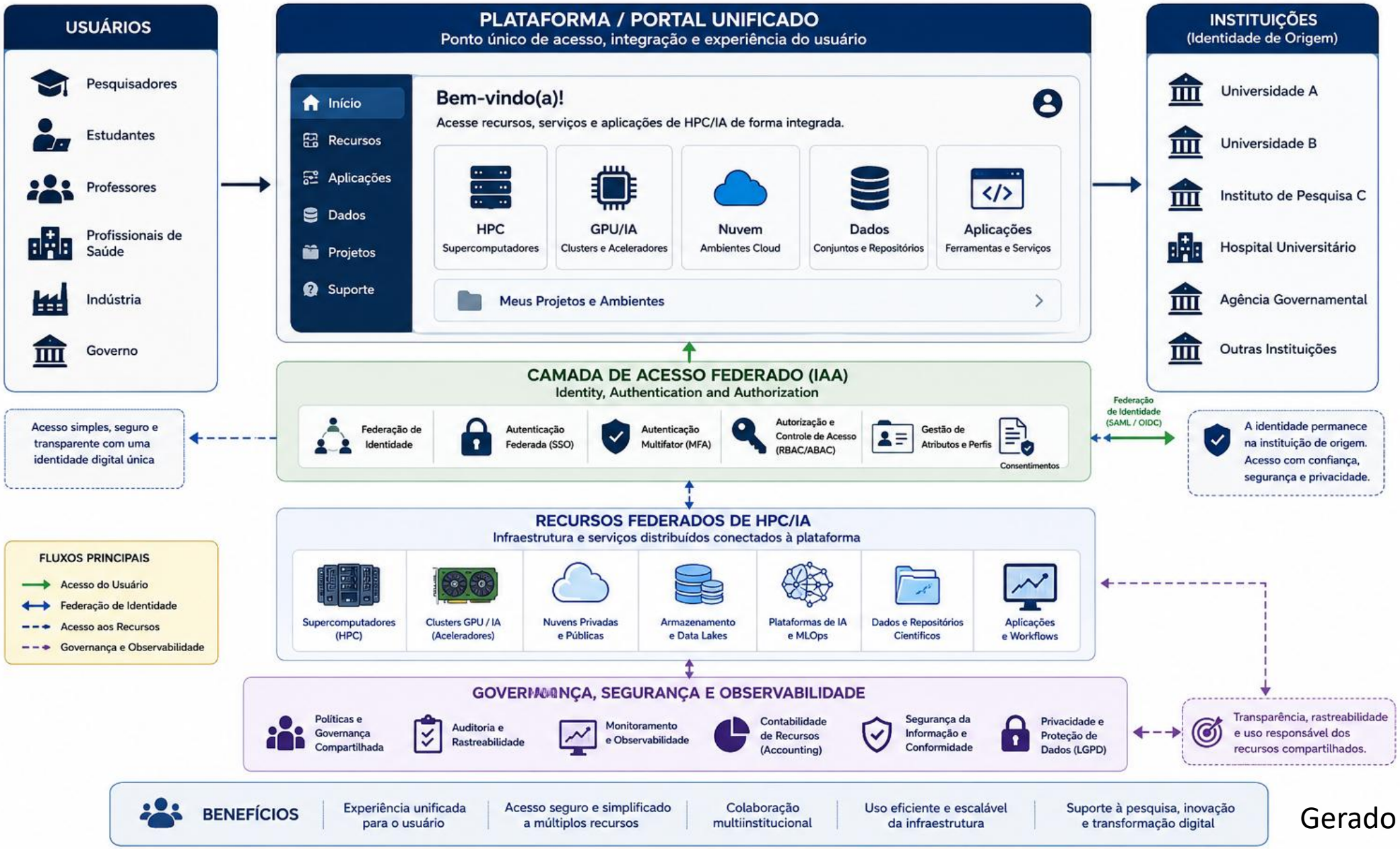
Fabrica de IA

- AI-Facitity a ser implantada no Rio de Janeiro
- Em parceria com a Huawei
- **1024 aceleradores** (NPUs)
- **560 Pflops** entregues em 2 fasses
- 4 MW de energia
- Desenvolvimento do modelo fundacional LLM em poprtugues e tres DLMs
- Inicio em 2027

Plataforma/Portal Unificado e Acesso Federado (IAA) (esboço)

Plataforma/Portal Unificado e Acesso Federado (IAA)

Um único ponto de entrada para múltiplos recursos de HPC/IA com identidade federada, segurança e governança compartilhada



Gerado por IA

RECURSOS FEDERADOS DE HPC/IA

Infraestrutura distribuída, integrada e interoperável

SUPERCOMPUTADORES (HPC)



- Partições multiusuário
- Interconexões de alta velocidade (InfiniBand)
- CPUs de última geração
- Ambientes de simulação e modelagem

CLUSTERS GPU / IA



- GPUs de alto desempenho (NVIDIA, AMD, Intel)
- Partições dedicadas e compartilhadas
- Suporte a multi-GPU e multi-nó
- Instâncias otimizadas para IA generativa

NUVEM HÍBRIDA FEDERADA



- Nuvem privada e pública federada
- Containers e Kubernetes
- Elasticidade sob demanda
- Ambientes efêmeros
- Integração com clouds comerciais (quando permitido)

ARMAZENAMENTO & DADOS



- Storage de alta performance (paralelo e distribuído)
- Data Lakes federados
- Compartilhamento seguro e colaborativo
- Hierarquias de dados (quente/morno/frio)

PLATAFORMAS DE IA



- Frameworks (PyTorch, TensorFlow, JAX, etc.)
- Ambientes MLOps
- Inferência em escala
- Fine-tuning e IA generativa
- Catálogo de modelos e datasets

FÁBRICAS DE IA (AI FACTORIES)



- Infraestrutura otimizada para treinamento de modelos em larga escala
- Ambientes para Fine-tuning e experimentação com dados sensíveis
- Apoio a startups, P&D e indústria
- Governança, segurança e conformidade integradas

TECNOLOGIAS E PADRÕES ABERTOS

Slurm (Workload Manager) · Kubernetes (Orquestração) · OpenStack (Cloud) · Ceph (Armazenamento) · KubeFlow (MLOps) · MLflow (Rastreamento de Experimentos) · Jupyter (Dev & Data Science) · PyTorch (IA) · TensorFlow (IA) · Horovod (Treinamento Distribuído) · OpenOnDemand (Acesso Web HPC) · Prometheus (Monitoramento) · Keycloak (Identidade) · S3/MinIO (Objetos) · ONNX (Modelos) · RStudio (IDE) · VS Code (IDE) · ... e outros

Gerado por IA

Iniciativas Internacionais (EU e EUA)

Resumo Comparativo das Iniciativas

Iniciativa	EuroHPC	ACCESS	NAIRR
Foco Principal	HPC/ Soberania	Pesquisa Geral	IA Nacional
Orquestrador	Waldur	Slurm/Globus	Kubernetes
Hardware	Super/Quântica	HPC Clusters	GPU Clusters
Financiamento	UE	NSF	Multi-Agência

Obrigado

iara.machado@rnp.br
leandro.Ciuffo@rnp.br
michelle.wangham@rnp.br
jose.rezende@rnp.br