27° Workshop RNP

Infraestrutura Federada de HPC/IA

José Rezende (RNP)

Antônio Tadeu (LNCC)

Michelle Wangham (RNP)

27° Workshop RNP

Motivação

Demanda crescente por processamento em IA

- Democratização do acesso
- Atendimento (ágil) a diferentes perfis de uso
- Redução da dependência de nuvens comerciais
- Fortalecimento da capacidade nacional em HPC e IA
  - fomento à pesquisa em IA e HPC
  - fomento à pesquisa com o uso de IA
  - desenvolvimento tecnológico (startups, PMEs)
  - formação de recursos humanos

27° Workshop RNP

Nuvem de GPUs da RNP

Infraestrutura Distribuída

3 sites

- CND Brasília - Lei de TICs
- CNDs São Paulo e Fortaleza – PBIA

Hardware de Referência

- 3x Servidores HGX: 8 GPUs NVidia H200
- 1x Storage Paralelo ~368 TB NVMe
- 4x Servidores Kubernetes (1 Head e 3 Cluster Nodes)
- 2x Switches Ethernet 200Gbps

Software

- NVidia AI Enterprise 4
- Run.AI

27° Workshop RNP

O que é a infraestrutura federada de HPC/IA?

Em construção

Arquitetura em múltiplos tiers

- Diversidade de serviços/ofertas/clientes
- Diversidade tecnológica

Governança para promover o uso eficiente dos recursos da federação

TIER 1 – LNCC (supercomputadores)

- Grandes Modelos / Treinamento em Larga Escala

TIER 2 – CENAPADs

- Modelos Médios / Treinamento Intermediário

TIER 3 – Nuvem de GPUs - RNP

- Pequenos Modelos / Inferência / Edge

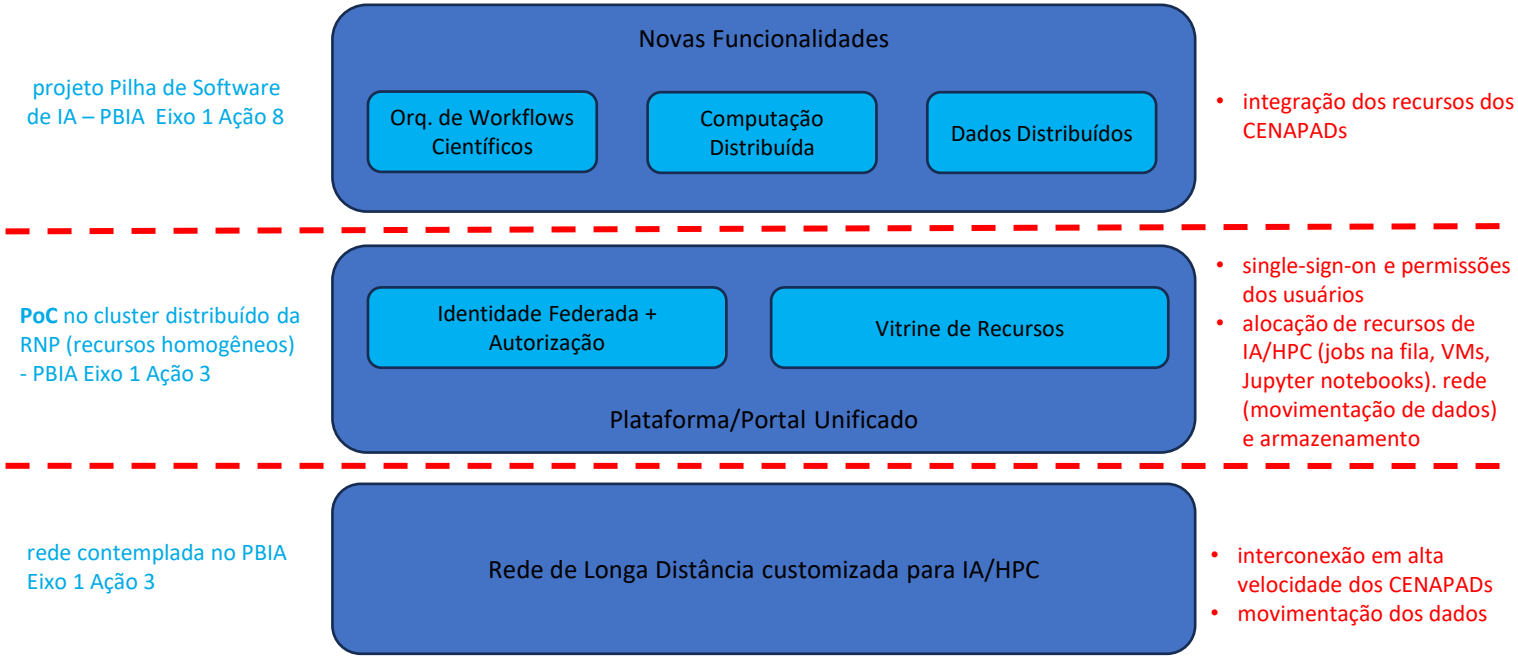
27 Workshop RNP

Resumo Comparativo

Supercomputador tradicional	Infraestrutura federada HPC/IA
Centralizada	Distribuída
Uma instituição	Múltiplas instituições
Recursos locais	Recursos compartilhados
Gestão isolada	Governança federada
HPC científico	HPC + IA + dados + cloud
Usuários locais	Ecossistema nacional/internacional

27° Workshop RNP

Alicerces da Infraestrutura Federada



27 Workshop RNP

Rede de Alto Desempenho para IA

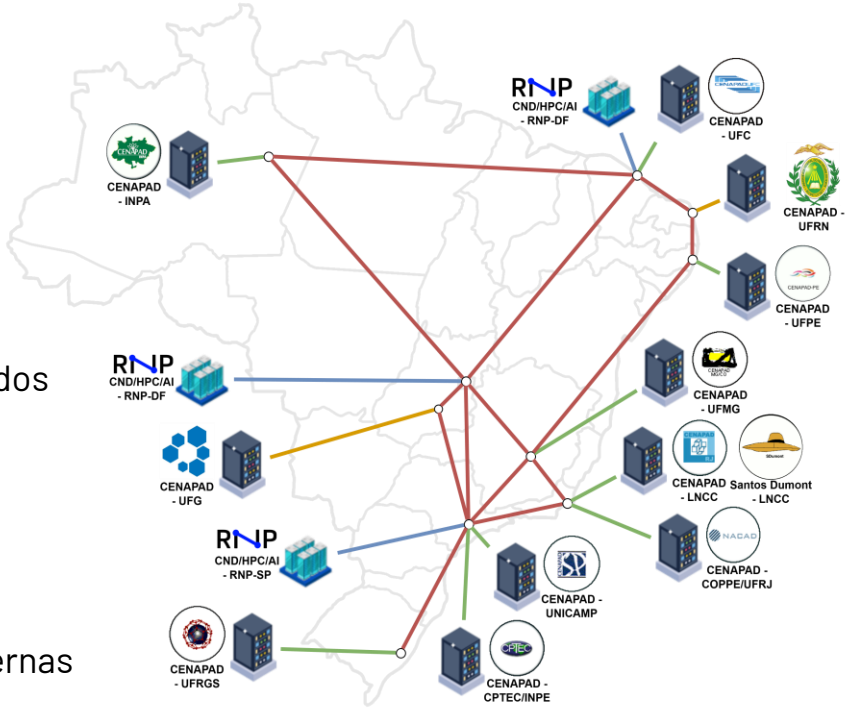
Rede de E-Ciência para IA

Interliga centros de supercomputação, CNDs e repositórios de dados

- movimentação otimizada dos dados

Fornece acesso seguro à Internet

- IA como Serviço (IAaaS)
- inferência em tempo real
- ingestão de dados de fontes externas



27° Workshop RNP

Diferenças entre as Redes da RNP

Rede Ipê	Rede de e-Ciência	Rede de e-Ciência para IA
+1.300 pontos conectados	+20 pontos conectados	**+30** pontos conectados
Com acesso aberto à Internet	Sem acesso à Internet. Tráfego de dados liberado entre instituições específicas	**Acesso seguro e restrito para serviços de IA à Internet**. Tráfego de dados liberado entre instituições específicas
Acessível globalmente a partir de qualquer ponto de conexão à Internet	Acessível apenas a partir dos servidores de transferência de dados homologados das instituições participantes da rede	**Acesso controlado à serviços de IA. Acesso restrito para transferência otimizadas** a partir dos servidores homologados das instituições participantes da rede
Conectividade até a borda da instituição	Conectividade da RNP até o servidor de armazenamento científico da instituição	Conectividade da RNP até o servidor de armazenamento científico da instituição
Não otimizada para transferência de dados. Configuração padrão para aplicações convencionais (e-mail, vídeo, serviços online)	Nativamente otimizada para transferências de dados com alto desempenho	Nativamente otimizada para transferências de dados com alto desempenho **e serviços para IA**
Disponível para todas as organizações usuárias do Sistema RNP	Requer o cumprimento de requisitos técnicos e políticas de segurança para a instituição fazer parte da rede	Requer o cumprimento de requisitos técnicos e políticas de segurança para a instituição fazer parte da rede

27° Workshop RNP

Requisitos de Rede para IA

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

27 Workshop RNP

Obrigado!

jose.rezende@rnp.br

27° Workshop RNP

SINAPAD e Infraestrutura Federada: uma perspectiva histórica

Antônio Tadeu Gomes

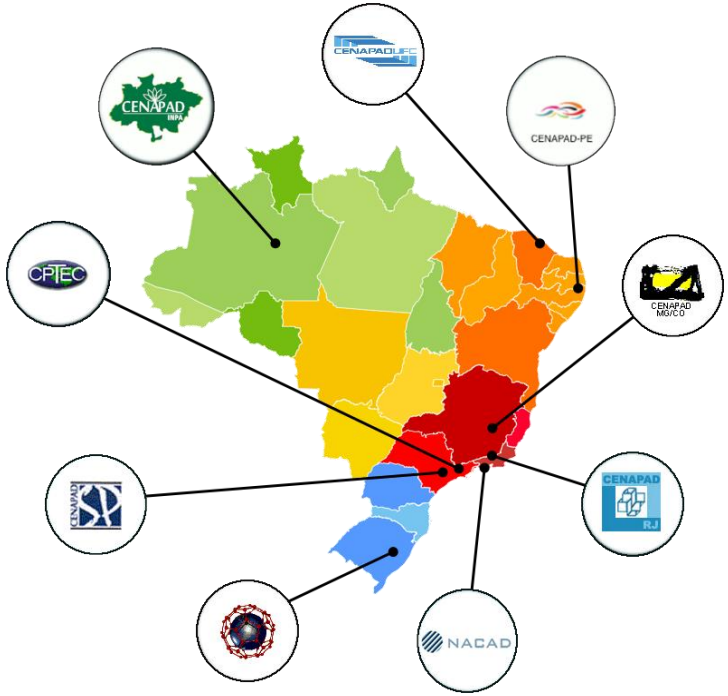
Pesquisador LNCC

27 Workshop RNP

Origens

Forte ligação com a criação da RNP

- Interconexão entre centros de supercomputação
- Aproximação entre infraestrutura e usuários finais



27 Workshop RNP

Origens

Fragilidade na governança

- Modelo financeiro
- Papel do MCTI
- Autonomia dos centros

27° Workshop RNP

Janela de oportunidade

Parceria Brasil-França (~2010)

- Revitalização do SINAPAD como "apêndice estratégico"
- Início do desenho do modelo de *tiers*

Brazil France Workshop

Cooperation for the establishment a petaflop HPC system under the SINAPAD framework

Date: Monday and Tuesday, December 14 and 15
Location: PETROPOLIS – RJ (LNCC Audithorium)

Agenda:

Part A: 1st Day Afternoon

12h30	14h00	Lunch	All
14h00	14h30	Welcome and introduction of the participants	Ademar Seabra Augusto C. Gadelha
14h30	14h45	Introduction of the workshop: summary of last contacts, objectives of the workshop	Didier Lamouche
14h45	15h30	Overview of HPC situation in Brazil	Pedro Dias
15h30	16h00	What is at stake with HPC	JF Lavignon
16h00	16h30	Bull added value in HPC	JF Lavignon
16h30	18h30	An example of industry – R&D HPC centre : CCRT	Catherine Rivière Jean Gonnord
18h30	19h00	Wrap up First Day	Pedro Dias, Ademar Seabra and All

27° Workshop RNP

Janela de oportunidade

Mobilização da comunidade de RESD

- Base tecnológica: sistema de arquivos distribuído (projeto PADBR)
- Workshop do SBAC-PAD em 2009
- Montagem da infraestrutura de portais do SINAPAD: GT-mc2

NOTÍCIAS LNCC

Aplicações científicas recebem suporte de projeto conjunto entre SINAPAD e RNP

Publicado em: 06/05/2013,00:00

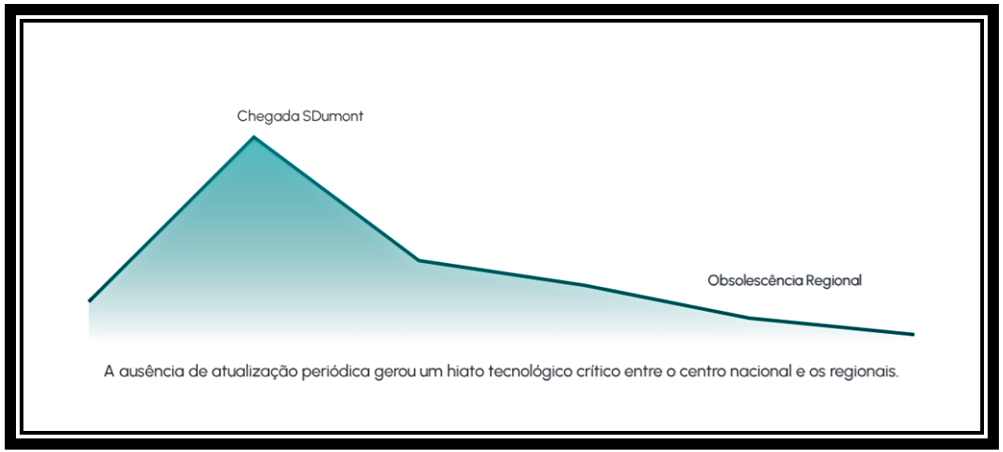
Aplicações científicas recebem suporte de projeto conjunto entre SINAPAD e RNP Projeto coordenado pelos professores Antônio Tadeu Gomes do LNCC e Francisco Brasileiro da UFCG é apresentado no 14o WRNP, que ocorre dias 6 e 7/5 em Brasília. Veja a programação em http://wrnp.rnp.br/programacao O projeto GT-mc² (Minha Cloud Científica), financiado pela RNP e com participação da UFBA, UFC e UFRGS, recebe também apoio do SINAPAD e da PUC-Rio. O objetivo desse projeto é implantar uma plataforma de computação na nuvem voltada para aplicações de e-ciência. Essa plataforma provê facilidades de: (i) acesso, em diferentes perfis de permissão, a uma variedade de recursos computacionais para a execução de experimentos, (ii) armazenamento, compartilhamento e publicação de resultados de experimentos, (iii) reprodutibilidade de experimentos, e (iv) controle de proveniência dos dados consumidos e gerados pelos experimentos. A plataforma segue um modelo PaaS (Platform-as-a-Service), permitindo o fácil desenvolvimento e implantação de ambientes de usuário personalizados, oferecidos em um modelo SaaS (Software-as-a- Service). A pilha de serviços também inclui um broker para mediar a alocação de recursos em diferentes tipos de provedores de IaaS (Infrastructure-as-a-Service), atendendo requisitos de segurança específicos de cada aplicação e demandas de carga de trabalho típicas de aplicações de e-ciência. Confira a apresentação do GT-mc² hoje 6/5 a partir das 10:40h em http://mconf.org/events/wrnp2013/ Para mais detalhes sobre o GT-mc², confira em http://gtmcc.lncc.br/

27 Workshop RNP

Marco: SDumont

Sucesso parcial da parceria Brasil-França

- Desequilíbrio no modelo de *tiers*
- Desconforto entre centros preteridos
- Inviabilidade técnica da federação plena



27° Workshop RNP

IA: "Killer application"

Característica transversal

- Demanda massiva de dados E de processamento
- Adoção multi-setorial
- Motor para uma nova economia digital

Obrigado!

atagomes@lncc.br

27° Workshop RNP

Infraestrutura Federada de HPC/IA

José Rezende (RNP)

Antônio Tadeu (LNCC)

Michelle Wangham (RNP)

27° Workshop RNP

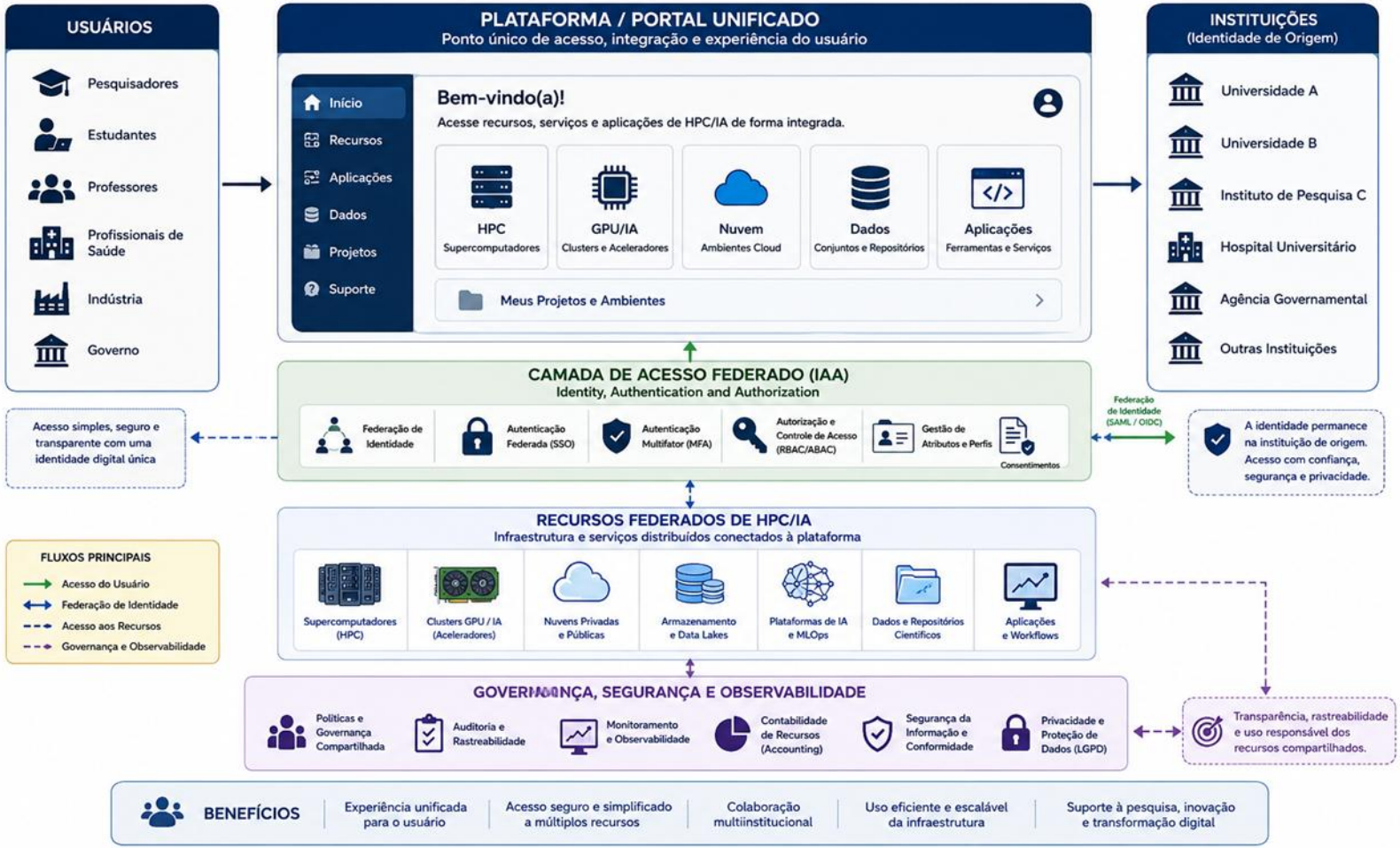
Plataforma/Portal Unificado e Acesso Federado (IAA) (esboço)

Michelle Wangham (RNP)

27 Workshop RNP

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

27° Workshop RNP

Iniciativas internacionais (EU e EUA)

EuroHPC Joint Undertaking

Objetivo: fortalecer a autonomia (soberania) tecnológica europeia por meio da criação de um ecossistema integrado de:

- supercomputadores
- fábricas de IA (AI Factories)
- plataformas federadas de HPC e IA
- armazenamento e espaços de dados científicos
- infraestrutura federada para experimentação
- formação de recursos humanos e inovação

Pilar estratégico:

- acesso federado baseado em identidade digital institucional

27 Workshop RNP

Iniciativas internacionais (EU e EUA)

EuroHPC Federation Platform (EFP)

Camada de AAI federada, inspirada no modelo acadêmico europeu operado pela GÉANT e pelo eduGAIN

- Single Sign-On federado para HPC e IA
- Uso do eduGAIN como *trust fabric*
- MyAccessID como **camada de harmonização**
  - proxy federado de identidade (interoperabilidade)
  - broker de autenticação
  - camada de confiança
  - harmonizador de atributos
  - federação precisa híbrida e *multi-stakeholder*
  - MFA e credenciais efêmeras (não se limita a login web)
  - **autorização federada**

27° Workshop RNP

Iniciativas internacionais (EU e EUA)

NAIR (National Artificial Intelligence Research Resource)

Objetivo: democratizar o acesso à infraestrutura avançada de IA para pesquisa, educação e inovação (*broker* nacional)

- compartilhamento federado de recursos de IA;
- acesso coordenado nacionalmente;
- parcerias público-privadas;
- integração de laboratórios nacionais;
- **acesso remoto e seguro a infraestruturas distribuídas**

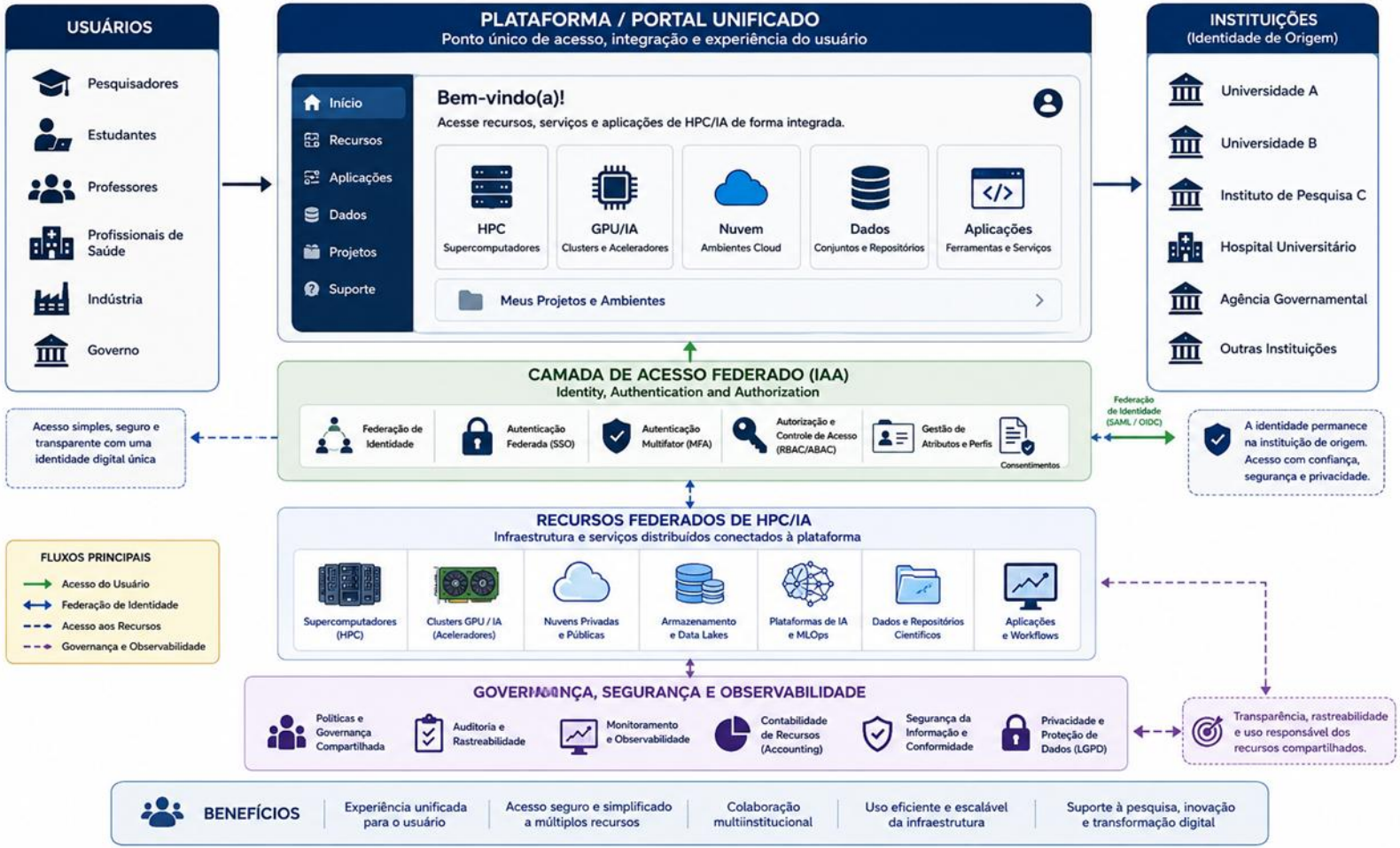
Diferença em relação ao euroHPC:

- foco em liderança de inovação e escala de IA
- **CILogon**: NSF ACCESS,Open Science Grid,

27 Workshop RNP

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

27 Workshop RNP

Obrigada!

michelle.wangham@rnp.br