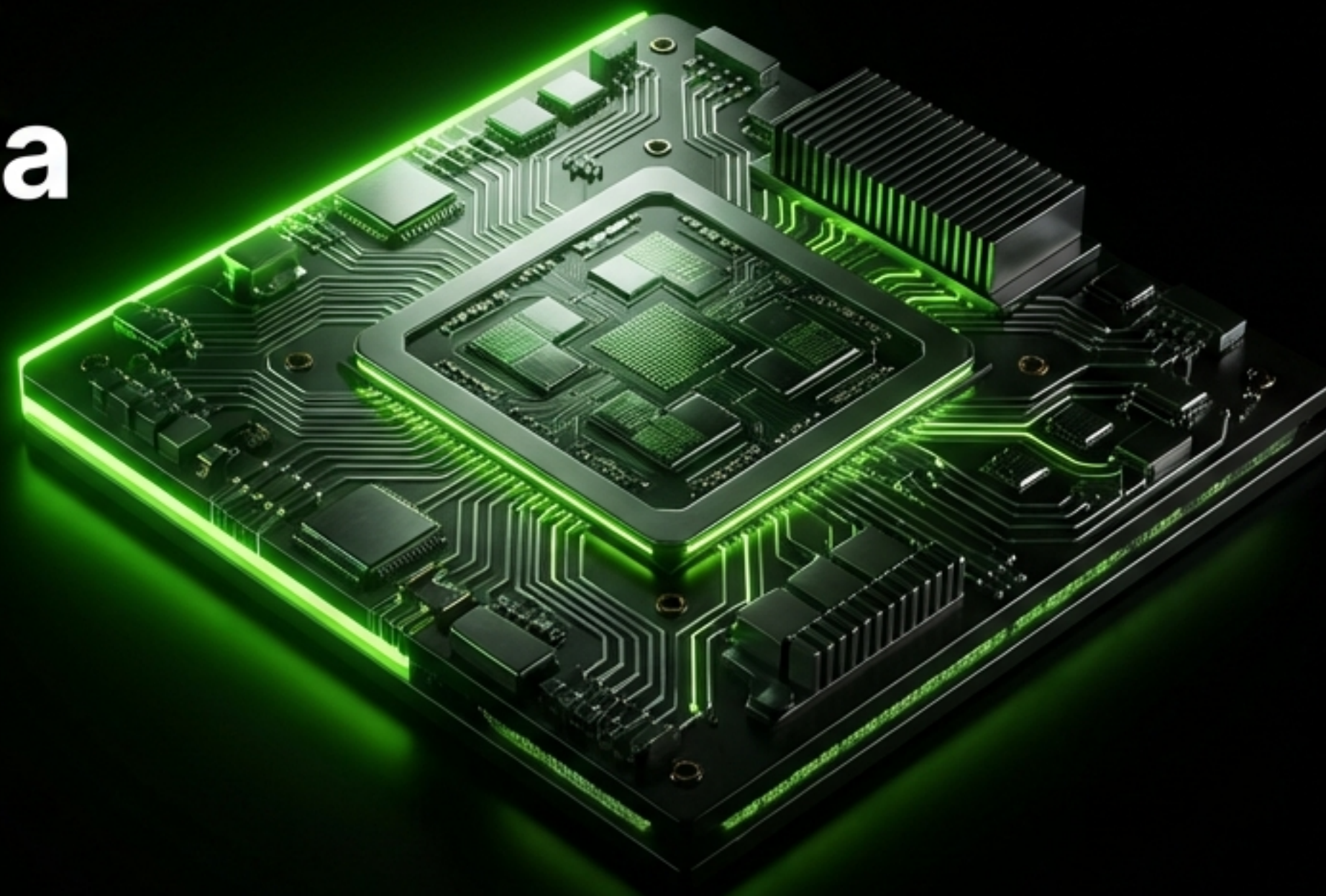


Construindo Sua AI Factory: O Guia Definitivo de Infraestrutura

Uma abordagem Full-Stack para
escalar Inteligência Artificial —
do Hardware à Simulação Física.



Baseado nas soluções Nvidia DGX, Run:ai, Omniverse e AI Enterprise.



A IA Não é Apenas Software — É Um Desafio de Infraestrutura



Latência de Rede

Dados presos em congestionamentos impedem o treinamento eficiente.



Ociosidade de Computação

GPUs poderosas ficam subutilizadas sem orquestração dinâmica.

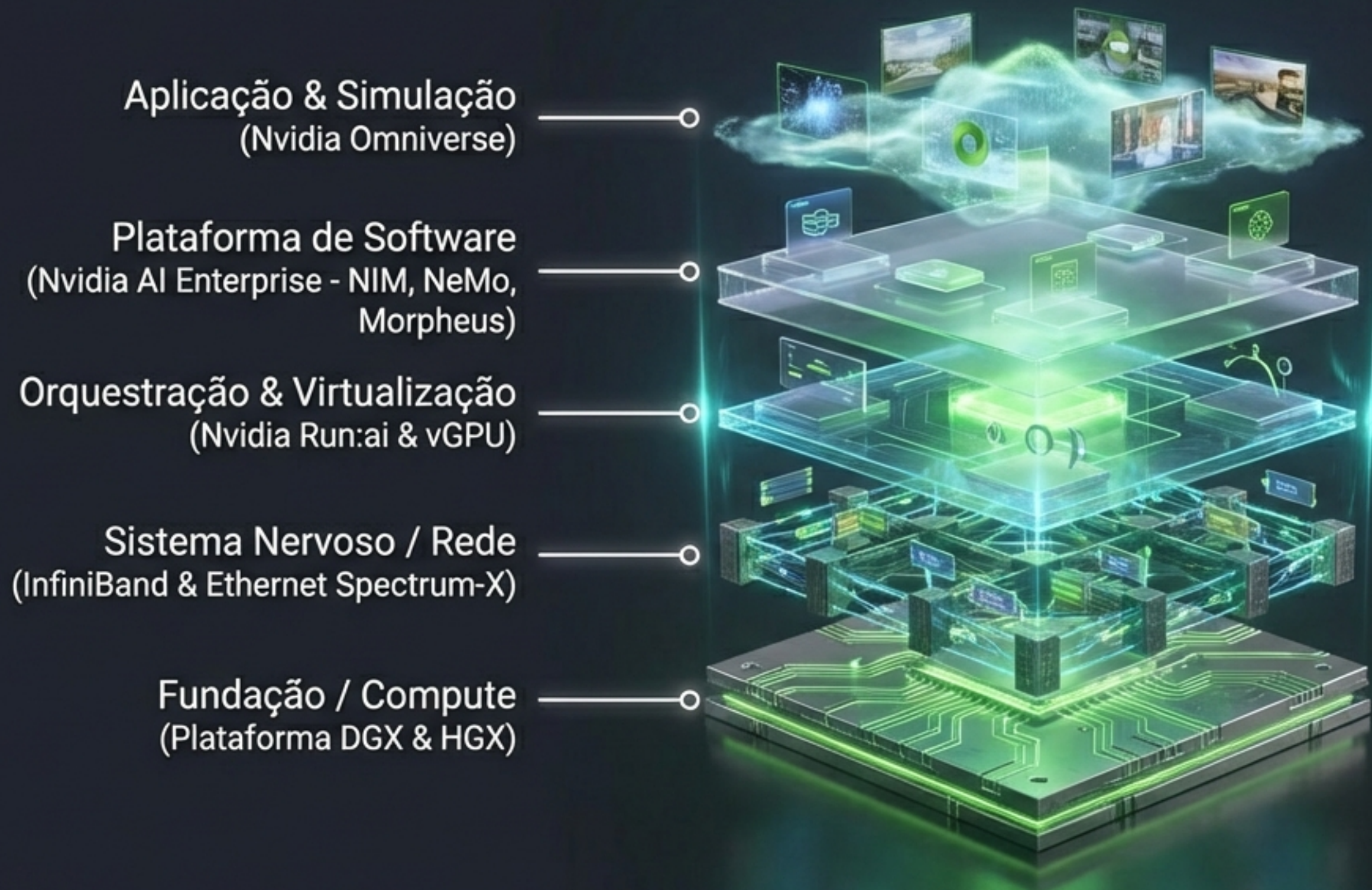


Silos de Dados

Dificuldade em conectar o mundo digital ao físico (Gêmeos Digitais).

Insight: Para vencer na era da inteligência, as empresas precisam evoluir de servidores isolados para **Fábricas** de IA integradas.

O Blueprint da AI Factory: Uma Arquitetura de Pilha Completa



Uma visão integrada onde cada camada acelera a superior, garantindo segurança, segurança, escala e eficiência operacional.



A Fundação de Computação: Plataforma NVIDIA DGX

O padrão definitivo para a IA empresarial.

DGX SuperPOD

Infraestrutura de nível de liderança para escala massiva. Transforma data centers em fábricas de inovação.

DGX BasePOD

Arquitetura de referência comprovada para infraestrutura empresarial, ideal para iniciar a jornada com parceiros de armazenamento líderes.

10 das 10 principais fabricantes de automóveis globais.
7 das 10 principais empresas farmacêuticas globais.

Projetado para executar o NVIDIA AI Enterprise.

O Sistema Nervoso: A Rede Define o Data Center

A conectividade acelerada é vital para RAG, treinamento de modelos e inferência distribuída.

NVIDIA Quantum InfiniBand

Desempenho máximo para supercomputação e treinamento de IA em larga escala.

NVIDIA Spectrum-X (Ethernet)

A primeira plataforma Ethernet construída especificamente para IA. Ideal para nuvens de IA e infraestrutura empresarial.

Acelerador de Infraestrutura: NVIDIA BlueField (DPU)

Descarrega, isola e acelera tarefas de infraestrutura e segurança.

Democratizando o Acesso: Virtualização de GPU (vGPU)

De VDI a AI – Potência em qualquer lugar.

- **Densidade:** Suporte para até 32 computadores virtuais por GPU física.
- **Flexibilidade:** Perfis ajustáveis para designers gráficos e cientistas de dados.
- **Migração Dinâmica:** Movimentação de VMs sem interrupção.



NVIDIA RTX vDWS:
Para workstations
criativas.

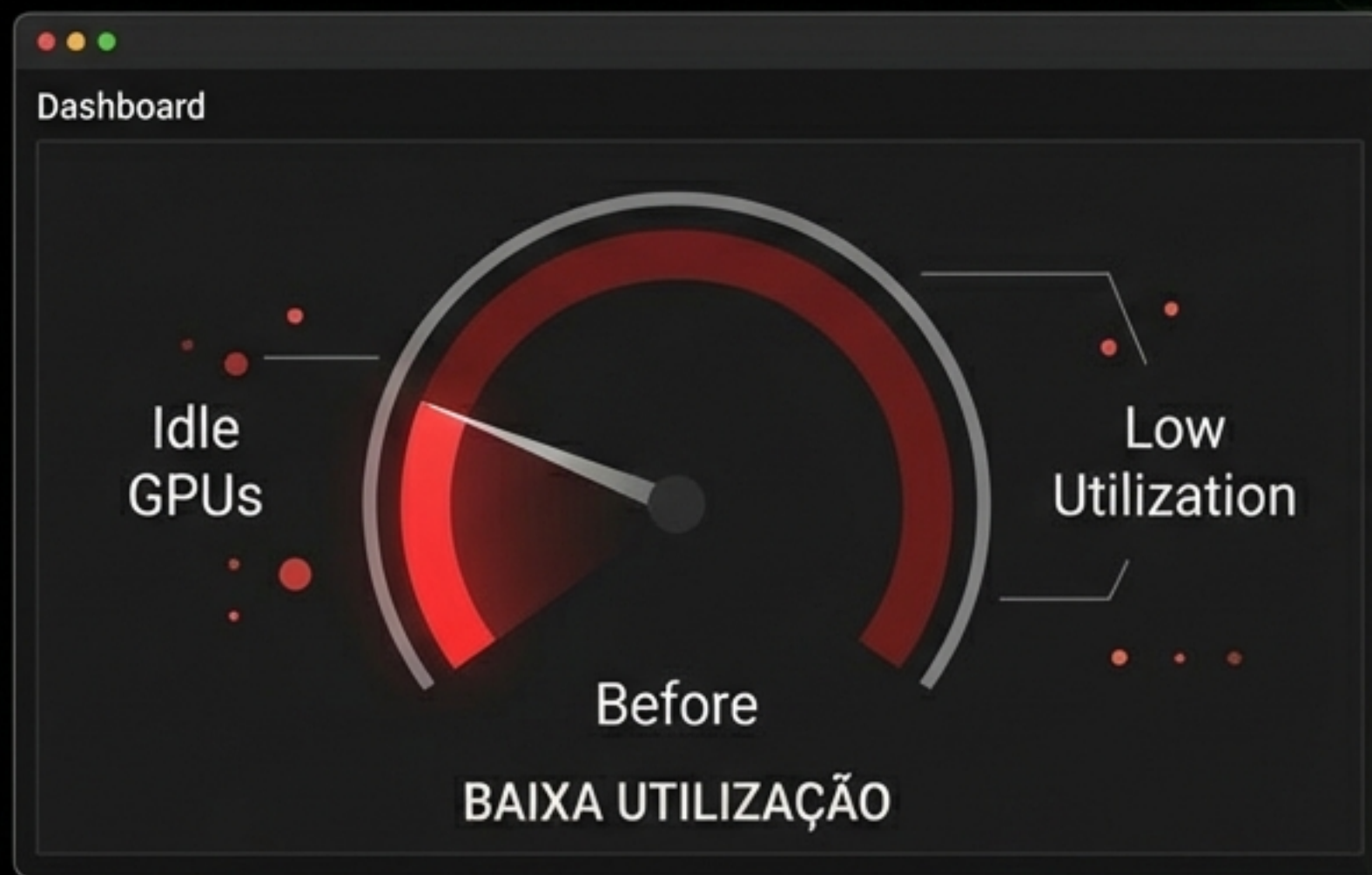


**NVIDIA Virtual Compute
Server:** Para cargas de
IA e Deep Learning.



Eficiência Máxima com Orquestração Dinâmica: **NVIDIA Run:ai**

Elimine GPUs ociosas e gargalos de infraestrutura.



Orquestração inteligente que reúne recursos em ambientes híbridos e aloca dinamicamente baseada em prioridades.



10x

Disponibilidade da GPU

5x

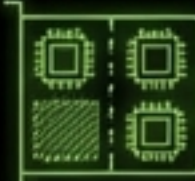
Utilização da GPU

20x

Cargas simultâneas

0

Intervenção manual



Recurso "Fractional GPU" — múltiplas cargas de trabalho compartilham uma única GPU com segurança.

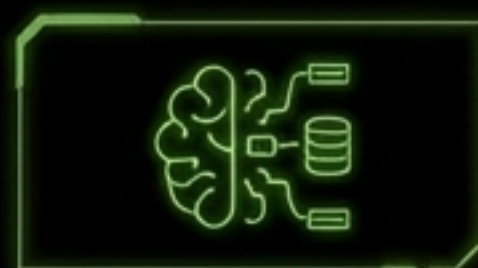
O Sistema Operacional da IA: NVIDIA AI Enterprise

Pacote nativo da nuvem de bibliotecas e frameworks que acelera o desenvolvimento e garante a segurança.



NVIDIA NIM

Microsserviços de inferência pré-construídos.



NVIDIA NeMo

Framework para IA generativa e RAG.



Segurança Empresarial

Verificação de CVEs e estabilidade de API.



Desenvolva uma vez, execute em qualquer lugar.

A Era da IA Baseada em Agentes e Raciocínio

Transição de chatbots simples para agentes que raciocinam, planejam e agem.

Principais Tecnologias

- **Modelos Llama Nemotron:** Destilados para alta capacidade de raciocínio.
- **NVIDIA Blueprints:** Workflows de referência personalizáveis.

Exemplos de Uso

- RAG Empresarial
- Analista de Segurança (via Morpheus)
- Humanos Digitais



IA Física e Gêmeos Digitais: **NVIDIA Omniverse**

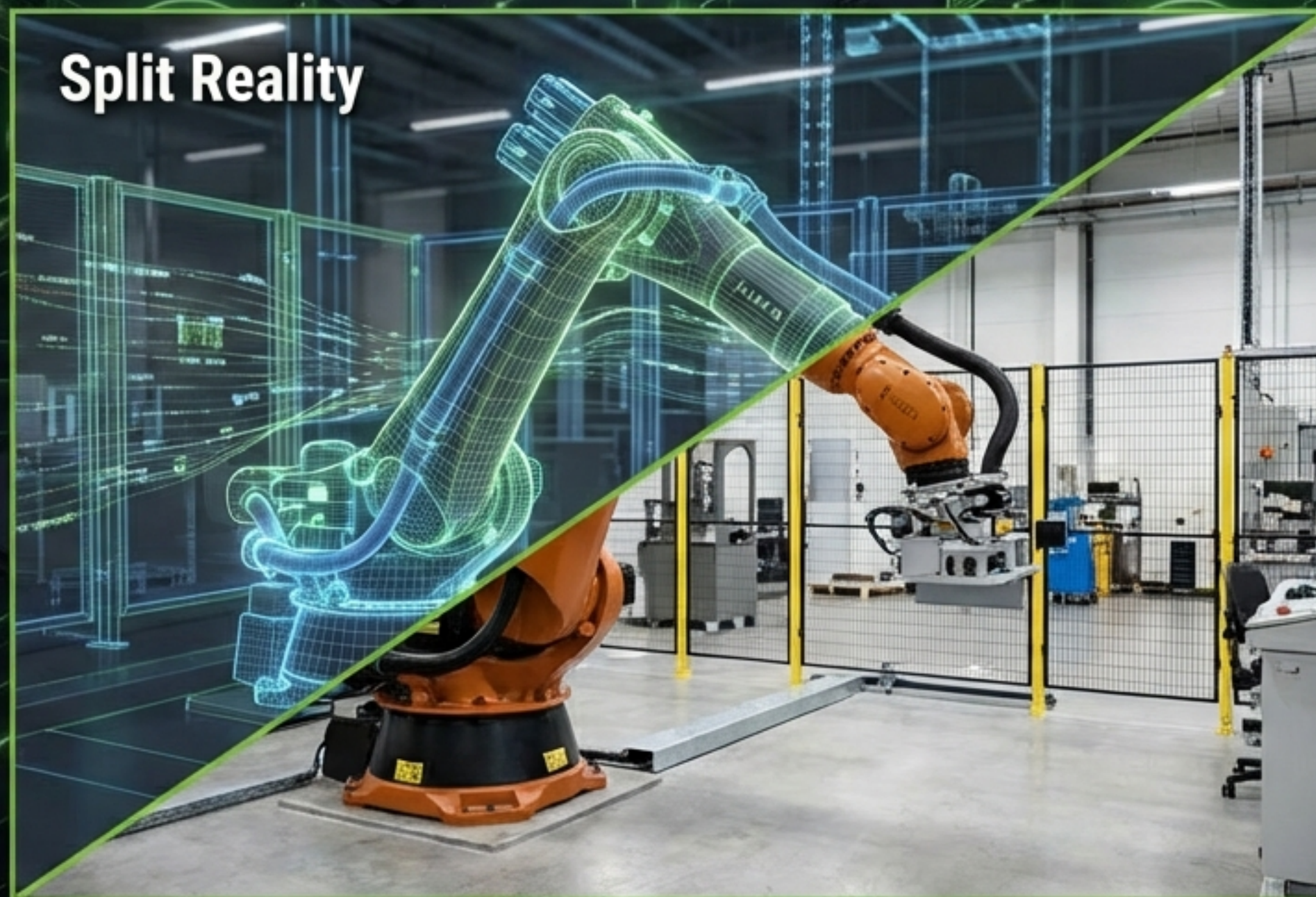
Plataforma para desenvolver aplicações de OpenUSD e simular o mundo físico.

Key Applications

- **Treinamento de Robôs:** Isaac Lab para ensino em simulação.
- **Dados Sintéticos:** NVIDIA Cosmos para criar cenários de treino.
- **Fábricas Digitais:** O metaverso industrial.



A simulação precede a implantação física para reduzir custos e riscos operacionais.



Transformação Real: Casos de Uso por Indústria



Automotivo

BMW: Aumento de 8x na produtividade de qualidade com dados sintéticos e DGX.



Saúde

Amgen: Uso do DGX Cloud e BioNeMo para acelerar descoberta de medicamentos.



Energia

Shell: Modelagem de reservatórios e materiais sustentáveis com DGX.




Defesa

Lockheed Martin: AI Factory centralizada para segurança de dados.

Modelos de Implantação: Flexibilidade Total



* Compatibilidade de desenvolvimento entre clouds e local.



O Futuro da Computação: Arquitetura Blackwell

O motor da nova revolução industrial.

- Preparada para modelos de trilhões de parâmetros.
- IA generativa em tempo real.
- **NVIDIA GB200 NVL72**: Escala massiva.
- **DGX Quantum**: Fusão de supercomputação quântica e clássica.

Um Ecossistema Integrado de Sucesso

Parceiros de Armazenamento

Validado para
DGX BasePOD



Colocation

DGX-Ready
Data Centers




nVIDIA.
AI Factory

Software & MLOps

Integrações certificadas
AI Enterprise



Serviços Gerenciados

Foco na inovação,
não na operação



Comece a Construir Sua AI Factory Hoje

A infraestrutura certa transforma a IA de experimento em vantagem competitiva.

1. **Avalie:** Explore os NVIDIA Blueprints para seu caso de uso.
2. **Experimente:** Solicite uma Licença de 90 Dias do NVIDIA AI Enterprise.
3. **Teste:** Acesse o NVIDIA LaunchPad ou DGX Cloud.



Saiba mais clicando aqui!

