



NVIDIA ENTERPRISE

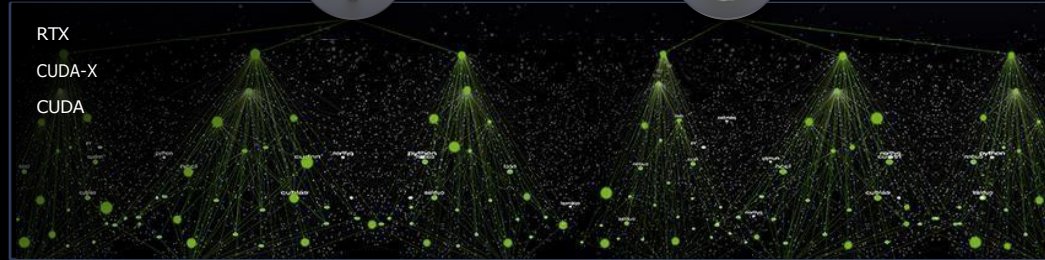
APPLICATION
FRAMEWORKS



PLATFORM



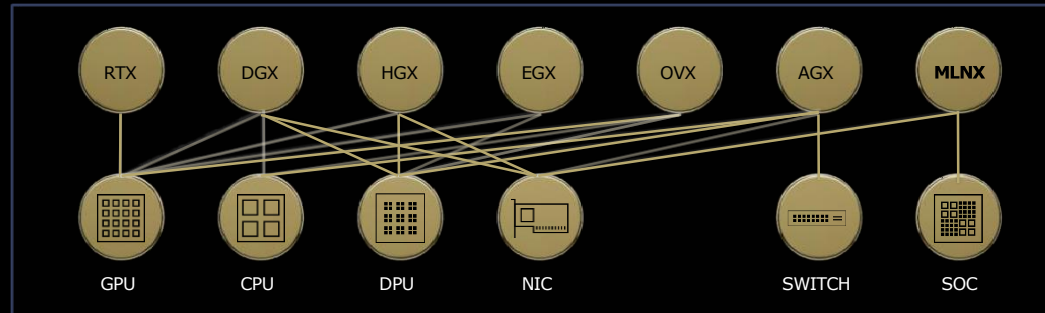
ACCELERATION
LIBRARIES



SYSTEM
SOFTWARE

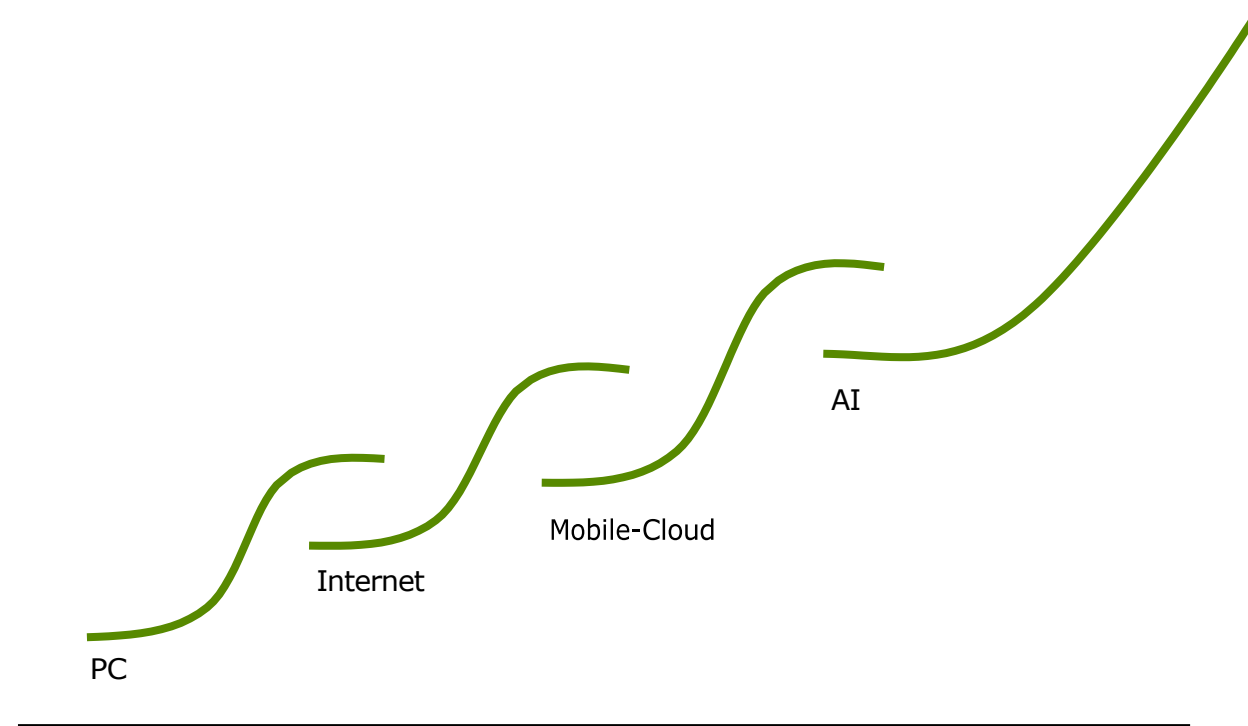


HARDWARE

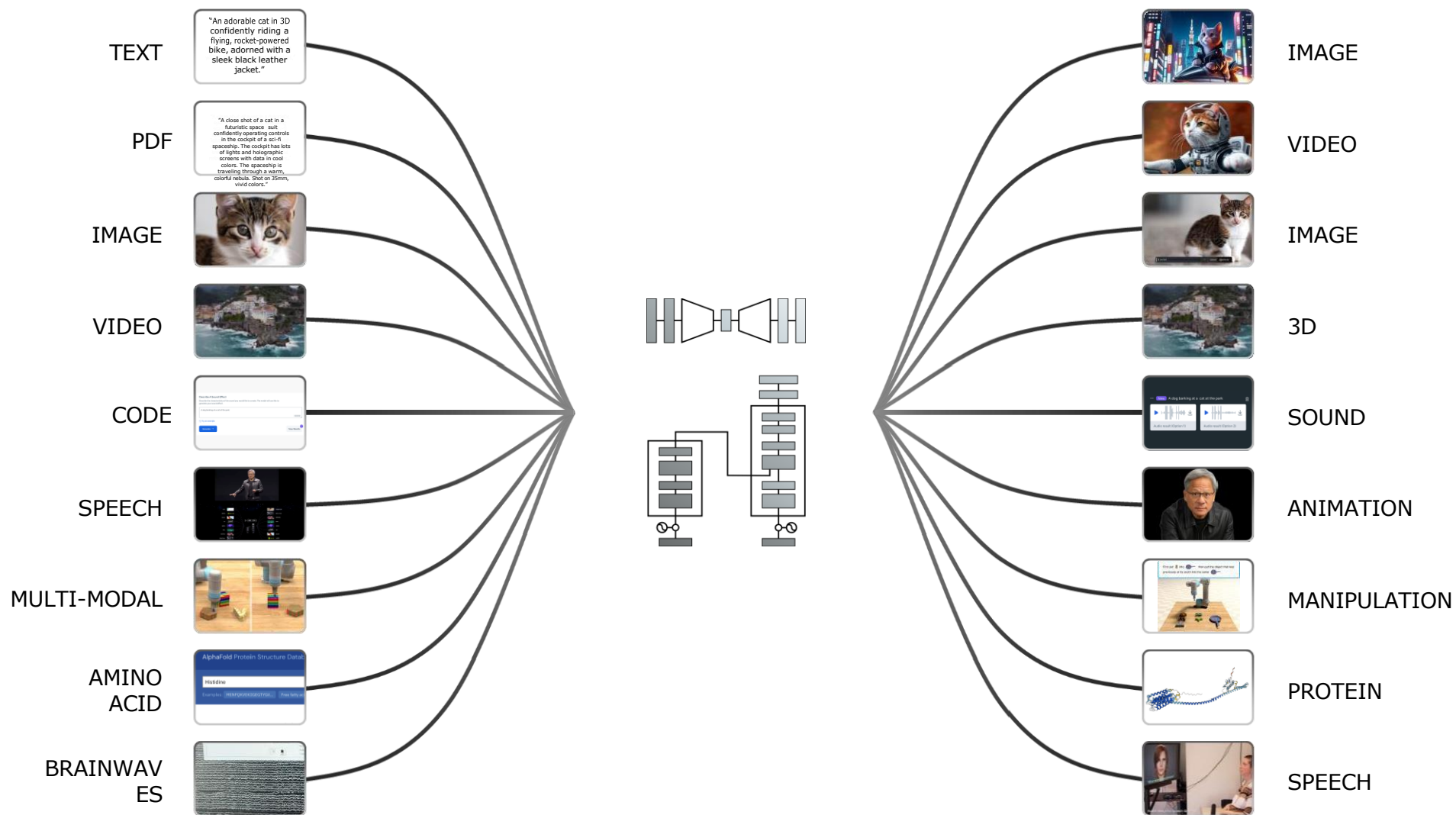


Generative AI

The iPhone moment of AI



Generative AI Can Learn and Understand Everything



ChatGPT Launch is the “iPhone Moment” for Generative AI

MIT
Technology
Review

Featured Topics Newsletters Events Podcasts

Sign in

Subscribe

TR EXPLAINS

ChatGPT is everywhere. Here's where it came from

OpenAI's breakout hit was an overnight sensation—but it is built on decades of research.

Events Special Issues

VentureBeat

Q

Subscribe

GamesBeat

Artificial Intelligence ▾

Security ▾

Data Infrastructure ▾

Automation ▾

Enterprise Analytics ▾

More ▾

Nvidia will bring AI to every industry, says CEO Jensen Huang in GTC keynote: 'We are at the iPhone moment of AI'

Sharon Goldman
@sharongoldman

March 21, 2023 9:36 AM

f t in

CNBC

Search quotes, news & videos

Q

WATCHLIST |

MARKETS BUSINESS INVESTING TECH POLITICS CNBC TV INVESTING CLUB PRO

MAKE IT SELECT

PRO

As ChatGPT mania takes hold, here's how companies beyond tech are using A.I.

FIERCE
Electronics

Electronics IoT & Wireless Sensors Embedded

Resources

Subscribe

Q

in

t

f

e

o

SENSORS

Update: ChatGPT runs 10K Nvidia training GPUs with potential for thousands more

By Matt Hamblen • Feb 11, 2023 11:29am

chatbots

ChatGPT

NVIDIA

GPU

NVIDIA

Home » Resources » News

NEWS

NVIDIA Delivers DGX-1 Supercomputer in a Box to OpenAI

Aug. 15, 2016

by [Jen-Hsun Huang](#)

OpenAI, a non-profit research company devoted to advancing artificial intelligence, has become one of the proud owners of a DGX-1, NVIDIA's so-called "supercomputer in a box," a server specifically designed for machine learning work. The system, which was hand-delivered to the company's headquarters in San Francisco by NVIDIA CEO Jen-Hsun Huang, will be used to run some of OpenAI's most computationally challenging applications.



Generative AI Impacts Every Function in a Bank

ENTERPRISE SEARCH

Optimizes information retrieval by evaluating multiple sources, and summarizing results

AML/KYC, TRANSACTION FRAUD

Improves accuracy and generates reports, reducing investigations and compliance risk

HPC PRICING & RISK

Summarizes news feeds, improving risk management and optimizing reserves

ALGORITHMIC TRADING

Summarizes real time data streams accelerating time to insight, improving market returns

DOCUMENT MANAGEMENT

Summarization and report generation will optimize middle/back office workflows

CUSTOMER SERVICE

Expands knowledge base beyond scripted answers increasing customer satisfaction

MARKETING

Creates personalized ad copy, email, etc. increasing click to conversion

WEALTH MANAGEMENT

Utilizes various data inputs to personalize wealth management plans to drive AUM

Generative AI Impacts Every Function in Retail

PERSONALIZED SEARCH

Multi-modal search with natural language understanding to deliver better recommendations

SHOPPING ADVISOR

Personalized recommendations and conversation style tailored to each customer's needs

PRODUCT DESCRIPTION GENERATOR

Generating catalogue information like descriptions, titles, and product images

EMPLOYEE PRODUCTIVITY

Tools like operator manual chatbots and translators to assist employees

PRODUCT DESIGN GENERATION

From ideation to creation, simulating products in various environments and use cases

CUSTOMER SERVICE

Using chatbots for simple inquiries and routing complex ones to human agents for improved, more efficient customer service

MARKETING & SENTIMENT ANALYSIS

Creating personalized ad copy, email, etc. increasing click to conversion

ADVERTISING

Generating personalized advertisements or product images

Enterprises are on the Generative AI Journey



Explosion

ChatGPT gets announced late in 2022, gaining over 100 million users in just two months. Users of all levels can experience AI and feel the benefits firsthand.



Experimentation

Enterprise application developers kick off POCs for generative AI applications with API services and open models including Llama 2, Mistral, NVIDIA, and others.

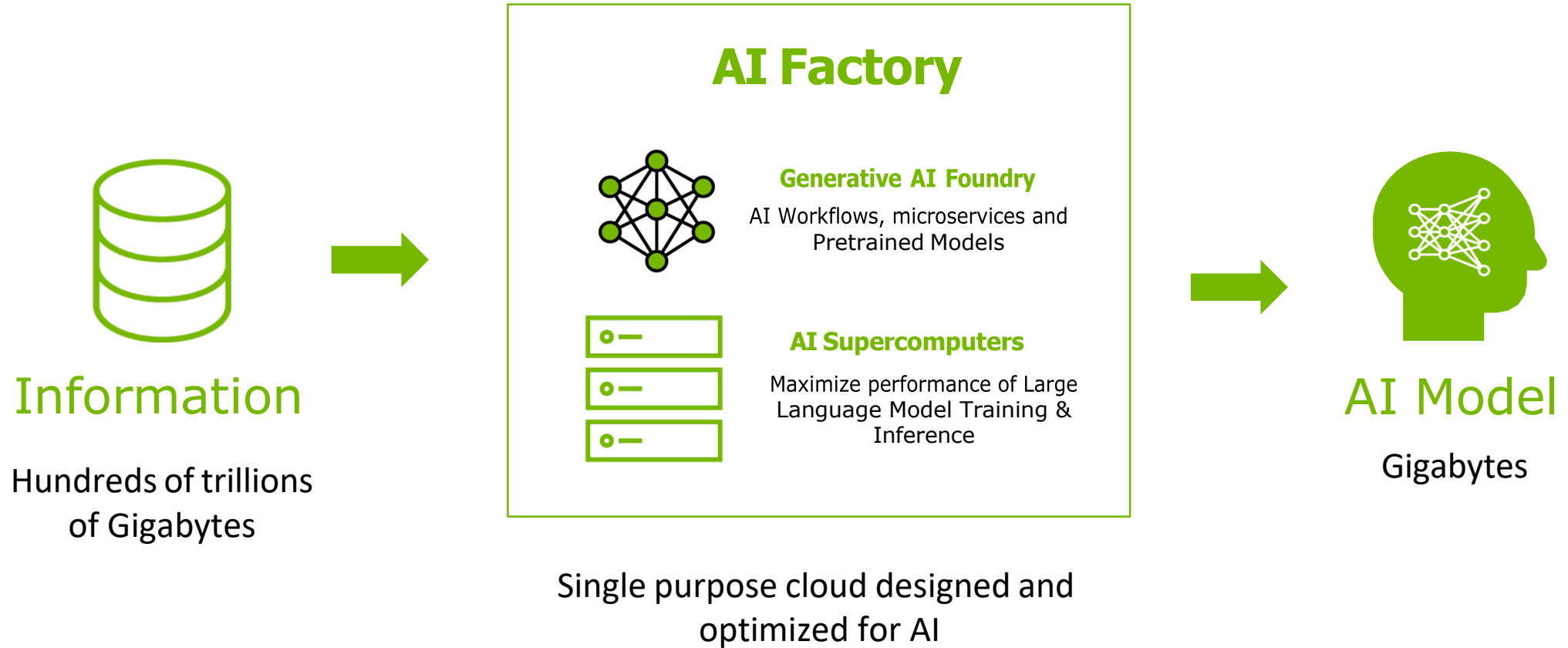


Production

Organizations have set aside budget and are ramping up efforts to build accelerated infrastructure to support generative AI in production.

AI Factory

Enabling Intelligence Manufacturing



Sovereign AI Factory

Demand waves

1

Startups



Provide regional innovators with access to AI Platform and Infrastructure

2

Nations



Infrastructure for research, upskill talent, competitive advantage

3

Enterprises



Racing to turn proprietary data into enhanced AI products and services

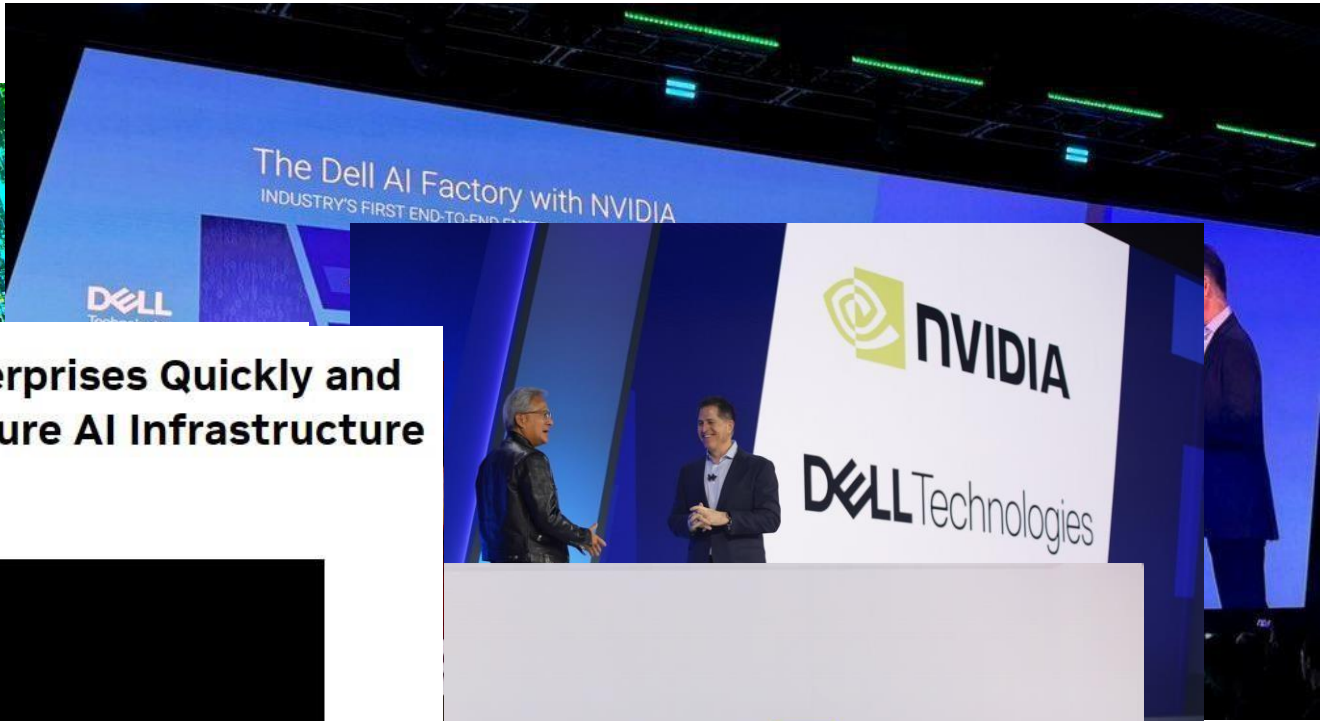
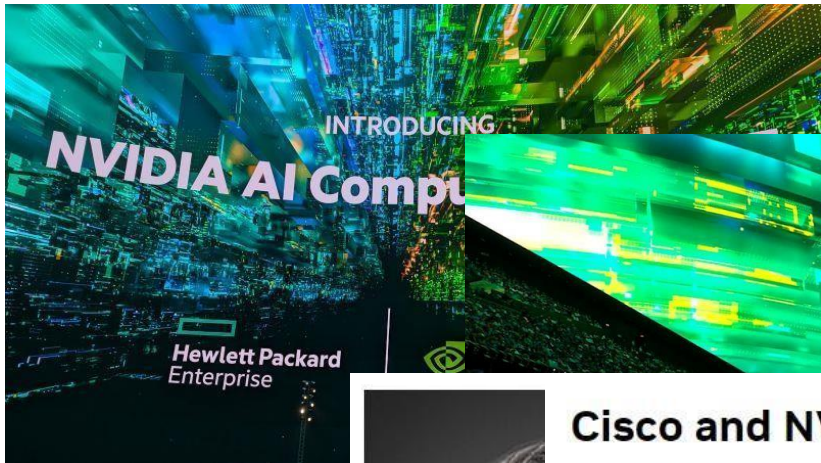
4

Generative AI Apps



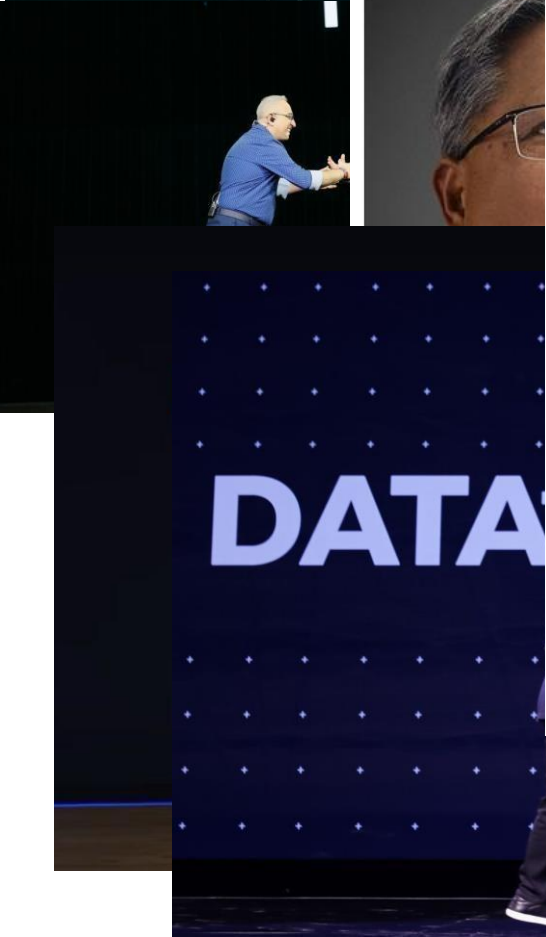
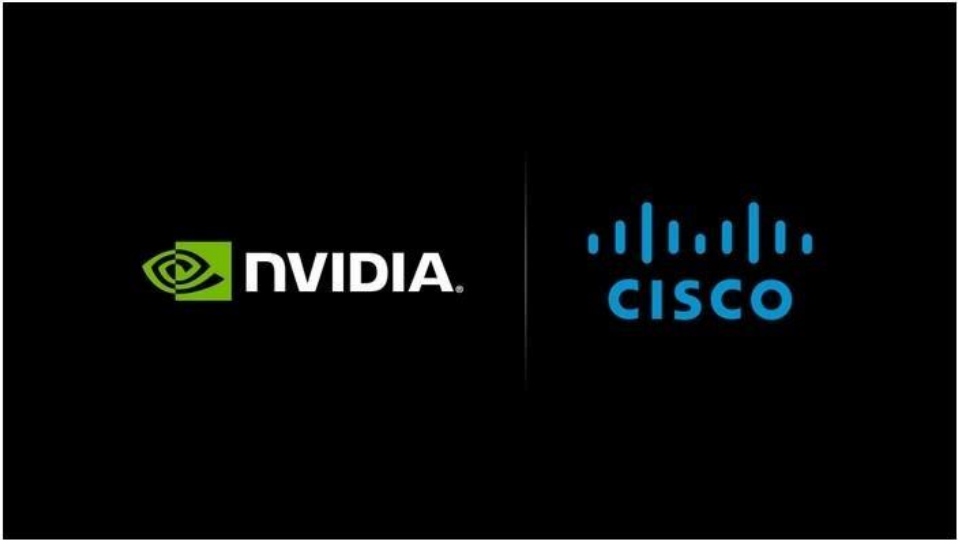
Inference is a new infrastructure business as every app adopts generative AI

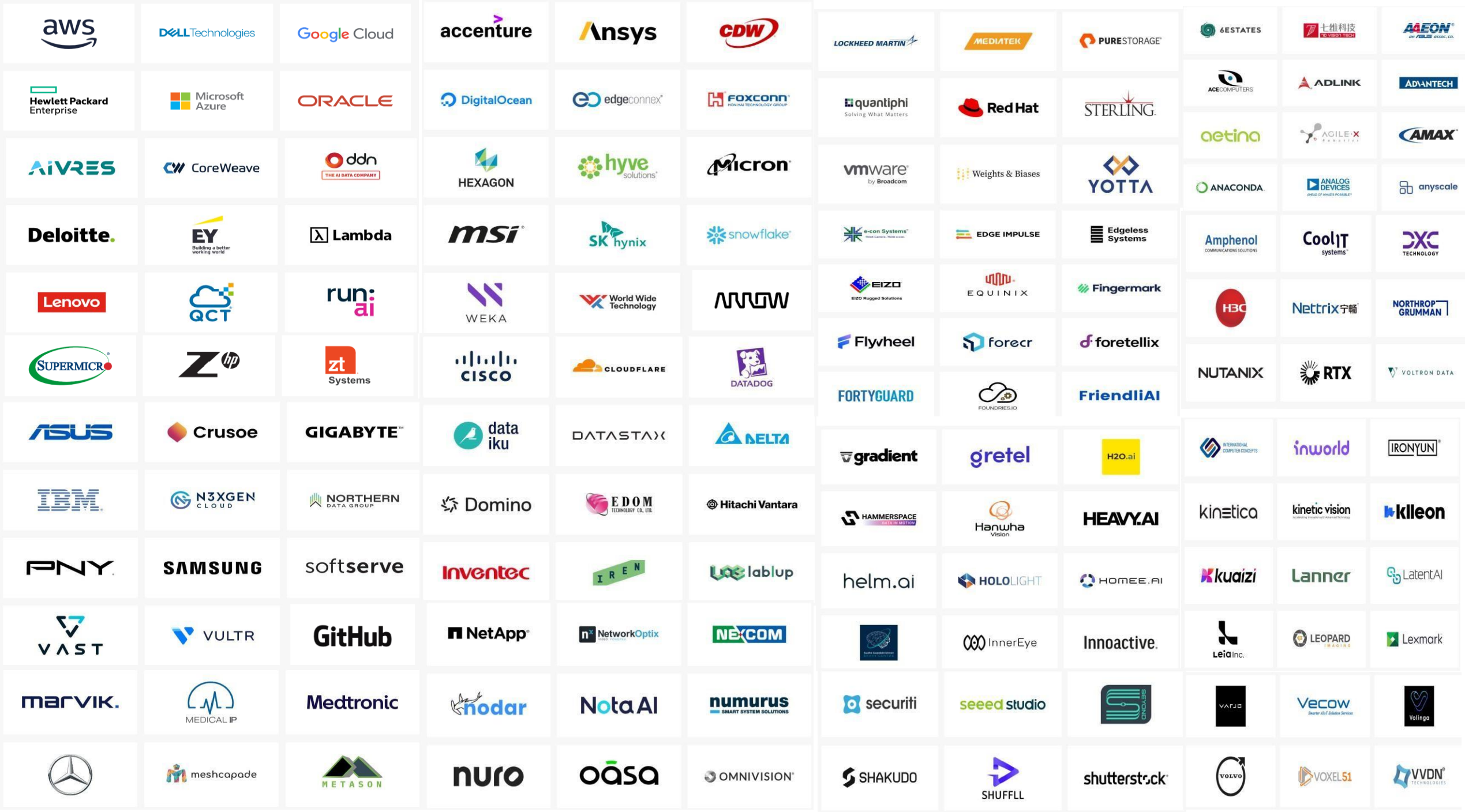
SOVEREIGN AI FACTORY



Cisco and NVIDIA to Help Enterprises Quickly and Easily Deploy and Manage Secure AI Infrastructure

February 6, 2024







Marcel Saraiva msaraiva@nvidia.com

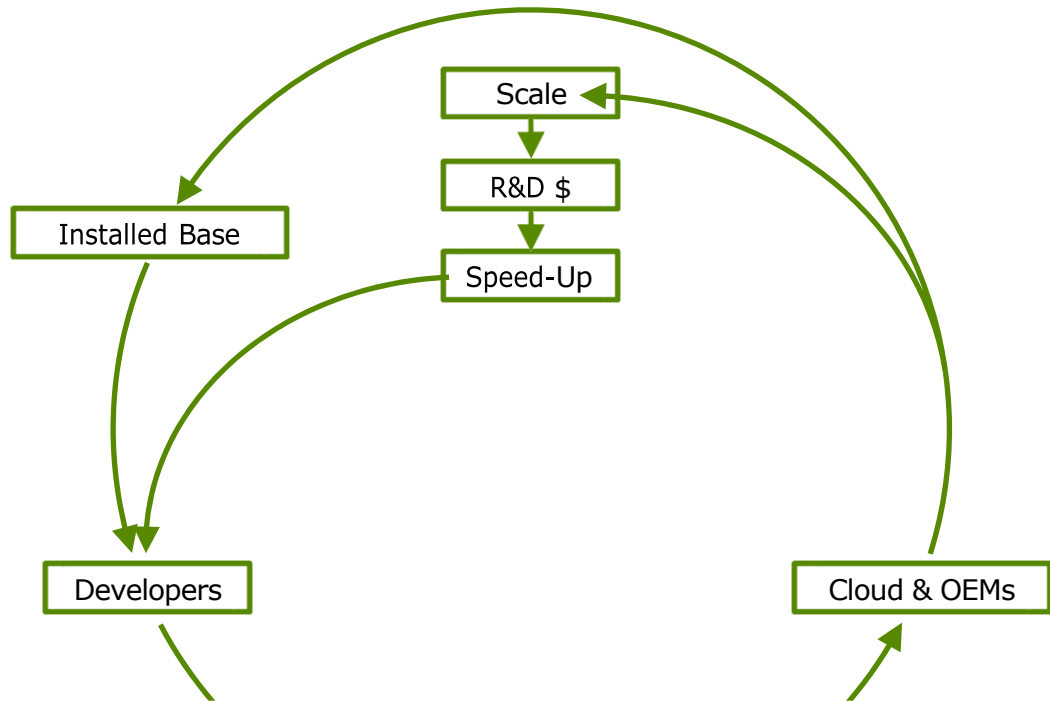


Gicon IT Partners & Solutions

Gilmar Pereira – gilmarpereira@giconit.com.br

OBRIGADO!

NVIDIA Accelerated Computing Virtuous Cycle



- 4 Million Developers
- 3000+ Accelerated Applications
- 40 Million+ CUDA Downloads – 25 Million in 2022
- 15,000+ Startups
- 40,000 Enterprises

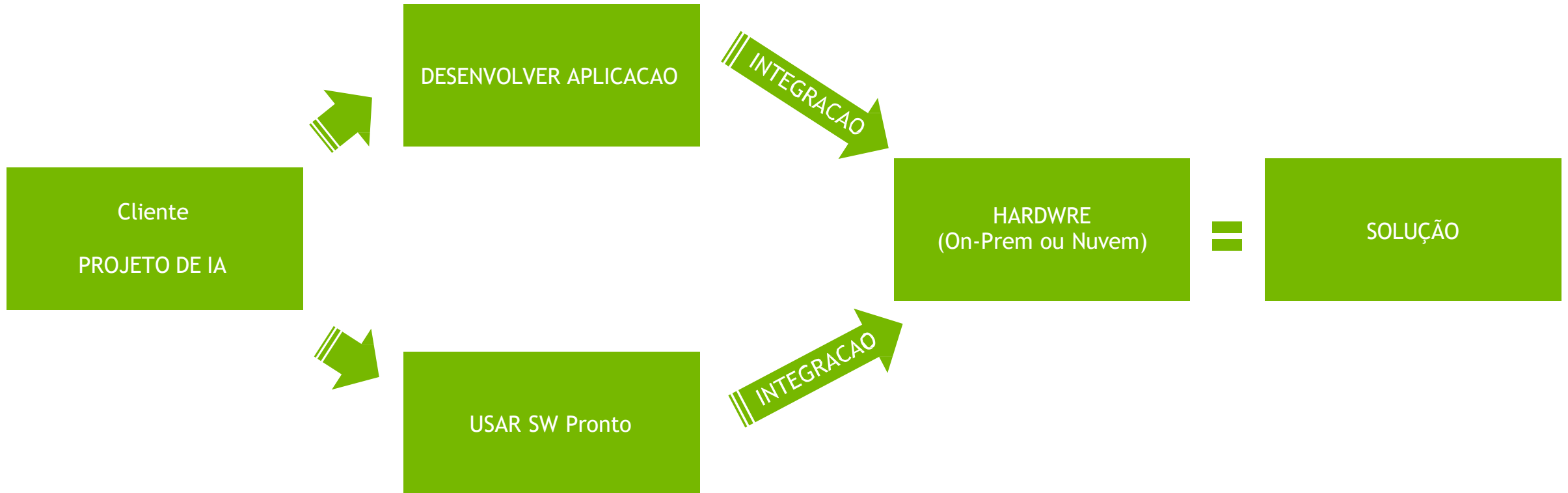
Projetos NVIDIA

Como funciona um projeto com NVIDIA

Primeria Conversa	Sugestao de Solucao	POC	Proposta	Execucao
Identificar os interlocutors Area de Negocios TI Identificar a oportunidade Desafios do cliente Qual o impacto do desafio na area/empresa	Validacao Tecnologica Identificar as tecnologias O que há de ferramenta pronta O que tera que ser desenvolvido	Prova de Conceito (Opcional) Definir o escopo, o timeline, os KPIs; Preparar POC e executar. Eventualmente desenvolver a solucao para POC	Preparacao da Proposta Custo do projeto (incluindo HW, SW e Servicos) Definicao dos prazos e etapas de entrega	Entrega do Projeto Implementacao do projeto
Lead Qualificado	Projeto Definido	POC	Proposta	Entrega
- AM Parceiro - AM NVIDIA	- AM + Pre vendas Parceiro - AM + SA NVIDIA	Pre Vendas e Ops Parceiro	Reseller Sales & Operacoes	Reseller

Caminho mais comum de um projeto de IA

Desenvolvendo ou Integrando



DESENVOLVENDO APLICACAO (80% dos Projetos)

Requisitos para Desenvolver Aplicacoes de IA com NVIDIA



- Certificacao NVIDIA DEEP LEARNING INSTITUTE
 - Machine Learning
 - Deep Learning
 - Visao Computacional
 - LLM
- Conhecimento de ferramentas complementares
 - Banco de Dados Vetorial
 - Phyton
 - Jupiter Notebook
 - Docker/Container
 - Cloud
 - Kubernetes
- Conhecimento para especificar o HW que ira suportar a aplicacao:
 - On-prem
 - Cloud
 - IOT (Jetson)
- Fazer a integracao com o Sistema do cliente + Hardware

O Brasil é o maior investidor em infraestrutura de IA na LATAM, com um gasto esperado de US\$ 240 milhões em 2024. Esses gastos são impulsionados principalmente por projetos de supercomputadores no setor de petróleo e gás e implantações generalizadas de borda em telecomunicações. 58% da liderança e ITBDMs acreditam que a IA terá um impacto significativo na segurança cibernética, detecção de ameaças, automação e eficiência.

Principais prioridades de negócios do CIO em 2024

- 1 Novas tecnologias (por exemplo, IA generativa, etc.)
- 2 Avanço da inovação nos negócios digitais
- 3 Aumento do crescimento da receita e dos lucros

Principais prioridades de investimento em tecnologia em 2024

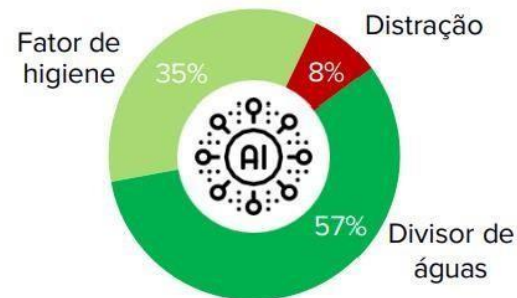
- 1 Aumento do crescimento da receita e dos lucros
- 2 IA generativa
- 3 Modernização de aplicativos antigos essenciais para os negócios

Desafios previstos em 2024

- 1 Cibersegurança e privacidade de dados
- 2 Transformação digital
- 3 Modernização de sistemas antigos



Impacto da IA nas organizações



Principais tecnologias impactadas/influenciadas pela IA

- 1 Segurança cibernética e detecção de ameaças
- 2 Automação e eficiência
- 3 Personalização e experiência do cliente

Aumento dos gastos com computação de borda



Casos de uso de borda mais impactados pela IA

- 1 Gerenciamento de dispositivos de IoT
- 2 Varejo e experiência do cliente
- 3 Automação industrial e manufatura

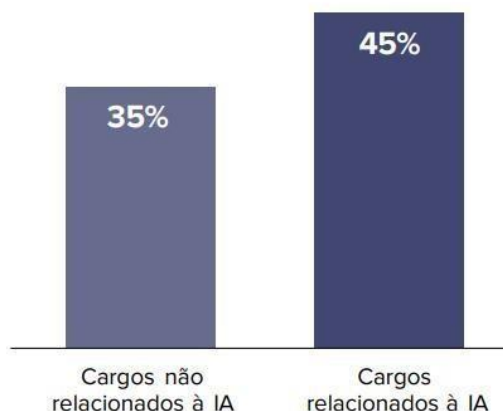
Os desafios encontrados na implementação de casos de uso de negócios de IA incluem falta de capacitação ao utilizar ferramentas de IA de autoatendimento, principalmente aplicativos low/no-code. Além disso, os desafios no longo prazo significam manter uma plataforma de dados confiável e garantir a segurança.

Principais desafios ao implantar a IA

- 1 Falta de capacitação para uso de ferramentas de IA de autoatendimento (aplicativos low/no-code)
- 2 Funcionários preocupados que a IA ameaçará seu emprego
- 3 Desafio em fazer um caso de negócio para tecnologia de IA

Dificuldade de recrutamento

Os respondentes selecionaram
"Extremamente/moderadamente difícil"



Investimento em IA generativa



Preparação para a IA generativa

- 1 Desenvolvimento de habilidades dos funcionários
- 2 Segurança de rede de ponta a ponta
- 3 Construção de alta capacidade de computação

Tipos de interesse em IA generativa

- 1 Business intelligence (encontrando padrões, revelando insights e facilitando as consultas *ad hoc*)
- 2 Produtividade (pessoal, desenvolvedor [geração de código] e analistas de dados)
- 3 Gerenciamento de interação com o cliente

Principais desafios ao usar a IA generativa

- 1 Plataforma de dados confiável
- 2 Segurança
- 3 Remodelagem das forças de trabalho



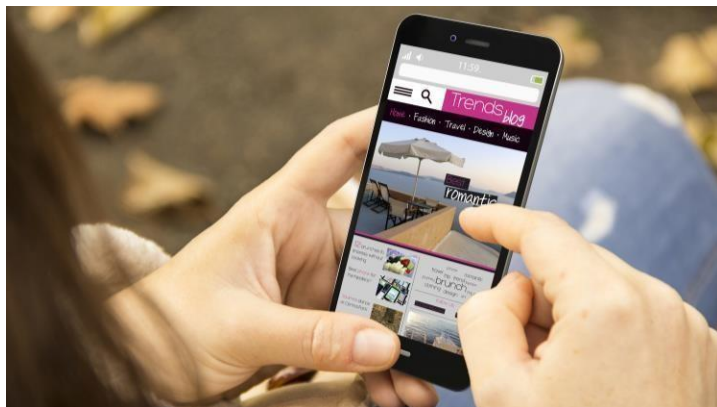
AI Driving Transformation Across Industries

Driving Deeper Insights | Accelerating Product Innovation | Powering Greater Efficiencies



Financial Services

Banking | Insurance | Payments



Consumer Internet

Ecommerce | Social Media | Video Conferencing



Healthcare and Life Sciences

Medical Devices | Smart Hospitals | Genomics



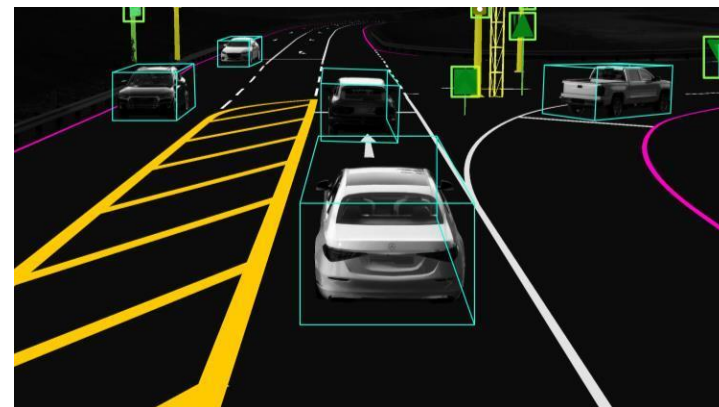
Retail

Intelligent Stores | Logistics | Recommendations



Manufacturing

Digital Twin | Quality Inspection | Maintenance



Transportation

Robotaxis | Driver Assistance | Last-Mile Delivery

Use Cases



[Home](#) [IA](#) [Data Center](#) [Gaming](#) [Gráficos Profissionais](#) [Robótica](#) [Área da Saúde](#) [Startups](#) [Podcast](#) [Casos de Uso](#)

Tag: Cases de Clientes



