## III FÓRUM DE INOVAÇÃO

# AIOT: AMPLIANDO AS FRONTEIRAS DA INTELIGÊNCIA E CONECTIVIDADE

Realização:

FÓRUM BRASILEIRO DE IOT

MCTI

SiDi 20 anos

Apoio:

nic.br

ABISEMI
ASSOCIAÇÃO BRASILEIRA DA INDÚSTRIA DE SEMICONDUTORES

Qualcomm

III FÓRUM DE INOVAÇÃO

AIOT: AMPLIANDO AS FRONTEIRAS DA INTELIGÊNCIA E CONECTIVIDADE

# A importância de HPC e IA para a pesquisa e o setor privado

SiDi – Carlos Pavarina

Realização:

FÓRUM BRASILEIRO DE IOT

MCTI

SiDi 20 anos

Apoio:

nic.br

ABISEMI ASSOCIAÇÃO BRASILEIRA DA INDÚSTRIA DE SEMICONDUTORES

Qualcomm

# Como surgiu o projeto IARA? O Supercomputador de IA!

**NLP, visão computacional e ciência dos dados** exigem modelos cada vez mais complexos

Infraestrutura computacional tradicional se torna **ineficaz** para atender essas necessidades.

Quando surgiu a oportunidade, NVIDIA estava anunciando o lançamento da linha DGX A100 (início de 2020)

Isso que ajudou a impulsionar a adoção da AI na computação de alto desempenho (HPC)

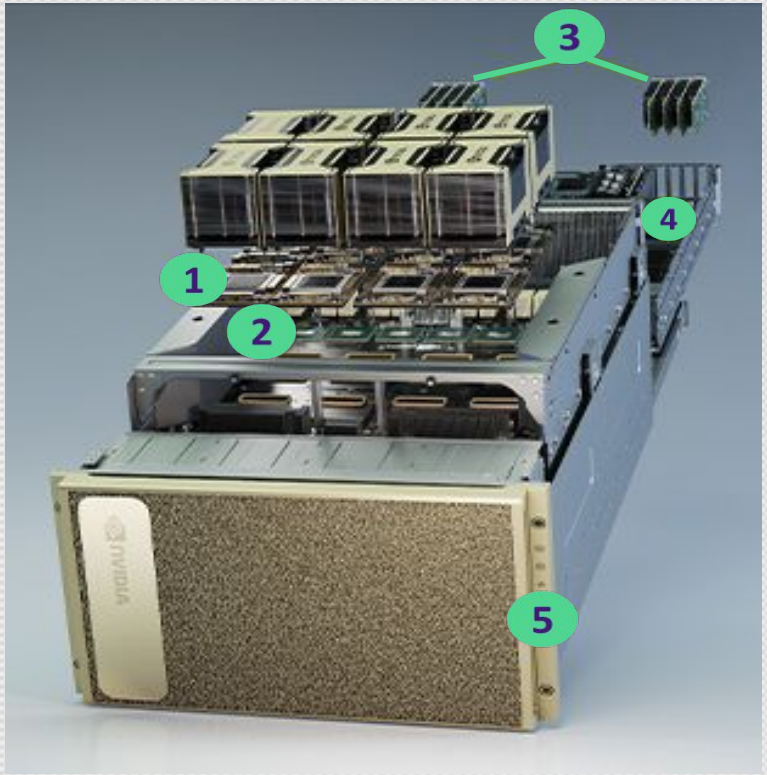
Iniciamos o contato e os estudos para implantar essa infraestrutura computacional para o Brasil

**IARA: Inteligência Artificial Revolucionando o Amanhã**

# NVIDIA - DGX-A100

*40 NVIDIA DGX A100 interconectados em uma rede NVIDIA Mellanox InfiniBand*

1. **8X GPUS NVIDIA A100 COM MEMÓRIA GPU TOTAL DE 320GB**
2. **6X NVIDIA NVSWITCHES**
3. **10x INTERFACE DE REDE MELLANOX CONNECTX-6 200Gb/S**
4. **CPUS AMD DUAL 128 CORE E MEMÓRIA DE SISTEMA DE 2TB**
5. **SSD NVME GEN4 DE 30TB**




Fonte: NVIDIA

# Lançamento Nacional

## Lei de Informática permitiu ao Brasil ter maior infraestrutura de IA na América Latina

Ana Paula Lobo* ... 25/03/2021 ... Convergência Digital

O SiDi, instituto privado de ciência e tecnologia do Brasil, e a Samsung são responsáveis pelo primeiro supercomputador no País concebido para contribuir na aceleração de projetos de pesquisa e desenvolvimento de soluções de Inteligência Artificial (IA).

Construído a partir de recursos da Lei da Informática, o supercomputador conta com tecnologia da NVIDIA Enterprise e se posiciona como a maior infraestrutura de IA da América Latina, com capacidade de fornecer 125 petaflops (cento e vinte cinco quadrilhões de operações matemáticas por segundo) de desempenho para tarefas específicas de Inteligência Artificial. Para se ter uma ideia, seriam necessários cerca de dois milhões de notebooks trabalhando em conjunto para realizar uma tarefa similar.

A máquina reúne 25 sistemas NVIDIA DGX A100 interconectados em uma rede NVIDIA Mellanox InfiniBand, que representa a capacidade de 8,7 milhões de smartphones atuais. Além de 1,5 petabytes de armazenamento de alto desempenho em FlashBlade®, a primeira plataforma do mercado para armazenamento rápido e unificado de file e objetos desenvolvida pela Pure Storage, emppresa desoluções para armazenamento de dados.

**top500.org**

| | | |
|---|---|---|
| 234 | **IARA** - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100 SXM4 40 GB, Infiniband, Nvidia<br>SiDi<br>Brazil | |

O IARA também é o 6º maior do Brasil e o 1º da América Latina focado em projetos de pesquisa e desenvolvimento de soluções de Inteligência Artificial

# Quais os benefícios que o projeto IARA trouxe?

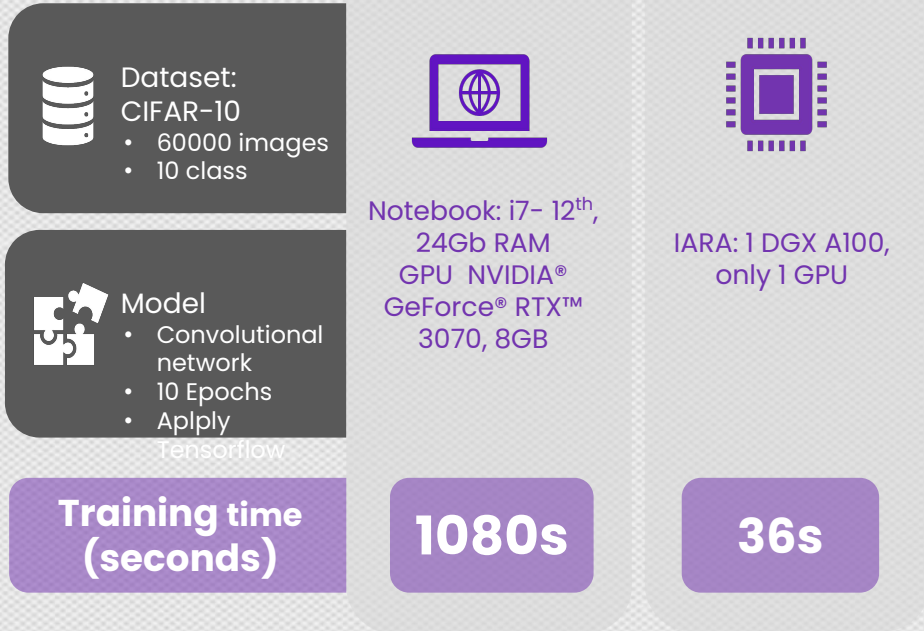
**Versatilidade:** Conseguimos executar vários projetos paralelamente, com centenas de profissionais.

**Agilidade:** Validamos hipóteses e treinamos modelos em muito menos tempo

**Escalabilidade:** Capacidade de treinar modelos cada vez maiores, e com maior número de parâmetros

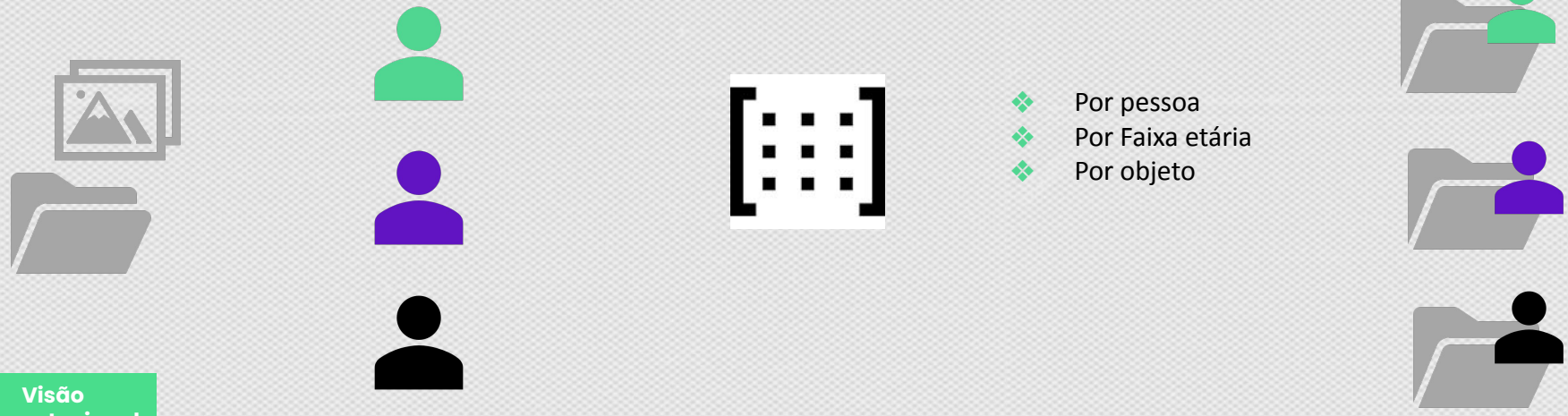
**Com 1 única GPU do IARA:**

- *é capaz de realizar treinamento modelo de rede neural com grandes dataset (referência CIFAR-10)*
- *é 30X mais rápido que um notebook gamer RTX 3070*



# Visão Computacional Avançada

Tecnologias de detecção facial e agrupamento facial para ajudar na identificação e organização de fotos agrupados por faces semelhantes em várias fotos ou vídeos disponibilizados.

- Por pessoa
- Por Faixa etária
- Por objeto

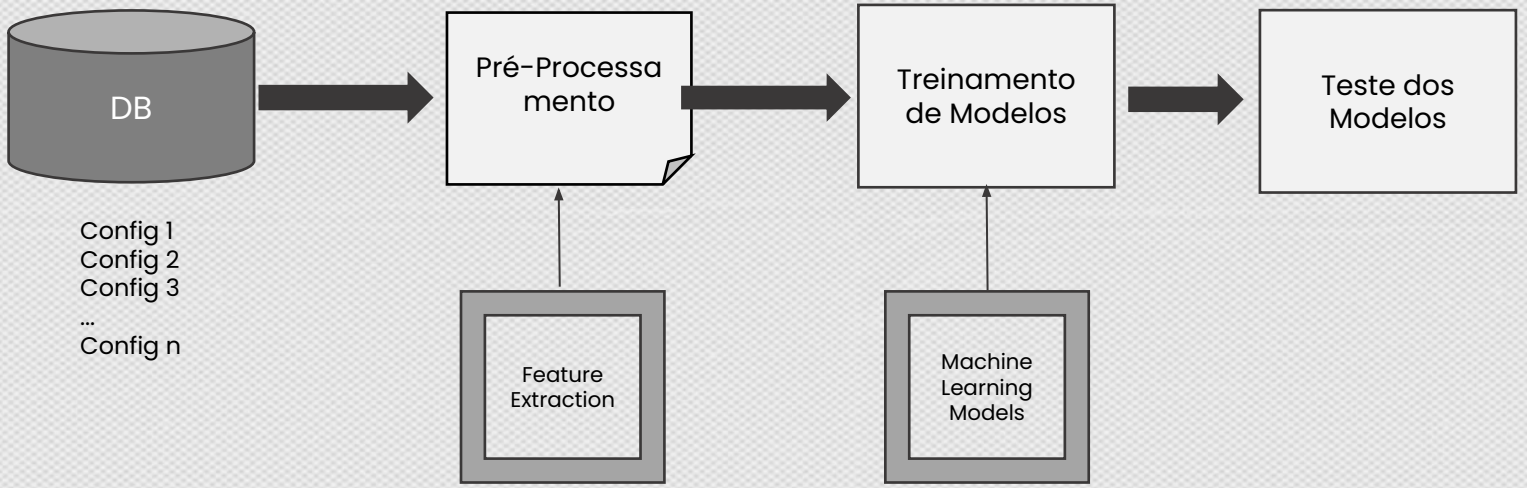
Visão Computacional

Processamento imagem

# Detecção anomalias de sinais

Framework desenvolvido para detecção de anomalias em sinais, possibilitando integração com sistemas industriais para indicação de falhas, manutenções em equipamentos e apoio a tomada de decisão.

## Framework




NLP

Analytics

# Grandes modelos de linguagem

LLM (Large Language Model ou Grandes modelos de linguagem) permite novas possibilidades de compreensão e geração de texto em sistemas de software.

Desenvolvimento de uma infraestrutura de treinamento distribuído (multi-gpu/multi-node) e treinamento de um LLM.

- 100 Bilhões de parâmetros;
- Multi language;
- Teste de stress em escalabilidade;

## Data

## Challenges

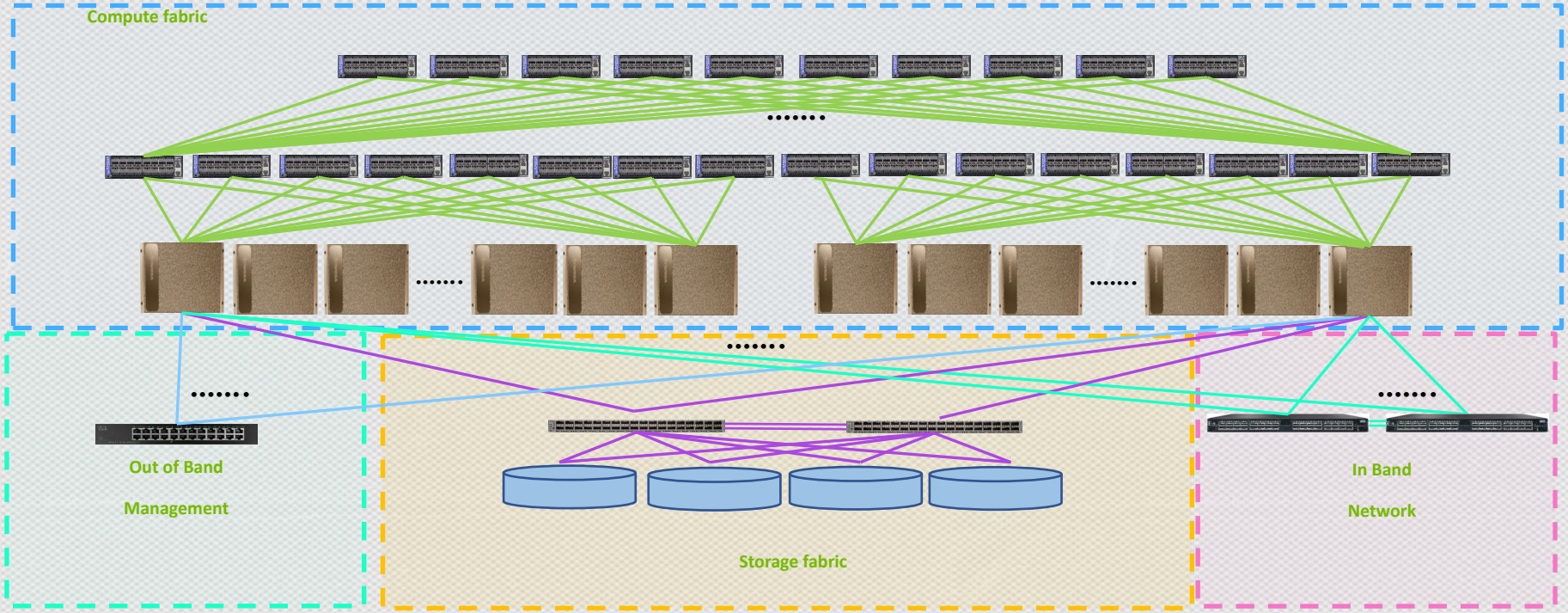
1. Understanding model limitations
2. Choosing your model's endpoint
3. Finetuning the model to your task
4. Choosing the right set of parameters
5. Designing prompt for the model

## Result

- Generate
- Summarize
- Rewrite
- Extract
- Search/Similarity
- Cluster
- Classify

LLM

# A Complexidade de um Supercomputador




- 40 DGXs-A100 com um total de 320 GPUs trabalhando em conjunto
- Sistemas de armazenamento massivo e de alta performance
- Redes de interconexão complexas
- Software especializado para gerenciamento e otimização

# Desafios e oportunidades

- **Complexidade Tecnológica:** Exige atualizações constantes e uma equipe altamente qualificada.
- **AI SW Stack:** Novas ferrametas de desenvolvimento e operação.
- **Gerenciamento de energia:** Otimização para reduzir custos e impacto ambiental.
- **Disponibilidade:** Garantir o funcionamento ininterrupto do Sistema.
- **Segurança:** Proteção contra ataques cibernéticos e perda de dados.
- **Otimização de Recursos:** Fundamental para maximizar o desempenho e evitar gargalos.
- **Escalabilidade:** Adaptar o sistema para atender às demandas crescentes
- **Atualizações constantes:** Acompanhar a evolução tecnológica.

# A Importância do Investimento em Supercomputadores de IA:

- **Inovação:** Impulsiona o desenvolvimento de novas aplicações e soluções baseadas em IA.
- **Competitividade:** Aumentar a competitividade das empresas em seus respectivos mercados.
- **Crescimento Econômico:** Potencial de gerar novos negócios e impulsionar o crescimento econômico.
- **Desenvolvimento Científico:** Financiar pesquisas e projetos que contribuam para o avanço tecnológico e retenção de talentos.
- **Solução de Problemas Sociais:** Resolver problemas sociais complexos, como a saúde, a educação e o meio ambiente.

**A gestão de um supercomputador de IA é um desafio complexo, mas com grandes recompensas, pois ao investirmos em infraestrutura, capacitação e em uma cultura de inovação, podemos nos beneficiar de todas essas vantagens que a IA pode oferecer.**

# Obrigado!

**Carlos Pavarina**

SiDi – HPC Manager

c.pavarina@sidi.org.br

Realização:

Apoio: