**Arquiteturas de Computação Especializadas**

# **GPUs**

Uma GPU (Unidade de Processamento Gráfico) é um circuito eletrônico especializado em realizar cálculos matemáticos complexos, especialmente aqueles relacionados à renderização de gráficos. Ela é fundamental para a geração de imagens em dispositivos eletrônicos, como computadores e consoles de jogos.

Responsável pelo processamento dos dados gráficos. Ela contém vários multiprocessadores que permitem o processamento paralelo de grandes quantidades de dados.

As GPUs são usadas para criar e treinar redes neurais, que são sistemas projetados para aprender e tomar decisões como um cérebro humano. Isso inclui desde a análise de grandes volumes de dados em serviços financeiros até a pesquisa em medicina.

# **TPUs** **O que são TPUs (Tensor Processing Units)?**

**As TPUs (Tensor Processing Units) são processadores especializados desenvolvidos pelo Google para acelerar as cargas de trabalho de machine learning. Elas foram criadas para executar cálculos de matrizes e tensores de forma extremamente rápida e eficiente, superando GPUs e CPUs em tarefas de IA.**

### **Por que as TPUs são importantes?**

A demanda por **treinamento e inferência de modelos de IA** cresceu muito nos últimos anos. No entanto, **CPUs são lentas** para essas tarefas e **GPUs consomem muita energia**. As TPUs foram desenvolvidas para resolver esse problema, trazendo vantagens como:

**Velocidade extrema** → TPUs treinam modelos de IA até **15 vezes mais rápido** que GPUs.

**Eficiência energética** → Consomem **menos energia** do que GPUs para a mesma carga de trabalho.

**Latência reduzida** → Executam inferências rapidamente, sendo ideais para **reconhecimento de imagem e voz em tempo real**.

**Escalabilidade** → São usadas em **data centers do Google** para treinar modelos gigantes de IA lidando com um volume crescente de dados e processamento sem perder desempenho.

### **Onde as TPUs são usadas?**

**Serviços do Google** → O Google usa TPUs em **Google Search, Google Photos, Google Translate e Google Assistant** para oferecer IA mais rápida e precisa.

**Treinamento de Modelos de IA** → Empresas e pesquisadores usam TPUs na **Google Cloud** para treinar redes neurais profundas, como GPT e BERT.

**Ciência e Pesquisa** → TPUs aceleram simulações científicas, como descoberta de novos medicamentos e previsões climáticas.

**Jogos de IA** → O Google usou TPUs para treinar o **AlphaGo**, a IA que derrotou campeões humanos no jogo Go.

### **Quando usar Cloud TPUs?**

As Cloud TPUs são otimizadas para treinar modelos de aprendizado profundo grandes e complexos que apresentam muitos cálculos de matriz, por exemplo, criando modelos de linguagem grandes (LLMs). As Cloud TPUs também têm o SparseCores, que são processadores de fluxo de dados que aceleram modelos que dependem de embeddings encontrados em modelos de recomendação. Outros casos de uso incluem assistência médica, como modelagem da dobra de proteínas e descoberta de medicamentos.

# **NPUs** **O que são NPUs (Neural Processing Units)?**

As NPUs (Neural Processing Units) são um tipo de microprocessador ou componente de hardware projetado para desempenhar de forma otimizada tarefas relacionadas a IA. A importância das NPUs é cada vez mais crescente com o aumento da demanda por inteligência artificial e machine learning com suas tarefas complexas e que exigem grandes capacidades de processamento e geram necessidade por otimização.

### **Por que as NPUs são importantes?**

* Otimizadas para as crescentes necessidades de IA
* Melhoria de desempenho em relação a CPUs e GPUs, liberando-as para outras tarefas
* Eficiência energética, vital para dispositivos que operam com baterias
* Capacidade de realizar cálculos complexos e intensivos
* Grande capacidade para processamento paralelo

### **Onde as NPUs são usadas?**

Mesmo que as NPUs sejam uma tecnologia muito nova no mercado, sua presença já pode ser observada em diversos outros componentes como:

* **TPU** (focado em Machine Learning)
* **NNP** (focado em processamento avançado de redes neurais)
* **IPU** (focado em atividades de inteligência computacional)
* **VPU** (focado em processamento de imagens e visão computacional)
* **GPU** (conhecida pela renderização gráfica e também utilizada em cálculos relacionados a IA)

Além de aplicações mais comuns no cotidiano:

* Big Data
* Carros autônomos
* LLMs
* Realidade Virtual e Aumentada
* Análise de imagens médicas e diagnósticos precisos
* Segurança e monitoramento em tempo real
* Dispositivos móveis para melhorar reconhecimento de voz e processamento de imagens economizando energia ao mesmo tempo

As NPUs estão sendo empregadas em Chipsets (conjunto de chips que funcionam em suporte ao processador) das principais empresas de tecnologia pelo mundo:

* Google TPU
* Apple Neural Engine
* Huawei Kirin Chips
* AWS Inferentia
* Intel NNP, Myriad, EyeQ e GNA
* Nvidia NVDLA e Xavier
* Samsung NPU
* Tesla FSD Chip

**A Situação Afetada pela Tecnologia (O Problema Tratado por Ela)**

As arquiteturas de computação especializadas, como GPUs, TPUs e NPUs, surgiram para resolver limitações das CPUs tradicionais quando se trata de tarefas altamente paralelizáveis, como aprendizado profundo, processamento gráfico e inteligência artificial. O problema central abordado por essas tecnologias é a necessidade crescente de desempenho computacional para aplicações modernas, que envolvem grandes volumes de dados e cálculos complexos.

### Processamento Paralelo e Desempenho

Os processadores convencionais (CPUs) são otimizados para executar um pequeno número de tarefas sequenciais com alta eficiência. No entanto, tarefas como treinamento de redes neurais, simulações científicas e processamento gráfico exigem uma enorme capacidade de computação simultânea. GPUs, TPUs e NPUs resolvem esse problema ao adotar arquiteturas especializadas para processar múltiplas operações em paralelo, tornando essas tarefas muito mais rápidas e eficientes.

### Avanços em Inteligência Artificial e Aprendizado de Máquina

O treinamento de modelos de aprendizado profundo requer milhões ou até bilhões de cálculos de multiplicação de matrizes e operações de otimização, tornando inviável o uso exclusivo de CPUs. Para enfrentar esse desafio:

* **GPUs** aceleram o processamento de redes neurais devido à sua capacidade de realizar cálculos paralelos em larga escala.
* **TPUs** são projetadas especificamente para computação em IA, reduzindo o consumo de energia e aumentando a velocidade de inferência e treinamento.
* **NPUs** possibilitam a execução de modelos de IA embarcada em dispositivos móveis e IoT, permitindo aplicações em tempo real, como reconhecimento de imagens e voz.

### Aplicações em Jogos e Realidade Virtual

A renderização gráfica em tempo real para jogos e realidade virtual depende de uma taxa de processamento altíssima para gerar imagens detalhadas e fluidas. GPUs foram projetadas para lidar com esse problema, possibilitando gráficos ultrarrealistas e experiências imersivas.

### 4. Eficiência Energética e Computação em Nuvem

À medida que a demanda por computação intensiva cresce, surge um novo problema: o consumo excessivo de energia. As TPUs e NPUs oferecem soluções mais eficientes ao consumir menos energia para as mesmas tarefas que GPUs e CPUs realizariam com maior custo energético. Isso tem um impacto direto em data centers, dispositivos móveis e sistemas embarcados.

# **Quais foram os avanços tecnológicos?**

Os avanços em GPUs, TPUs e NPUs transformaram diversas áreas da tecnologia. As TPUs, são mais especializadas em inteligência artificial, viabilizando o treinamento de redes neurais avançadas, impulsionando modelos como ChatGPT. Já as NPUs otimizaram dispositivos móveis, melhorando recursos como fotografia computacional e assistentes virtuais. Essas tecnologias também impactaram a medicina, tornando diagnósticos por imagem mais precisos, além de viabilizar veículos autônomos, robótica avançada e realidade virtual. As GPUs, por sua vez, revolucionaram principalmente os gráficos, tornando animações e jogos mais realistas, além de acelerar simulações científicas. Com isso, continuam impulsionando a inovação e expandindo as fronteiras da computação.

# **Link slides**

<https://www.canva.com/design/DAGgMrcKeHc/nr4q5ulZCWPOyPwdliwD4w/edit?utm_content=DAGgMrcKeHc&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton>