GPU GPGPU

HW GRÁFICO

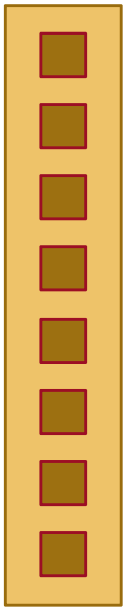
Evolução Tecnológica – GPUs

Evolução das GPUs ao longo dos anos

https://www.youtube.com/watch?v=wHTdnIviZTE

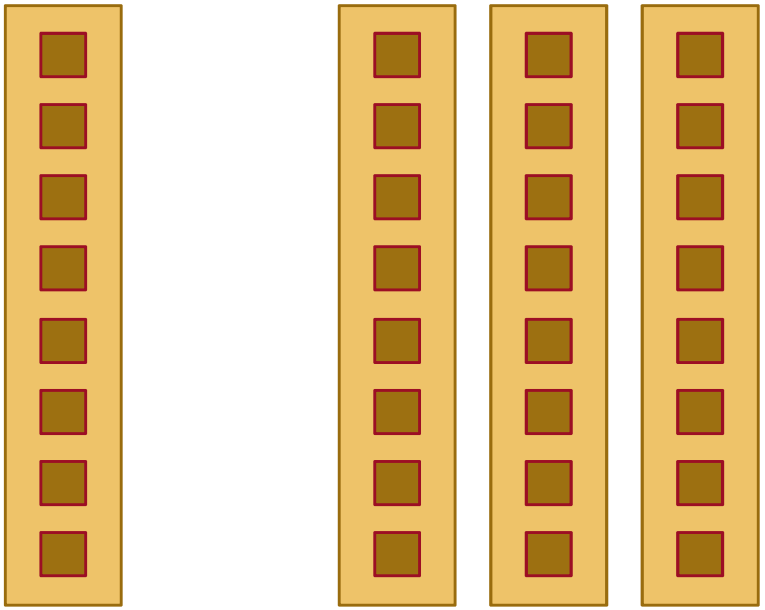
https://www.techspot.com/article/650-history-of-the-gpu/

Conceito
evolução – tipos de paralelismo



Pipeline
Especialização Funcional / Linha de Produção
- processadores especializados para cálculos de cada etapa
- processamento em fluxo contínuo (streaming)

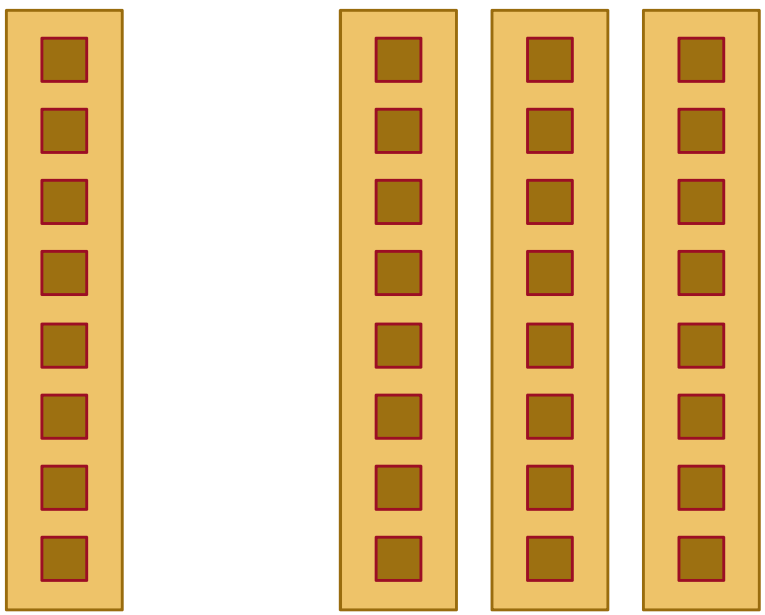
Conceito
evolução – tipos de paralelismo

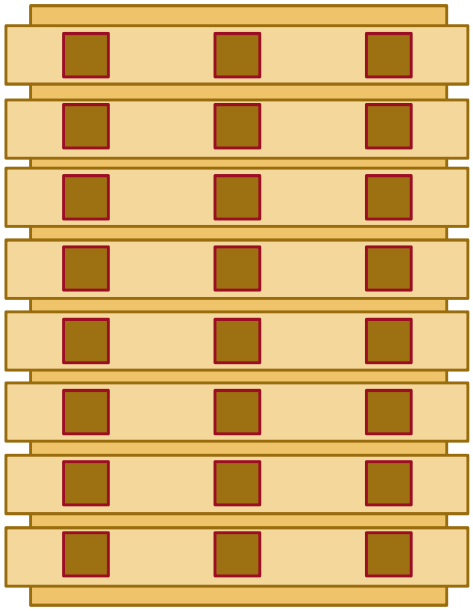


Múltiplas Pipelines
Replicação da Linha de Produção

- permite processamento simultâneo de diferentes objetos / partes da cena
- processamento paralelo

Conceito
evolução – tipos de paralelismo

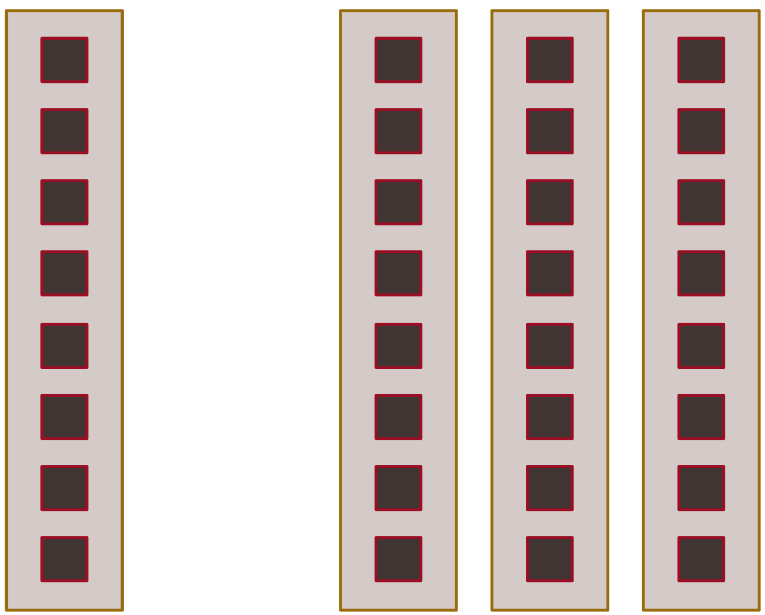


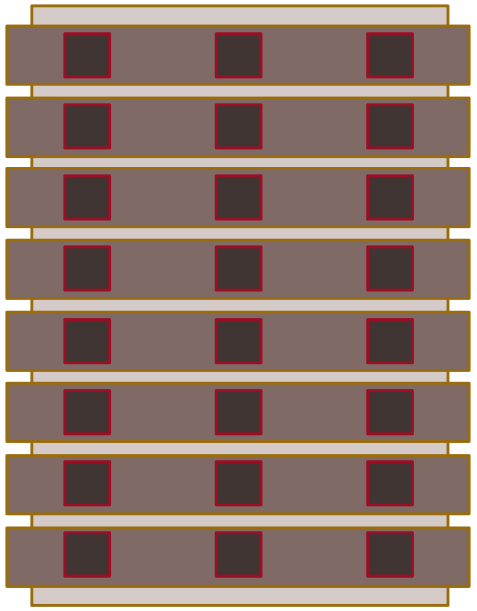


Reorganização

- Paralelismo em cada nível
- melhor ajuste dinâmico da carga computacional

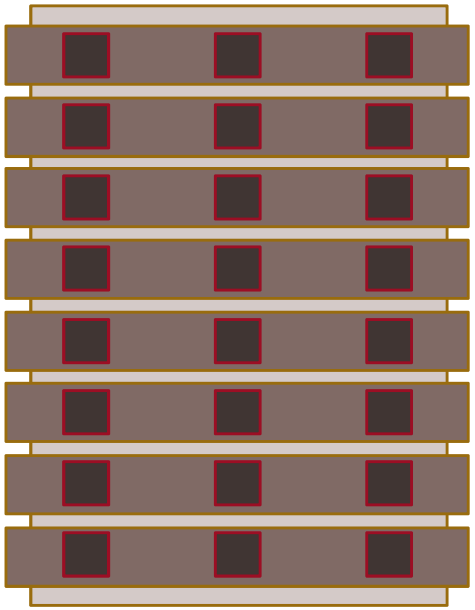
Conceito
evolução – programabilidade

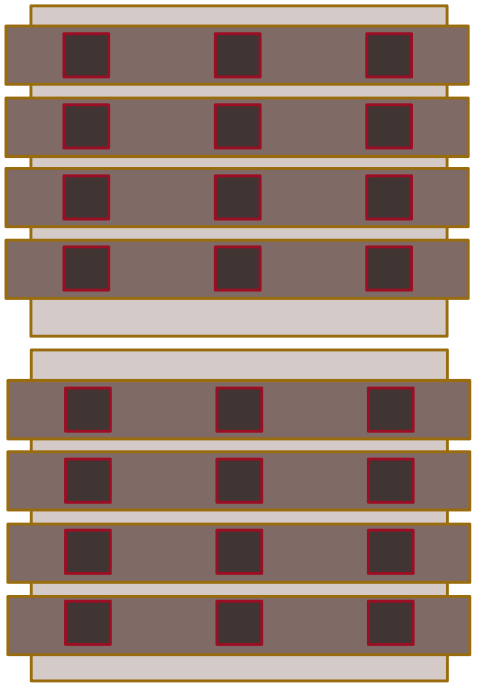


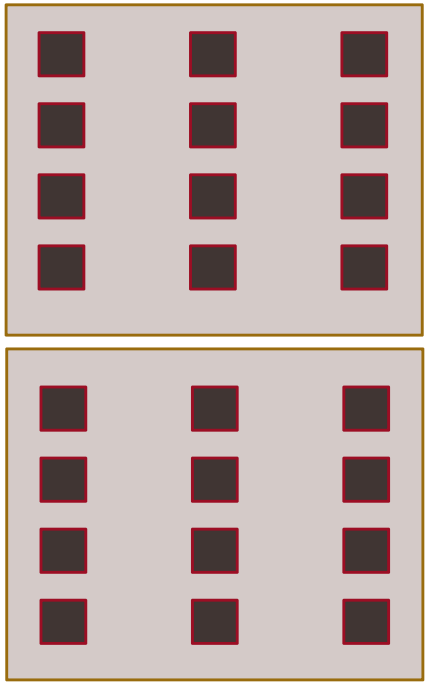


Unidades programáveis
- ajuste funcional

Conceito
evolução – segmentação X uniformidade



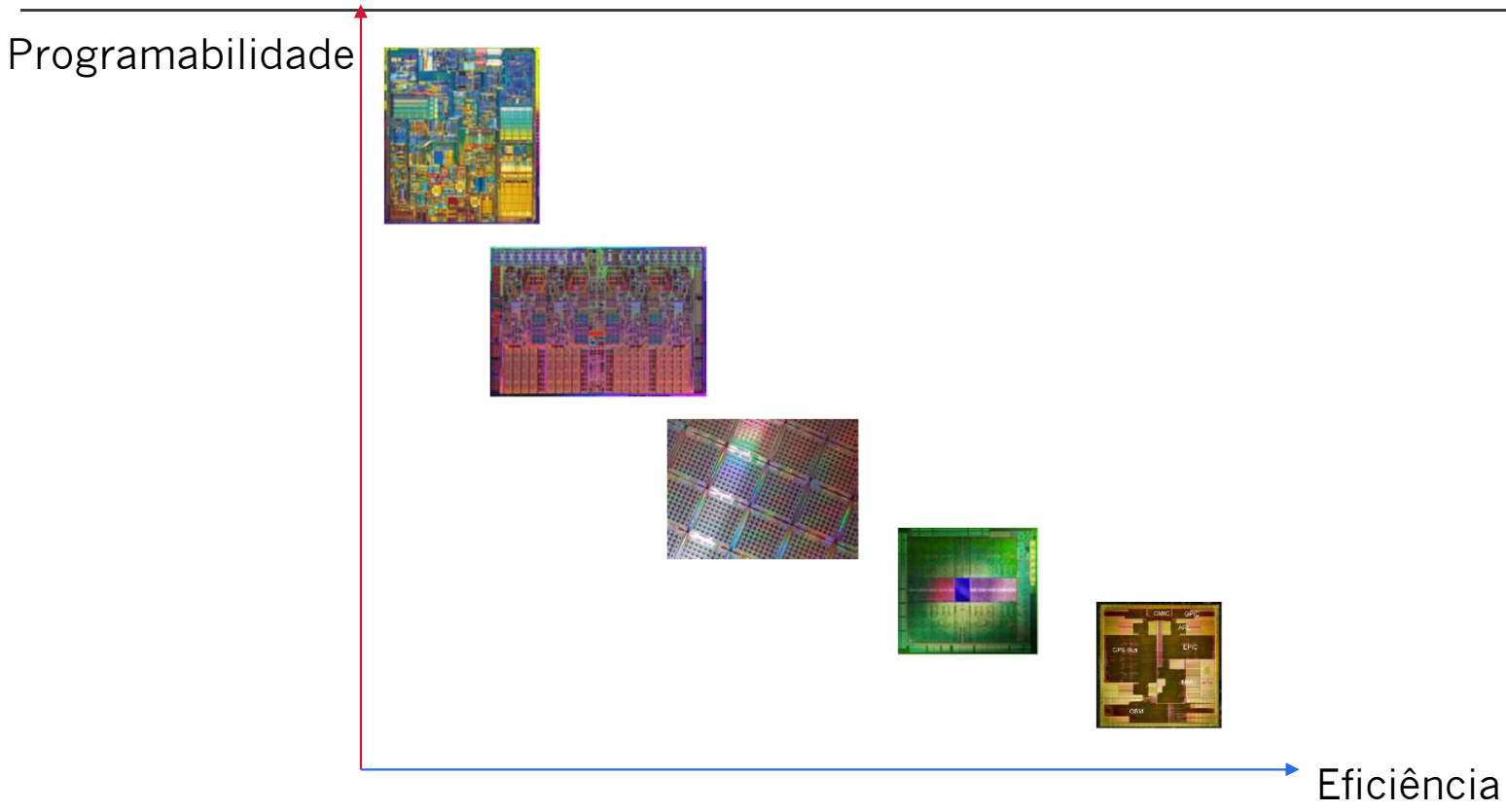




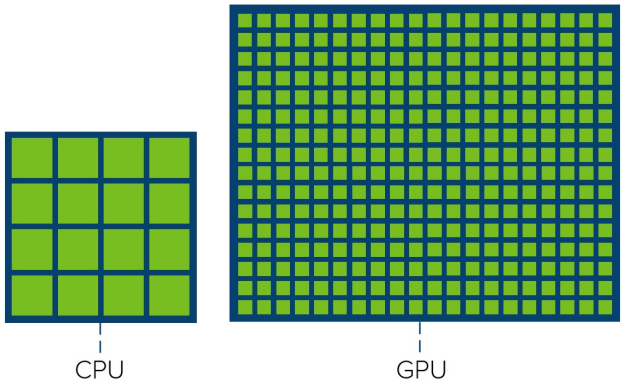
Especialização
GPU

Generalização
GPGPU

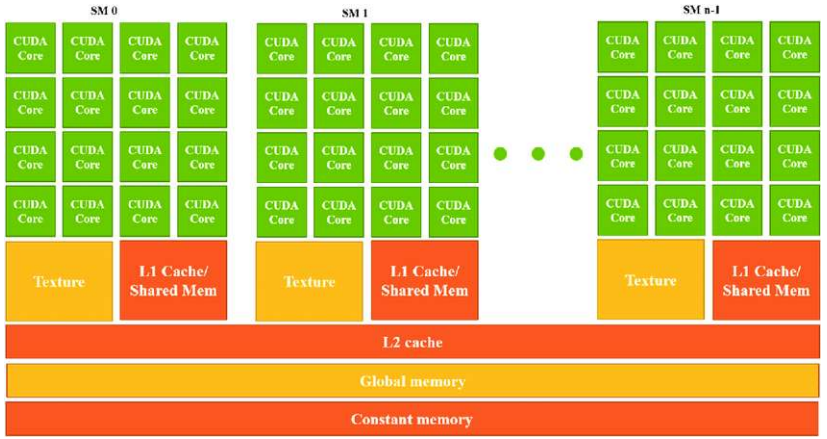
Compromisso flexibilidade X desempenho



Arquiteturas



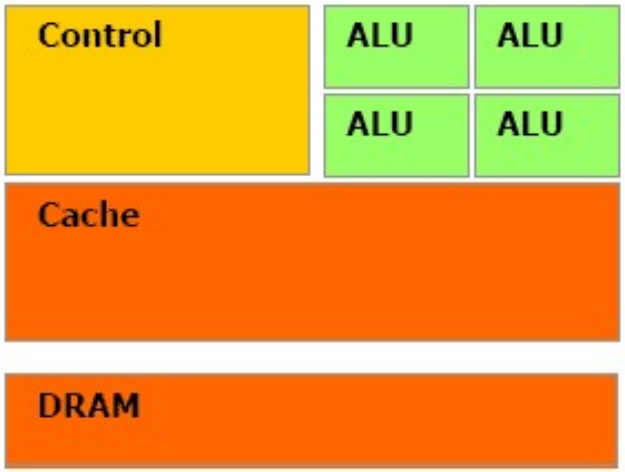
CPU x GPU

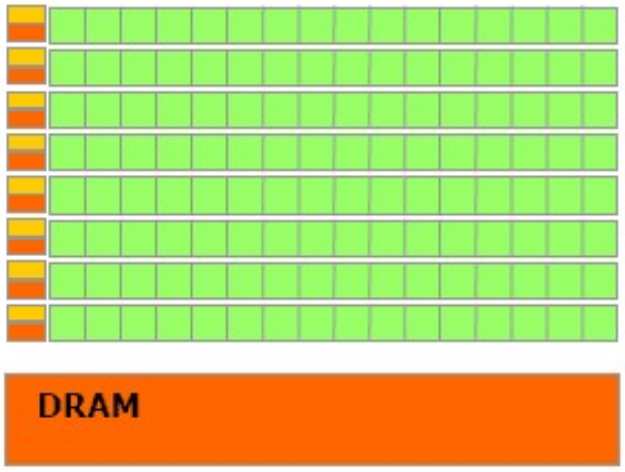


Unidade Processamento
CORE

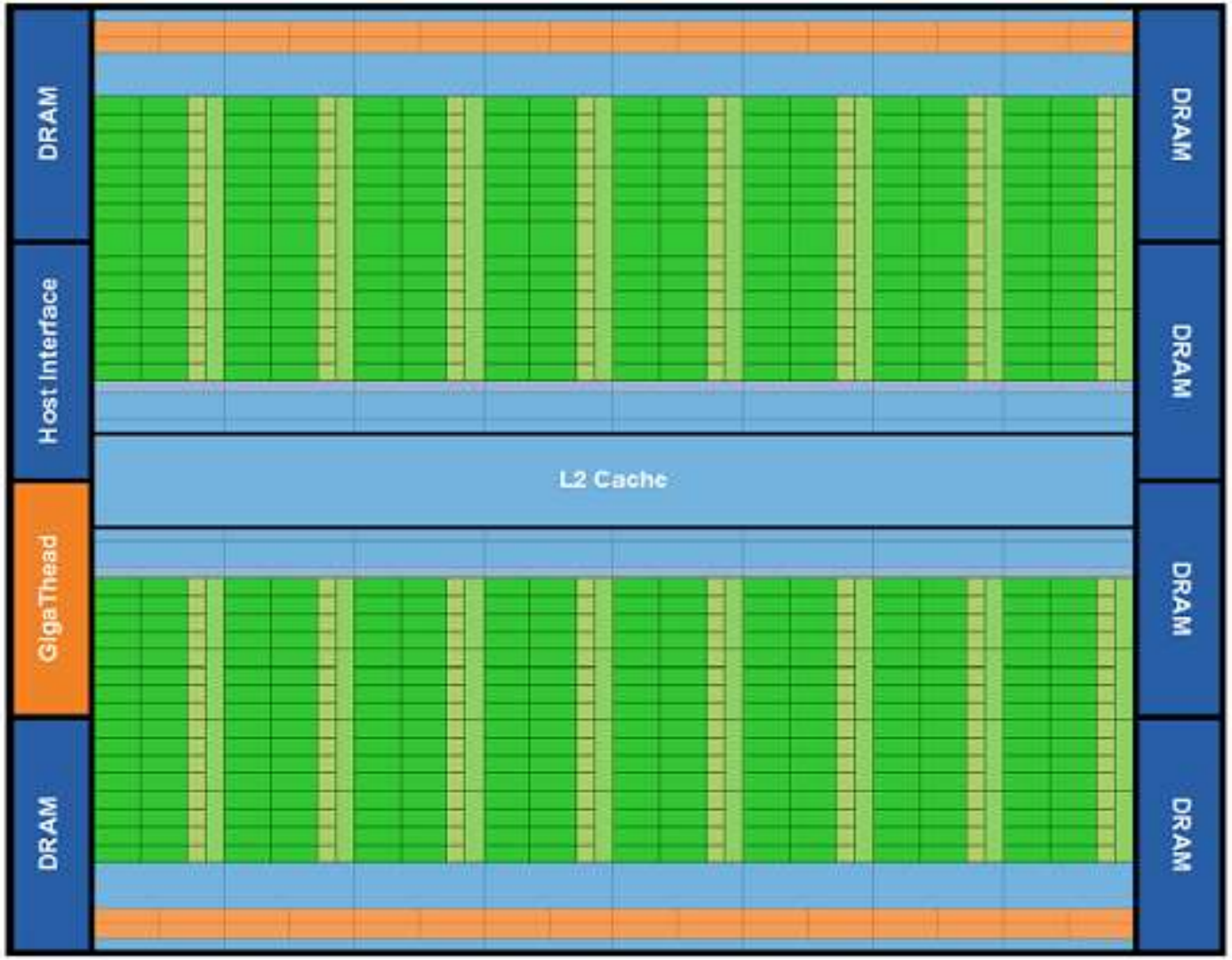
Memória (R / C-L1 / C-L2 / ML / MG)

Arquiteturas





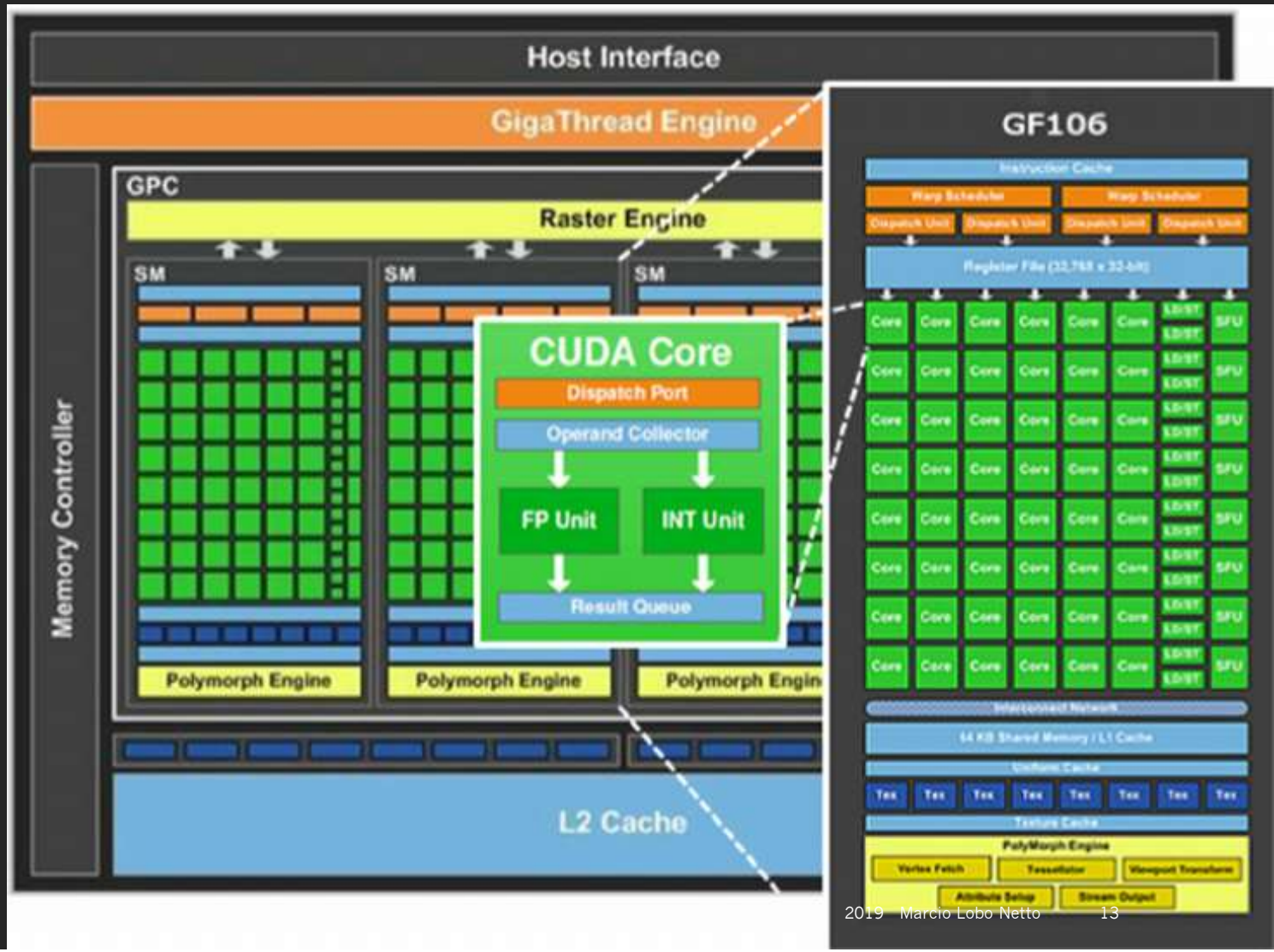
arquiteturas



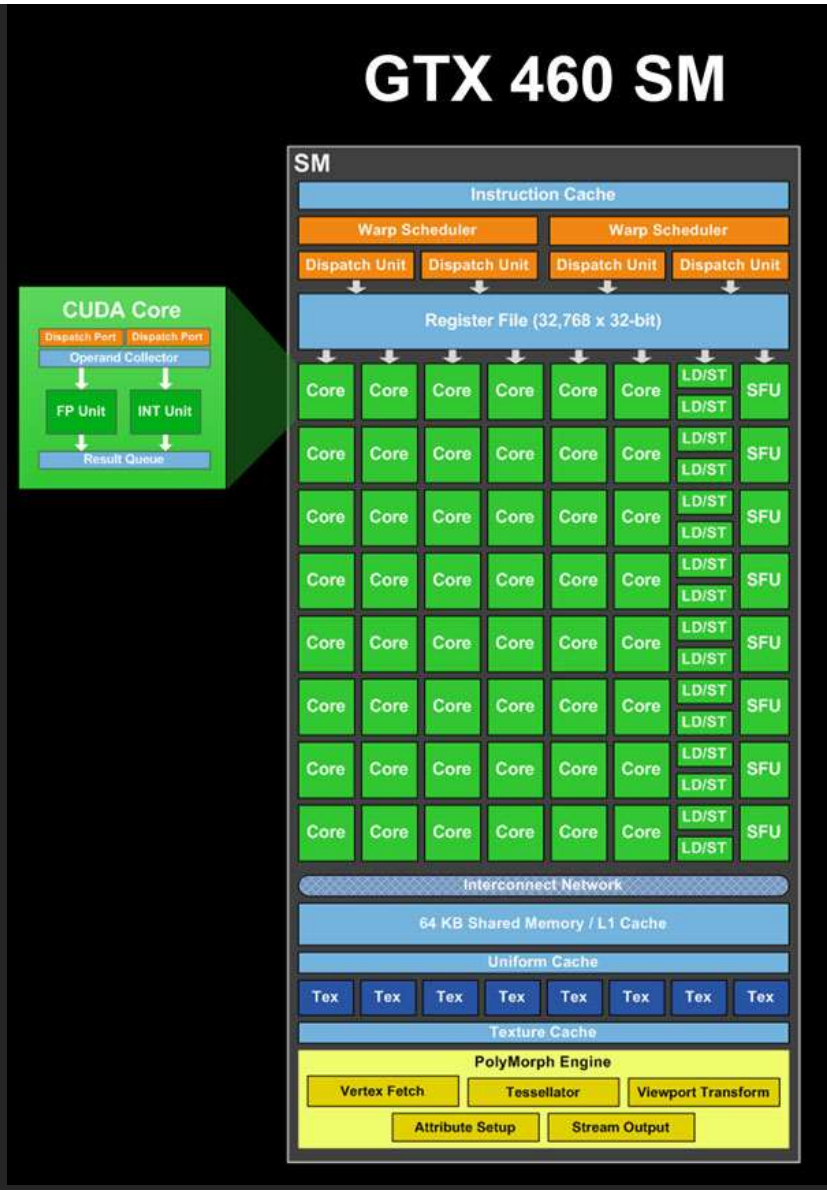
Fermi's 16 SM are positioned around a common L2 cache. Each SM is a vertical rectangular strip that contain an orange portion (scheduler and dispatch), a green portion (execution units), and light blue portions (register file and L1 cache).

arquiteturas

arquiteturas



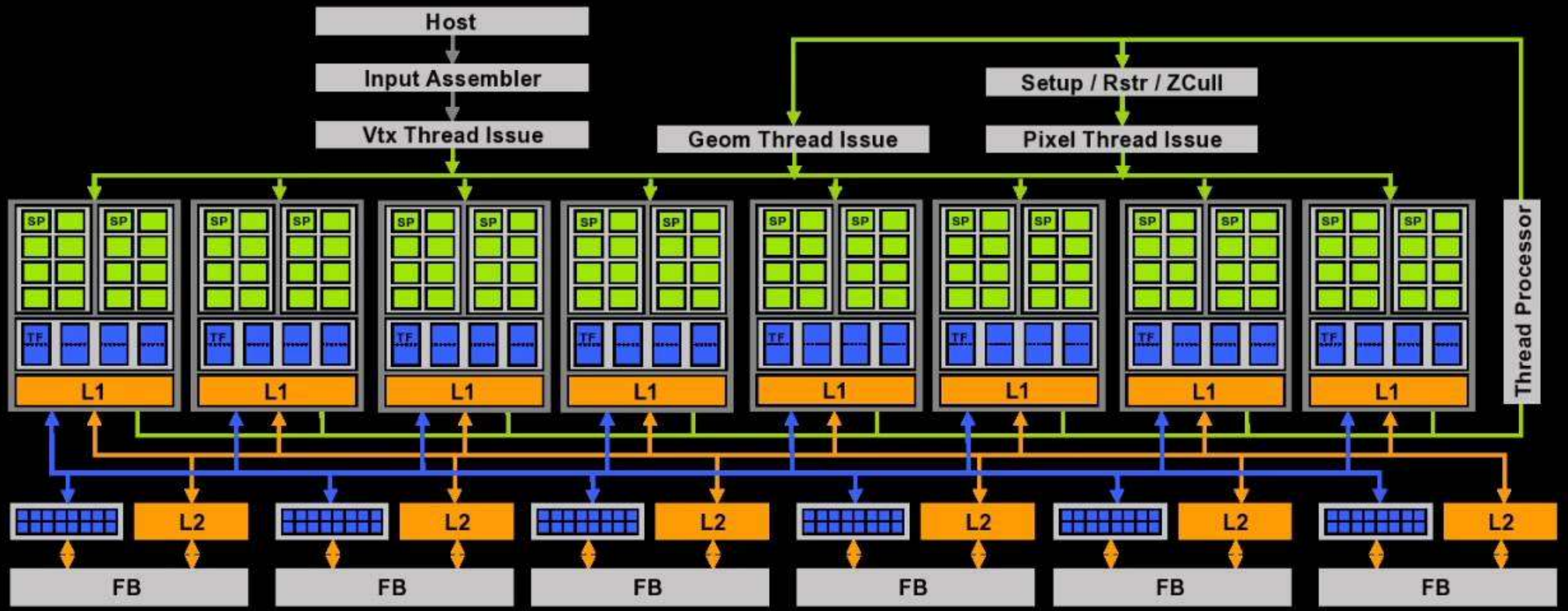
arquiteturas

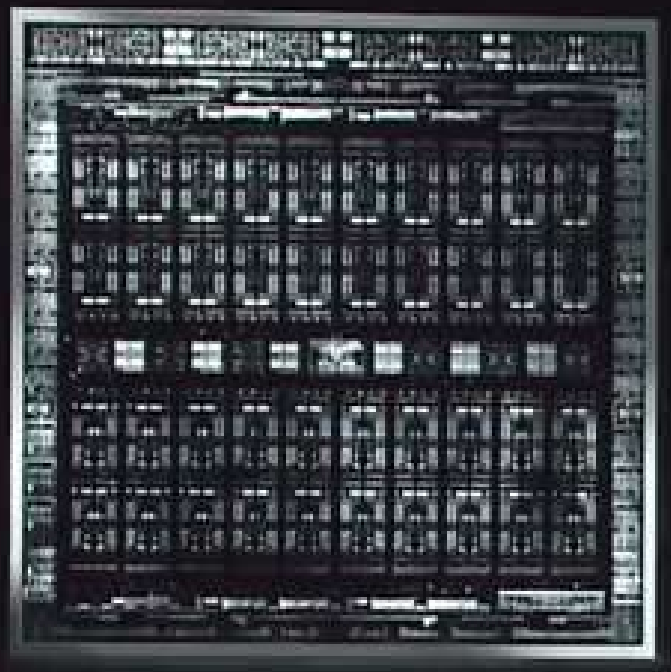


NVIDIA G80 Architecture

Unified Compute Architecture

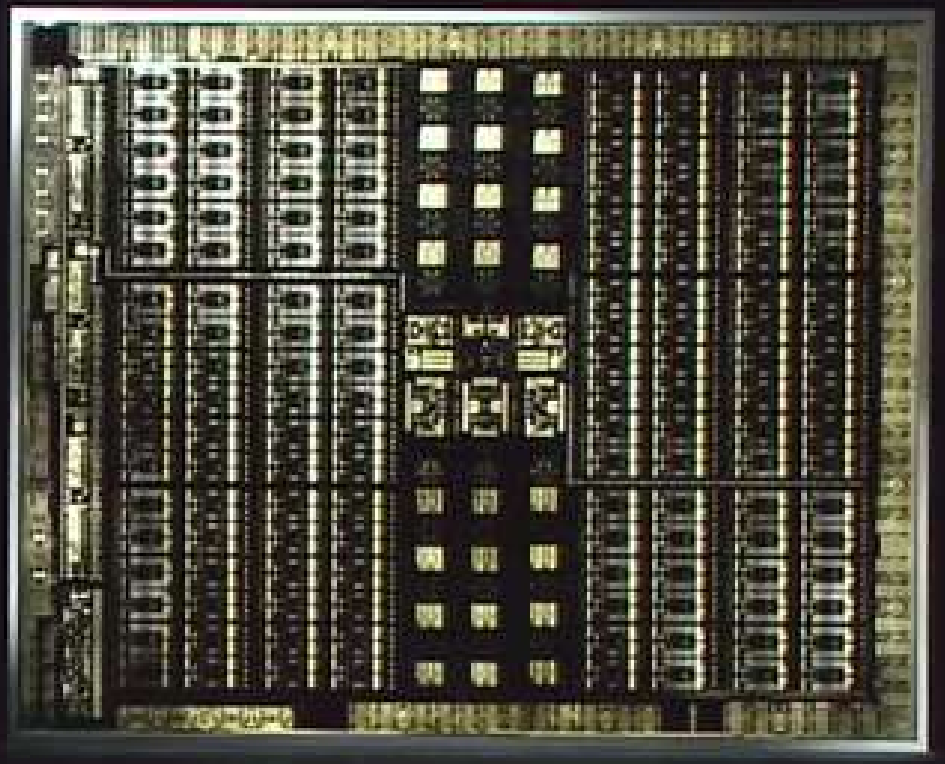
- Unified processor types
- Unified access to mem structures
- SIMT, shared memory
- DirectX 10 & SM 4.0





PASCAL

11.8 Billion xtors | 471 mm² | 24 GB 10GHz



TURING

18.6 Billion xtors | 754 mm² | 48+48 GB 14GHz

Famílias tecnológicas

Pascal	1080 series
Turing	1016 / 1020 series
Fermi	
Titan	
Tesla	

conceitos arquiteturas

TensorFlow | Homepage | DeepMind | Placas de Vídeo com Arquitetura | Shot com GeForce Galleria Ansel

nvidia.com/pt-br/geforce/turing/

Pausada

GEFORCE PRODUTOS DRIVERS GEFORCE EXPERIENCE NOTÍCIAS GAMES COMUNIDADE SUPORTE

NVIDIA TURING

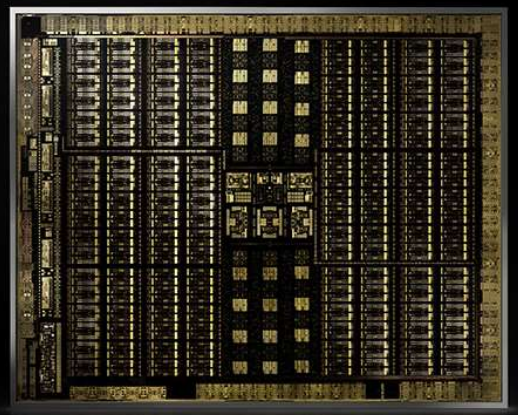
RAY TRACING AI NOS GAMES SOMBREAMENTO AVANÇADO

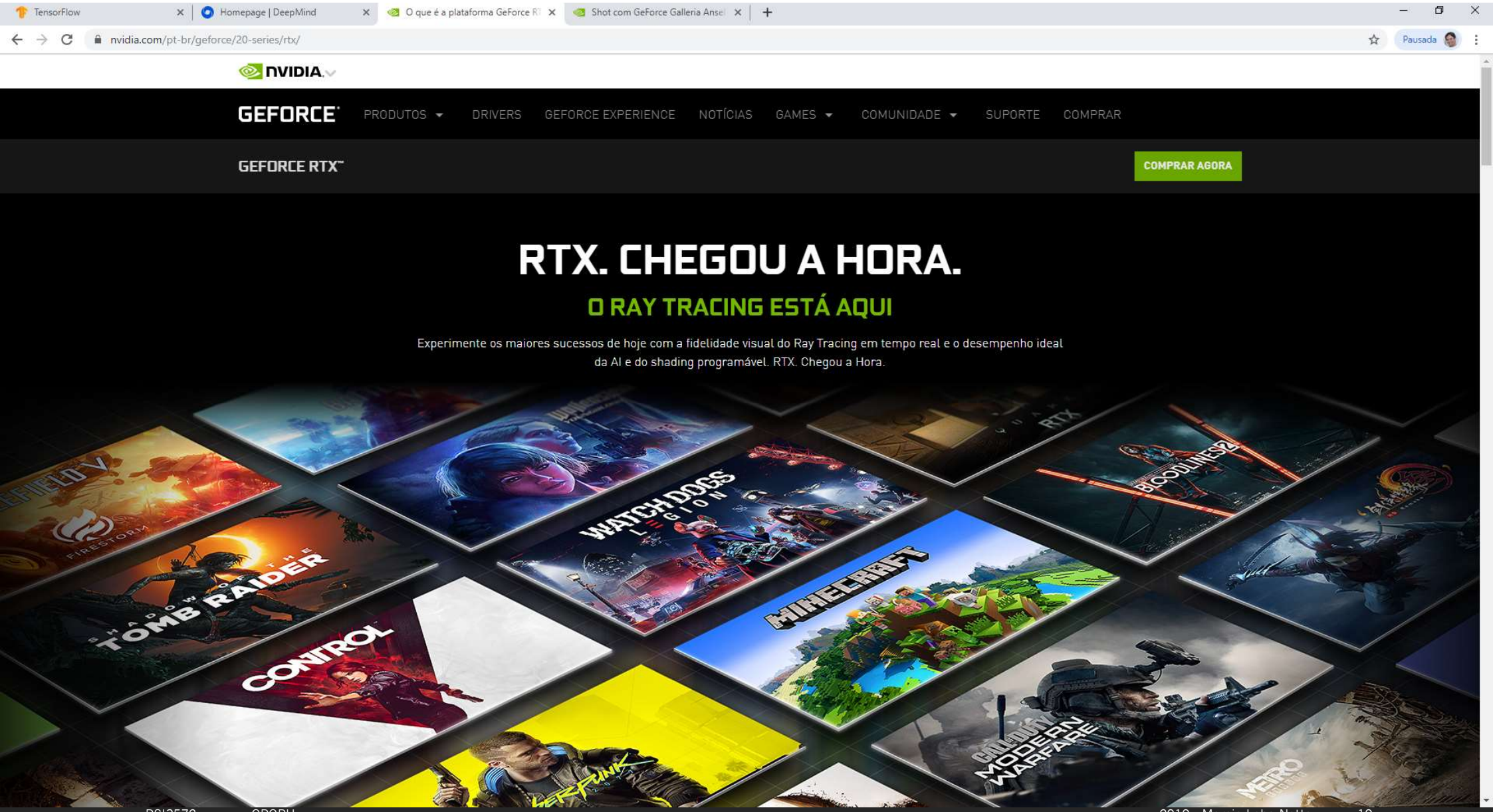
RTX. CHEGOU A HORA.

NVIDIA TURING

A NVIDIA está reinventando a computação gráfica.

A revolucionária arquitetura NVIDIA Turing™, junto com nossa nova plataforma GeForce RTX™, combina o Ray Tracing em tempo real, a inteligência artificial e o sombreamento programável para oferecer uma experiência totalmente nova nos games.




TensorFlow
Homepage | DeepMind
O que é a plataforma GeForce RT
Shot com GeForce Galleria Ansel
nvidia.com/pt-br/geforce/20-series/rtx/
Pausada
NVIDIA
GEFORCE
PRODUTOS
DRIVERS
GEFORCE EXPERIENCE
NOTÍCIAS
GAMES
COMUNIDADE
SUPORTE
COMPRAR
GEFORCE RTX™
COMPRAR AGORA
RTX. CHEGOU A HORA.
O RAY TRACING ESTÁ AQUI
Experimente os maiores sucessos de hoje com a fidelidade visual do Ray Tracing em tempo real e o desempenho ideal da AI e do shading programável. RTX. Chegou a Hora.
FIRESTORM
SHADOW OF THE TOMB RAIDER
WATCH DOGS LEGION
CONTROL
MINECRAFT
BOODLINES2
CYBERPUNK
CALL OF DUTY MODERN WARFARE
METRO EXODUS

TensorFlow | Homepage | DeepMind | Placa de Vídeo GeForce RTX 208 | Computador Gamer, T-power Sp | Shot com GeForce Galleria Ansel

nvidia.com/pt-br/geforce/graphics-cards/rtx-2080-ti/

Pausada

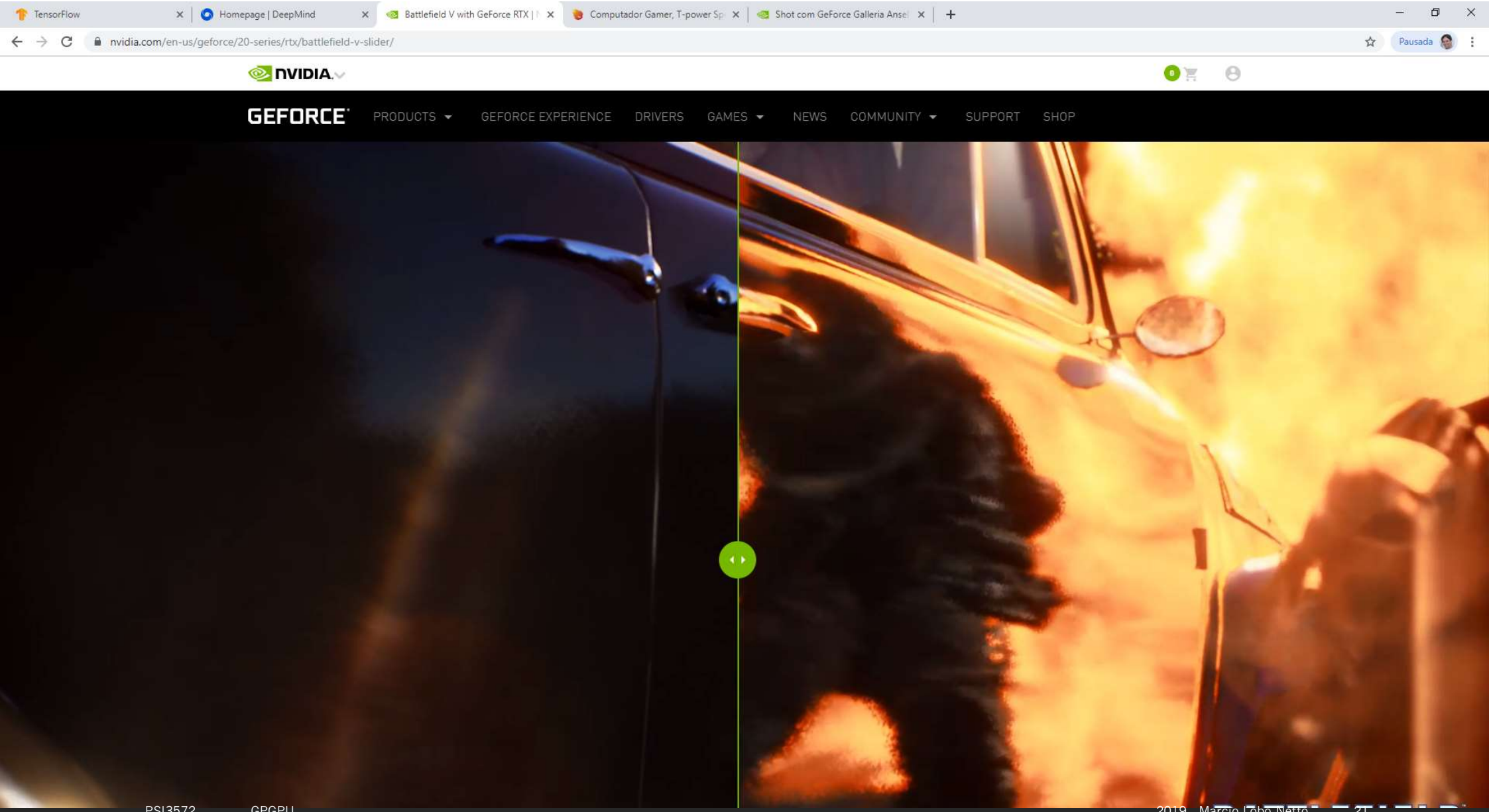
GEFORCE PRODUTOS DRIVERS GEFORCE EXPERIENCE NOTÍCIAS GAMES COMUNIDADE SUPORTE

que proporciona um funcionamento ultra frio e silencioso.

TURING VER RTX DESEMPENHO DESIGN GALERIA ESPECIFICAÇÕES NVLINK COMPRAR

GPU NVIDIA Turing

TA arquitetura Turing e a nova plataforma gráfica RTX proporcionam um desempenho até três vezes superior ao das placas de vídeo da geração anterior e trazem o poder do Ray Tracing em tempo real e da Inteligência Artificial aos games.


TensorFlow
Homepage | DeepMind
Battlefield V with GeForce RTX |
Computador Gamer, T-power Sp
Shot com GeForce Galleria Ansel
nvidia.com/en-us/geforce/20-series/rtx/battlefield-v-slider/
Pausada
NVIDIA
GEFORCE
PRODUCTS
GEFORCE EXPERIENCE
DRIVERS
GAMES
NEWS
COMMUNITY
SUPPORT
SHOP

TensorFlow | Homepage | DeepMind | Placas de Vídeo GeForce GTX Sé | Shot com GeForce Galleria Ansel

nvidia.com/pt-br/geforce/graphics-cards/gtx-1650/

Pausada

GEFORCE PRODUTOS DRIVERS GEFORCE EXPERIENCE NOTÍCIAS GAMES COMUNIDADE SUPORTE •••

TURING DESEMPENHO GEFORCE EXPERIENCE ANSEL CREATORS GALERIA ESPECIFICAÇÕES COMPRAR

GALERIA

https://www.nvidia.com/pt-br/geforce/graphics-cards/gtx-1650/#image1

Digite aqui para pesquisar

2019 Marcio Lobo Netto

22 POR PTB2 16:47 22/10/2019

TensorFlow | Homepage | DeepMind | Placa de Vídeo GeForce RTX 208 | Shot com GeForce Galleria Ansel

nvidia.com/pt-br/geforce/graphics-cards/rtx-2080-ti/

Pausada

GEFORCE PRODUTOS DRIVERS GEFORCE EXPERIENCE NOTÍCIAS GAMES COMUNIDADE SUPORTE •••

TURING VER RTX DESEMPENHO DESIGN GALERIA ESPECIFICAÇÕES NVLINK COMPRAR

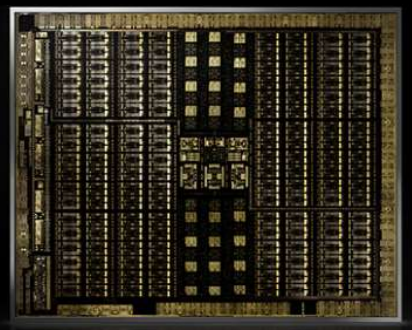
NVIDIA TURING

As placas de vídeo GeForce RTX™ contam com a arquitetura Turing e a nova plataforma gráfica RTX. Isso proporciona um desempenho até seis vezes superior ao das placas de vídeo da geração anterior e leva aos games o poder do Ray Tracing em tempo real e da AI.

ATÉ
6X
MAIS DESEMPENHO

SUPORTE A
RAY TRACING
EM TEMPO REAL NOS GAMES

A MAIS ATUAL
AI
PARA GRÁFICOS APRIMORADOS



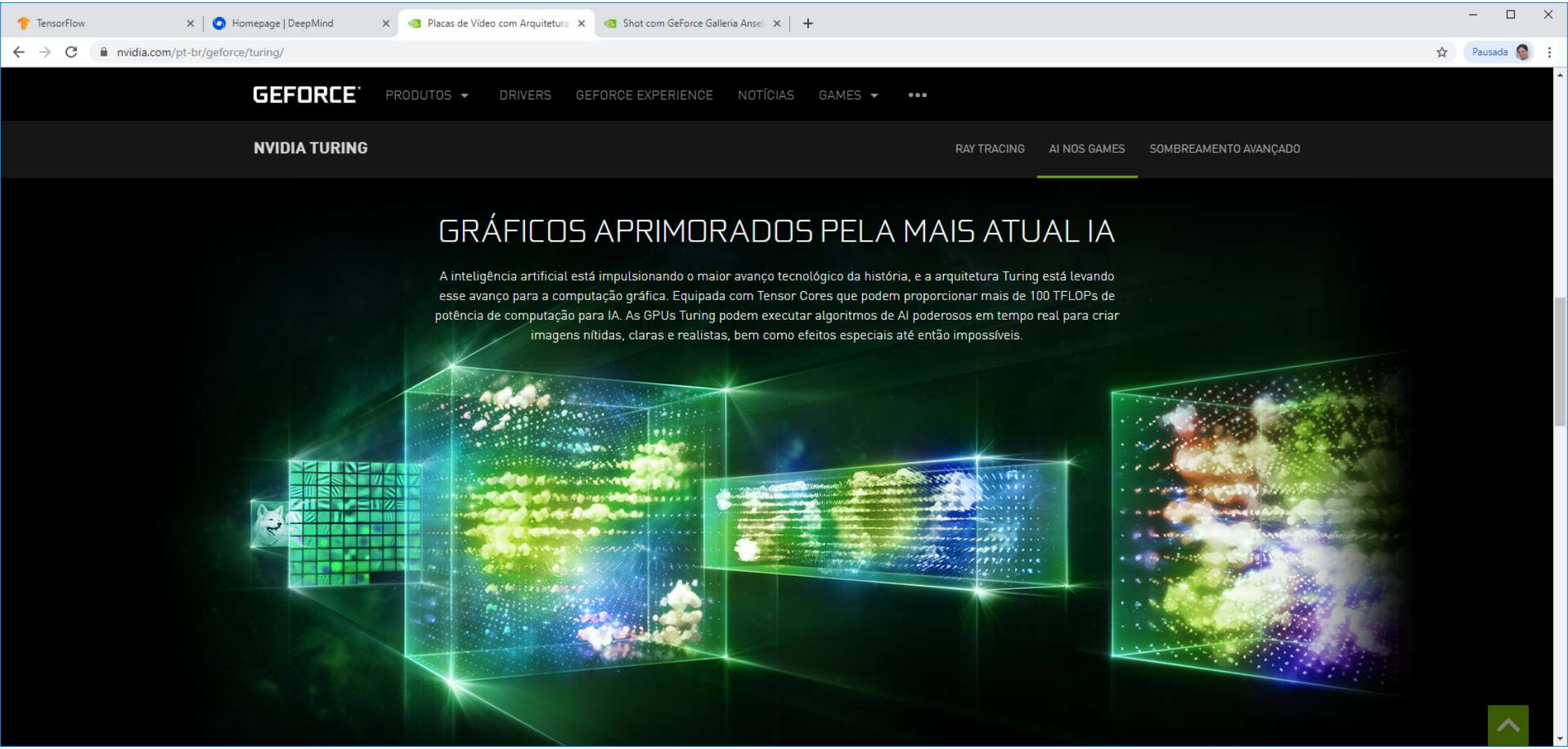
SAIBA MAIS

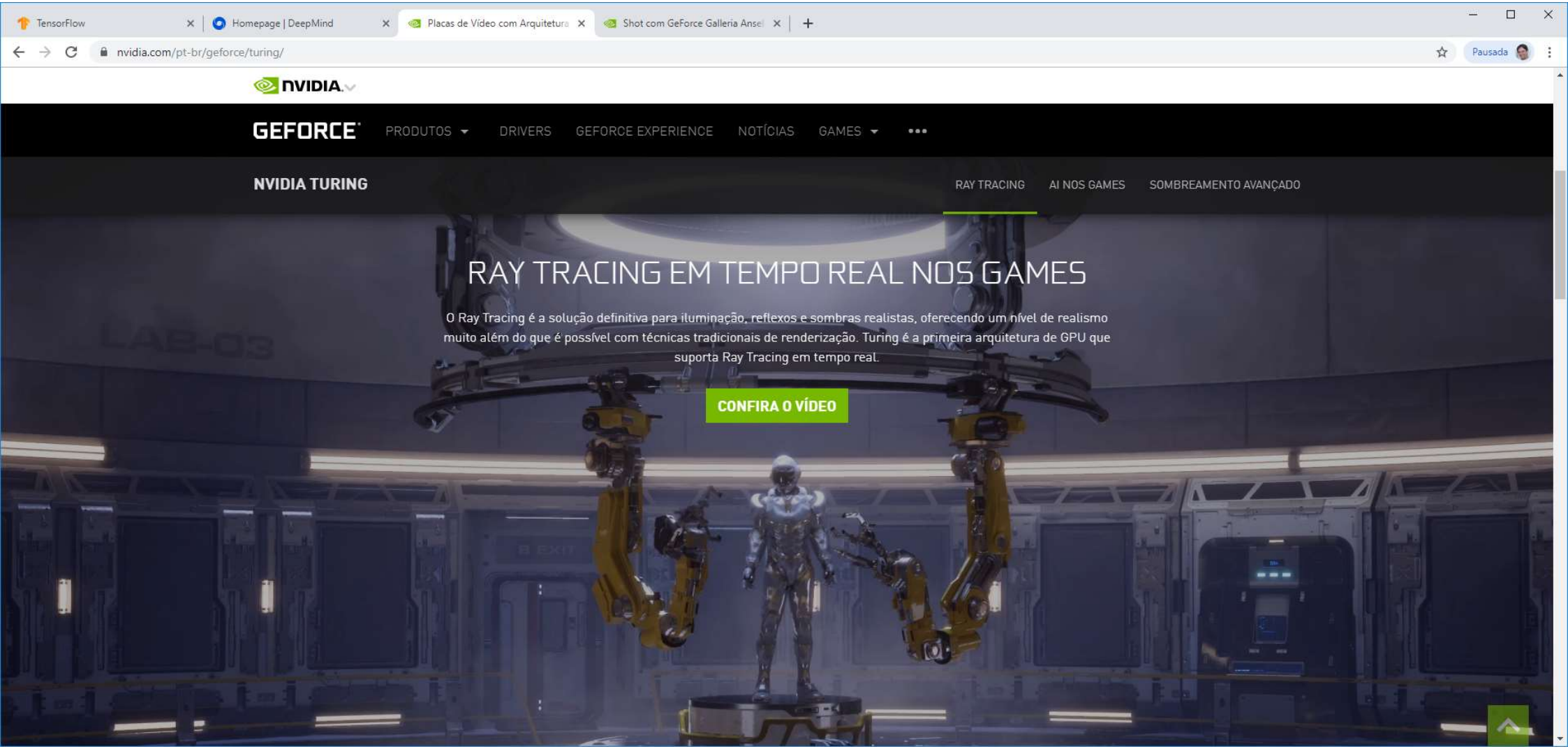
A PRIMEIRA PLACA DE VÍDEO PARA GAMES COM RAY TRACING

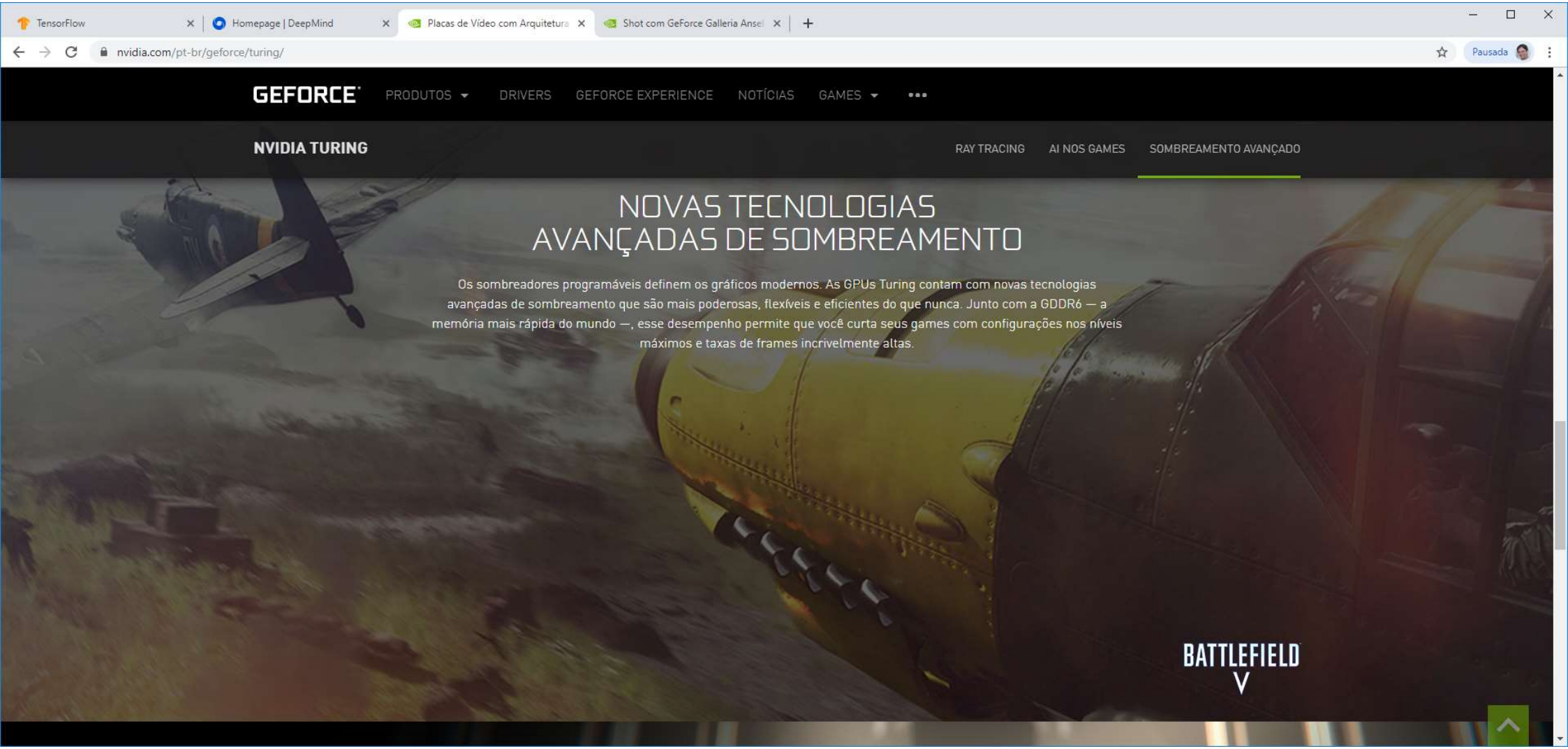
Quando o assunto são os games da próxima geração, tudo gira em torno do realismo. A GeForce RTX 2080 Ti está anos-luz à frente das outras placas, oferecendo tecnologias verdadeiramente únicas de Ray Tracing em tempo real para garantir gráficos hiper-realistas de ponta.

Digite aqui para pesquisar

POR PTB2 16:49 22/10/2019


TensorFlow
Homepage | DeepMind
Placas de Vídeo com Arquitetura
Shot com GeForce Galleria Ansel
nvidia.com/pt-br/geforce/turing/
Pausada
GEFORCE
PRODUTOS
DRIVERS
GEFORCE EXPERIENCE
NOTÍCIAS
GAMES
NVIDIA TURING
RAY TRACING
AI NOS GAMES
SOMBREAMENTO AVANÇADO
GRÁFICOS APRIMORADOS PELA MAIS ATUAL IA
A inteligência artificial está impulsionando o maior avanço tecnológico da história, e a arquitetura Turing está levando esse avanço para a computação gráfica. Equipada com Tensor Cores que podem proporcionar mais de 100 TFLOPs de potência de computação para IA. As GPUs Turing podem executar algoritmos de AI poderosos em tempo real para criar imagens nítidas, claras e realistas, bem como efeitos especiais até então impossíveis.


TensorFlow
Homepage | DeepMind
Placas de Vídeo com Arquitetura
Shot com GeForce Galleria Ansel
nvidia.com/pt-br/geforce/turing/
Pausada
NVIDIA
GEFORCE
PRODUTOS
DRIVERS
GEFORCE EXPERIENCE
NOTÍCIAS
GAMES
NVIDIA TURING
RAY TRACING
AI NOS GAMES
SOMBREAMENTO AVANÇADO
RAY TRACING EM TEMPO REAL NOS GAMES
O Ray Tracing é a solução definitiva para iluminação, reflexos e sombras realistas, oferecendo um nível de realismo muito além do que é possível com técnicas tradicionais de renderização. Turing é a primeira arquitetura de GPU que suporta Ray Tracing em tempo real.
CONFIRA O VÍDEO


TensorFlow
Homepage | DeepMind
Placas de Vídeo com Arquitetura
Shot com GeForce Galleria Ansel
nvidia.com/pt-br/geforce/turing/
Pausada
GEFORCE
PRODUTOS
DRIVERS
GEFORCE EXPERIENCE
NOTÍCIAS
GAMES
NVIDIA TURING
RAY TRACING
AI NOS GAMES
SOMBREAMENTO AVANÇADO
NOVAS TECNOLOGIAS
AVANÇADAS DE SOMBREAMENTO
Os sombreadores programáveis definem os gráficos modernos. As GPUs Turing contam com novas tecnologias avançadas de sombreamento que são mais poderosas, flexíveis e eficientes do que nunca. Junto com a GDDR6 — a memória mais rápida do mundo —, esse desempenho permite que você curta seus games com configurações nos níveis máximos e taxas de frames incrivelmente altas.
BATTLEFIELD V

CUDA

linguagem programação GPU

Linguagem tipo C

Com diretivas para explorar bem as possibilidades de paralelismo das GPUs

TensorFlow
Keras

ambiente programação GPU


Definição Topológica Rede Neural (camadas e neurônios)

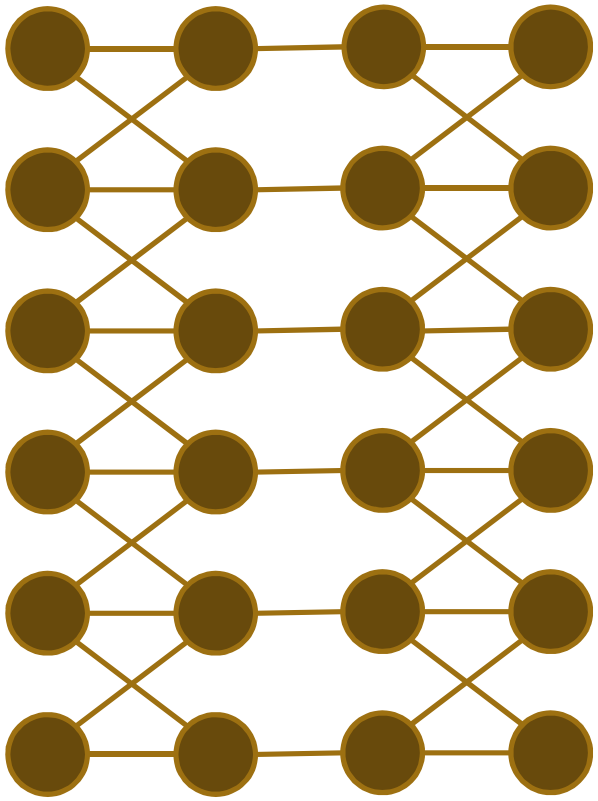

Método Aprendizagem


Treinamento


Uso

TensorFlow
Keras

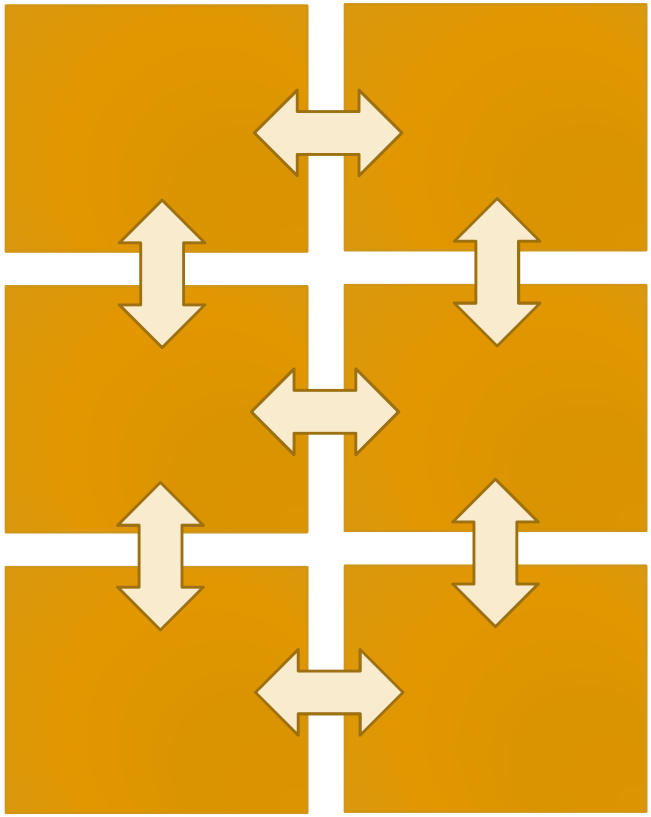
ambiente programação GPU



Rede Neurônios (unidades e conexões)

TensorFlow
Keras

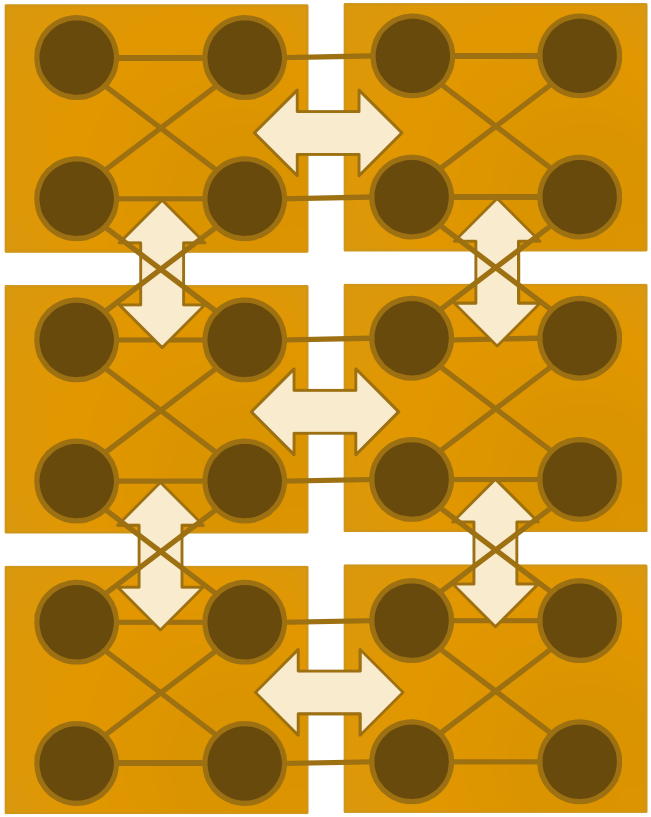
ambiente programação GPU



Rede de Processadores (GPU) (núcleos e interconexões)

TensorFlow
Keras

ambiente programação GPU



Mapeamento
neurônios ⇔ núcleos

Programação de GPUs

Como fazer bom uso do potencial computacional oferecido pelas GPUs

- Usando ferramentas disponibilizadas pelos fabricantes
  - Lingugem CUDA
  - Ambiente KERAS e TENSORFLOW

Programação de GPUs (1)

Conceitos referenciais

Em todos os casos, exploram possibilidades de programação paralela, de diferentes formas e seguindo diferentes paradigmas

Arquiteturas Específicas

- Menores possibilidades de programação
  - código dos núcleos pré programado (ex: Open GL)
- O desenvolvedor da aplicação usa comandos disponíveis na biblioteca sem maiores possibilidades de intervir na sua modificação
- O modelo de programação é mais limitado ao conceito de fluxo contínuo (streaming) onde o acesso a memórias globais / compartilhadas é mais restrito

Programação de GPUs (2)

Conceitos referenciais

Em todos os casos, exploram possibilidades de programação paralela, de diferentes formas e seguindo diferentes paradigmas

Arquiteturas Genéricas

- Maiores possibilidades de programação
- O desenvolvedor da aplicação usa comandos disponíveis na biblioteca, mas há maior liberdade para criar códigos que façam uso de memórias compartilhadas
  - As aplicações podem ser mais genéricas / modelos podem fazer uso de memórias compartilhadas em maior grau
- O modelo de programação é mais geral (como na programação de CPUs) onde o acesso a memórias globais / compartilhadas é mais aberto. A diferença é naturalmente o conceito de paralelismo, que pode exigir atenções especiais do programador

Programação de GPUs (3)

Conceitos referenciais

Em todos os casos, exploram possibilidades de programação paralela, de diferentes formas e seguindo diferentes paradigmas

Arquiteturas Programáveis

- Maiores possibilidades de programação
  - código dos núcleos podem ser programados (ex: shaders)
- O desenvolvedor da aplicação usa comandos disponíveis na biblioteca, mas pode ajusta-los aos seus propósitos ou mesmo incluir outros que venha a desenvolver

GP-GPUs
Síntese de Imagens / RayTracing

Arquiteturas Genéricas

- Família Turing RTX
- Adequadas para aplicações como RayTracing
  - Recursos adequados para cálculos globais
    - Cálculos com complexidade O2

Arquiteturas Específicas

- Fluxo contínuo (sentido único de processamento)
- Com propósito pré definido (ex: rendering – OpenGL)

GP-GPUs
Deep Learning

Arquiteturas Específicas

Fluxo contínuo (sentido único de processamento)

Com propósito geral / programável (ex: aprendizagem máquina – ANN)

- A topologia da Rede Neural pode ser mapeada na rede de núcleos de processamento (cores) disponíveis na GPU
- A função computacional pode ser atribuída aos nós / núcleos (em arquiteturas programáveis)

- Alto paralelismo intrínseco a GPU e a flexibilidade de conectividade entre núcleos (caso das arquiteturas mais gerais) as torna ótimas plataformas para computação de Redes Neurais
  - Machine Learning e Deep Learning

Conclusão

Discussão