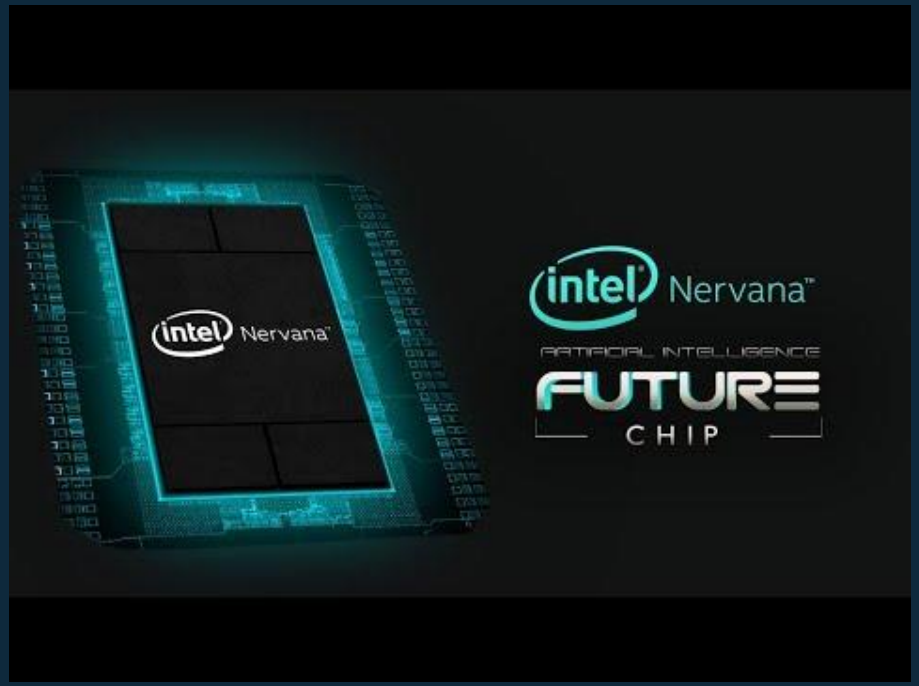
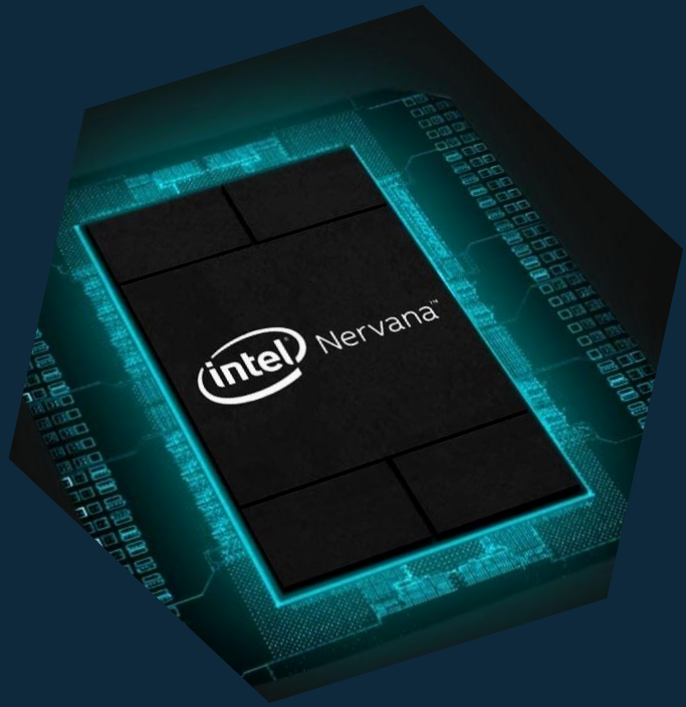
Intel Nervana

Nervana Neural Processor?



Começo do NNP



- Anunciado em um evento organizado pela própria Intel, o *Nervana Neural Processor* (NNP), anteriormente chamado de Lake Crest, feito a base de silício, faz parte da primeira geração da Intel para Redes Neurais.

- O NNP almeja ser poderoso o bastante para lidar com os requisitos computacionais intensivos da execução de redes de *deep learning*.

200bi$

Fatia de Mercado Prevista pela Intel

100x Performance

Prometem entregar 100x treino de performance de IA que placas graficas concorrentes

2019

Lançamento previsto para o final de 2019

“Antes do final da década, a Intel oferecerá um desempenho 100x maior, que vai turbinar o ritmo de inovação em aprendizagem de máquina”

- Diane Bryant, vice-presidente executiva de DataCenter da Intel

1.

Conceitos Base

Conceitos base para melhor entendimento das especificações e arquitetura

GPU para Deep Learning

GPUs se provaram mais apropriadas para *deep learning*. Elas foram inicialmente projetadas para *video games*, e o movimento de objetos na tela se dão através de vetores e álgebra linear, conceitos básicos de redes neurais.

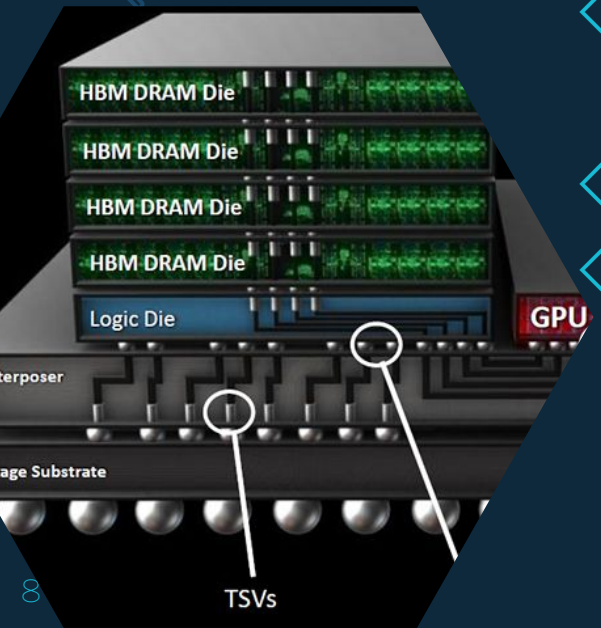
CPU

Possui poucos núcleos, próprios para computações complexas.

GPU

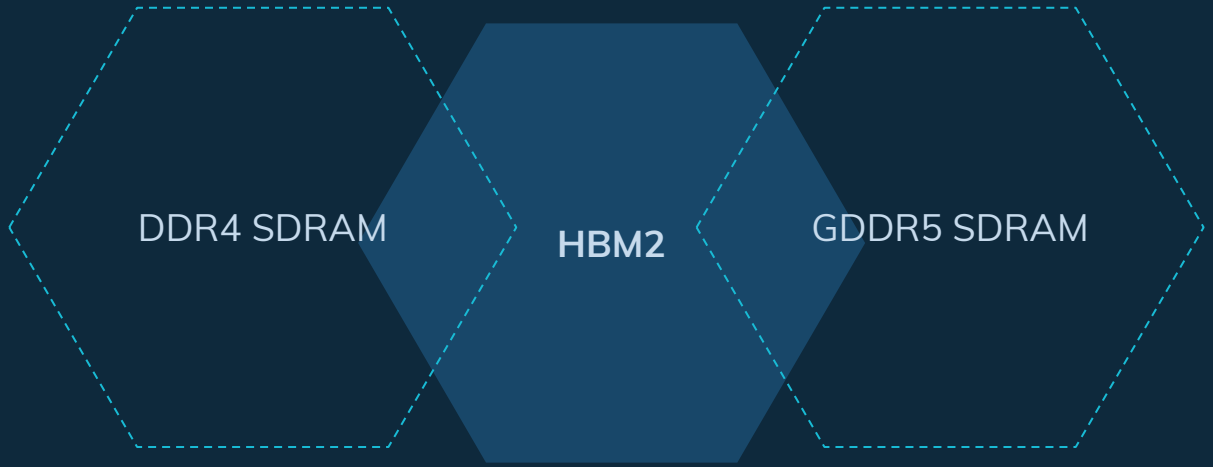
Possui uma grande quantidade de núcleos, apesar de menores. Ideal para numerosas operações, simples e similares, em paralelo

HBM2



- Utiliza menos energia e disponibiliza maior banda.
- Maior barramento de memória.
- Utiliza 2 canais de 128 bits por dado

HBM2

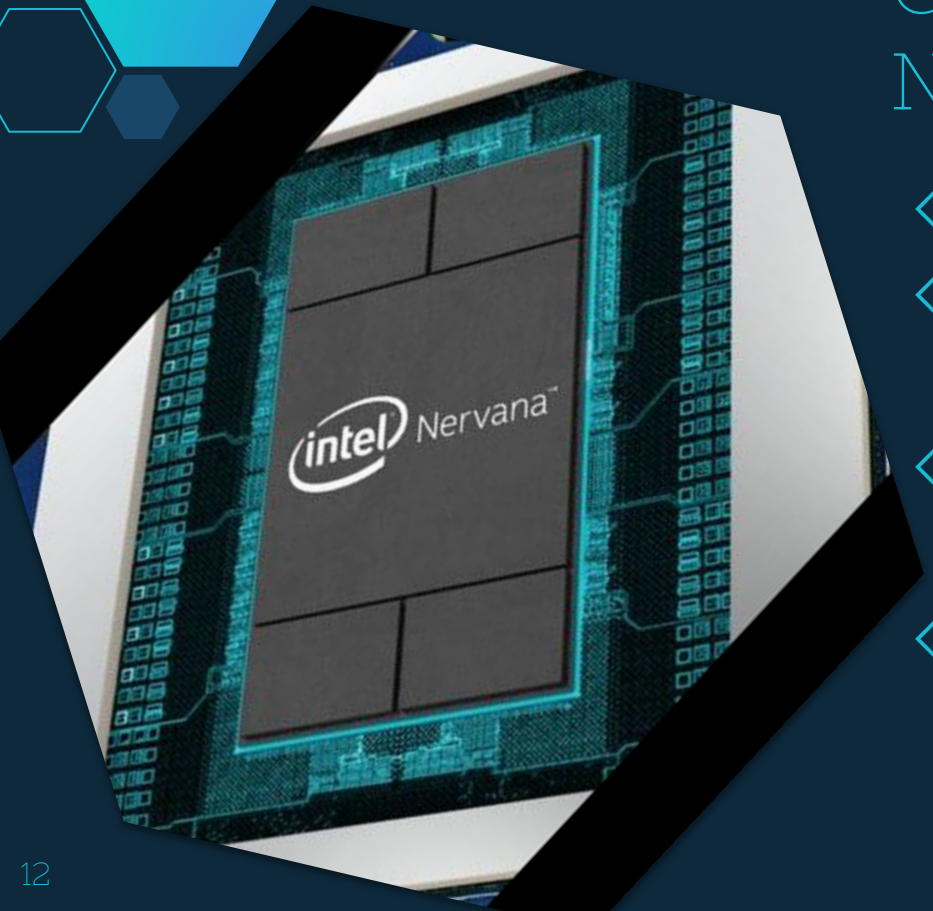


2. Especificações

Dimensões Físicas e questões de Processamento

Dimensões e Curiosidades

- Aproximadamente 102mm x 165mm.
- Projetado com circuito integrado de aplicação 28nm com espaço para mais!
- Consumindo menos de 210 Watts.
- Desenvolvido para facilitar sua integração com outros processadores



Configuração do Nervana

- Chip equipado com 12 núcleos
- Cada núcleo possui 2MB de memória cache.
- Possui 4 HBM2 totalizando 32GB de memória
- Banda de 1TB/s

3.

Arquitetura

Estruturação do *Nervana Neural Processor*

Caracteristicas

- Cada núcleo contém 2 unidades matemáticas de *deep learning*
- FlexPoint maximiza a precisão em 16 bits
- Desenvolvido para minimizar questões energéticas

Caracteristicas

- Ausência de hierarquia de memória cache
- Poder de Processamento de 40 TOps
- Pipelines separadas
- “Paralelismo”

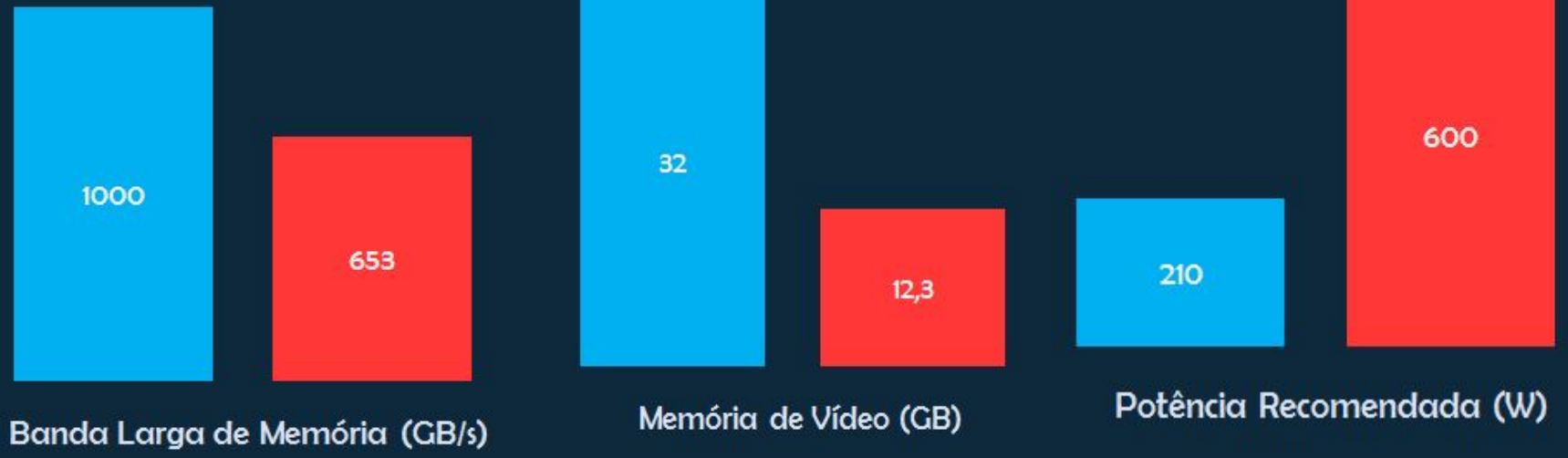
Estratégias Efetivadas

- Redução da necessidade de transferência de dados
- Memoria HBM alocada proximo aos núcleos
- Tensores divisiveis
- Novo Processo 2.5D *manufacturing*

Paralelismo

- 6 links bidimensionais de alta banda larga
- Permite expansão de seus módulos sem perda de velocidade significativa
- Não precisa de dispositivos terceiros para comunicação entre máquinas

Intel Nervana x Titan V



4. Considerações Finais

Expectativas para o NNP?

Obrigado!

Perguntas?

Equipe:

Raphael de Jesus	16.00378-0
Júlia Catarina	16.00645-3
Lucas Primati Menezes	16.00683-6
Guilherme Tagliati	17.00375-0
Breno Thomaz	17.00815-8

Referências

Essa apresentação leva em conta dados e fatos retirados das fontes a seguir, os créditos das imagens vão ao seus respectivos autores.

https://www.intel.ai/nervana-engine-delivers-deep-learning-at-ludicrous-speed/#gs.22xqzd

https://www.servethehome.com/intel-nervana-nnp-l-1000-oam-and-system-topology/

https://newsroom.intel.com.br/news-releases/intel-aposta-em-tecnologia-para-inteligencia-artificial/#gs.25ugzh

https://www.intel.ai/nervana-engine-delivers-deep-learning-at-ludicrous-speed/#gs.1yd09j

https://youtu.be/zEzm-rMwyVo