SSC0510 - Arquitetura de Computadores

Nervana Intel Chip: Deep Learning Chip Architecture

Letícia Lumi Asano	9292841
Leonardo Xavier de Almeida	8658395
Rafael Rozendo	8670020

O projeto Nervana

A Nervana Systems, foi fundada em **2014** pelo **CEO Naveen Rao**, pelo **CTO Amir Khosrowshahi** e pelo **VP Algorithms Arjun Bansal**.

É uma empresa de **software de inteligência artificial**, e fornece uma plataforma completa de software como serviço chamada **Nervana Cloud**, que permite que as empresas desenvolvam **software de *deep learning* personalizado**.

Foi adquirida pela **Intel em 2016** por aproximadamente **US$ 400 milhões**

Limitações

Os grandes desafios da Inteligência Artificial estão relacionados com:

- O **poder de processamento** que é necessário para executar certos programas que trabalham com **grandes quantidades de dados**.

- Tempo de resposta para exibir resultados

Limitações

No início, a computação teve grandes investimentos em hardware, que se tornou saturado. Com isso, o enfoque passou a ser o investimento em software.

O Nervana trouxe uma proposta diferente, voltada novamente para o hardware e sua melhor interação com o software, a fim de oferecer maior eficiência em IA e Deep Learning

Objetivo do Processador Nervana

Possui arquitetura de propósito específico para deep learning.

O objetivo dessa nova arquitetura é fornecer a flexibilidade necessária para oferecer suporte a todas as primitivas de deep learning e, ao mesmo tempo, tornar os componentes de hardware essenciais o mais eficiente possível.

As necessidades da I.A e hardware

Necessidade:
Multiplicação de matrizes e Convolução.

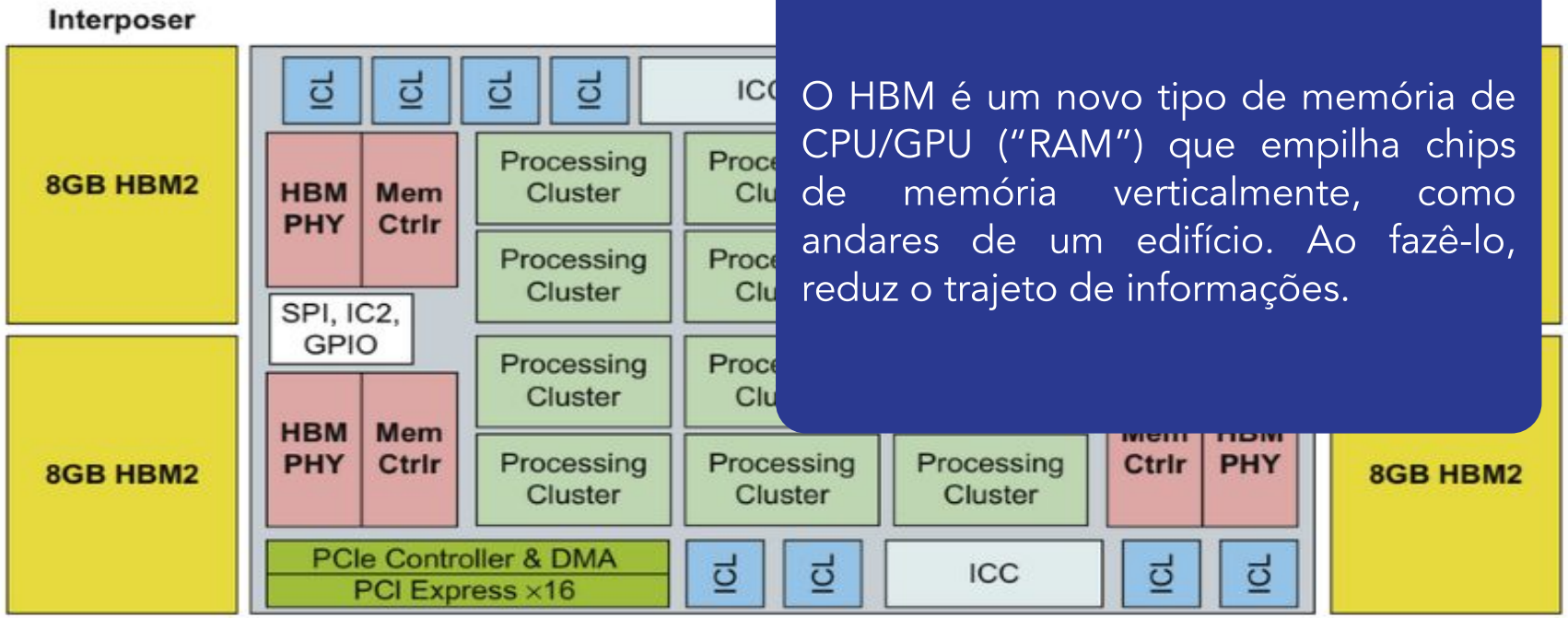
Impacto:

O NNP da Intel Nervana não possui uma hierarquia de cache padrão e a memória no chip é gerenciada diretamente pelo software.

Estrutura de um Processador IA Nervana

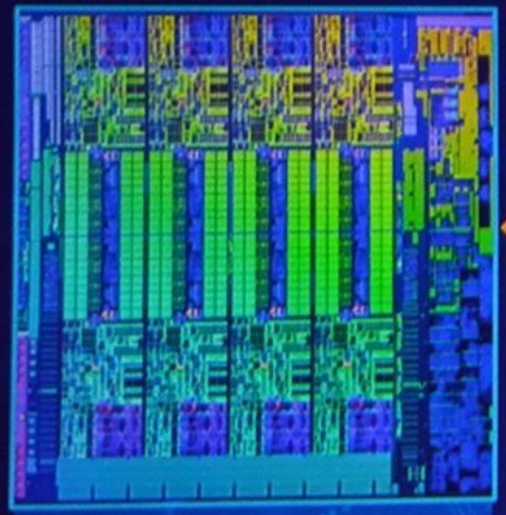
- Suporta o paralelismo de modelo verdadeiro
- Cada nó do computador tem interface de memória própria
- Memória de Entrada e Saída aumentada

Estrutura de um Processador IA Nervana



O HBM é um novo tipo de memória de CPU/GPU ("RAM") que empilha chips de memória verticalmente, como andares de um edifício. Ao fazê-lo, reduz o trajeto de informações.

MODEL AND SUBSTRATE FOR COMPUTATION

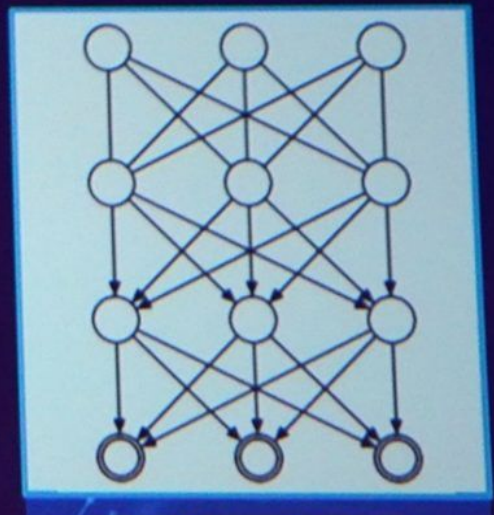


WORKLOAD OPTIMIZED HARDWARE

?

```
model: &obj:models.MLP {
  num_epochs: 30,
  batch_size: &bs 100,
  layers: [
    &datalayer !obj:layers.DataLayer {
      name: d0,
      nout: 784,
    },
    !obj:layers.FCLayer {
      name: h0,
      nout: 100,
      lrule_init: *gdr,
      weight_init: *at_init,
      activation: !obj:transforms.RectLin {},
    },
    &lastlayer !obj:layers.FCLayer {
      name: output,
      nout: 10,
      lrule_init: *gdr,
      weight_init: *at_initb,
      activation: !obj:transforms.Logistic {},
    },
    &costlayer !obj:layers.CostLayer {
      name: cost,
      ref_layer: *datalayer,
      cost: !obj:transforms.CrossEntropy {},
    },
  ],
},
```

?



DEEP LEARNING MODEL

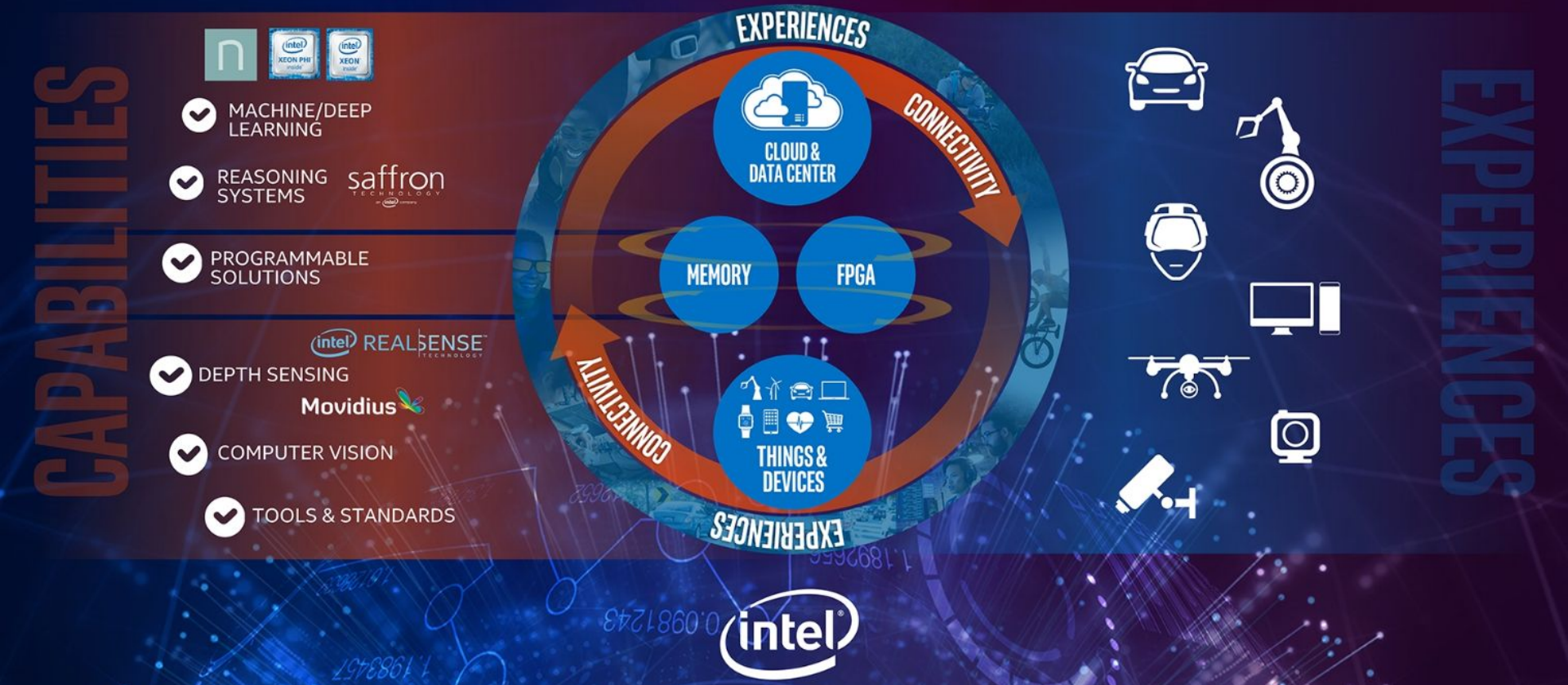
- MODEL DESCRIPTION LANGUAGE
- HARDWARE ABSTRACTION LAYER
- DISTRIBUTED PRIMITIVES
- COMPILERS, DRIVERS

Aplicações

O NNP (Nervana Neural Network Processor) tem um impacto em áreas como:

- Saúde (diagnósticos mais rápidos e certeiros para doenças como mal de Alzheimer)
- Mídias sociais (mirar o público mais adequado para um ação)
- Indústria automotiva (veículos autônomos)
- Clima (previsões com base em dados - big data)

DELIVERING INTELLIGENCE IN THE AI AGE



Evolução

O processador Nervana, além de oferecer mapeamento na casa de teraflops para teraops para o NNP, se mostrou de 5x a 6x mais rápido e eficiente quando comparado com a arquitetura Pascal, utilizada para operações de IA. “Conhecemos pessoas usando o Pascal agora para *deep learning*, então temos um bom senso dos números. Eles introduziram o NVLink, mas a partir do zero, projetamos uma solução com vários chips. Nosso chip (Nervana) tem uma velocidade muito maior e um conjunto dedicado de links seriais entre os chips. Também temos construções de software para permitir que os chips atuem como um grande chip. Nós jogamos fora a bagagem em vez de mergulhar em uma hierarquia de memória existente e construímos algo apenas para esse problema. ”

INTEL® NERVANA™ PORTFOLIO

Common architecture for AI implementations

Most widely deployed machine learning solution

High performance, general purpose machine learning

Programmable, low-latency inference

Best in class neural network performance[1]

TARGETED ACCELERATION

Other names and brands may be claimed as the property of others

1. Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance. Competitive performance data based on publicly available projections.

INTEL® NERVANA™ PLATFORM
FOR DEEP LEARNING

ANNOUNCING

LAKE CREST

Discrete accelerator
First silicon 1H'2017

ANNOUNCING

KNIGHTS CREST

Bootable Intel Xeon Processor
with integrated acceleration

DELIVERING 100X REDUCTION IN TIME TO TRAIN BY 2020

COMPARED TO TODAY'S FASTEST SOLUTION[1]

1 Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance

Obrigado!

Referências

1. Introdução à Inteligência Artificial - IME/USP - https://www.ime.usp.br/~slago/IA-introducao.pdf
2. Nervana: Intel prepara lançamento de processador para IA - https://goo.gl/WrR1GT
3. "Deep learning vai muito além de reconhecer gatos e cachorros em fotos" - https://goo.gl/kb1zWq
4. INTEL, NERVANA SHED LIGHT ON DEEP LEARNING CHIP ARCHITECTURE - https://goo.gl/4VgFvo
5. 5 requisitos para aplicar inteligência artificial e machine learning no futuro - http://enginebr.com.br/5-requisitos-para-aplicar-inteligencia-artificial-e-machine-learning-no-futuro.html