NVIDIA TESLA V100 GPU - Arquitetura Volta

Kleber Yuji Inoue - 8604297
Oton Papa - 9292883
Victor França - 9790781

Acrônimo de Compute Unified Device Architecture - *CUDA*

- Extensão para a linguagem de programação C,C++,Java,Fortran e Python
- Possibilita o uso de computação paralela, possibilitando com que a GPU consiga realizar algumas operações com mais rapidez, como processamento de vídeo, análises sísmicas, simulações de dinâmica dos fluidos e previsão do tempo.
- Desvantagens:
  - Renderização de texturas não é suportada pela CUDA.
  - Diversas cópias realizadas entre uma memória e outra pode causar impactos no desempenho geral das aplicações, mas varia de acordo com o barramento
  - Funciona apenas em placas da NVIDIA.

Evolução Histórica

- NVIDIA Tesla C870 GPU
  - Lançado em 2 de maio de 2007
  - Possui um núcleo de computação com 128 processadores
  - Frequência (Clock): 1.35 GHz
  - Ambiente de desenvolvimento para a GPU baseado na linguagem C (CUDA)
  - Memória dedicada 1.5 GB
  - Largura de banda: 76.8 GB/s
  - Possui 681 milhões de transistores

Fonte: https://cnet1.cbsistatic.com/img/upg5P49FKJ3dayctbsq3FivQmQM=/830x467/2010/04/28/7dd31920-f8a0-40e4-a290-aad191866f46/7dd31920-f8a0-40e4-a290-aad191866f46.jpg

Evolução Histórica

- NVIDIA Tesla D870 GPU
  - Lançado em 2 de maio de 2007
  - Dois NVIDIA Tesla C870 GPU com 128 processadores em cada GPU
  - Frequência (Clock): 1.35 GHz
  - Ambiente de desenvolvimento para a GPU baseado na linguagem C (CUDA)
  - Memória dedicada: 1.5 GB em cada GPU (3 GB)
  - Largura de banda: 2x 76.8 GB/s

Fonte: https://www.techpowerup.com/gpu-specs/images/c/1543.jpg

Evolução Histórica

- NVIDIA Tesla P100 GPU
  - Lançado em junho de 2016
  - Possui 15.3 bilhões de transistores
  - Frequência (Clock): 1.12 GHz
  - Largura de banda: 732 GB/s

Fonte: https://images-na.ssl-images-amazon.com/images/I/41LnQd8YRQL._SX425_.jpg

Estado da Arte

- NVIDIA Tesla V100 GPU (PCIe)
  - Lançado em 21 de junho de 2017
  - Possui 21.1 bilhões de transistores em um espaço de 815mm²
  - Frequência (Clock): 1.37 GHz
  - 640 Tensor Cores
  - 5020 CUDA Cores
  - Largura de Banda 900GB/s

Fonte: https://www.nvidia.com/content/dam/en-zz/es_em/Solutions/Data-Center/tesla-v100/data-center-tesla-v100-pcie-625-ud@2x.jpg

Estado da Arte

- NVIDIA Tesla V100 GPU (NVLINK)

Otimização da versão PCIe versão IBM POWER x86;

- Integrado a PCI Express - barramento encontrado em placas-mãe utilizados em expansão gráfica, som , rede;
- Interconexão Nvlink - otimização do tempo de processamento, desbloqueio total da GPU;
- Aumento de throughput;
- Recursos de memória unificada;
- GPUS ⅓ do tamanho padrão;
- Otimização do gargalo da versão Pcle;

Fontes: https://www.em360tech.com/wp-content/uploads/2017/09/nvidia-tesla-v100-nvlink.jpg
https://videocardz.com/50056/nvidia-launches-nvlink-worlds-first-high-speed-gpu-interconnect

Comparativo de Performance:

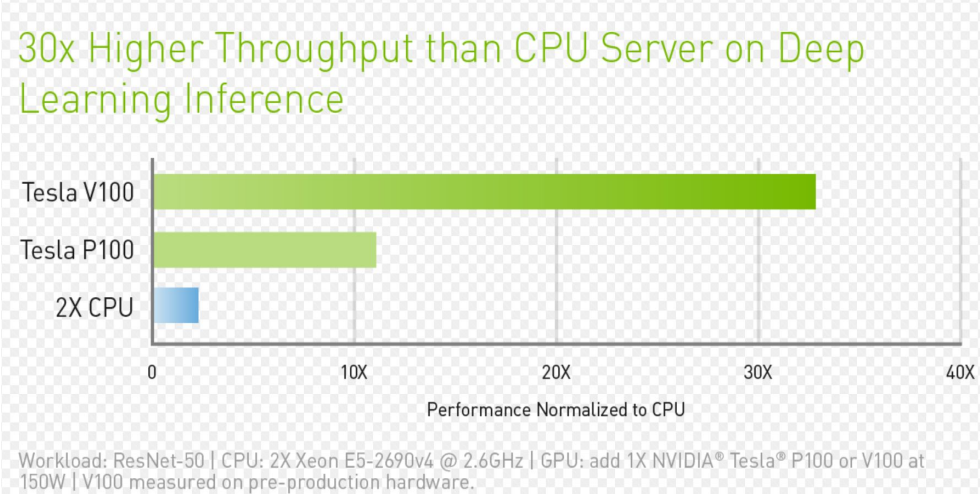
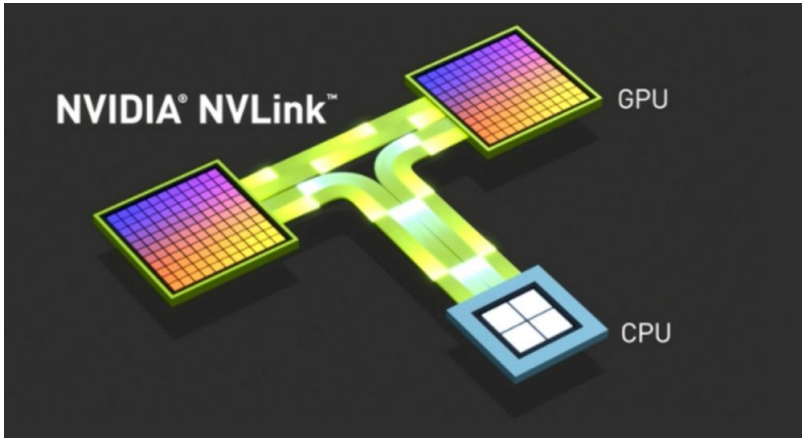


Ilustração da NVLink com 2 GPUs:



Fontes:https://www.fudzilla.com/news/graphics/41420-nvidia-nvlink-2-0-arrives-in-ibm-servers-next-year

NVidia arquitetura Volta

- Volta, um codinome de microarquitetura GPU desenvolvido pela Nvidia, sucede a Pascal. Foi anunciado pela primeira vez como uma futura ambição de roadmap em março de 2013. No entanto, o primeiro produto não foi anunciado até maio de 2017.
- A arquitetura é nomeada após Alessandro Volta. Foi o primeiro chip da NVIDIA a apresentar núcleos Tensor.
- Núcleos especialmente projetados que possuem desempenho de aprendizado profundo superior ao dos núcleos regulares CUDA.
- Inicialmente focado para a utilização de carros autônomos.

NVIDIA Arquitetura Volta - V100 Composição

A arquitetura de computação de placa de vídeo mais poderosa do mundo, criada para impulsionar a próxima onda de avanços em inteligência artificial e computação de alto desempenho.

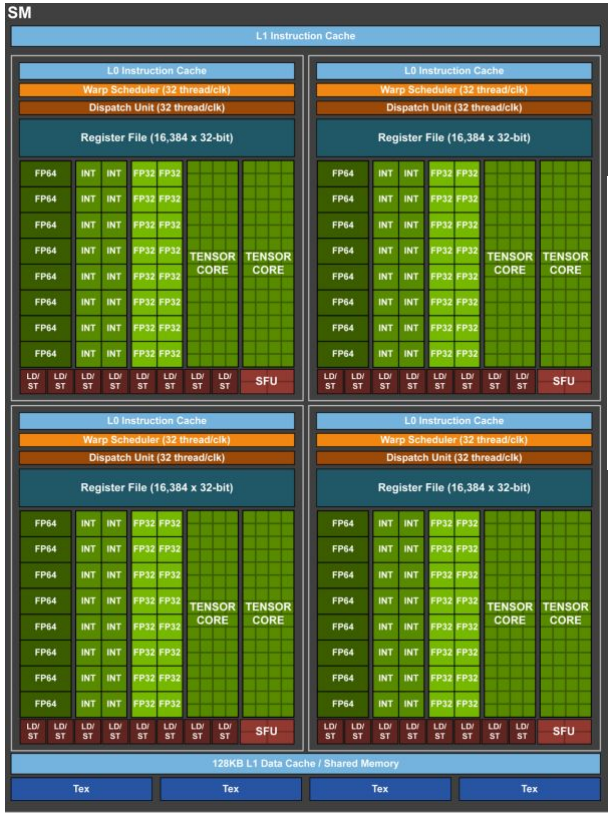
Tecnologias avançadas

A placa de vídeo Tesla V100 supera as gerações anteriores de placas de vídeo NVIDIA com tecnologias inovadoras que permitem quebrar a barreira de 100 teraflops de desempenho de deep learning. Ela inclui:

- **Tensor Cores** combinada a CUDA cores, projetada para acelerar as cargas de trabalho de AI. Equipada com 640 Tensor Cores, a V100 oferece 120 teraflops de desempenho de deep learning, equivalente ao desempenho de 100 CPUs.
- **Nova arquitetura de placa de vídeo** com mais de 21 bilhões de transistores. Combina núcleos CUDA e Tensor Cores em uma arquitetura unificada, fornecendo o desempenho de um supercomputador de AI em uma única placa de vídeo.
- **NVLink™** fornece a próxima geração de placas de vídeo de link de interconexão com alta velocidade e placas de vídeo para CPUs, com até 2 vezes o throughput da geração anterior NVLink.
- **900 GB/s HBM2 RAM**, desenvolvida em colaboração com a Samsung, atinge 50% mais largura de banda de memória do que as placas de vídeo da geração anterior, essencial para suportar a extraordinária produção computacional da Volta.
- **Software otimizado para Volta**, incluindo software CUDA, cuDNN e TensorRT ™, cujos frameworks e aplicativos principais podem facilmente ser usados para acelerar a AI e a pesquisa.

Fonte: https://www.nvidia.com.br/object/prbr_08072017b.html

Volta Tensor Cores



Volta GV100 Streaming Multiprocessor (SM)

$$D = \begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix} \begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{pmatrix} + \begin{pmatrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{pmatrix}$$

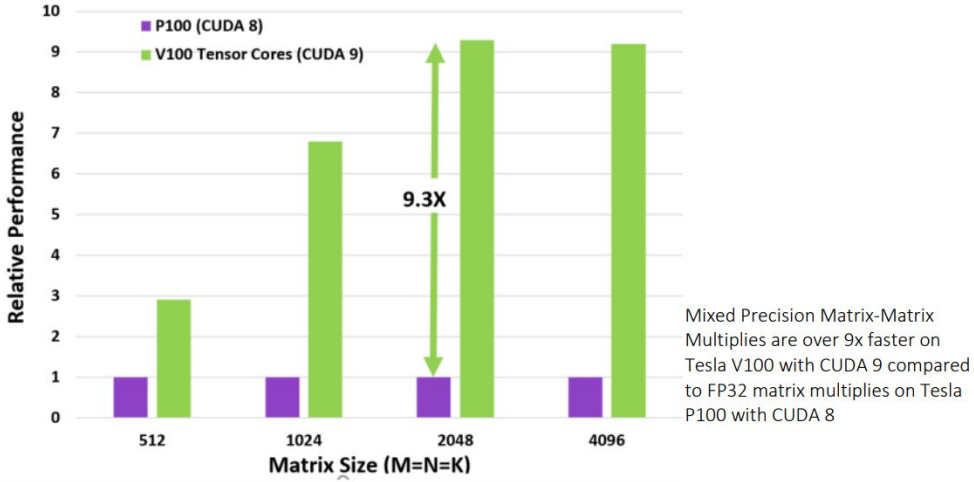
FP16 or FP32 | FP16 | FP16 | FP16 or FP32

BLAS Level 3

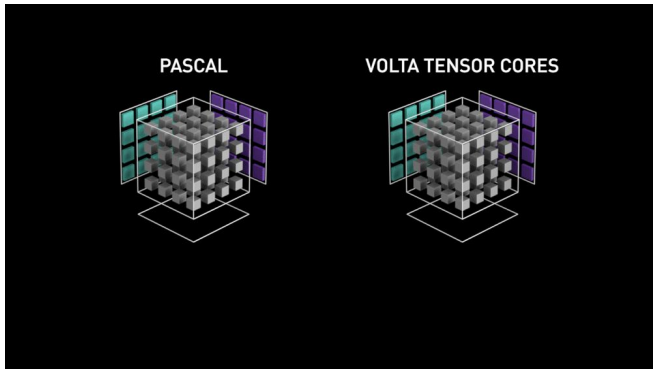
Fontes:
http://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf

Volta Tensor Cores

- Voltados para o processamento de rotinas direcionadas a aplicações de IA e deep learning.
- Cada Tensor Core fornece uma matriz de 4x4 que realiza a seguinte operação:
  - D = (A * B) + C
- Volta Tensor Core realiza as operações em paralelo, realizando as operações ao mesmo tempo.

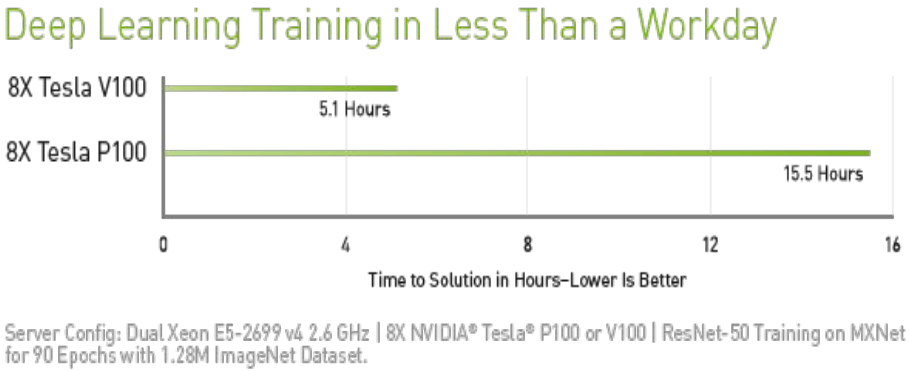


Mixed Precision Matrix-Matrix Multiplies are over 9x faster on Tesla V100 with CUDA 9 compared to FP32 matrix multiplies on Tesla P100 with CUDA 8



Fontes: https://www.nvidia.com/en-us/data-center/tensorcore/
https://devblogs.nvidia.com/programming-tensor-cores-cuda-9/

Principais resultados alcançados



A GPU entrega 125 TFLOPS (Tera Floating-points Operations Per Second). Portanto os desenvolvedores podem executar algoritmos de deep learning usando FP16 e FP32.

Por utilizar o Volta Tensor Core, a GPU é 3x mais eficiente que a versão anterior (NVIDIA Tesla P100).

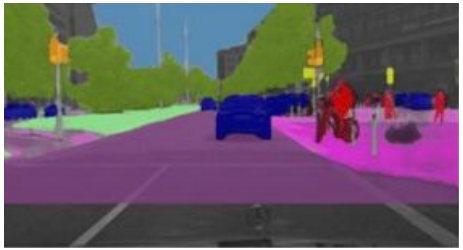
Fonte: https://www.nvidia.com/en-us/data-center/tensorcore/

Contribuição Para Sociedade

O princípio de criação está atrelado a deep learning e contribuições para as tarefas do dia a dia, que requerem muitas análises para a tomada de decisão mais eficiente.

Podemos destacar:

- Treinamento de carros autônomos;
- Assistente pessoal virtual, exemplo: Cortana, Siri;
- Reagir o mais próximo a realidade;
- “ Refazer “ imagens danificadas/baixa resolução;
- Utilizada para auxílio na saúde, exemplo: como curar o câncer;
- Aplicativos de Inteligência Artificial, e auxílio de data centers;



*Vídeo:https://www.youtube.com/watch?time_continue=87&v=5TUk5BtM0Bc

Fonte: http://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf

Referências Bibliográficas

https://images.nvidia.com/content/technologies/volta/pdf/tesla-volta-v100-datasheet-letter-fnl-web.pdf

https://www.nvidia.com/en-us/data-center/tesla-p100/

https://la.nvidia.com/docs/IO/43395/D870-SystemSpec-SP-03718-001_v01.pdf

https://www.nvidia.com.br/object/tesla_d870_br.html

https://la.nvidia.com/docs/IO/43395/c870-BoardSpec_BD-03399-001_v04.pdf

https://www.nvidia.com.br/object/tesla_c870_br.htm

https://canaltech.com.br/hardware/O-que-e-a-CUDA/ https://www.oficinadanet.com.br/post/14818-o-que-e-cudal

https://www.tecmundo.com.br/computacao-grafica/10507-nvidia-cuda-o-que-e-e-como-funciona.htm

Referências Bibliográficas

https://www.tecmundo.com.br/computacao-grafica/10507-nvidia-cuda-o-que-e-e-como-funciona.htm

https://canaltech.com.br/hardware/O-que-e-a-CUDA/ https://www.oficinadanet.com.br/post/14818-o-que-e-cuda

https://images.nvidia.com/content/pdf/v100-application-performance-guide.pdf

https://www.nvidia.com/en-us/data-center/volta-gpu-architecture/ https://www.nvidia.com.br/object/prbr_08072017b.html

https://en.wikipedia.org/wiki/Nvidia_Tesla