Uso de FPGAs na implementação e treinamento de redes neurais

Raphael Costa
Rafael Corsi

Insper 2019 - Eng. Computação

Outline

- Objetivos desta iniciação
- Por que abordar o tema FPGAs
- Empecilhos iniciais (não tão iniciais assim)
- Ferramentas
- Como acelerar código e OpenCL
- A entrega final
- Dúvidas

A iniciação Tecnológica

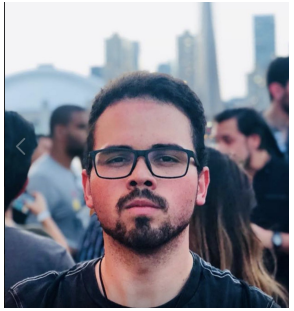
01 de Agosto de 2018
á
13 de Agosto de 2019

Quem somos nós

Rafa Corsi

- Professor Adjunto Insper desde 2016
- Especialidade em sistemas embarcados

- É graduado em Engenharia Elétrica com ênfase em eletrônica pelo Instituto Mauá de Tecnologia e mestre em Engenharia Elétrica pela Escola Politécnica da USP



Raphael Costa (Eu)

- Engenharia de Computação (7o semestre)

- Previsão de formação: 2020

- Já trabalhou como assistente de pesquisa na área de mineração de dados

- Áreas de interesse: IoT, Segurança da informação e Computação embarcada.

Tema original

- Uso de FPGAs na implementação e treinamentos de redes neurais
- Foco inicial: 50% FPGAs e 50% Machine Learning
- FPGAs exigiam um maior cuidado, enxergamos uma potencial entrega que fazia mais sentido que a inicial

Entrega

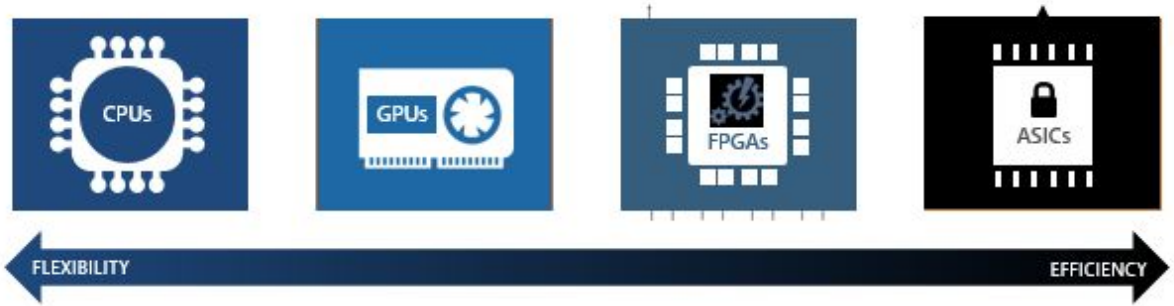
- Mudança para: Intro à aceleração de código utilizando OpenCL com FPGAs.
- Como? Tutoriais

Por que falar de FPGAs?

Silicon alternatives

TRAINING
CPUs and GPUs, limited FPGAs,
ASICs under investigation

EVALUATION
CPUs and FPGAs,
ASICs under investigation



Aplicações

https://www.xilinx.com

Últimas atualizações na área

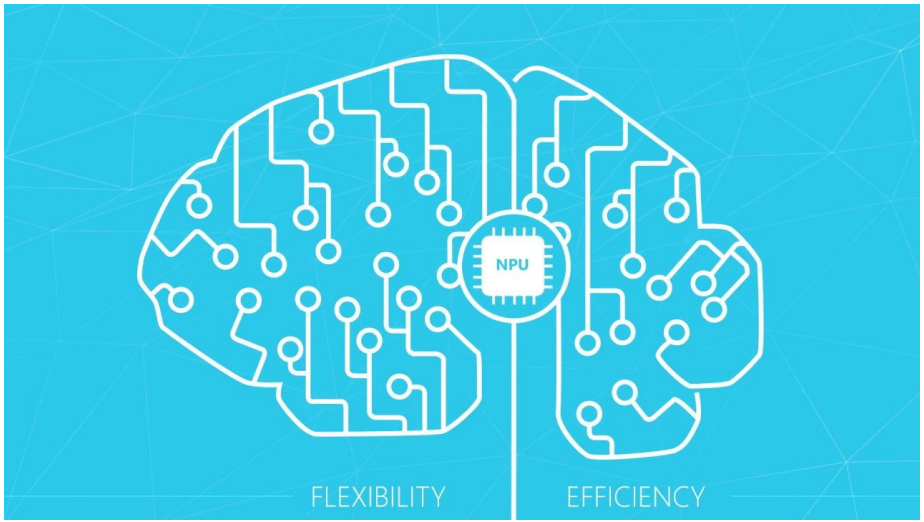
Instâncias F1 do Amazon EC2

Habilitam o desenvolvimento e a implantação com maior rapidez do acelerador de FPGA na nuvem

https://aws.amazon.com/pt/ec2/instance-types/f1/

Project Brainwave, Microsoft

O Projeto Brainwave é uma plataforma para deep learning que fornece "Inteligências Artificiais" (soluções) diretamente da nuvem.



https://www.microsoft.com/en-us/research/project/project-brainwave/

Apple - Afterburner - Mac Pro 2019

The difference between Apple's Afterburner and "ordinary" GPUs

The Afterburner is not an ordinary GPU, but an FPGA based accelerator.

A Field-programmable gate array (often shortened to FPGA) is an electronic component used to build reconfigurable digital circuits. That means that an FPGA is different from a logic gate because a logic gate has a fixed function. Before the FPGA can be used in a circuit, it must be programmed, that is, reconfigured. In simple words, an FPGA could be programmed to do a specific task. In more simple words, it's like a utilized GPU for editing purposes. It's not a GPU per se because the architecture and roles are different. However, you can think of it as a more efficient version of the RED Rocket (and hopefully more affordable) that is committed to high-resolution native rendering.

https://www.apple.com/newsroom/2019/06/apple-unveils-powerful-all-new-mac-pro-and-groundbreaking-pro-display-xdr/

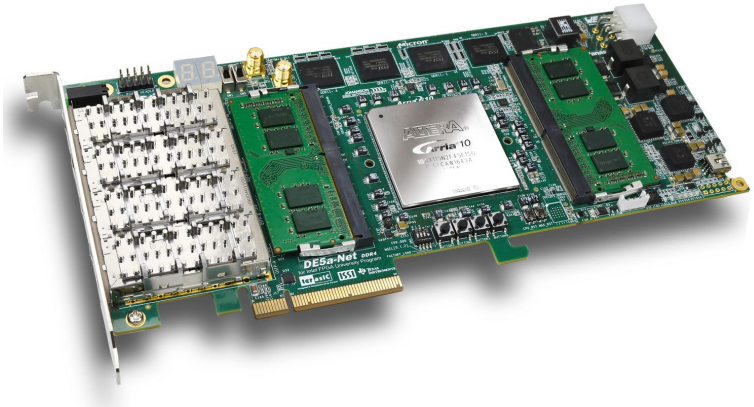
Projeto

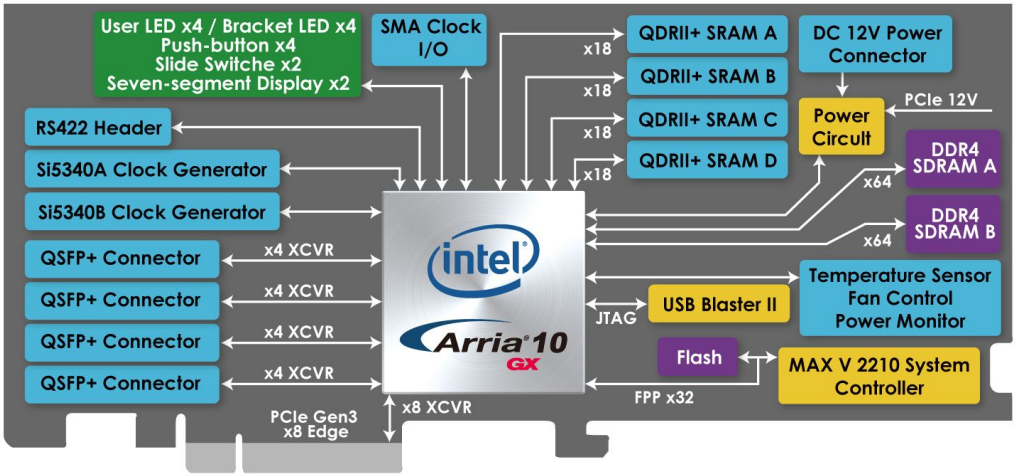
Início do projeto e dificuldades encontradas

- Instalação do ambiente (Fonte)
- Cada software utiliza um kernel e uma distro diferente do Linux (O driver da placa utiliza o CentOS com kernel 3.14 enquanto que o OpenCL utiliza o Ubuntu com kernel 4.14).
- **State of Art - Tutoriais muito novos e comunidade ainda em crescimento**
- **OpenVINO (1a versão em Agosto/2018)**
- Mercado dominado por dois grandes players: Intel e Xilinx

Enfim, o projeto

DE5a-Net-DDR4



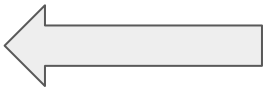


Workstation

Dell Precision T5600

Como desenvolver com FPGA

- RTL(VHDL) - Otimizado, vendor independent, porém demorado e longe de software
- MATLAB/Labview -> FPGA (nenhum controle sobre o código gerado)
- **OpenCL**
  - **OpenVINO**



- HLS (High Level Synthesis)

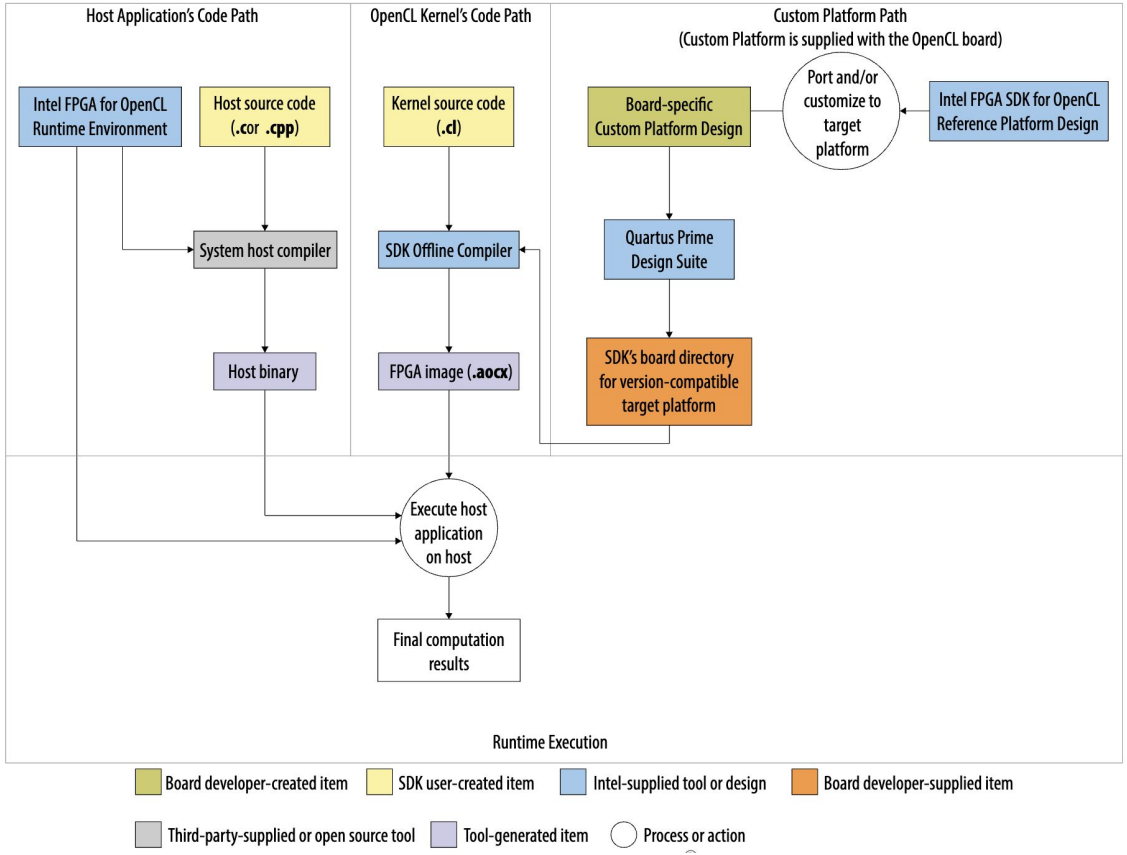
Estudo OpenCL
Multiplicação de matrizes

OpenCL

- Framework para programação paralela, que, através de sua linguagem, API, bibliotecas e sistemas runtime, possibilita desenvolver programas que sejam portáveis e ainda sejam eficientes.
- Vantagem de se programar em C (e não em HDL)
- A Intel disponibiliza um SDK que facilita a programação de OpenCL para FPGAs.

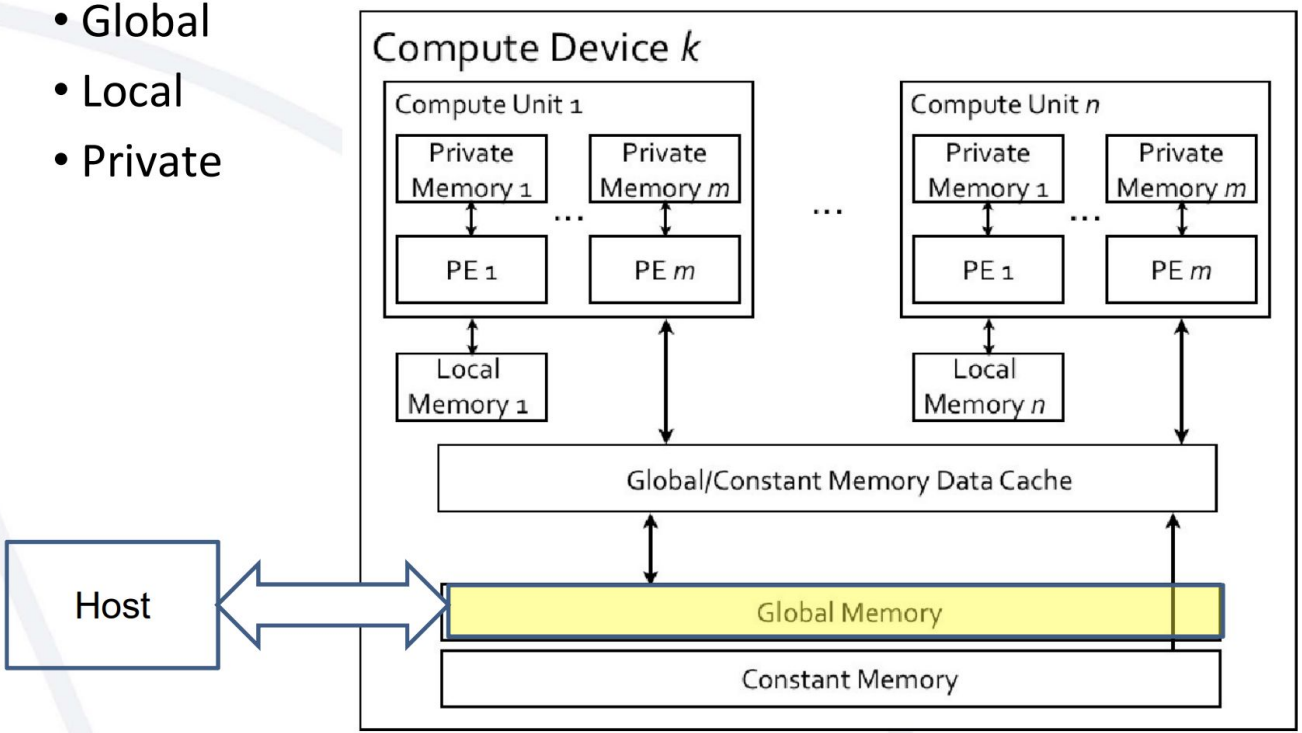
Conceitos básicos de OpenCL

Figure 1. Schematic Diagram of the Intel® FPGA SDK for OpenCL™ Programming Model

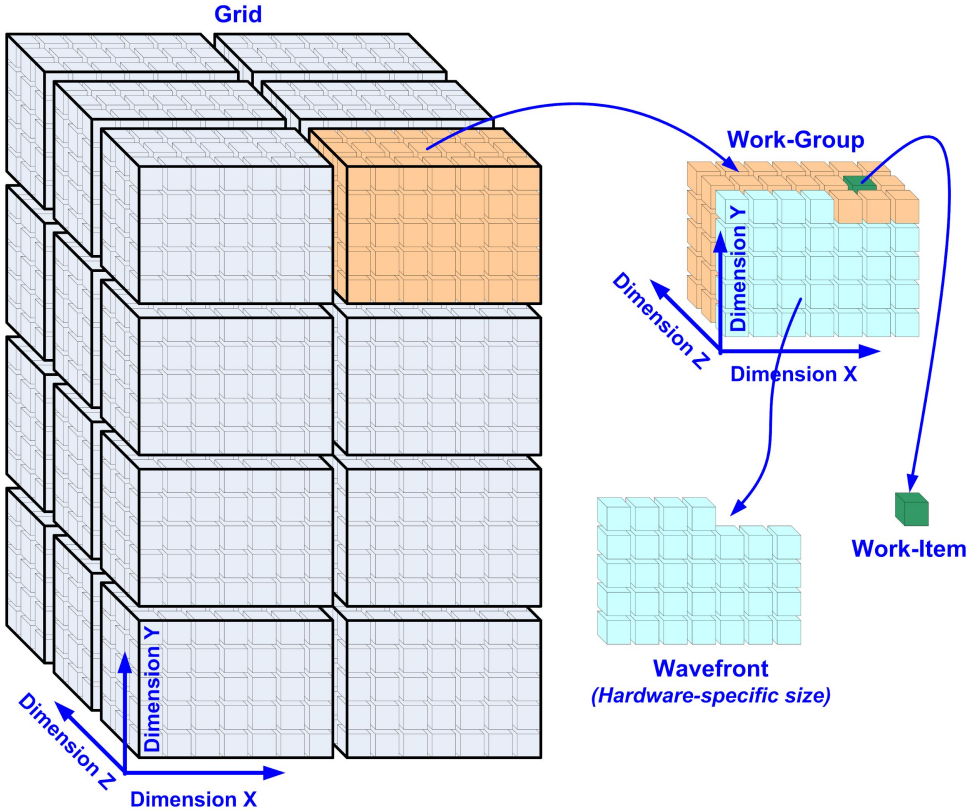


Tipos de memória

- Global
- Local
- Private



Work Groups e Work Items



Exemplo de código: multiplicação de matrizes

- A[n x k] x B[k x m] = C[n x m]
- no-intervaling e fp-relaxed (Divisão de memória e alteração da ordem de multiplicação)
- Cada work-item resolve uma posição da matriz C
- Faz a multiplicação na memória local e depois copia para memória global
- Utiliza operações SIMD
- Unrolling do loop principal que calcula cada posição da matriz C

CPU

Tempo de execução em um programa de C sequencial:

18292.205 ms

- Sem otimizações
- Sem contar o tempo de gerar os valores

hardware: i7 7820HQ - 21.03

FPGA

```
➜ bin ./host
Matrix sizes:
  A: 2048 x 1024
  B: 1024 x 1024
  C: 2048 x 1024
Initializing OpenCL
Platform: Intel(R) FPGA SDK for OpenCL(TM)
Using 1 device(s)
  de5a_net_ddr4 : Arria 10 Reference Platform (aclde5a_net_ddr40)
Using AOCX: matrix_mult.aocx
Reprogramming device [0] with handle 1
Generating input matrices
Launching for device 0 (global size: 1024, 2048)

Time: 30.557 ms
Kernel time (device 0): 30.466 ms

Throughput: 140.56 GFLOPS

Computing reference output
Verifying
Verification: PASS
```

Por que mudamos para a entrega atual

- Caminho das pedras para aceleração de código
- Tutoriais tem muitas falhas e cada passo essencial está espalhado em diversos tutoriais, o que torna a utilização mais difícil.
- Contribuição com a comunidade, do Insper principalmente, para que alguém que queira iniciar neste assunto tenha por onde começar.
- (jun/2019) email da fabricante tardio avisando que a placa ainda não era suportada pelo OpenVINO.

Git/Wiki

https://github.com/Insper/FPGA-Acceleration

Tutoriais:

- Acesso remoto
- Setup FPGA
- Setup OpenCL (FPGA)
- Demos OpenCL (FPGA)
- Intro OpenCL

FPGA-Acceleration

Estudos sobre aceleração de códigos com FPGA

Temas de interesse

- Execução de redes neurais
- Treinamento de redes neurais
- Virtualização/ Cloud
- Aceleração de algorítimos

Histórico

- 2018-2019 Iniciação Tecnológica Raphael Costa

Placas Insper - Arq. de Computadores

- DE5a-NET-DDR4
- Arria 10 SoC

Em processo de compra

- Intel® Programmable Acceleration Card with Intel Arria® 10 GX FPGA

Aplicações futuras com a entrega

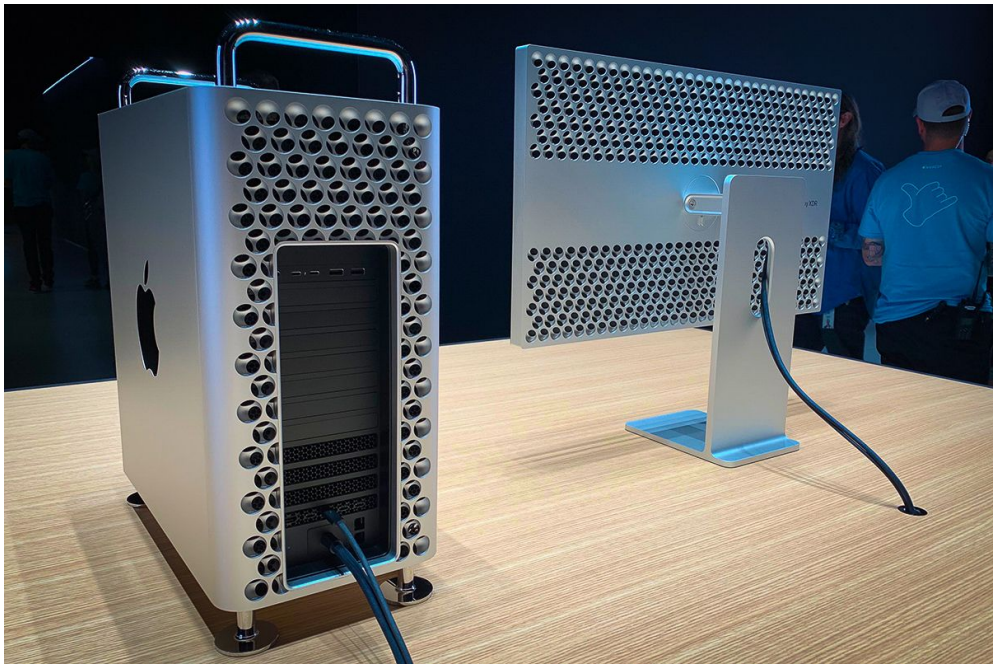
Artigo sobre comparação de desempenho com devices diferentes:

- CPUs
- GPUs
- FPGAs
- PAC (Programmable Acceleration Card)

Este tipo de artigo, com um programa fixado e diversos tamanhos, testando para vários tipos de devices e montando um gráfico de desempenho ainda não existe, somente para aplicações específicas, ou com testes somente FPGAxCPU.

O futuro das FPGAs

- Programar FPGAs por Software
- Virtualização (AWS e Azure)
- CPU + GPU + FPGA (Mac Pro 3)



Obrigado!