Arquitetura e Organização de Computadores II

Unidades de Processamento Gráfico
NVIDIA

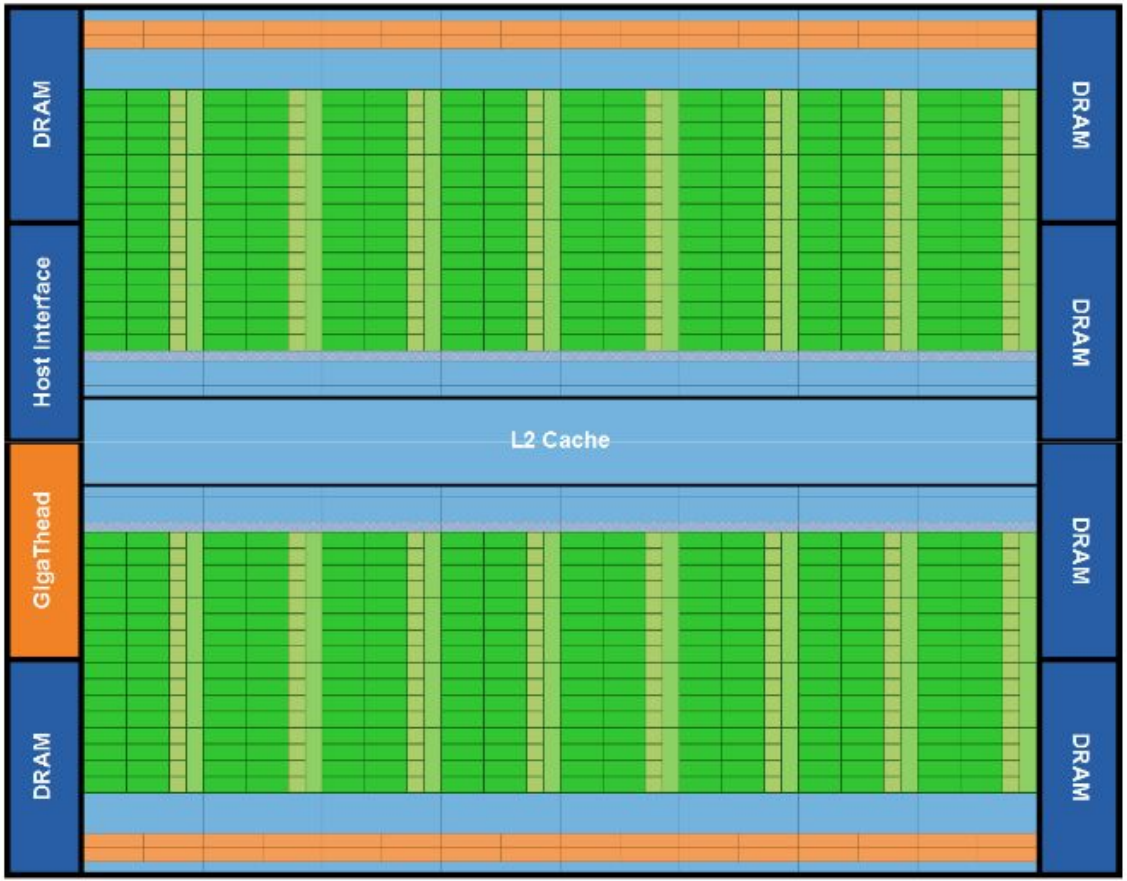
Prof. Nilton Luiz Queiroz Jr.

Arquitetura NVIDIA

- As arquiteturas mais recentes da NVIDIA são um pouco mais complexas que as arquiteturas de GPUs estudadas anteriormente;
  - Geralmente possuem mais de uma unidade de despacho de instrução
  - Possuem mais de um escalonador de warp;
  - Unidades de funções especiais;
    - Funções como seno, cosseno, raiz quadrada podem ser calculadas nessas unidades;
  - Entre outros detalhes
- Algumas dessas inovações se tornaram mais comuns a partir da arquitetura Fermi;

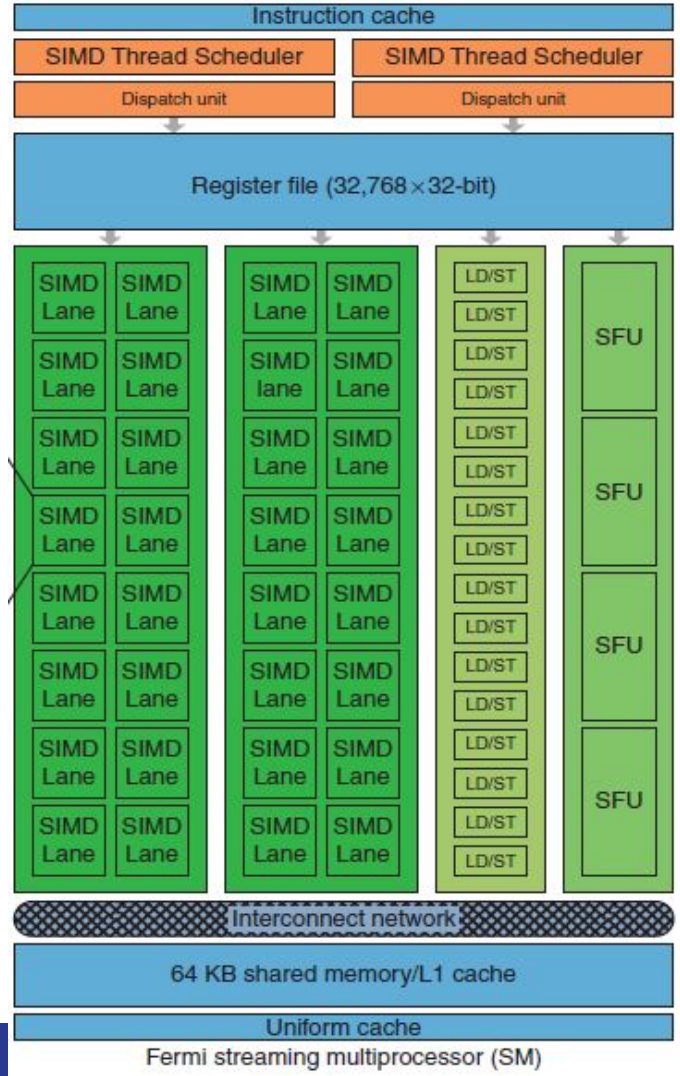
Arquitetura Fermi

- 16 processadores SIMD multithreaded (Streaming multiprocessor - SM) ;
  - SMs de terceira geração

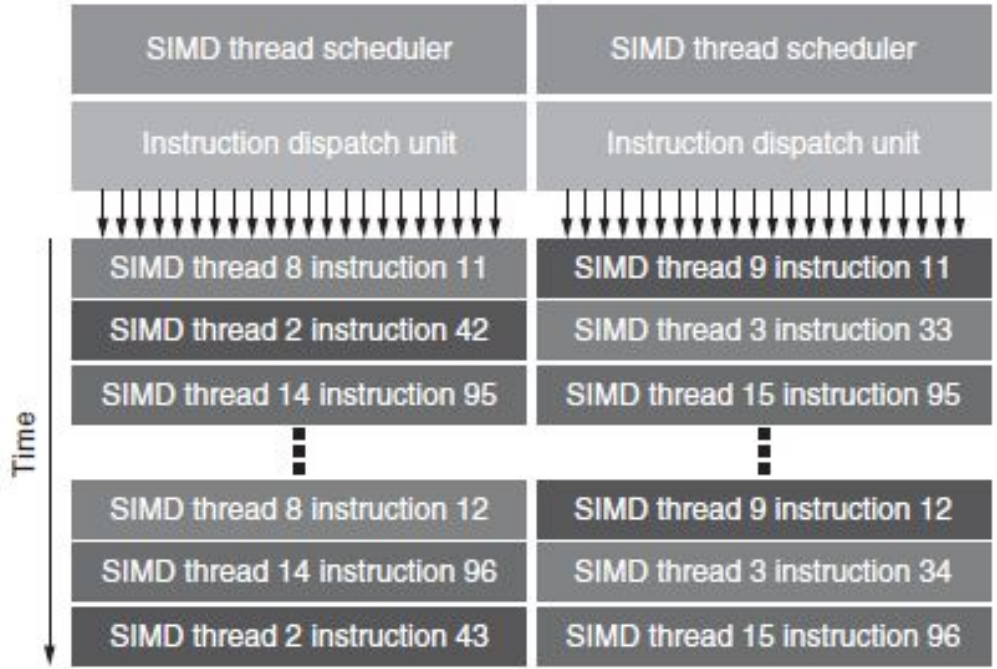


SM arquitetura Fermi

- 32 Lanes SIMD
  - Também chamados de CUDA core
- 4 unidades de funções especiais;
  - Cálculos de seno, cosseno, etc;
- Dois escalonadores de thread;
  - Cada escalonador emite uma instrução para grupos de 16 CUDA cores;
  - Duas SIMD threads (warps) podem ser executadas concorrentemente;

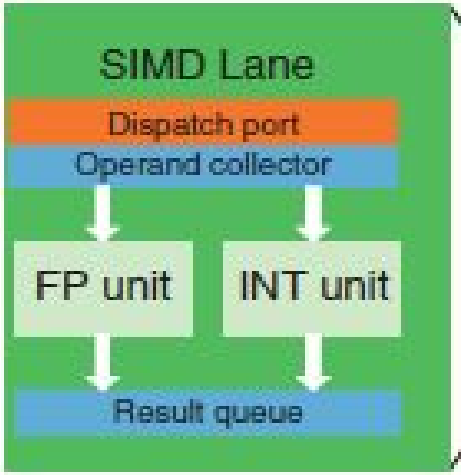


SM arquitetura Fermi



Pista SIMD

- As pistas SIMD possuem ULAs e Unidades de ponto flutuante pipelined;

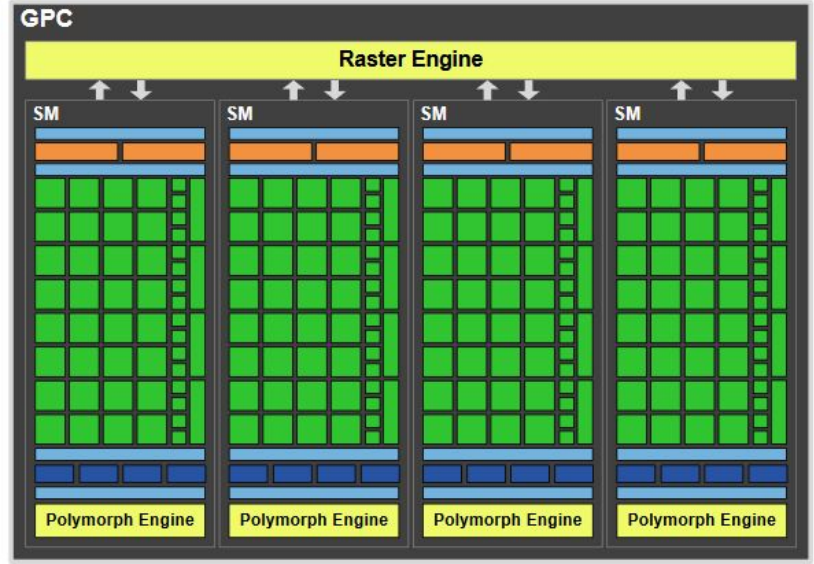


Arquitetura Fermi

- Introdução de memórias:
  - Cache L1;
    - Cada SM tem 64KB de memória para dividir entre a memória compartilhada e memória cache;
      - A divisão é feita em 48 e 16KB;
  - Cache L2;

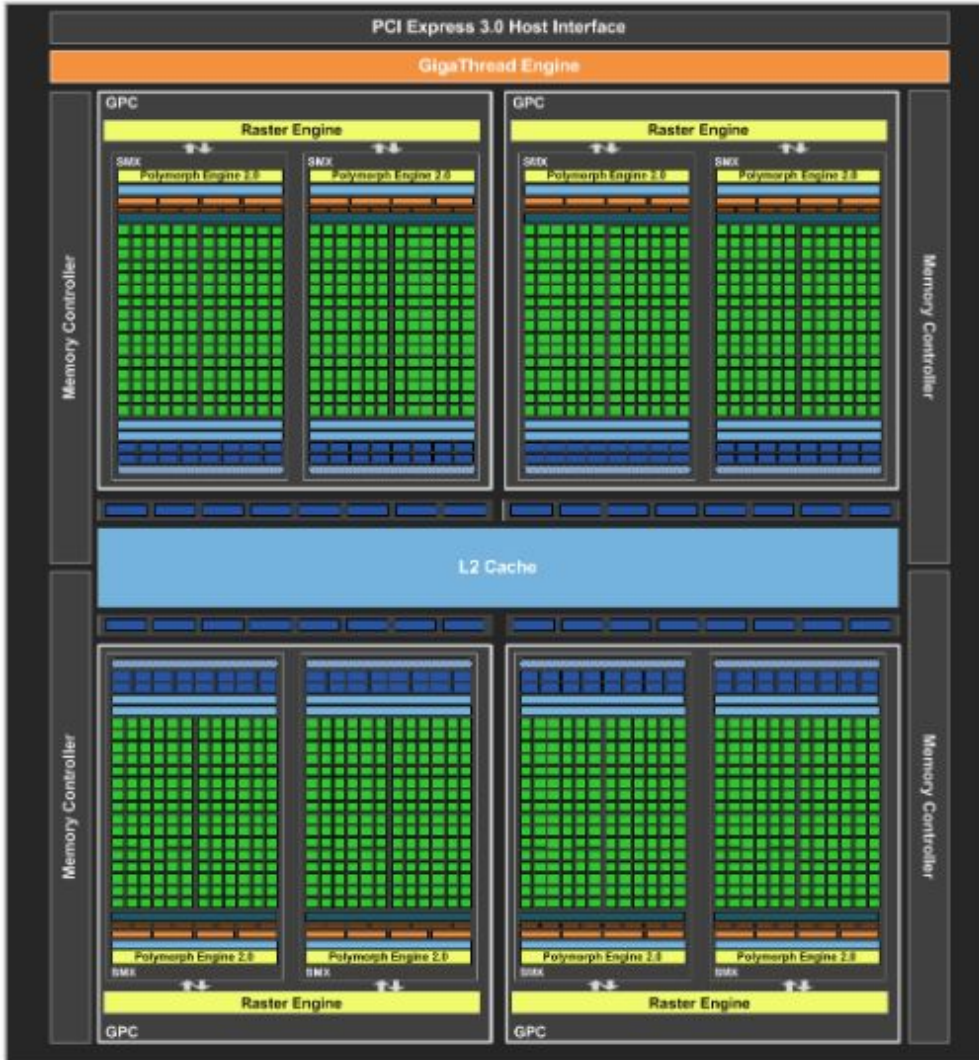
Arquitetura Fermi

- Introduz o GPC(Graphic processing cluster) ;
  - Focado em processamento gráfico;
  - Conjunto balanceado de unidades de processamento geométrico, de textura, de pixels, etc.
  - Composto por 4 SMs



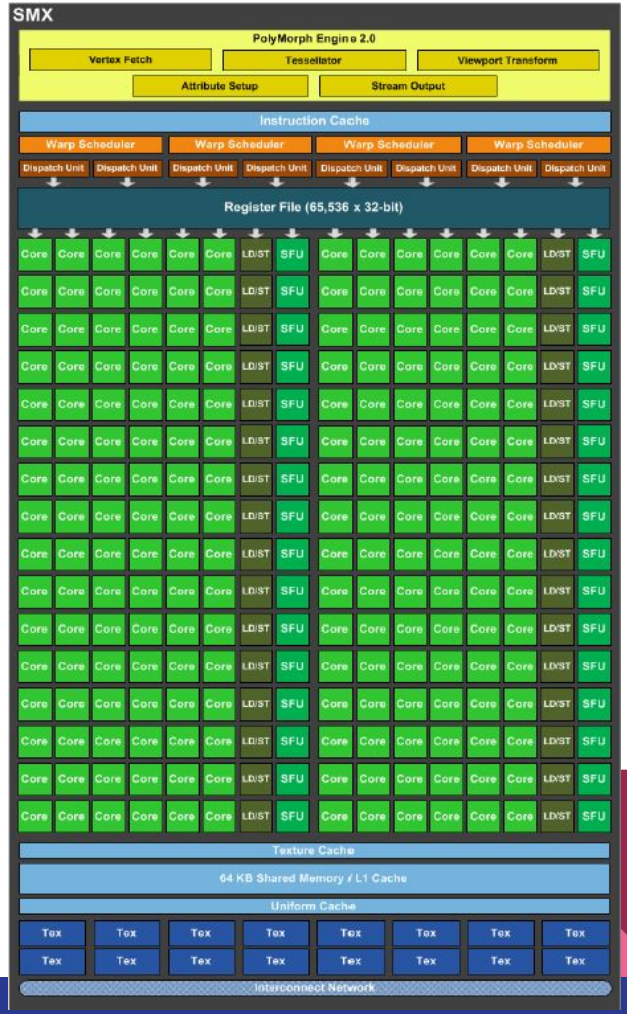
Kepler

- Uso de GPC (introduzidos na arquitetura Fermi GF100);
  - Graphic process cluster;
- Novo design de Streaming Multiprocessor, chamado de SMX



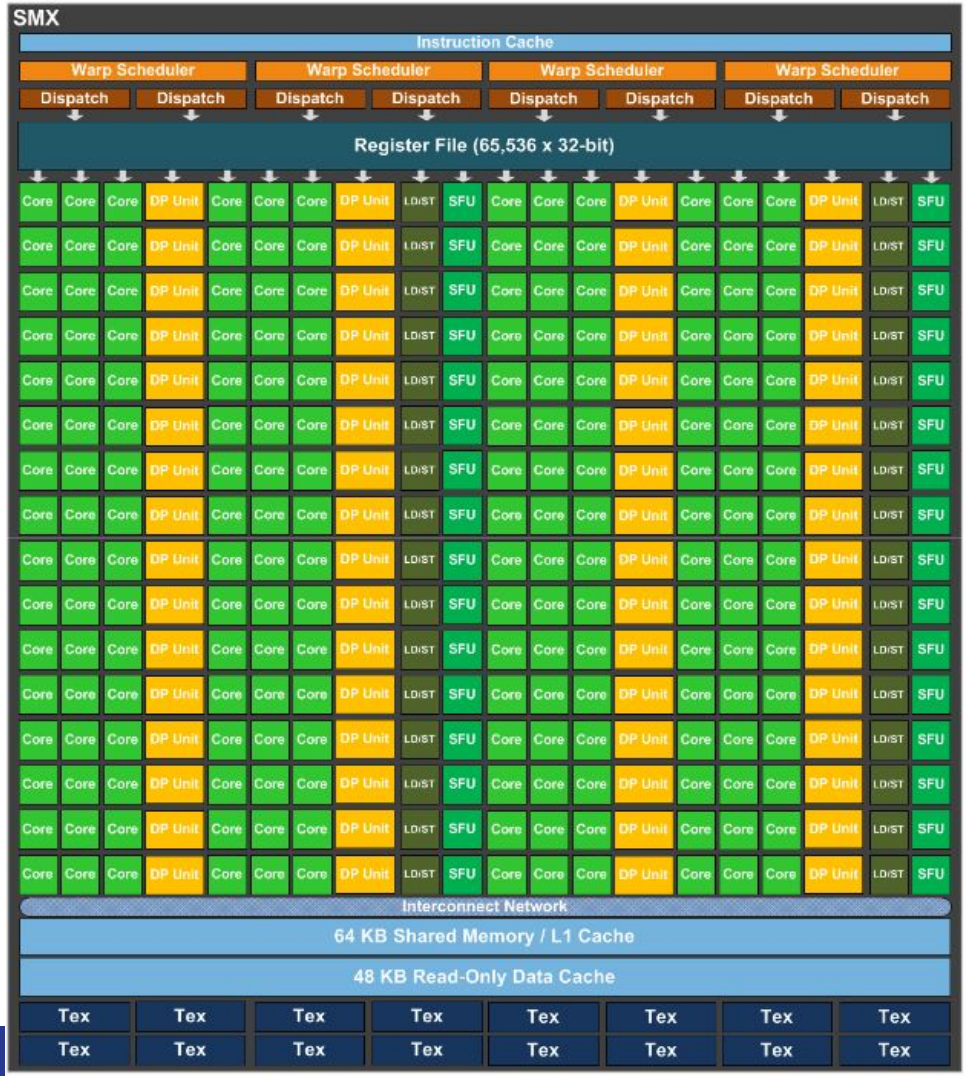
SMX da arquitetura Kepler

- 192 Lanes SIMD (CUDA cores);
- 32 unidades de load/store;
- 32 unidades de função especial;
- 4 escalonadores de Thread SIMD (escalonadores de warp);
    - Cada uma pode despachar 2 instruções por warp a cada clock;



Kepler

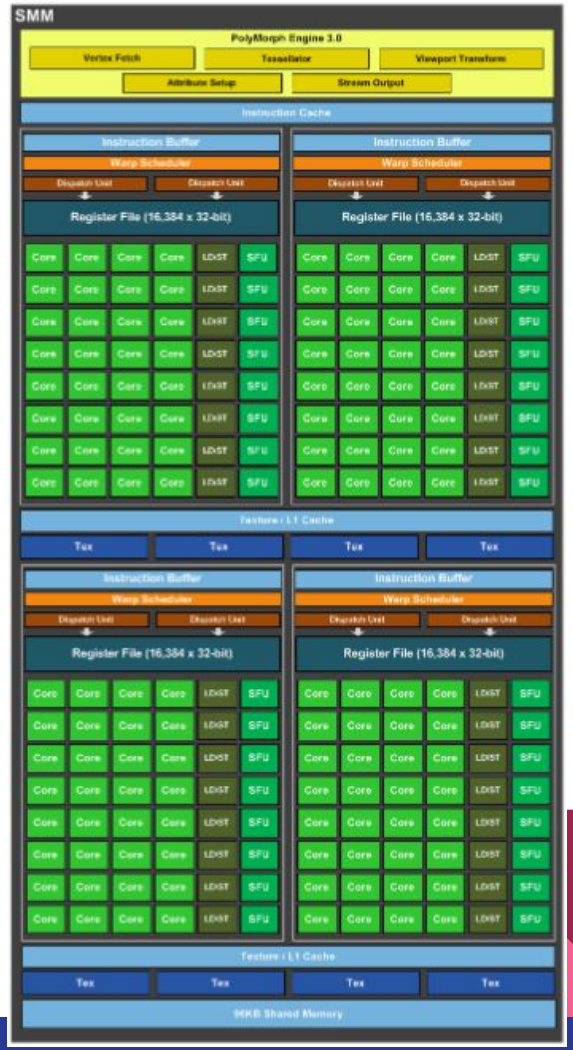
- Alguns SMX de arquiteturas kepler possuem unidades para cálculo de ponto flutuante de precisão dupla



Maxwell

- A arquitetura maxwell trouxe um novo design de Streaming Multiprocessor, chamado de SMM;
  - Particionado em 4 blocos de 32 pistas (CUDA cores) cada;
    - Cada bloco com seu próprio recurso para escalonar as warps e duas unidades de despacho de instruções;
  - Sem divisão entre cache L1 e memória compartilhada;
    - L1 é compartilhada com cache de textura;
- Desempenho de precisão dupla não é tão focado quanto na geração anterior;


PCI Express 3.0 Host Interface
GigaThread Engine
Memory Controller
GPC
Raster Engine
SMM
L2 Cache


SMM
PolyMorph Engine 3.0
Vertex Fetch
Tessellator
Viewport Transform
Attribute Setup
Stream Output
Instruction Cache
Instruction Buffer
Warp Scheduler
Dispatch Unit
Register File (16,384 x 32-bit)
Core
LD/ST
SFU
Texture / L1 Cache
Tex
64KB Shared Memory

Pascal

- Mais desempenho com double precision FP;
- Novo modelo de SM;
  - Particionado em 2 blocos de processamento;
  - Cada bloco com um escalonador de warp e duas unidades de despacho

Pascal


PCI Express 3.0 Host Interface
GigaThread Engine
GPC
TPC
SM
Instruction Cache
Memory Controller
High Bandwidth Memory 2
L2 Cache
High-Speed Hub
NVLink
NVLink
NVLink
NVLink


SM
Instruction Cache
Instruction Buffer
Warp Scheduler
Dispatch Unit
Dispatch Unit
Register File (32,768 x 32-bit)
Core
DP Unit
LD/ST
SFU
Texture / L1 Cache
Tex
64KB Shared Memory

Referências

NVIDIA. Whitepaper NVIDIA Tesla P100. Disponível em: <https://images.nvidia.com/content/pdf/tesla/whitepaper/pascal-architecture-whitepaper.pdf> acessado em 26/07/2017.

NVIDIA. Whitepaper NVIDIA GeForce GTX 980. Disponível em: <https://international.download.nvidia.com/geforce-com/international/pdfs/GeForce_GTX_980_Whitepaper_FINAL.PDF> acessdo em 26/07/2017.

NVIDIA. Whitepaper NVIDIA's next generation CUDA compute architecture: Kepler GK110. Disponível em: <https://www.nvidia.com/content/PDF/kepler/NVIDIA-Kepler-[illegible]0-Archite[illegible]-Whitepaper.pdf>. Acessado em 26/07/2017.

Referências

NVIDIA. Whitepaper NVIDIA GeForce GTX 680. Disponível em: <http://la.nvidia.com/content/PDF/product-specifications/GeForce_GTX_680_Whitepaper_FINAL.pdf> acessdo em 26/07/2017.

NVIDIA. Whitepaper NVIDIA's next generation CUDA compute architecture: Fermi. Disponível em: <https://www.nvidia.com/content/PDF/fermi_white_papers/NVIDIA_Fermi_Compute_Architecture_Whitepaper.pdf> acessado em 26/07/2017.

HENNESSY, John L.;PATTERSON, David A. Computer architecture: a quantitative approach. Elsevier, 2011.