FPGAs, GPUs e Xeon Phi

Pedro Bruel
phrb@ime.usp.br
Marcos Amarís
amaris@ime.usp.br

28 de Março de 2017

Instituto de Matemática e Estatística
Universidade de São Paulo

Sobre

Pedro Bruel
phrb@ime.usp.br

Alfredo Goldman
gold@ime.usp.br

Marcos Amarís
amaris@ime.usp.br

Roteiro

1. FPGAs
    - Computação Reconfigurável
    - Arquitetura de FPGAs: Lógica Programável
2. GPGPUs
    - História e Computação de Propósito Geral
    - GPGPUs Nvidia
    - GPGPUs AMD
3. Intel Xeon Phi

Slides

Os slides estão no GitHub:

- `github.com/phrb/aula-fpgas-gpus-xeonphi`

Computação Reconfigurável

Computação Reconfigurável

Field-Programmable Gate Arrays (FPGAs):

- Dispositivos feitos de semicondutores
- Funcionalidade definível após fabricação
- Reconfiguráveis mesmo após instalação
- Adaptáveis a diferentes aplicações
- Consumo eficiente de energia

Computação Reconfigurável

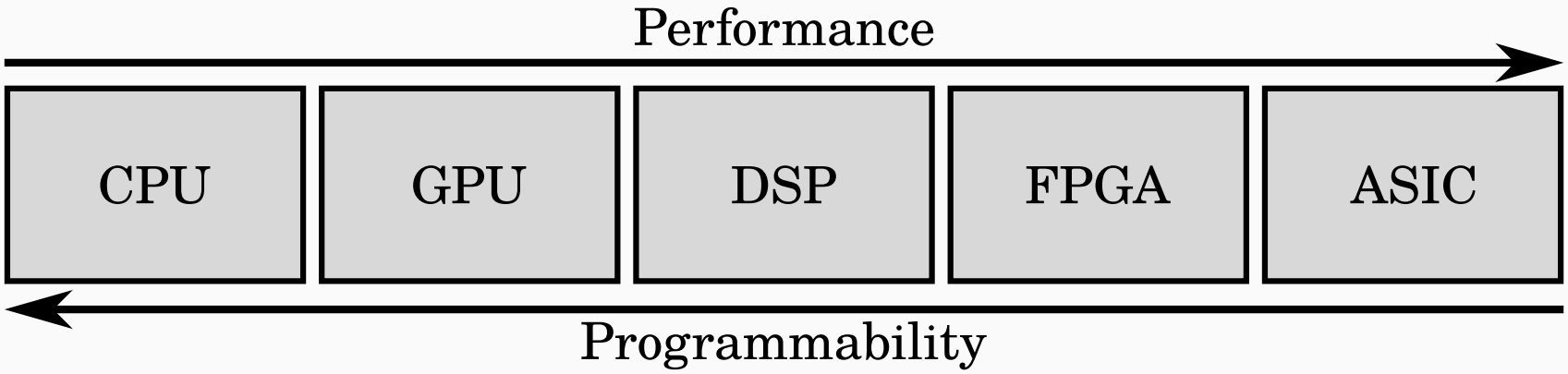
Field-Programmable Gate Arrays (FPGAs):

- Dispositivos feitos de semicondutores
- Funcionalidade definível após fabricação
- Reconfiguráveis mesmo após instalação
- Adaptáveis a diferentes aplicações
- Consumo eficiente de energia

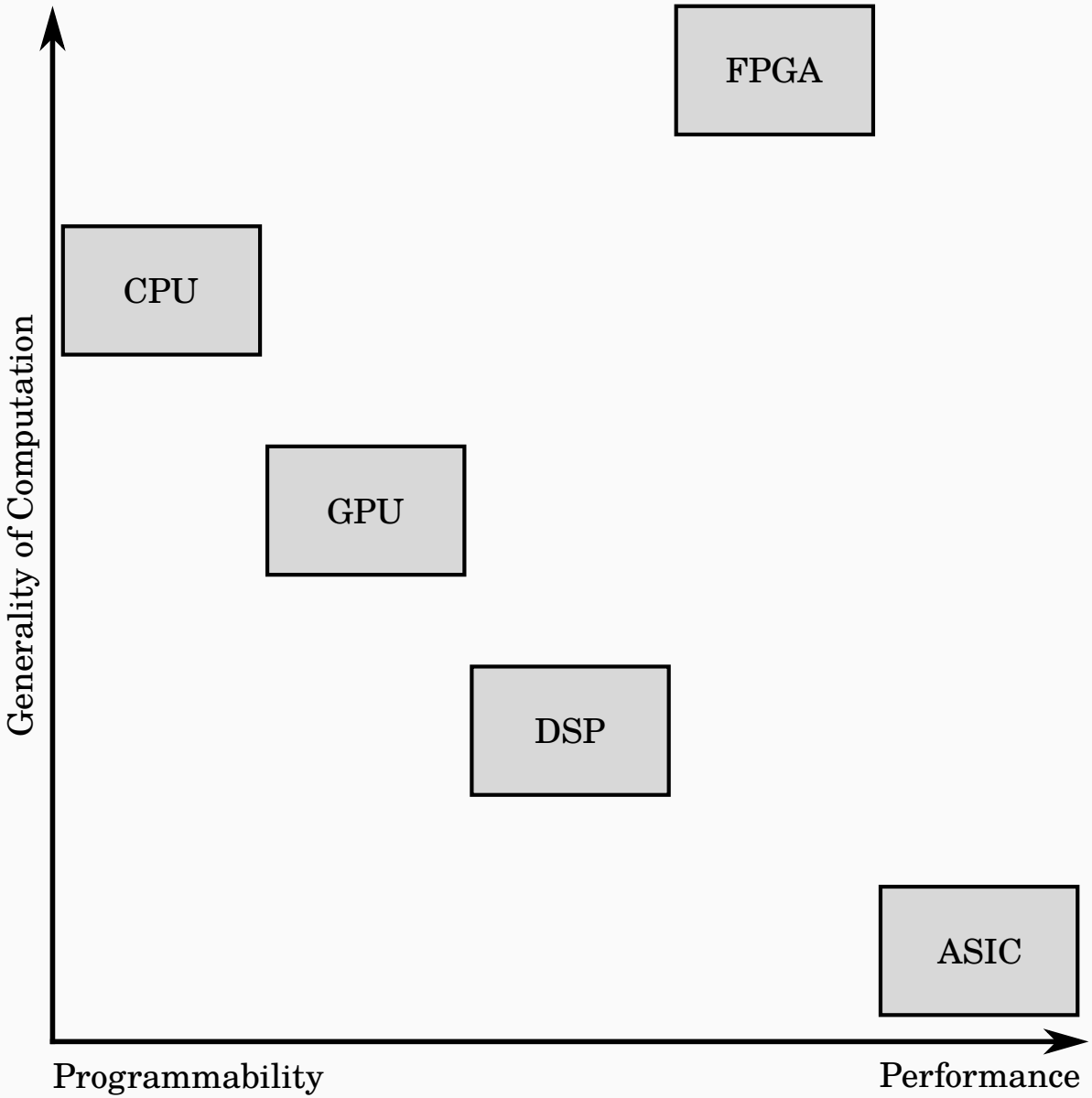
Compostas de uma matriz de elementos lógicos (Gate Array):

- Elementos Lógicos: Portas Lógicas, Transistores (LEs)
- Conexões entre LEs: Interconnect
- Tabelas de Verdade: *Lookup Tables* (LUTs)
- 2D Gate Arrays: LEs + LUTs + Interconnect

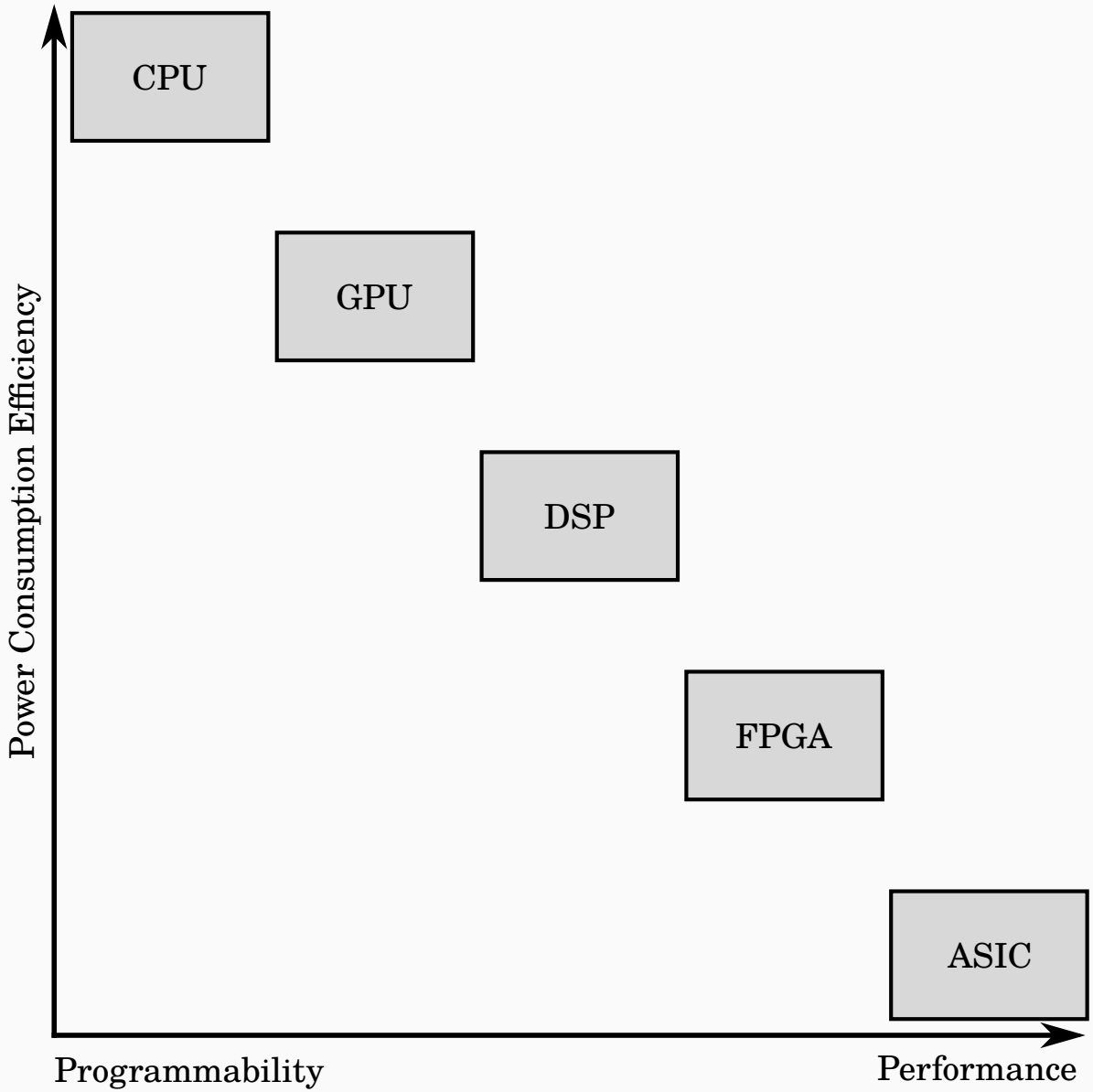
Programabilidade vs. Desempenho

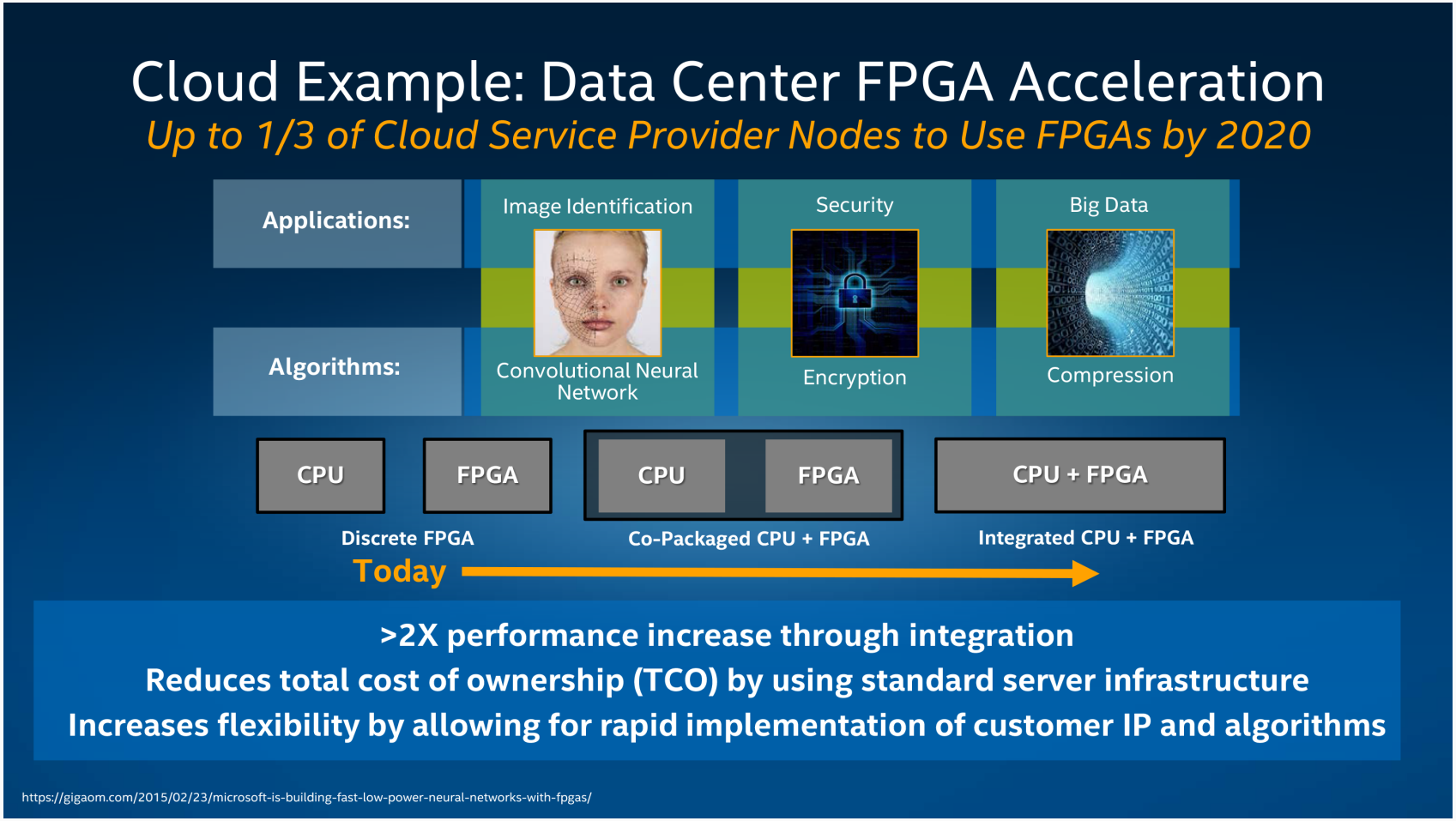


Programabilidade vs. Desempenho vs. Generalidade



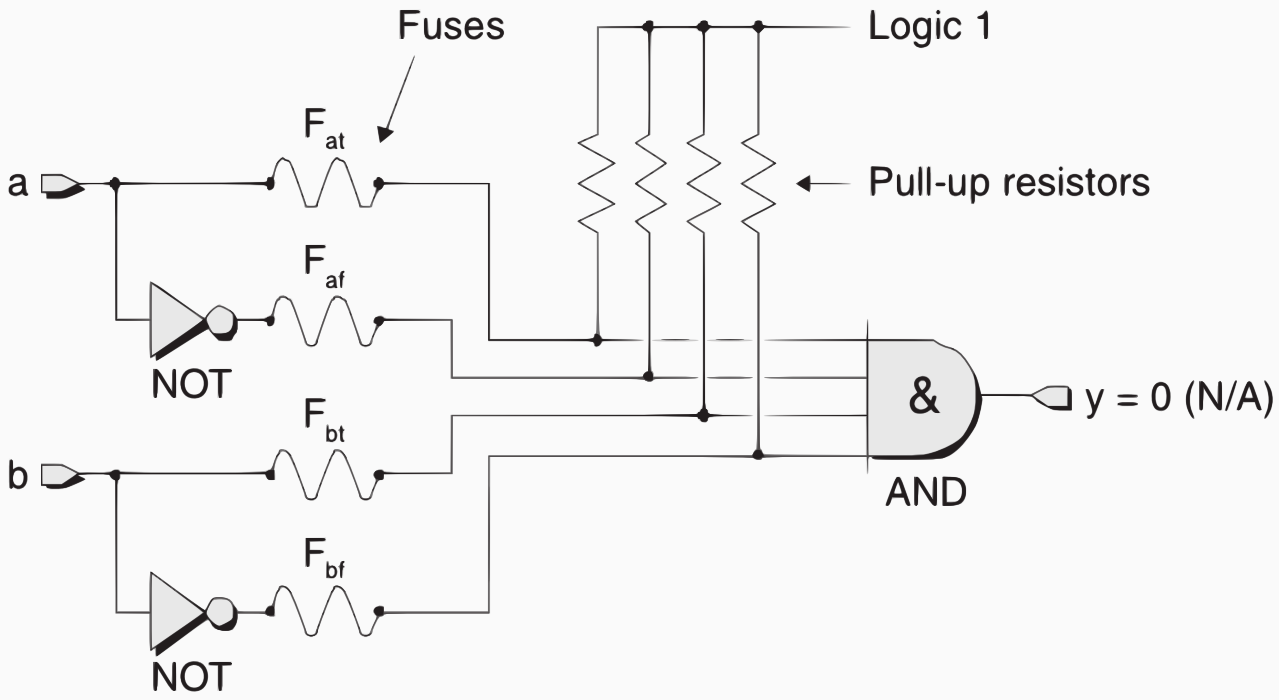
Programabilidade vs. Desempenho vs. Eficiência





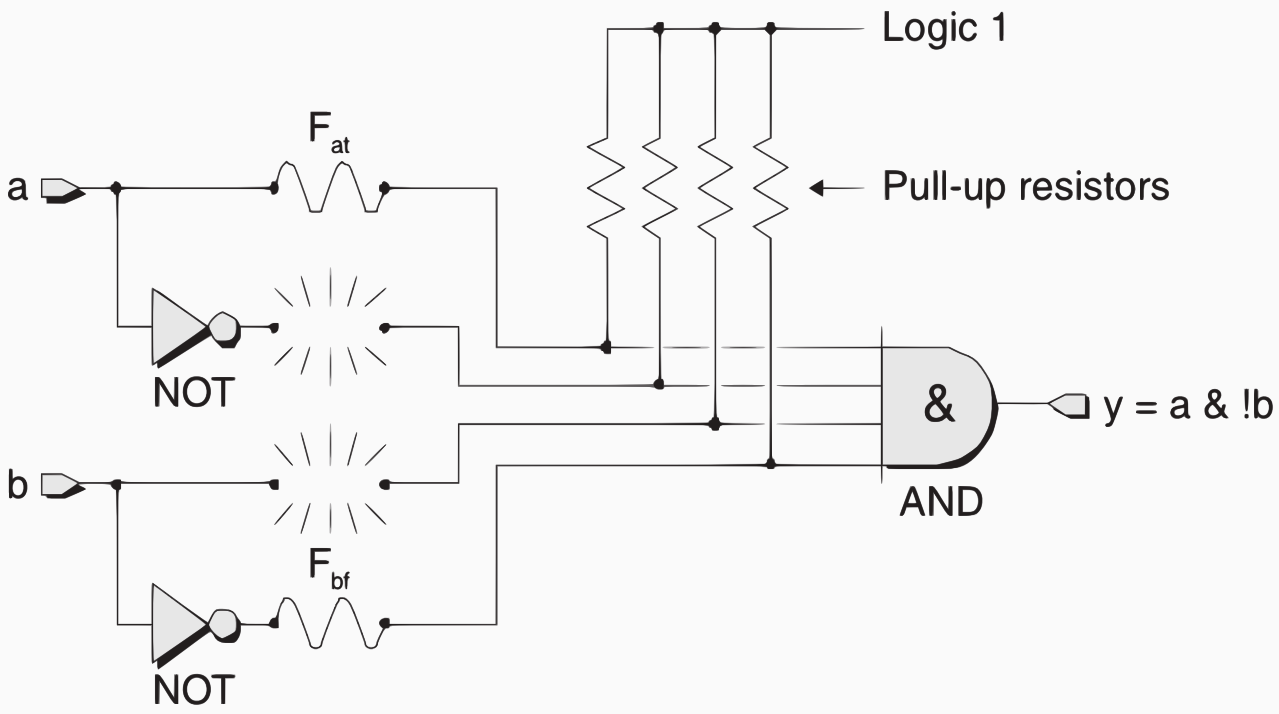
Fonte: anandtech.com/show/9321/intel-to-acquire-fpgaspecialist-altera-for-167-billion

Arquitetura de Memória: Lógica Programável



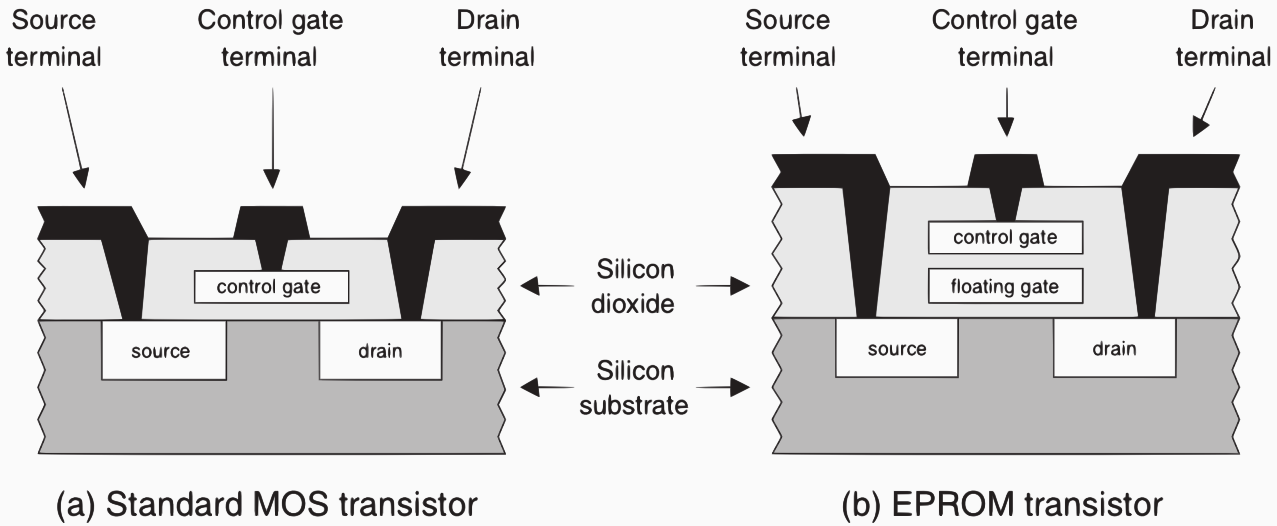
Fonte: Maxfield, Clive. The design warrior's guide to FPGAs: devices, tools and flows. Elsevier, 2004.

Arquitetura de Memória: Lógica Programável



Fonte: Maxfield, Clive. The design warrior's guide to FPGAs: devices, tools and flows. Elsevier, 2004.

Arquitetura de Memória: EPROM



(a) Standard MOS transistor (b) EPROM transistor

Fonte: Maxfield, Clive. The design warrior's guide to FPGAs: devices, tools and flows. Elsevier, 2004.

Arquitetura de Memória:

- Lógica Programável: Static Random Access Memory (SRAM)
- Memória: Synchronous Dynamic Random Access Memory (SDRAM)

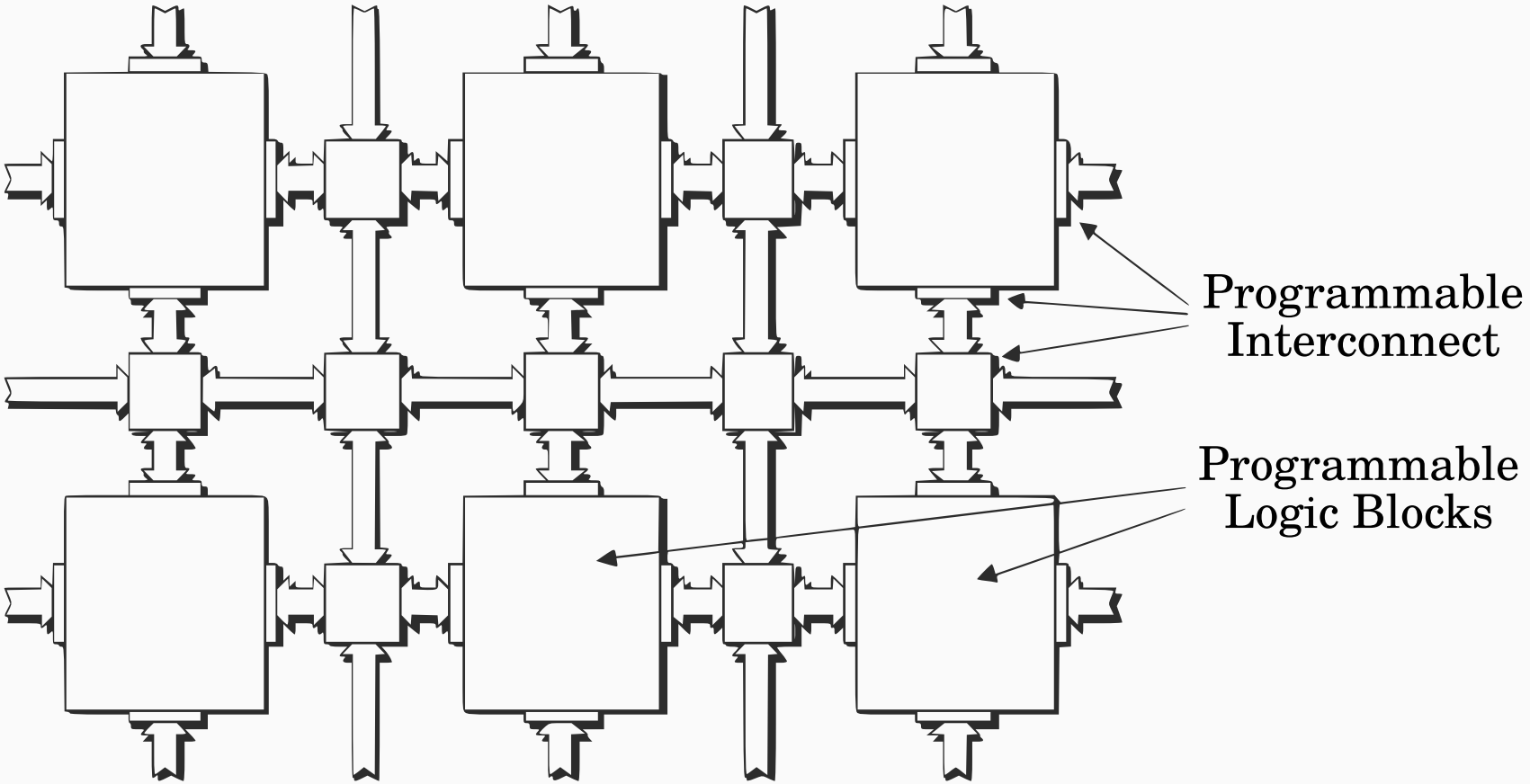
Arquitetura de Memória:

- Lógica Programável: Static Random Access Memory (SRAM)
- Memória: Synchronous Dynamic Random Access Memory (SDRAM)

2D Gate Arrays:

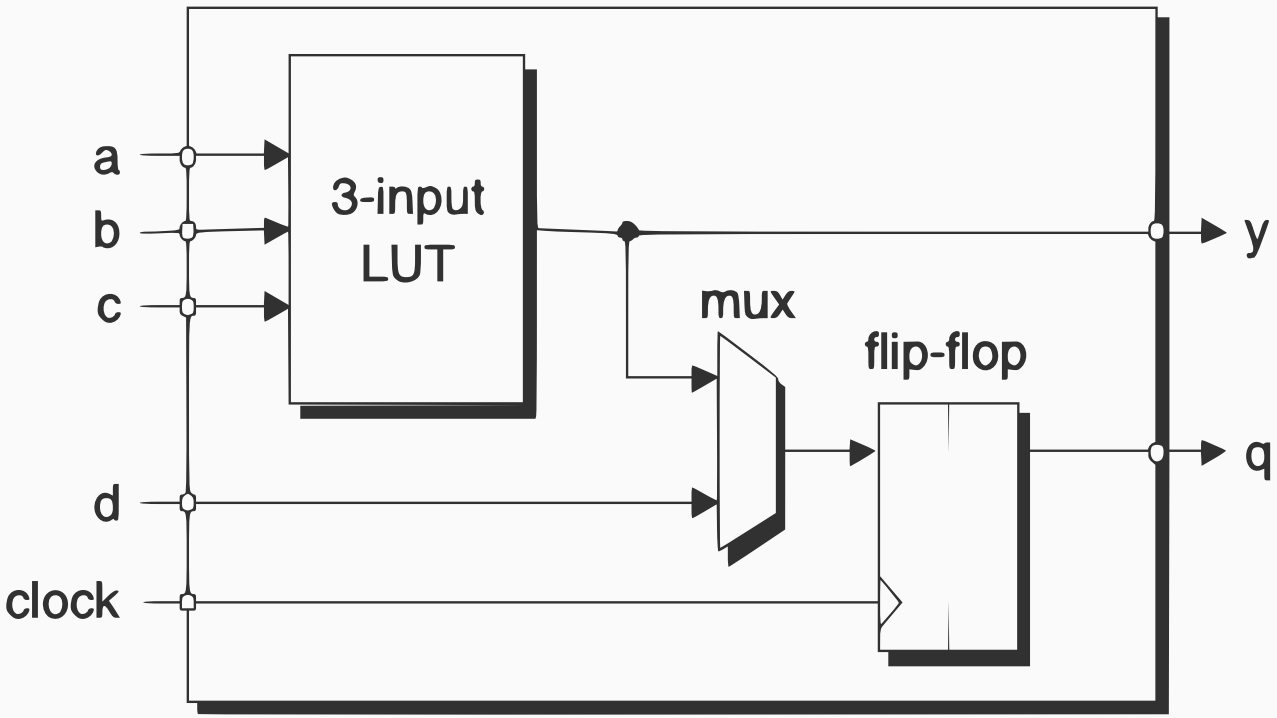
- Elementos Lógicos: Portas Lógicas, Transistores (LEs)
- Conexões entre LEs: Interconnect
- Tabelas de Verdade: *Lookup Tables* (LUTs)

FPGAs: Visão Simplificada de Alto Nível



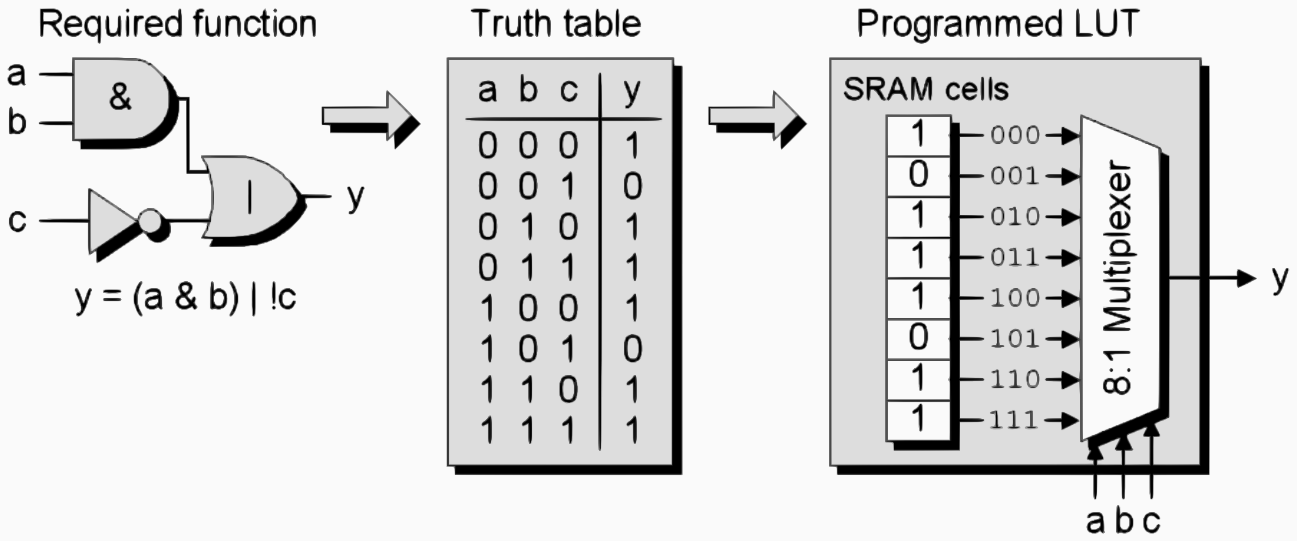
Fonte: Maxfield, Clive. The design warrior's guide to FPGAs: devices, tools and flows. Elsevier, 2004.

FPGAs: Visão Simplificada de Baixo Nível



Fonte: Maxfield, Clive. The design warrior's guide to FPGAs: devices, tools and flows. Elsevier, 2004.

FPGAs: Visão Simplificada de Baixo Nível



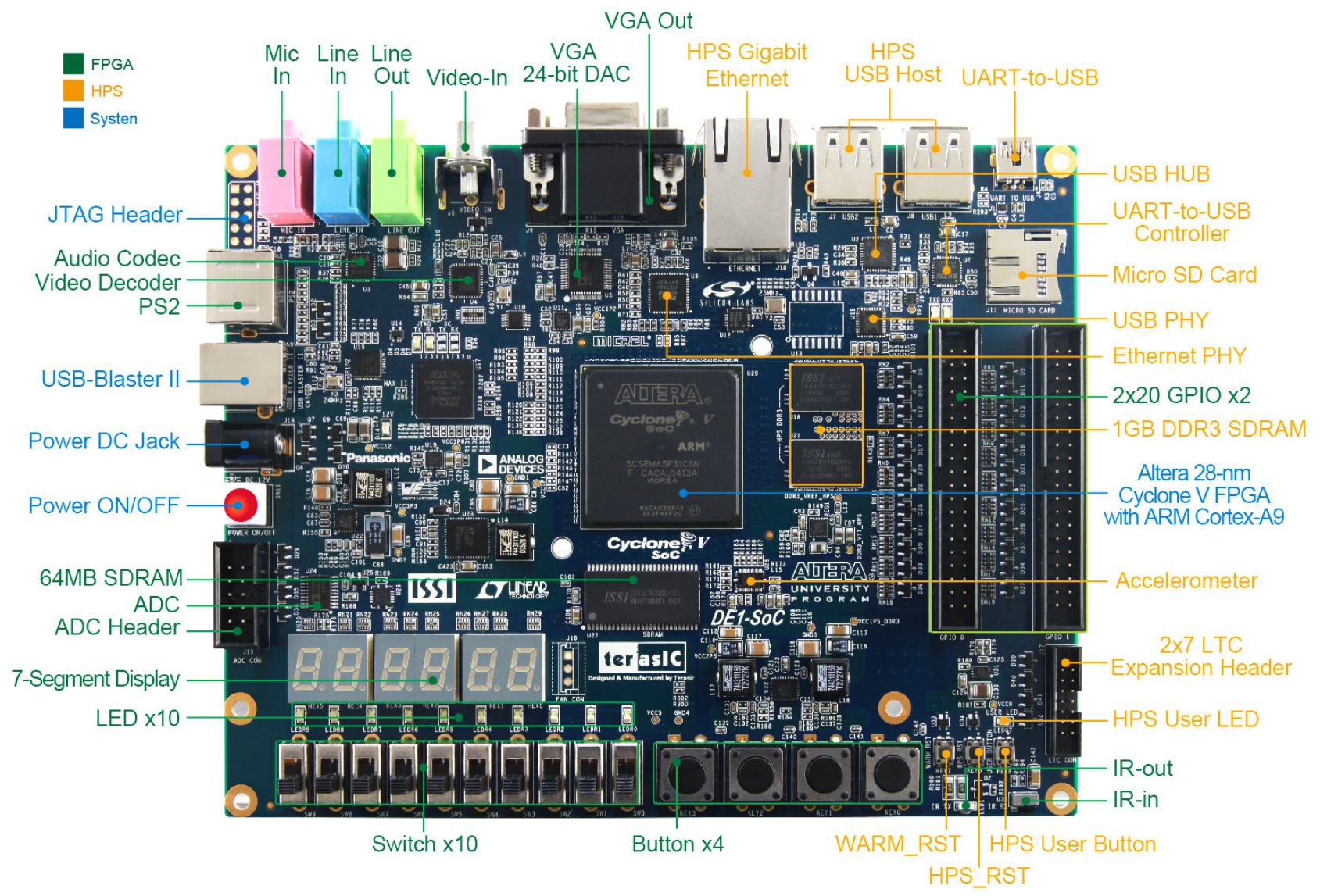
Fonte: Maxfield, Clive. The design warrior's guide to FPGAs: devices, tools and flows. Elsevier, 2004.

FPGAs Hoje

Arquitetura de FPGAs contemporânea:

- Associada a uma CPU
- Static Random Access Memory (SRAM)
- Synchronous Dynamic Random Access Memory (SDRAM)
- I/O
- Elementos Lógicos (Logic Elements, LEs)
- Interconexões

FPGAs SoC: Cyclone V



Programando FPGAs

Hardware Description Languages (HDL):

- Definição de *clock*
- Definição de circuitos e operações simples
- Hoje, muito já vem pré-definido

Programando FPGAs

Hardware Description Languages (HDL):

- Definição de *clock*
- Definição de circuitos e operações simples
- Hoje, muito já vem pré-definido

High-Level Synthesis (HLS):

- Gerar HDL a partir de código em C
- OpenCL

FPGAs: Saiba Mais

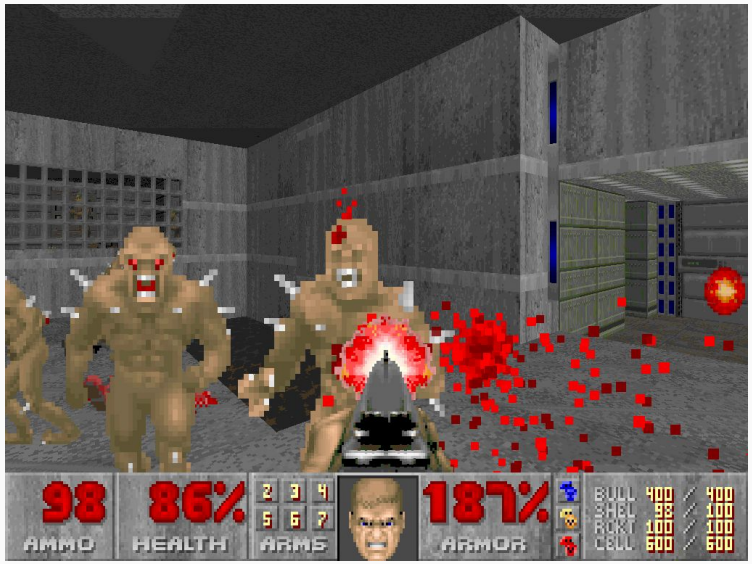
Livros:

- Andrew Moore and Ron Wilson. FPGAs for Dummies. Intel/ Wiley, 2017
- Maxfield, Clive. The design warrior's guide to FPGAs: devices, tools and flows. Elsevier, 2004
- Vanderbauwhede, Wim, and Khaled Benkrid, eds. High-performance computing using FPGAs. New York: Springer, 2013 (Avançado)

GPUs: História e Computação de Propósito Geral

Placas de Vídeo

- 80's: Primeiro controlador de vídeo
- Evolução dos jogos 3D
- Aplicação de texturas, iluminação, sombras, . . .
- Animação, simulação computacional, design gráfico, física da luz, . . .



Unidades de Processamento Gráfico (GPUs)

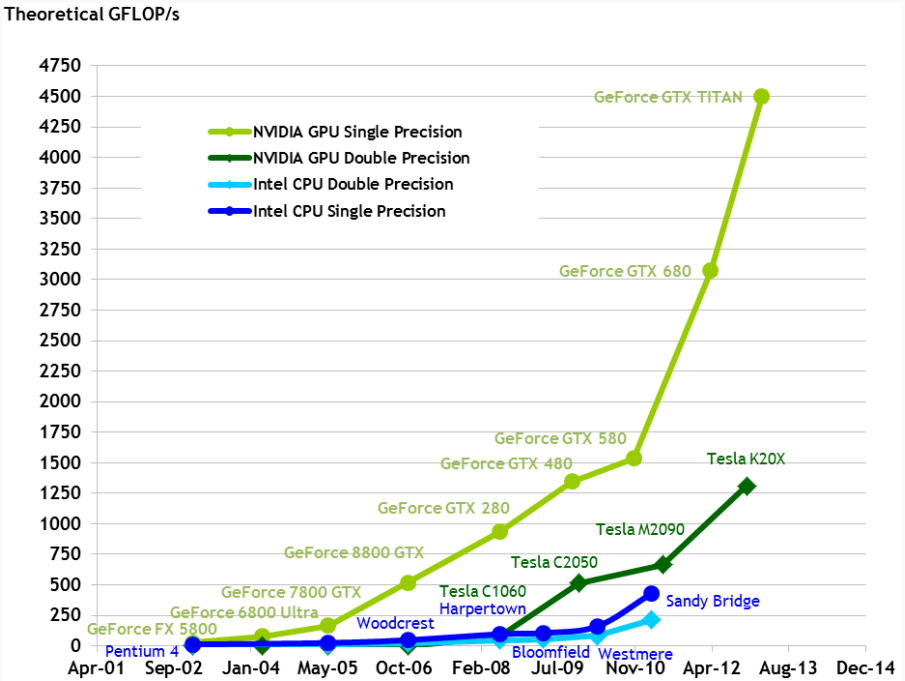
- O termo GPU foi popularizado pela Nvidia em 1999, quando criou a GeForce 256
- Em 2002 foi lançada a primeira GPU de Propósito Geral (GPGPU)
- Os principais fabricantes de GPUs são a Nvidia e a AMD
- Em 2005 a Nvidia lançou CUDA
- Carros autônomos, Realidade Virtual, Inteligencia Artificial, . . .

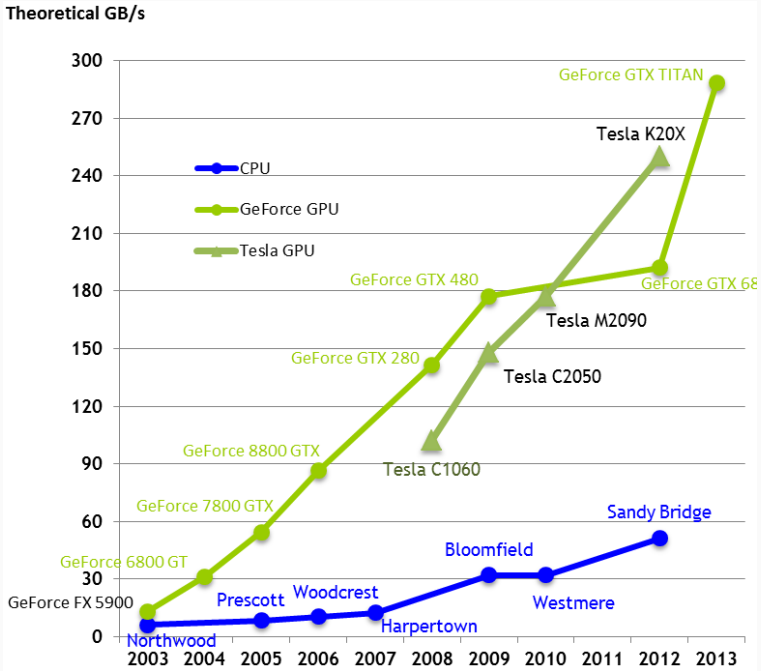
GeForce 256 Series

Release date	October 11, 1999
Codename	NV10
Cards	
Mid-range	GeForce 256 SDR
High-end	GeForce 256 DDR
Rendering support	
Direct3D	Direct3D 7.0
OpenGL	OpenGL 1.3 (T&L)
History	
Predecessor	Pre-GeForce
Successor	GeForce 2 Series

GPU vs. CPU

Hoje em dia, GPUs são capazes de fazer em paralelo mais operações computacionais que CPUs multi-core

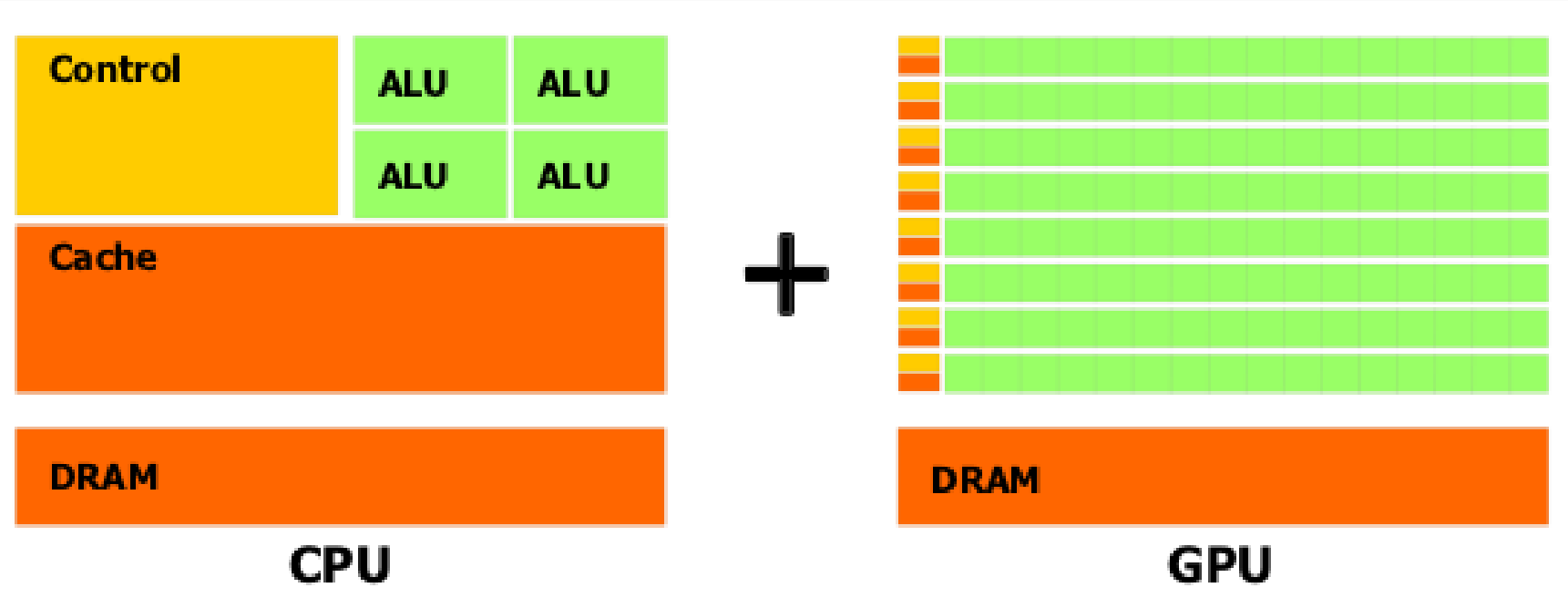




GPU de Propósito Geral (GPGPU)

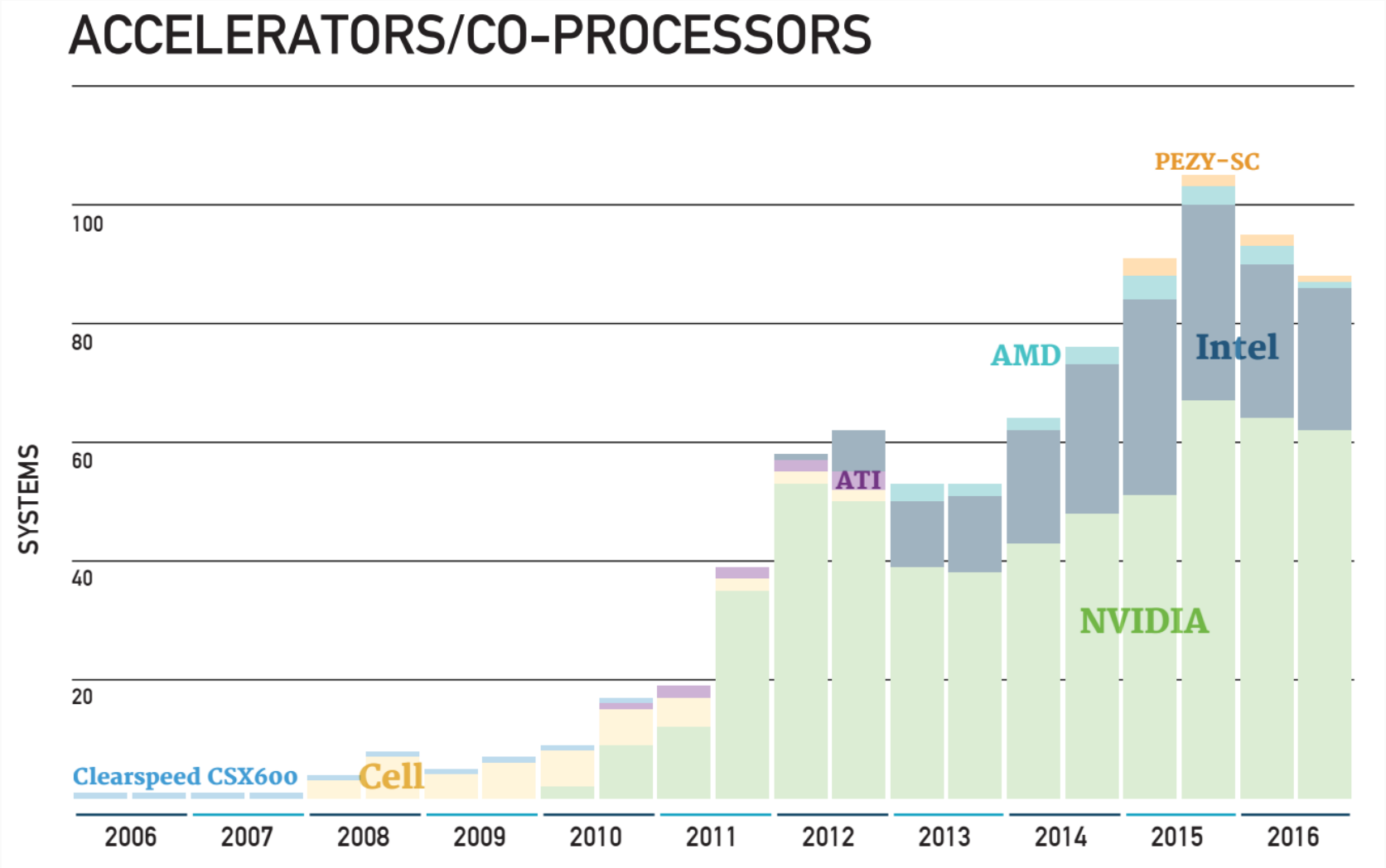
O programa principal é executado na CPU (host), que é responsável por iniciar a execução na GPU (device).

As GPUs têm sua própria herarquia de memória e os dados devem ser transferidos através do barramento PCI Express.



Supercomputadores com Aceleradores de Hardware

Supercomputadores na Top 500 com aceleradores e co-processadores:



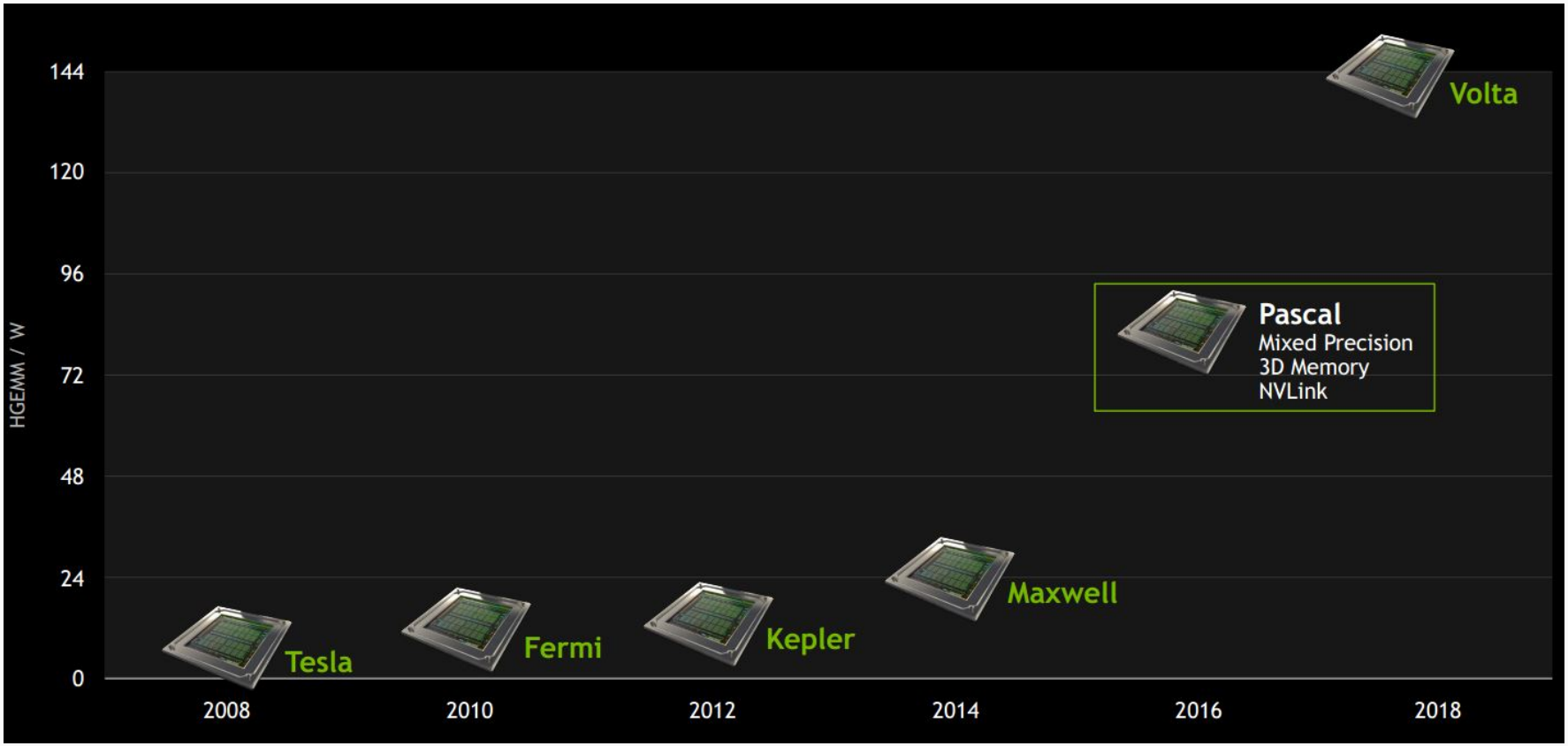
Green 500: Lista de supercomputadores mais eficientes em consumo de energia

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	4,503.17	GSIC Center, Tokyo Institute of Technology	TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x	27.78
2	3,631.86	Cambridge University	Wilkes - Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20	52.62
3	3,517.84	Center for Computational Sciences, University of Tsukuba	HA-PACS TCA - Cray 3623G4-SM Cluster, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband QDR, NVIDIA K20x	78.77
4	3,185.91	Swiss National Supercomputing Centre (CSCS)	Piz Daint - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Level 3 measurement data available	1,753.66
5	3,130.95	ROMEO HPC Center - Champagne-Ardenne	romeo - Bull R421-E3 Cluster, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR, NVIDIA K20x	81.41
6	3,068.71	GSIC Center, Tokyo Institute of Technology	TSUBAME 2.5 - Cluster Platform SL390s G7, Xeon X5670 6C 2.930GHz, Infiniband QDR, NVIDIA K20x	922.54
7	2,702.16	University of Arizona	iDataPlex DX360M4, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR14, NVIDIA K20x	53.62
8	2,629.10	Max-Planck-Gesellschaft MPI/IPP	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	269.94
9	2,629.10	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	55.62
10	2,358.69	CSIRO	CSIRO GPU Cluster - Nitro G16 3GPU, Xeon E5-2650 8C 2.000GHz, Infiniband FDR, Nvidia K20m	71.01

GPGPUs Nvidia

Roadmap para Arquiteturas de GPUs NVIDIA

A próxima arquitetura será a Volta, com 12-10 nm FinFET.



Espaços de Memória em GPGPUs

O acesso à memória global é custoso, a latência é 100*x* mais lenta que na memória compartilhada.

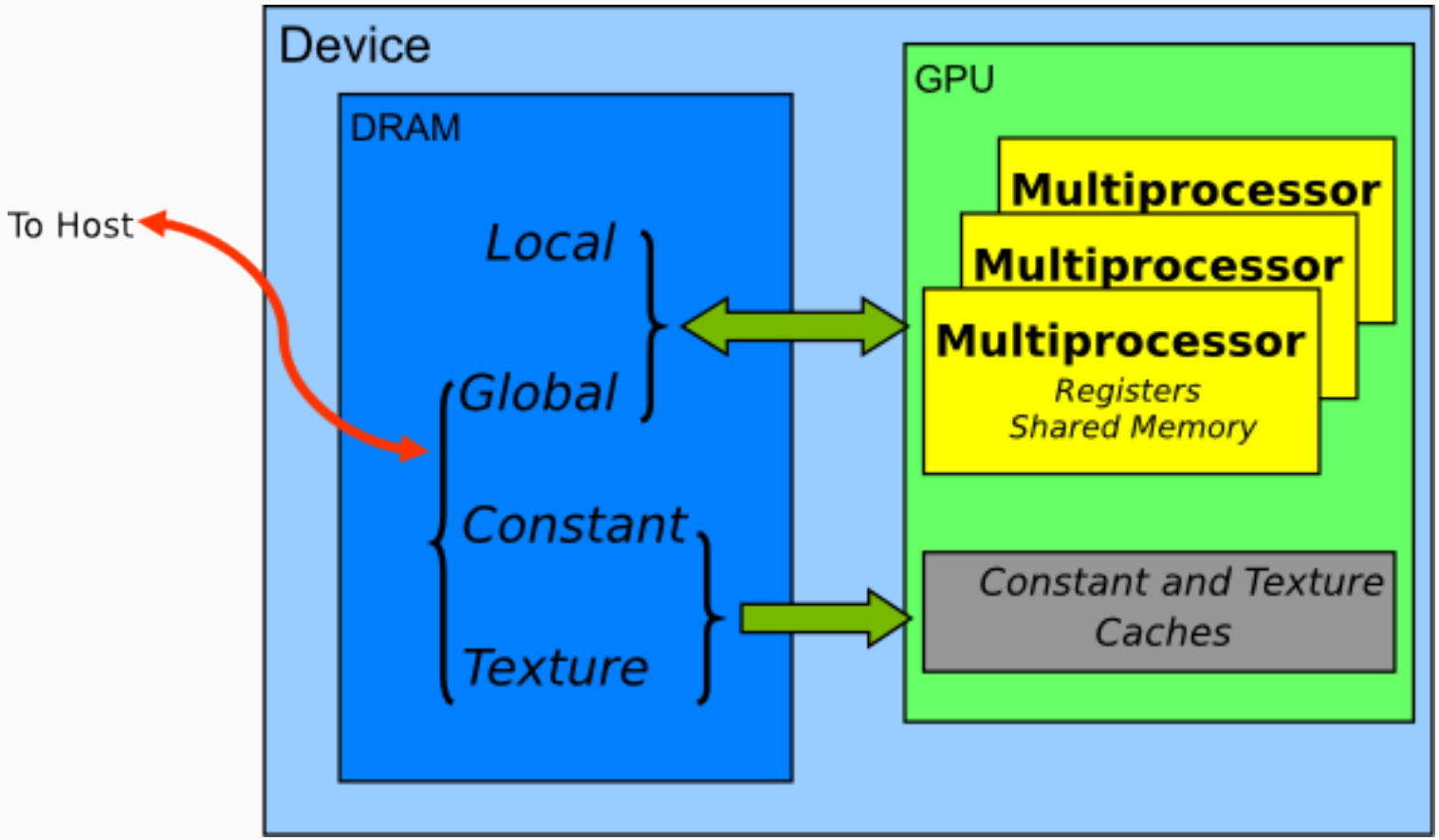


Figura 1: Espaços de memória em dispositivos CUDA

Compute Capability de GPUs NVIDIA

Compute Capability é uma diferenciação entre arquiteturas e modelos de GPUs NVIDIA.

Architecture	Tesla G80	Tesla GT200	Fermi GF100	Fermi GF104	Kepler GK104	Kepler GK110
Time frame	2006-07	2008-09	2010	2011	2012	2013
CUDA Compute Capability (CCC)	1.0	1.2	2.0	2.1	3.0	3.5
N (multiprocs.)	16	30	16	7	8	14
M (cores/multip.)	8	8	32	48	192	192
Number of cores	128	240	512	336	1536	2688

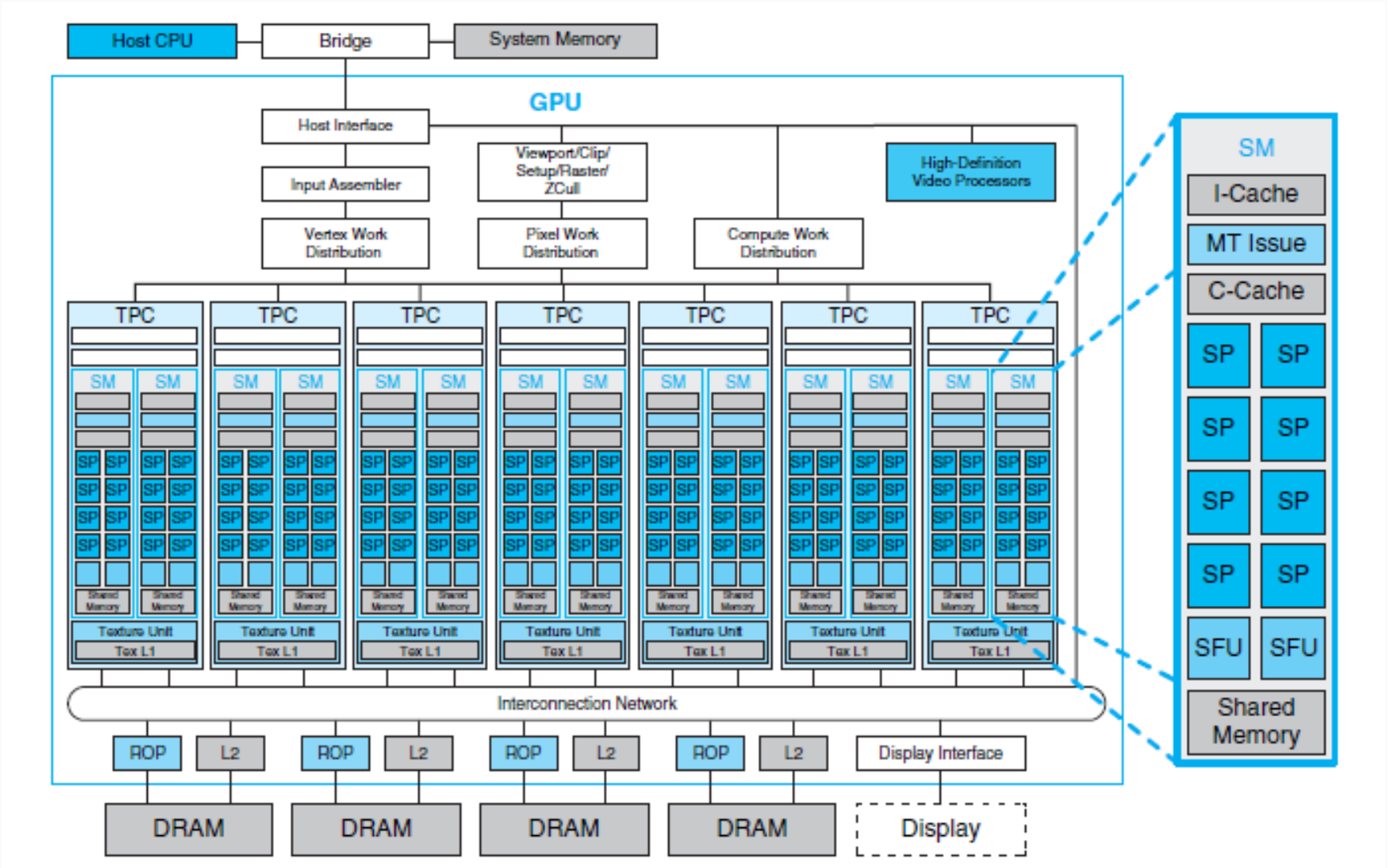
GPUs Tesla da NVIDIA

- Multiprocessadores
- FP 32
- FP 64
- Interface mem.
- TDP
- Fabricação

Tesla Products	Tesla K40	Tesla M40	Tesla P100
GPU	GK110 (Kepler)	GM200 (Maxwell)	GP100 (Pascal)
SMs	15	24	56
TPCs	15	24	28
FP32 CUDA Cores / SM	192	128	64
FP32 CUDA Cores / GPU	2880	3072	3584
FP64 CUDA Cores / SM	64	4	32
FP64 CUDA Cores / GPU	960	96	1792
Base Clock	745 MHz	948 MHz	1328 MHz
GPU Boost Clock	810/875 MHz	1114 MHz	1480 MHz
Peak FP32 GFLOPs[1]	5040	6840	10600
Peak FP64 GFLOPs[1]	1680	210	5300
Texture Units	240	192	224
Memory Interface	384-bit GDDR5	384-bit GDDR5	4096-bit HBM2
Memory Size	Up to 12 GB	Up to 24 GB	16 GB
L2 Cache Size	1536 KB	3072 KB	4096 KB
Register File Size / SM	256 KB	256 KB	256 KB
Register File Size / GPU	3840 KB	6144 KB	14336 KB
TDP	235 Watts	250 Watts	300 Watts
Transistors	7.1 billion	8 billion	15.3 billion
GPU Die Size	551 mm^2	601 mm^2	610 mm^2
Manufacturing Process	28-nm	28-nm	16-nm FinFET

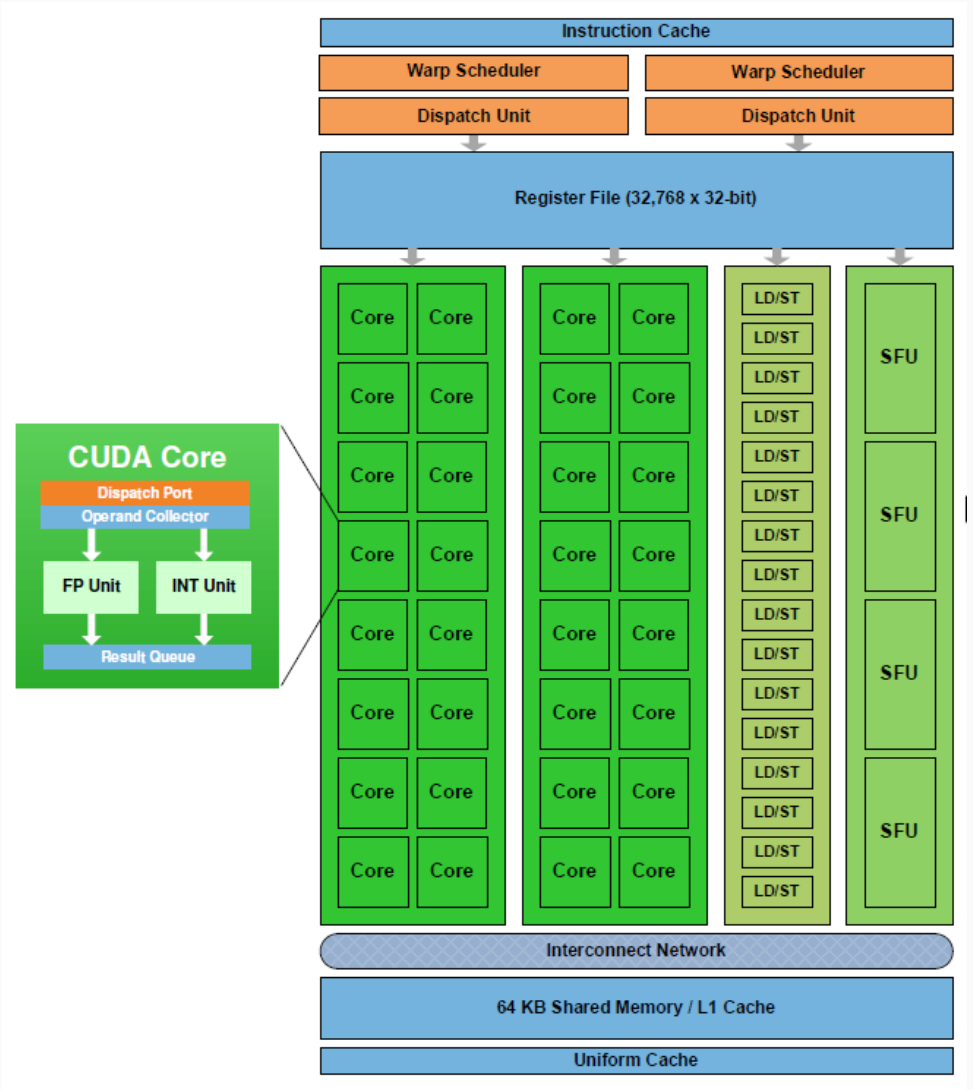
[1] The GFLOPS in this chart are based on GPU Boost Clocks.

Arquitetura Tesla: Primeiras GPGPU

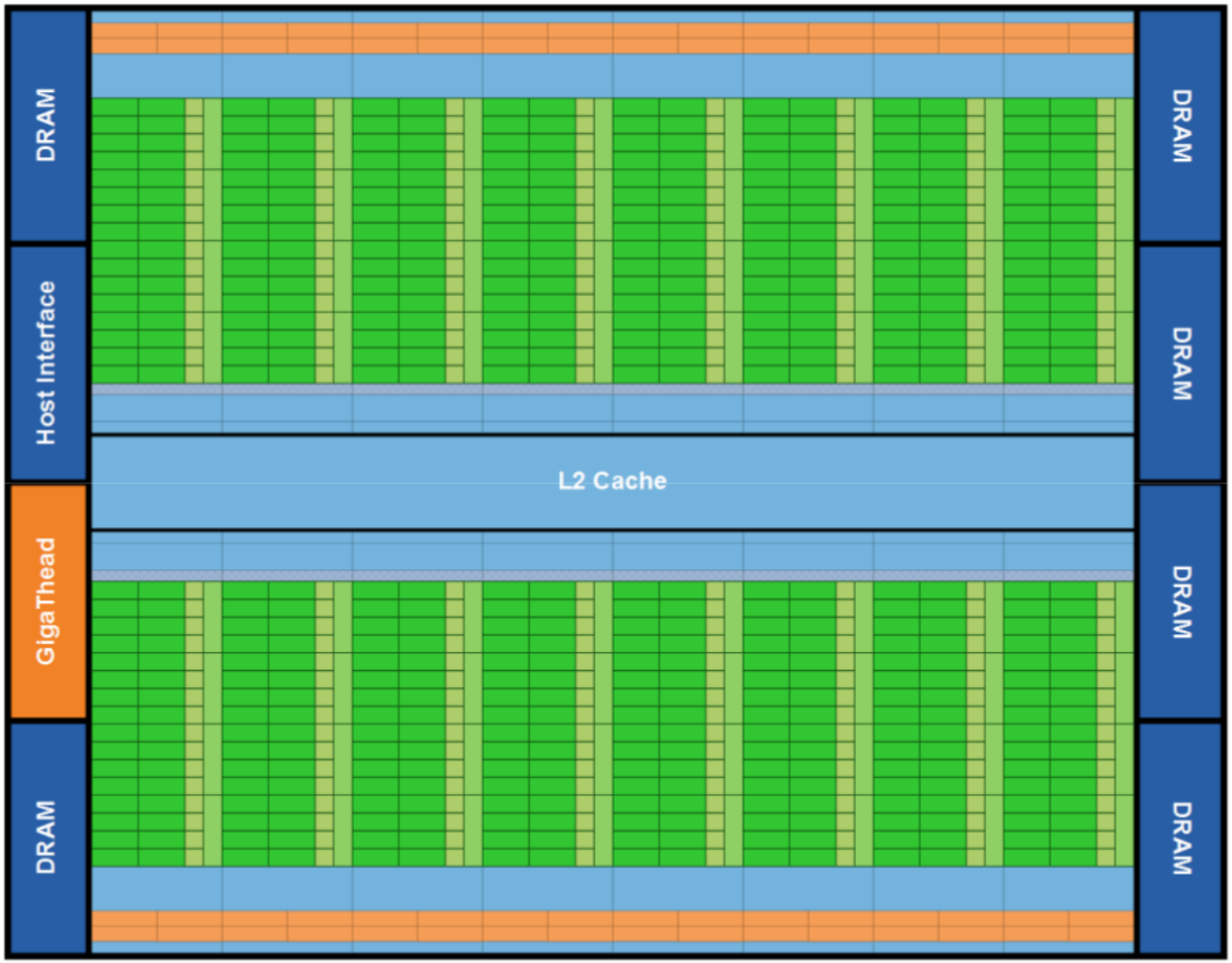


Multiprocessador da Arquitetura Fermi

- Processo de fabricação 40 nm
- 16 Multiprocessadores
- 32 processadores
- 16 unidades leitura/escrita
- 4 unidades funções especiais
- 48 kB Memória compartilhada
- 32768 registradores
- 2 escalonadores de warps

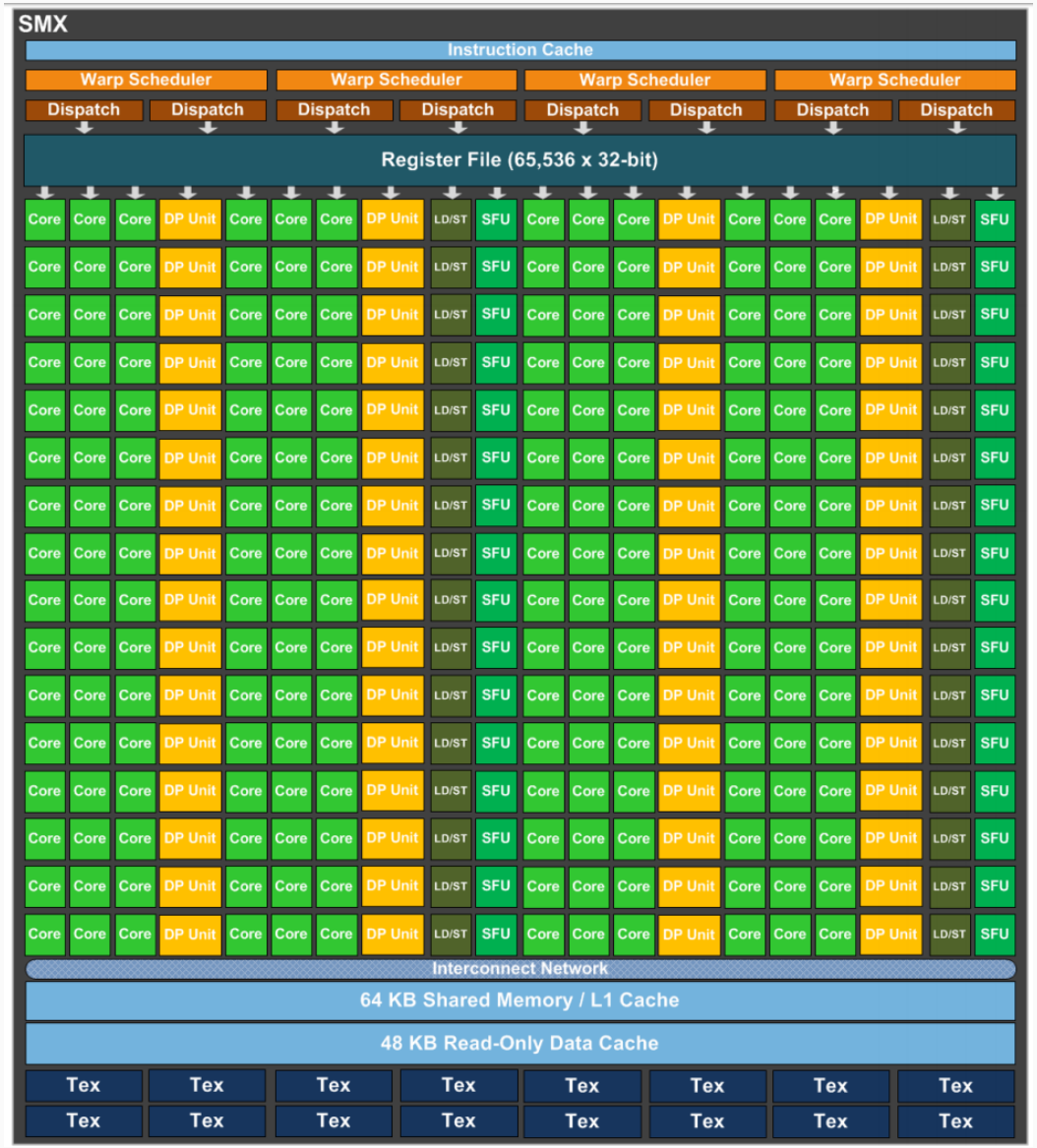


Arquitetura Fermi



Multiprocessador da Arquitetura Kepler

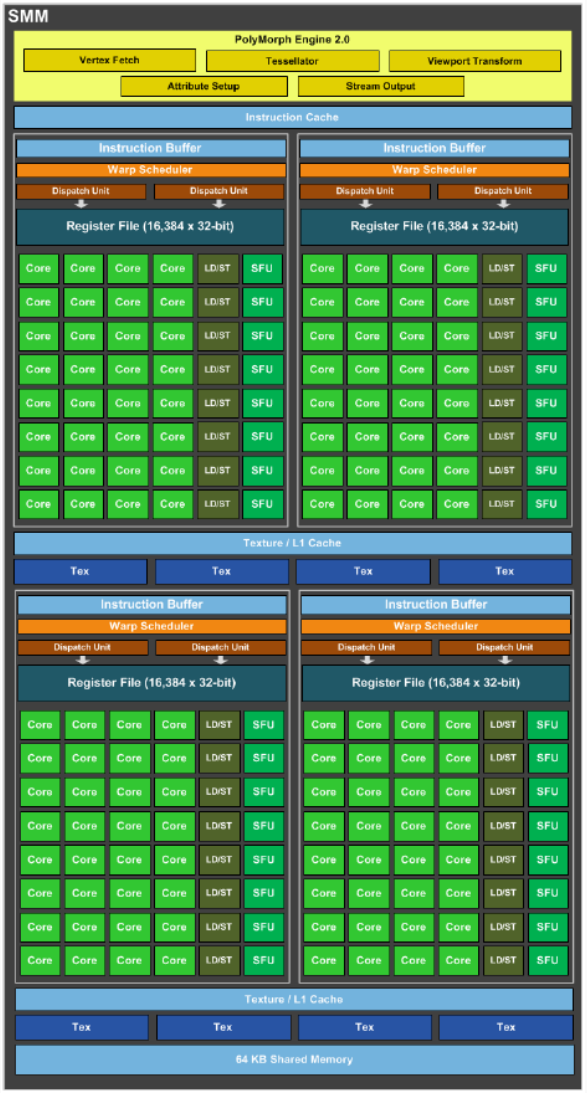
- Processo de fabricação 28 nm
- 15 SMX de 192 cores
- 2880 CUDA Cores
- 65 KB do arquivo de registradores
- 3072 KB de cache L2
- Até 24 GB de memória GDDR5
- Hyper-Q
- Paralelismo dinâmico
- 4 escalonadores de warps



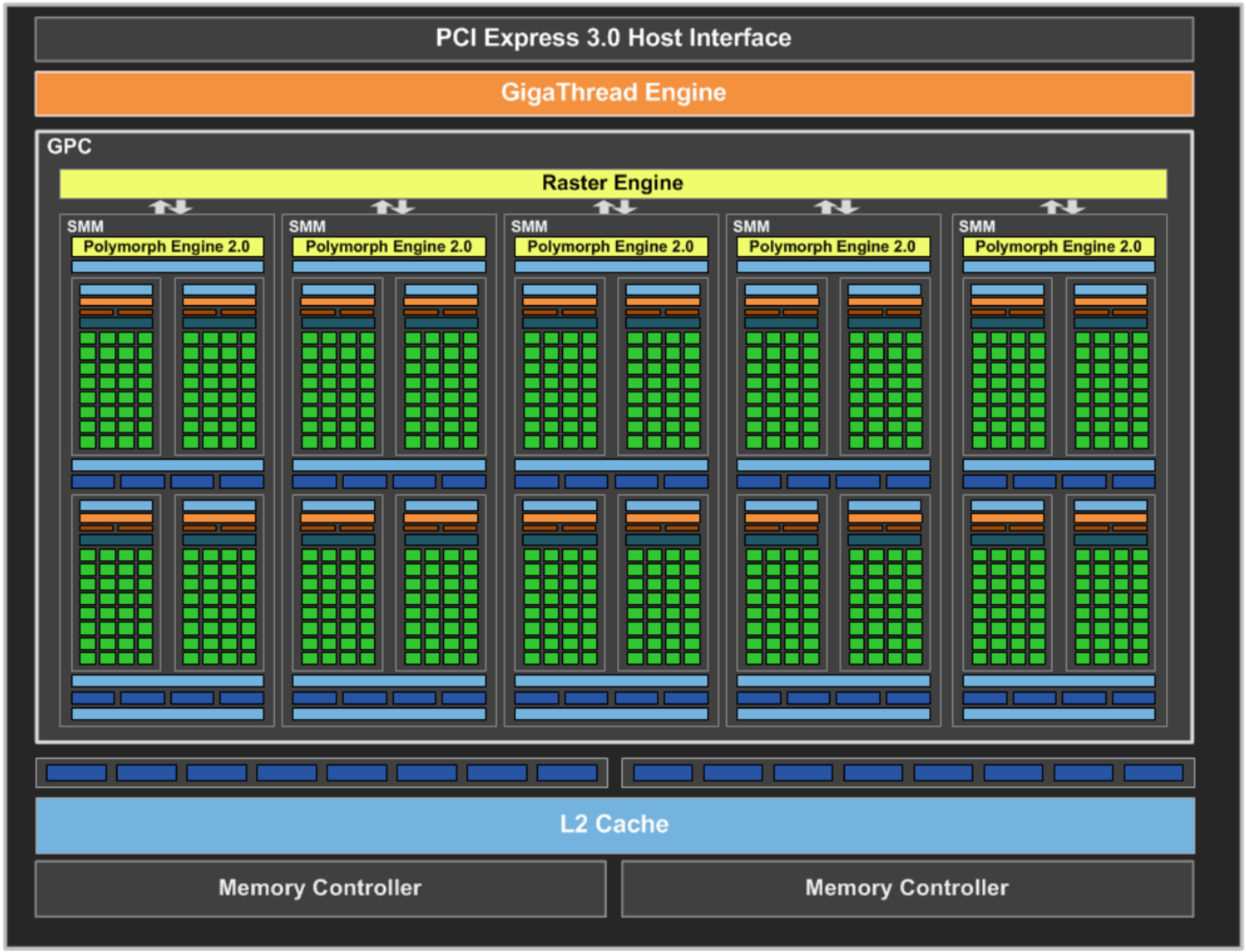
Arquitetura Kepler

Multiprocessador da Arquitetura Maxwell

- Clusters de Processamento Gráfico (GPC)
- 128 CUDA cores por SMM
- Dividido em 4 distintos 32 cores FP
- 96 KB de memória compartilhada dedicada
- 48 KB de cache L1/cache textura (unificado)
- 256 KB do arquivo de registradores
- 2 MB de memória cache L2
- 4 escalonadores de warps
- 128 cores de Mawell têm 90% do desempenho de 192 cores da Kepler



Arquitetura Maxwell



Multiprocessador da Arquitetura Pascal

Figura 2: Multiprocessador da arquitetura Pascal

Arquitetura Pascal

- SMs of 64 cores de precisão simples
- 32 unidades de precisão dupla
- Cada SM está dividido em 2 blocos de processamento
- 256 KB do arquivo de registradores
- 4096 KB de cache L2
- 4096-bit do barramento de interface de memória
- Um escalonador de warps
- Preempção na computação
- HBM Stacked DRAM
- NVLink

Arquitetura Pascal: NVLink

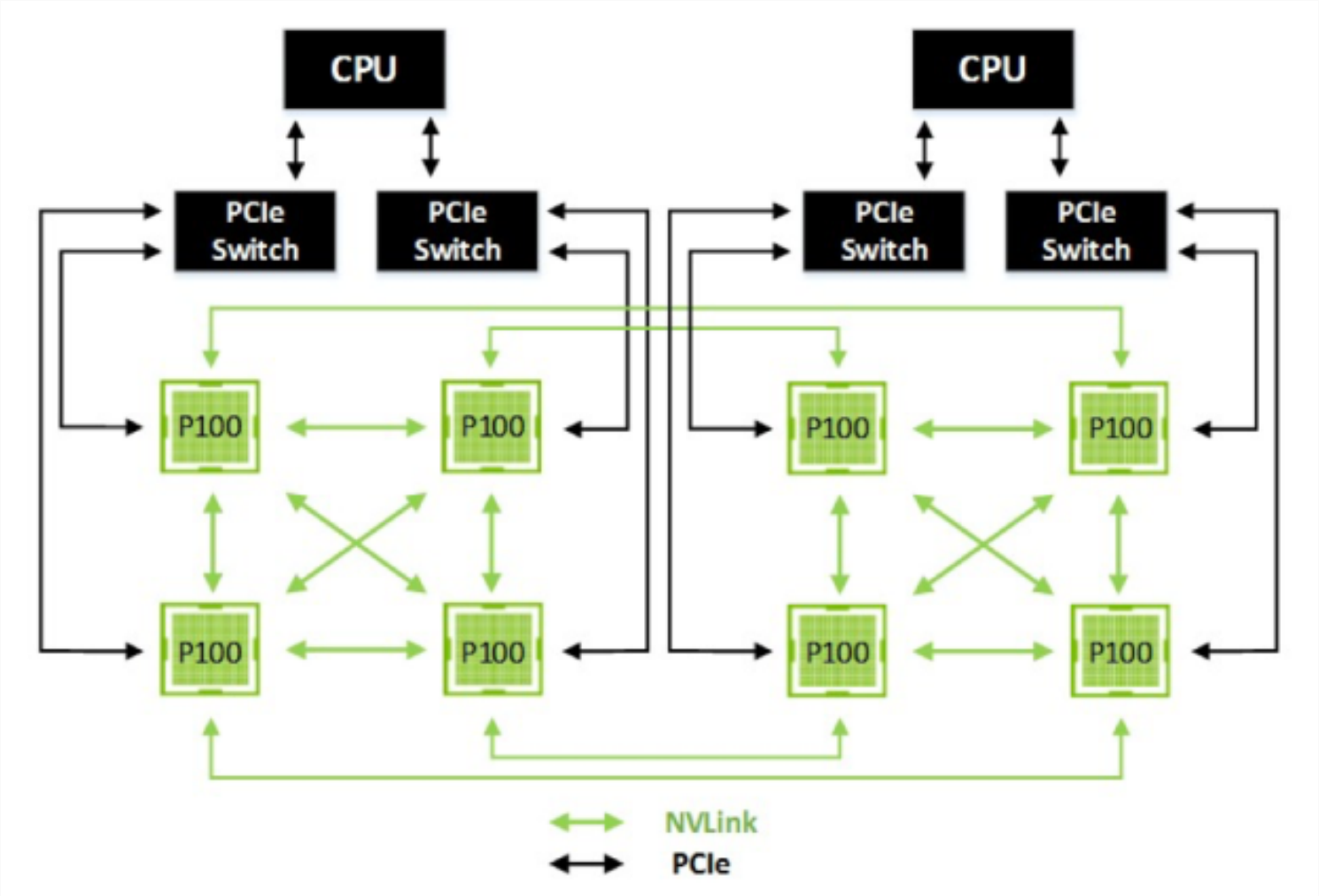


Figura 3: NVLink ligando 8 Tesla P100 em uma topologia Hibrida Cubo malha

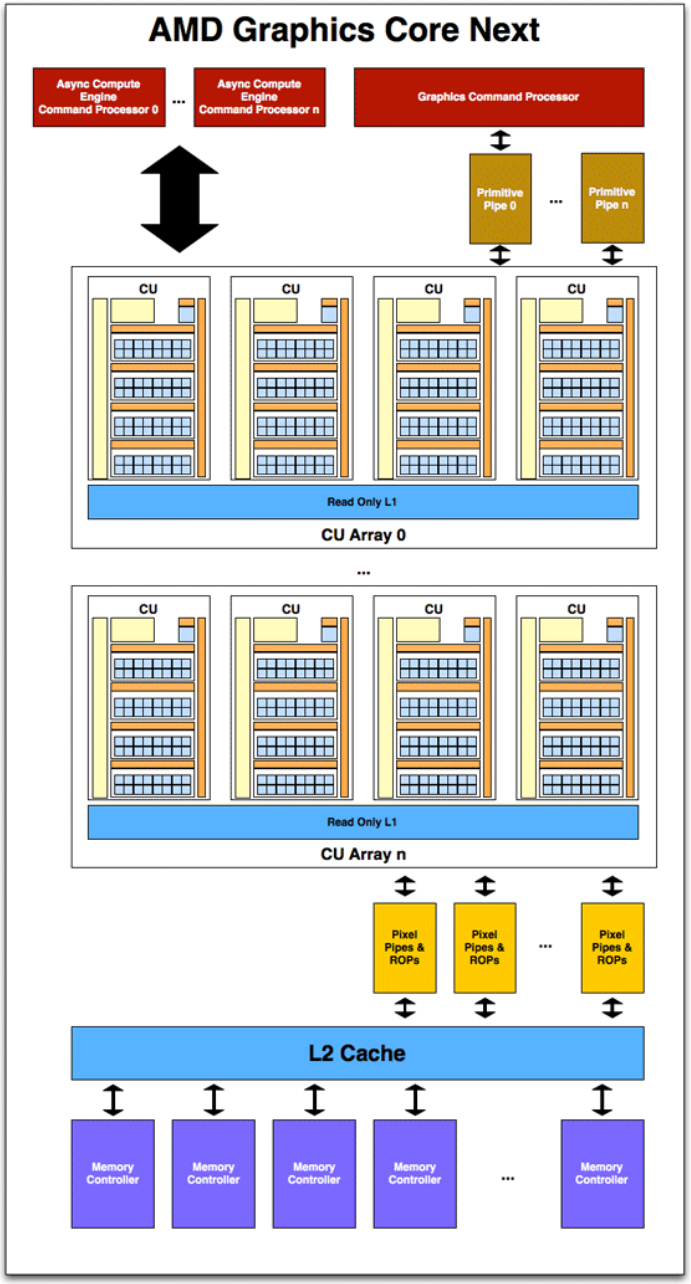
GPGPUs AMD

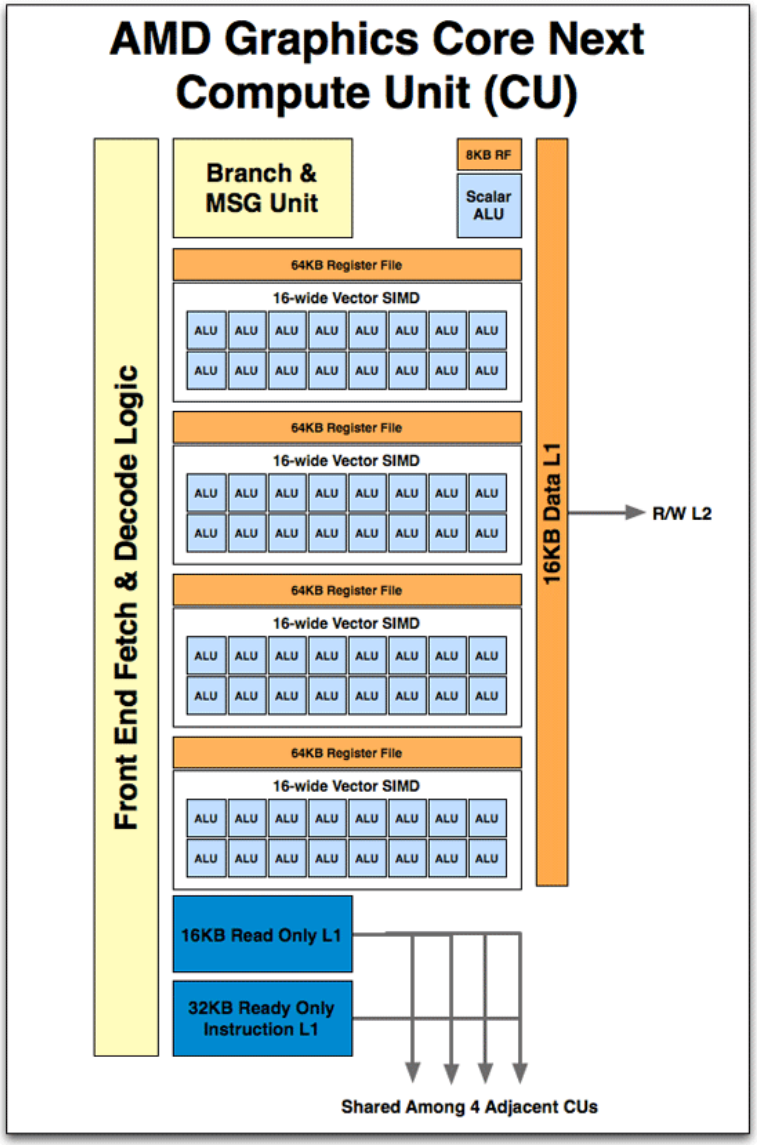
GPUs de Próposito Geral da AMD

Evolução das GPUs AMD:

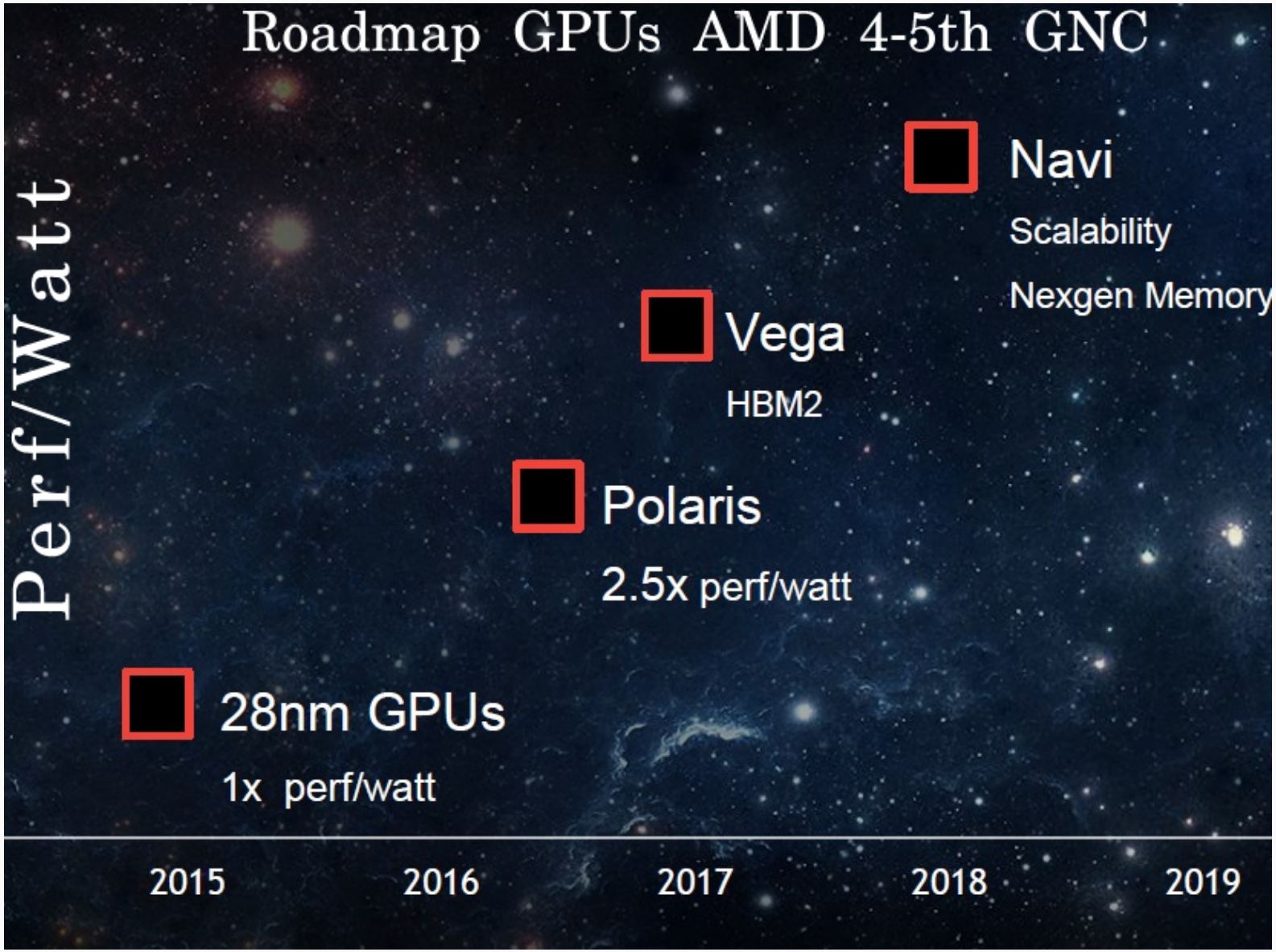
- Array Technology Inc. (ATI): 1985
- Visual Processing Unit (VPU), da ATI
- Advanced Micro Devices (AMD) adquire a ATI: 2006
- Graphic Core Next (GCN): 2011
- AMD Radeon HD 7790, usa GCN 1.1: 2012
- GCN: 5a. Geração

Graphic Core Next

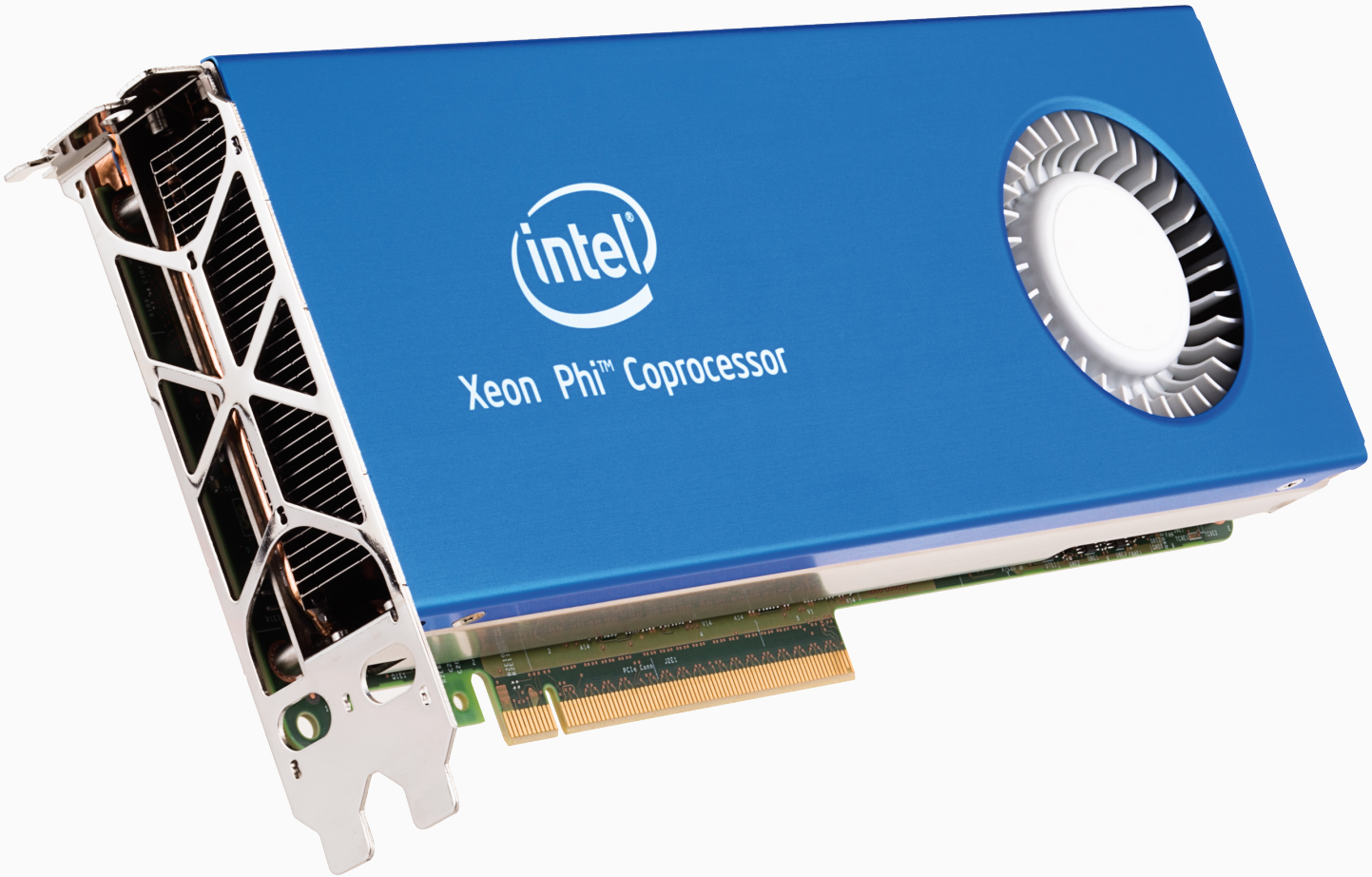




Roadmap para as GPUs AMD



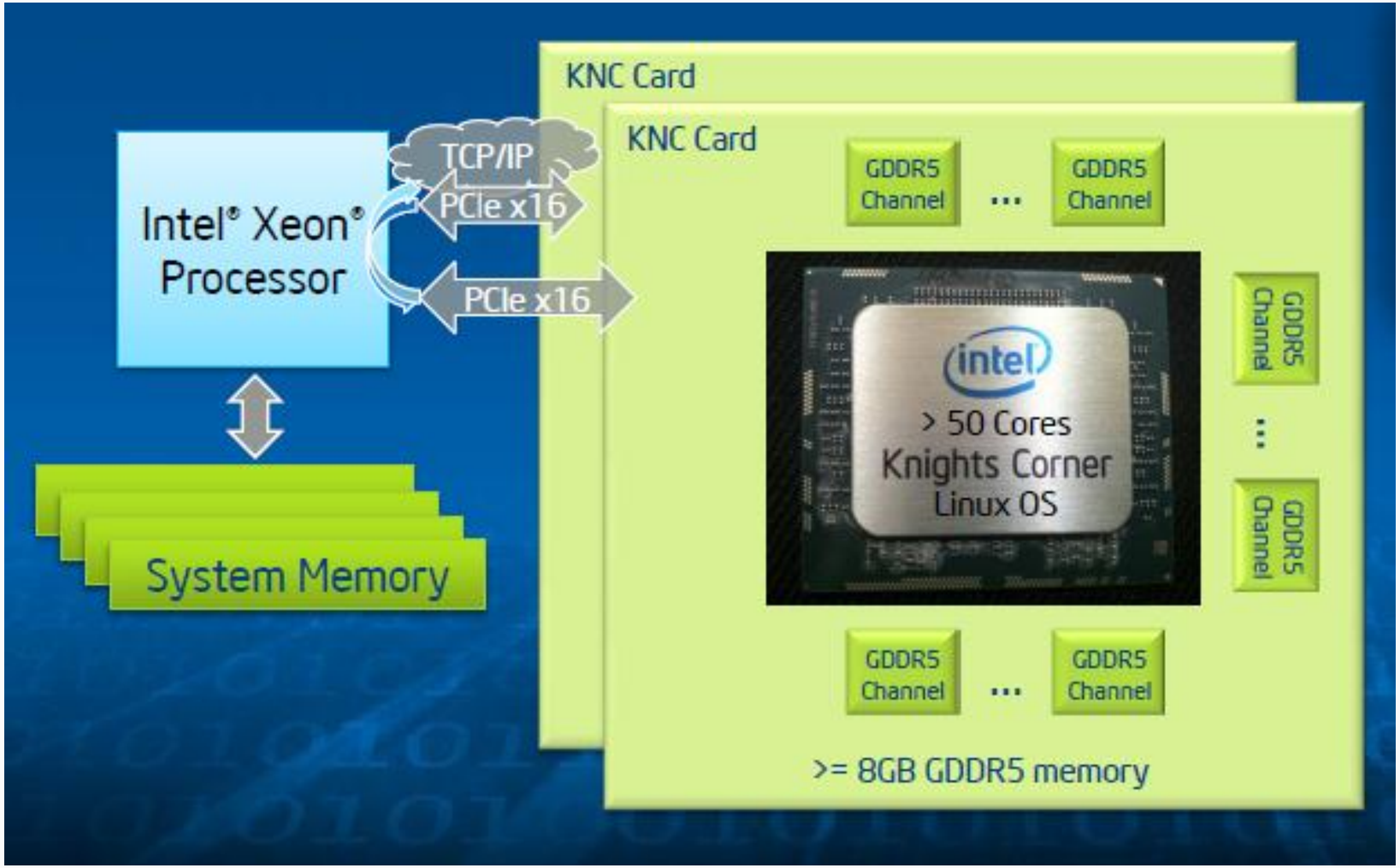
Intel Xeon Phi



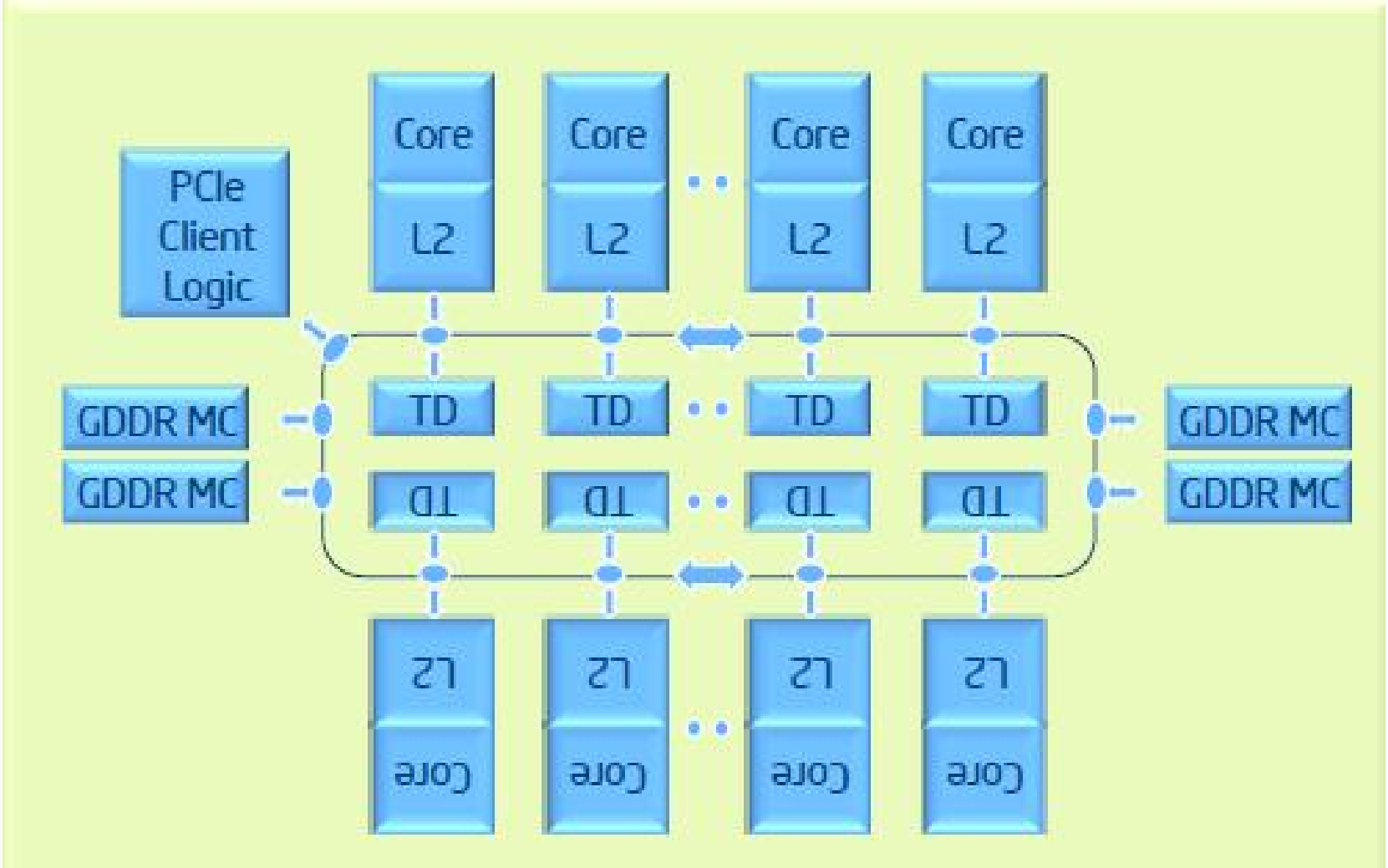
Roadmap para Intel Xeon Phi

- Larrabee: 2006-2010 $\rightarrow$ Knights Ferry: 2010
- 1a. geração x100 Knights Corner: 22 nm
- 2a. geração x200 Knights Landing: 14 nm
- 3a. geração Knights Hill: 10 nm

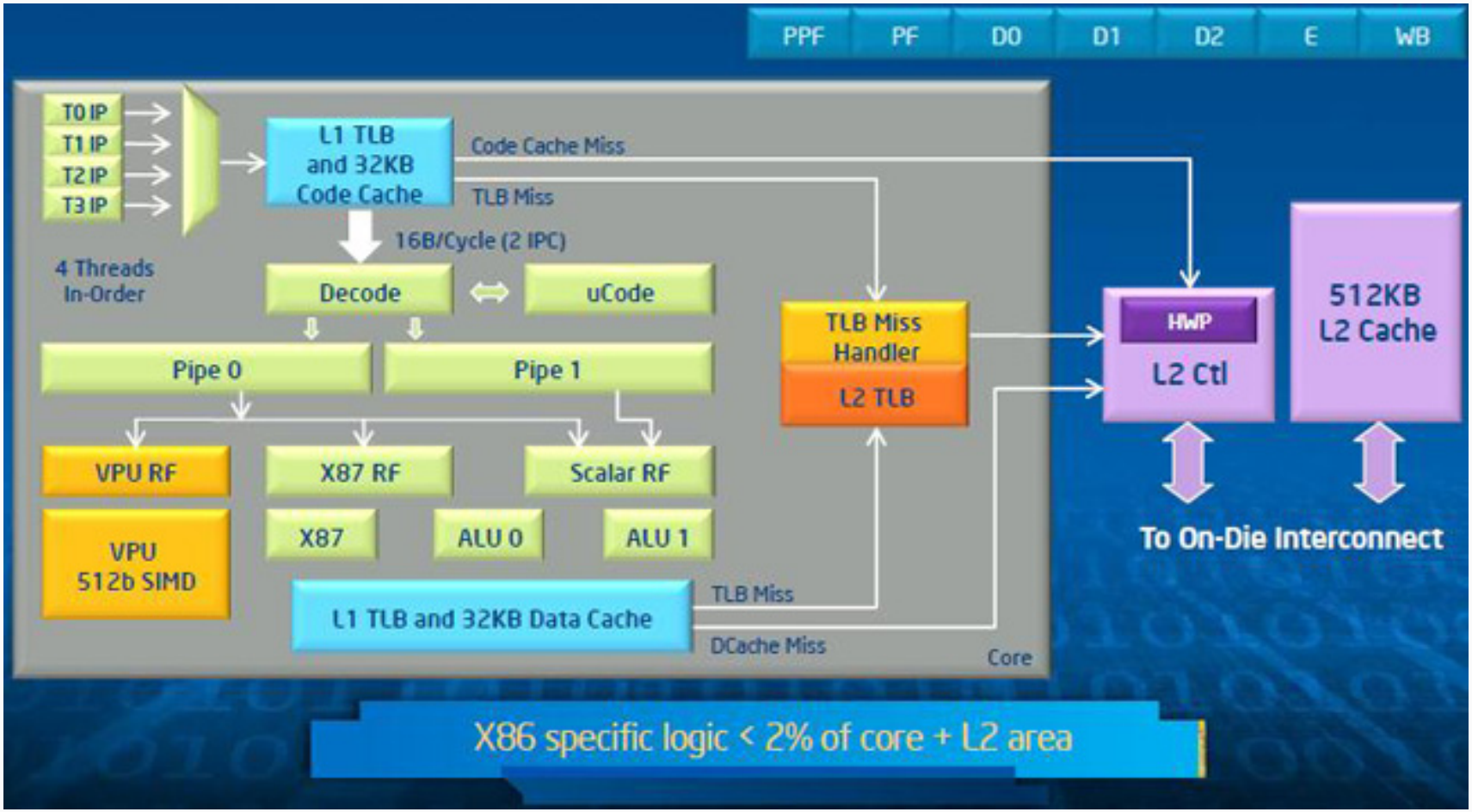
Knights Corner: Arquitetura



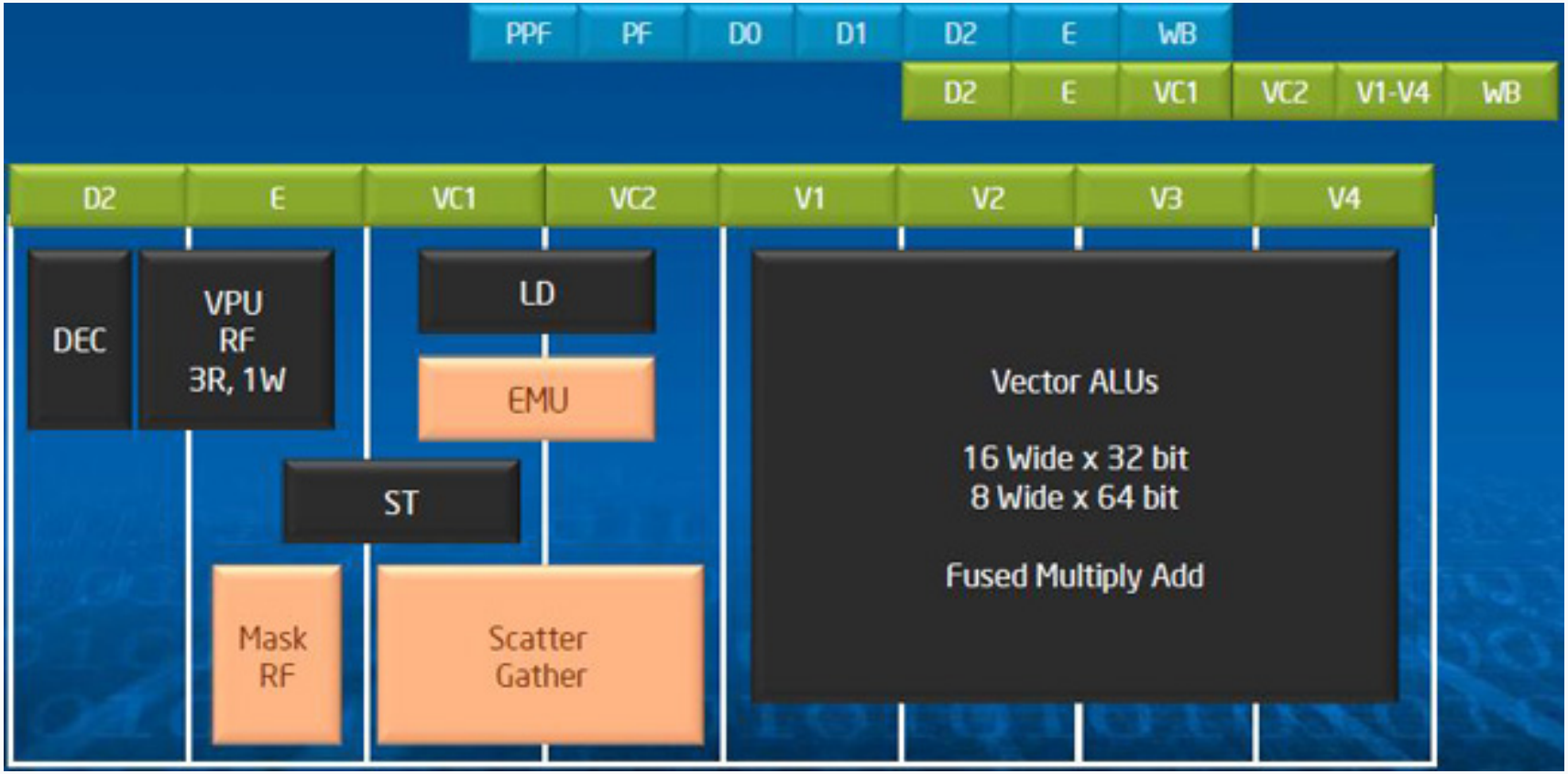
Anel de Interconexão Bidirecional



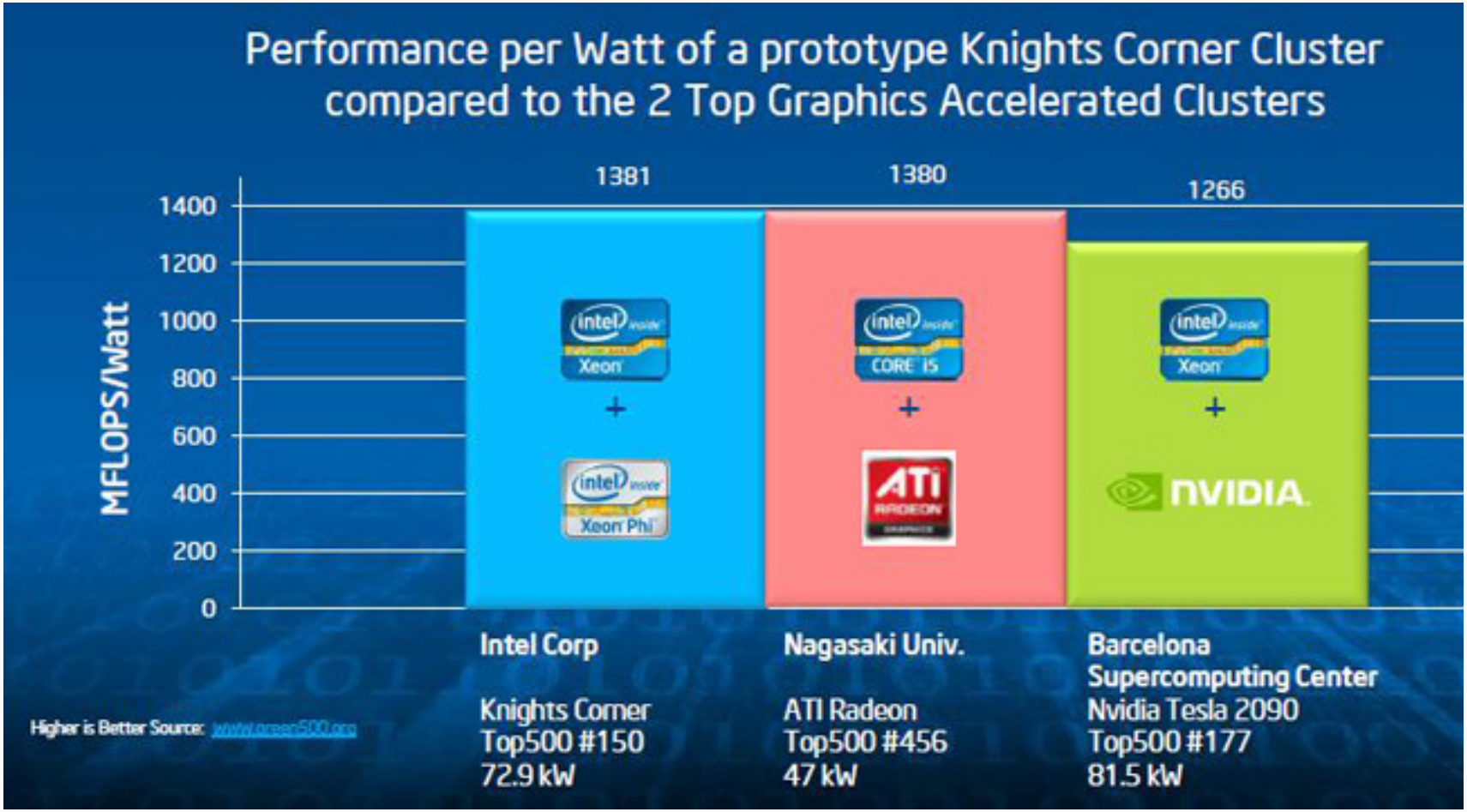
Core do Xeon Phi



Unidade de Processamento Vetorial



Desempenho por Watts de alguns aceleradores



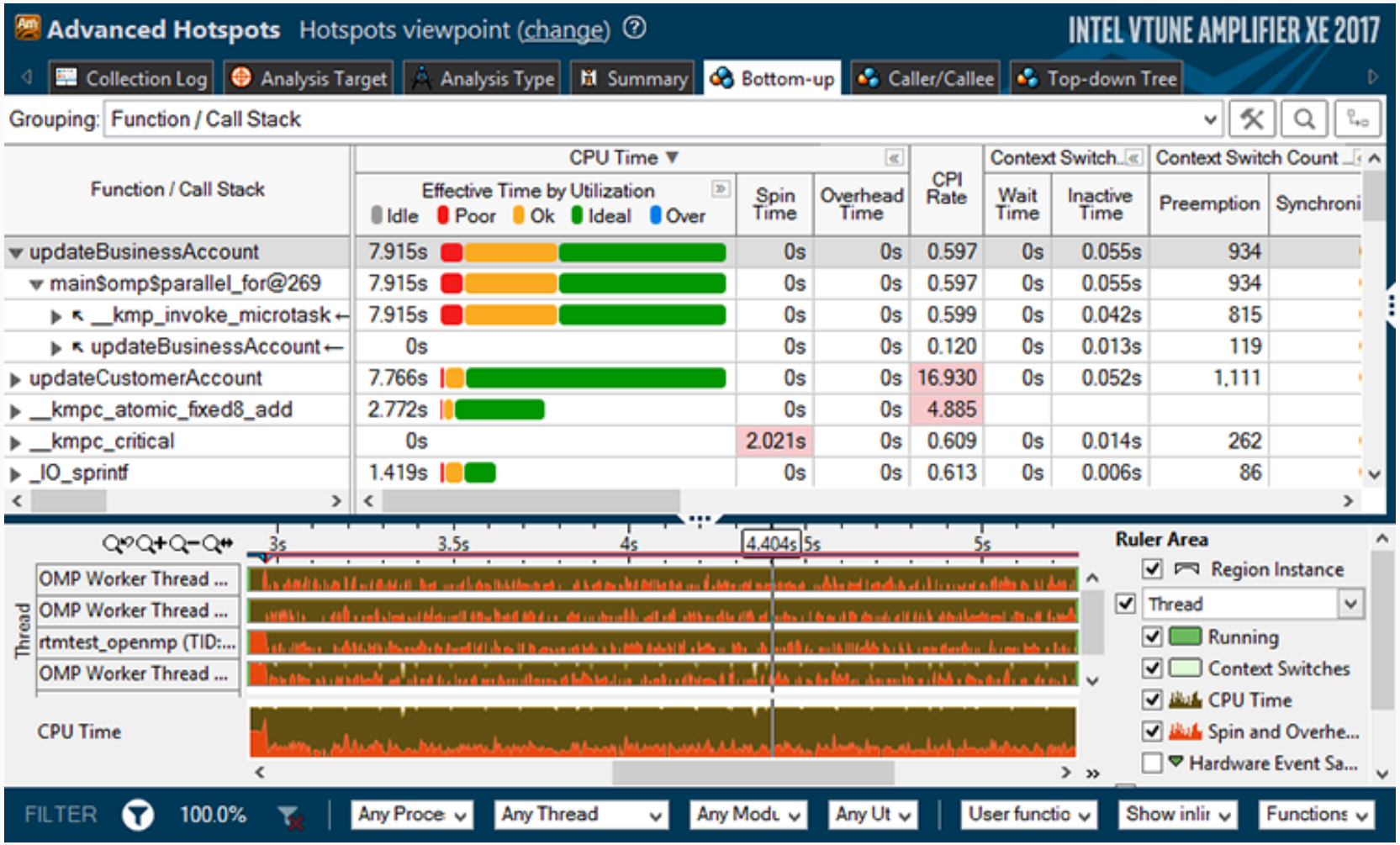


Figura 4: Intel Vtune Amplifier

OpenACC

OpenACC:

- Padrão para Programação Paralela
- Baseado no compilador PGI (Portland Group)
- Fornece coleção de diretivas para especificar laços e regiões de código paralelizáveis

FPGAs, GPUs e Xeon Phi

Pedro Bruel
phrb@ime.usp.br
Marcos Amarís
amaris@ime.usp.br

28 de Março de 2017

Instituto de Matemática e Estatística
Universidade de São Paulo