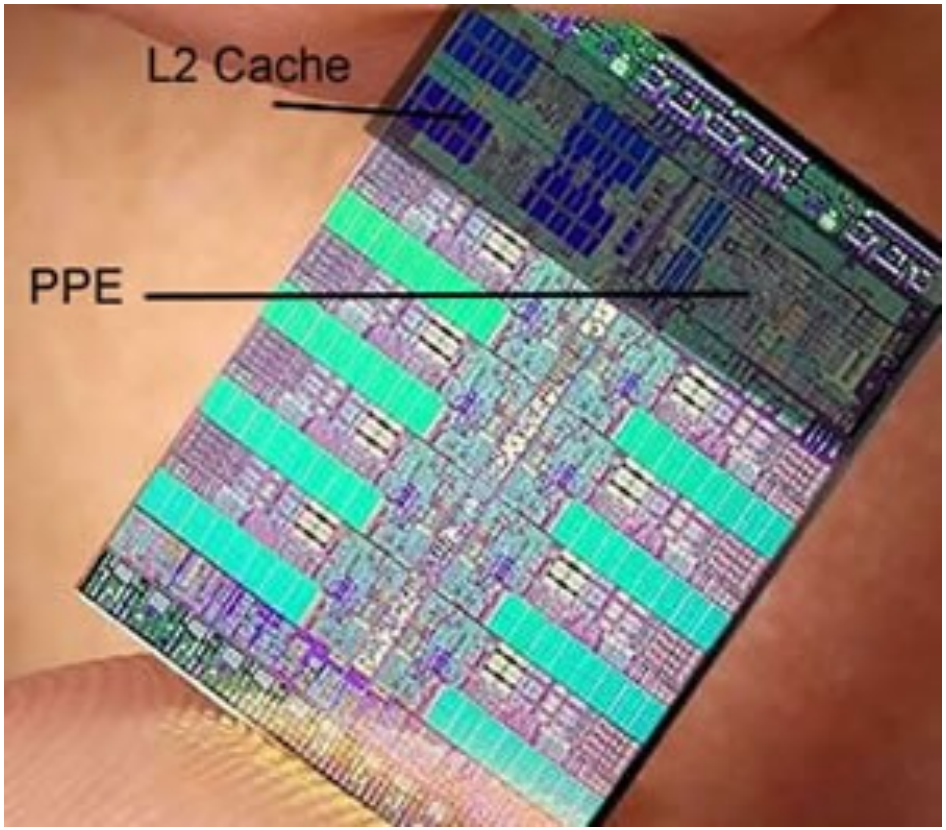
Multi-core/Many-core, GPU Architectures (2)

Overview of the Cell B.E. (I)

- 9 cores in total
  - Heterogenous architecture

- 1 PowerPC Processing Element (PPE)

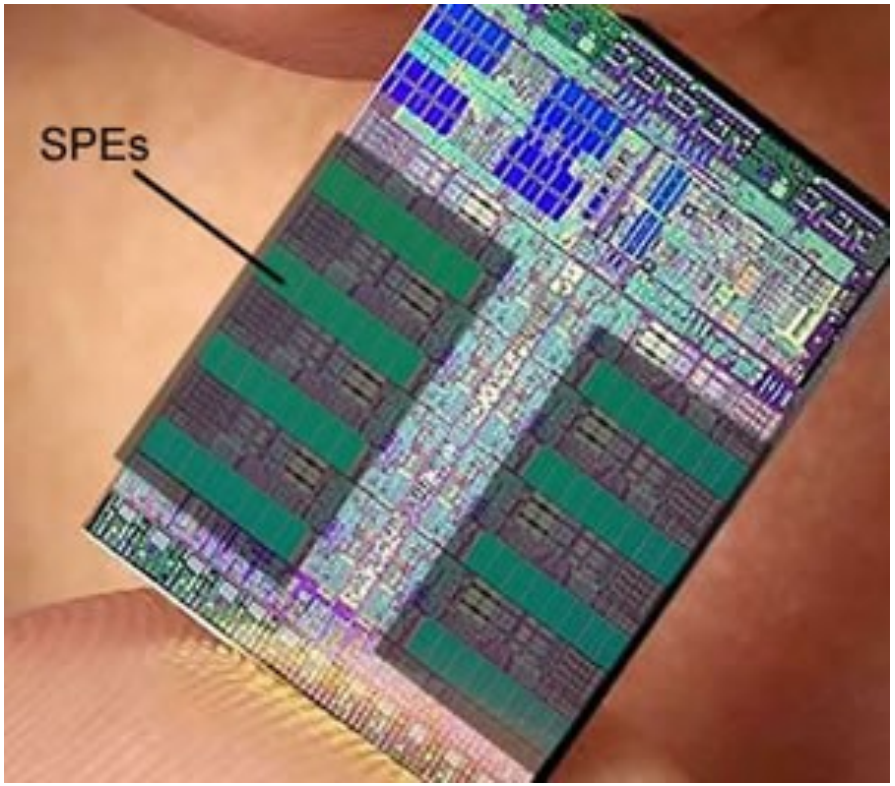
- 64 KB L1 cache
- 512 KB L2 cache
- Dual issue core
- In-Order execution



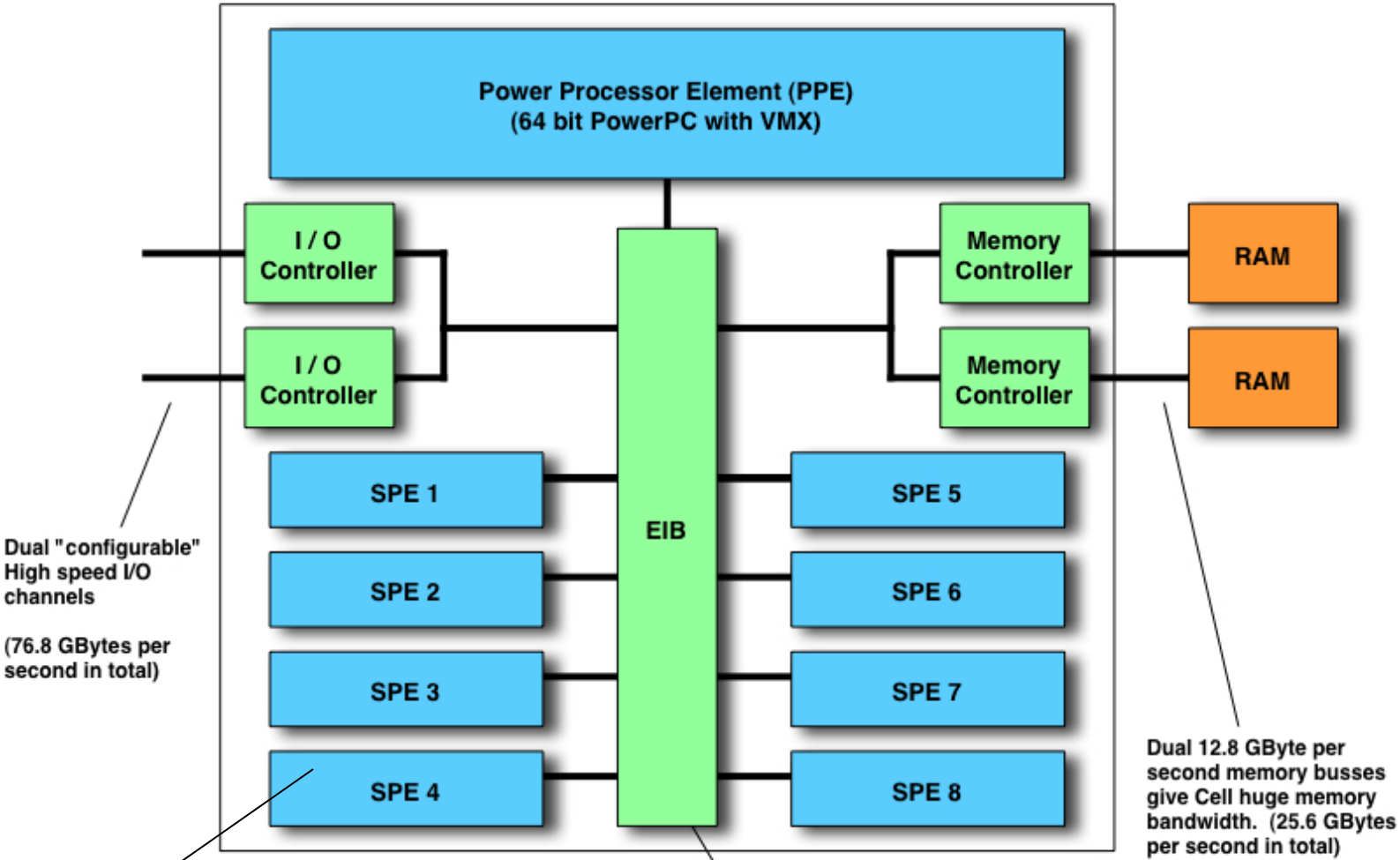
Overview of the Cell B.E. (II)

- 8 Synergistic Processing Elements

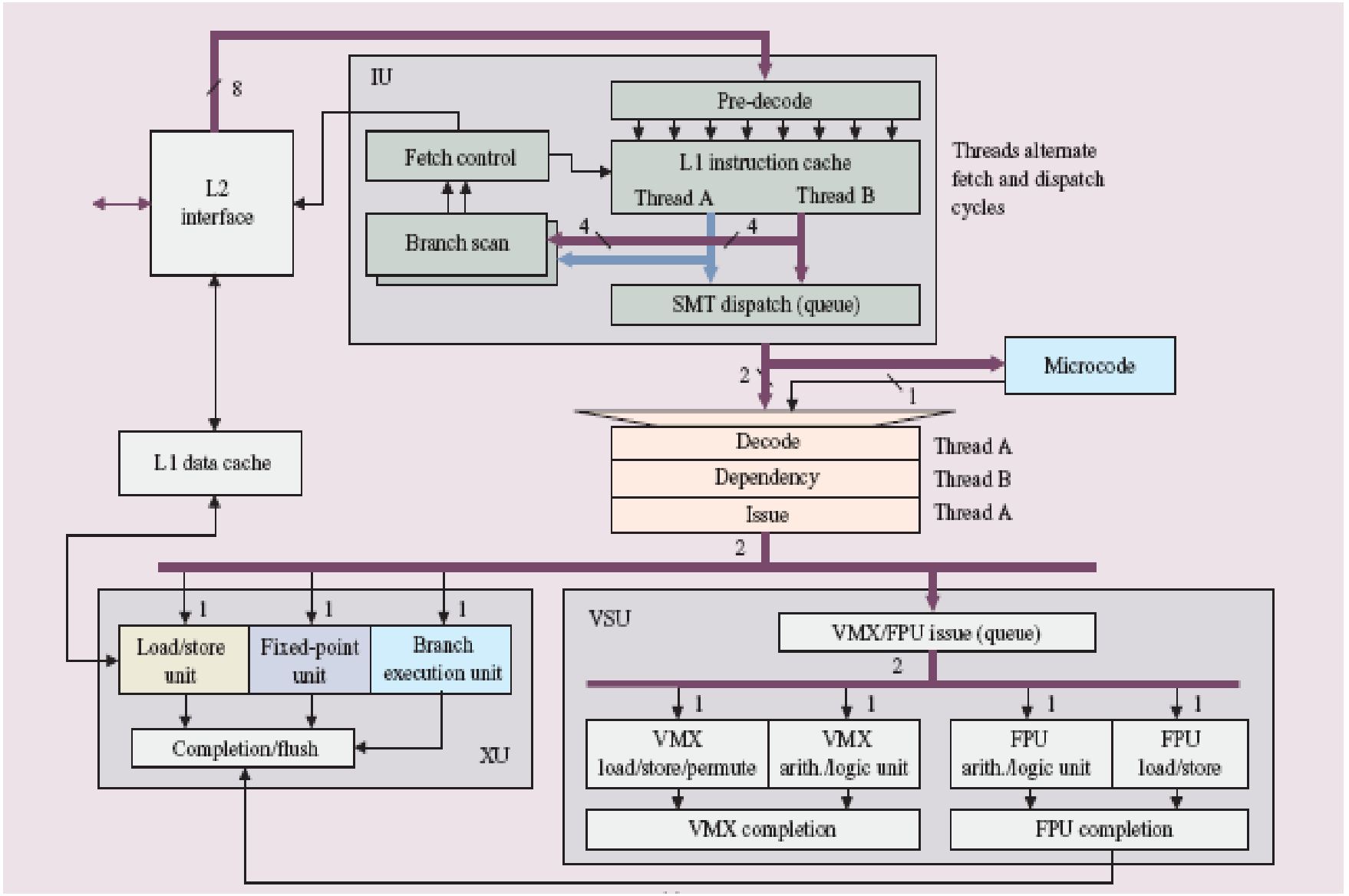
- Each with 256 KB local memory (not cache)
- Dual issue core
- In-Order execution
- No branch prediction



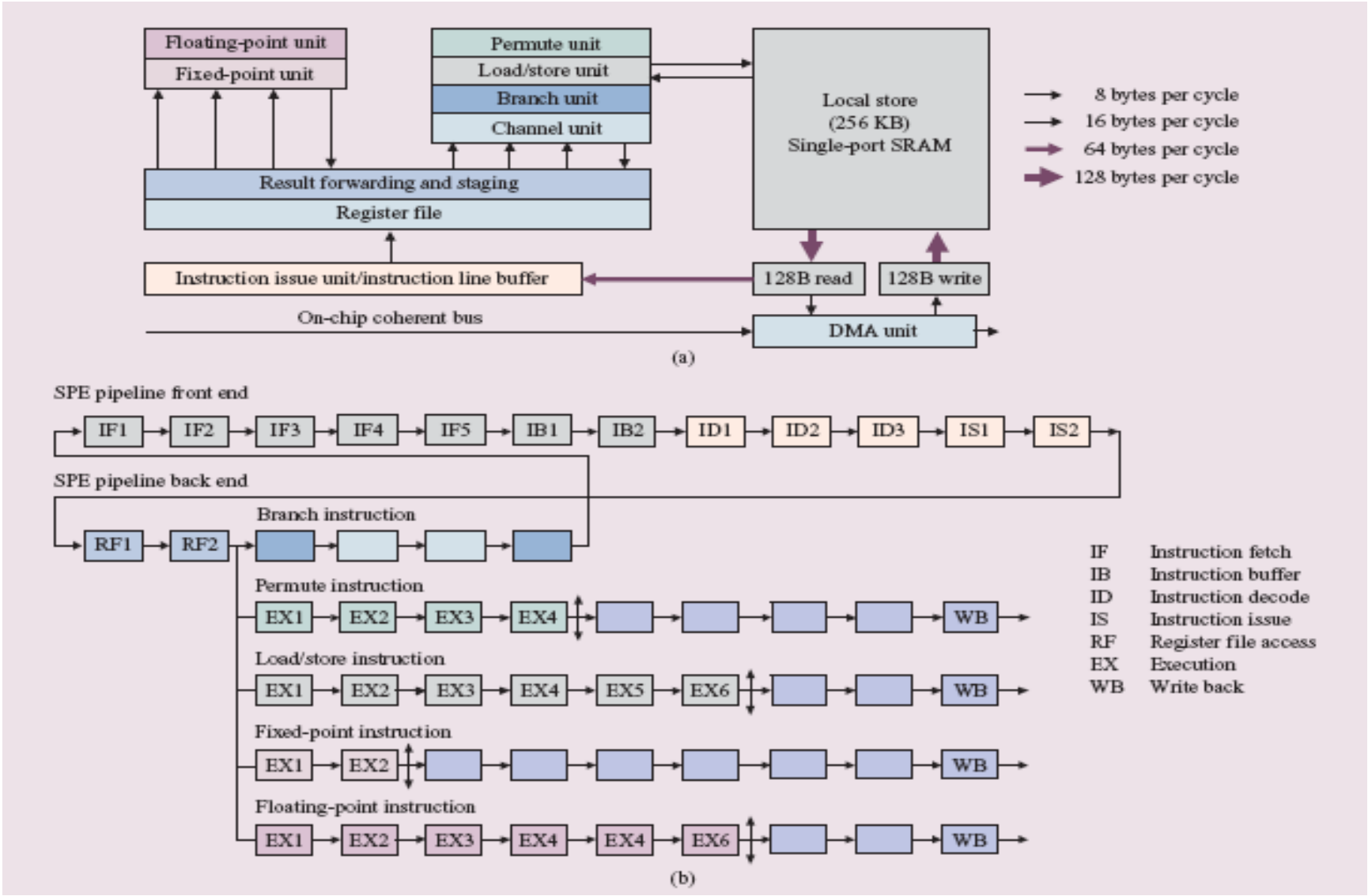
Cell Processor Architecture



Power Processor Element (PPE)

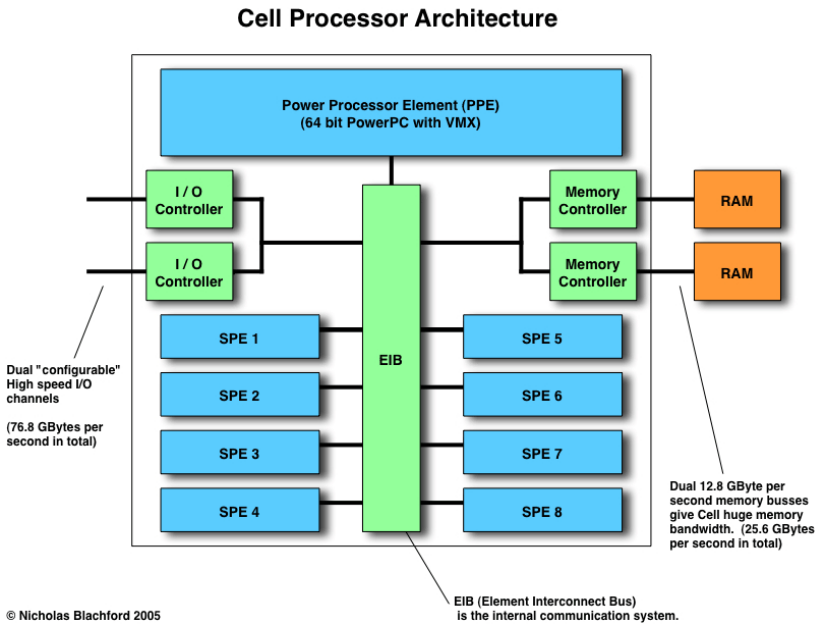


Synergistic Processing Elements



Element Interconnect Bus: NoC

- 9 cores are interconnected
  - PPE , 8 SPEs, memory controller (MIC) & off-chip I/O interfaces
- Ring topology
  - 4 rings
  - Bus width is 16 bytes working at 1.6GHz



Programming the Cell is challenging

Issues

- Dividing program among different cores
- Creating instructions in a different language for the 8 SPEs than for the PowerPC core.
- Need to think in terms of SIMD nature of dataflow to get maximum performance from SPUs
- SPU local store needs to perform coherent DMA access for accessing system memory

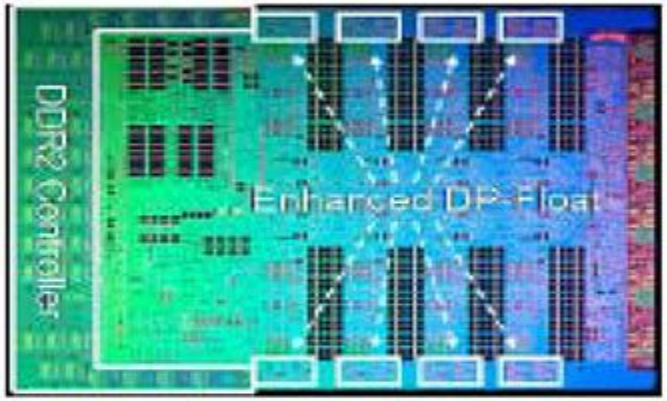
PowerXCell 8i processor

新IBM PowerXCell™ 8i プロセッサー

新 **PowerXCell 8i processor** は Cell Broadband Engine アーキテクチャと倍精度/単精度演算に強化最適化された8個のsynergistic processing element (SPE)を持つ汎用のPower Architecture™ コアの組み合わせ

- **性能を更新**
  - 倍精度演算性能: 108.8GFLOPSに向上 (eDP)
  - 1W当たりの計算能力強化
- **高いフレキシビリティ**

  広いアプリケーション分野をカバーするために下記の能力を備えている
  - 浮動小数点演算と整数演算
  - データ・ストリーミング/スループット サポート
  - リアルタイム サポート
  - C/C++, Fortran プログラミングモデルのサポート
- **セキュリティの拡張**
  - Virtual trusted computing environment for security



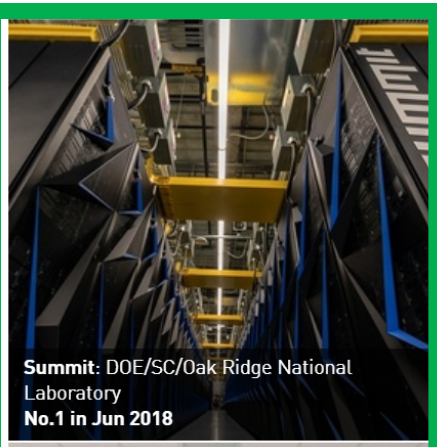
PowerXCell 8i processor

- 65 nm
- 9 cores, 10 threads
- 230.4 GFlops peak (SP) at 3.2GHz
- 108.8 GFlops peak (DP) at 3.2GHz
- Up to 25 GB/s memory bandwidth
- Up to 75 GB/s I/O bandwidth
- Top frequency >4GHz (observed in lab)

TOP #1 SYSTEMS

In the last 20 years, the following systems made it to the top of the TOP500 lists:

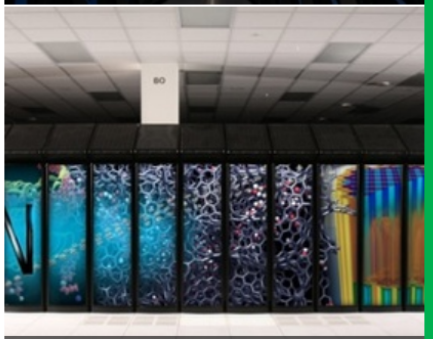
GPU



Summit: DOE/SC/Oak Ridge National Laboratory
No.1 in Jun 2018

Sunway TaihuLight: National Supercomputing Center in Wuxi
No.1 from Jun 2016 until Nov 2017

Tianhe-2 (MilkyWay-2) : National University of Defense Technology
No.1 from Jun 2013 until Nov 2015



Titan: Oak Ridge National Laboratory
No.1 in Nov 2012

Sequoia: Lawrence Livermore National Laboratory
No.1 in Jun 2012

K Computer: RIKEN Advanced Institute for Computational Science
No.1 from Jun 2011 until Nov 2011

Tianhe-1A: National Supercomputing Center in Tianjin
No.1 in Nov 2010

Jaguar: Oak ridge National Laboratory
No.1 from Nov 2009 until Jun 2010

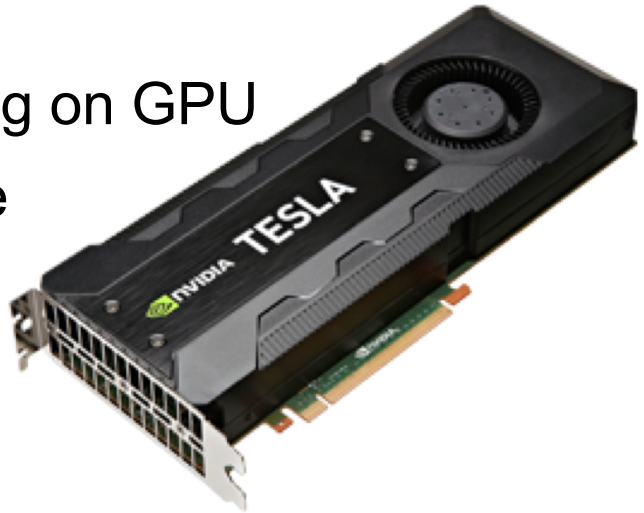
Roadrunner: Los Alamos National Laboratory
No.1 from Jun 2008 until Jun 2009

Cell

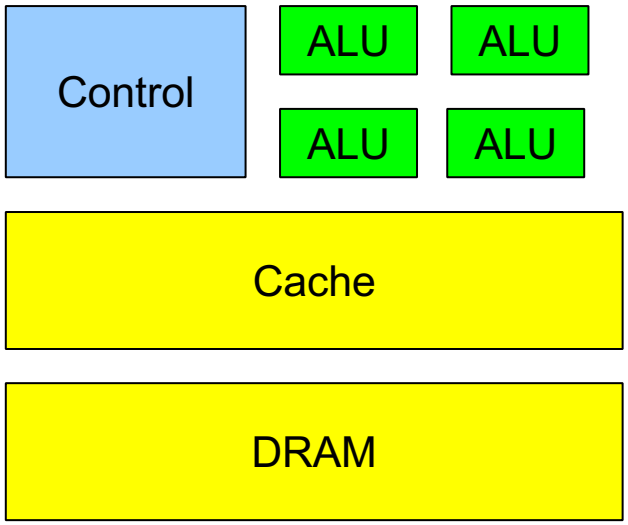
GPU

- GPU – Graphics Processing Unit
  - Special-purpose co-processors designed for high-end graphics processing in computers
  - Modern GPU architectures are many-core and multi-threaded
  - High computational power and memory bandwidth

- GPGPU – General Purpose computing on GPU
  - Utilizing GPUs for general-purpose computation

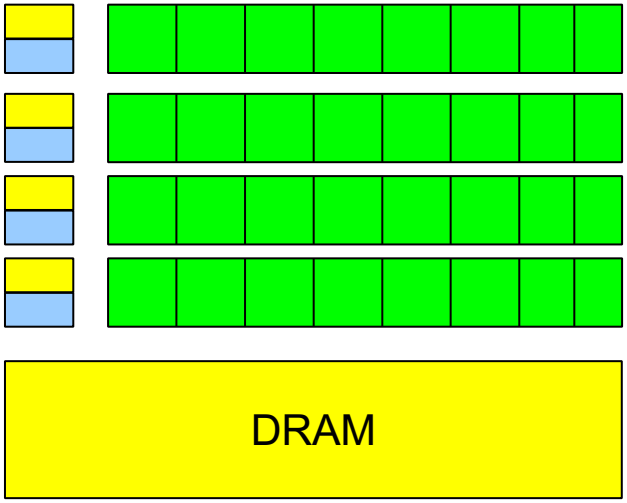


CPU vs GPU



CPU

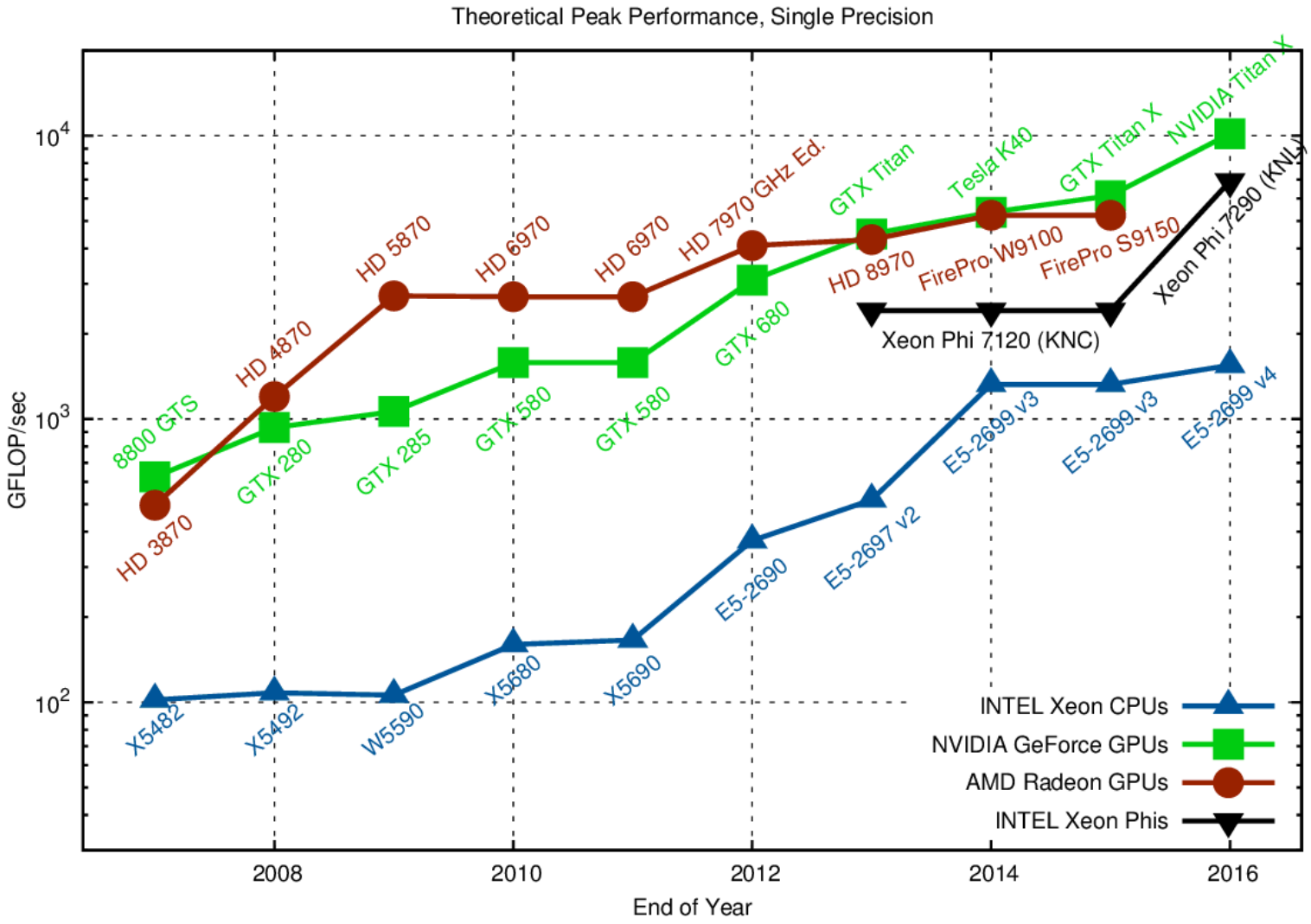
- Multi-core (2-16)
- Sophisticated arithmetic logic unit (ALU)
- Large cache memories



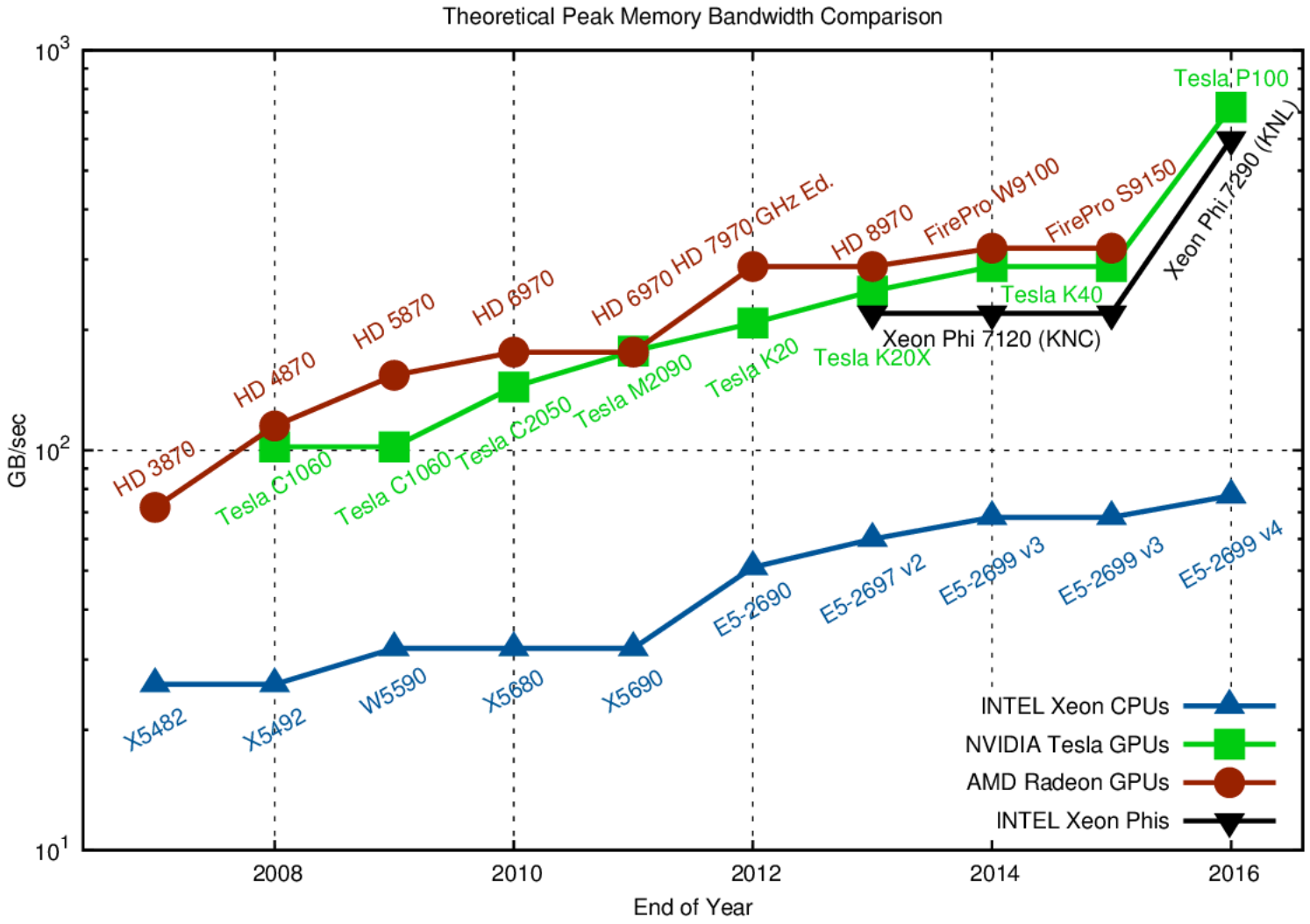
GPU

- Many-core (1000s)
- Minimized ALU
- High power efficiency

Performance Evolution



Memory Bandwidth Evolution

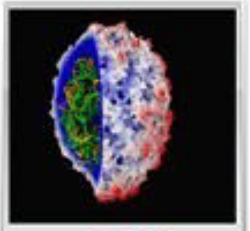


https://www.karlrupp.net/2013/06/cpu-gpu-and-mic-hardware-characteristics-over-time/

CPU vs GPU

146X

Interactive visualization of volumetric white matter connectivity

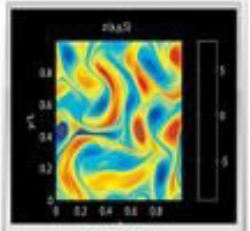


36X

Ionic placement for molecular dynamics simulation on GPU

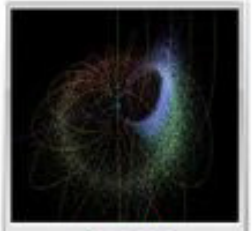
18X

Transcoding HD video stream to H.264 for portable video



17X

Simulation in Matlab using .mex file CUDA function

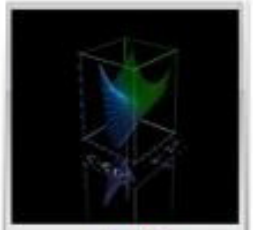


100X

Astrophysics N-body simulation

149X

Financial simulation of LIBOR model with swaptions



47X

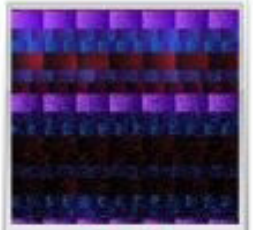
GLAME@lab: M-script API for linear Algebra operations on GPU

20X

Ultrasound medical imaging for cancer diagnostics

24X

Highly optimized object oriented molecular dynamics



30X

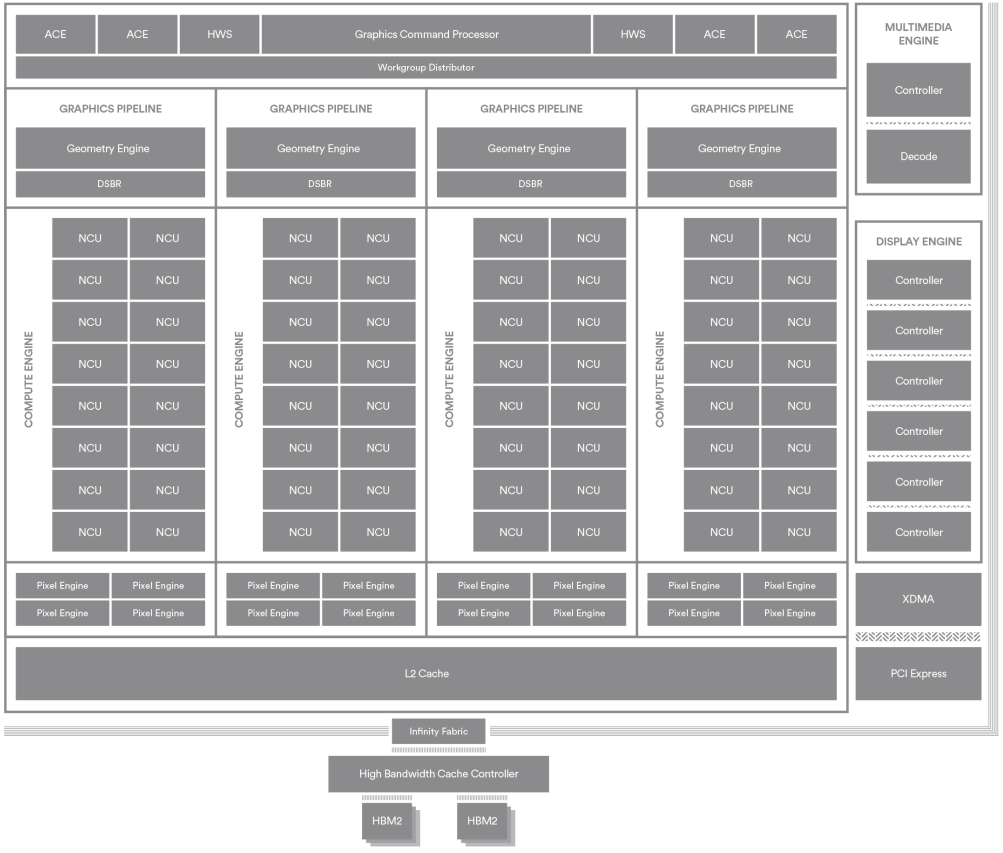
Cmatch exact string matching - find similar proteins & gene sequences

Modern GPU (AMD Vega)

"Vega 10" by the numbers

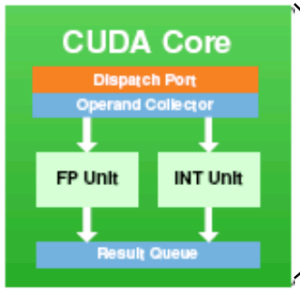
- **1** Graphics Engine
- **4** Asynchronous Compute Engines
- **4** Next-Gen Geometry Engines
- **64** Next-Gen Compute Units
- **4096** Stream Processors
- **256** Texture Units
- **64** Render Back-Ends
- **4 MB** L2 Cache
- **2048-bit** HBM2

AMD SIGGRAPH'17 Tech Day | Confidential – Under Embargo Until 7/30, 7:30pm PDT

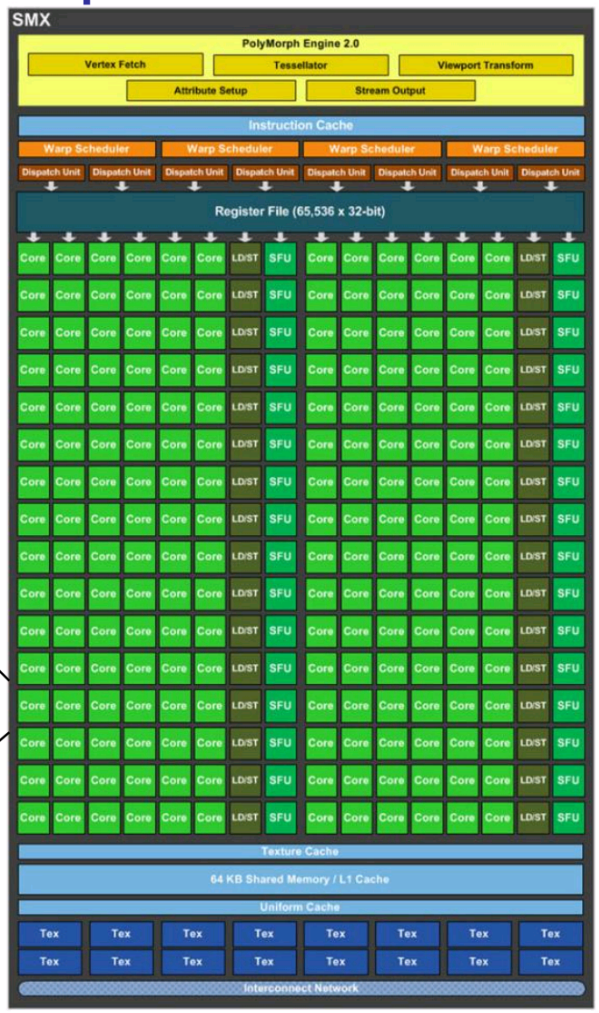


CUDA Core Architecture

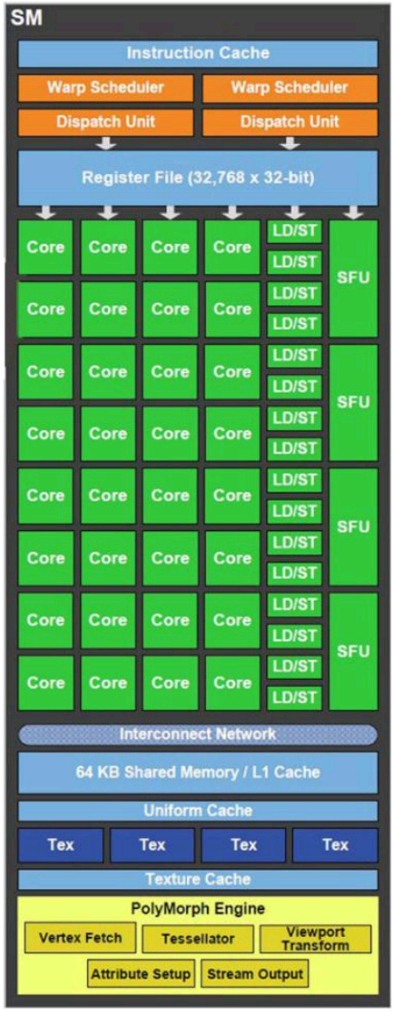
- SMX — Streaming Multiprocessor eXtreme
  - Multi-threaded processor core
  - Fundamental processing unit for CUDA thread block



Kepler SMX



Fermi SM



Modern GPU (NVIDIA Volta)

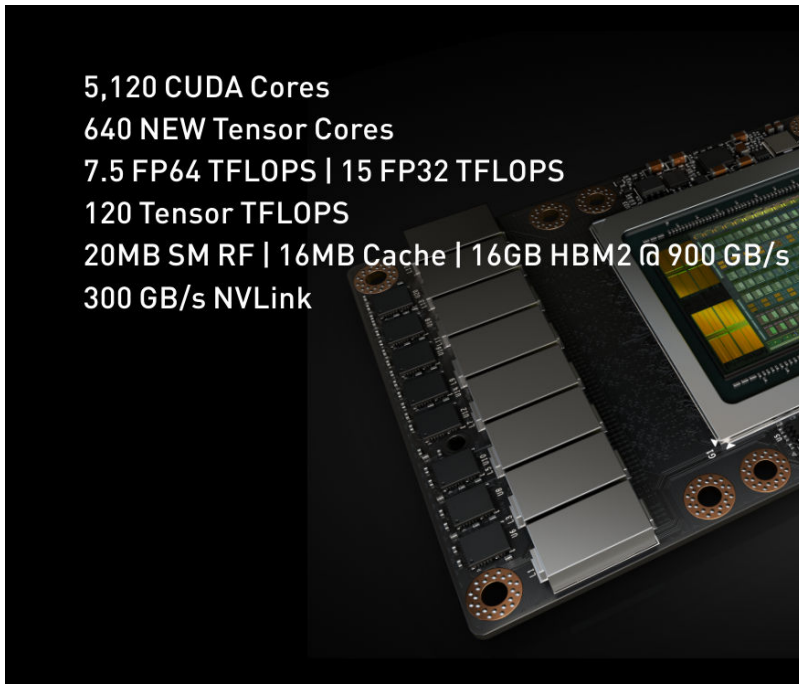
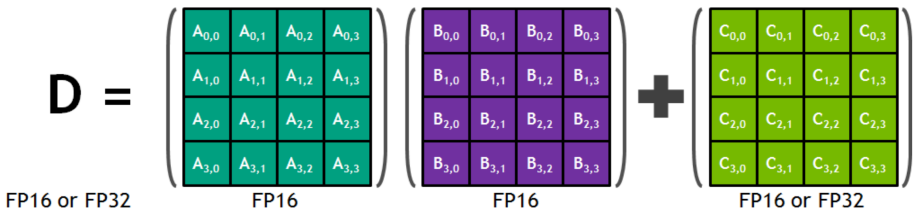


Figure 5. Volta GV100 Streaming Multiprocessor (SM)

Tensor core



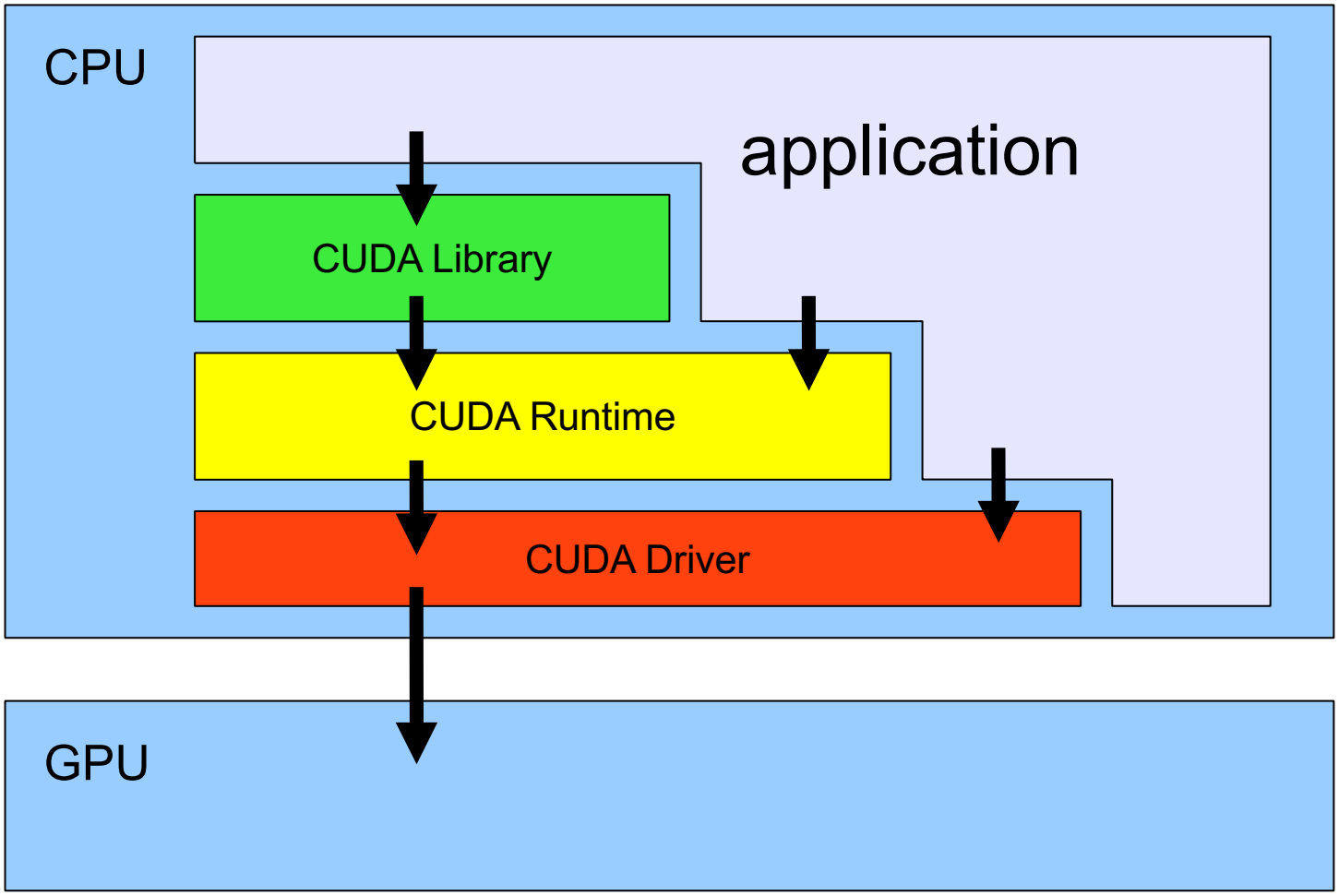
Generalized GPU programming

- **CUDA** (Compute Unified Device Architecture) is the first to drop graphics API; allows the GPU to be treaded as a co-processor to the CPU
  - Run thousands of threads on separate cores (with limitations)
  - High theoretical/achieved performance for data-parallel applications

CUDA Basics

- Proprietary technology for GPGPU programming from NVIDIA
- Not just API and tools, but name for the whole architecture
- Targets NVIDIA hardware only
- First SDK released Feb 2007
- SDK and tools available to 32- and 64-bit Windows, Linux and Mac OS.
- Tools and SDK are available for free from NVIDA's website.

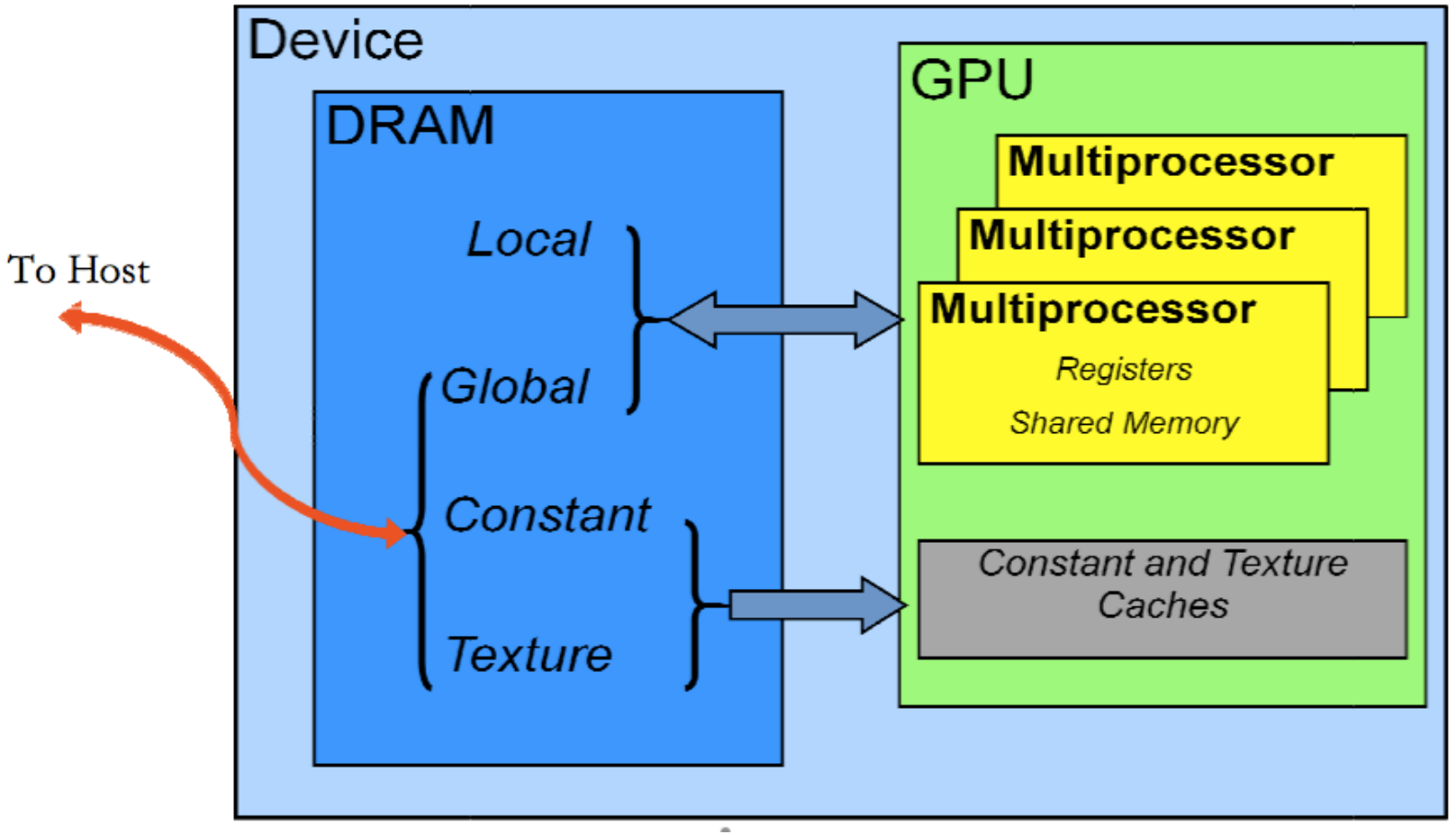
CUDA - Compute Unified Device Architecture



Definitions in CUDA

- **Thread** — execution unit of instruction flow
- **Thread block** — logical set of **threads** (<1024).
- **Warp** — group of threads inside thread block which are physically executed concurrently (32 **threads**)
- **Grid** — set of **thread blocks**.

CUDA Memory Model

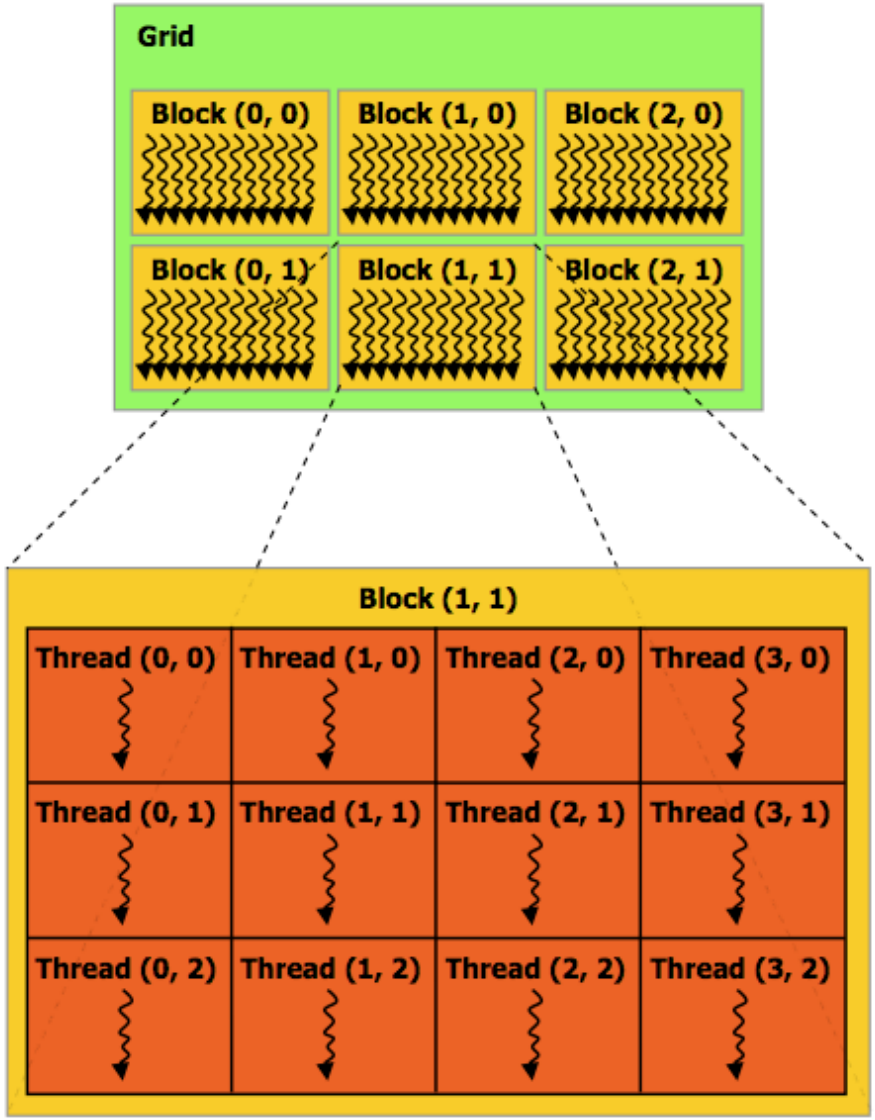


Programming model

- GPU has its own memory
- Program is split into threads and executed on SP
- SP has an access to the shared memory inside SM and GPU memory
- Thread synchronizaation allowed only inside SM
- Execution organized as a GRID of thread blocks
- Program on GPU is called «kernel».

CUDA Execution Model

- Thread block is execution on SM (Stream Multiprocessor)
  - One thread block can use only one SM
  - Schedule of thread block execution is not defined
- Number of thread blocks processed by SM defines by number of registers and shared memory available
- Active thread block is a block currently executed



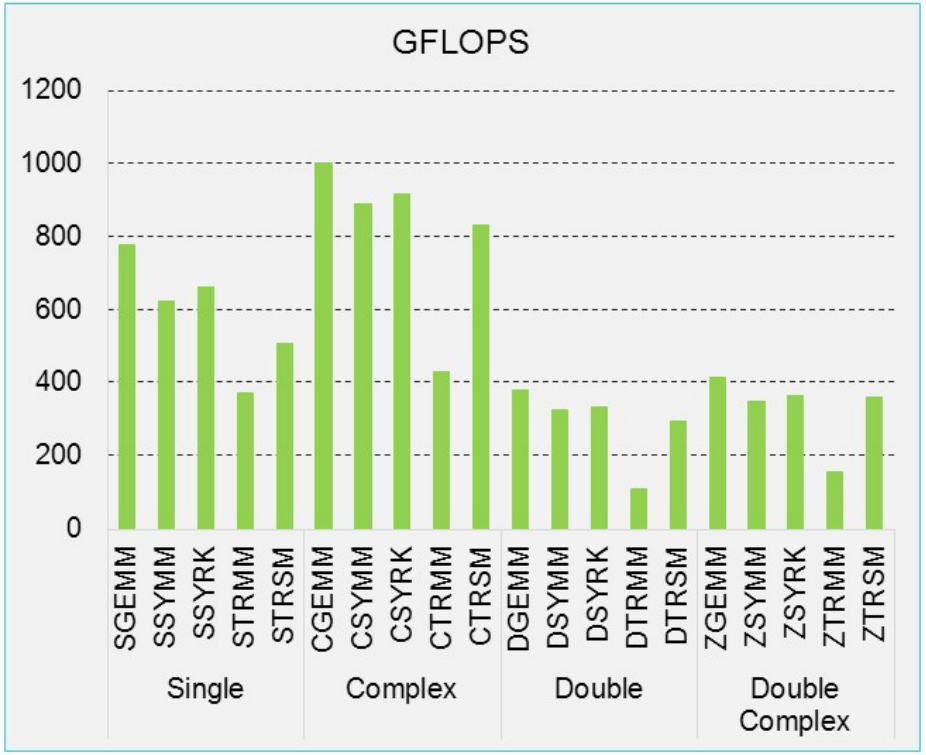
BLAS Functionality

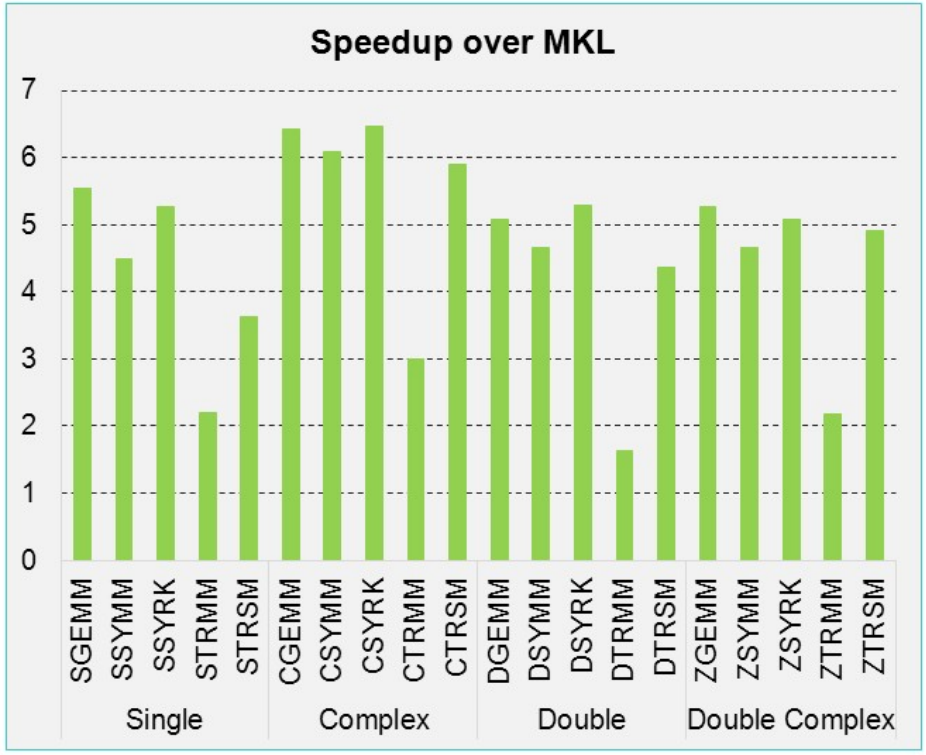
- Level 1: vector operations
  - $y = ax + y$
  - $a = x'y$
- Level 2: vector-matrix operations
  - $y = aAx + by$
  - $A = axy' + A$
- Level 3: matrix operations
  - $C = aAB + C$

BLAS Implementations

- refblas – C/Fortran77, netlib
- ATLAS – C/Fortran77, netlib
- uBLAS – C++, Boost
- **cuBLAS – C, NVIDIA**
- ACML – C/Fortran77, AMD
- MKL – C/Fortran77, Intel

Level-3 cuBLAS Performance

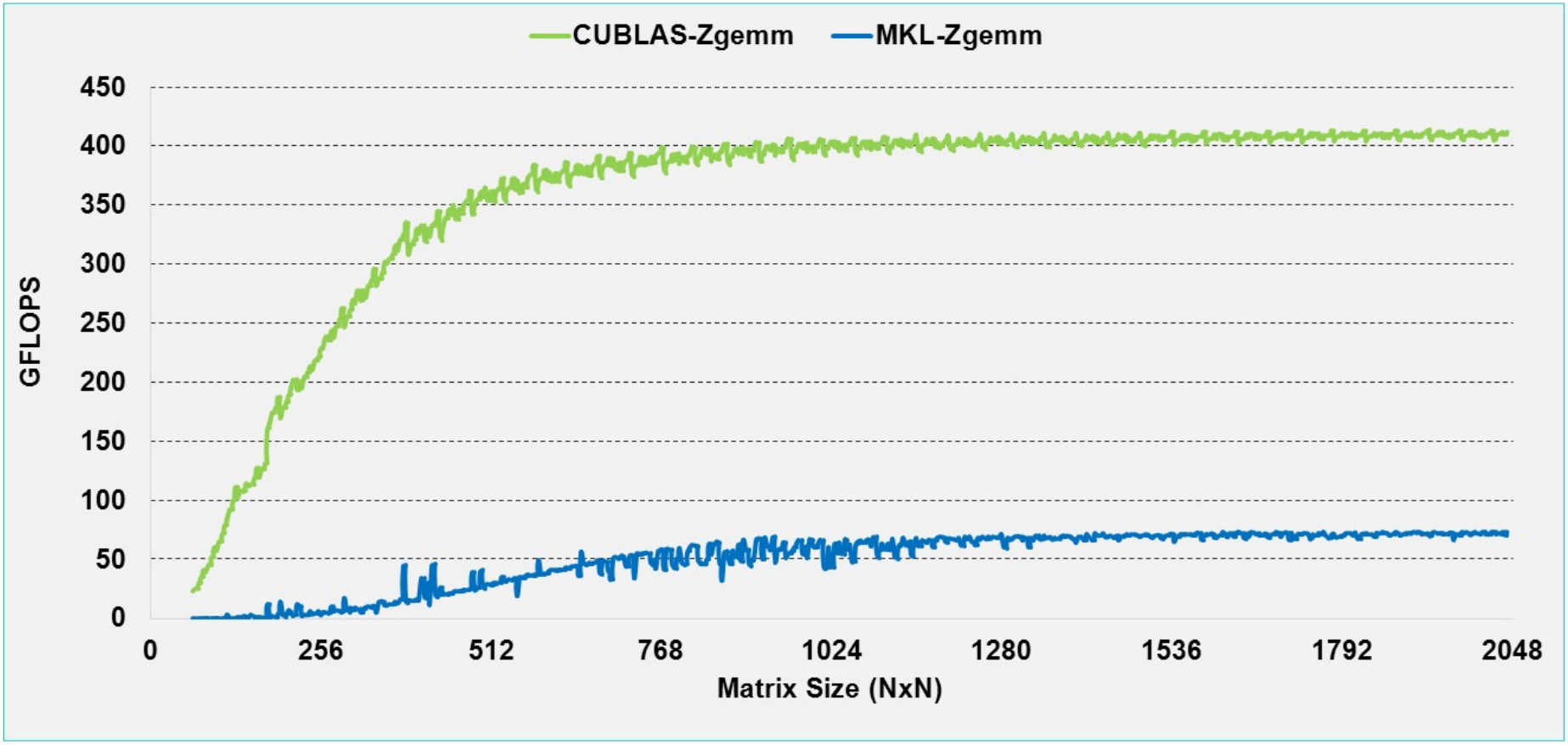




- 4Kx4K matrix size
- cuBLAS 4.1, Tesla M2090 (Fermi), ECC on

- MKL 10.2.3, TYAN FT72-B7015 Xeon x5680 Six-Core @ 3.33 GHz
- Performance may vary based on OS ver. and motherboard config

ZGEMM performance



- cuBLAS 4.1 on Tesla M2090, ECC on
- MKL 10.2.3, TYAN FT72-B7015 Xeon x5680 Six-Core @ 3.33 GHz

- Performance may vary based on OS ver. and motherboard config.

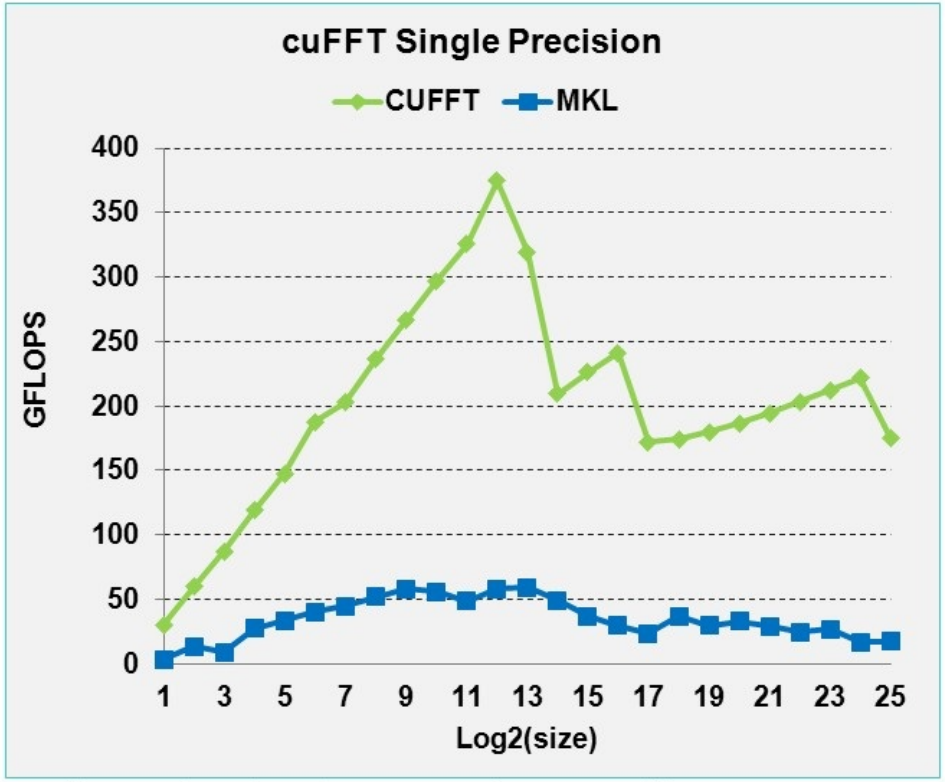
FFT

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \qquad k = 0, \ldots, N-1$$

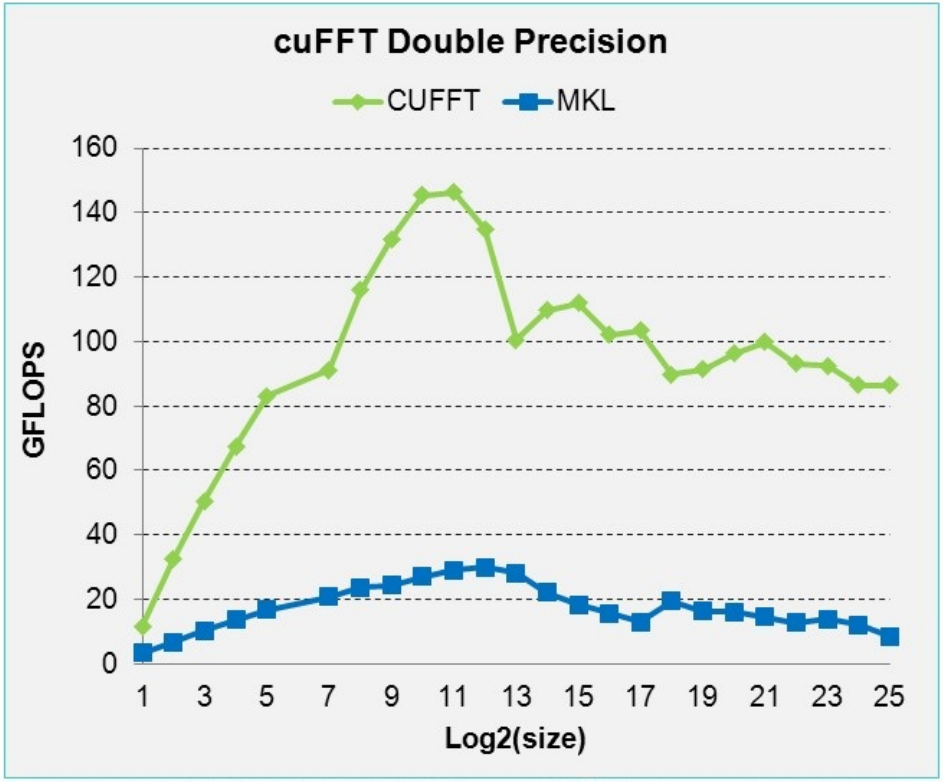
$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{\frac{2\pi i}{N} kn} \qquad n = 0, \ldots, N-1.$$

- Discrete Fast Fourie Transform
  - Real/complex data
  - 1D, 2D, 3D

cuFFT Performance

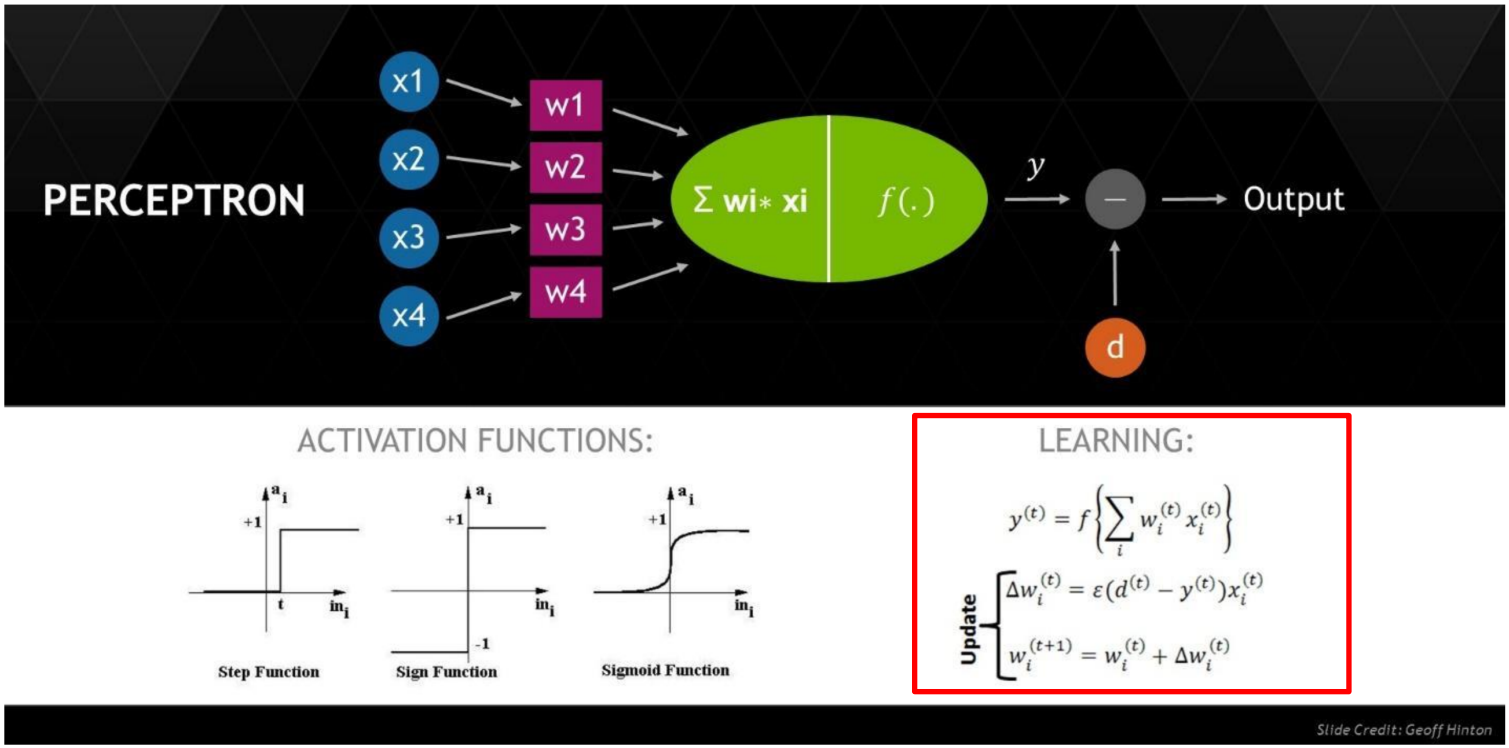


- Measured on sizes that are exactly powers-of-2
- cuFFT 4.1 on Tesla M2090, ECC on
- MKL 10.2.3, TYAN FT72-B7015 Xeon x5680 Six-Core @ 3.33 GHz



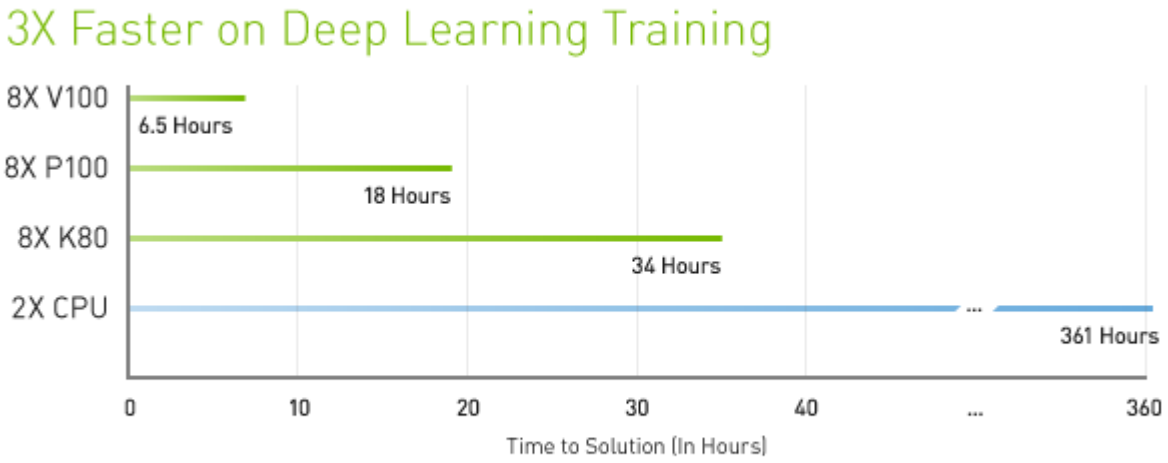
- MKL 10.2.3, TYAN FT72-B7015 Xeon x5680 Six-Core @ 3.33 GHz
- Performance may vary based on OS version and motherboard configuration

Machine Learning (1)

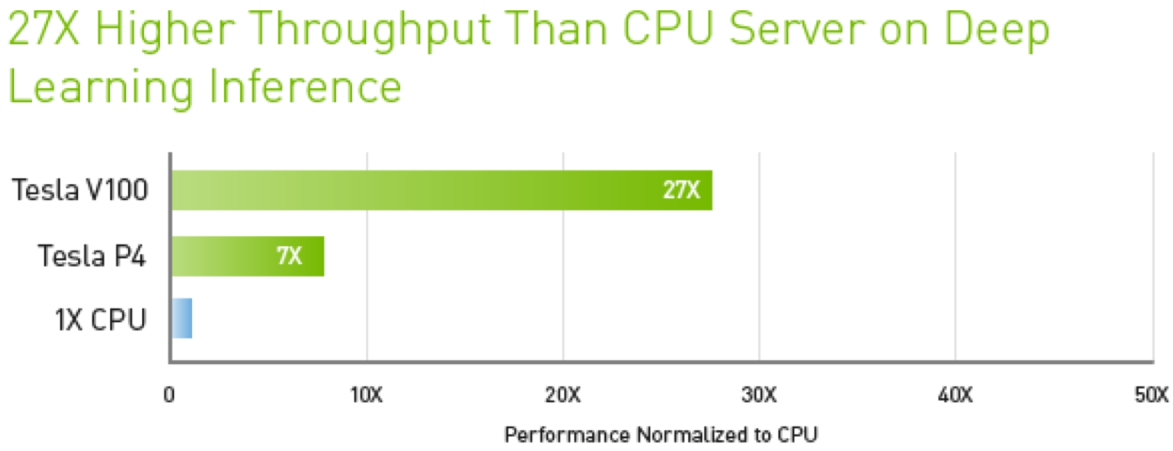


To compute the perceptron, we need to compute many dot-products

Machine Learning (2)



CPU Server: Dual Xeon E5-2699 v4, 2.6GHz | GPU Servers add 8X Tesla K80, Tesla P100 or Tesla V100 | V100 measured on pre-production hardware | Workload: NMT, 13 epochs to solution.



Workload: ResNet-50 | CPU: 1X Xeon E5-2690v4 @ 2.6 GHz | GPU: Add 1X Tesla P4 or V100