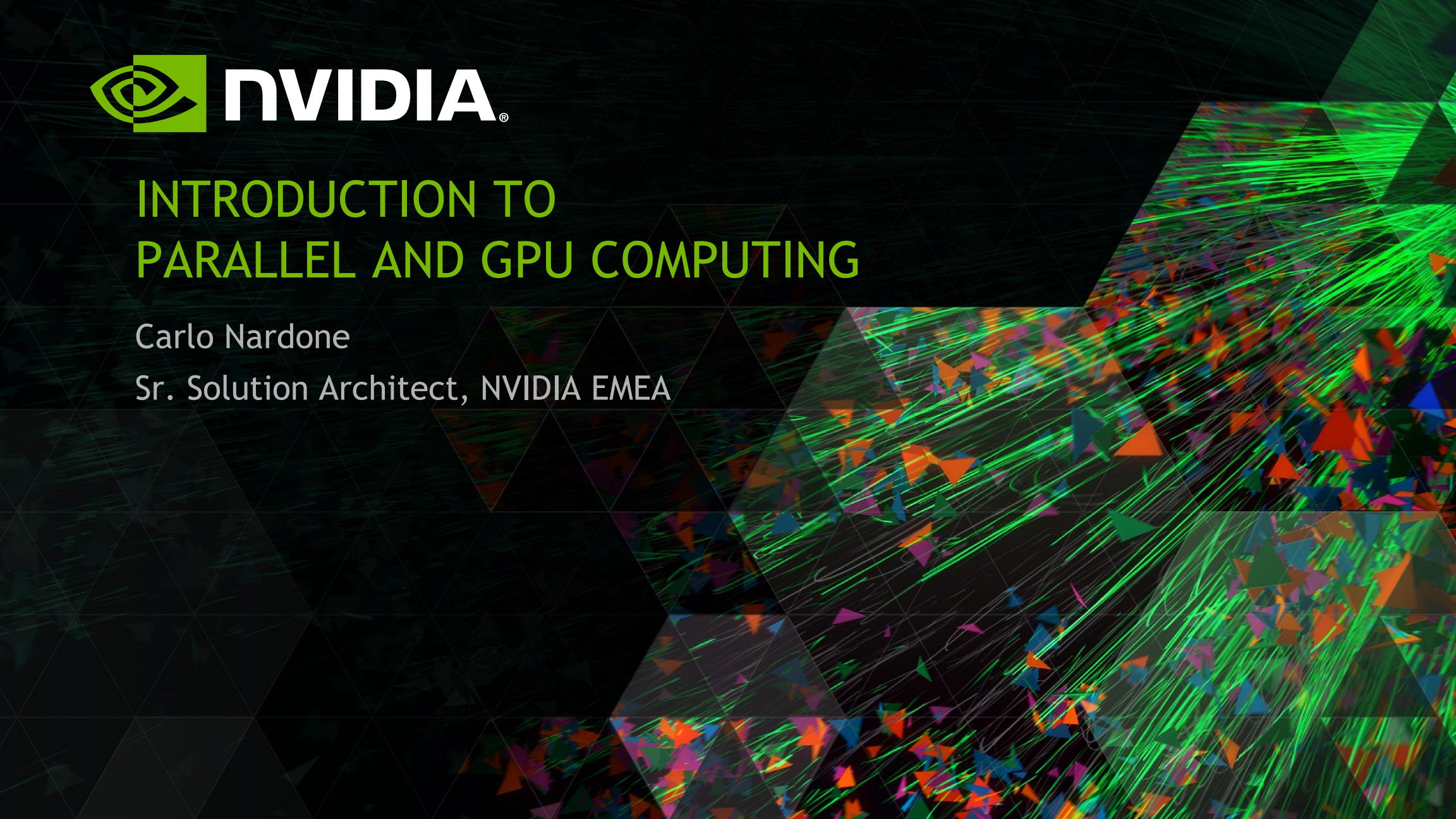


INTRODUCTION TO PARALLEL AND GPU COMPUTING

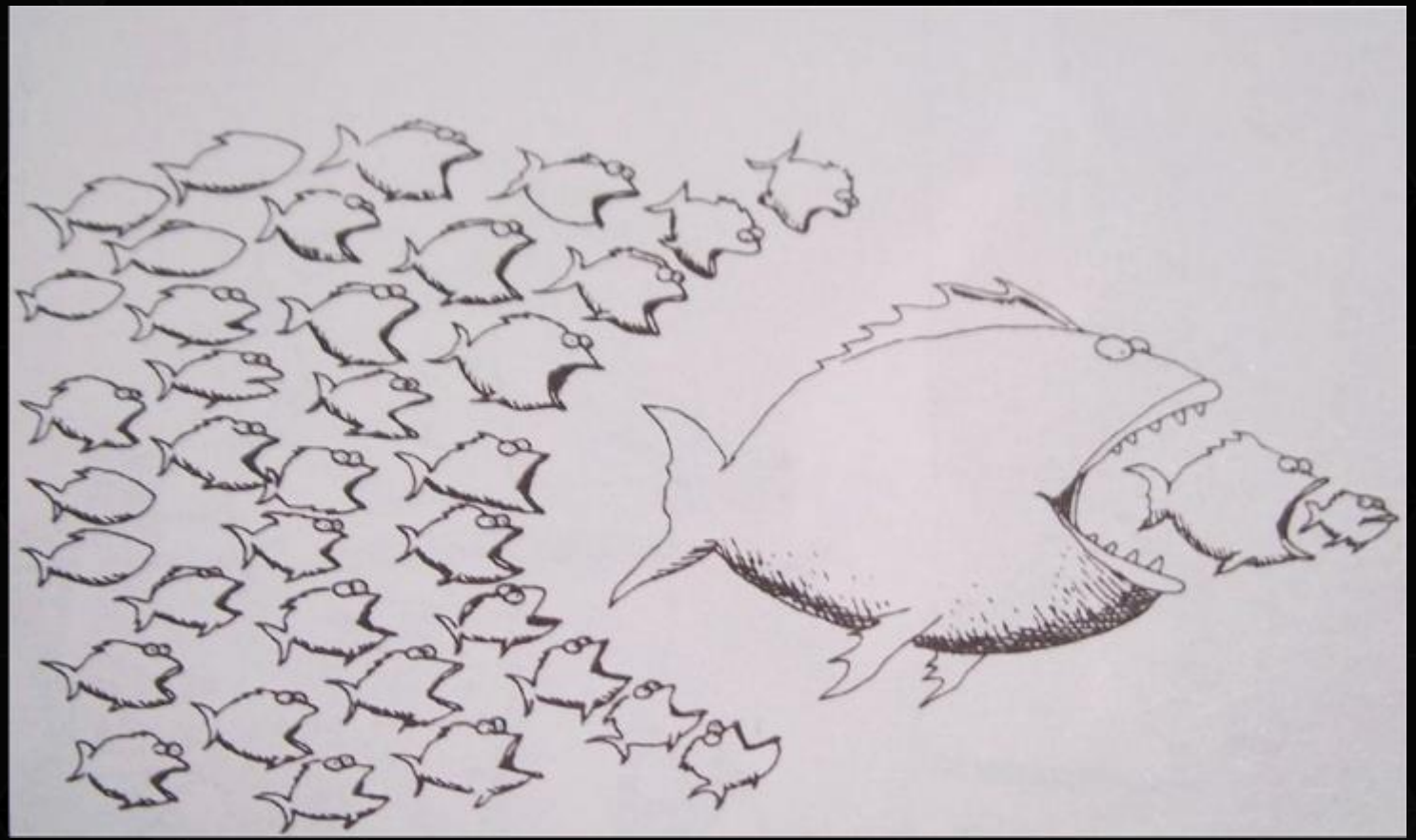
Carlo Nardone

Sr. Solution Architect, NVIDIA EMEA

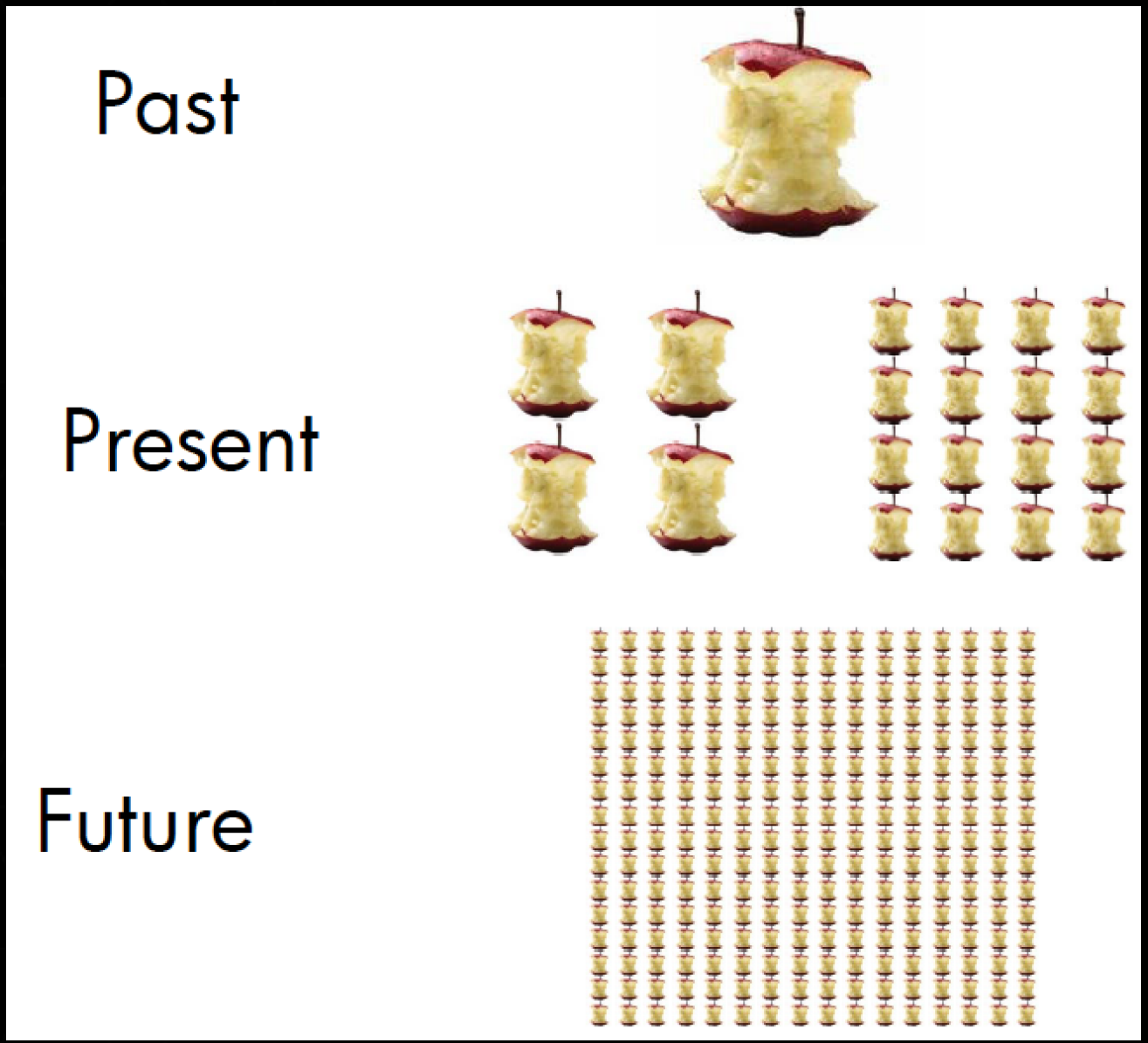
AGENDA

1. Intro
2. Processors Trends
3. Multiprocessors
4. GPU Computing

PARALLEL COMPUTING, ILLUSTRATED

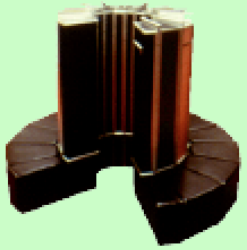


FROM MULTICORE TO MANYCORE



PARALLEL COMPUTING IS OLD

Vector Computers

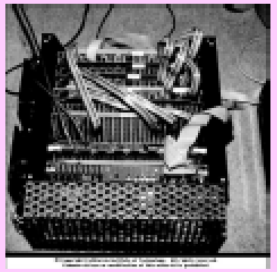


Cray 1 (1976)

Cray 2 (1985)

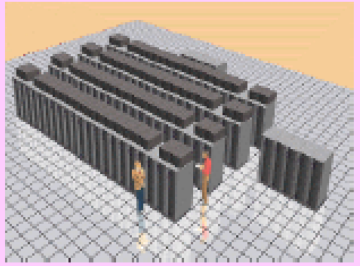
Cray C-90 (1991)

Massively Parallel Processors (MPP)



Cosmic cube (1983)

Paragon (1993)



ASCI Red (1997)

Cluster Computers

Clusters (late 80's)

Linux PC Clusters
(~1995)

Source: Tim Mattson

Late 70's

Late 80's

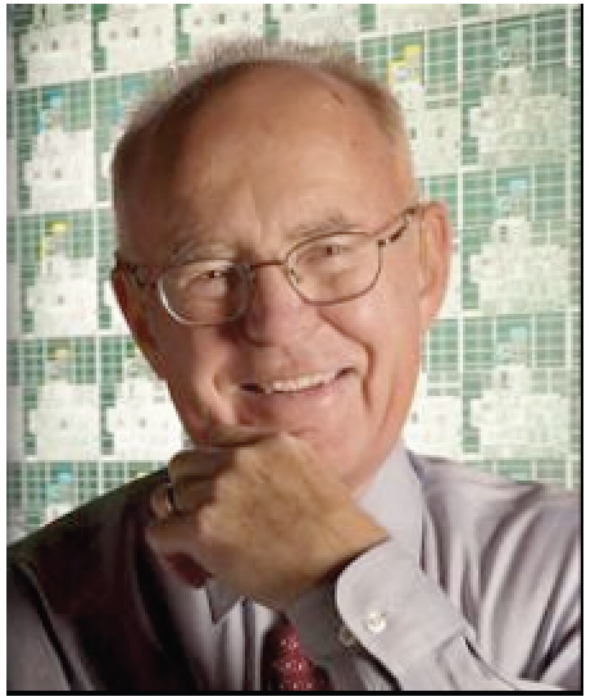
Late 90's

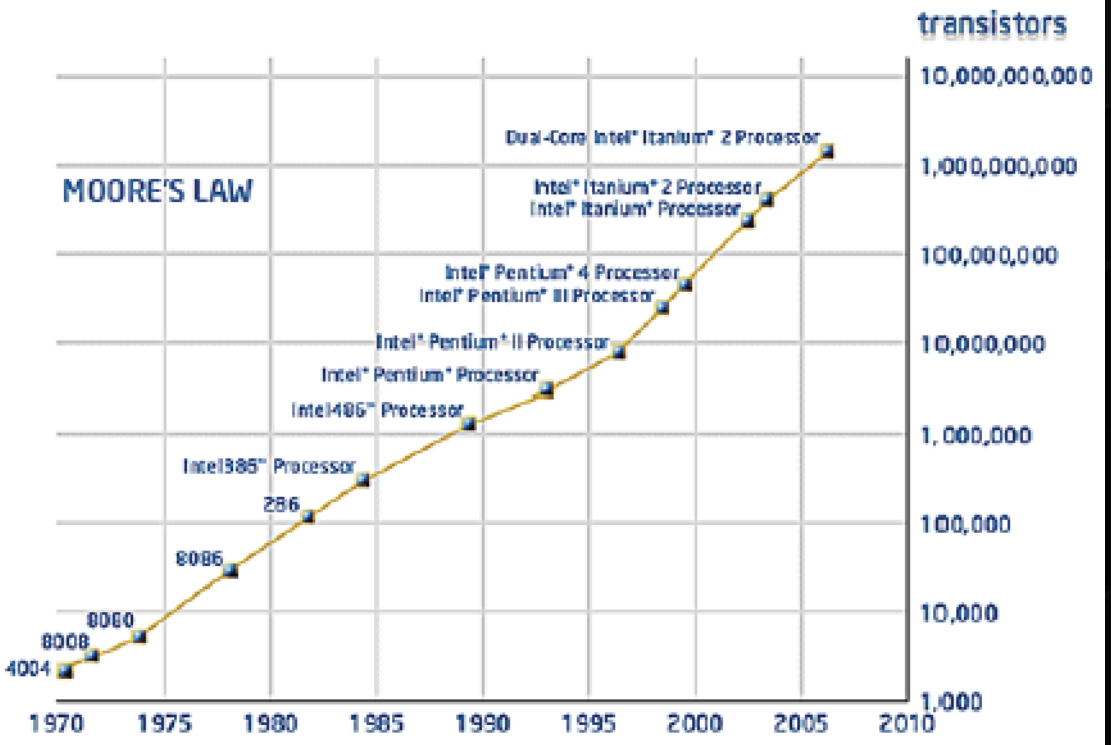
15

AGENDA

1 Intro
2 Processors Trends
3 Multiprocessors
4 GPU Computing

MOORE'S LAW



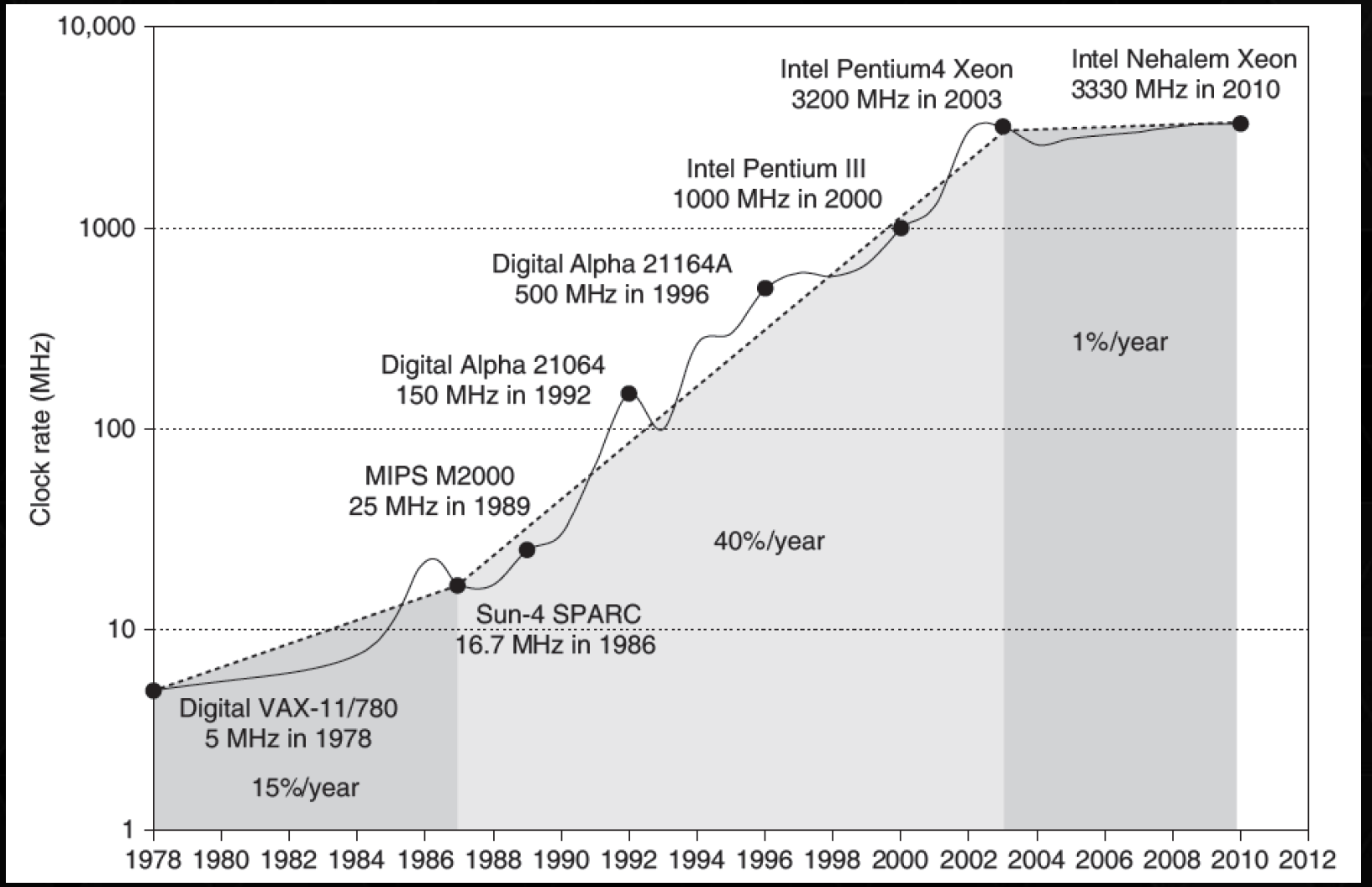


Gordon Moore (Intel co-founder) predicted in 1965 that transistor density of semiconductor chips would double every 18 months.

Moore's Law

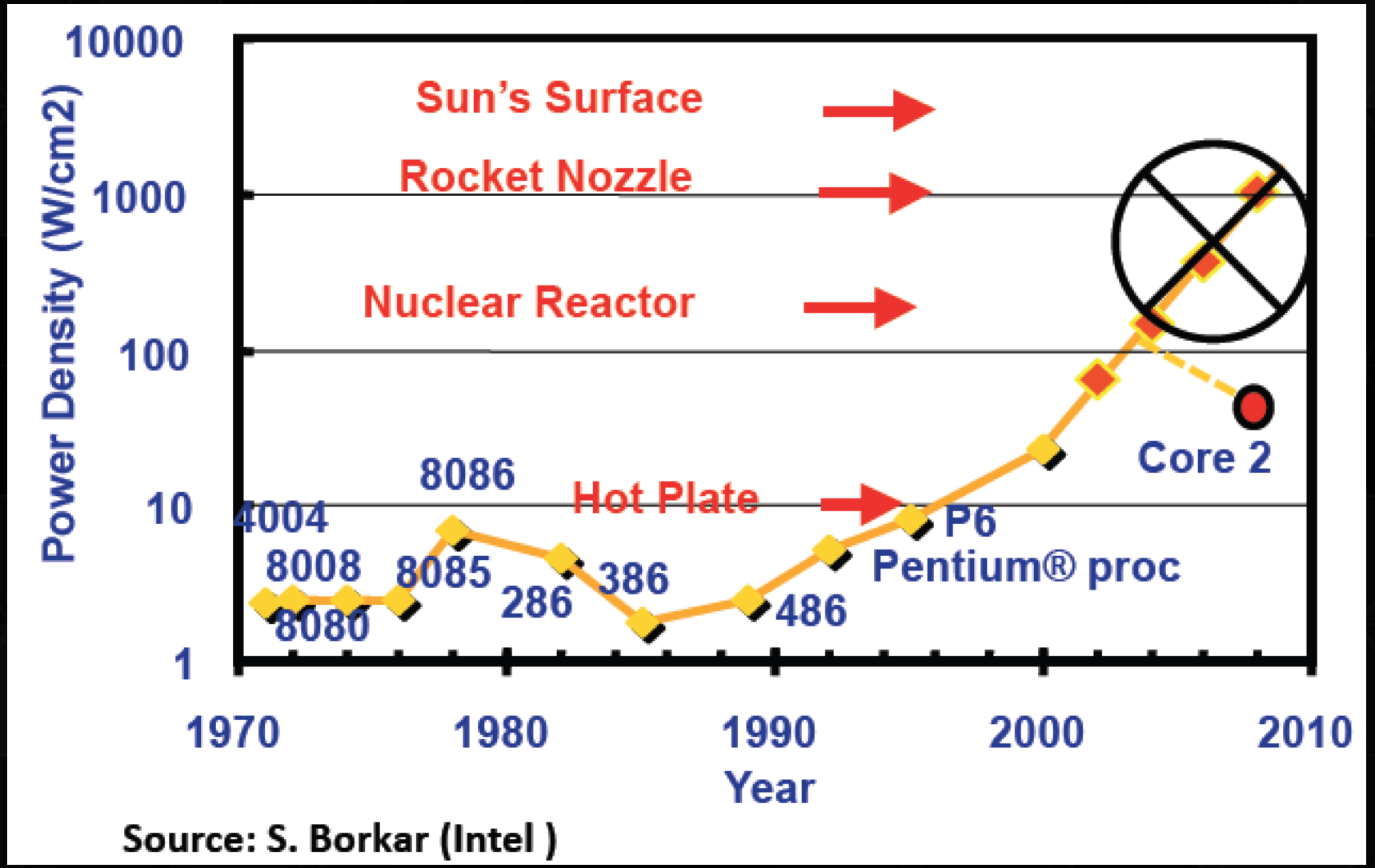
Source: U.Delaware CISC879 / J.Dongarra

MICROPROCESSOR FREQUENCY TREND



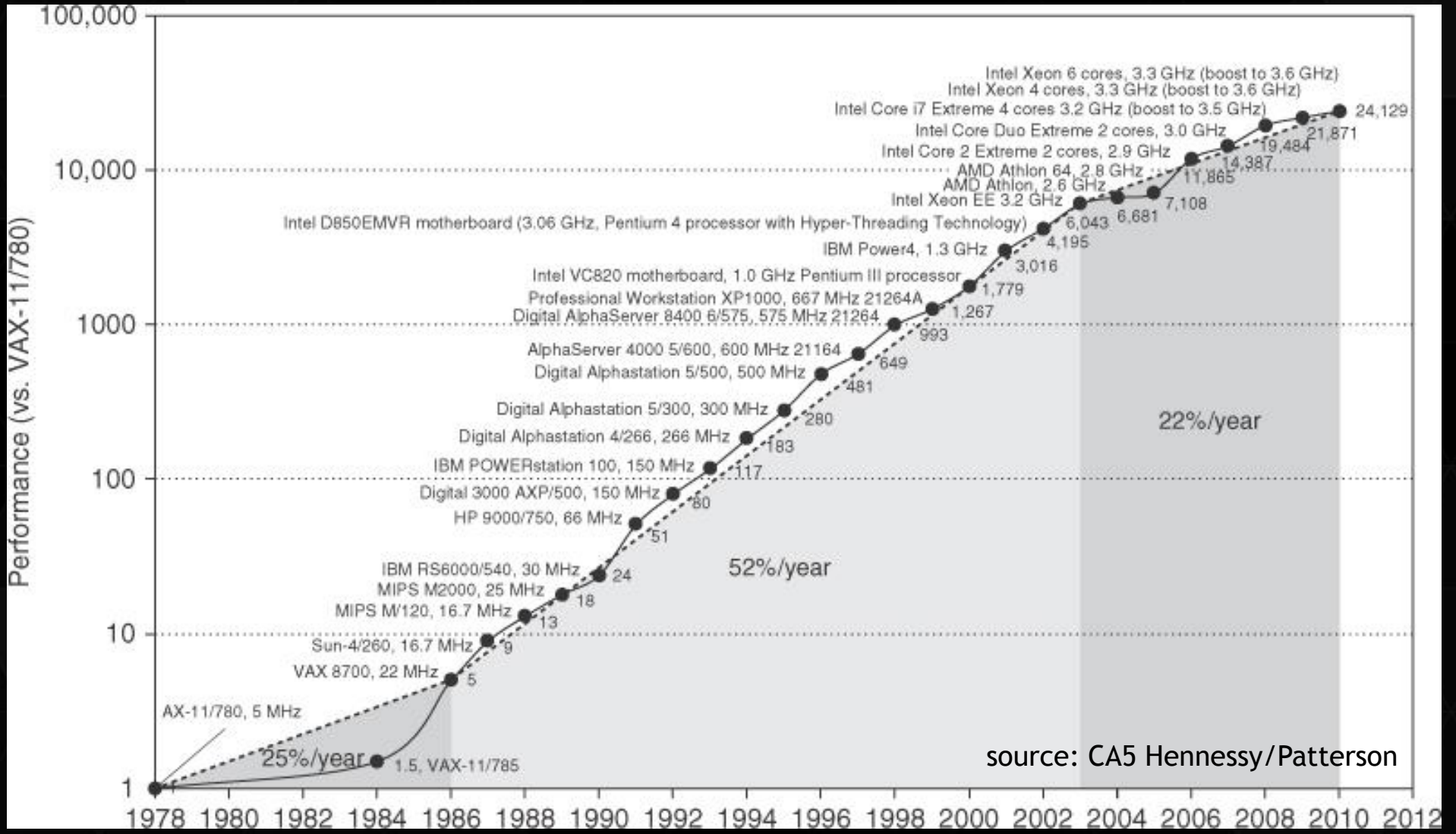
source: CA5
Hennessy/Patterson

THE POWER WALL



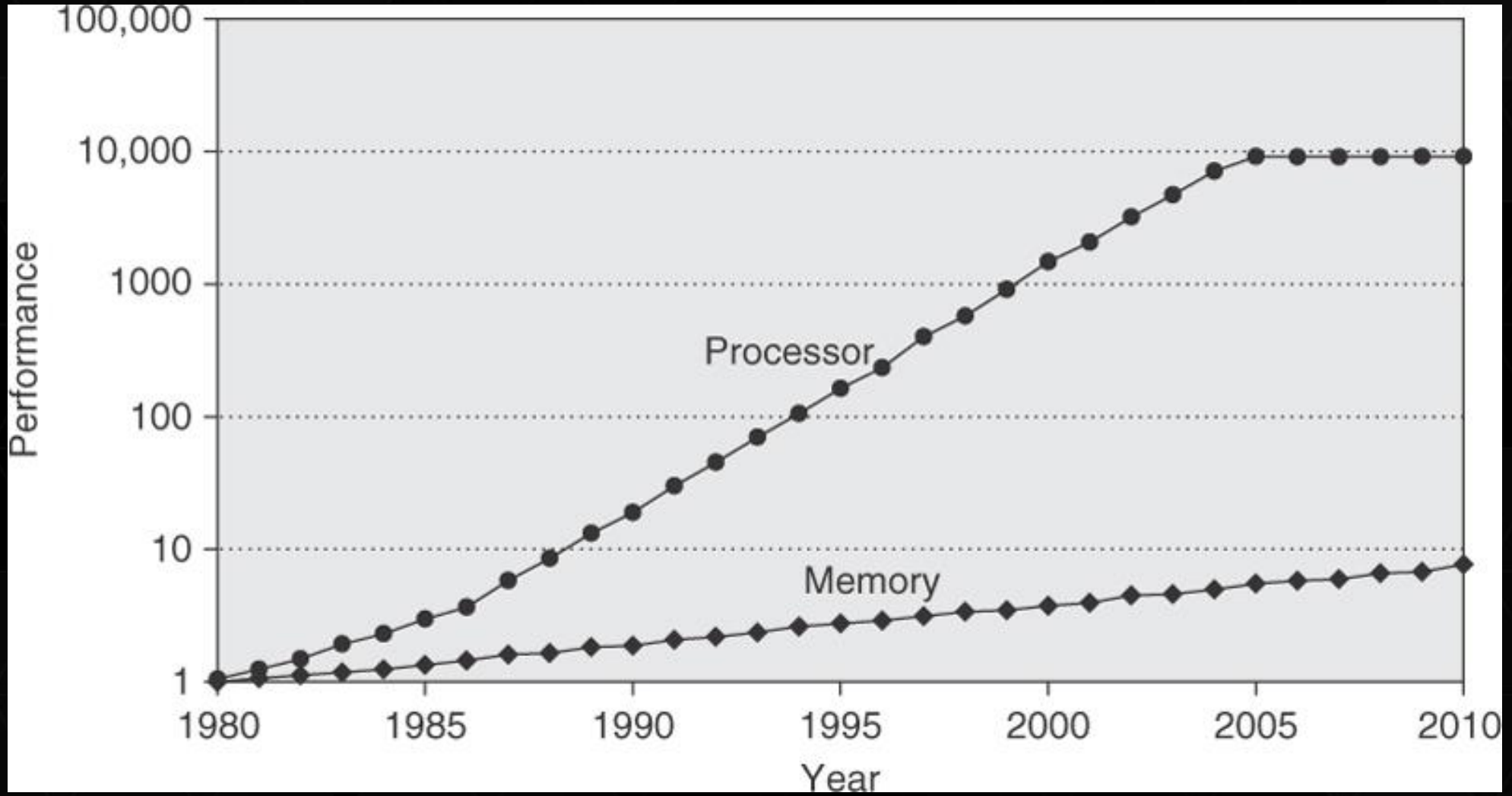
Source: S. Borkar (Intel)

MICROPROCESSORS PERFORMANCE TREND



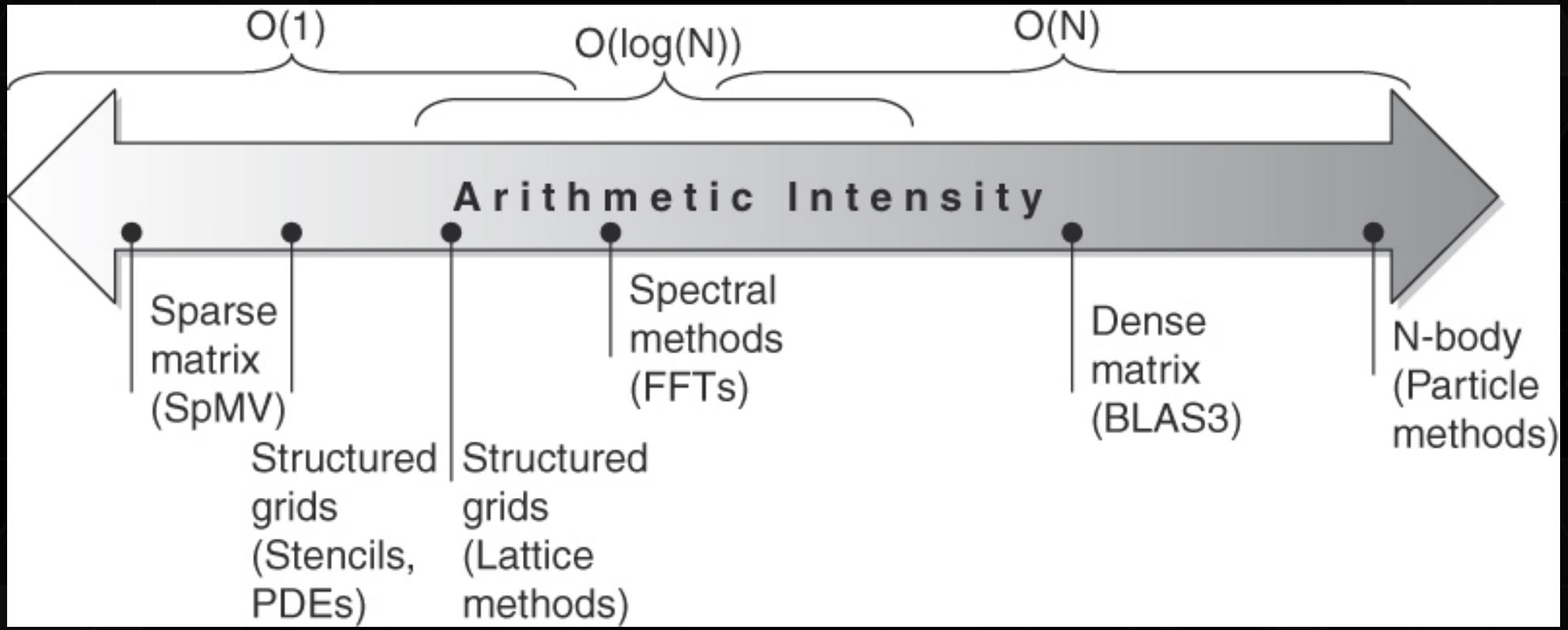
source: CA5 Hennessy/Patterson

THE PROCESSOR-MEMORY GAP



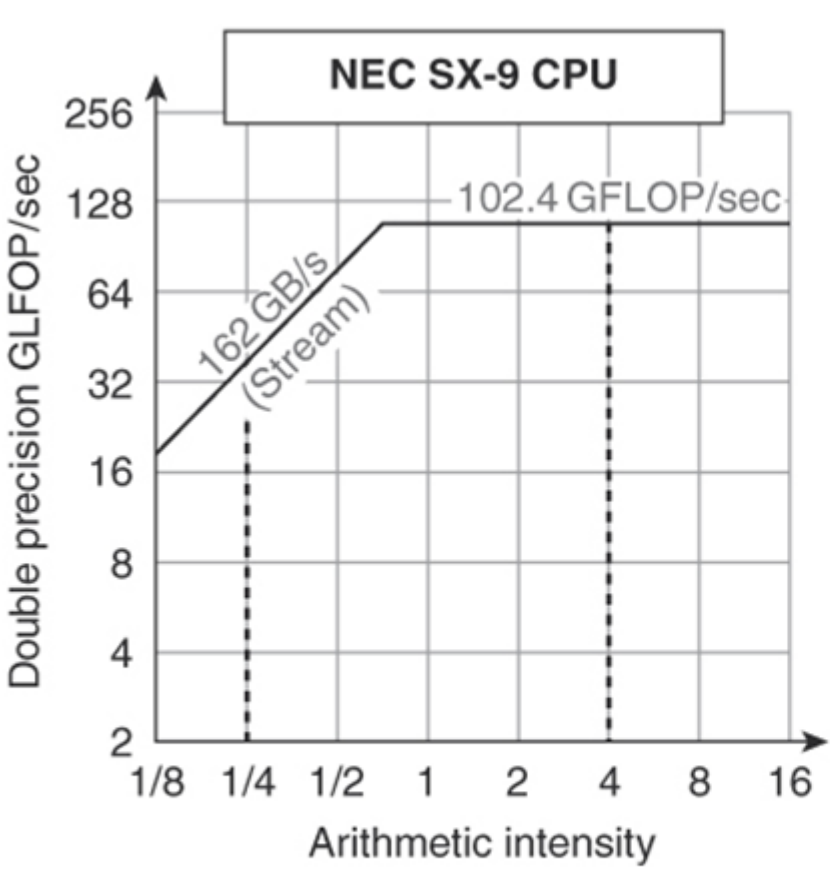
source: CA5 Hennessy & Patterson

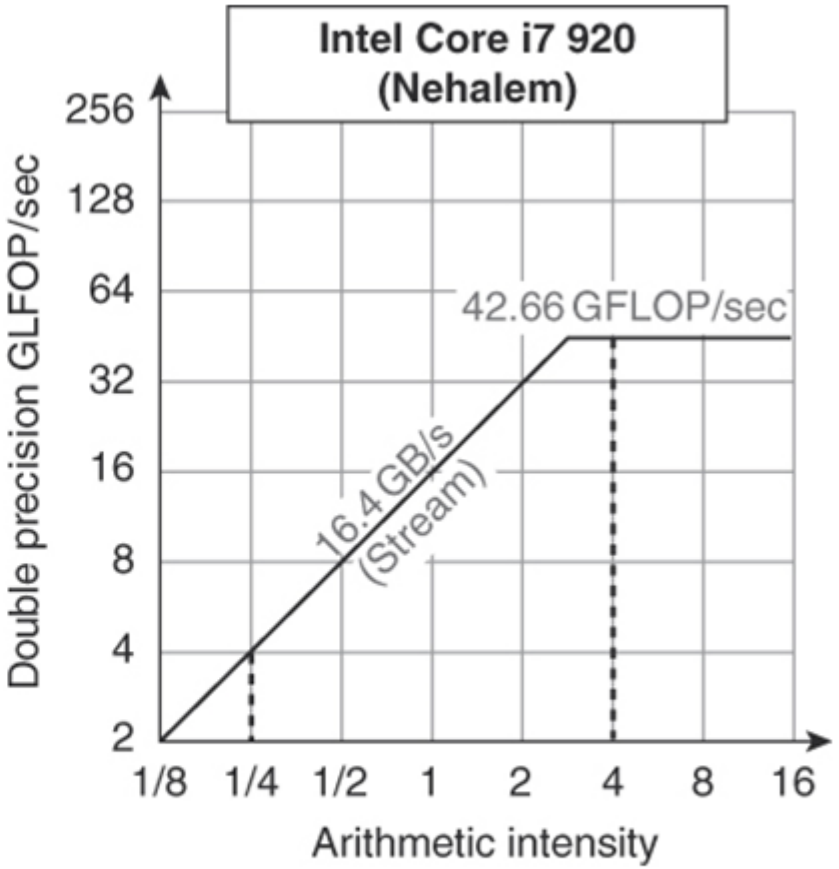
ARITHMETIC INTENSITY



Arithmetic intensity, specified as the number of floating-point operations to run the program divided by the number of bytes accessed in main memory [Williams et al. 2009]. Some kernels have an arithmetic intensity that scales with problem size, such as dense matrix, but there are many kernels with arithmetic intensities independent of problem size. Source: CA5 Hennessy & Patterson

BANDWIDTH AND THE ROOFLINE MODEL





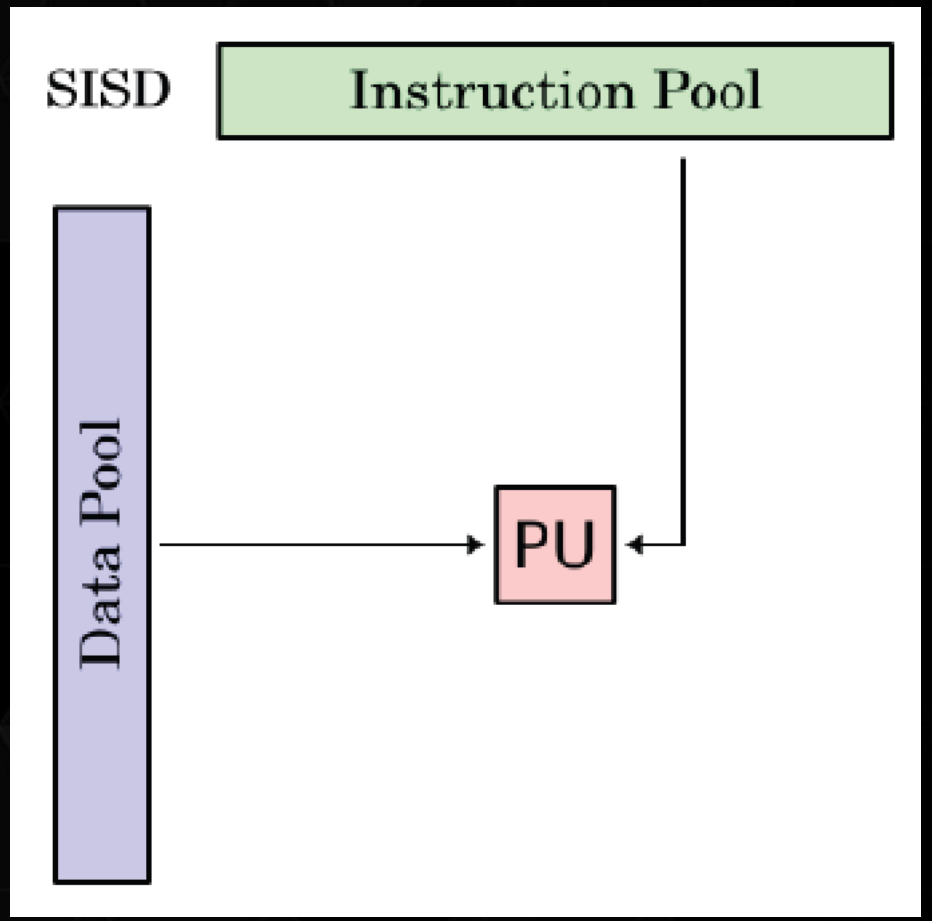
Source: HPCA5

Figure 4.11 Roofline model for one NEC SX-9 vector processor on the left and the Intel Core i7 920 multicore computer with SIMD Extensions on the right [Williams et al. 2009]. This Roofline is for unit-stride memory accesses and double-precision floating-point performance. .

AGENDA

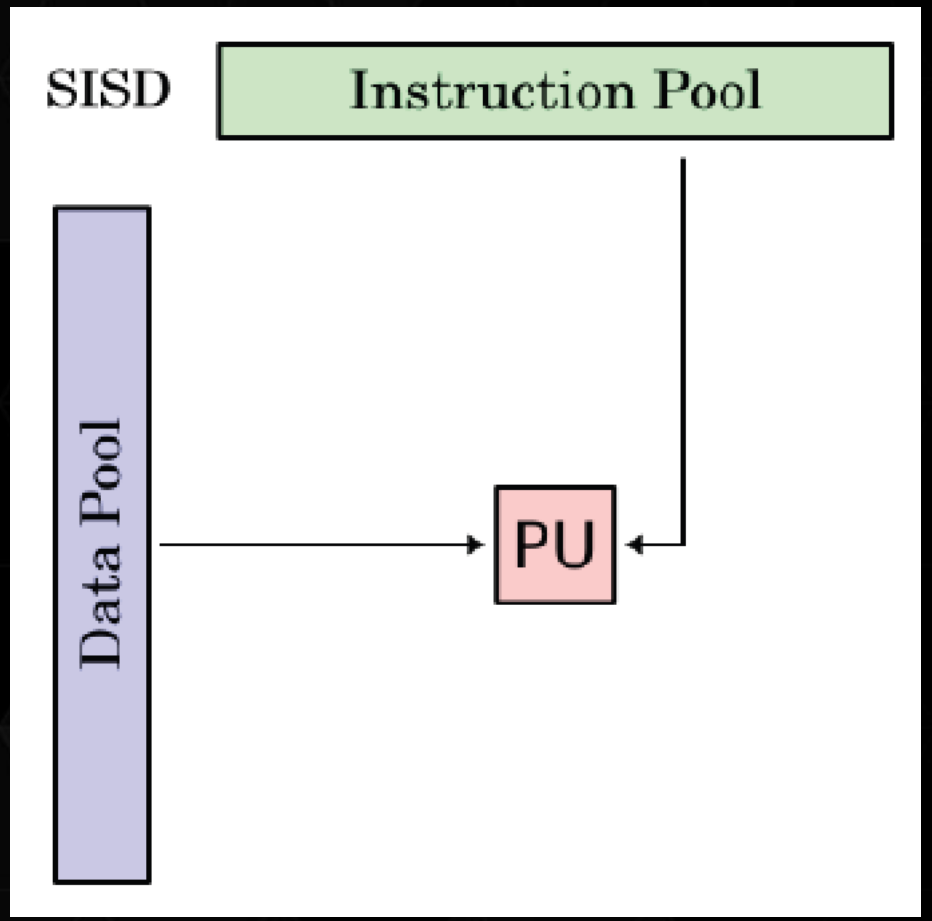
1 Intro
2 Processors Trends
3 Multiprocessors
4 GPU Computing

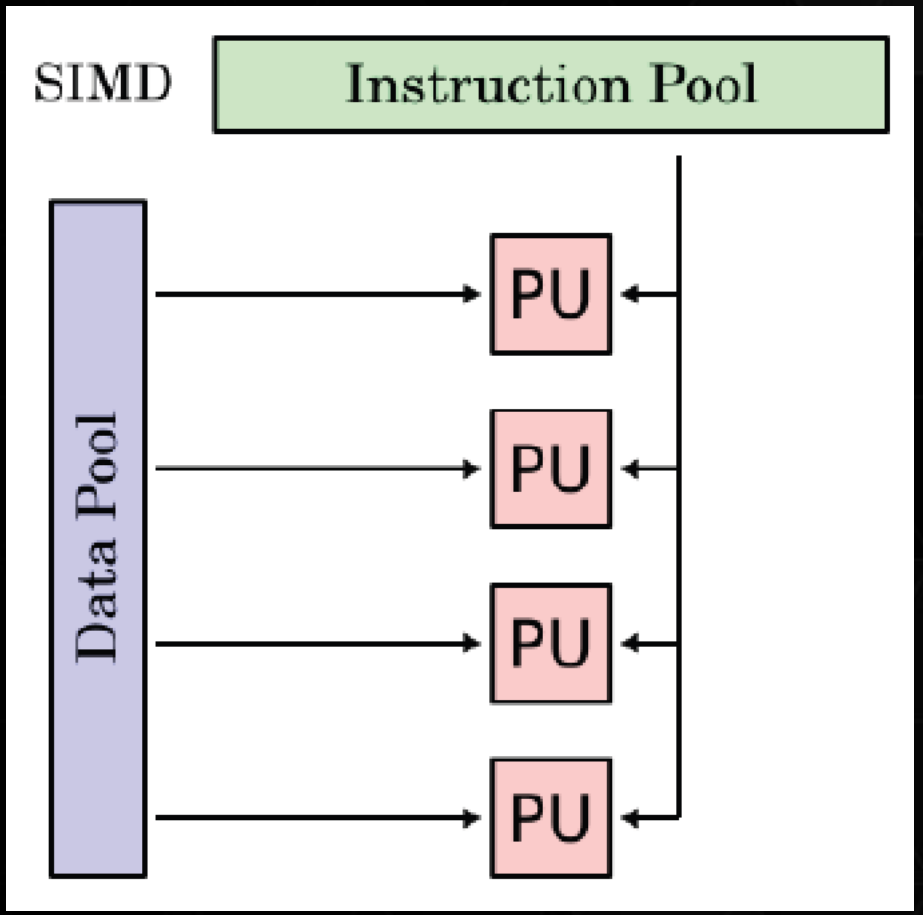
FLYNN'S TAXONOMY: SISD



source: CA5 Hennessy & Patterson

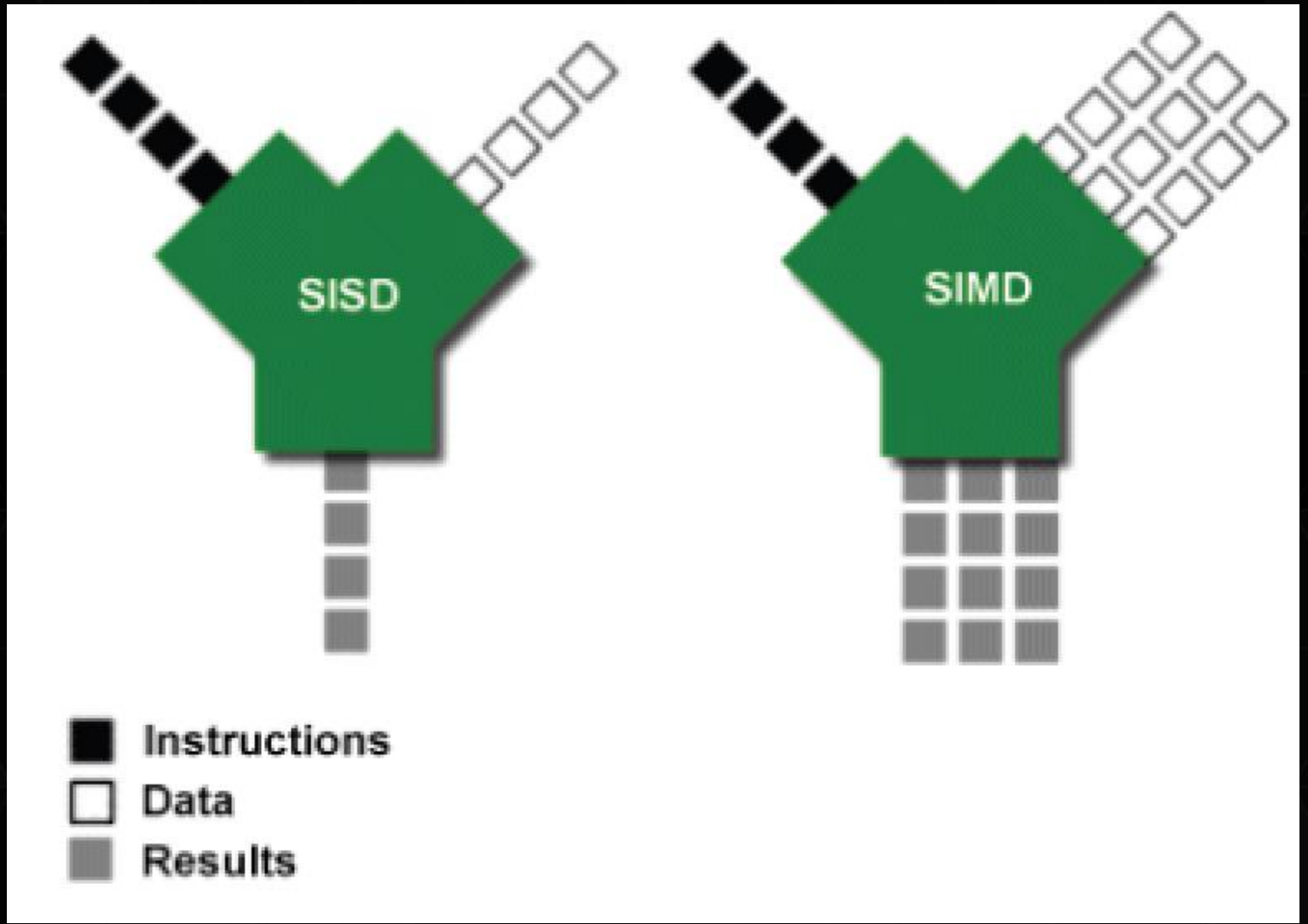
FLYNN'S TAXONOMY: SIMD





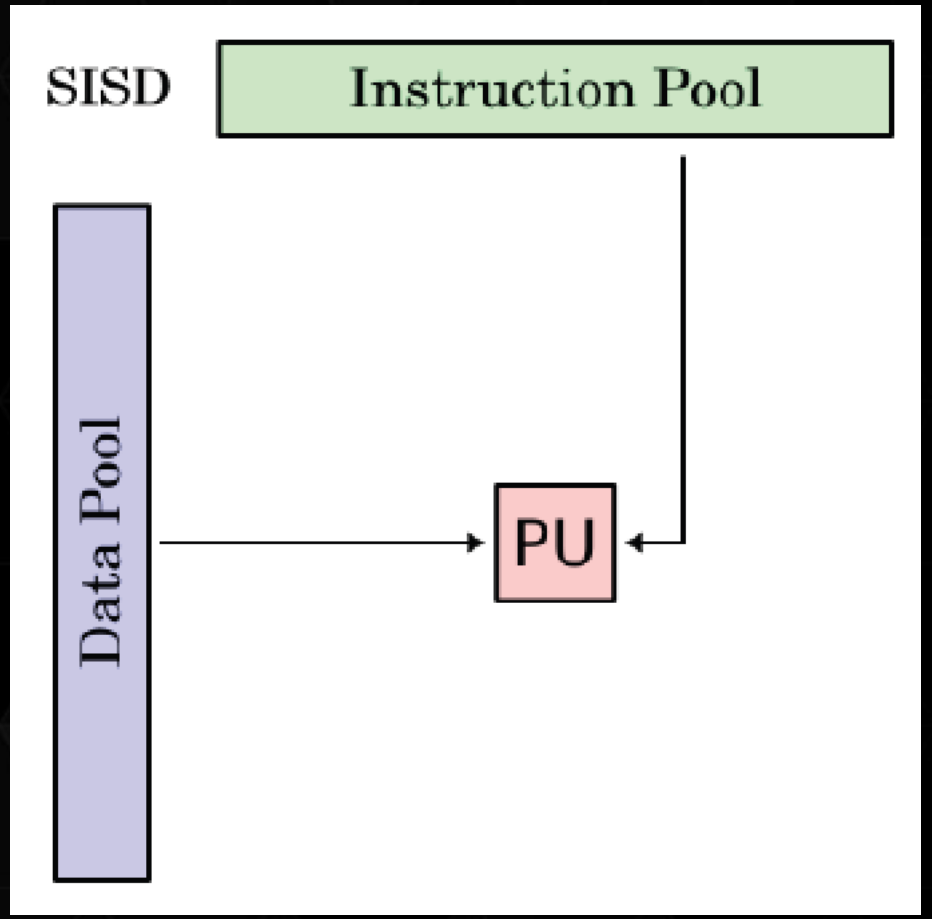
source: CA5 Hennessy & Patterson

FLYNN'S TAXONOMY: SIMD (2)



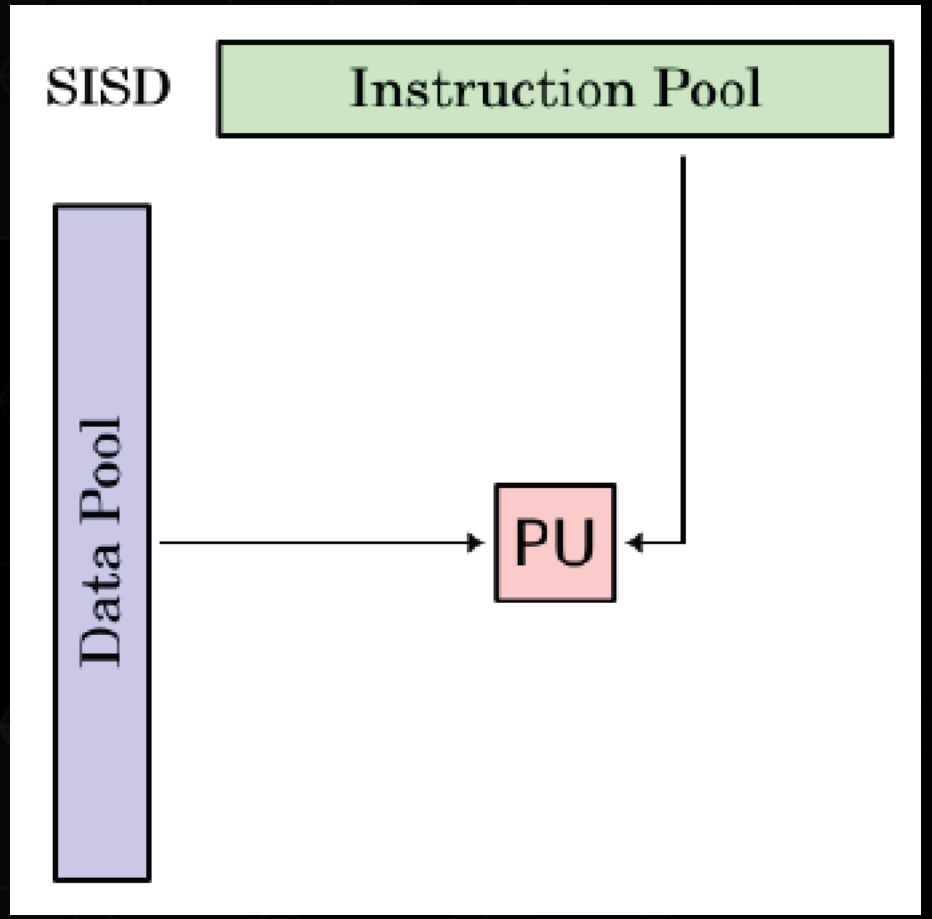
source: CA5 Hennessy & Patterson

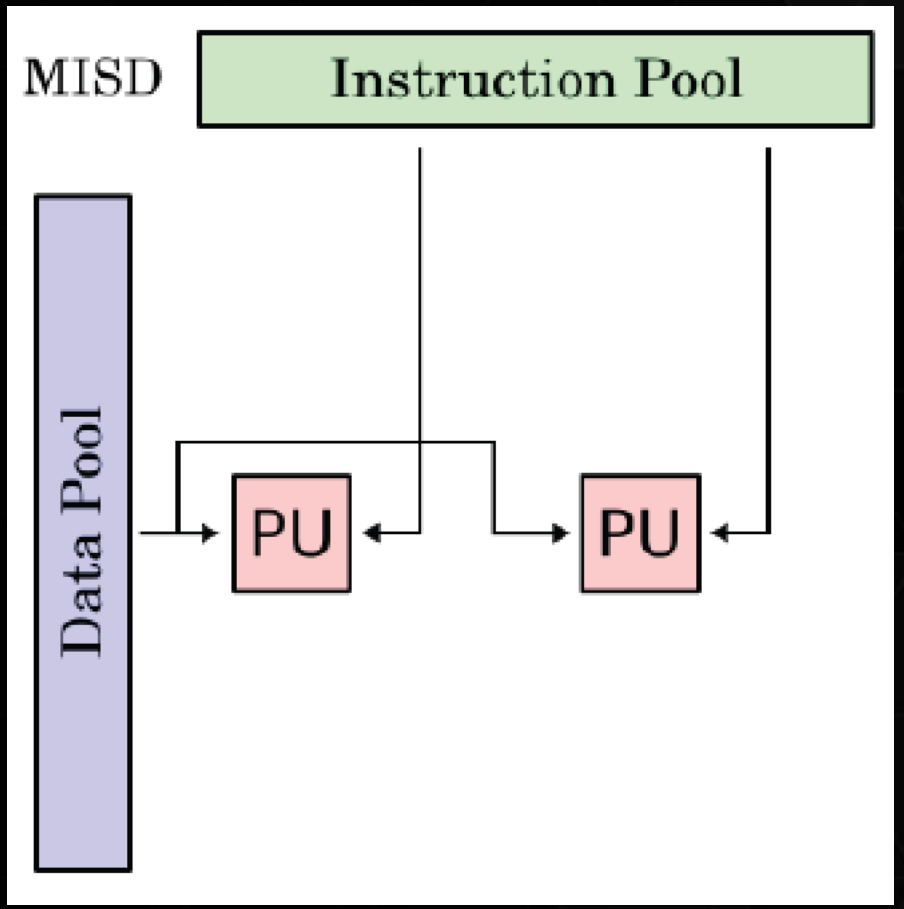
FLYNN'S TAXONOMY



source: CA5 Hennessy & Patterson

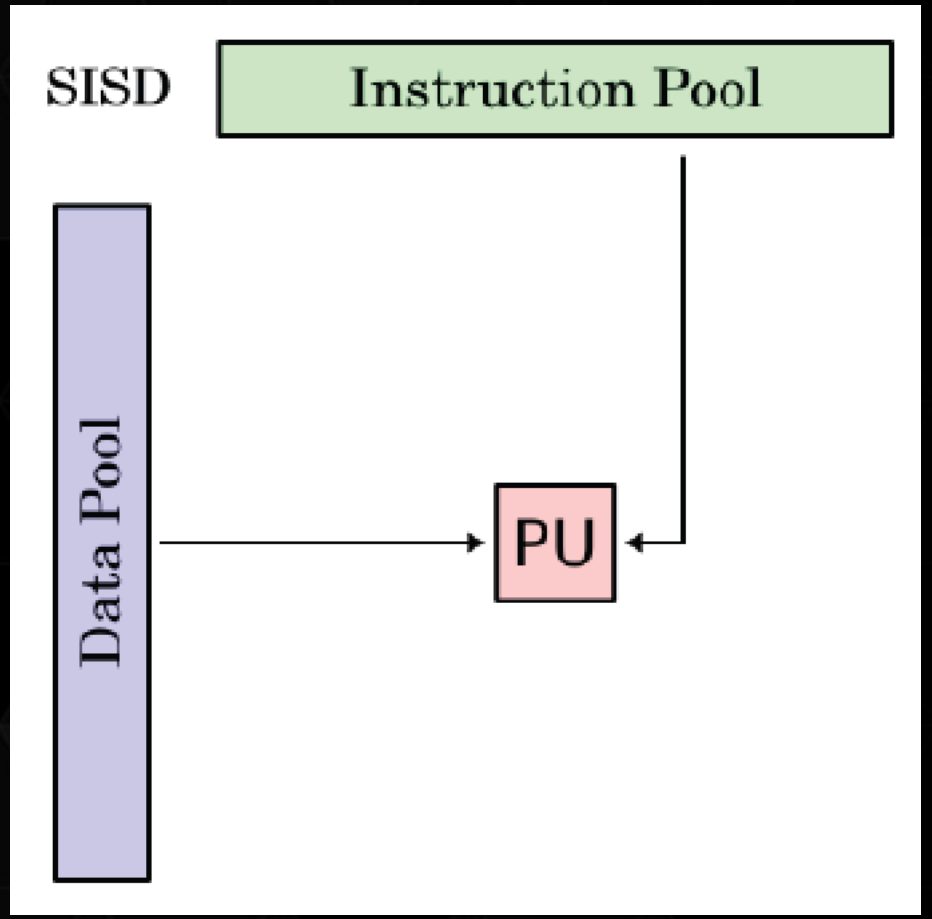
FLYNN'S TAXONOMY: MISD (?)

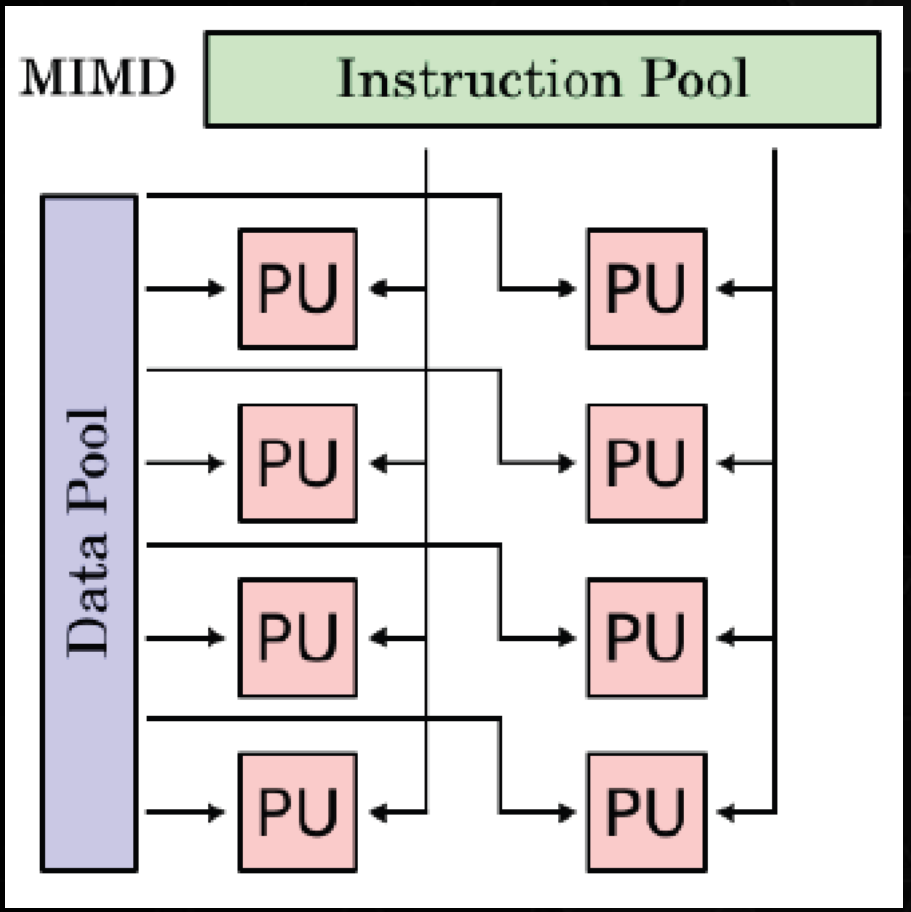




source: CA5 Hennessy & Patterson

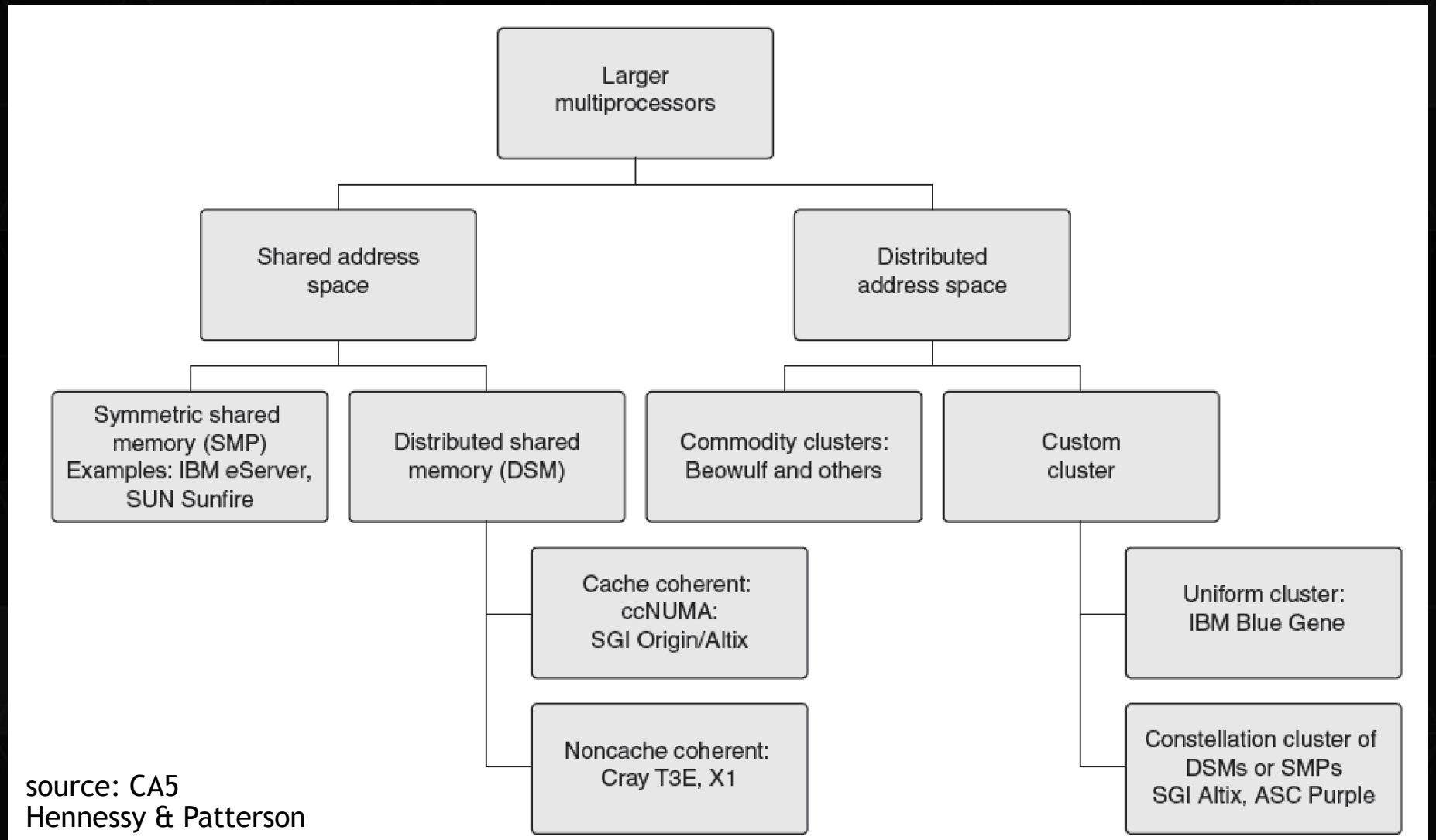
FLYNN'S TAXONOMY: MIMD





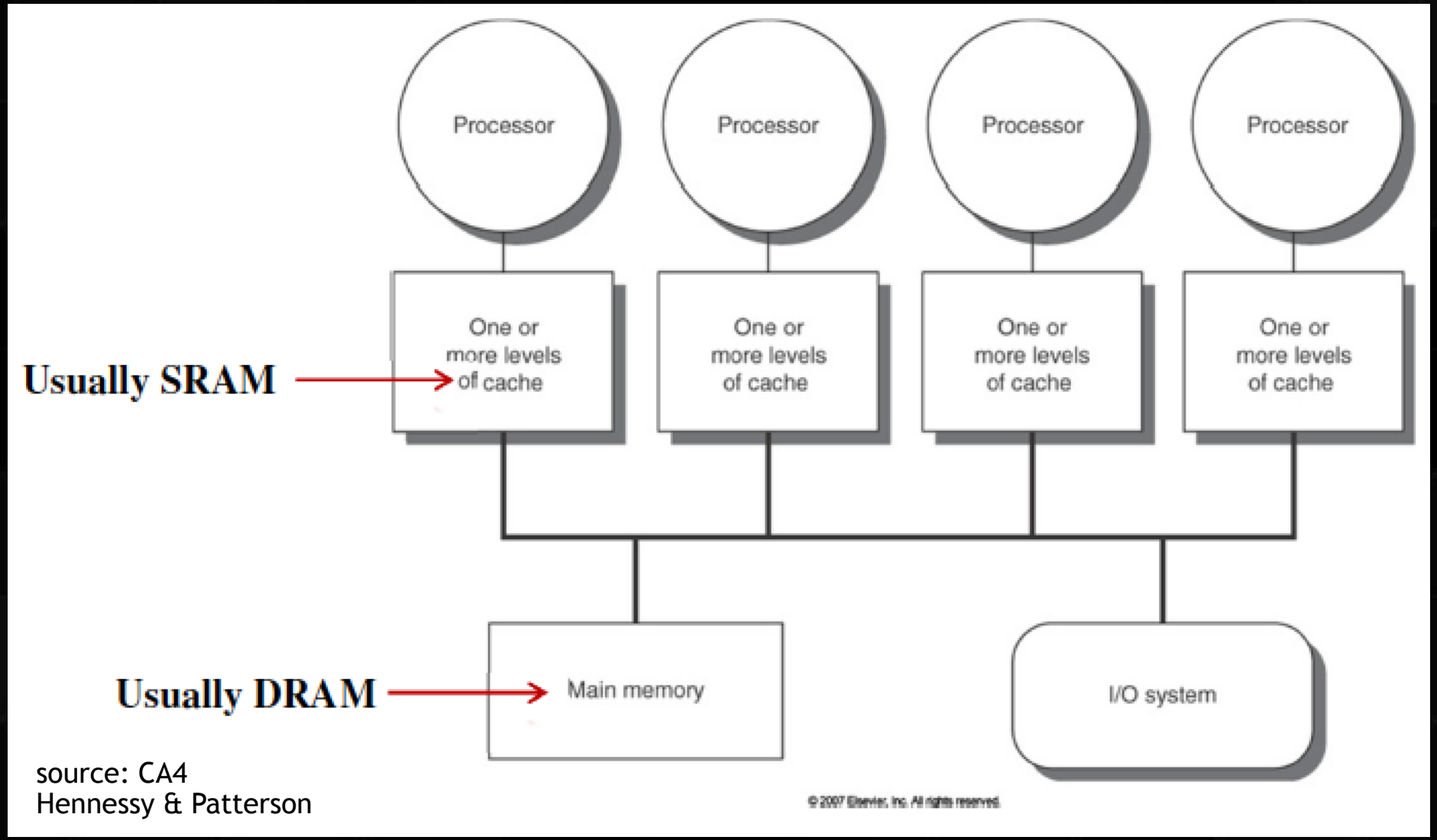
source: CA5 Hennessy & Patterson

MULTIPROCESSORS



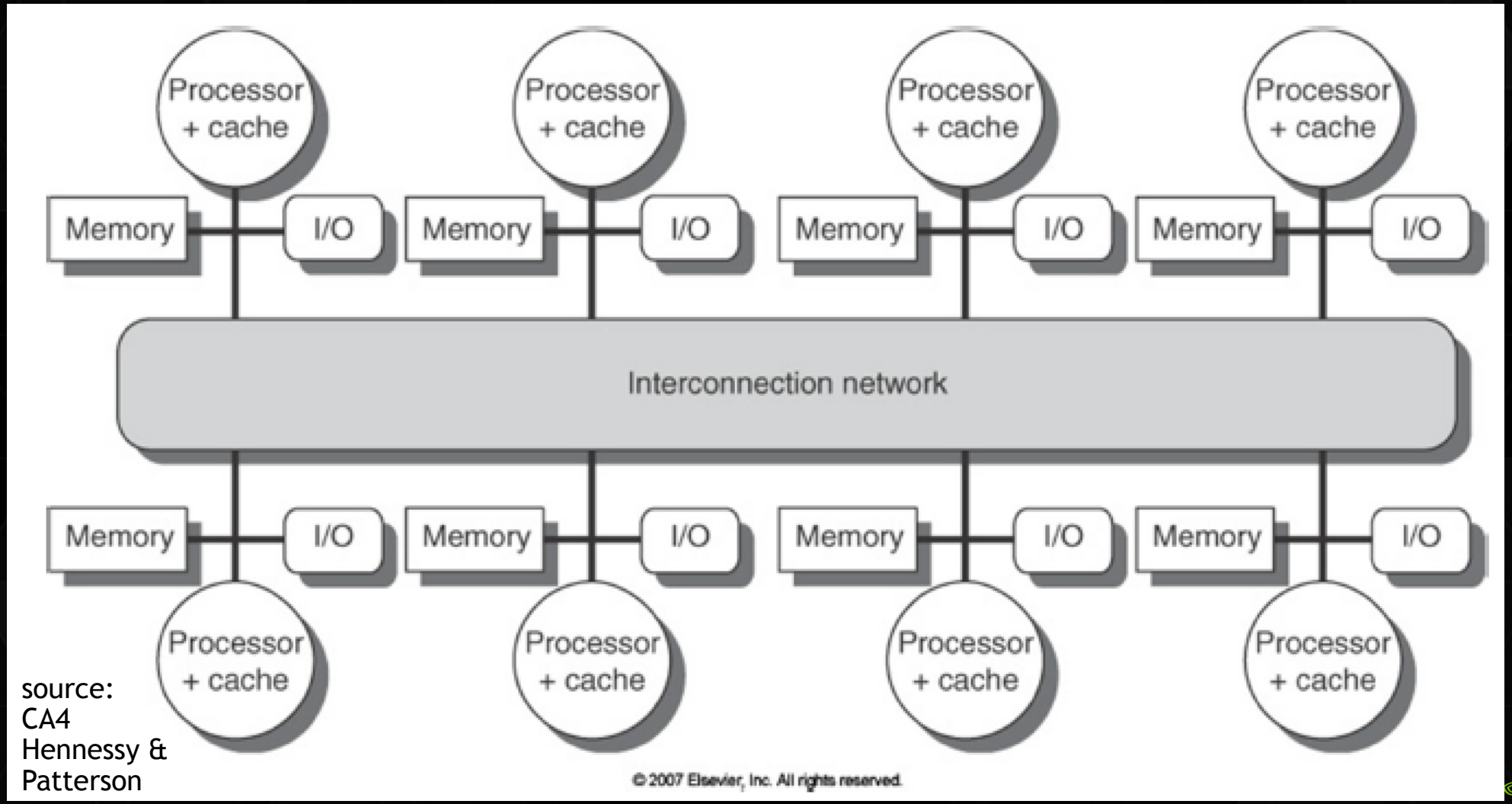
source: CA5
Hennessy & Patterson

MULTIPROCESSOR SMP CLASS



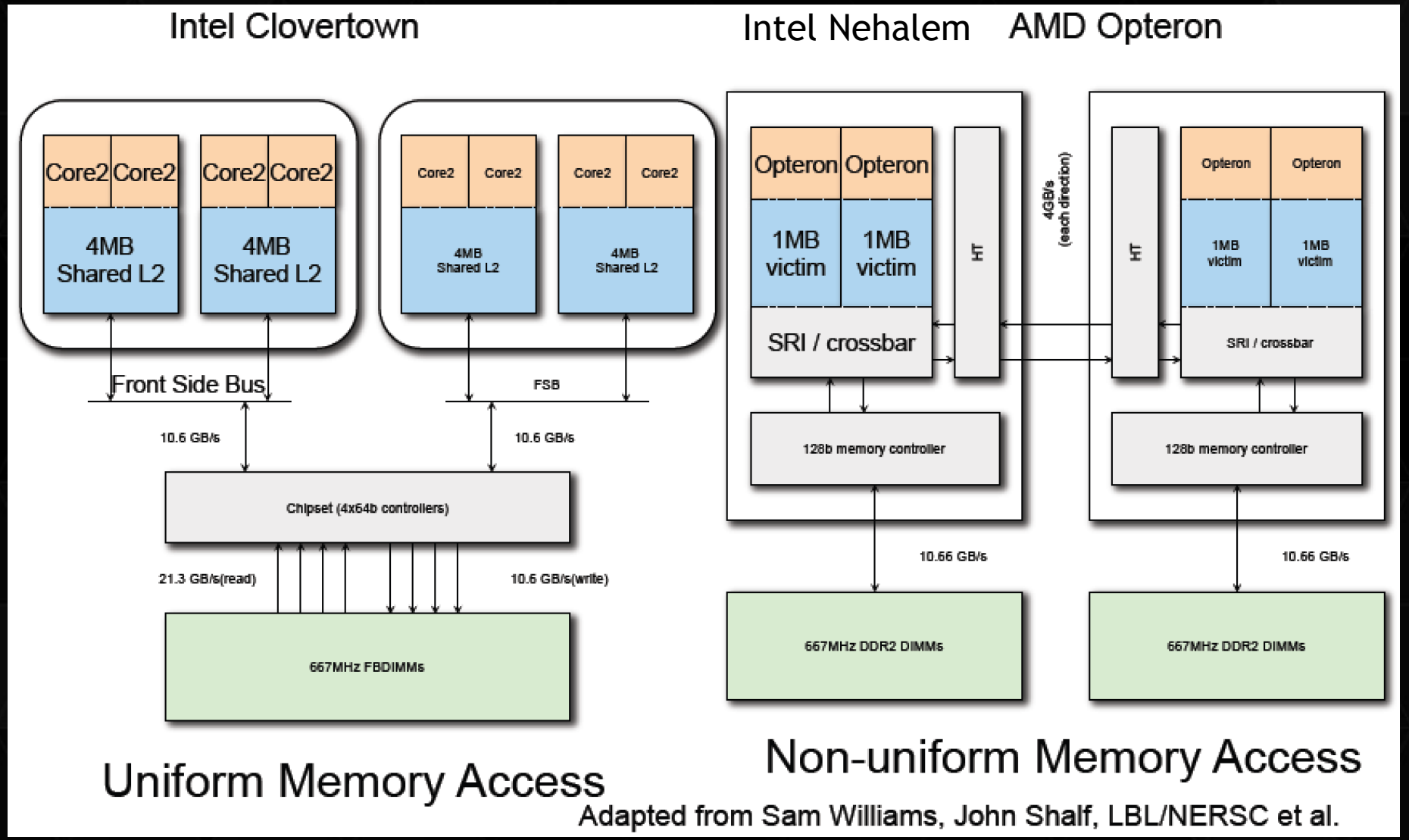
source: CA4
Hennessy & Patterson

MULTIPROCESSOR DISTRIBUTED MEM CLASS



source:
CA4
Hennessy &
Patterson

X86 ARCHITECTURE

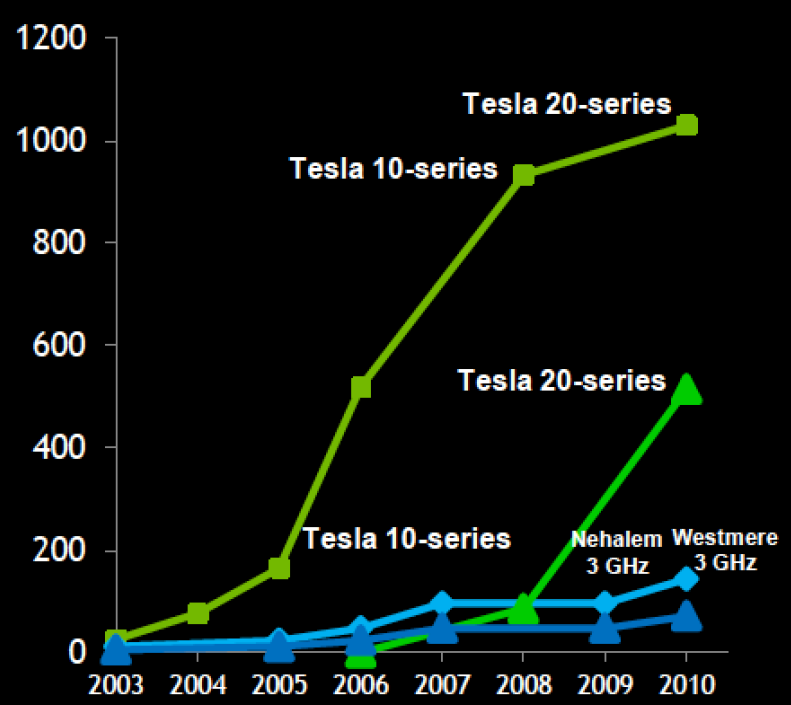


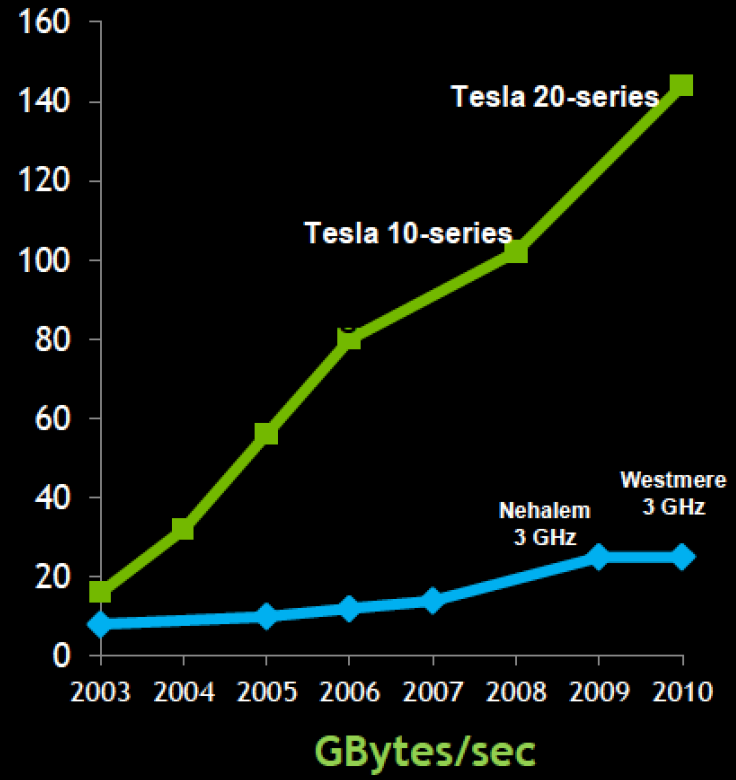
Adapted from Sam Williams, John Shalf, LBL/NERSC et al.

AGENDA

1 Intro
2 Processors Trends
3 Multiprocessors
4 GPU Computing

Why GPU Computing?

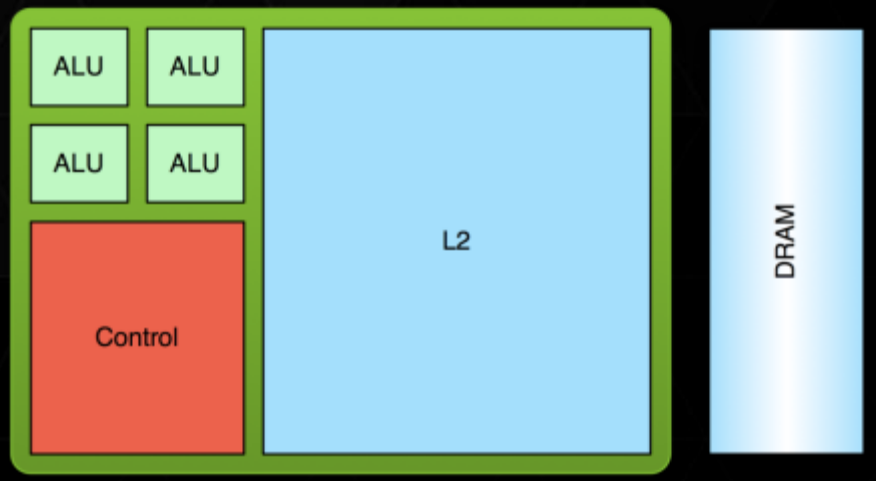




Single Precision: NVIDIA GPU
Double Precision: NVIDIA GPU
Single Precision: x86 CPU
Double Precision: x86 CPU

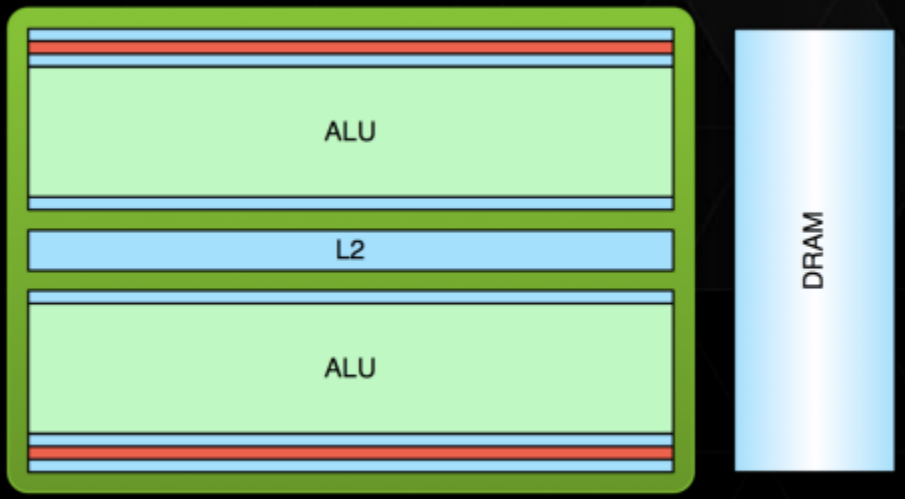
NVIDIA GPU
X86 CPU
ECC off

LOW LATENCY OR HIGH THROUGHPUT?



CPU

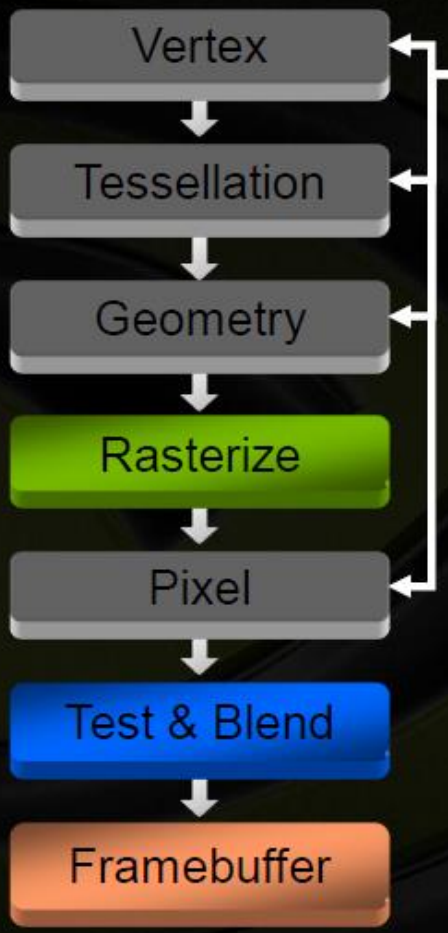
- Optimised for low-latency access to cached data sets
- Control logic for out-of-order and speculative execution



GPU

- Optimised for data-parallel, throughput computation
- Architecture tolerant of memory latency
- More transistors dedicated to computation

The Graphics Pipeline

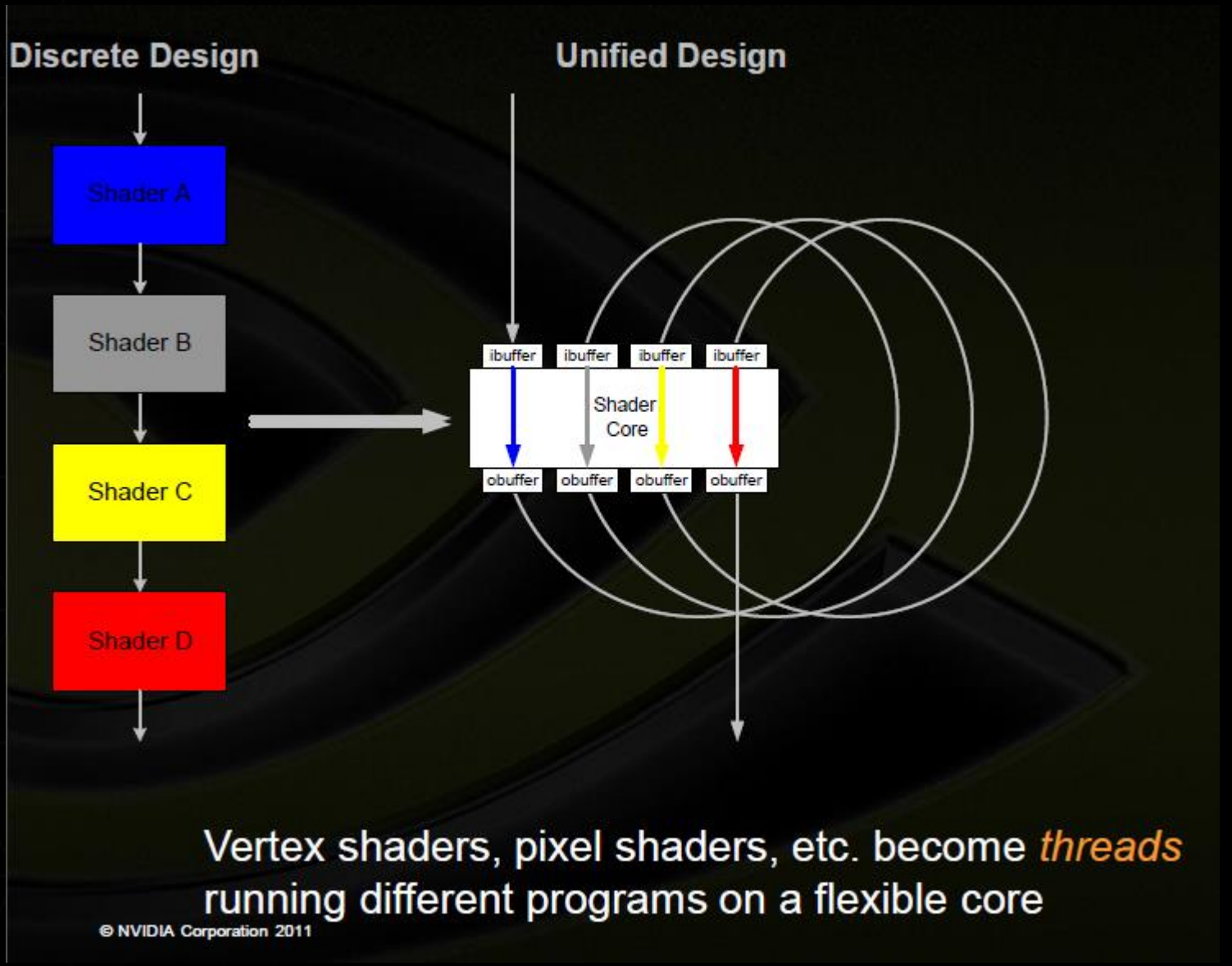



```
pixel_out
main(uniform sampler2D texture : TEXUNIT 0, pixel_in
{
  pixel_out OUT;
  float d= clamp(1.0 - pow(dot(IN.lightdist, IN.light
  float3 color = tex2D(texture, IN.texcoord).rgb;
  OUT.color = color * (d + 0.4);
  return OUT;
}
```

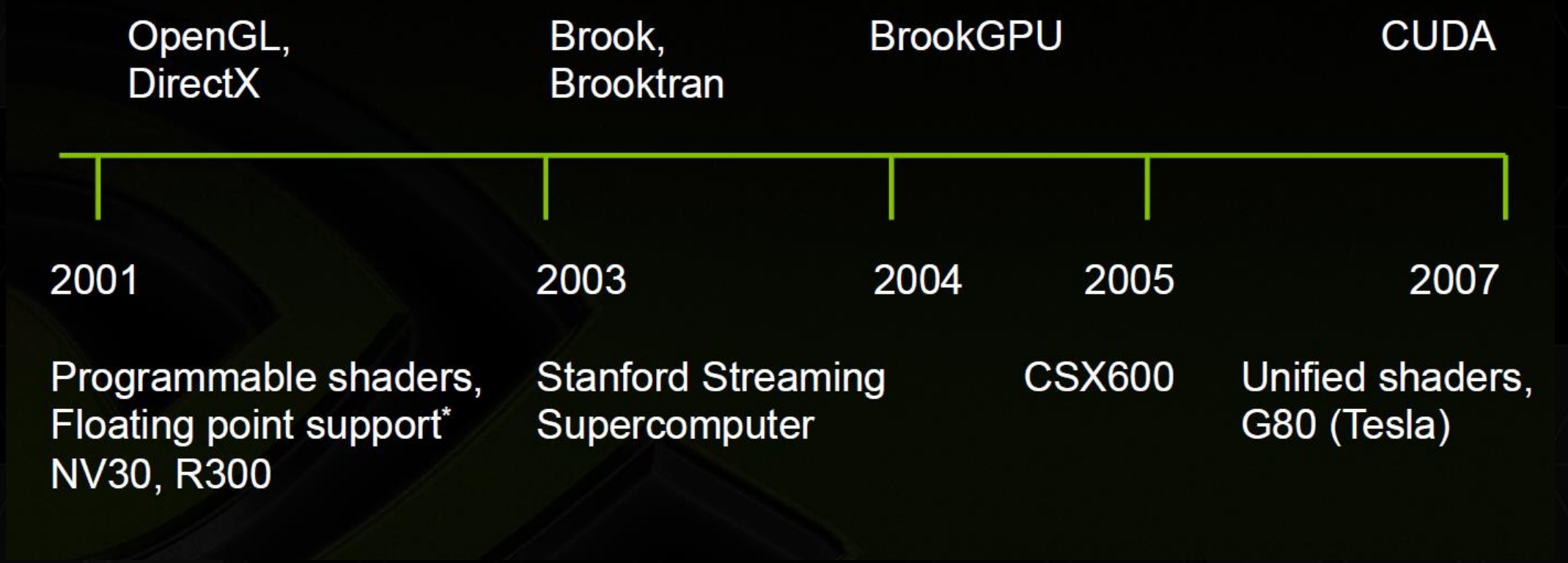
- Hardware used to look like this
  - Vertex, pixel processing became programmable
  - New stages added

GPU architecture increasingly centers around shader execution

GPU
UNIFIED DESIGN



GPGPU COMPUTING, B.C. ERA

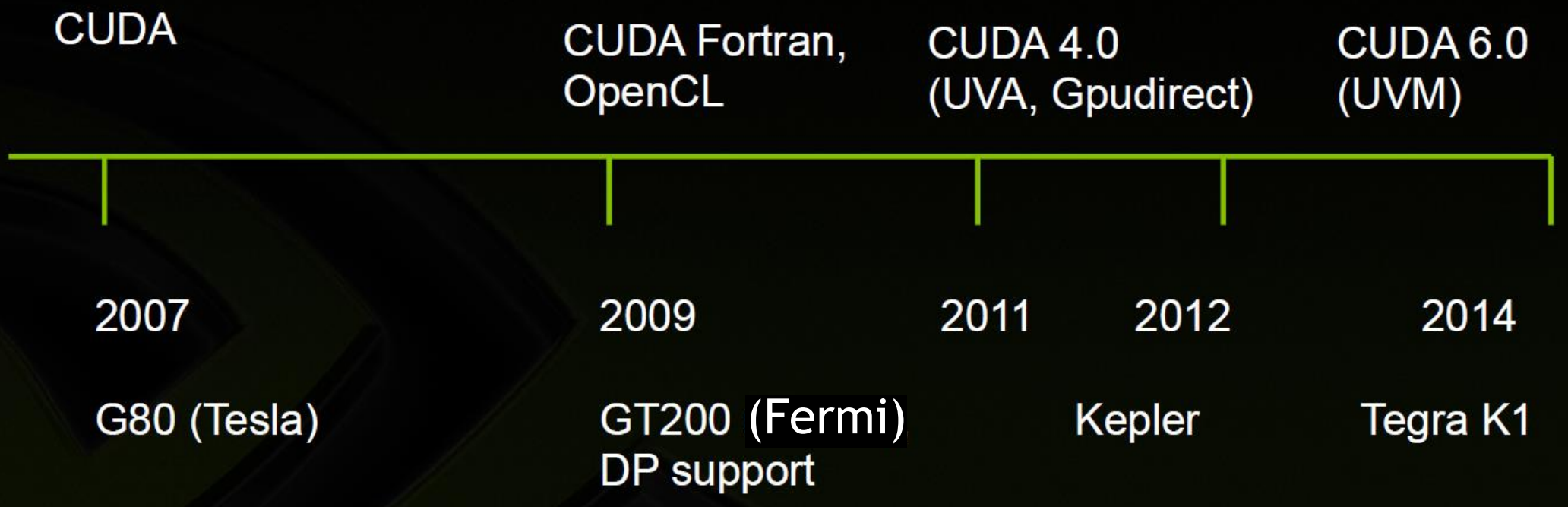


WHY GPU ARE ATTRACTIVE FOR HPC?

- Massive multithreaded manycore chips
- High flops count (SP and later DP)
- High memory bandwidth, (later) including ECC
- Lots of programming languages
- Lots of programming tools

- Widely used in Computational Physics

GPGPU COMPUTING, A.C. ERA



GROWTH IN GPU COMPUTING

2008

150,000
CUDA Downloads

27
CUDA Apps

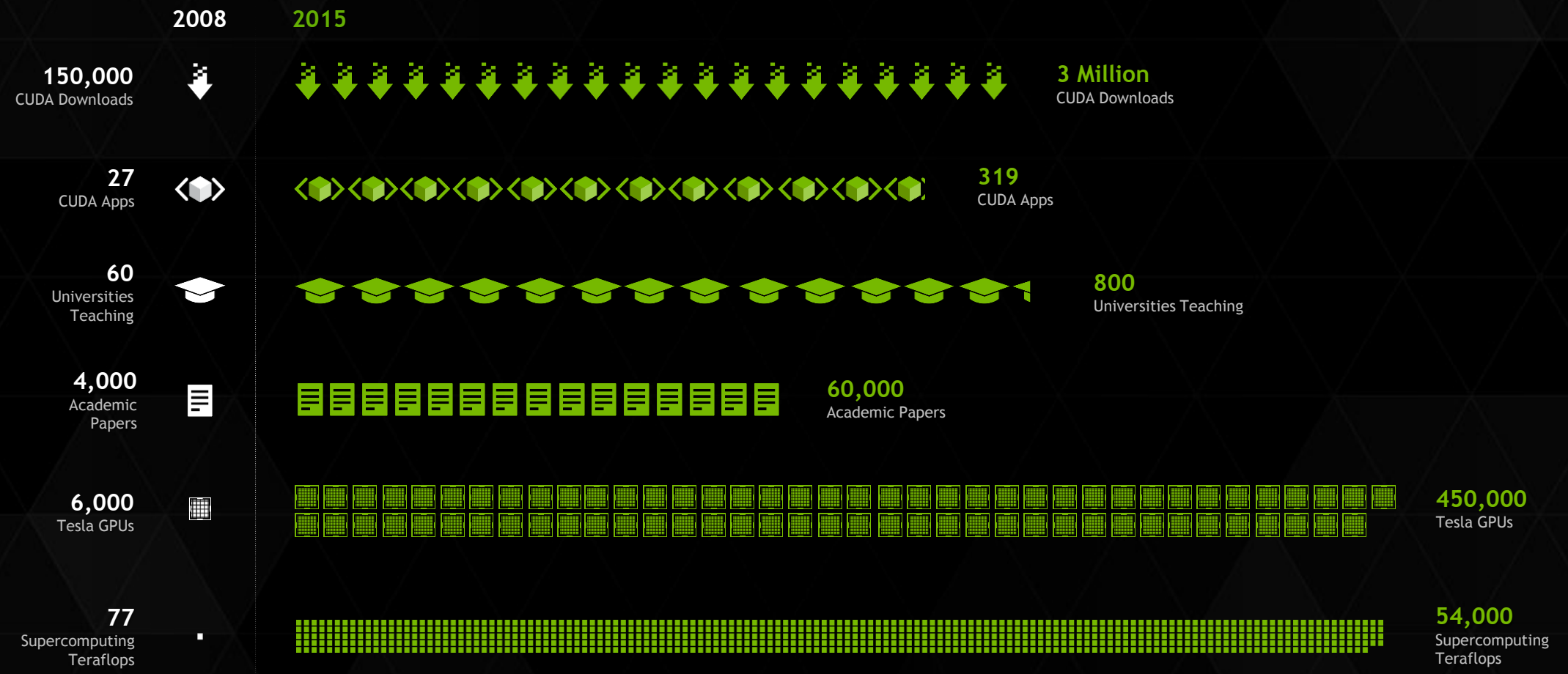
60
Universities
Teaching

4,000
Academic
Papers

6,000
Tesla GPUs

77
Supercomputing
Teraflops

GROWTH IN GPU COMPUTING

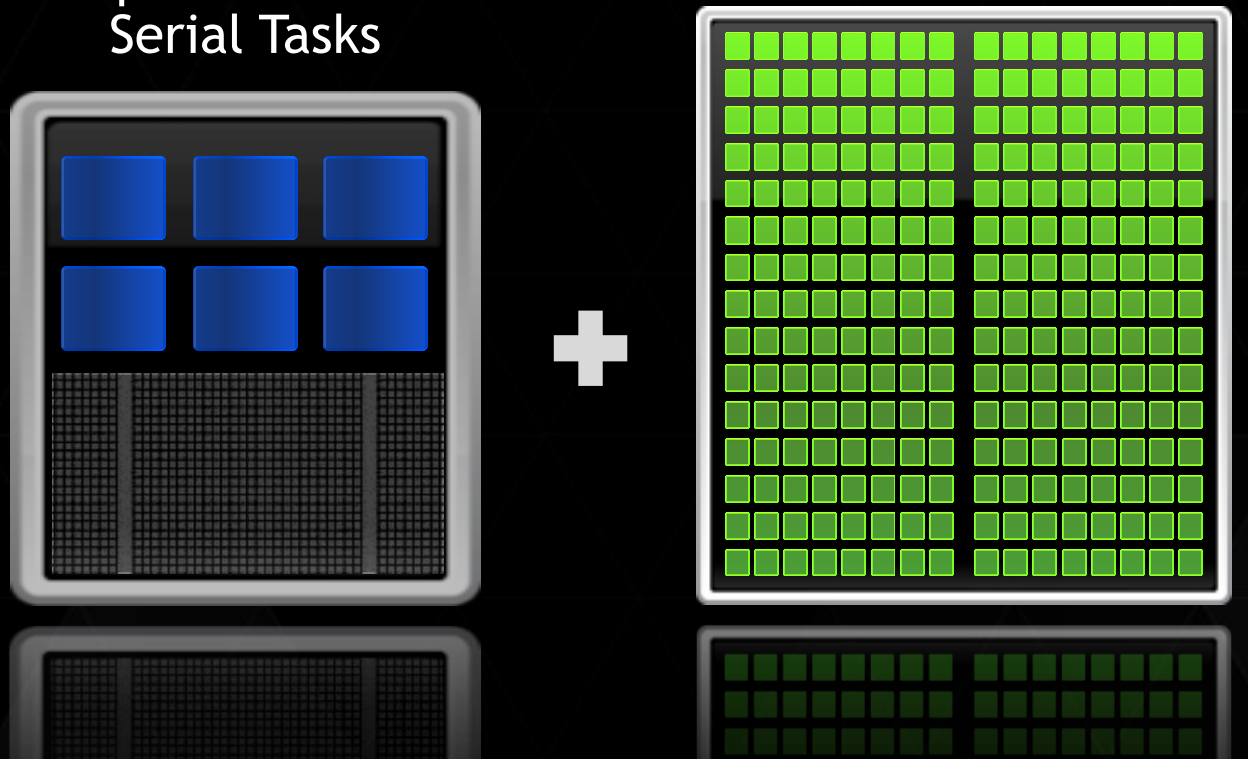


ACCELERATED, HETEROGENEOUS COMPUTING

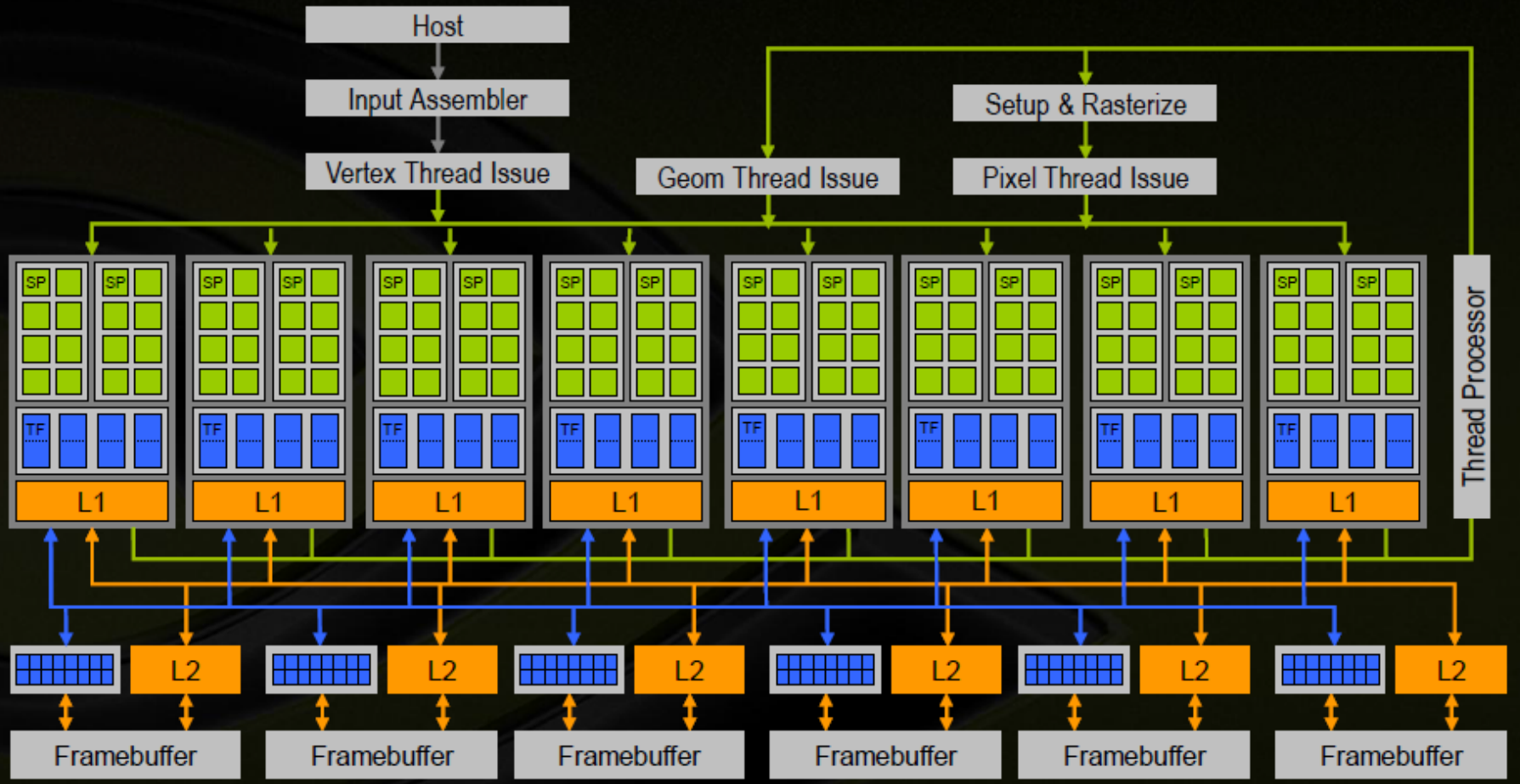
10x Performance & 5x Energy Efficiency for HPC

CPU
Optimized for
Serial Tasks

GPU Accelerator
Optimized for
Parallel Tasks

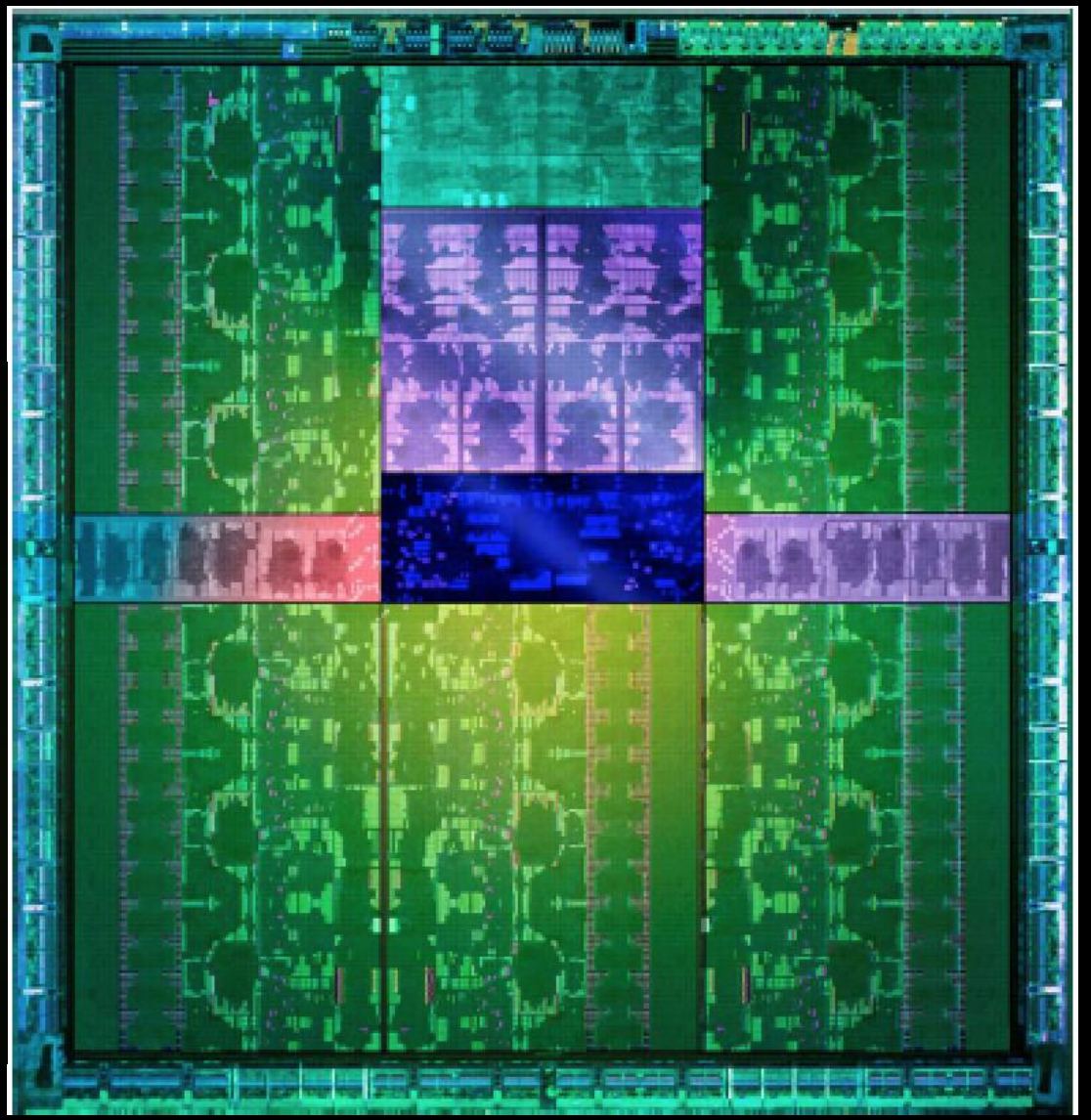


GEFORCE 8: 1ST MODERN GPU ARCH



NVIDIA KEPLER GK110 PROCESSOR

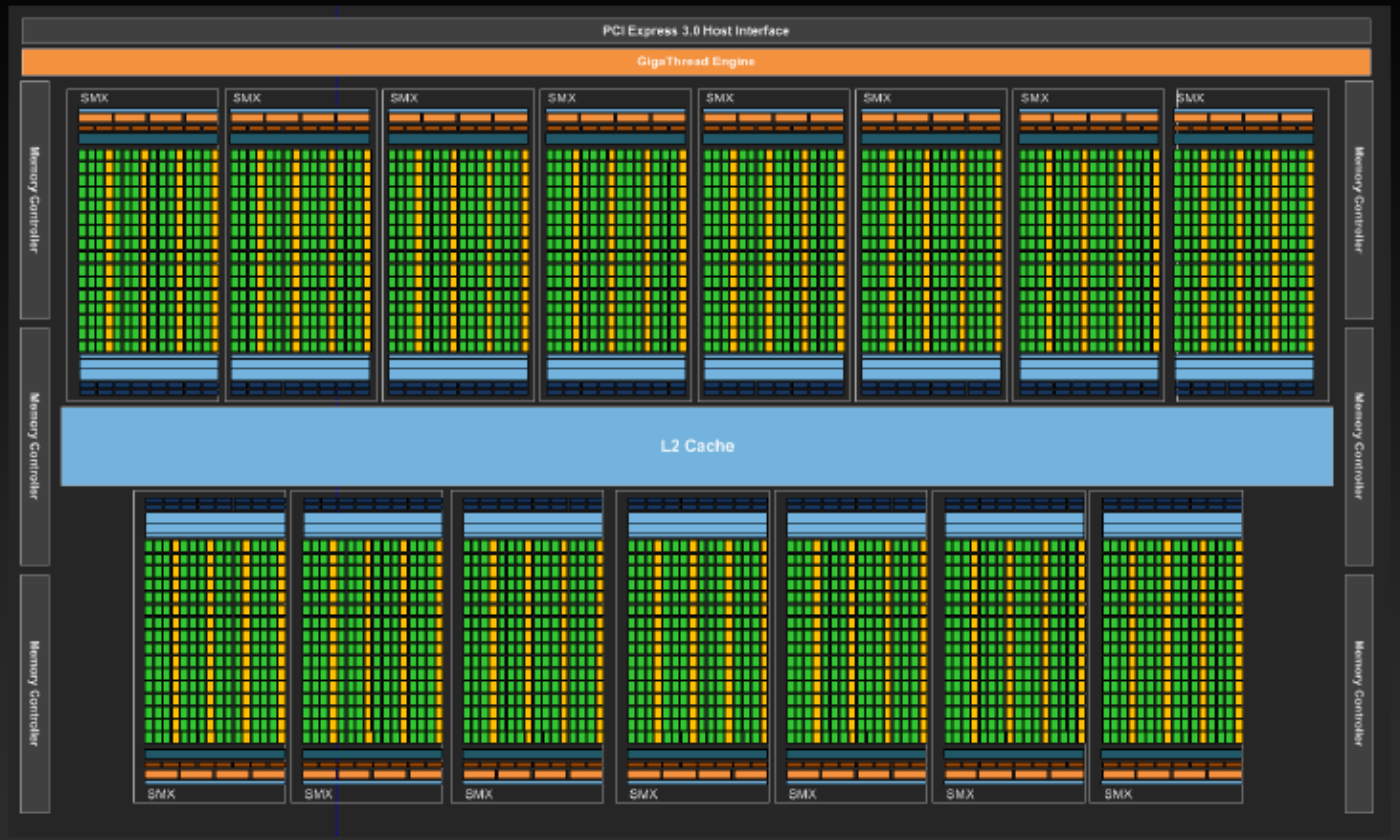
SMX	3x Performance per Watt
Hyper-Q	Easy Speed-up for Legacy MPI Apps
Dynamic Parallelism	Parallel Programming Made Easier than Ever



Kepler GK110 Block Diagram

Architecture

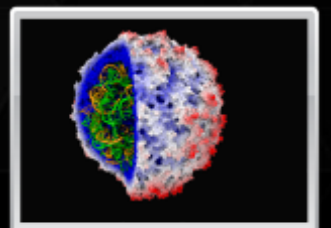
- 7.1B Transistors
- 15 SMX units
- > 1 TFLOP FP64
- 1.5 MB L2 Cache
- 384-bit GDDR5



KEPLER SMX DIAGRAM

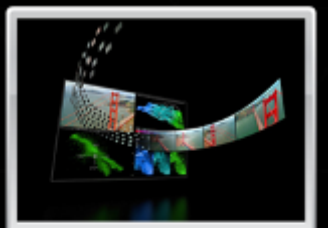
146X

Medical Imaging
U of Utah



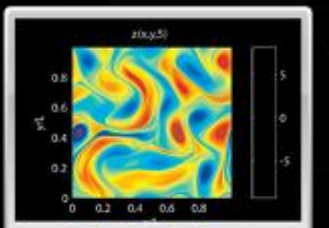
36X

Molecular Dynamics
U of Illinois, Urbana



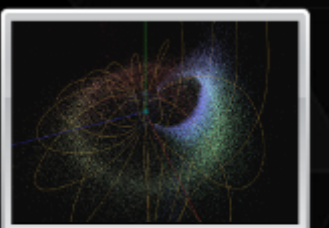
18X

Video Transcoding
Elemental Tech



50X

Matlab Computing
AccelerEyes



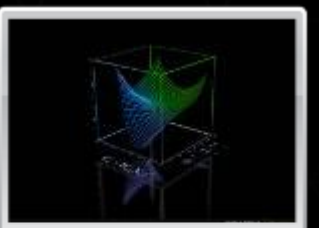
100X

Astrophysics
RIKEN

GPUs Accelerate Science

149X

Financial Simulation
Oxford



47X

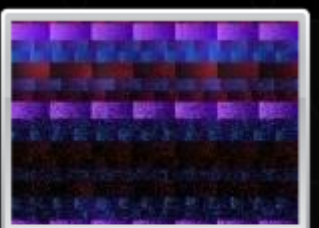
Linear Algebra
Universidad Jaime

20X

3D Ultrasound
Techniscan

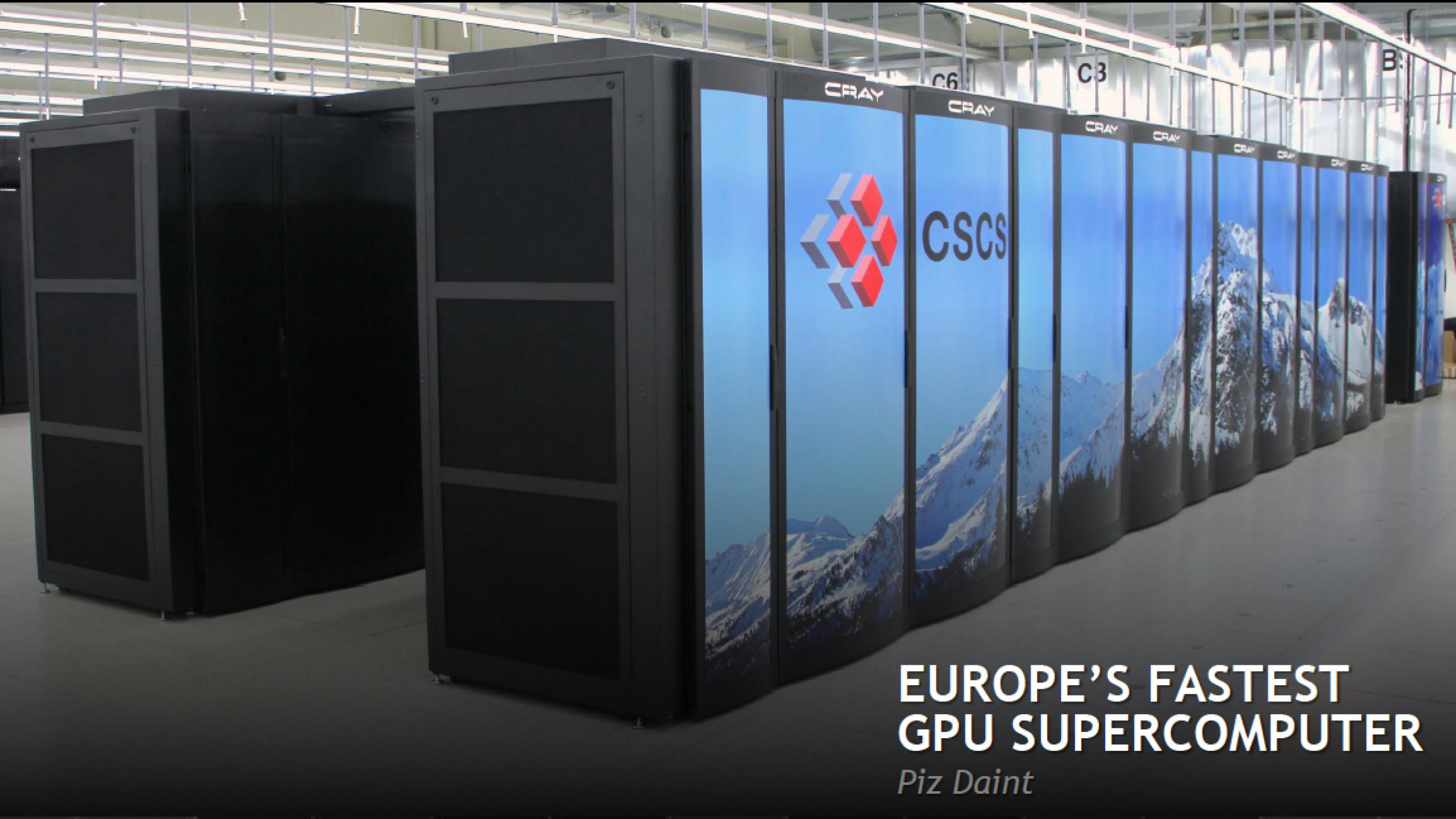
130X

Quantum Chemistry
U of Illinois, Urbana



30X

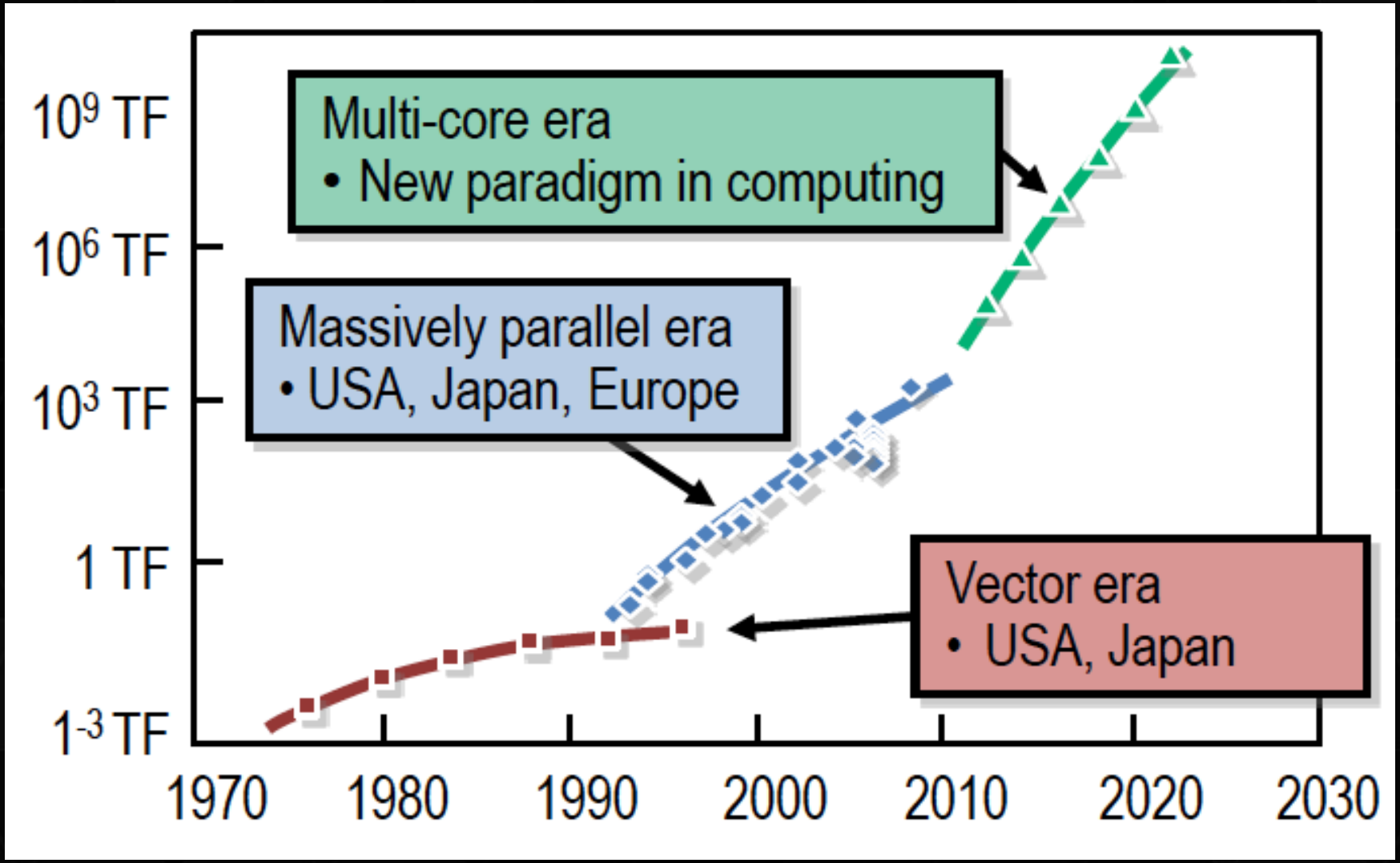
Gene Sequencing
U of Maryland



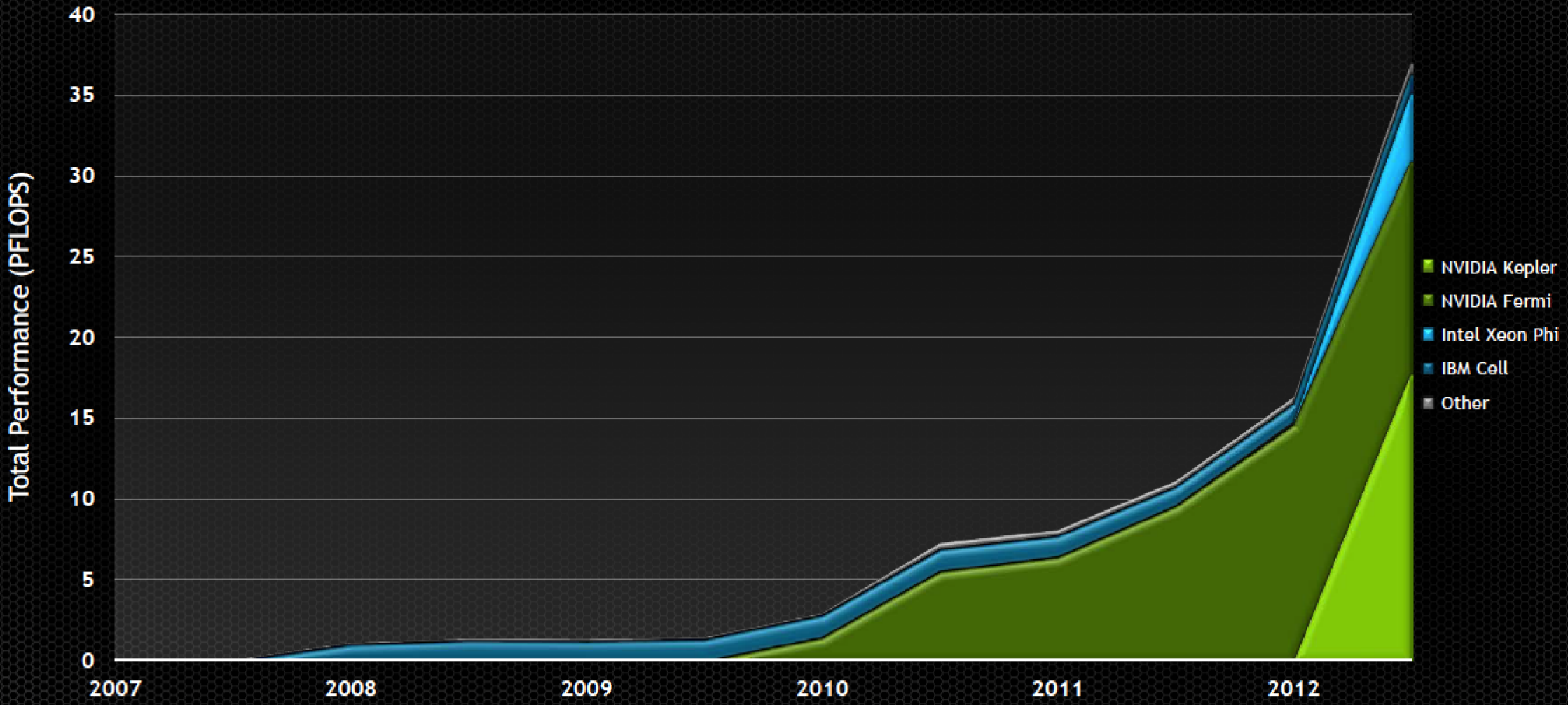
EUROPE'S FASTEST GPU SUPERCOMPUTER

Piz Daint

SUPERCOMPUTING ERA

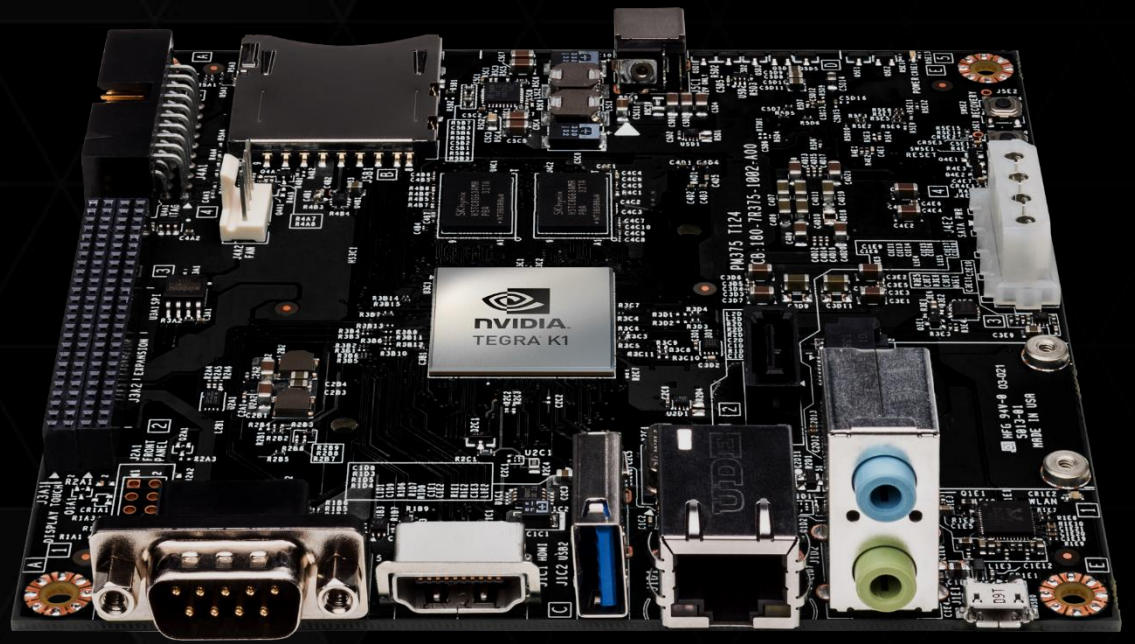


Top500: Performance from Accelerators



JETSON TK1

THE WORLD'S 1st EMBEDDED SUPERCOMPUTER



Development Platform for Embedded
Computer Vision, Robotics, Medical

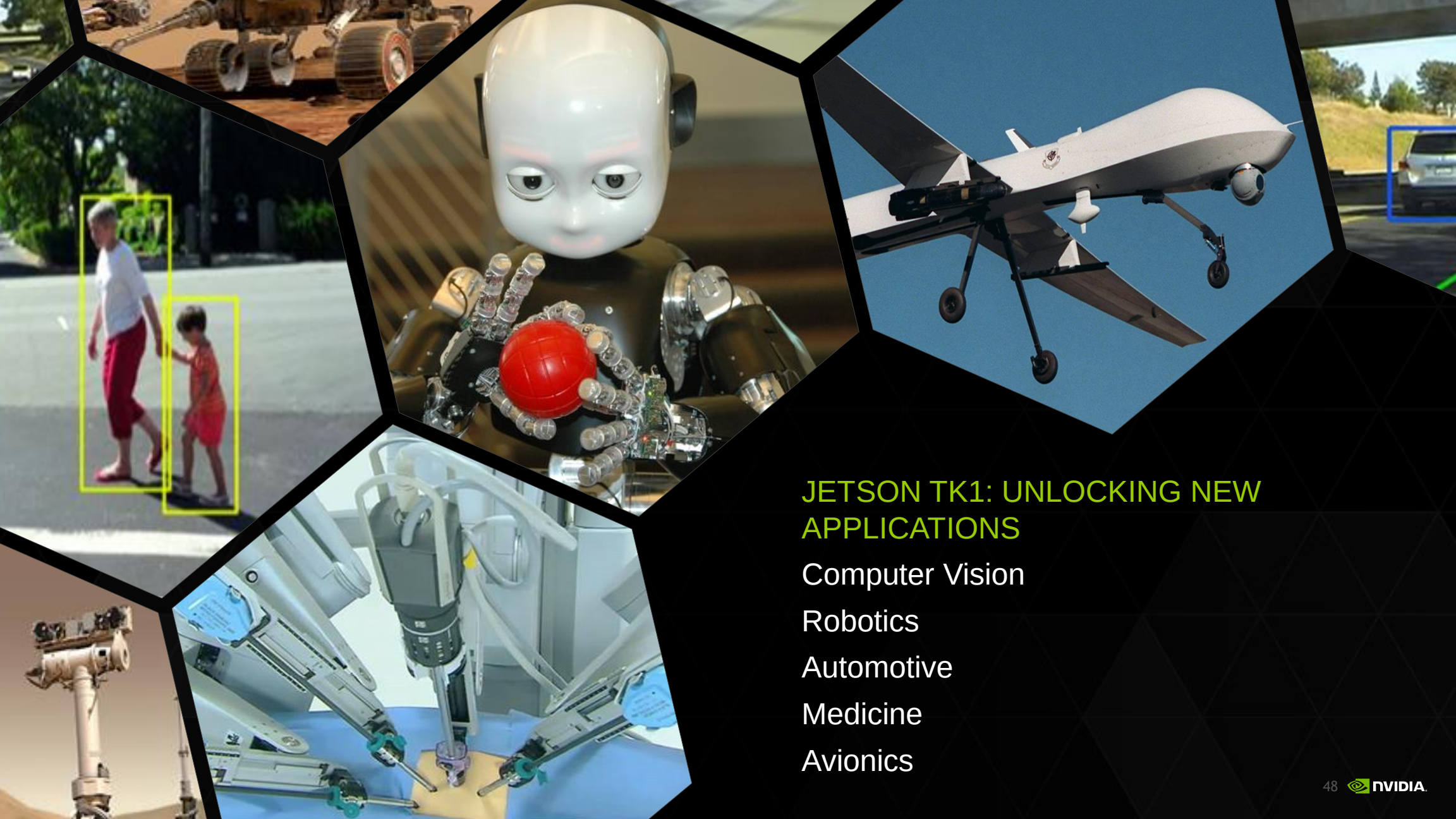
Tegra K1 SoC

Quad core A15 + Kepler GPU

192 CUDA Enabled cores

326 Gflops @ 5 Watt

$192



JETSON TK1: UNLOCKING NEW APPLICATIONS

Computer Vision

Robotics

Automotive

Medicine

Avionics

US TO BUILD TWO FLAGSHIP SUPERCOMPUTERS

OAK RIDGE National Laboratory

SUMMIT

150-300 PFLOPS
Peak Performance

Lawrence Livermore National Laboratory

SIERRA

> 100 PFLOPS
Peak Performance

IBM POWER9 CPU + NVIDIA Volta GPU

NVLink High Speed Interconnect

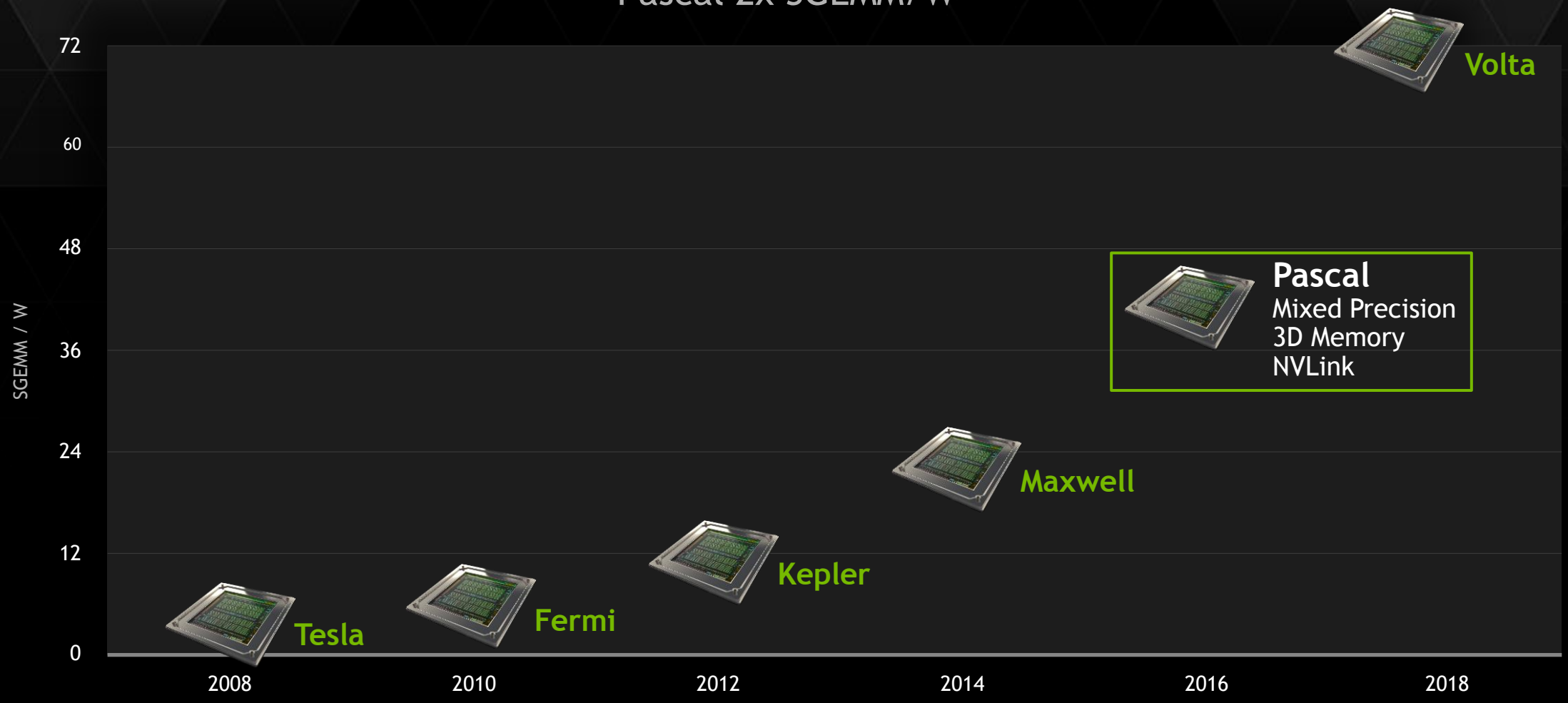
>40 TFLOPS per Node, >3,400 Nodes

2017

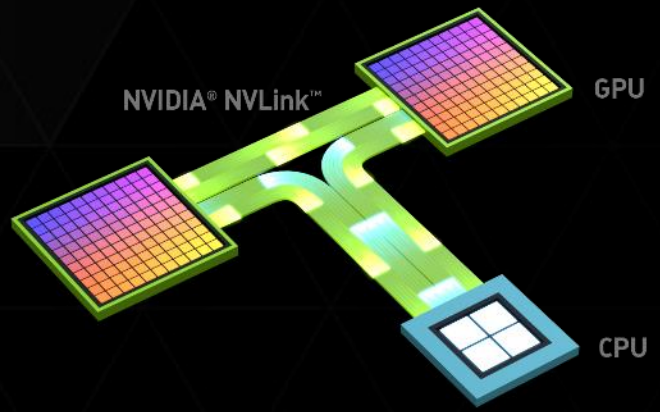
Major Step Forward on the Path to Exascale

GPU ROADMAP

Pascal 2x SGEMM/W

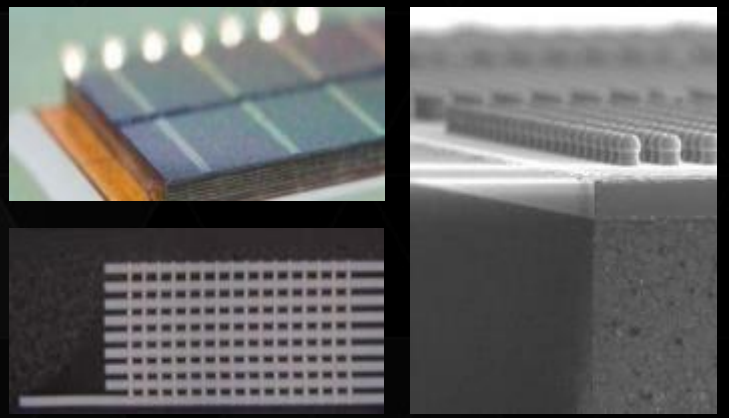


PASCAL GPU FEATURING NVLINK AND STACKED MEMORY



NVLINK

- GPU high speed interconnect
- 80-200 GB/s



3D Stacked Memory

- 4x Higher Bandwidth (~1 TB/s)
- 3x Larger Capacity
- 4x More Energy Efficient per bit

cnardone@nvidia.com
+39 335 5828197