Evolution of GPU Programming

20.03.23

Evolution of GPU Programming

- Where does the Graphics Processing Unit (GPU) come from?
- Why is it relevant outside of graphics?
- Tradeoffs in modern-day computing

Graphics Workloads

- 4k gaming at 120 Hz:

  $3840 \times 2160 \times 120\ \text{Hz} \approx 1\ \text{Gpx/s}$

  at 2 GHz clock: ≈ 2 cycles per pixel

- Full-HD stereo for HMDs at 120 Hz:

  $1920 \times 1080 \times 2 \times 120\ \text{Hz} \approx 500\ \text{Mpx/s}$

  at 2 GHz clock: ≈ 4 cycles per pixel

- And that's just the *output* data!

Graphics Workloads

How can we compute millions of pixels

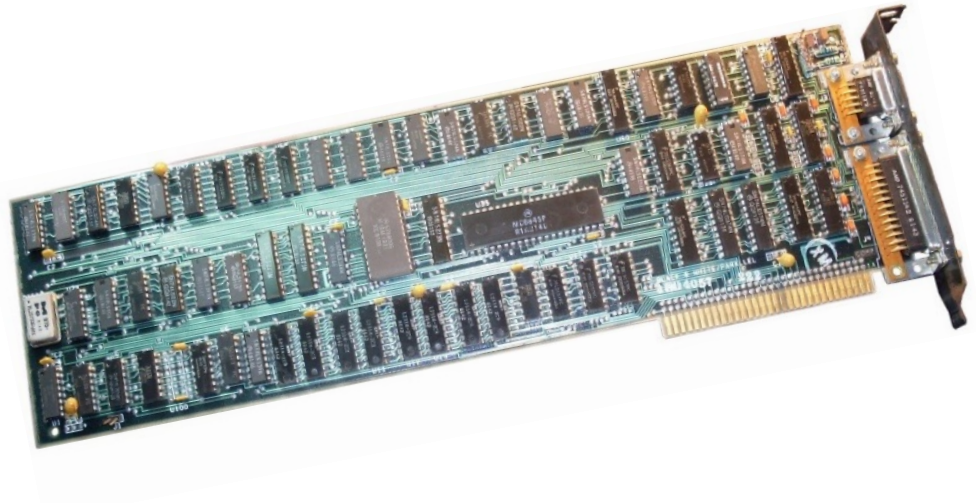
- forming an image of a complex 3D scene
- in real-time?

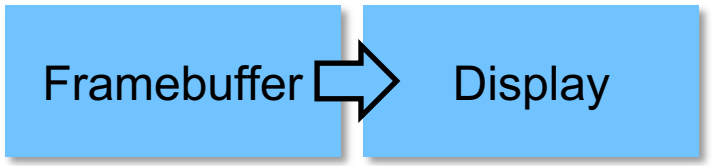
→ parallelization

→ specialized hardware

PC Graphics 1981

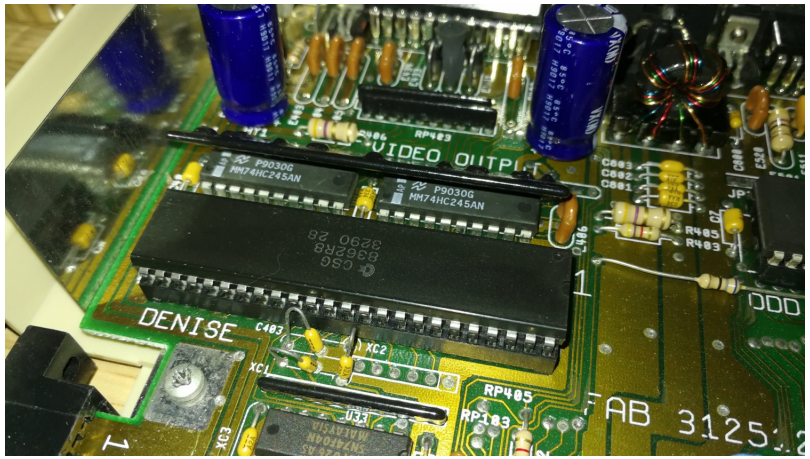
- Color graphics card
- IBM Color Graphics Adapter (CGA)
  - Display a color buffer
- Direct drive CRT monitor or TV
- Resolution: $160\text{x}100\ (4\text{bit}), 320\text{x}200\ (2\text{bit}), 640\times200\ (1\text{bit})$
- 16 colors (4bit)



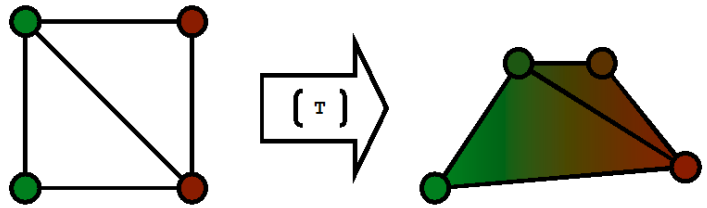


PC Graphics 1987

- Video processor (Denise.- Display ENabler)
- Commodore Amiga
- Blitter co-processor for image bitmap copy
- Process: 5 µm (NMOS)
- Resolution
  - 320x200 to 640x400 for NTSC (704x848 overscan)
  - 320x256 to 640x512 for PAL (704x576 overscan)
- 4096 colors



3D Graphics Pipeline



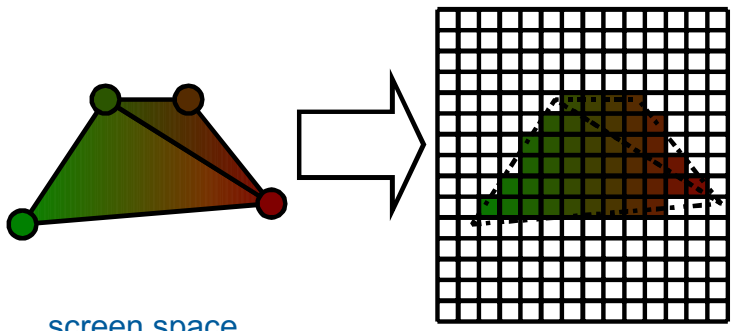
object space vertices

screen space vertices

texture sampling + blending

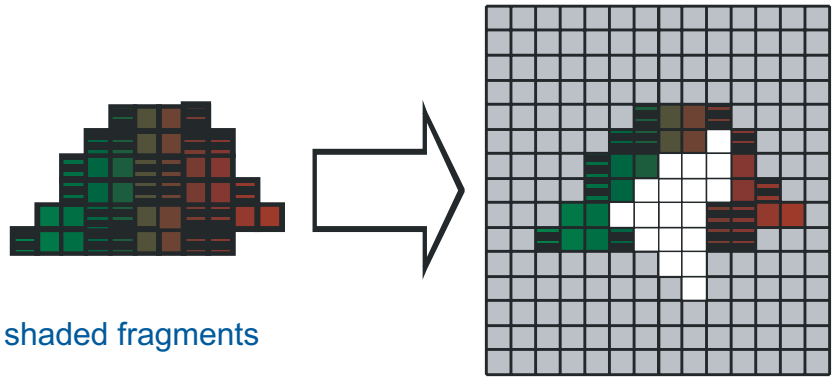
fragments

shaded fragments



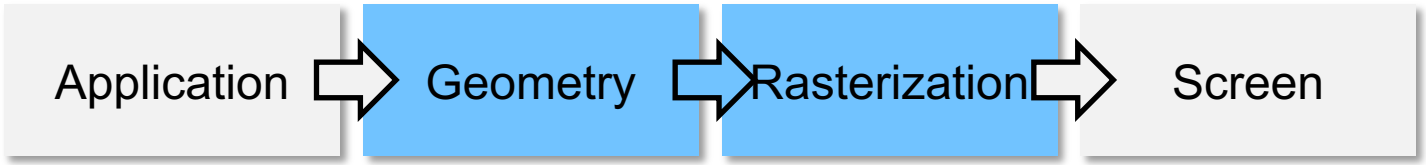
screen space primitives

fragments



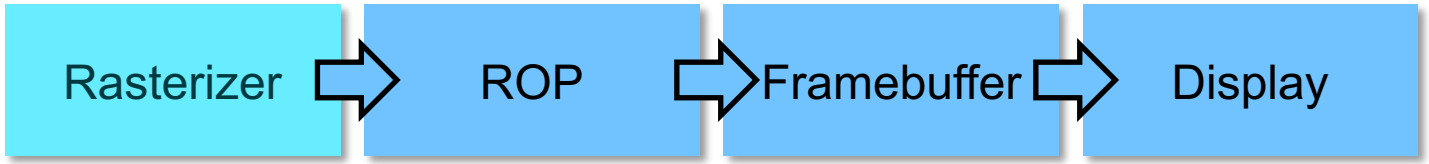
shaded fragments

framebuffer



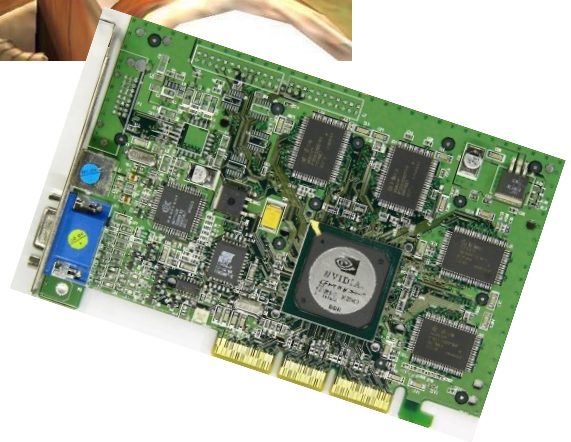
PC Graphics 1996

- 3dfx Voodoo 1
- Triangle rasterization
- 3D only, no 2D support
- Resolution 640×480
- Process: 500 nm
- 50 MHz, 45 Mpx/s, 4 MB RAM
- 1 texture unit, 1 raster unit



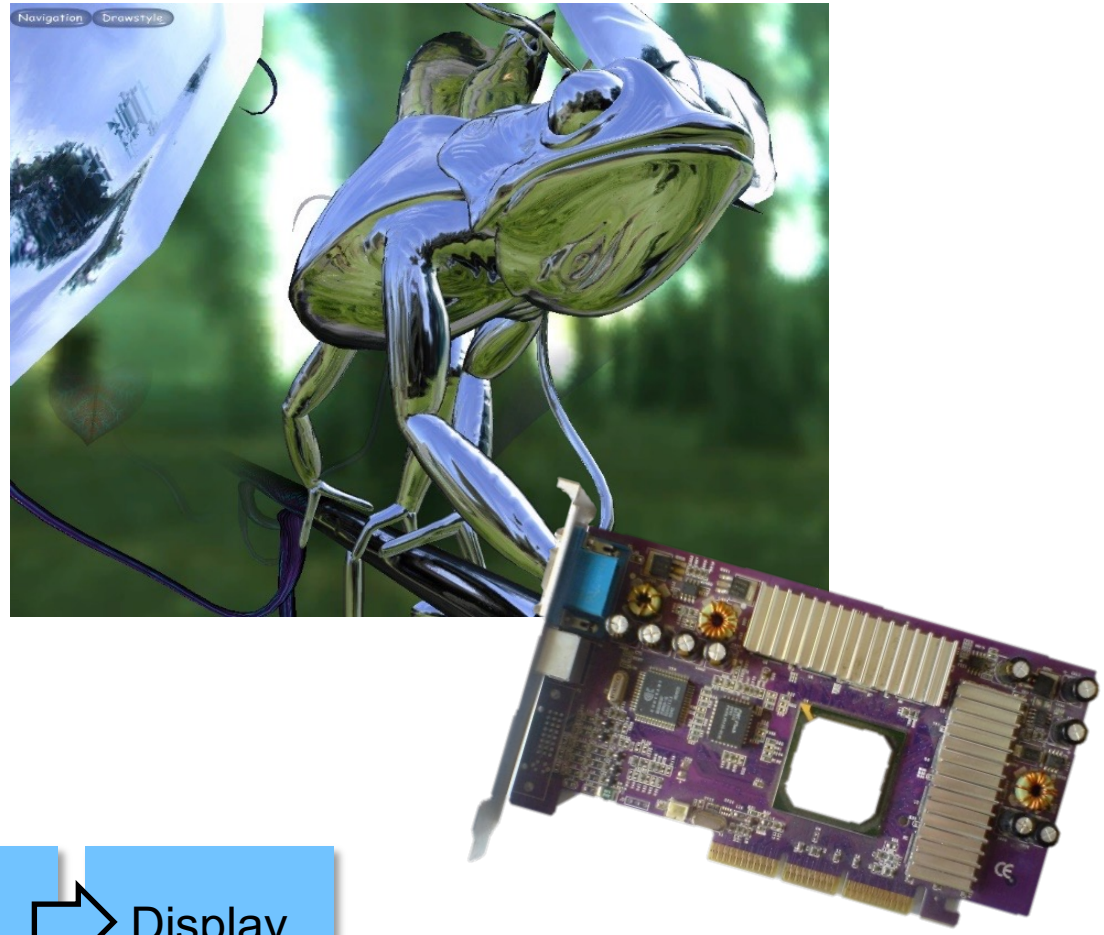
PC Graphics 1999

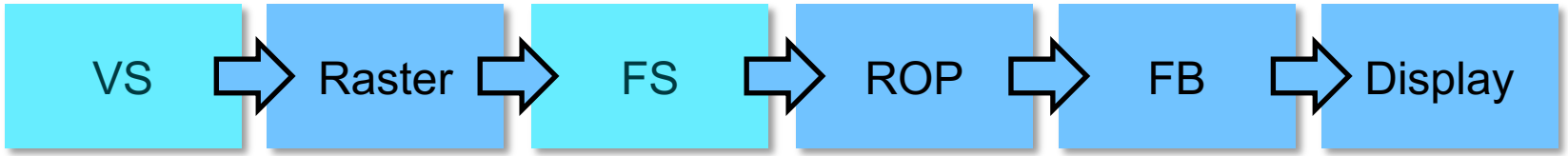
- NVIDIA GeForce 256
- Hardware transform and lighting
- First ‘GPU’
- Process: 220 nm
- 120 MHz, 480 Mpx/s, 64 MB RAM
- 1 geometry processor, 4 pixel processors,
  4 texture units, 4 ROP units



PC Graphics 2001

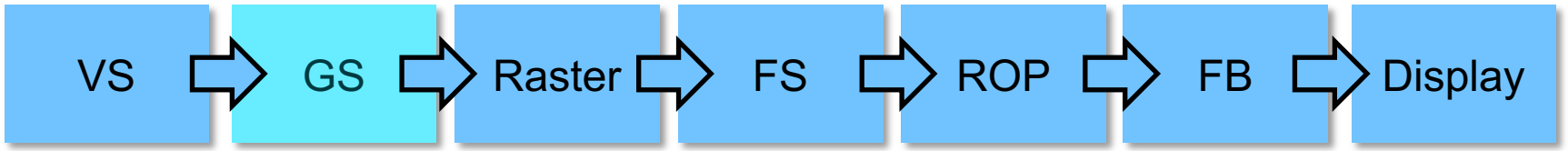
- NVIDIA GeForce 3
- Programmable vertex and fragment shader
- Process: 150 nm
- 200 MHz, 800 Mpx/s, 64 MB RAM
- 1 geometry processor, 4 pixel processors,
  8 texture units, 4 ROP units





PC Graphics 2006

- NVIDIA GeForce 8
- Geometry shader
- Unified shader model
- **CUDA 1.0**
- Process: 90 nm
- Shader clock 1500 MHz
- 576 GFLOPs



PC Graphics 2009

- ATI Radeon HD 5000
- Tessellation
- Process: 40 nm
- 850 MHz
- 2720 GFLOPs

PC Graphics 2013

- NVIDIA GeForce GTX TITAN
- Process: 28 nm
- 890 MHz
- 5100 GFLOPs @ 250 W

PC Graphics 2015

- NVIDIA GeForce GTX 980 Ti
- Process: 28 nm
- 1075 MHz
- 6050 GFLOPs @ 250 W

PC Graphics 2017

- NVIDIA GeForce GTX 1080 Ti
- Process: 16 nm
- 1480 (*1582*) MHz
- 10609 (*11340*) GFLOPs @ 250 W

PC Graphics 2018

- NVIDIA GeForce RTX 2080 Ti
- Raytracing cores
- Process: 12 nm
- 1350 (*1545*) MHz
- 11750 (*13448*) GFLOPs @ 250 W

PC Graphics 2020

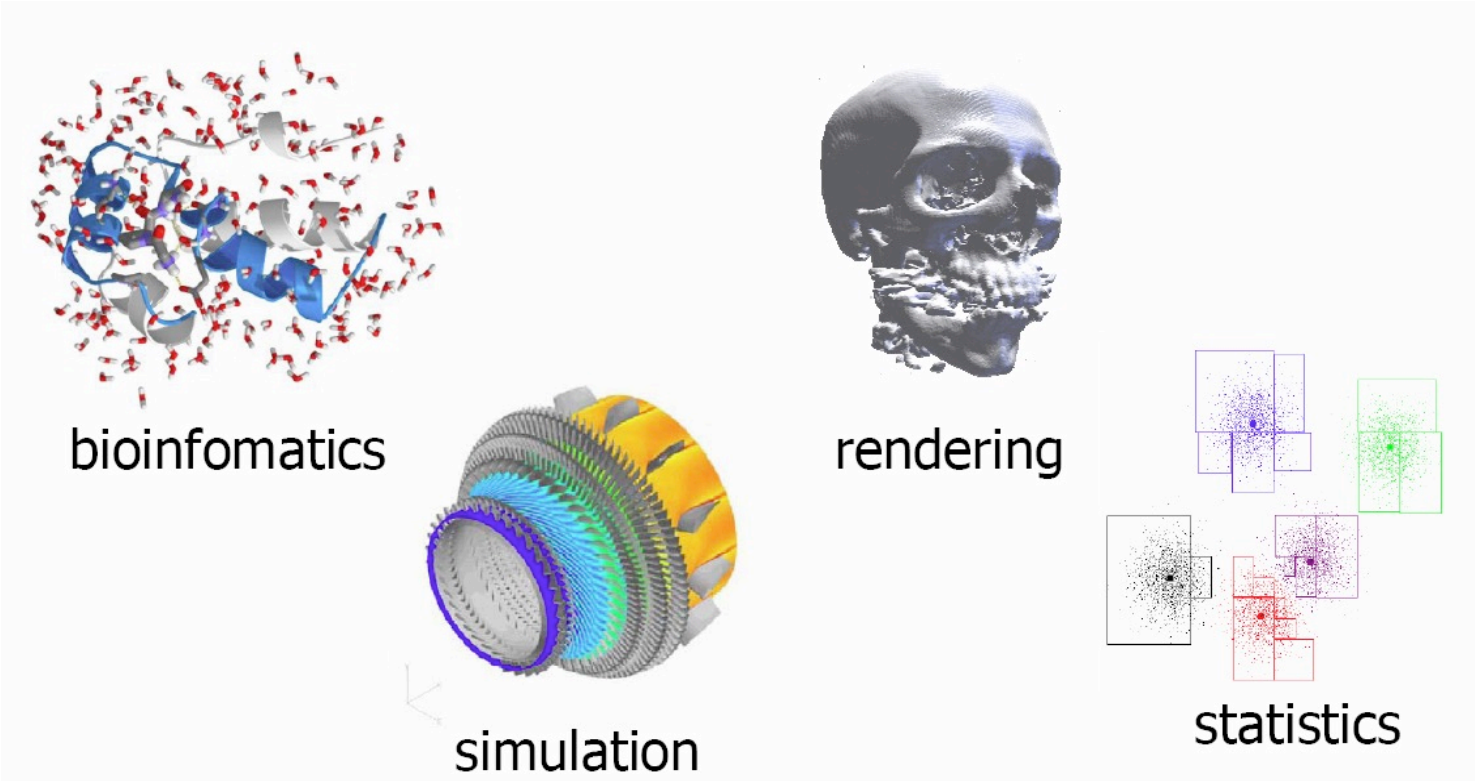
- NVIDIA GeForce RTX 3080
- Process: 8 nm
- 1440 (*1710*) MHz
- 25068 (*29768*) GFLOPs @ 320 W

PC Graphics 2020

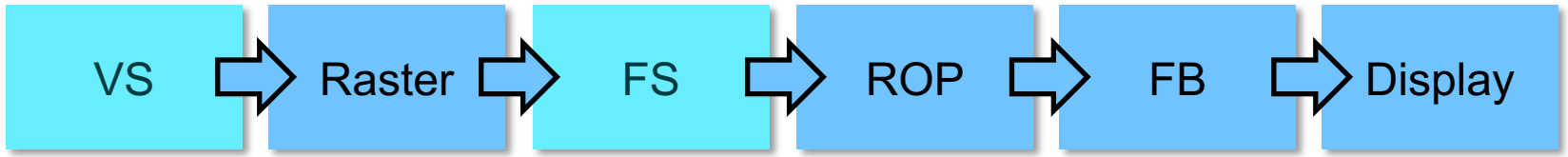
- AMD Radeon RX 6800
- Process: 7 nm
- 1825 (*2250*) MHz
- 16819 (*20736*) GFLOPs @ 300 W

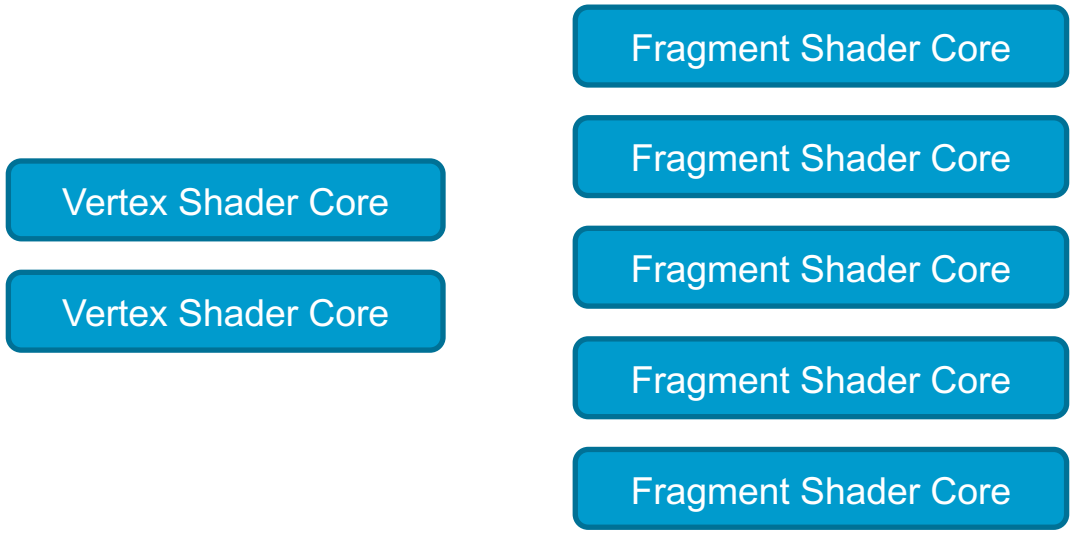
General Purpose Computation on GPUs (GPGPU)

- Abuse pixel shaders for parallel programming
- Shader metaprogramming, Brook for GPUs

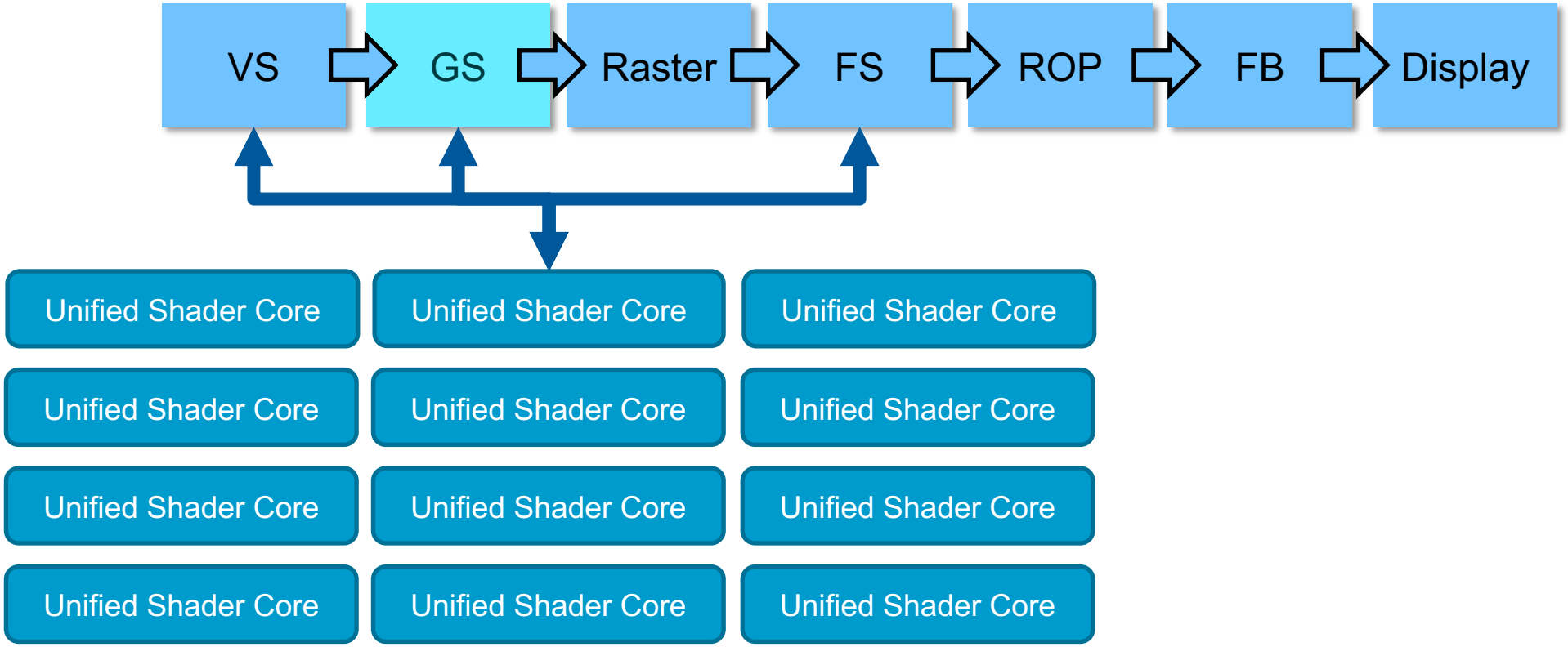


GPGPU: Program Vertex / Fragment Shaders

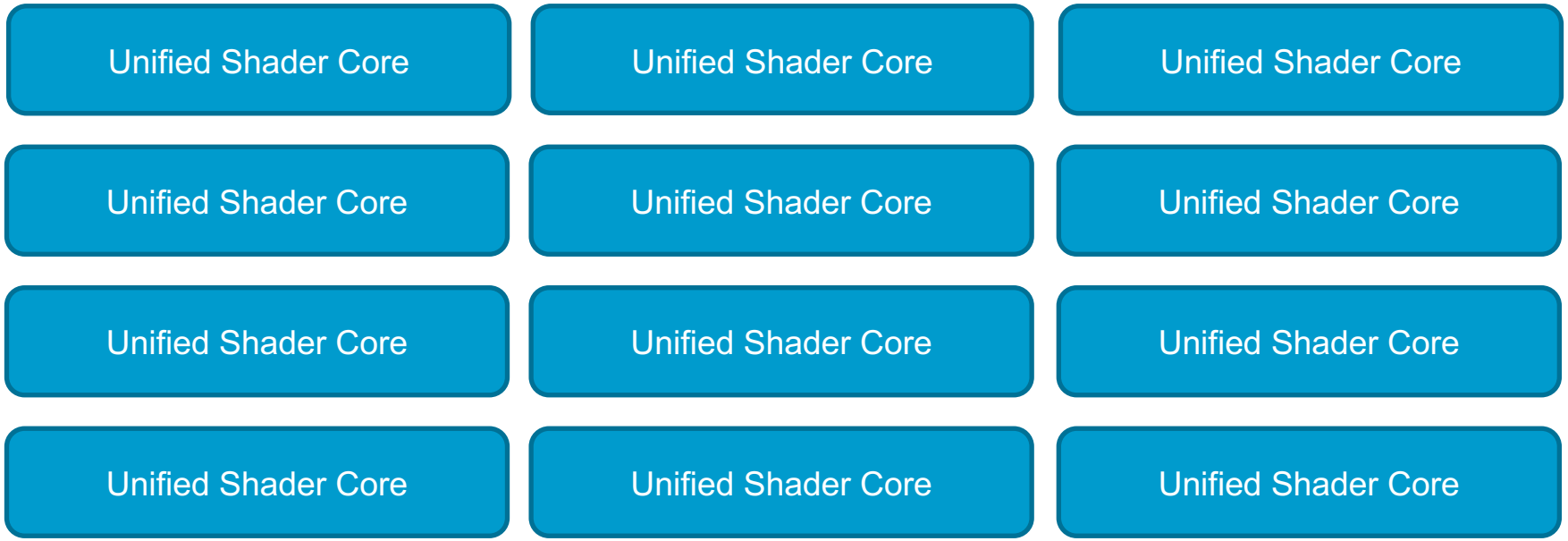




Unified Shader Architecture



GPU Computing: Compute Mode

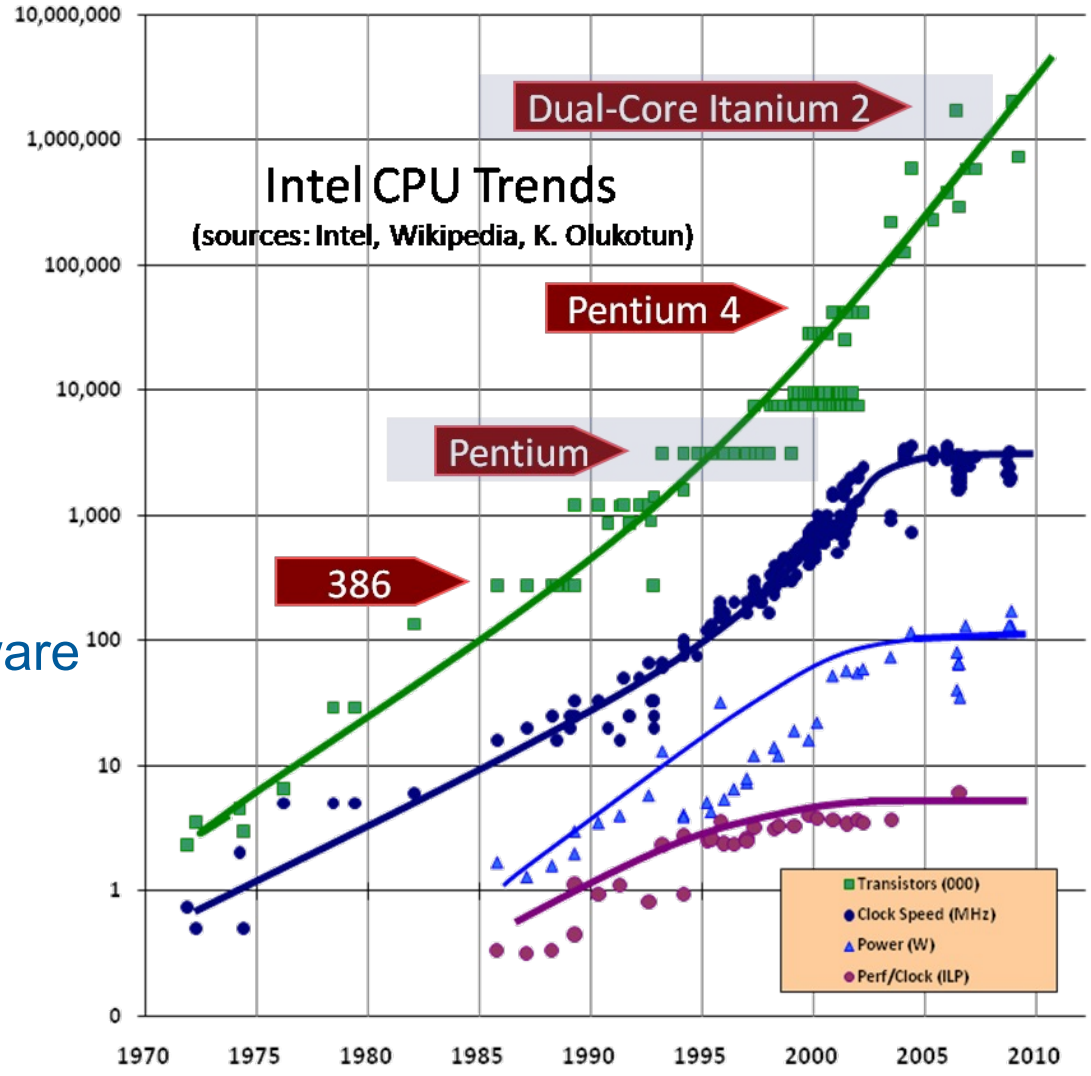


Computing 1970 – 2009

Herb Sutter in 2004: “The free lunch is over”

- Performance won’t increase with transistor count
- Processors hit the power wall
- Clock rates cannot be increased significantly

➔ A fundamental turn toward concurrency in software



Power Wall

the good old days

- Limited by number of transistors
- Use as much power as you want

today

- Limited by energy consumption and heat
- Transistors basically “free”

  (put as many on a chip as you can afford to turn on)

We might have many specialized cores on one chip, but only use the ones best suited for the current problem.

ILP Wall

the good old days

- Huge performance gains from superscalar design

today

- Diminishing returns on spending more hardware on instruction-level parallelism (ILP)

We need to provide parallelism at a higher level to get more computations done.

Memory Wall

the good old days

- Instructions are costly
- Memory access is fast

  (precompute and load whenever possible)

today

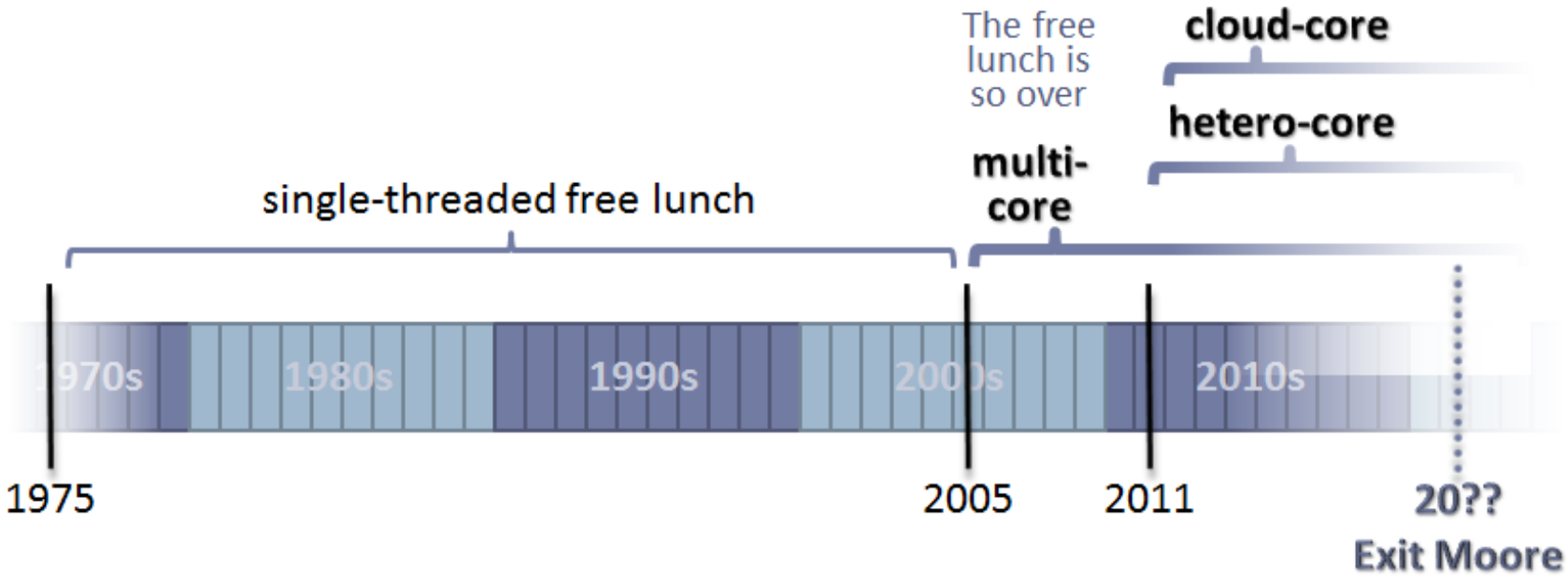
- Memory is slow

  (>200 cycles to go to DRAM)
- Instructions are fast

We need to avoid memory transaction and rather compute things on demand.

Herb Sutter in 2011: “Welcome to the Jungle”

- Transition of all mainstream form factors to parallel computing

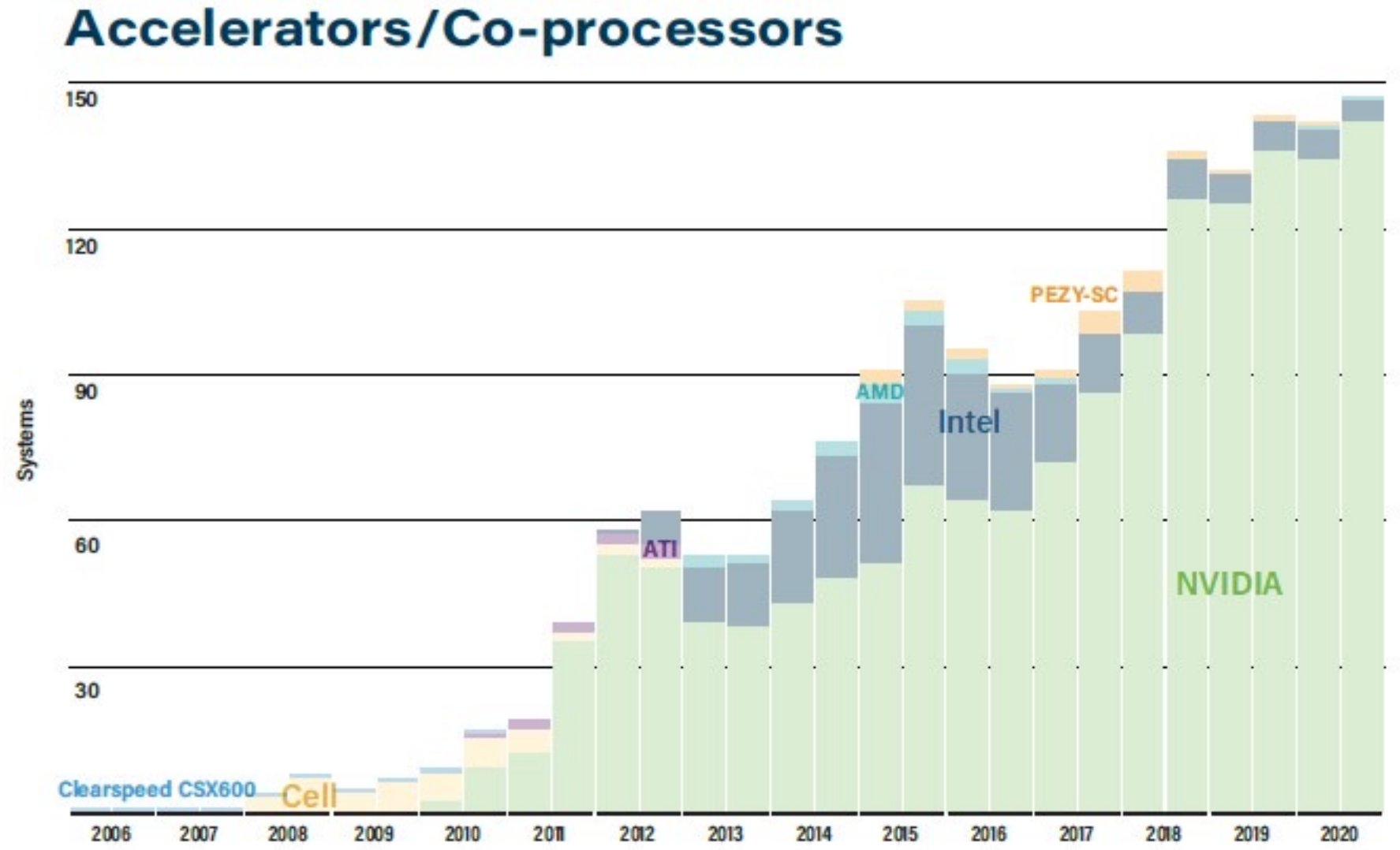


TOP500 Supercomputer

TOP500 Rank	GFLOPS/W	Rmax [TFlop/s]	Name Site	Computer
1	15,42	442.010,00	Supercomputer Fugaku RIKEN Center for Computational Science	Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D
2	14,72	148.600,00	Summit DOE/SC/Oak Ridge National Laboratory	IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband
3	12,72	94.640,00	Sierra DOE/NNSA/LLNL	IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband
4	6,05	93.014,60	Sunway TaihuLight National Supercomputing Center in Wuxi	Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway
5	23,98	63.460,00	Selene NVIDIA Corporation	NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband
6	3,32	61.444,50	Tianhe-2A National Super Computer Center in Guangzhou	TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000
7	25,01	44.120,00	JUWELS Booster Module Forschungszentrum Juelich (FZJ)	Bull Sequana XH2000 , AMD EPYC 7402 24C 2.8GHz, NVIDIA A100, Mellanox HDR InfiniBand/ParTec ParaStation ClusterSuite
8	15,74	35.450,00	HPC5 Eni S.p.A.	PowerEdge C4140, Xeon Gold 6252 24C 2.1GHz, NVIDIA Tesla V100, Mellanox HDR Infiniband
9	n/a	23.516,40	Frontera Texas Advanced Computing Center (TACC)	Dell C6420, Xeon Platinum 8280 28C 2.7GHz, Mellanox InfiniBand HDR
10	n/a	22.400,00	Dammam-7 Saudi Aramco	Cray CS-Storm, Xeon Gold 6248 20C 2.5GHz, NVIDIA Tesla V100 SXM2, InfiniBand HDR 100

TOP500 November 2020

TOP500 Supercomputer



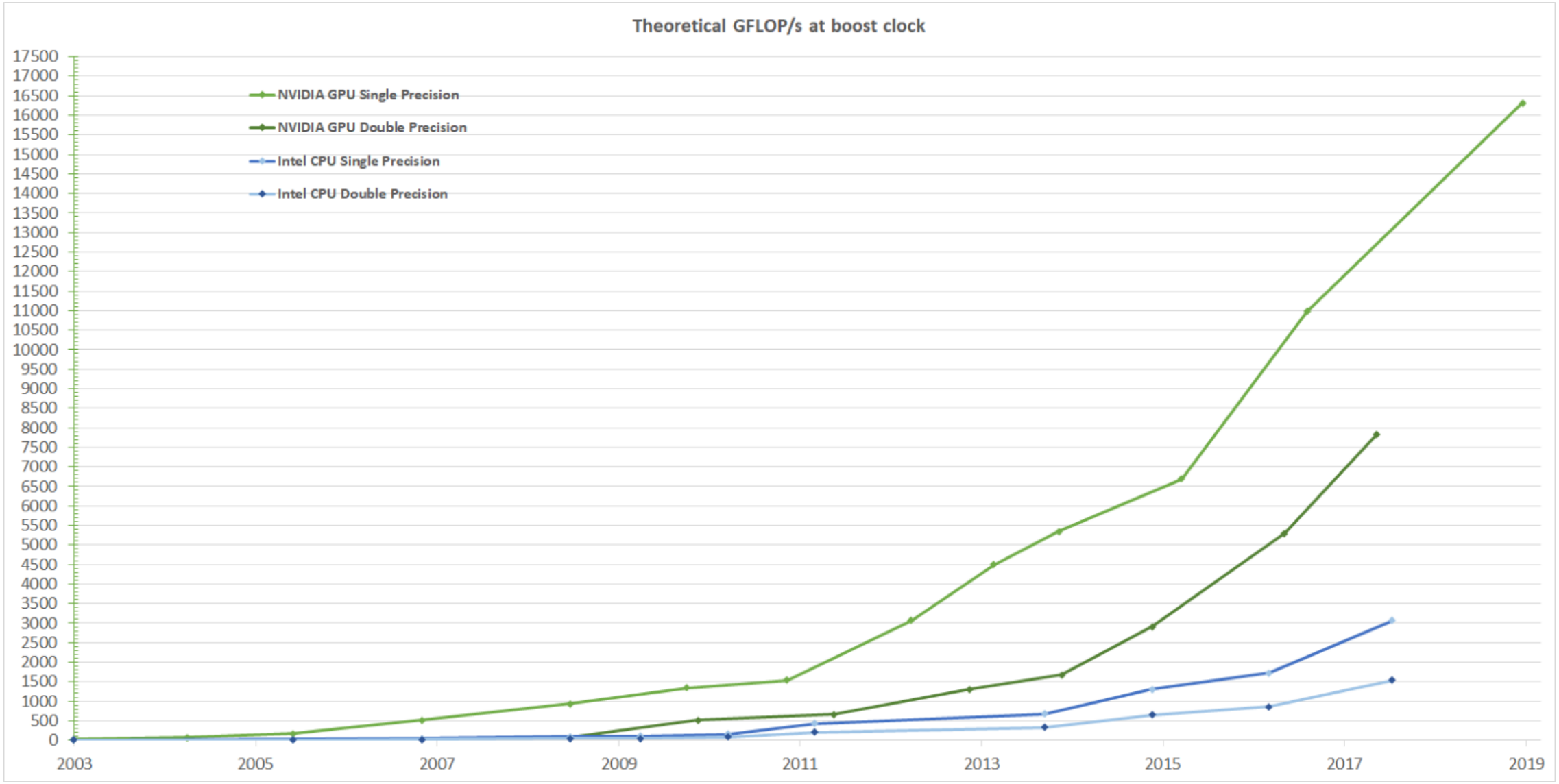
TOP500 November 2020

Green500 Supercomputer

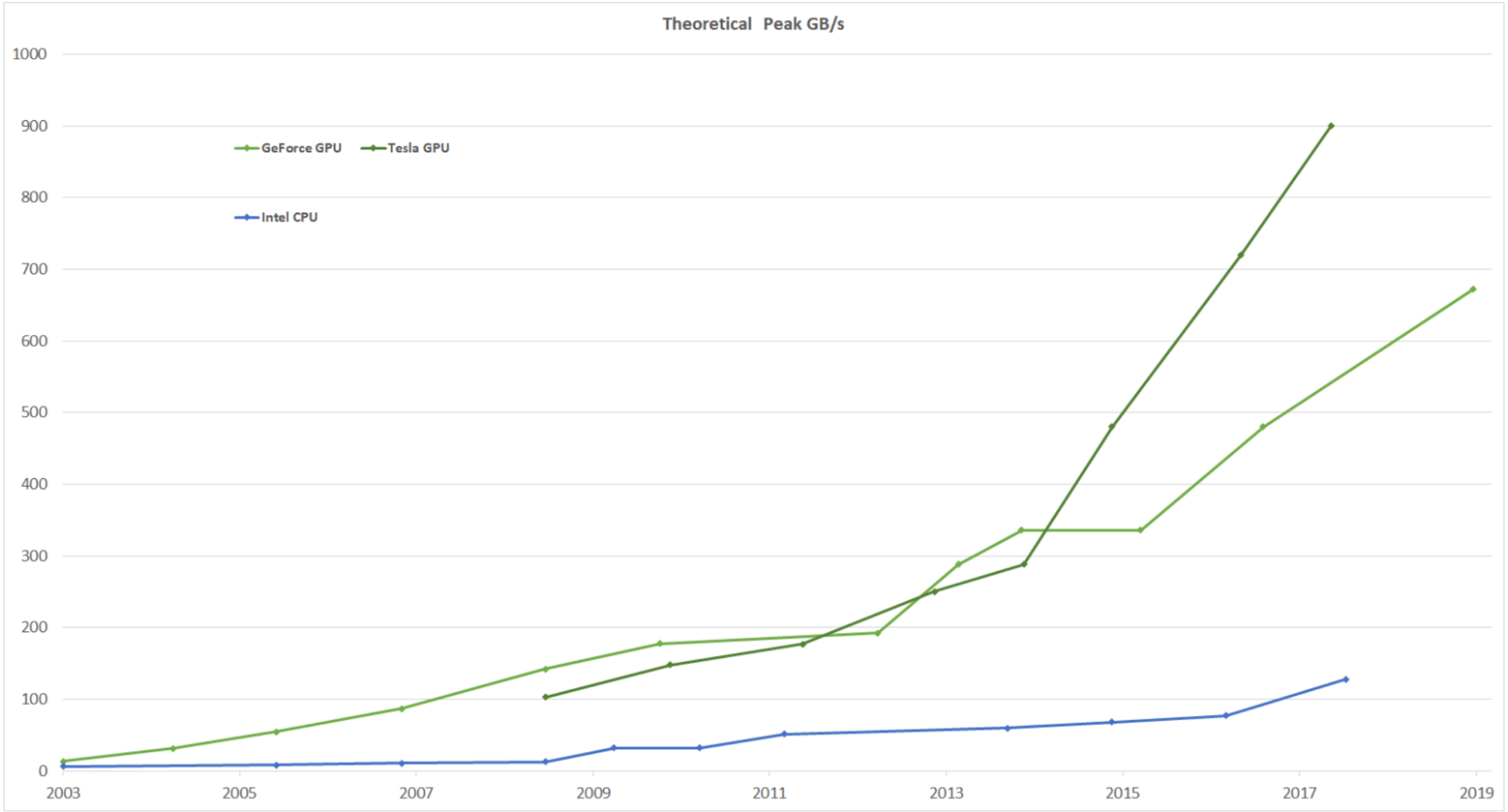
Green500 Rank	GFLOPS/W	Rmax [TFlop/s]	Name Site	Computer
1	26,20	2.356,00	NVIDIA DGX SuperPOD NVIDIA Corporation	NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband
2	26,04	1.652,90	MN-3 Preferred Networks	MN-Core Server, Xeon Platinum 8260M 24C 2.4GHz, Preferred Networks MN-Core, MN-Core DirectConnect
3	25,01	44.120,00	JUWELS Booster Module Forschungszentrum Juelich (FZJ)	Bull Sequana XH2000 , AMD EPYC 7402 24C 2.8GHz, NVIDIA A100, Mellanox HDR InfiniBand/ParTec ParaStation ClusterSuite
4	24,26	2.566,00	Spartan2 Atos	Bull Sequana XH2000 , AMD EPYC 7402 24C 2.8GHz, NVIDIA A100, Mellanox HDR Infiniband
5	23,98	63.460,00	Selene NVIDIA Corporation	NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband
6	16,88	1.999,50	A64FX prototype Fujitsu Numazu Plant	Fujitsu A64FX, Fujitsu A64FX 48C 2GHz, Tofu interconnect D
7	16,28	8.339,00	AiMOS Rensselaer Polytechnic Institute Center for Computational Innovations (CCI)	IBM Power System AC922, IBM POWER9 20C 3.45GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband
8	15,74	35.450,00	HPC5 Eni S.p.A.	PowerEdge C4140, Xeon Gold 6252 24C 2.1GHz, NVIDIA Tesla V100, Mellanox HDR Infiniband
9	15,57	1.464,00	Satori MIT/MGHPCC Holyoke, MA	IBM Power System AC922, IBM POWER9 20C 2.4GHz, Infiniband EDR, NVIDIA Tesla V100 SXM2
10	15,42	442.010,00	Supercomputer Fugaku RIKEN Center for Computational Science	Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D

Green500 November 2020

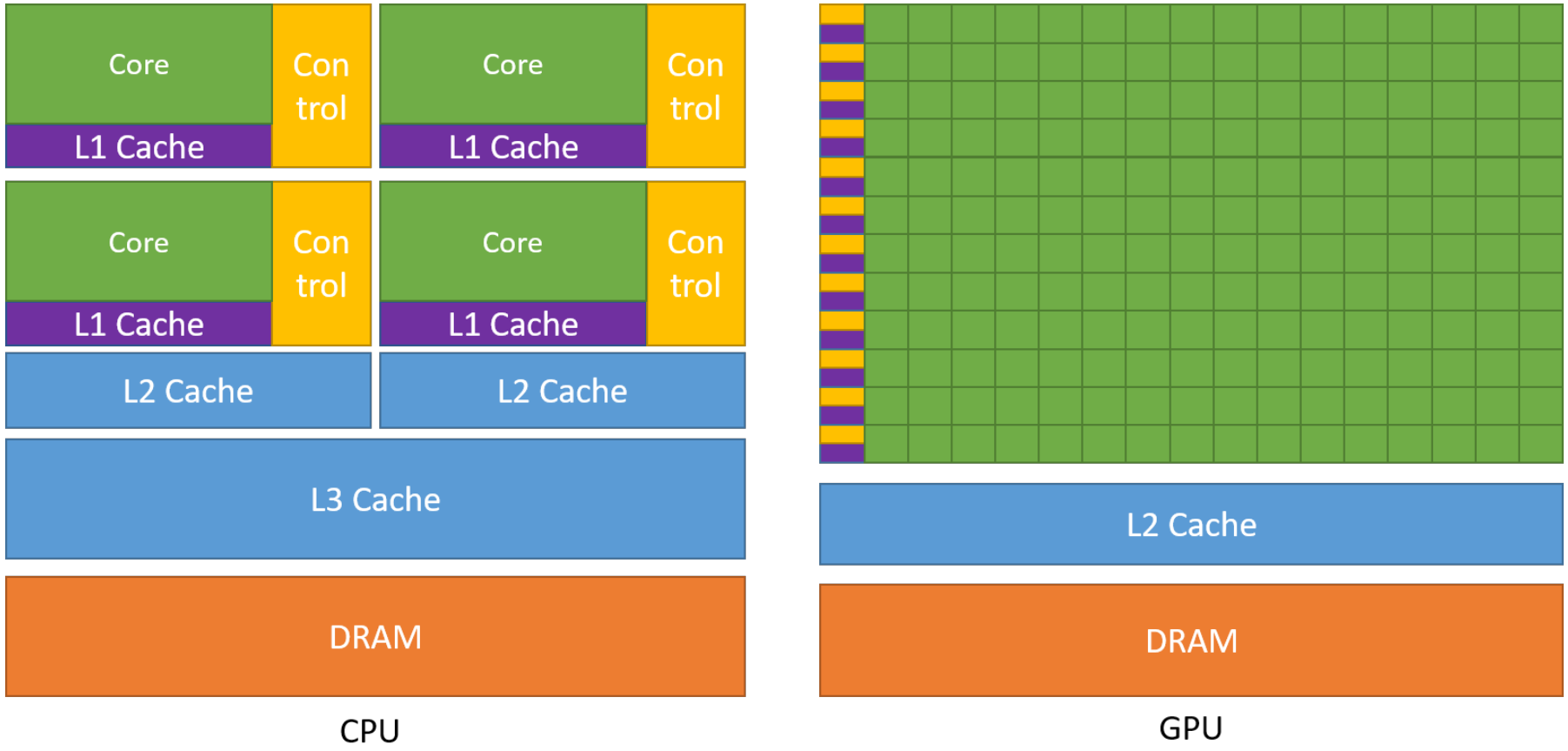
CPU vs. GPU Performance



CPU vs. GPU Memory Bandwidth



CPU vs. GPU Hardware



Summary

- GPU hardware
  - Driven by commodity GPU hardware (gaming, VR/AR)
  - Accelerator for HPC, DL, autonomous driving, …
- GPU programming
  - Early days: fixed-function graphics pipelines
  - Nowadays: fully programmable cores