King Abdullah University of Science and Technology

KAUST

CS 380 - GPU and GPGPU Programming
Lecture 2: Introduction, Pt. 2

Markus Hadwiger, KAUST

Reading Assignment #1 (until Sep 8)

Read (required):

- Programming Mass. Parallel Proc. book, 4th ed., Chapter 1 (*Introduction*)
- Programming Mass. Parallel Proc. book, 2nd ed., Chapter 2 (*History of GPU Computing*)
- OpenGL Shading Language (orange) book, Chapter 1 (*Review of OpenGL Basics*)

Read (optional):

- OpenGL Shading Language 4.6 (current: Aug 14, 2023) specification: Chapter 2

  `https://www.khronos.org/registry/OpenGL/specs/gl/GLSLangSpec.4.60.pdf`

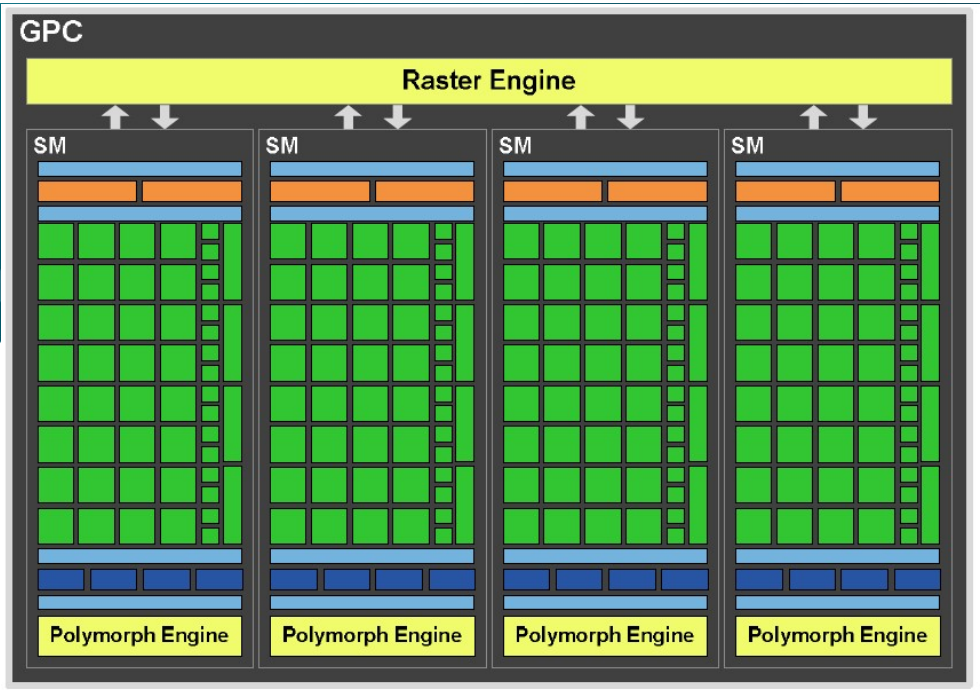
- Download OpenGL 4.6 (current: May 5, 2022) specification

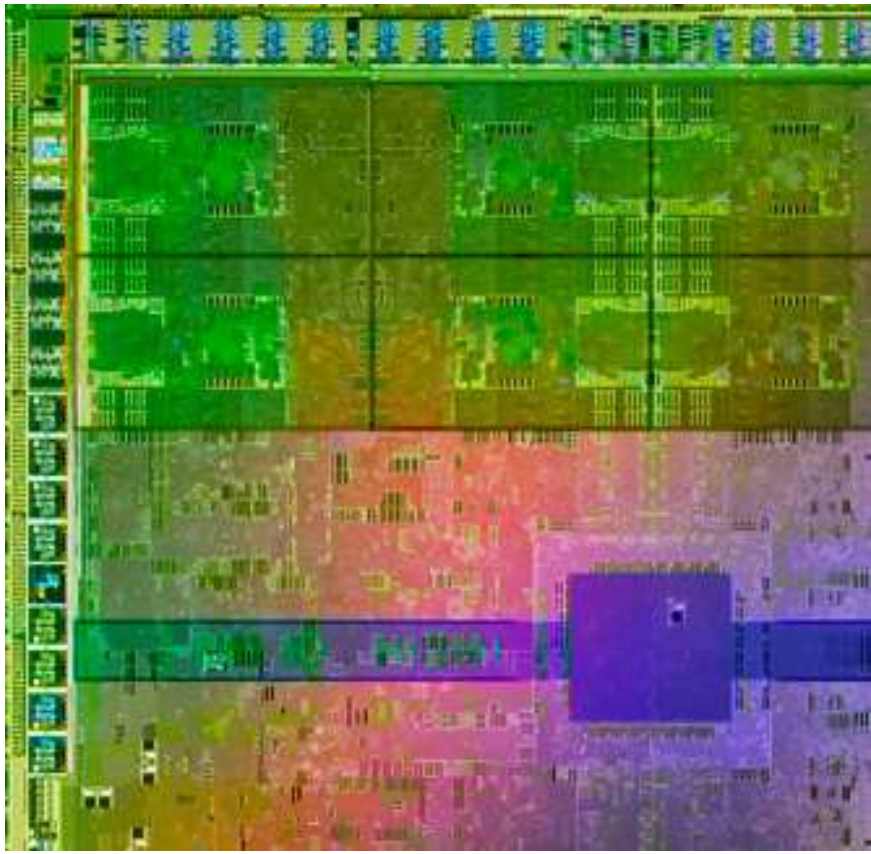
  `https://www.khronos.org/registry/OpenGL/specs/gl/glspec46.core.pdf`

Syllabus (1)

GPU Basics and Architecture
(~September, early October)

- Introduction
- **GPU architecture**
- How compute/shader cores work
- GPU shading and GPU compute APIs
  - General concepts and overview
  - Learn syntax details on your own !
    - CUDA book
    - GLSL book
    - Vulkan tutorial
    - online resources, ...

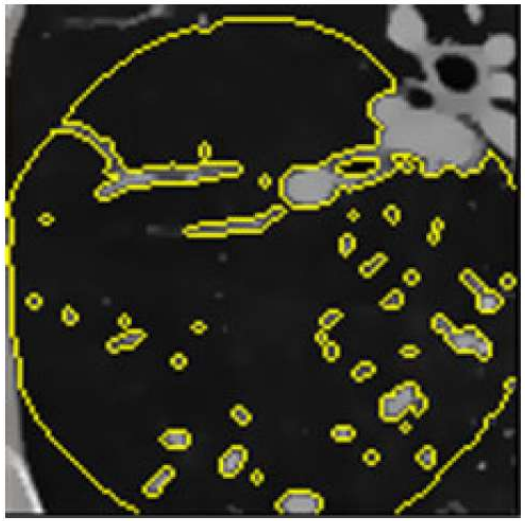




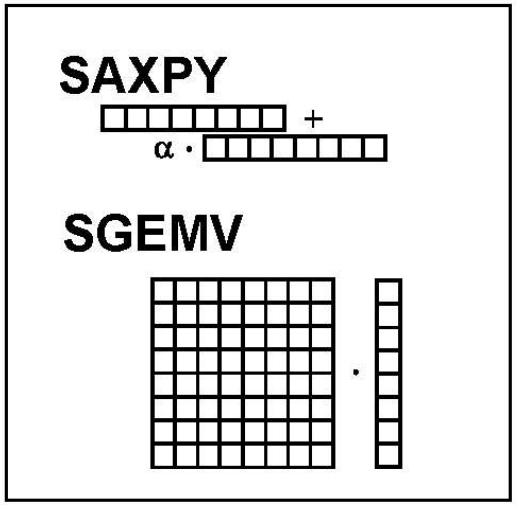
Syllabus (2)

More GPU Computing (~October)

- GPGPU, important parallel programming concepts
- CUDA memory access
- Reduction, scan
- Linear algebra on GPUs
- Deep learning on GPUs
- Combining graphics and compute
  - Display the results of computations
  - Interactive systems (fluid flow, ...)



segmentation

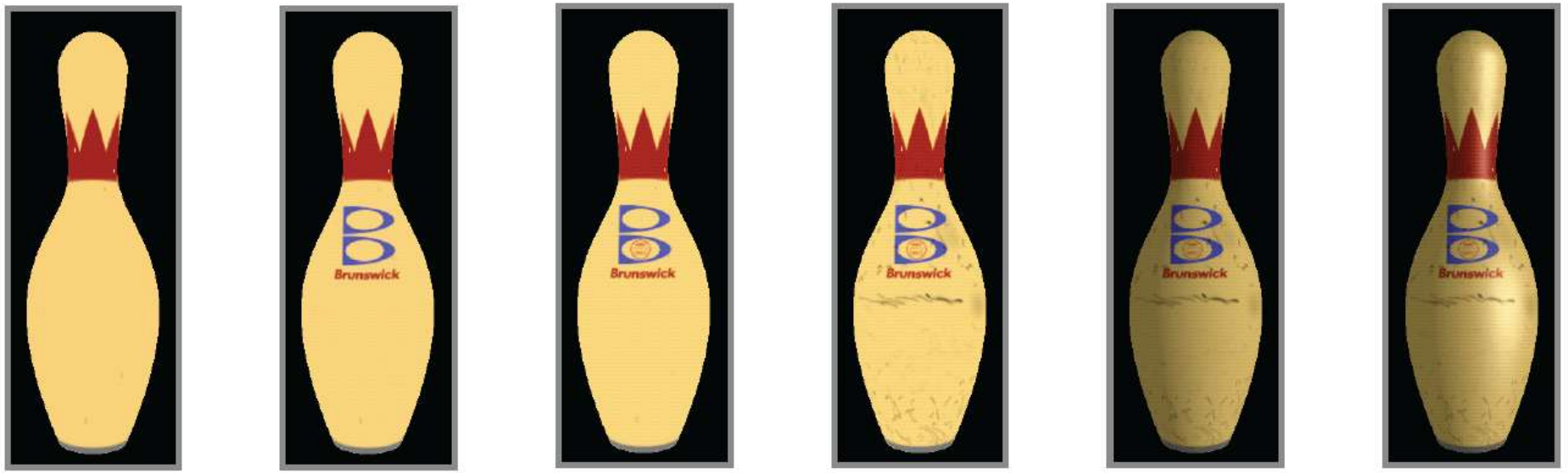


linear algebra

Syllabus (3)

GPU Graphics (~November)

- GPU (virtual) texturing, filtering
- GPU (texture) memory management
- Neural rendering, neural shading
- Modern game engine technologies

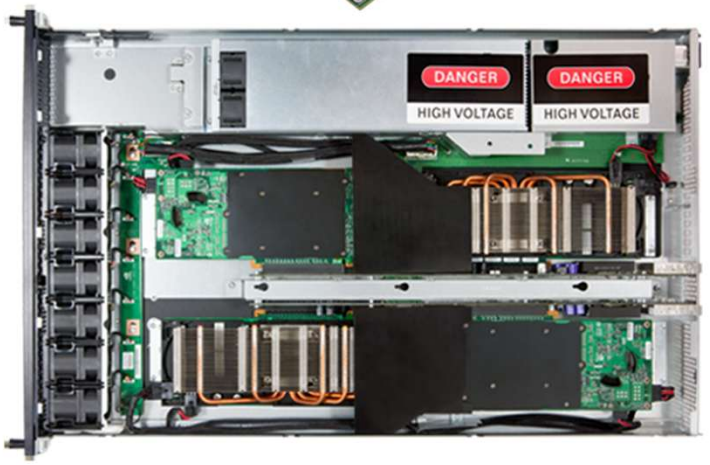


What are GPUs?

Graphics Processing Units

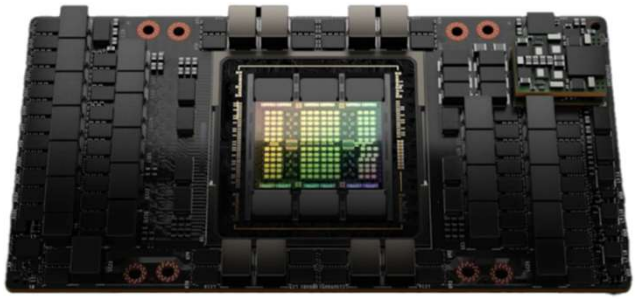
But evolved toward

- Very flexible, massively parallel floating point co-processors
- But not entirely programmable!
- Fixed-function parts have definite advantages (e.g., texture filtering, z-buffering)

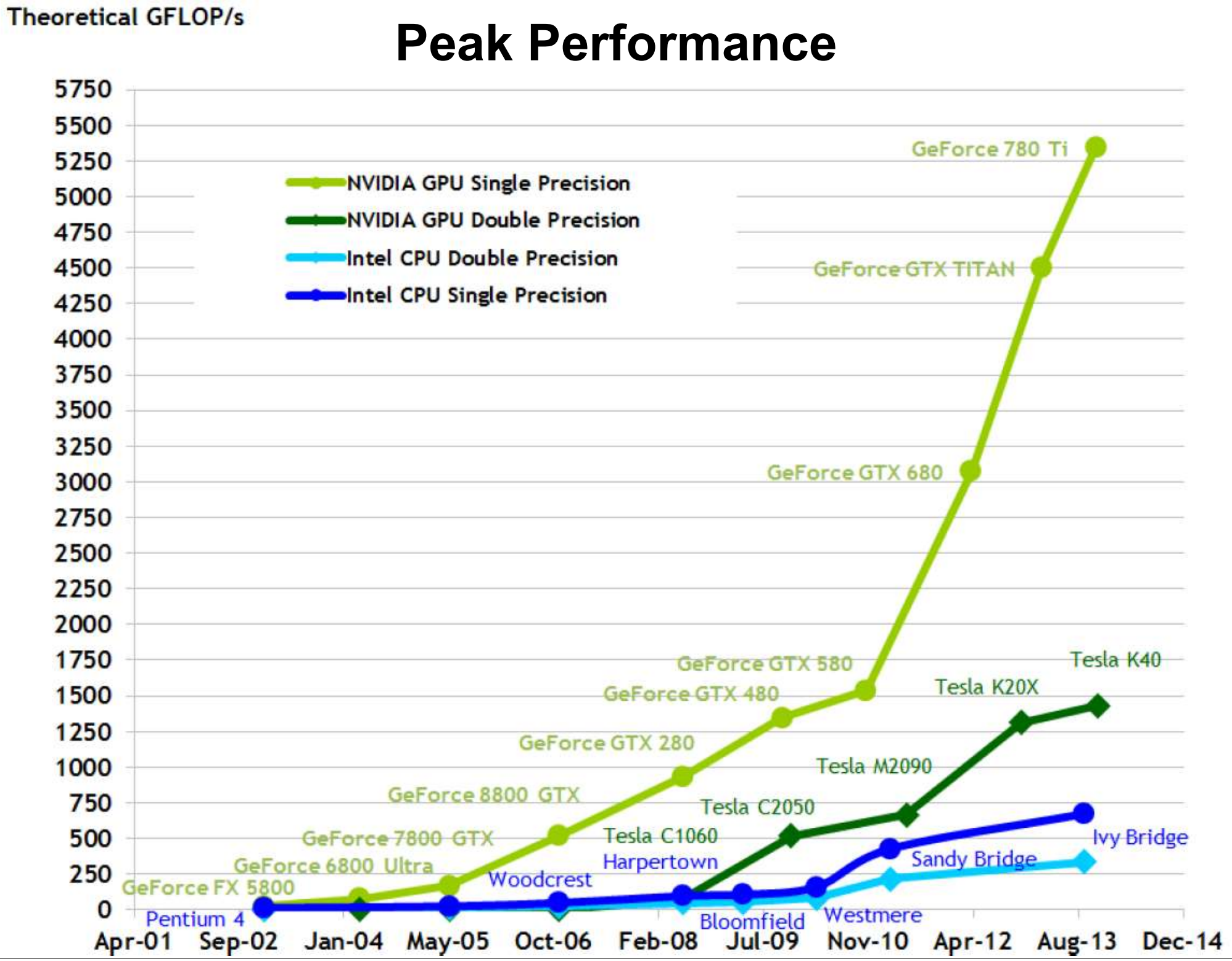


We will cover both perspectives

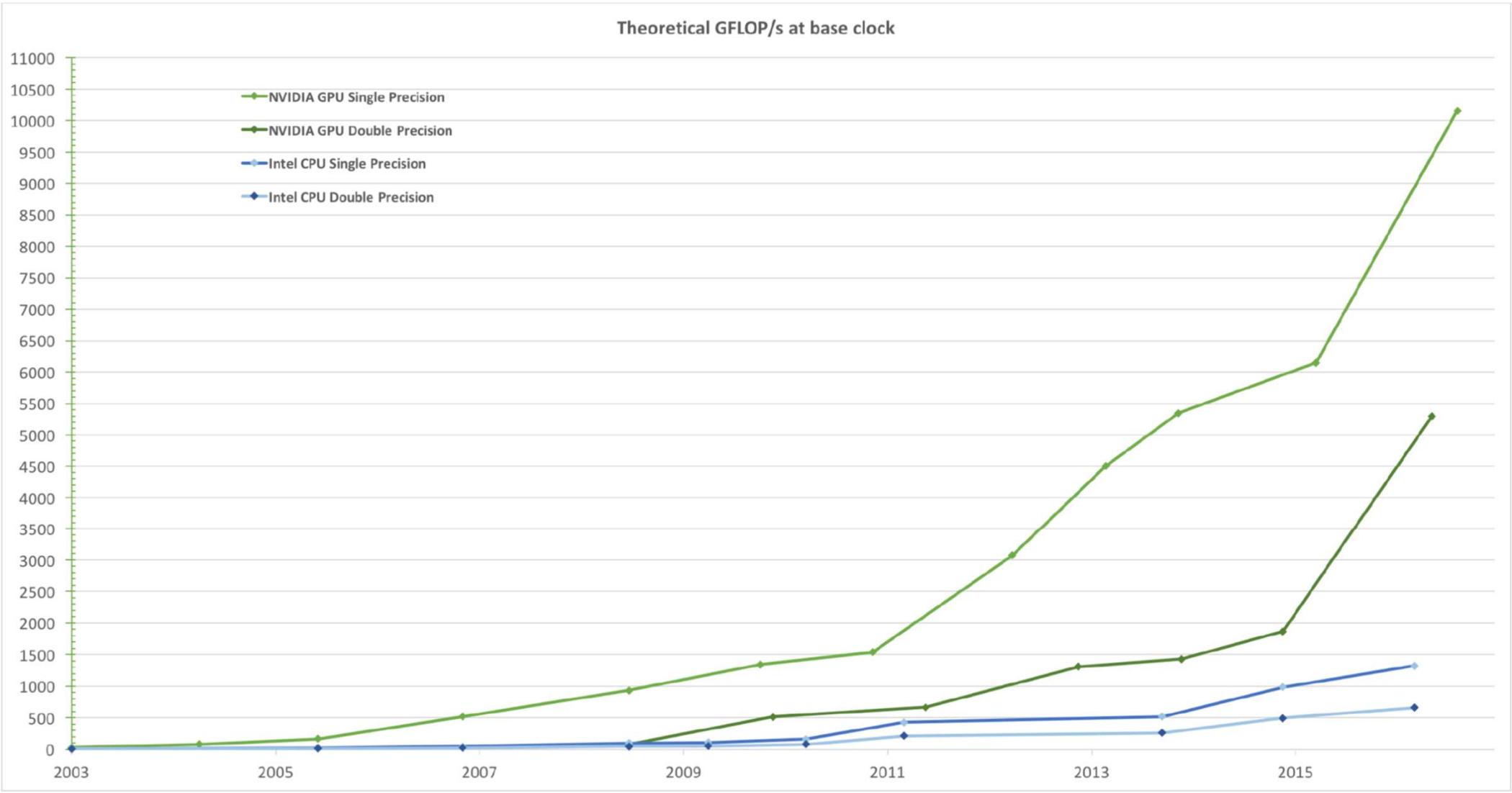
- GPUs for graphics
- GPU for compute (GPGPU – general purpose computation on GPU)



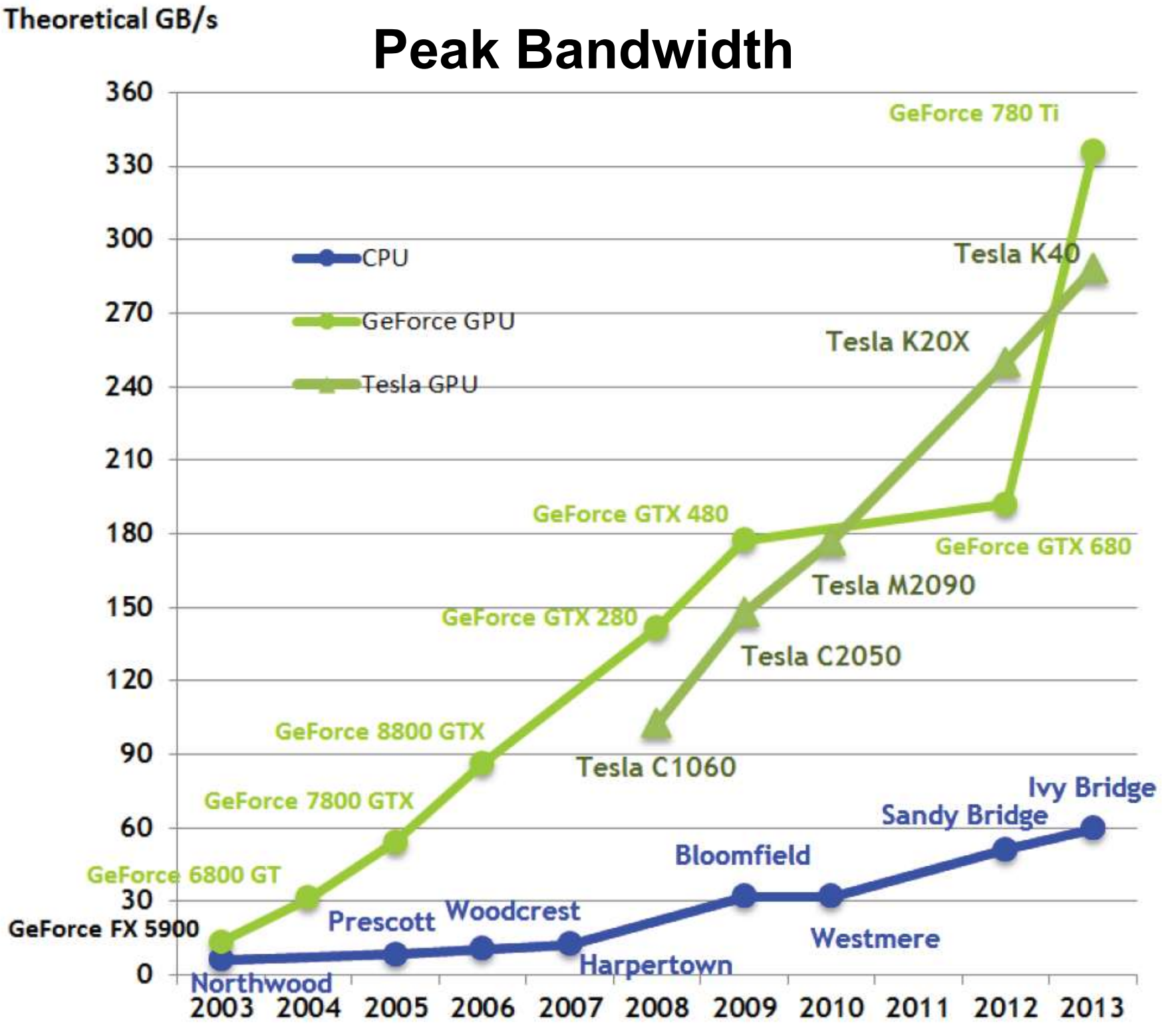
Peak Performance

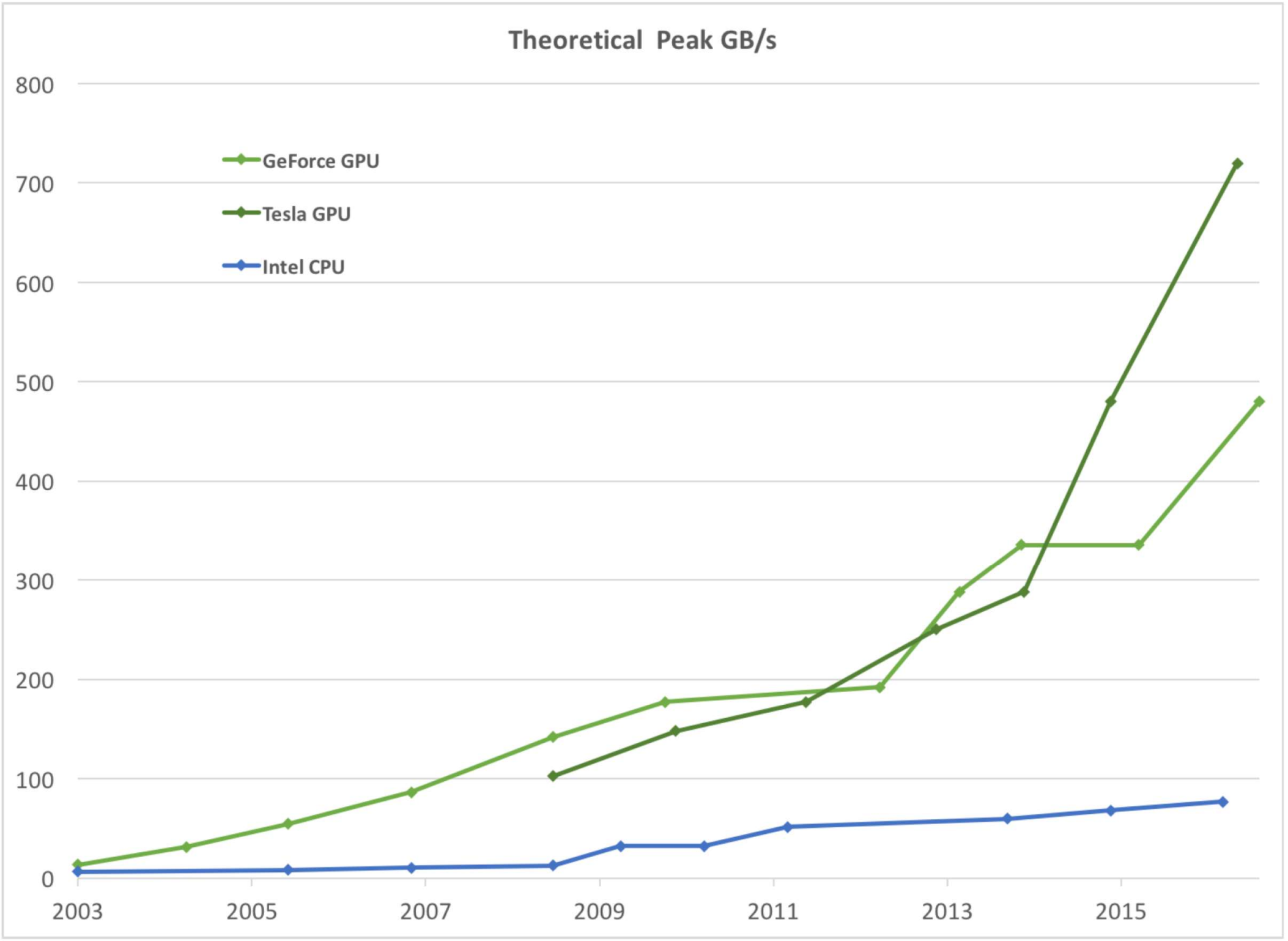


Peak Performance

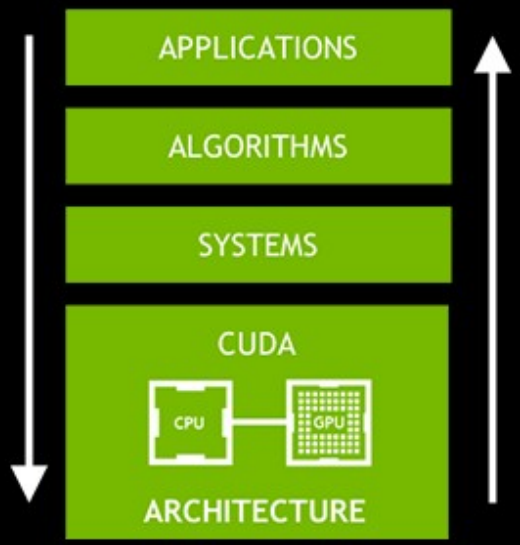


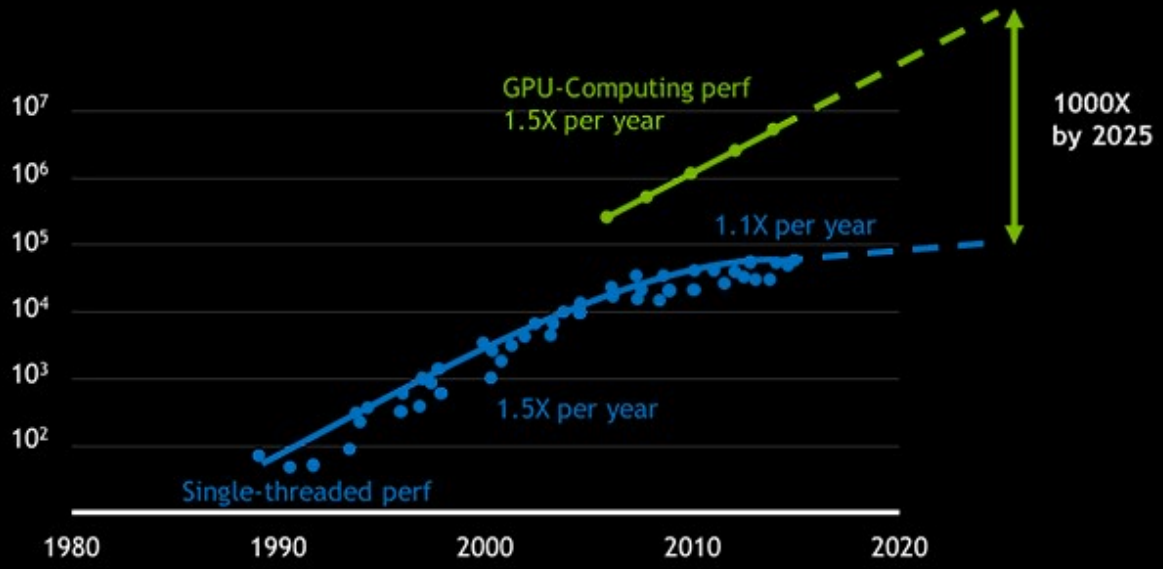
Peak Bandwidth




Theoretical Peak GB/s
800
700
600
500
400
300
200
100
0
2003
2005
2007
2009
2011
2013
2015
GeForce GPU
Tesla GPU
Intel CPU

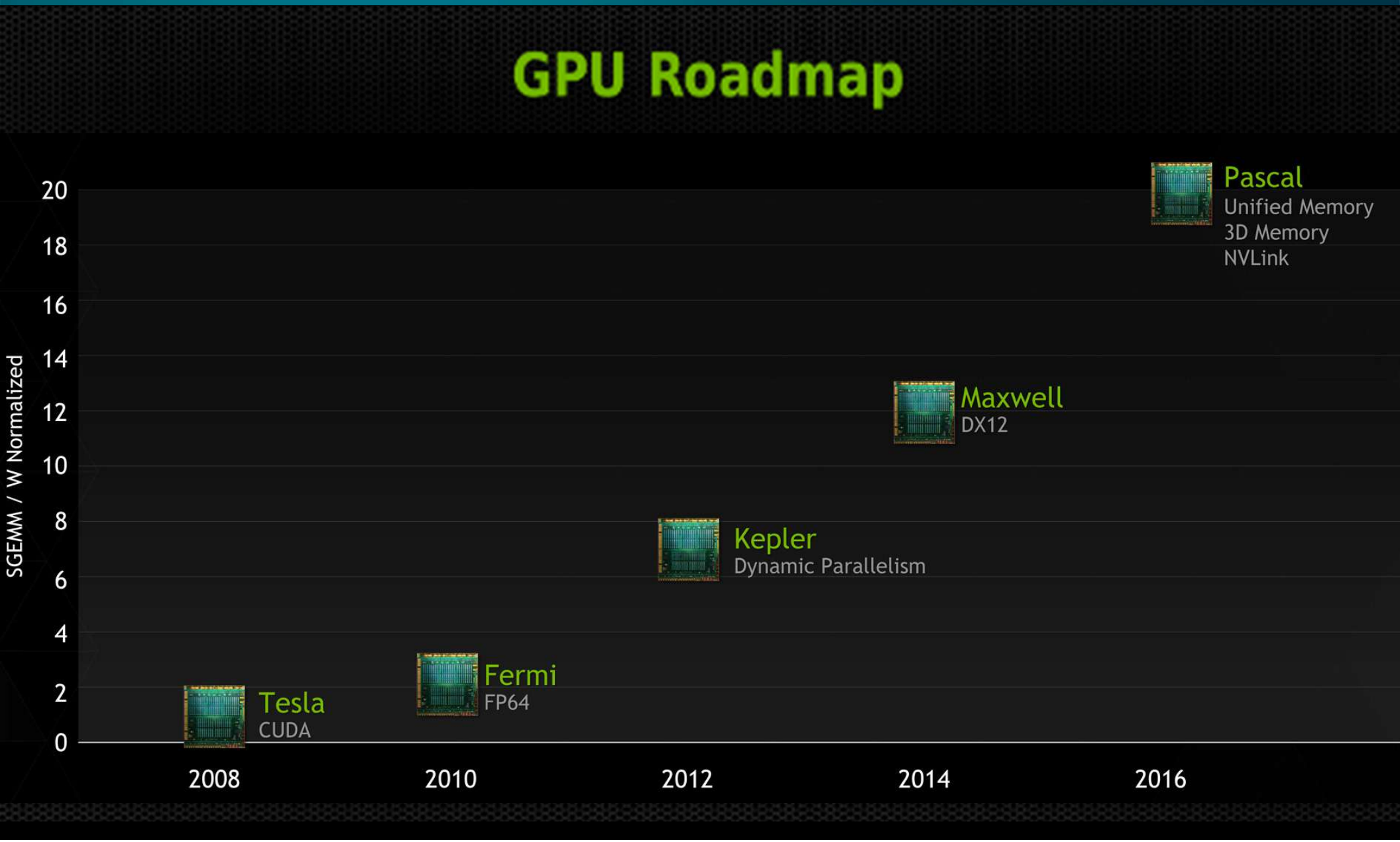
RISE OF GPU COMPUTING



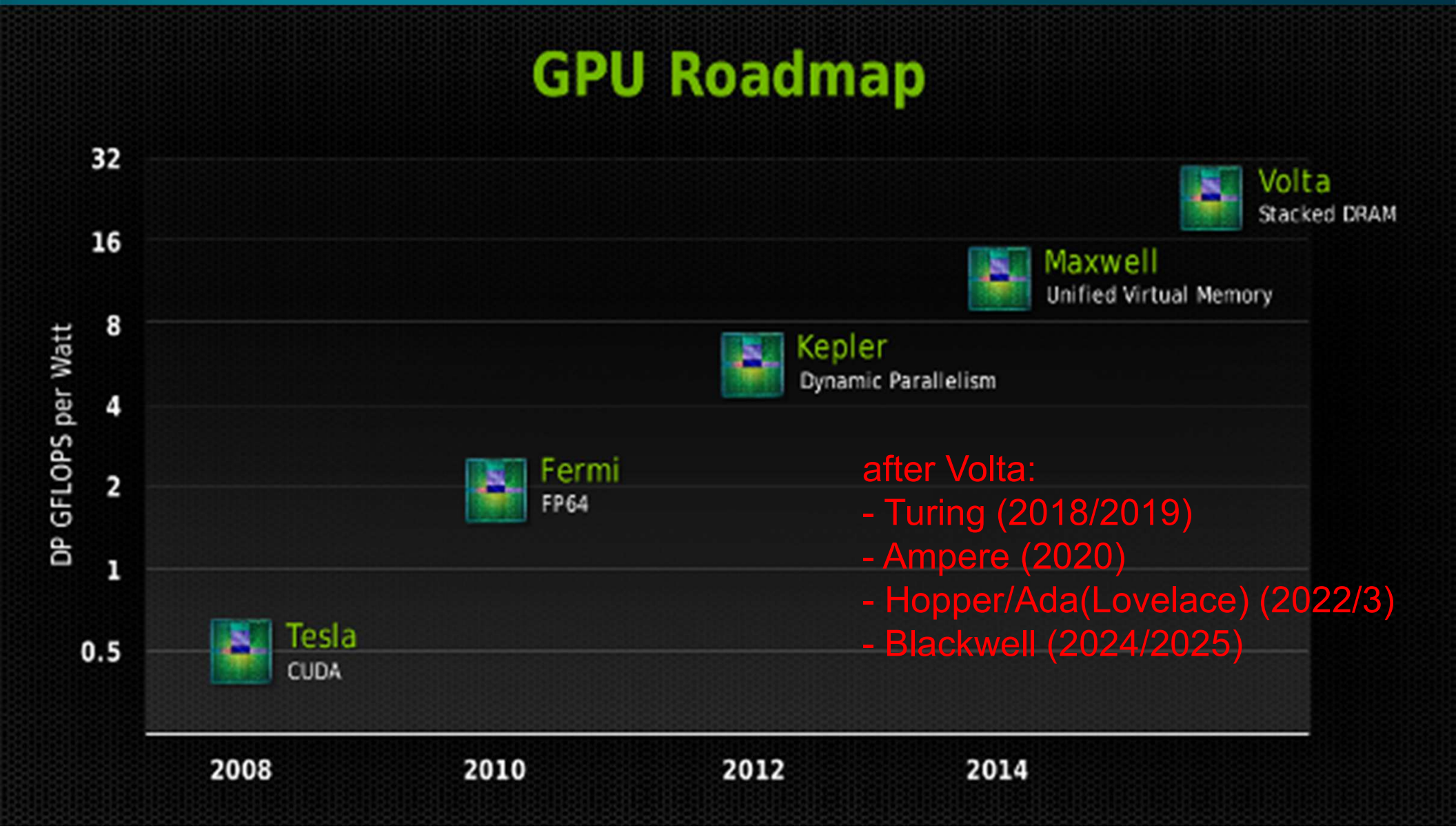


Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

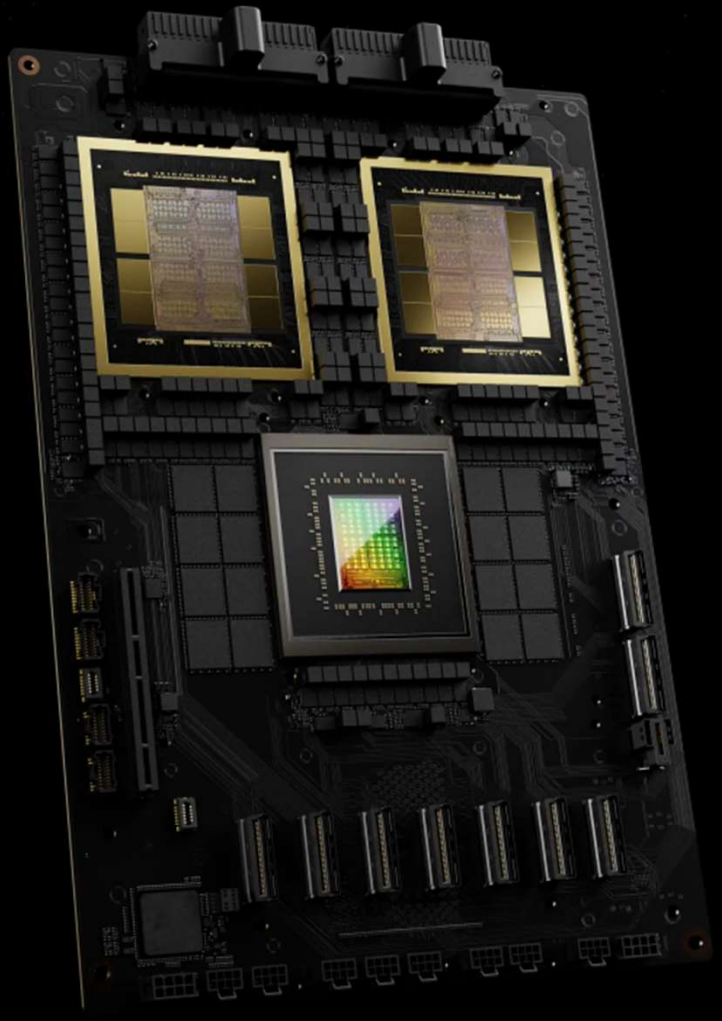
GPU Architectures Over the Years

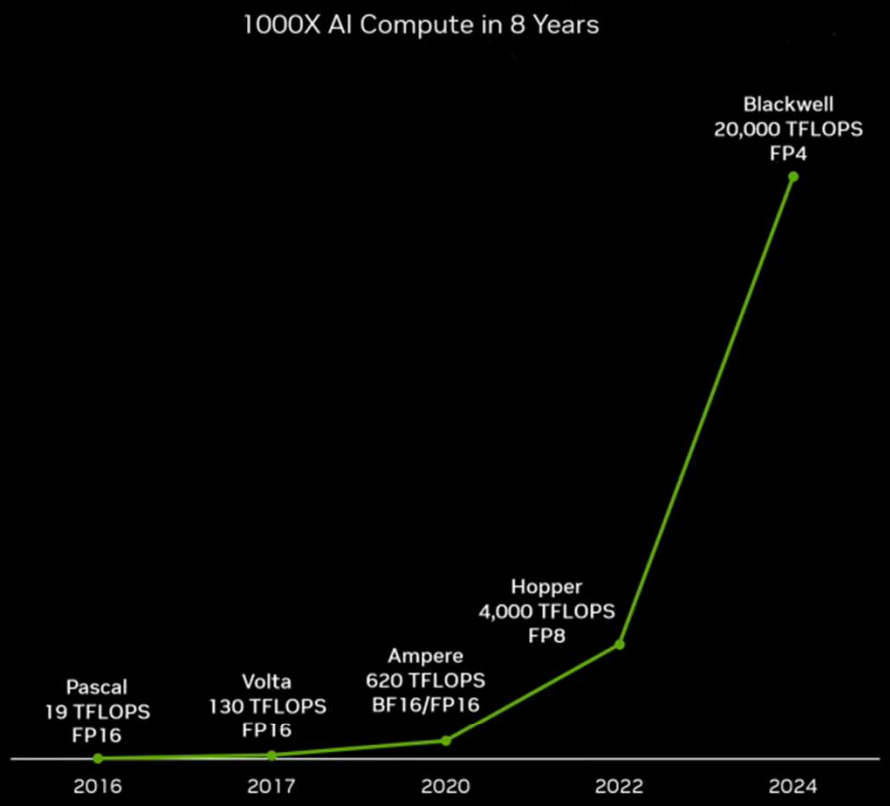


GPU Architectures Over the Years



GPU Architectures Over the Years

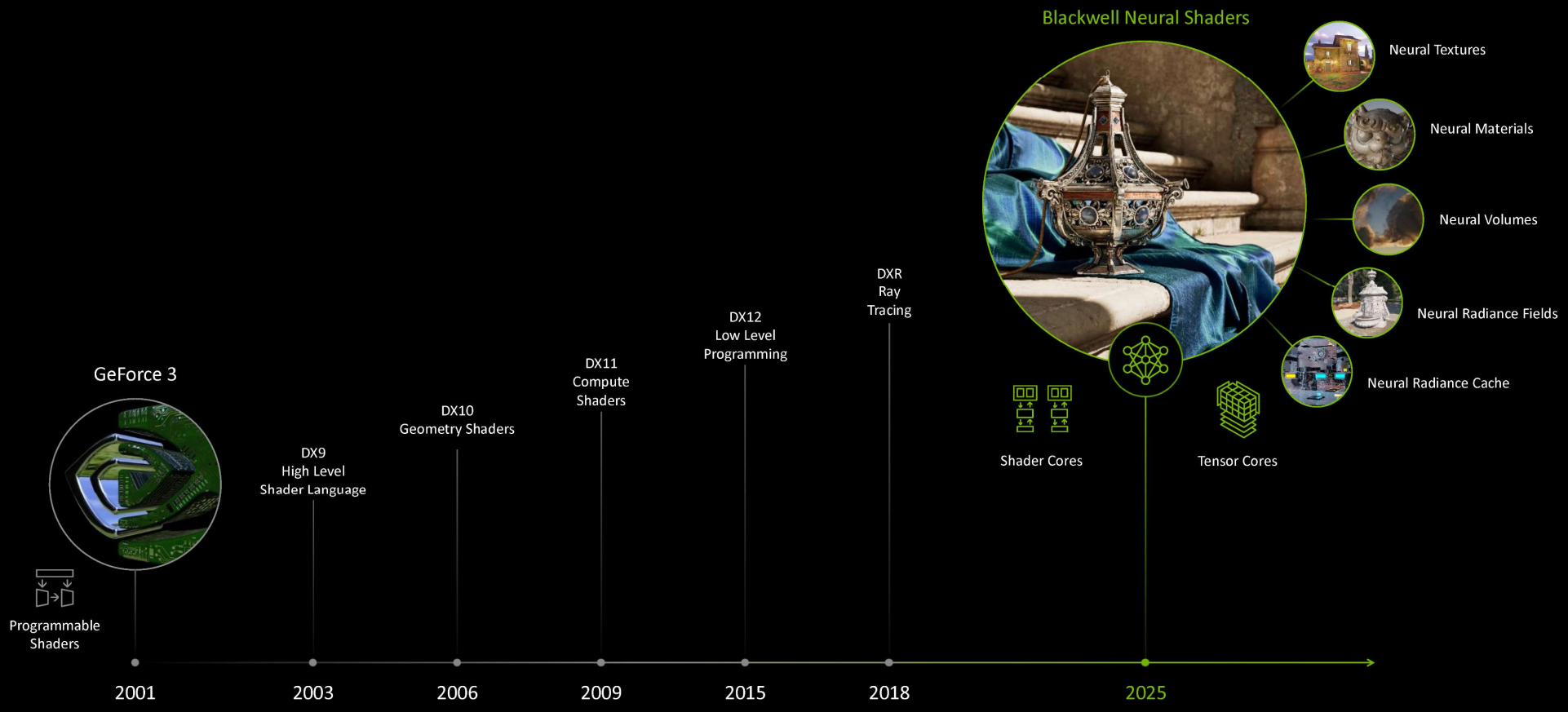




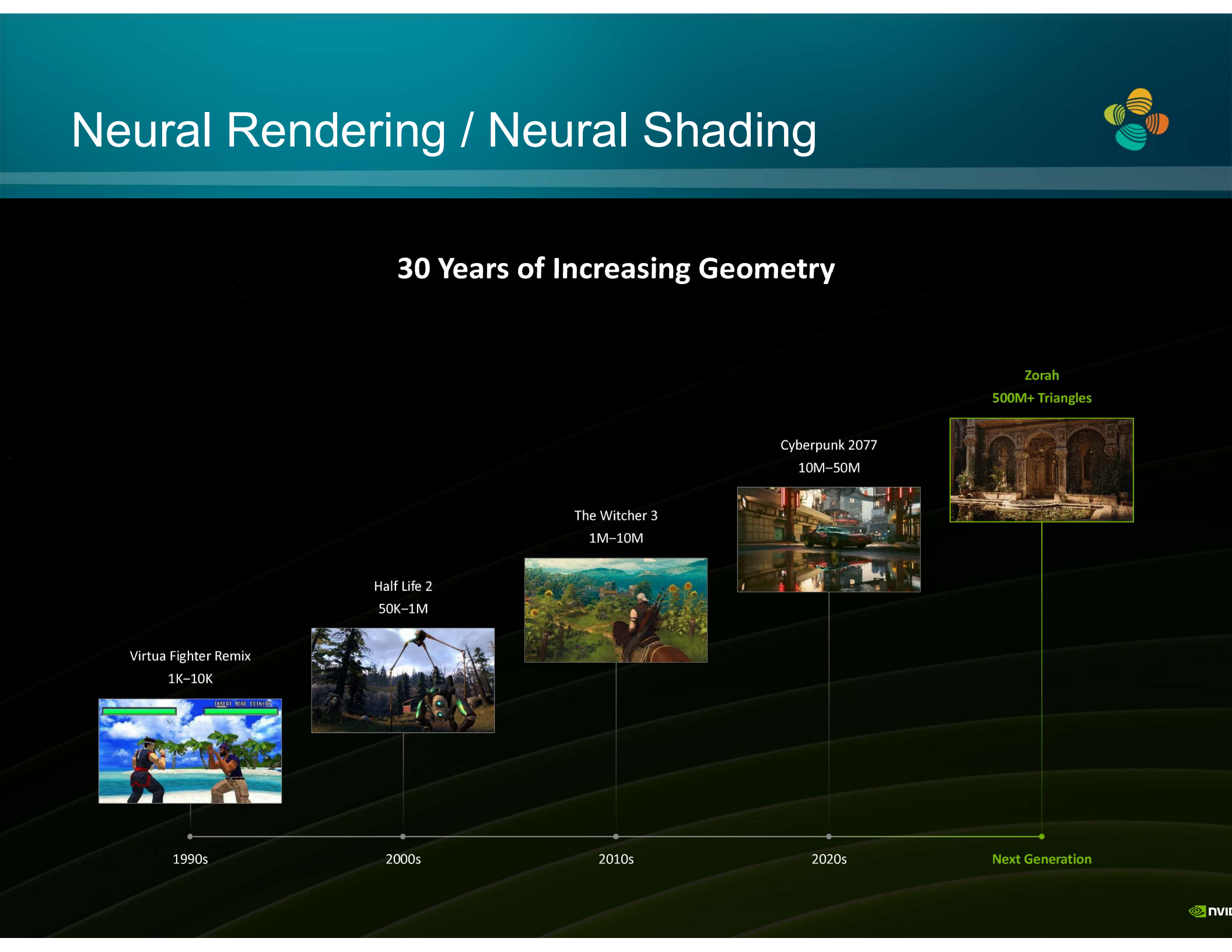
Neural Rendering / Neural Shading

Blackwell Brings AI to Shaders

Unlocking the next two decades of graphics innovation



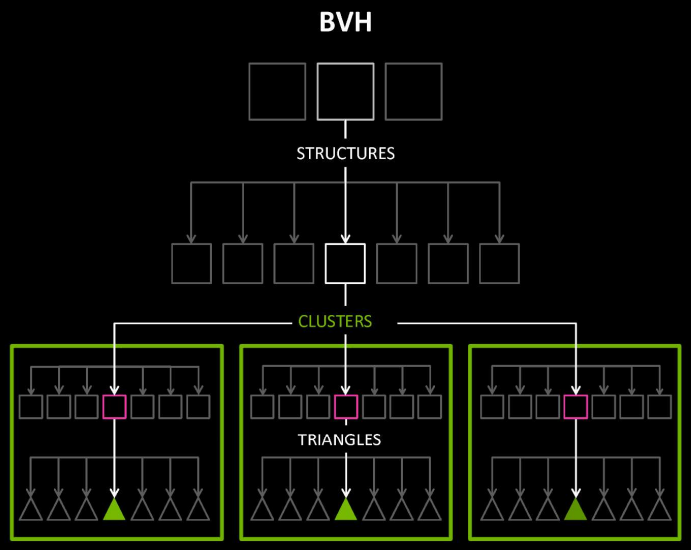
Neural Rendering / Neural Shading



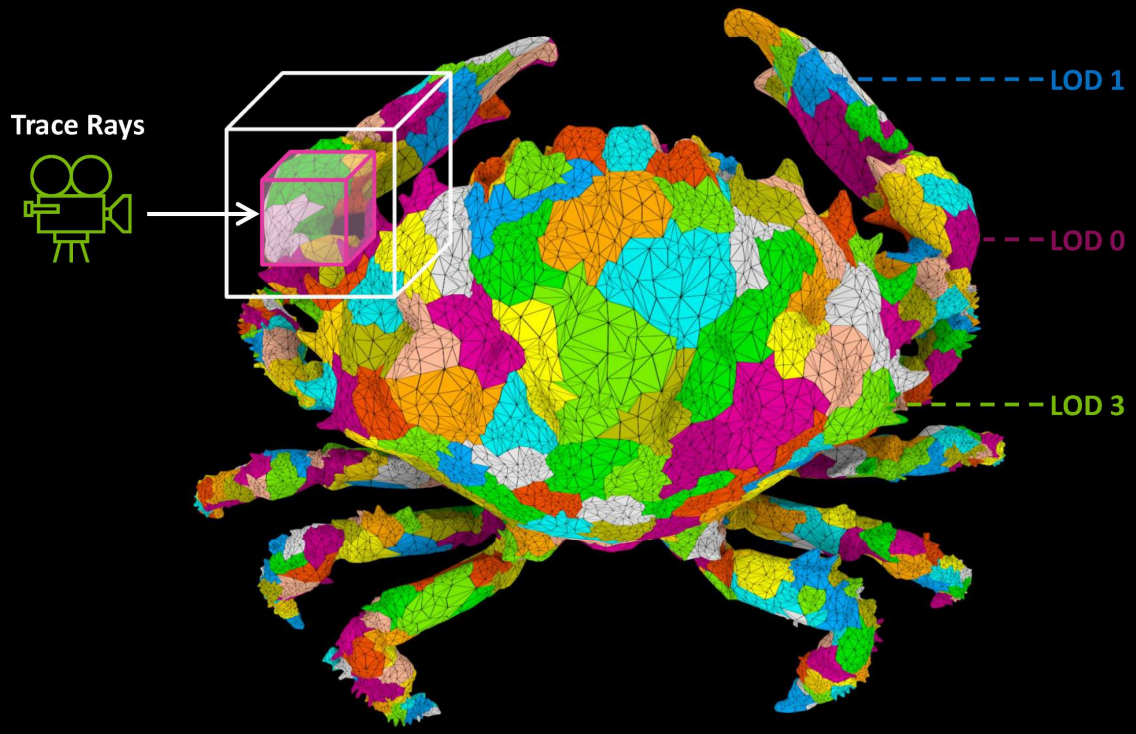
Neural Rendering / Neural Shading

RTX Mega Geometry

Accelerate BVH updates for cluster-based systems like Nanite



Compress and Cache Clusters Over Many Frames



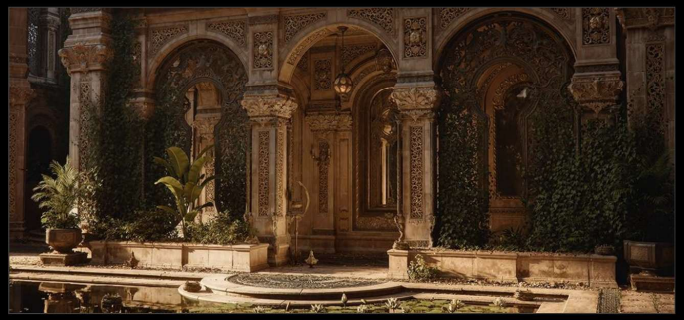
Mesh Composed of Clusters at Different LOD Levels

Neural Rendering / Neural Shading

NVIDIA RTX Kit

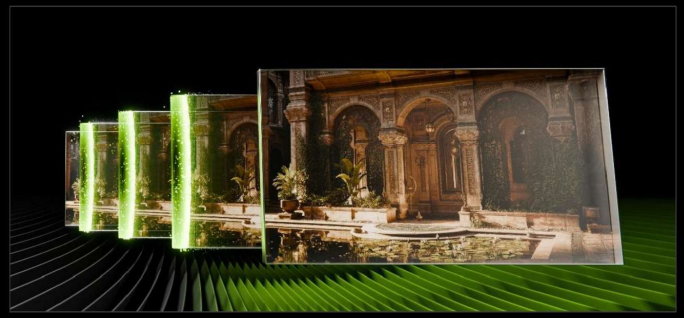
Delivering neural rendering to developers

RTX Neural Shaders



RTX Mega Geometry

RTX Hair & Skin



DLSS 4

Reflex 2

RTX Remix

NVIDIA Architectures (since first CUDA GPU)

Tesla [CC 1.x]: 2007-2009

- G80, G9x: 2007 (Geforce 8800, ...)
  GT200: 2008/2009 (GTX 280, ...)

Fermi [CC 2.x]: 2010 (2011, 2012, 2013, …)

- GF100, ... (GTX 480, ...)
  GF104, ... (GTX 460, ...)
  GF110, ... (GTX 580, ...)

Kepler [CC 3.x]: 2012 (2013, 2014, 2016, …)

- GK104, ... (GTX 680, ...)
  GK110, ... (GTX 780, GTX Titan, ...)

Maxwell [CC 5.x]: 2015

- GM107, ... (GTX 750Ti, ...); [Nintendo Switch]
  GM204, ... (GTX 980, Titan X, ...)

Pascal [CC 6.x]: 2016 (2017, 2018, 2021, 2022, …)

- GP100 (Tesla P100, ...)
- GP10x: x=2,4,6,7,8, ...
  (GTX 1060, 1070, 1080, Titan X *Pascal*, Titan Xp, ...)

Volta [CC 7.0, 7.2]: 2017/2018

- GV100, ...
  (Tesla V100, Titan V, Quadro GV100, ...)

Turing [CC 7.5]: 2018/2019

- TU102, TU104, TU106, TU116, TU117, ...
  (Titan RTX, RTX 2070, 2080 (Ti), GTX 1650, 1660, ...)

Ampere [CC 8.0, 8.6, 8.7]: 2020

- GA100, GA102, GA104, GA106, ...; [Nintendo Switch 2]
  (A100, RTX 3070, 3080, 3090 (Ti), RTX A6000, ...)

Hopper [CC 9.0], Ada Lovelace [CC 8.9]: 2022/23

- GH100, AD102, AD103, AD104, ...
  (H100, L40, RTX 4080 (12/16 GB), 4090, RTX 6000, ...)

Blackwell [CC 10.0, 10.1, 10.3, 12.0, 12.1] : 2024/2025

- GB100/102, GB200/202/203/205/206/207, ...
  (RTX 5080/5090, GB200 NVL72, HGX B100/200, ...)

see `https://en.wikipedia.org/wiki/List_of_Nvidia_graphics_processing_units`
and `https://en.wikipedia.org/wiki/CUDA`

Thank you.