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

Markus Hadwiger, KAUST

Reading Assignment #1 (until Sep 4)

Read (required):

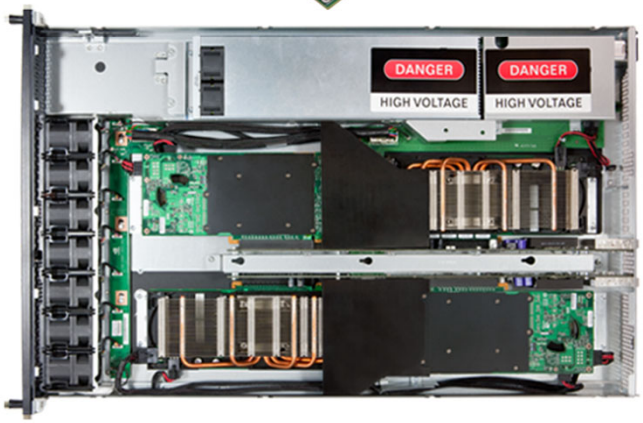
- Orange book, chapter 1 (*Review of OpenGL Basics*)
- Orange book, chapter 2 (*Basics*)

What are GPUs?

Graphics Processing Units

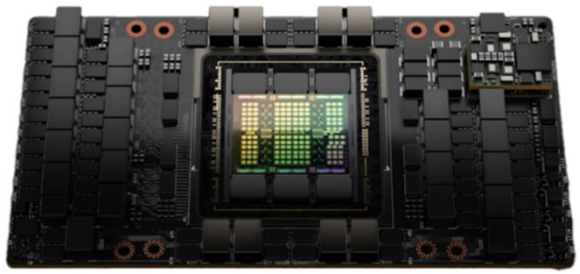
But evolved toward

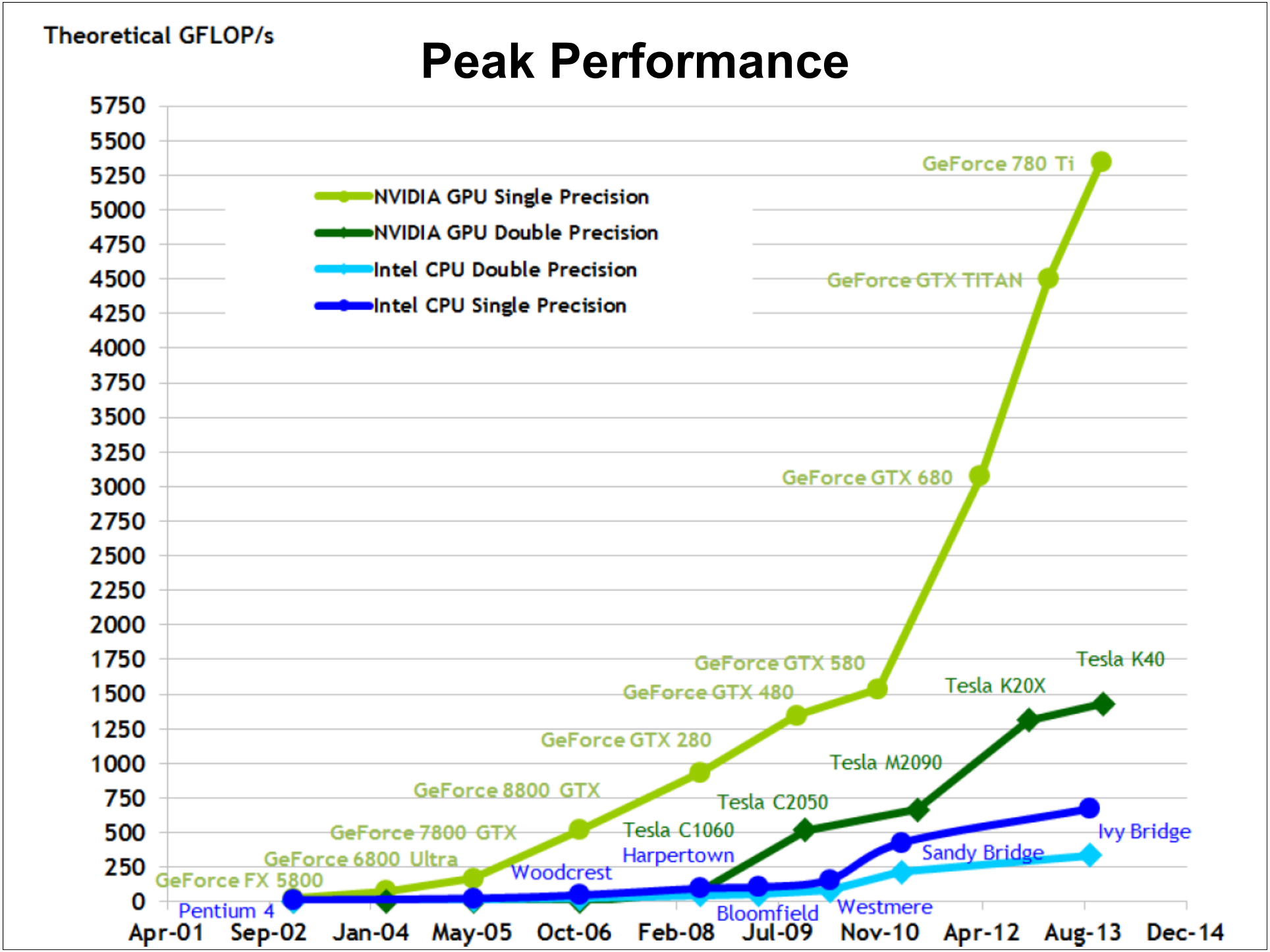
- Very flexible, massively parallel floating point co-processors
- But not entirely programmable!
- Fixed-function parts have definite advantages (e.g., texture filtering, z-buffering)



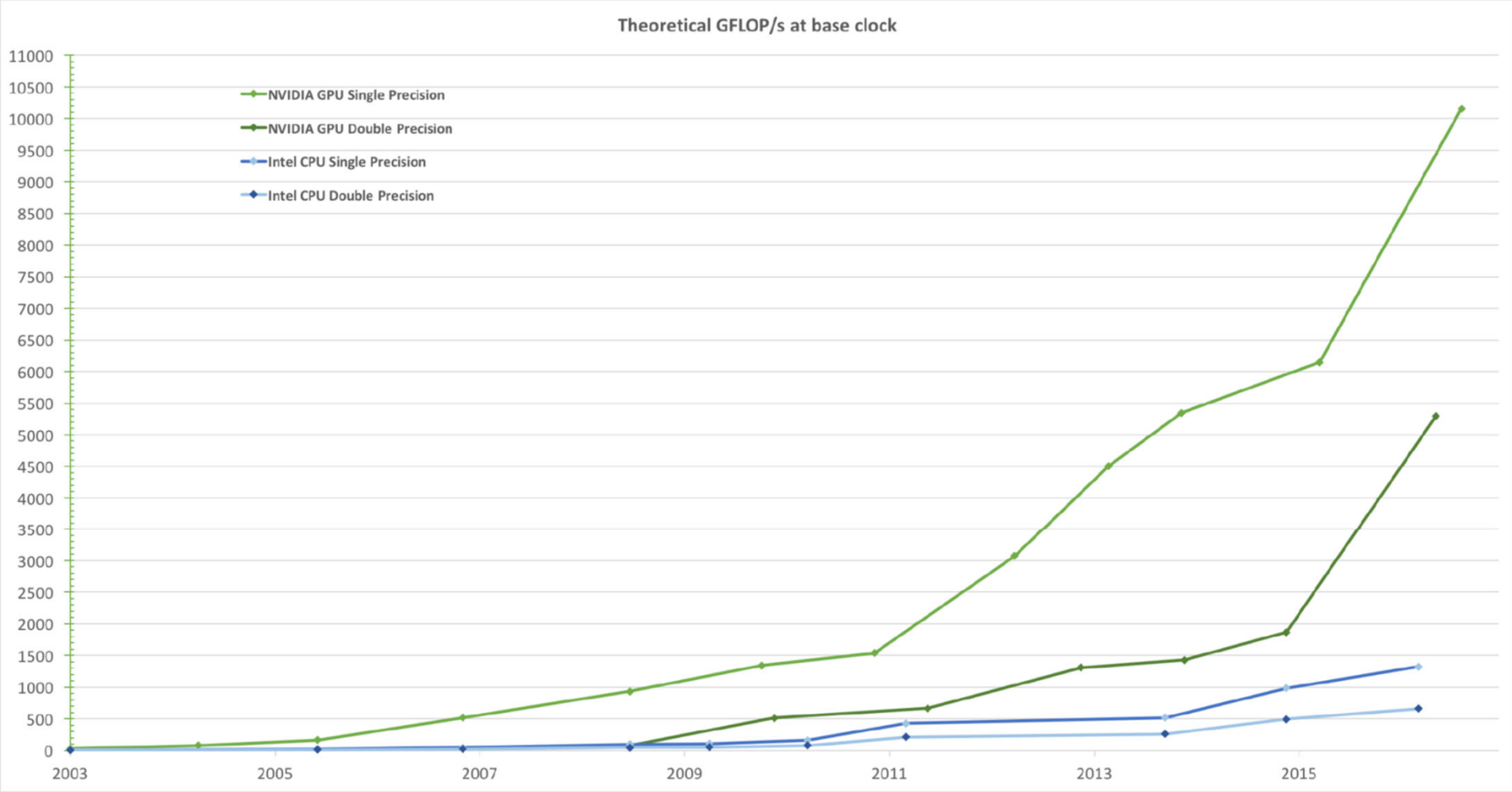
We will cover both perspectives

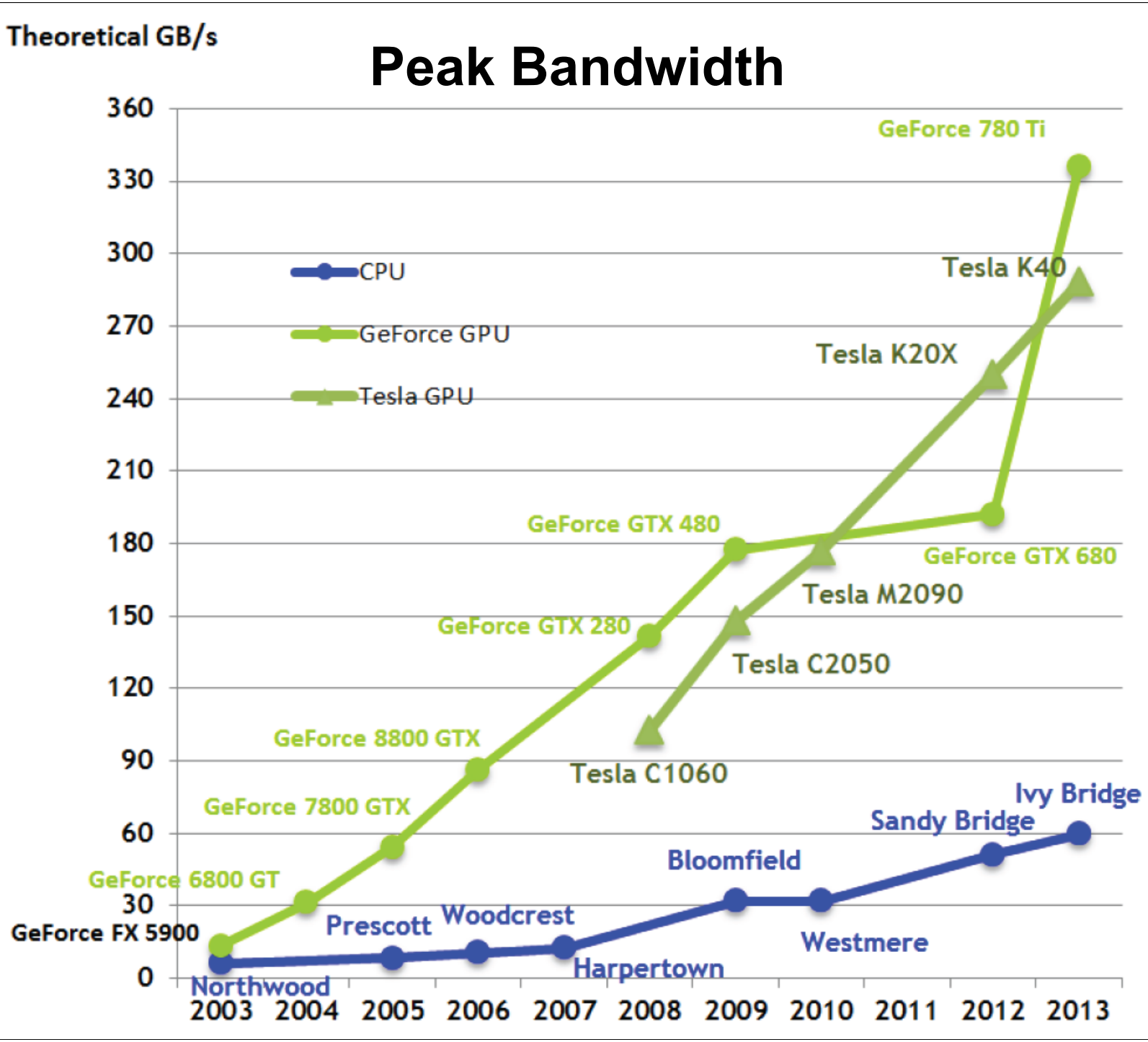
- GPUs for graphics
- GPU computing (GPGPU – general purpose computation on GPU)

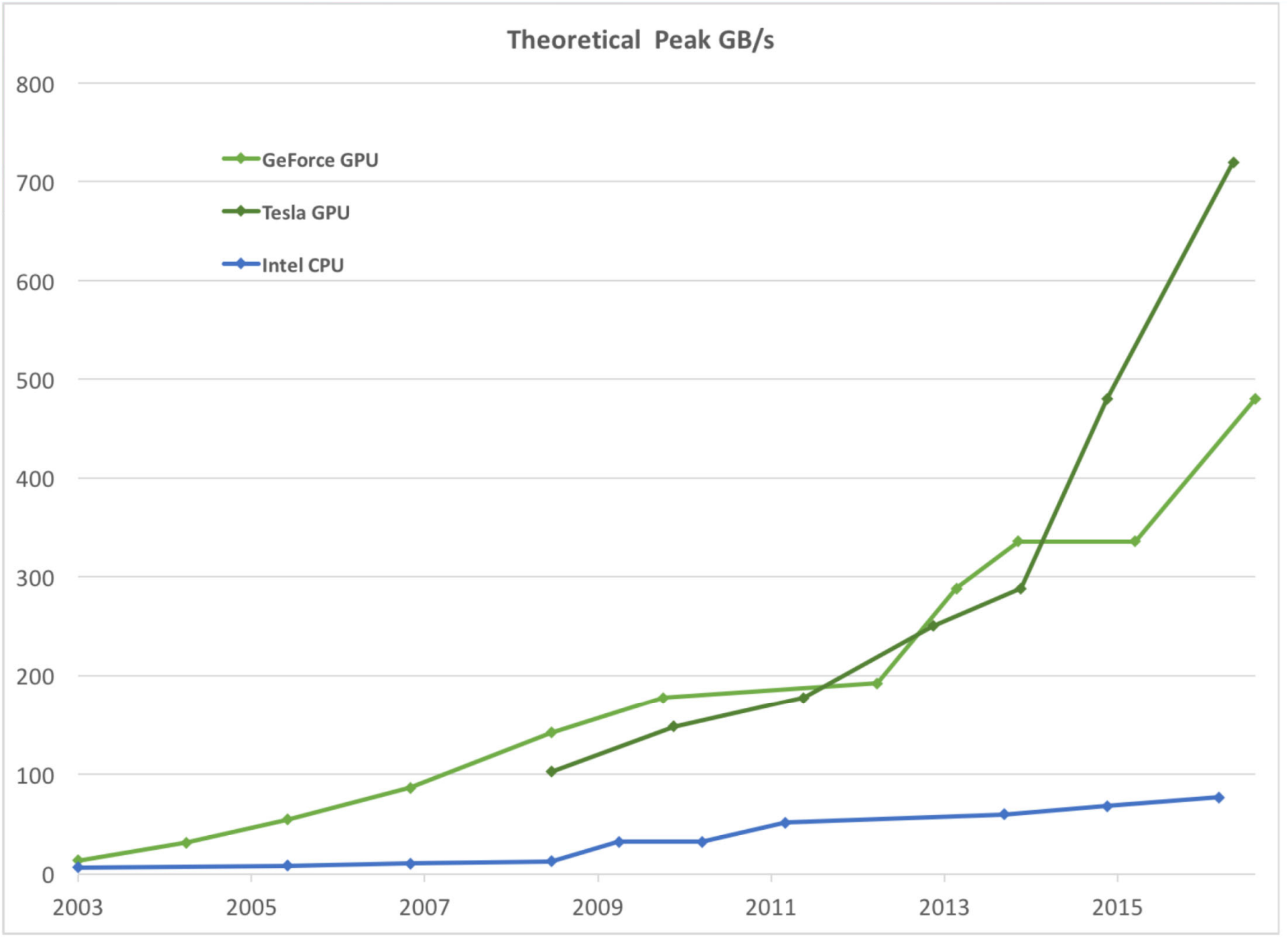


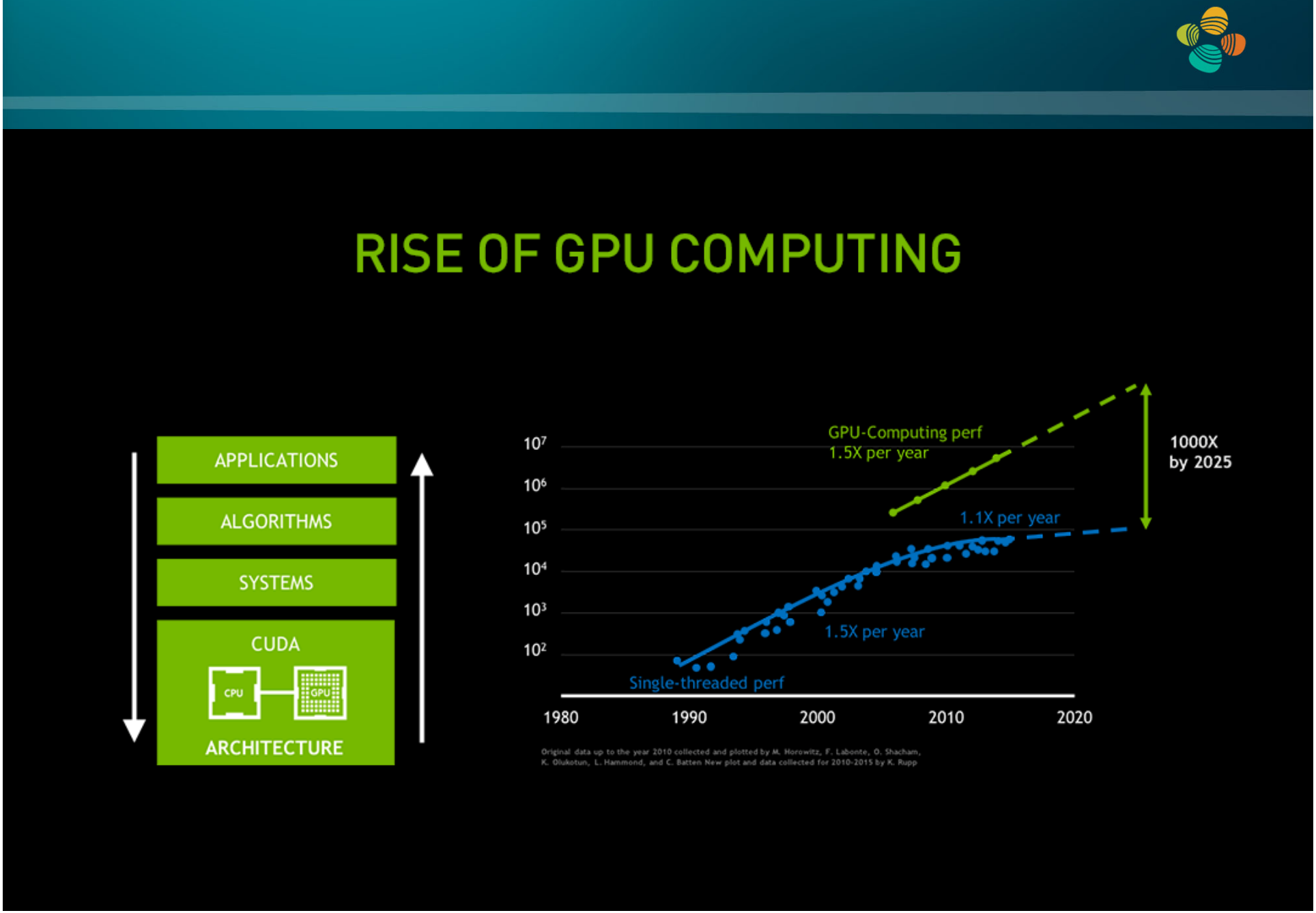

Theoretical GFLOP/s
Peak Performance
5750
5500
5250
5000
4750
4500
4250
4000
3750
3500
3250
3000
2750
2500
2250
2000
1750
1500
1250
1000
750
500
250
0
NVIDIA GPU Single Precision
NVIDIA GPU Double Precision
Intel CPU Double Precision
Intel CPU Single Precision
GeForce 780 Ti
GeForce GTX TITAN
GeForce GTX 680
GeForce GTX 580
GeForce GTX 480
GeForce GTX 280
GeForce 8800 GTX
GeForce 7800 GTX
GeForce 6800 Ultra
GeForce FX 5800
Tesla K40
Tesla K20X
Tesla M2090
Tesla C2050
Tesla C1060
Harpertown
Woodcrest
Pentium 4
Bloomfield
Westmere
Sandy Bridge
Ivy Bridge
Apr-01
Sep-02
Jan-04
May-05
Oct-06
Feb-08
Jul-09
Nov-10
Apr-12
Aug-13
Dec-14

Peak Performance

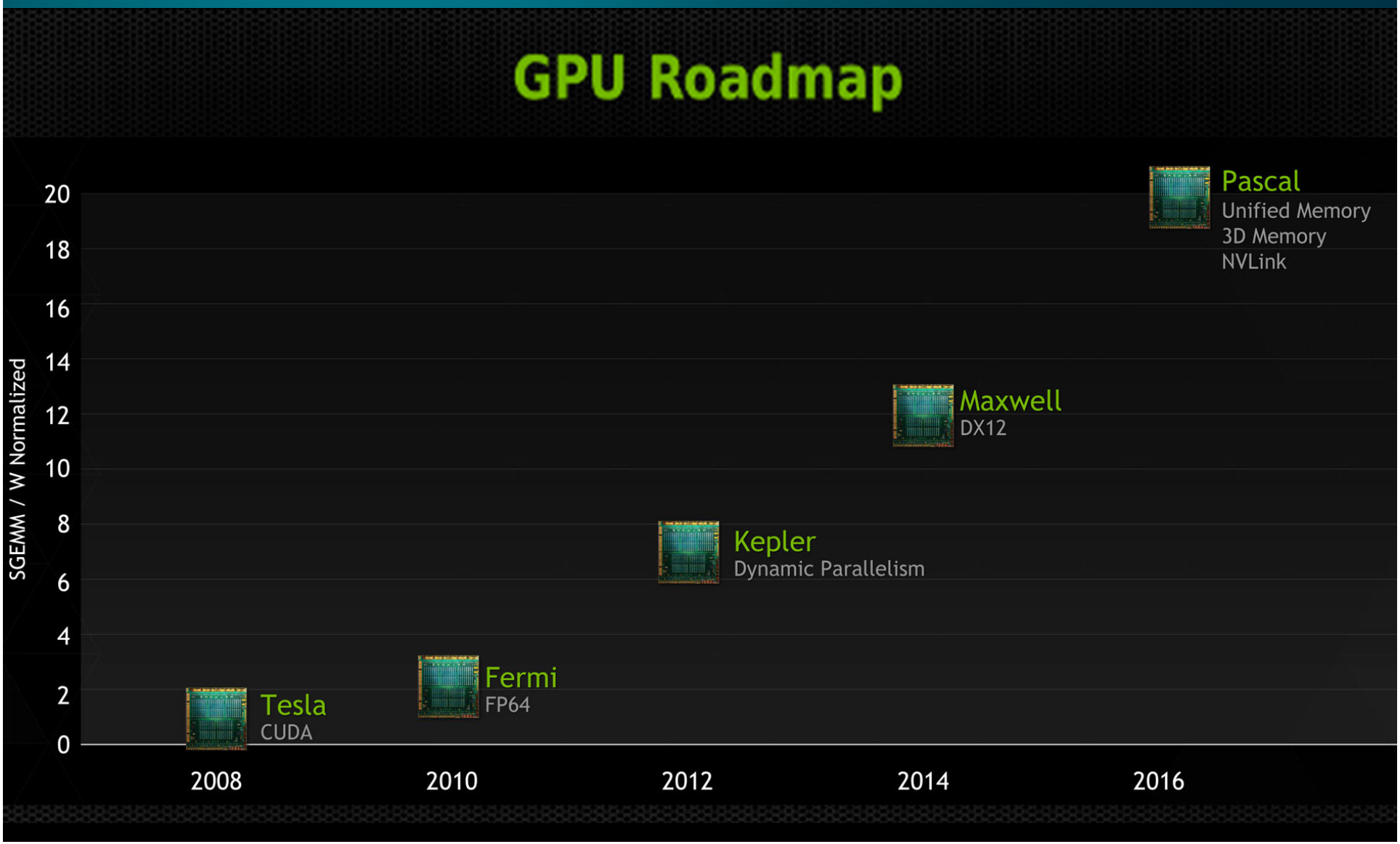



Peak Bandwidth
Theoretical GB/s
360
330
300
270
240
210
180
150
120
90
60
30
0
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
CPU
GeForce GPU
Tesla GPU
GeForce 780 Ti
Tesla K40
Tesla K20X
GeForce GTX 480
GeForce GTX 680
Tesla M2090
GeForce GTX 280
Tesla C2050
GeForce 8800 GTX
Tesla C1060
GeForce 7800 GTX
GeForce 6800 GT
GeForce FX 5900
Ivy Bridge
Sandy Bridge
Bloomfield
Westmere
Prescott
Woodcrest
Harpertown
Northwood

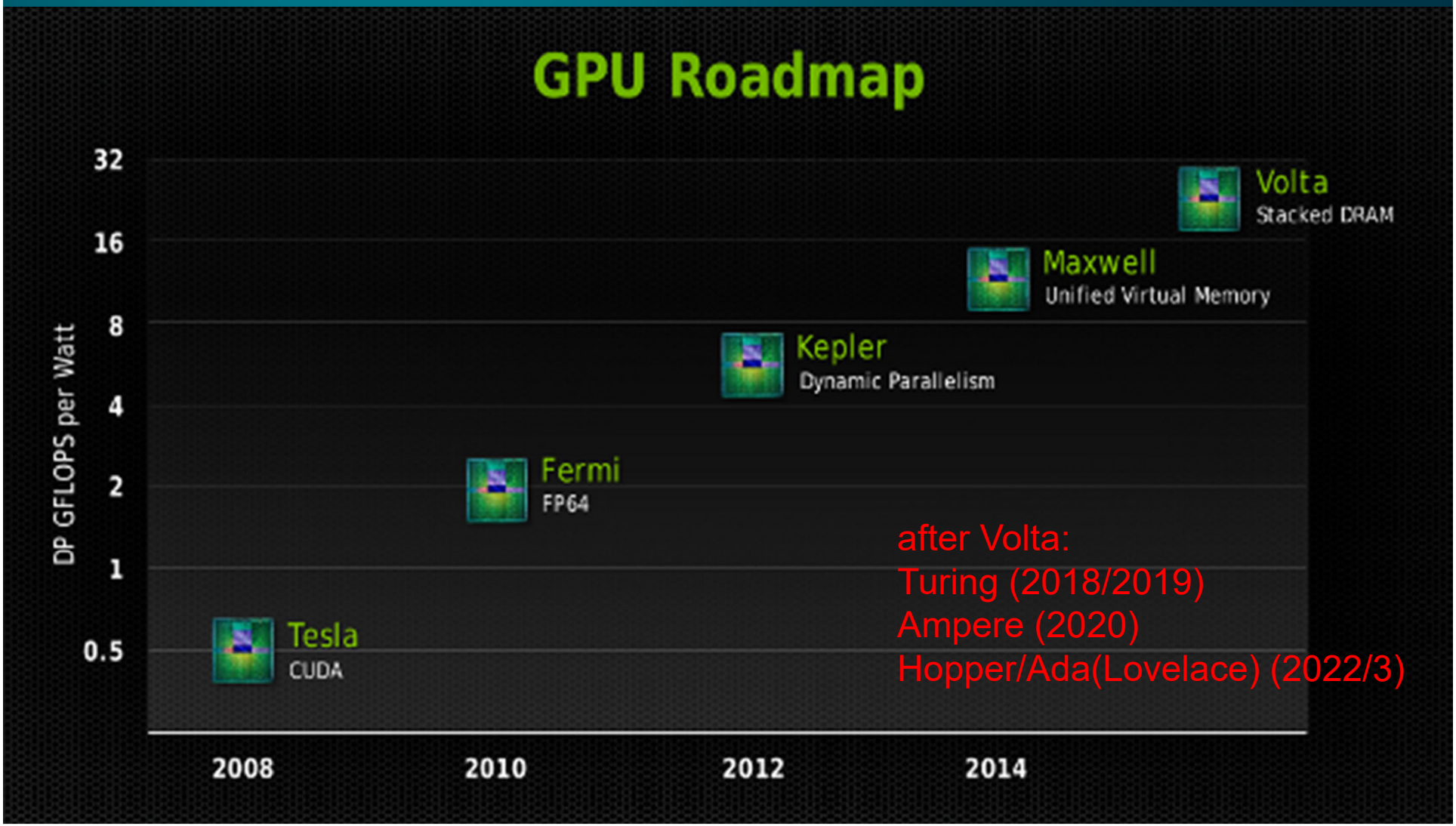

Theoretical Peak GB/s
800
700
600
500
400
300
200
100
0
GeForce GPU
Tesla GPU
Intel CPU
2003
2005
2007
2009
2011
2013
2015


RISE OF GPU COMPUTING
APPLICATIONS
ALGORITHMS
SYSTEMS
CUDA
CPU
GPU
ARCHITECTURE
GPU-Computing perf
1.5X per year
1000X
by 2025
1.1X per year
1.5X per year
Single-threaded perf
10^7
10^6
10^5
10^4
10^3
10^2
1980
1990
2000
2010
2020
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



Recent Updates (1)

NVIDIA Ampere architecture (2020)

https://en.wikipedia.org/wiki/Ampere_(microarchitecture)

Initial presentation 2020:

https://blogs.nvidia.com/blog/2020/05/14/gtc-2020-keynote/

Geforce 30-series (Ampere):

https://nvidia.com/en-us/geforce/graphics-cards/30-series/

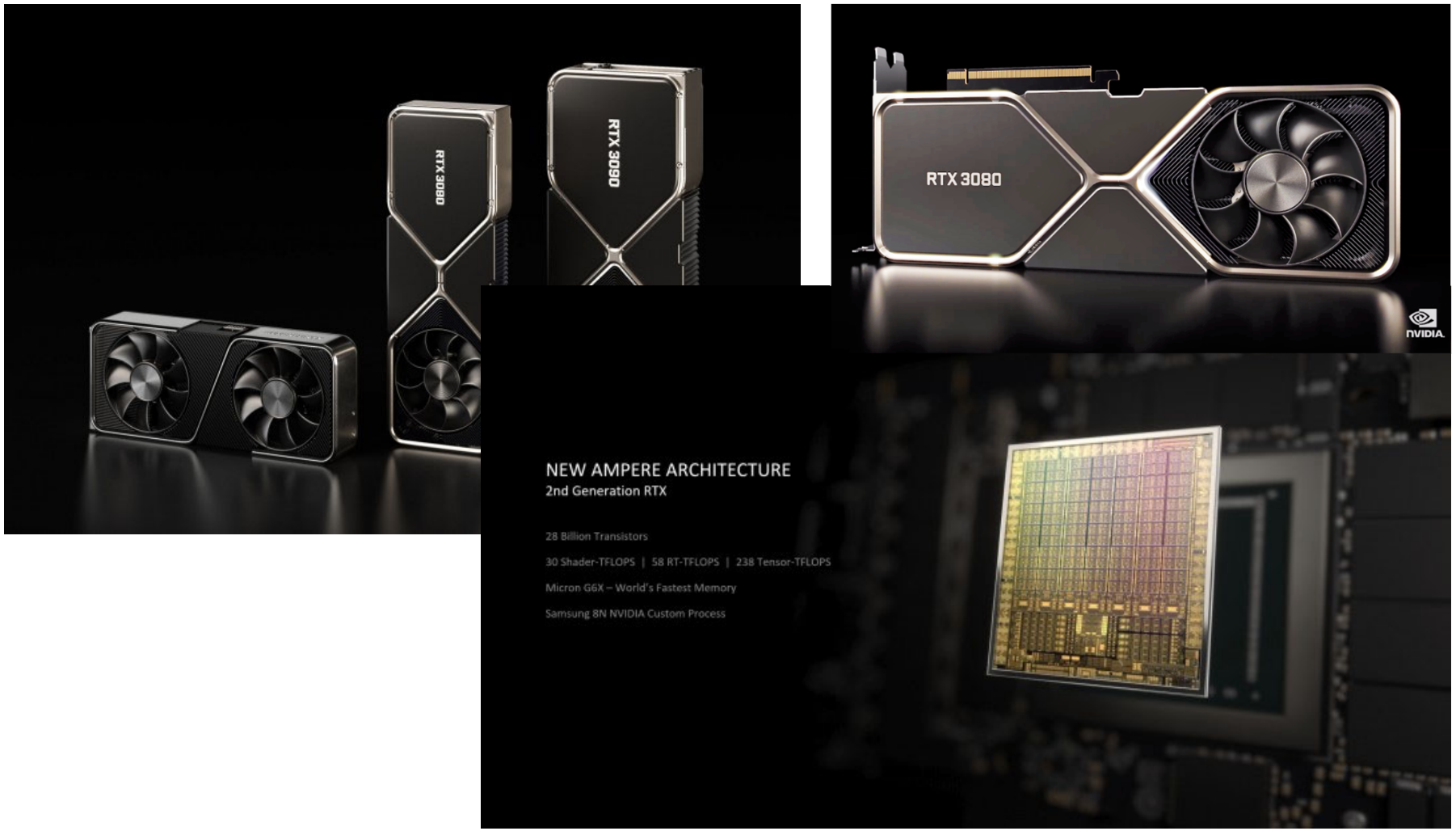
RTX 3090 has 10,496 FP32 CUDA cores (max arch would be 10,752)

A100 (Ampere):

https://www.nvidia.com/en-us/data-center/a100/

A100 has 6,912 FP32 CUDA cores (max arch would be 8,192)

Recent Updates (1)



Recent Updates (2)

NVIDIA Hopper architecture (2022)

https://en.wikipedia.org/wiki/Hopper_(microarchitecture)

Presentation [27:02] from NVIDIA GTC keynote, March 2022:

https://www.youtube.com/watch?v=39ubNuxnrK8

Hopper Whitepaper:

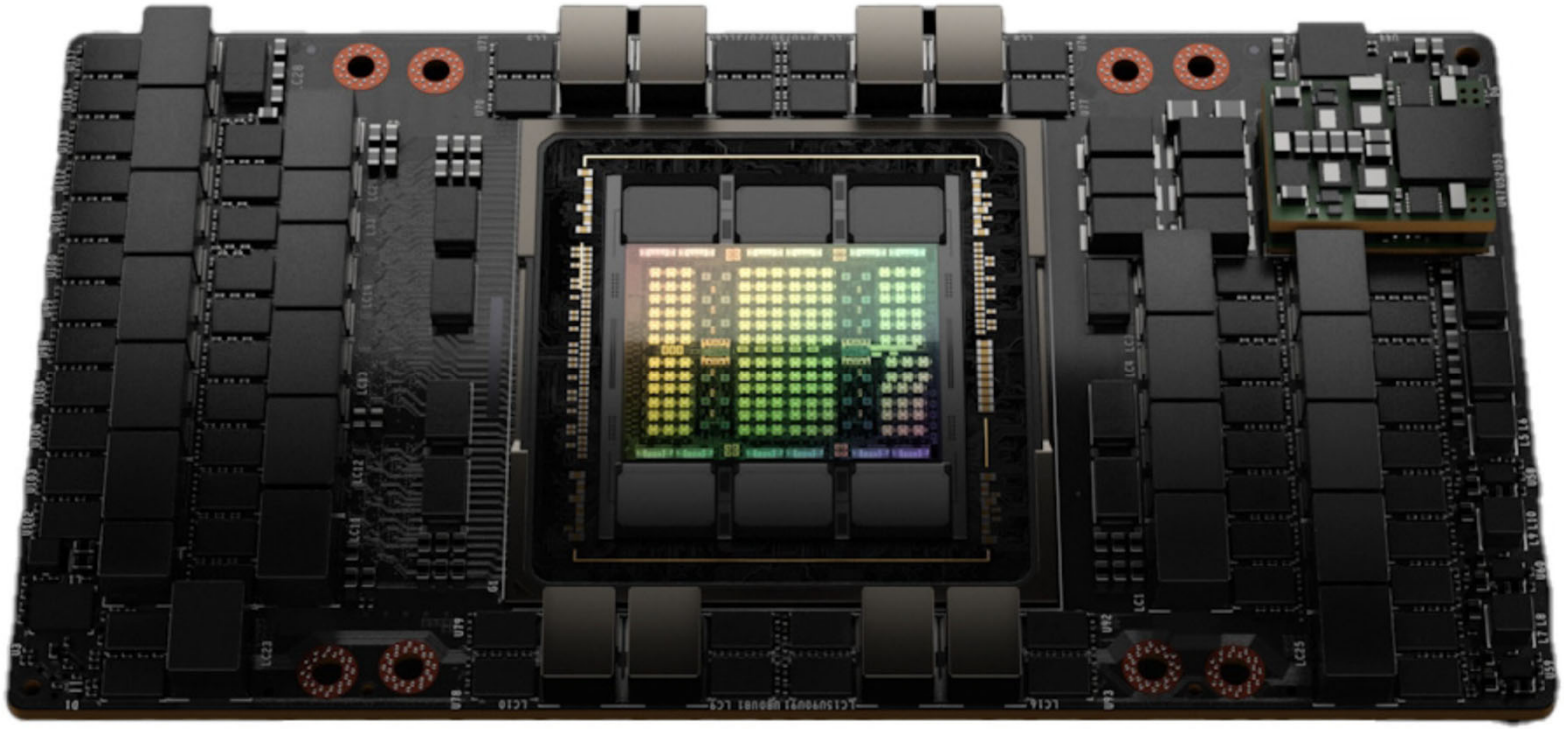
https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper/

H100 (Hopper):

https://www.nvidia.com/en-us/data-center/h100/

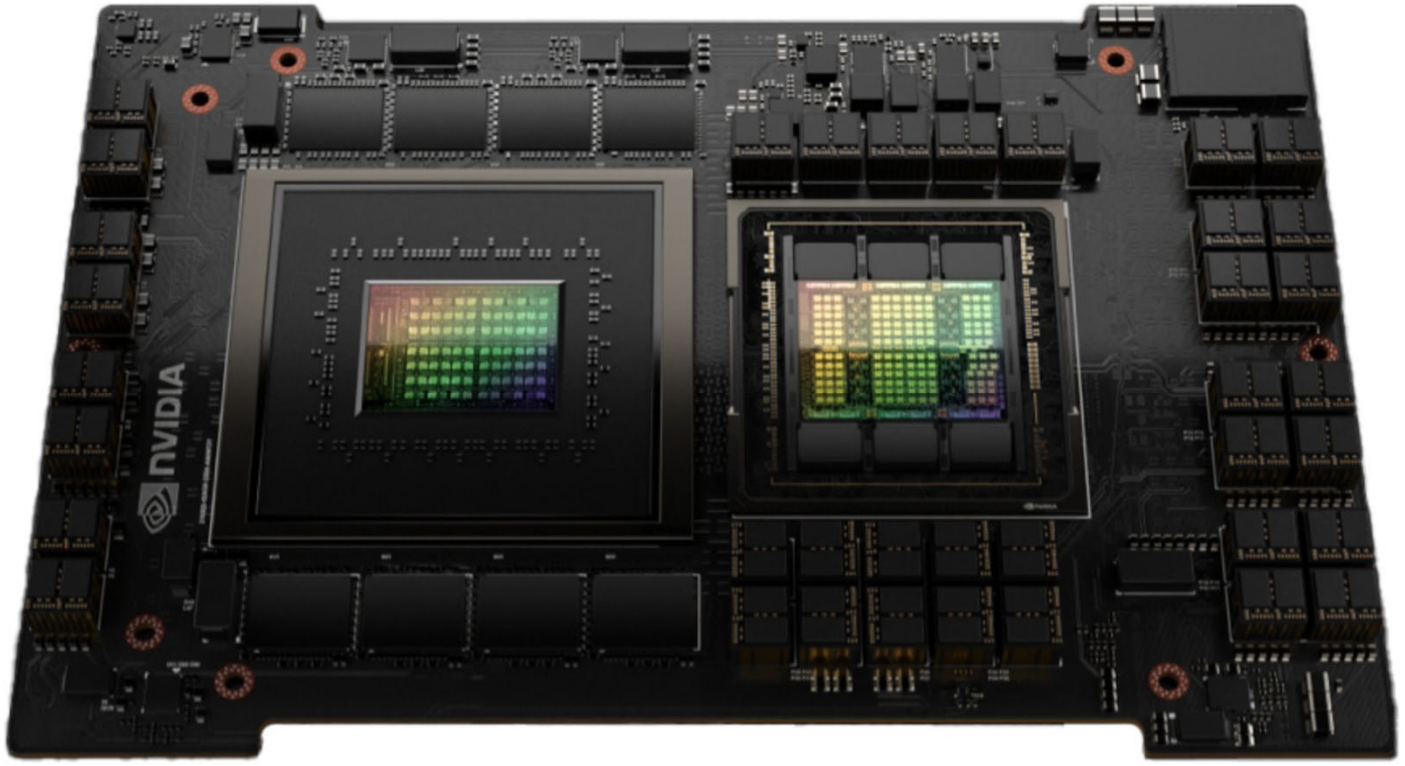
H100 has up to 18,432 FP32 CUDA cores (max arch)
(H100 SXM5: 16,896; H100 PCIe: 14,592)

Recent Updates (2)



NVIDIA H100 SXM5

Recent Updates (2)



NVIDIA Grace Hopper Superchip (Grace CPU + Hopper GPU)

NVIDIA Hopper GH100 Architecture (2022)

GH 100 (H100 Tensor Core GPU) Full GPU: 144 SMs (in 8 GPCs/72 TPCs)

NVIDIA GH100 SM

CC 9.0 Multiprocessor

- 128 FP32 + 64 INT32 cores
- 64 FP64 cores
- 4x 4th gen tensor cores
- ++ thread block clusters, DPX insts., FP8, TMA

4 partitions inside SM

- 32 FP32 + 16 INT32 cores
- 16 FP64 cores
- 8x LD/ST units each
- 1x 4th gen tensor core each
- Each has: warp scheduler, dispatch unit, 16K register file

NVIDIA Hopper GH100 Architecture (2022)

GH 100 (H100) Full GPU: 144 SMs (in 8 GPCs/72 TPCs)

- 64K 32-bit registers / SM = 256 KB register storage per SM
- 256 KB shared memory / L1 per SM

For 144 SMs on full GPU [SXM5: 132; PCIe: 114]

- 36 MB register storage, 36 MB shared mem / L1 storage = **72 MB context+"shared context" storage !**
- L2 cache size on H100: 50 MB
- 18,432 FP32 cores (128 FP32 cores per SM)
- 294,912 max threads in flight (max warps / SM = 64)

Recent Updates (3)

NVIDIA Ada (Lovelace) (2022/2023)

Upcoming GTC presentation September 19-22, 2022:

`https://www.nvidia.com/gtc/`

Geforce 40-series (speculative for now):

`https://en.wikipedia.org/wiki/Draft:GeForce_40_series/`

RTX 40xx has (not announced yet) FP32 CUDA cores

Overviews and Specs

Wikipedia has many comprehensive lists of architectures and specs

https://en.wikipedia.org/wiki/
List_of_Nvidia_graphics_processing_units

https://en.wikipedia.org/wiki/
List_of_AMD_graphics_processing_units

Thank you.