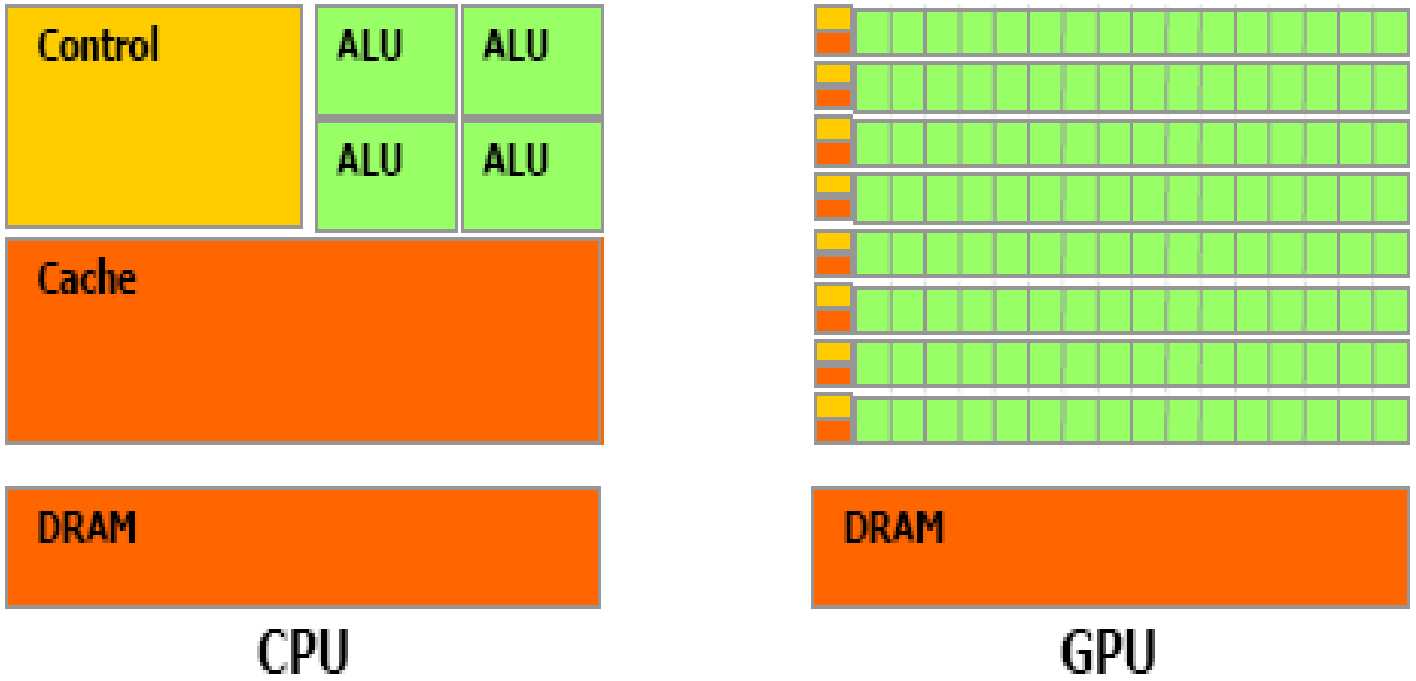
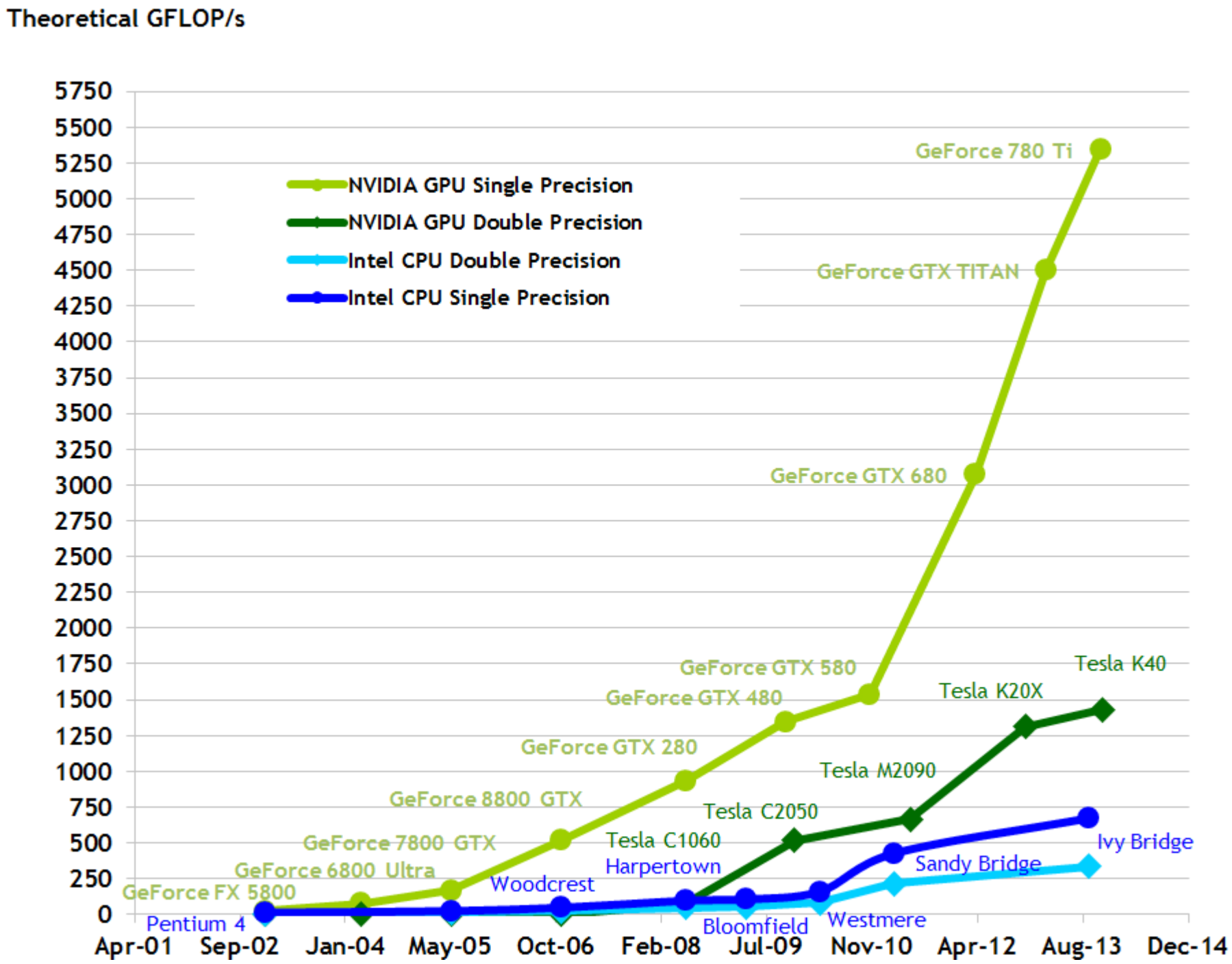
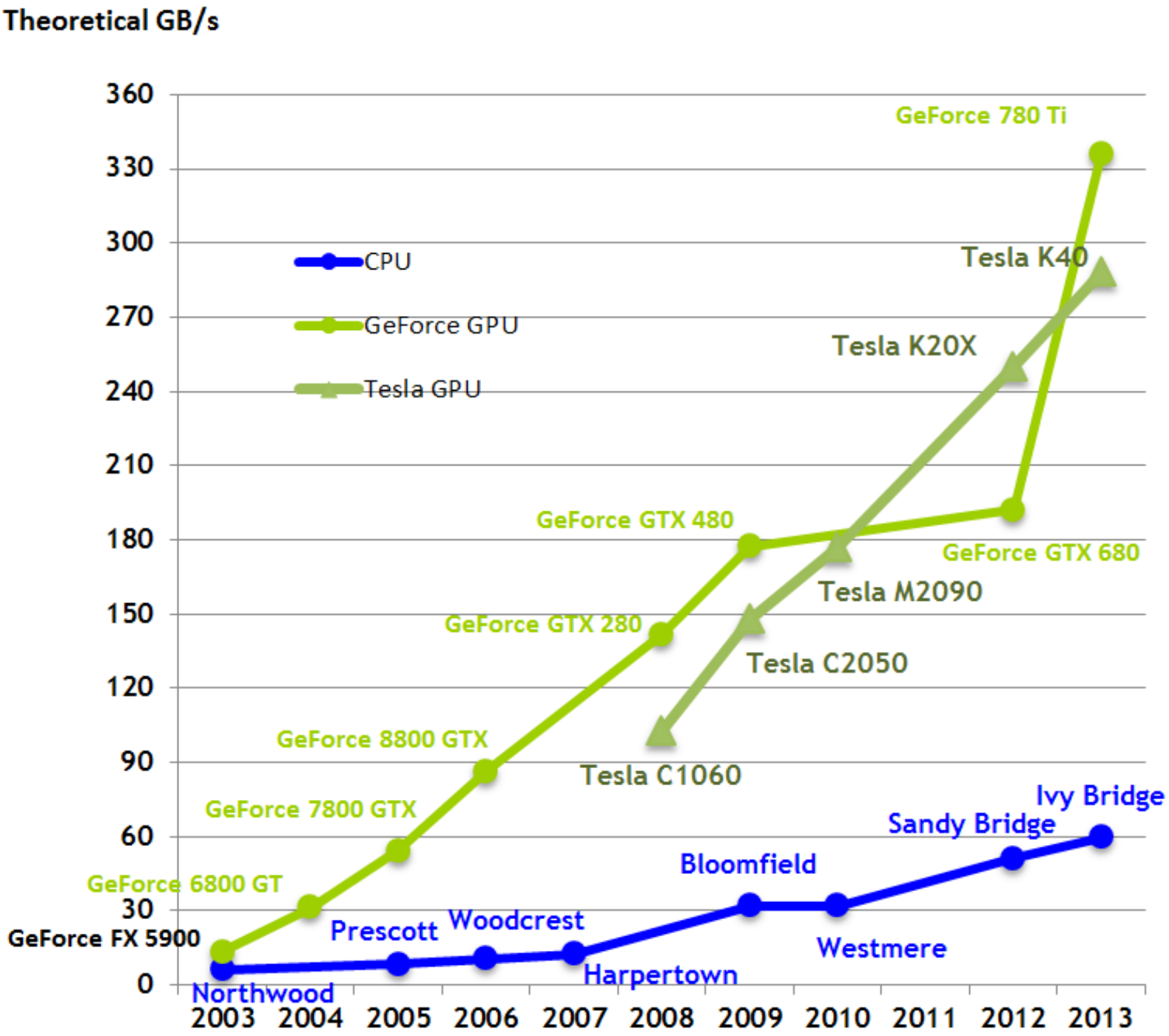
CUDA Architecture

CPU v/s GPU

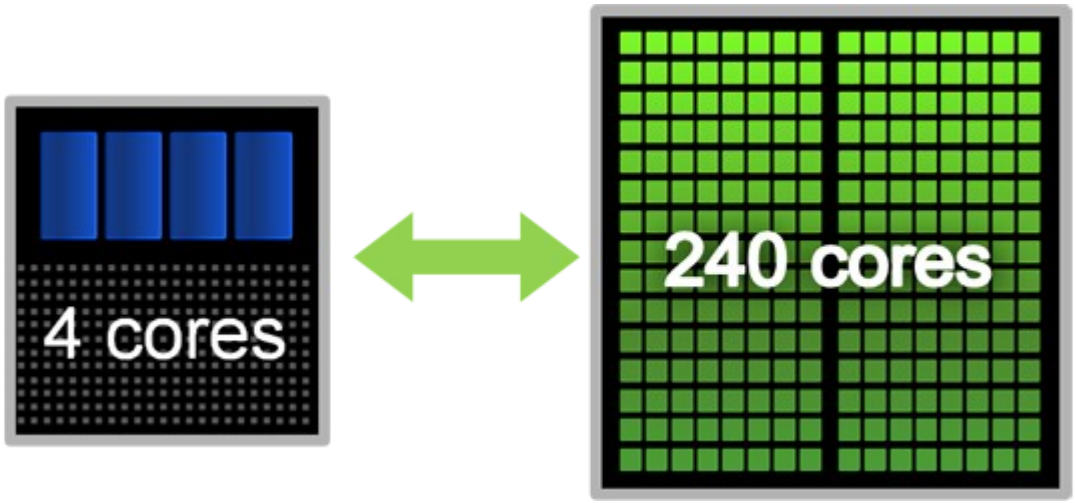



Theoretical GFLOP/s
5750
5500
5250
5000
4750
4500
4250
4000
3750
3500
3250
3000
2750
2500
2250
2000
1750
1500
1250
1000
750
500
250
0
NVIDIA GPU Single Precision
NVIDIA GPU Double Precision
Intel CPU Double Precision
Intel CPU Single Precision
GeForce 780 Ti
GeForce GTX TITAN
GeForce GTX 680
GeForce 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
Ivy Bridge
Sandy Bridge
Westmere
Bloomfield
Apr-01
Sep-02
Jan-04
May-05
Oct-06
Feb-08
Jul-09
Nov-10
Apr-12
Aug-13
Dec-14

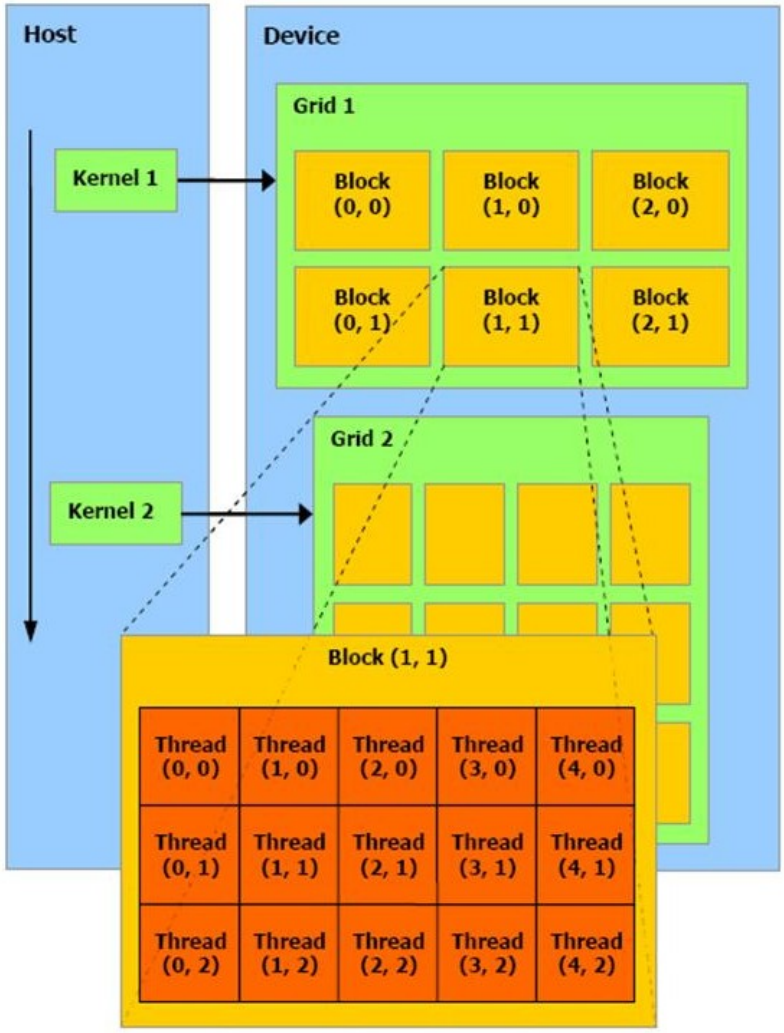

Theoretical GB/s
360
330
300
270
240
210
180
150
120
90
60
30
0
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
CPU
GeForce GPU
Tesla GPU
GeForce 780 Ti
Tesla K40
Tesla K20X
GeForce GTX 480
GeForce GTX 680
Tesla M2090
GeForce GTX 280
Tesla C2050
GeForce 8800 GTX
Tesla C1060
GeForce 7800 GTX
Ivy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

GPU and CPU

- Typically GPU and CPU coexist in a heterogeneous setting
- “Less” computationally intensive part runs on CPU (coarse-grained parallelism), and more intensive parts run on GPU (fine-grained parallelism)
- NVIDIA’s GPU architecture is called CUDA (Compute Unified Device Architecture) architecture, accompanied by CUDA programming model, and CUDA C language



Programming Model

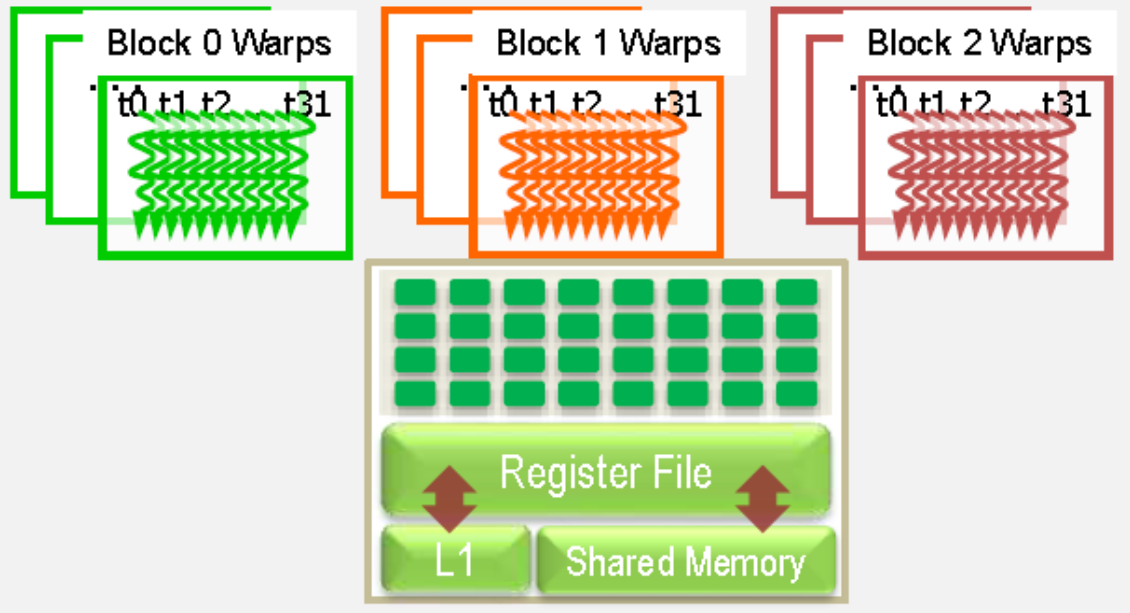


Programming Model

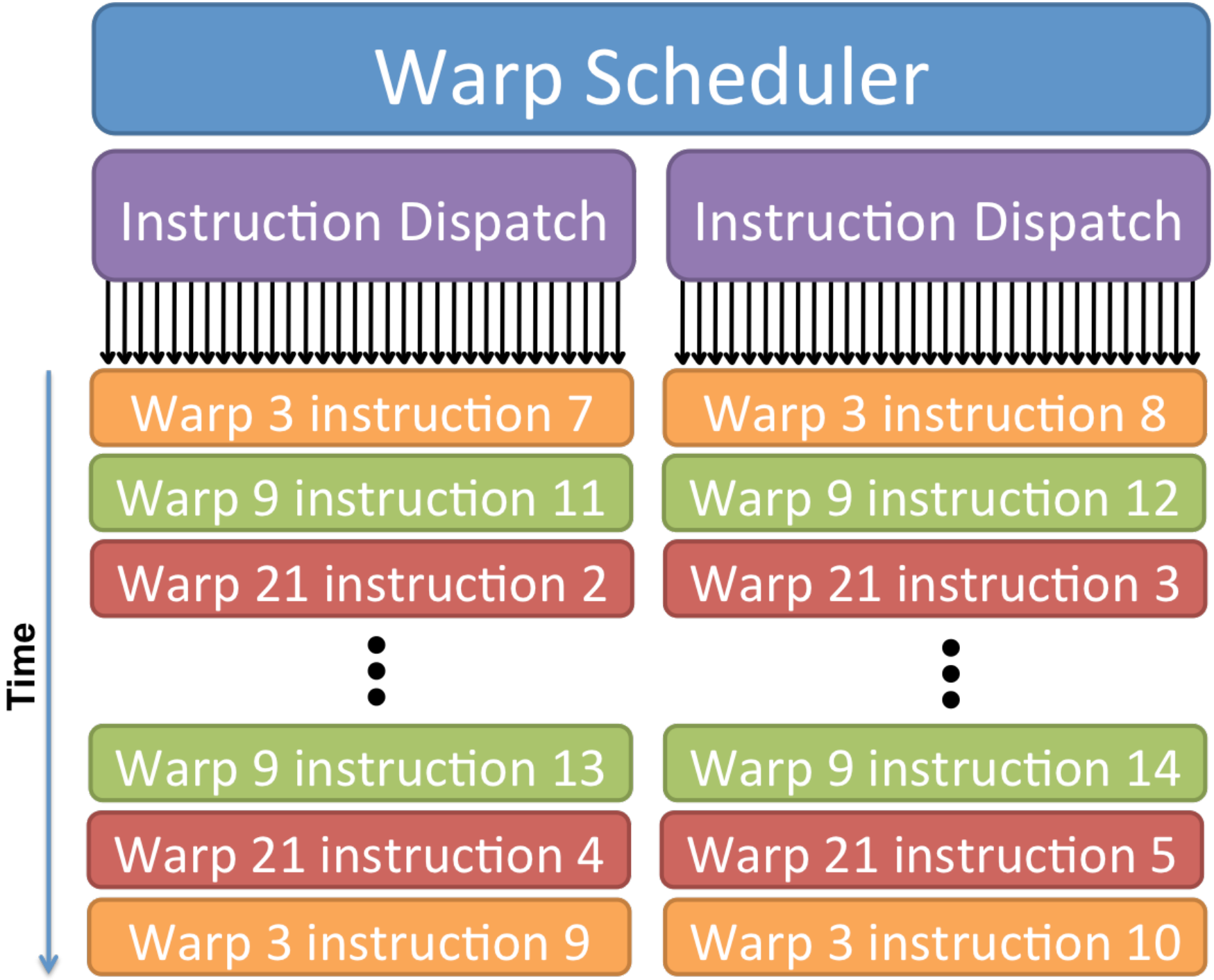
- All threads in a grid run the same kernel code (SPMD)
- Each thread has indexes that it uses to compute memory addresses and make control decisions.
    - blockIdx.x*blockDim.x+threadIdx.x
- Threads run in groups of 32 called warps
    - Threads in a warp execute in SIMD
- Every thread in a warp executes the same instruction at a time

Warp example

- If 3 blocks are assigned to an SM and each block has 256 threads, how many Warps are there in an SM?
  - Each Block is divided into 256/32 = 8 Warps
  - There are 8 * 3 = 24 Warps



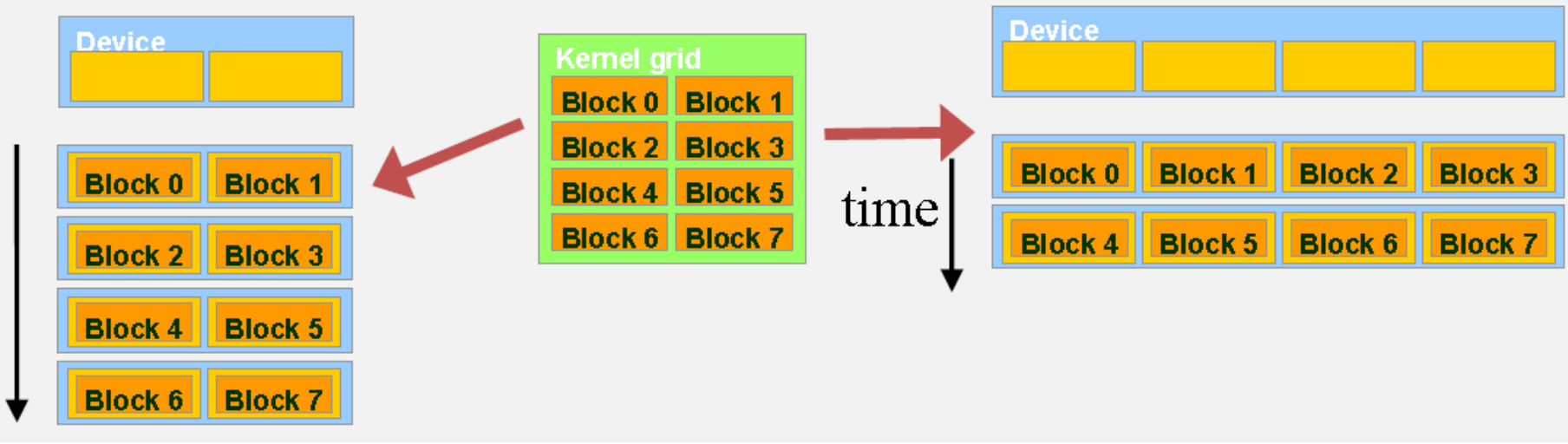
Onde os blocos são executados


Warp Scheduler
Instruction Dispatch
Instruction Dispatch
Warp 3 instruction 7
Warp 3 instruction 8
Warp 9 instruction 11
Warp 9 instruction 12
Warp 21 instruction 2
Warp 21 instruction 3
Warp 9 instruction 13
Warp 9 instruction 14
Warp 21 instruction 4
Warp 21 instruction 5
Warp 3 instruction 9
Warp 3 instruction 10
Time

Onde os blocos são executados

Transparent Scalability

- Each block can execute in any order relative to others.
  - Hardware is free to assign blocks to any processor at any time
  - A kernel scales to any number of parallel processors

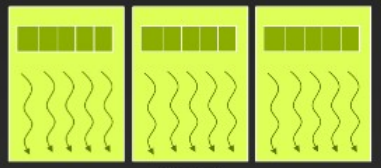


Execution Model

Software

Thread

Thread Block

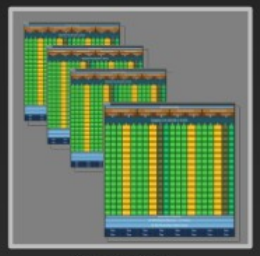


Grid

Hardware

CUDA Core

Multiprocessor



Device

Threads are executed by scalar CUDA Cores

Thread blocks are executed on multiprocessors

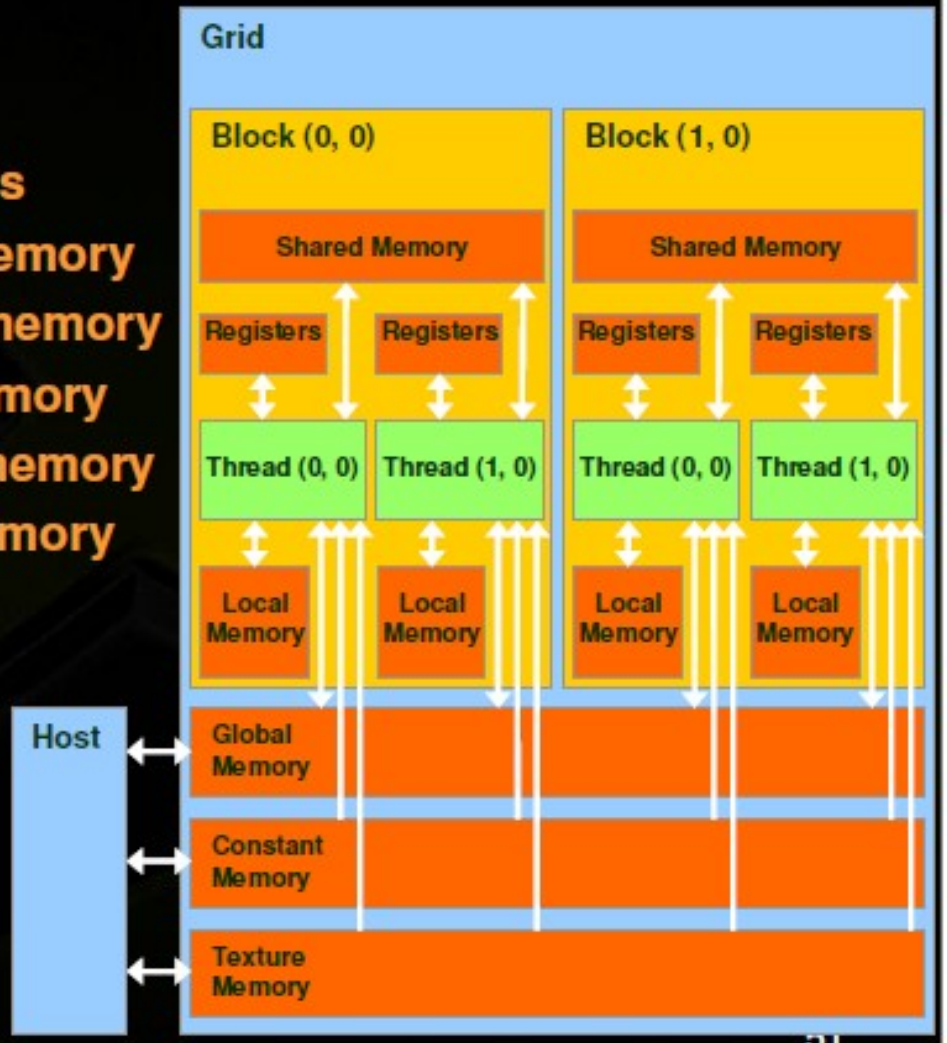
Thread blocks do not migrate

Several concurrent thread blocks can reside on one multiprocessor - limited by multiprocessor resources (shared memory and register file)

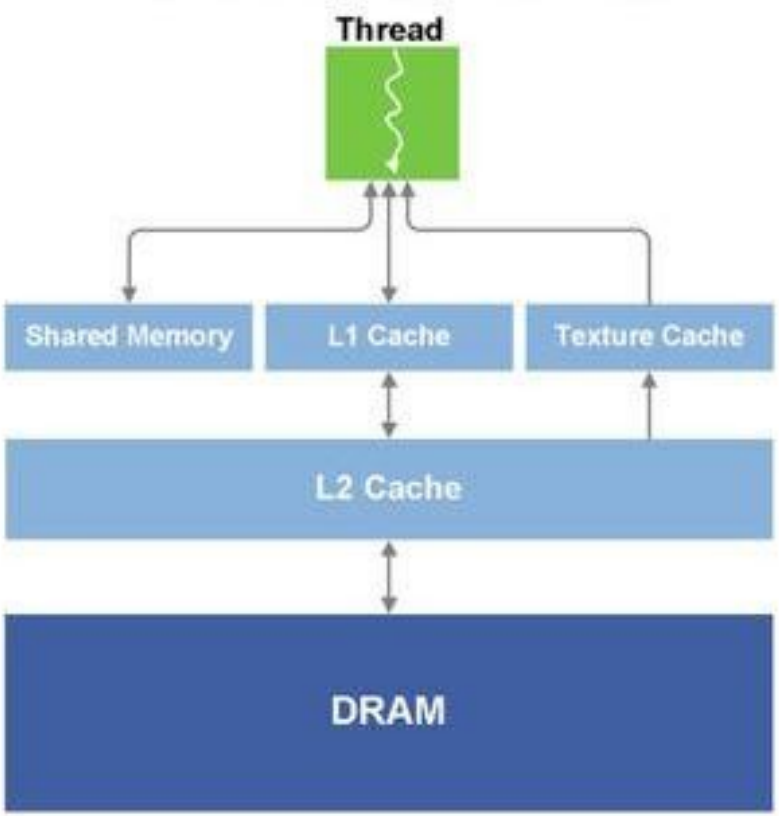
A kernel is launched as a grid of thread blocks

CUDA Memory Spaces

- Each thread can:
  - Read/write per-thread registers
  - Read/write per-thread local memory
  - Read/write per-block shared memory
  - Read/write per-grid global memory
  - Read only per-grid constant memory
  - Read only per-grid texture memory
- The host can read/write global, constant, and texture memory (stored in DRAM)



Thread Memory Access



NVIDIA Kepler K40

- 2880 streaming processors/cores (SPs) organized as 15 streaming multiprocessors (SMs)
- Each SM contains 192 cores
- Memory size of the GPU system: 12 GB
- Clock speed of a core: 745 MHz

Technical specifications	Compute capability (version)									
	1.0	1.1	1.2	1.3	2.x	3.0	3.5	3.7	5.0	5.2
Maximum dimensionality of grid of thread blocks	2				3					
Maximum x-dimension of a grid of thread blocks	65535					2^{31}-1				
Maximum y-, or z-dimension of a grid of thread blocks	65535									
Maximum dimensionality of thread block	3									
Maximum x- or y-dimension of a block	512				1024					
Maximum z-dimension of a block	64									
Maximum number of threads per block	512				1024					
Warp size	32									
Maximum number of resident blocks per multiprocessor	8					16			32	
Maximum number of resident warps per multiprocessor	24		32		48	64				
Maximum number of resident threads per multiprocessor	768		1024		1536	2048				
Number of 32-bit registers per multiprocessor	8 K		16 K		32 K	64 K		128 K	64 K	
Maximum number of 32-bit registers per thread	128				63	255				
Maximum amount of shared memory per multiprocessor	16 KB				48 KB			112 KB	64 KB	96 KB
Number of shared memory banks	16				32					
Amount of local memory per thread	16 KB				512 KB					
Constant memory size	64 KB									
Cache working set per multiprocessor for constant memory	8 KB								10 KB	
Cache working set per multiprocessor for texture memory	Device dependent, between 6 KB and 8 KB				12 KB	Between 12 KB and 48 KB			24 KB	
Maximum width for 1D texture reference bound to a CUDA array	8192				65536					
Maximum width for 1D texture reference bound to linear memory	2^{27}									

Tesla Products	Tesla K40	Tesla M40	Tesla P100
GPU	GK110 (Kepler)	GM200 (Maxwell)	GP100 (Pascal)
SMs	15	24	56
TPCs	15	24	28
FP32 CUDA Cores / SM	192	128	64
FP32 CUDA Cores / GPU	2880	3072	3584
FP64 CUDA Cores / SM	64	4	32
FP64 CUDA Cores / GPU	960	96	1792
Base Clock	745 MHz	948 MHz	1328 MHz
GPU Boost Clock	810/875 MHz	1114 MHz	1480 MHz
Compute Performance - FP32	5.04 TFLOPS	6.82 TFLOPS	10.6 TFLOPS
Compute Performance - FP64	1.68 TFLOPS	0.21 TFLOPS	5.3 TFLOPS
Texture Units	240	192	224
Memory Interface	384-bit GDDR5	384-bit GDDR5	4096-bit HBM2
Memory Size	Up to 12 GB	Up to 24 GB	16 GB
L2 Cache Size	1536 KB	3072 KB	4096 KB
Register File Size / SM	256 KB	256 KB	256 KB
Register File Size / GPU	3840 KB	6144 KB	14336 KB
TDP	235 Watts	250 Watts	300 Watts
Transistors	7.1 billion	8 billion	15.3 billion
GPU Die Size	551 mm^2	601 mm^2	610 mm^2
Manufacturing Process	28-nm	28-nm	16-nm

P100

Tesla P100 accelerators have four 4-die HBM2 stacks, for a total of 16 GB of memory, and **720** GB/s peak bandwidth, which is 3 times higher than the Tesla M40 memory bandwidth