Introduction to GPU architecture

Sylvain Collange
Inria Rennes – Bretagne Atlantique

http://www.irisa.fr/alf/collange/
sylvain.collange@inria.fr

ADA - 2017

Graphics processing unit (GPU)

or

- Graphics rendering accelerator for computer games
  - Mass market: low unit price, amortized R&D
  - Increasing programmability and flexibility
- Inexpensive, high-performance parallel processor
  - GPUs are everywhere, from cell phones to supercomputers
- General-Purpose computation on GPU (GPGPU)

GPUs in high-performance computing

- GPU/accelerator share in Top500 supercomputers
  - In 2010: 2%
  - In 2017: 18%

- 2016+ trend:
  Heterogeneous multi-core processors influenced by GPUs

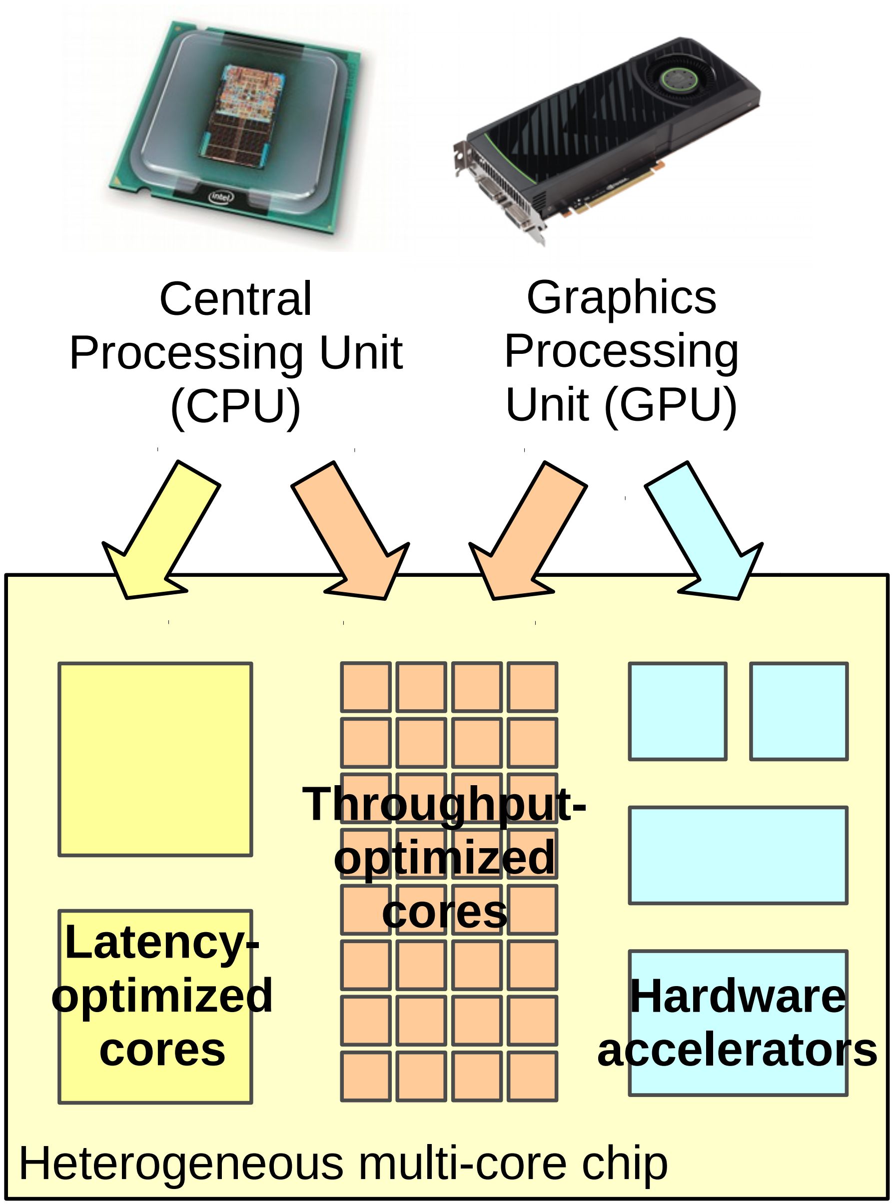
#1 Sunway TaihuLight (China)
40,960 × SW26010 (4 big + 256 small cores)

#2 Tianhe-2 (China)
16,000 × (2×12-core Xeon + 3×57-core Xeon Phi)

Being upgraded to Matrix-2000 coprocessors

GPGPU in the future?

- Yesterday (2000-2010)
  - Homogeneous multi-core
  - Discrete components
- Today (2011-...)
  Chip-level integration
  - Many embedded SoCs
  - Intel Sandy Bridge
  - AMD Fusion
  - NVIDIA Denver/Maxwell project…
- Tomorrow
  Heterogeneous multi-core
  - GPUs to blend into throughput-optimized cores?

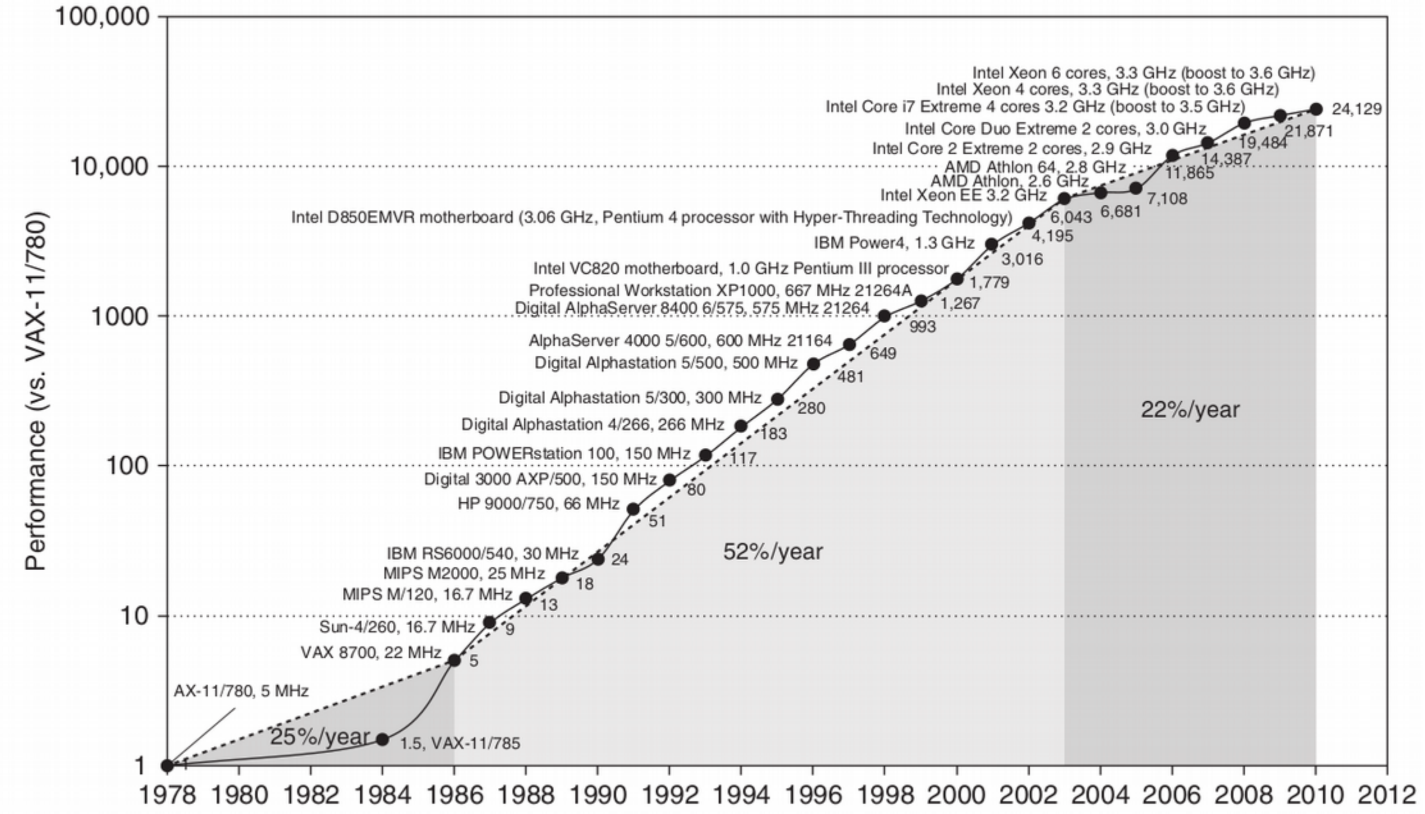


Outline

- GPU, many-core: why, what for?
  - Technological trends and constraints
  - From graphics to general purpose
  - Hardware trends
- Forms of parallelism, how to exploit them
  - Why we need (so much) parallelism: latency and throughput
  - Sources of parallelism: ILP, TLP, DLP
  - Uses of parallelism: horizontal, vertical
- Let's design a GPU!
  - Ingredients: Sequential core, Multi-core, Multi-threaded core, SIMD
  - Putting it all together
  - Architecture of current GPUs: cores, memory

The free lunch era... was yesterday

- 1980's to 2002: *Moore's law*, *Dennard scaling*, micro-architecture improvements
  - Exponential performance increase
  - Software compatibility preserved

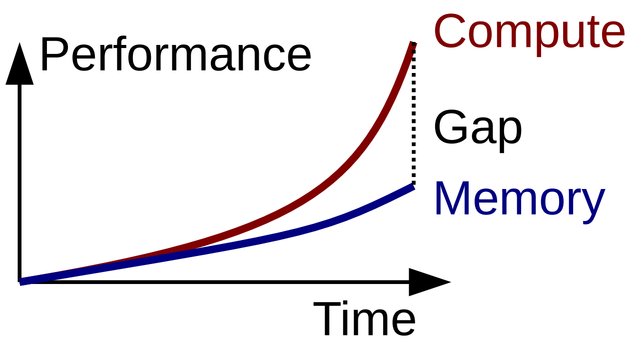


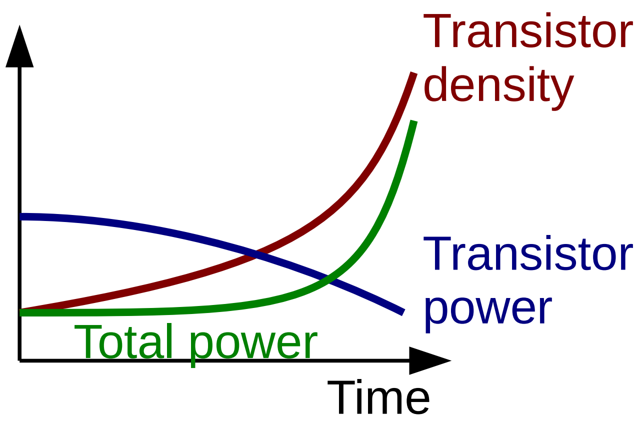
Hennessy, Patterson. Computer Architecture, a quantitative approach. 5th Ed. 2010

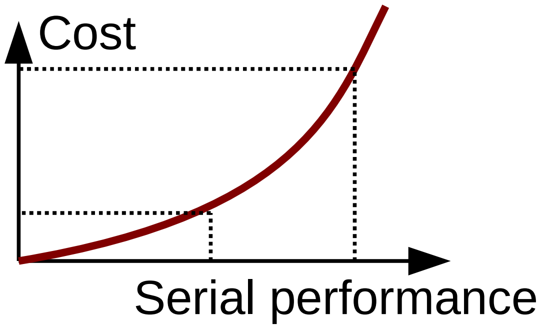
- Do not rewrite software, buy a new machine!

Technology evolution

- Memory wall
  - Memory speed does not increase as fast as computing speed
  - Harder to hide memory latency
- Power wall
  - Power consumption of transistors does not decrease as fast as density increases
  - Performance is now limited by power consumption
- ILP wall
  - Law of diminishing returns on Instruction-Level Parallelism
  - Pollack rule: cost ≃ performance²







Usage changes

- New applications demand **parallel processing**
  - Computer games : 3D graphics
  - Search engines, social networks… “big data” processing
- New computing devices are **power-constrained**
  - Laptops, cell phones, tablets…
    - Small, light, battery-powered
  - Datacenters
    - High power supply and cooling costs

Latency vs. throughput

- **Latency**: time to solution
  - Minimize time, at the expense of power
  - Metric: time
    e.g. seconds
- **Throughput**: quantity of tasks processed per unit of time
  - Assumes unlimited parallelism
  - Minimize energy per operation
  - Metric: operations / time
    e.g. Gflops / s
- CPU: optimized for latency
- GPU: optimized for throughput

Amdahl's law

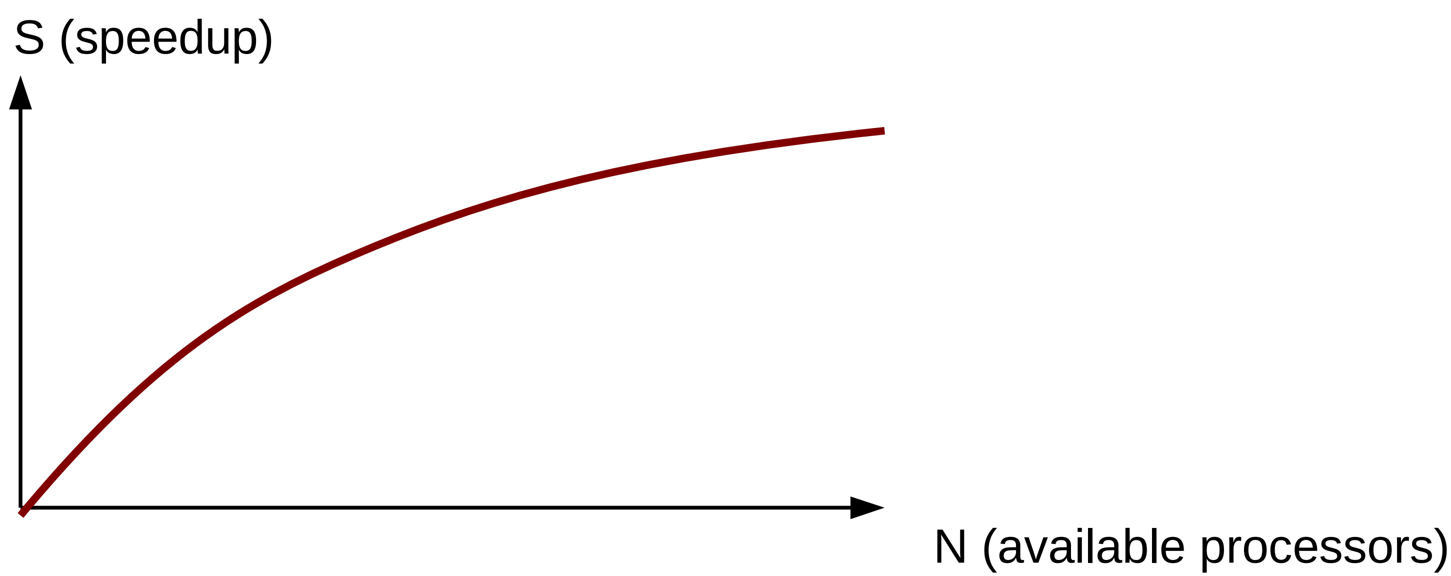
- Bounds speedup attainable on a parallel machine

Time to run sequential portions

$$S = \frac{1}{(1-P) + \frac{P}{N}}$$

Time to run parallel portions

S	Speedup
P	Ratio of parallel portions
N	Number of processors



G. Amdahl. Validity of the Single Processor Approach to Achieving Large-Scale Computing Capabilities. AFIPS 1967.

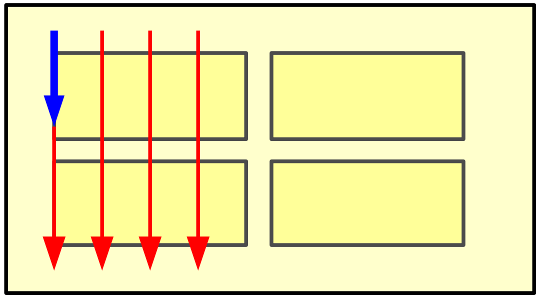
Why heterogeneous architectures?

$$S = \frac{1}{(1-P) + \frac{P}{N}}$$

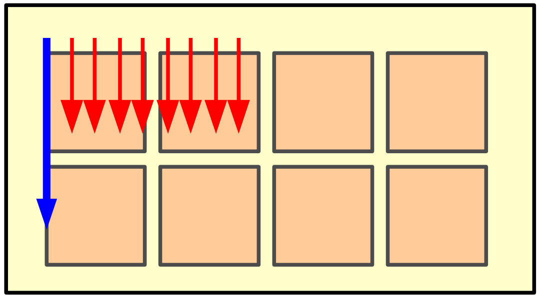
Time to run sequential portions

Time to run parallel portions

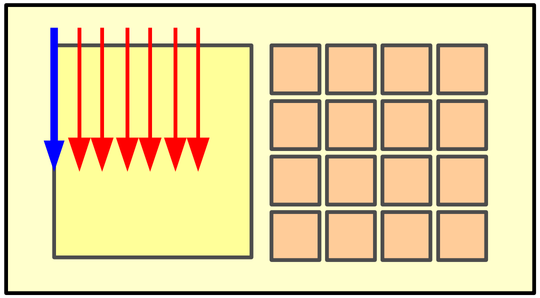
- Latency-optimized multi-core (CPU)
  - Low efficiency on parallel portions: spends too much resources



- Throughput-optimized multi-core (GPU)
  - Low performance on sequential portions

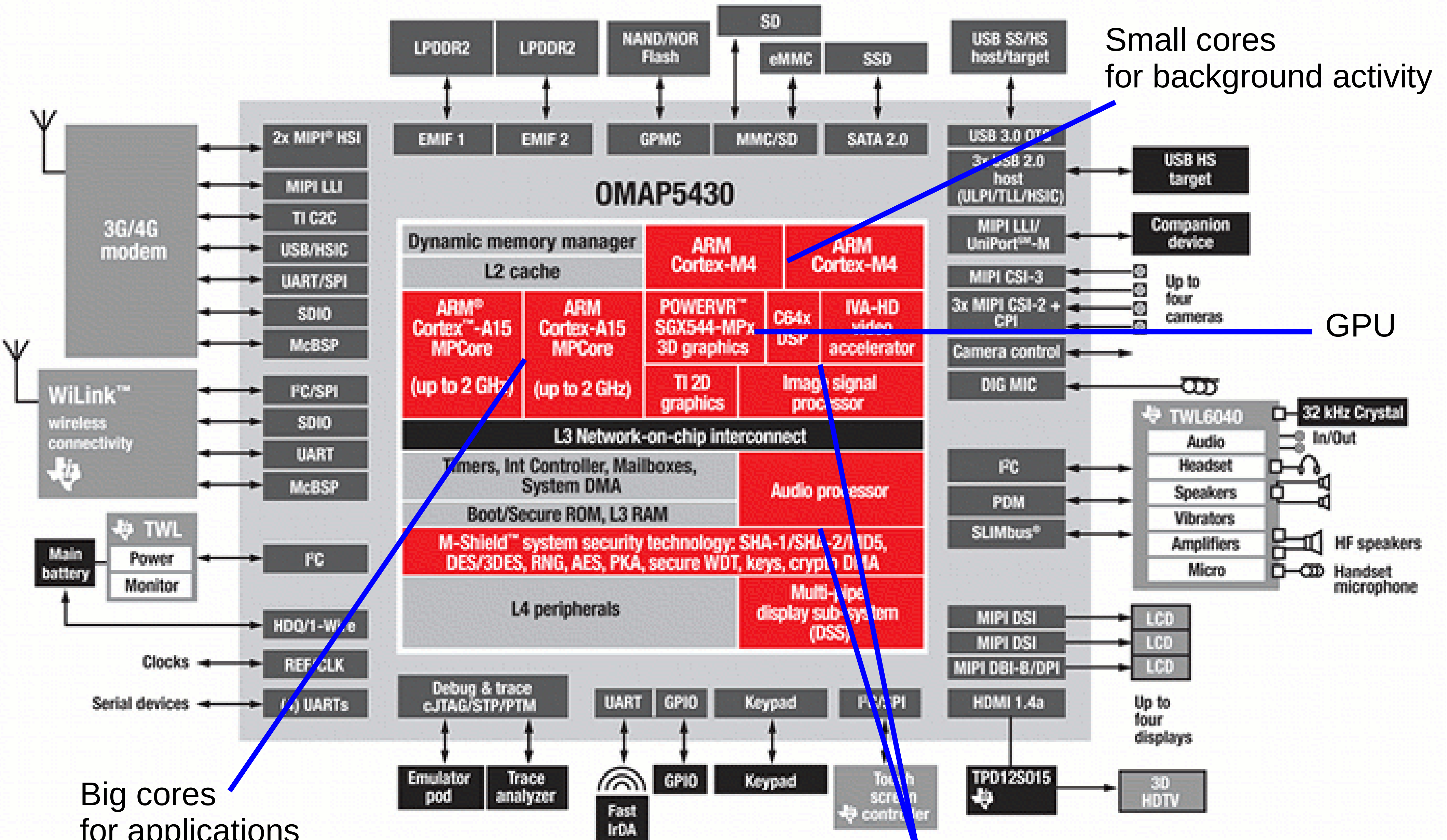


- Heterogeneous multi-core (CPU+GPU)
  - Use the right tool for the right job
  - Allows aggressive optimization for latency **or** for throughput



M. Hill, M. Marty. Amdahl's law in the multicore era. IEEE Computer, 2008.

Example: System on Chip for smartphone



Small cores for background activity

GPU

Big cores for applications

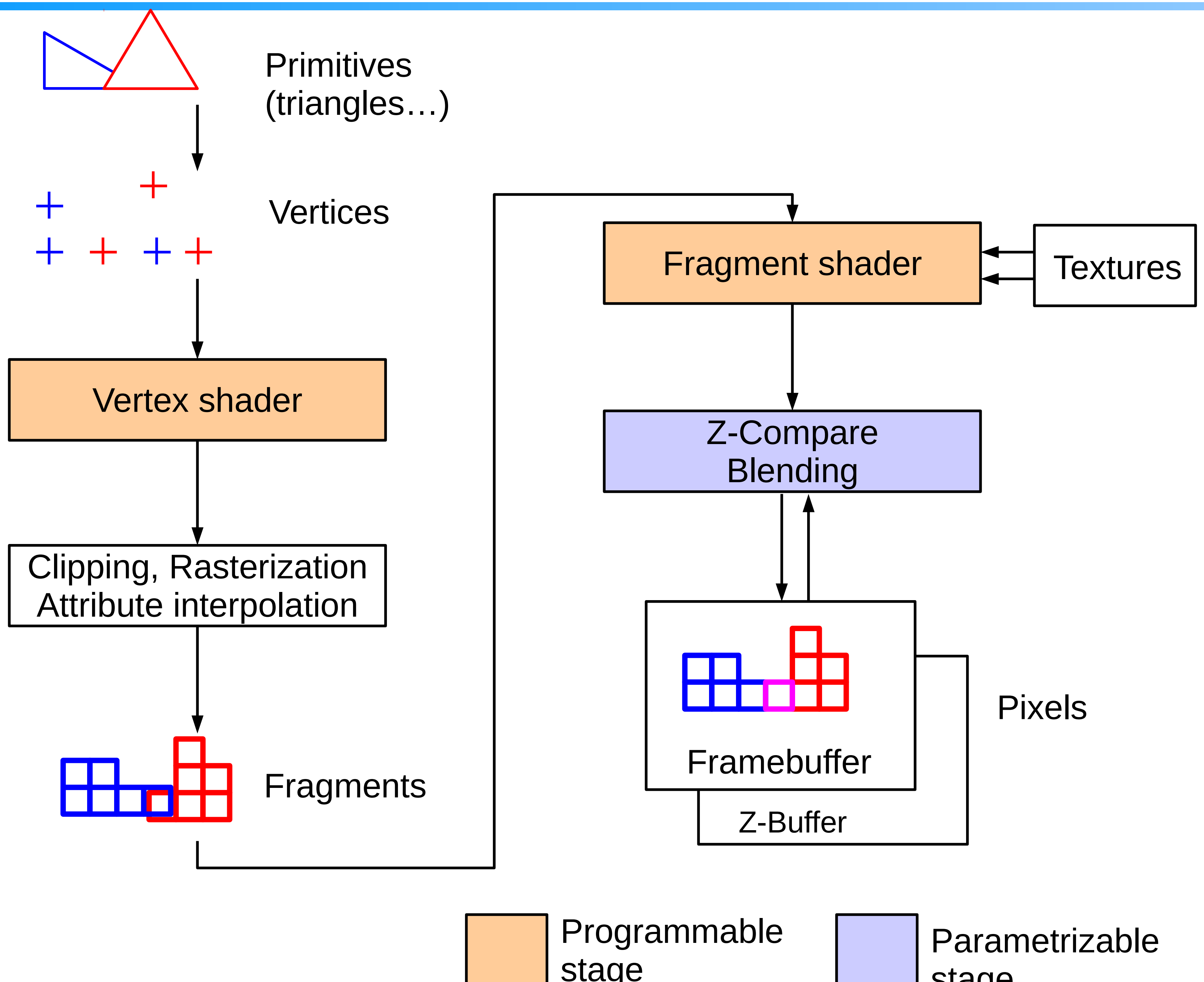
Lots of interfaces

Special-purpose accelerators

Outline

- GPU, many-core: why, what for?
  - Technological trends and constraints
  - From graphics to general purpose
  - Hardware trends
- Forms of parallelism, how to exploit them
  - Why we need (so much) parallelism: latency and throughput
  - Sources of parallelism: ILP, TLP, DLP
  - Uses of parallelism: horizontal, vertical
- Let's design a GPU!
  - Ingredients: Sequential core, Multi-core, Multi-threaded core, SIMD
  - Putting it all together
  - Architecture of current GPUs: cores, memory

The (simplest) graphics rendering pipeline



How much performance do we need

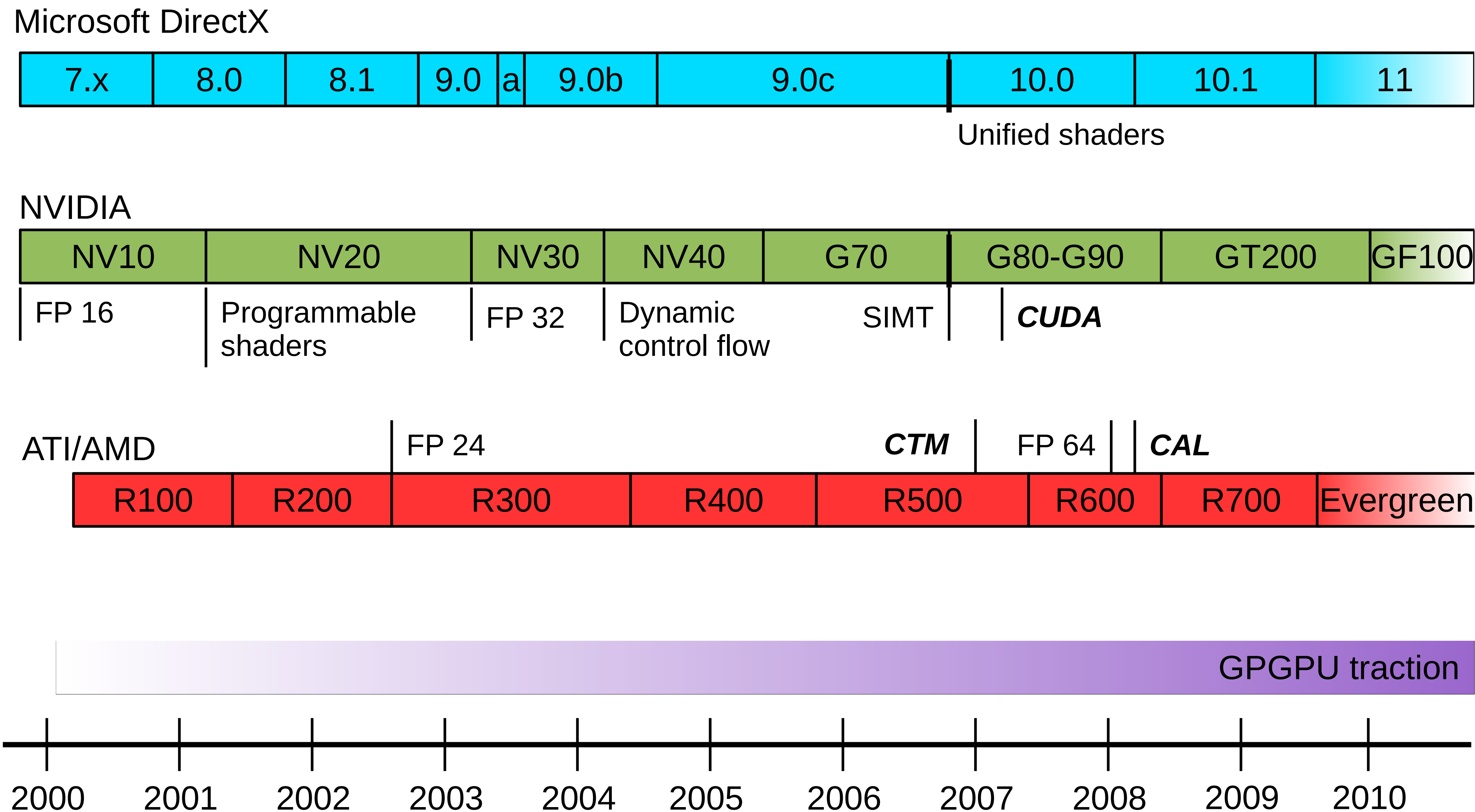
- … to run 3DMark 11 at 50 frames/second?

Element	Per frame	Per second
Vertices	12.0M	600M
Primitives	12.6M	630M
Fragments	180M	9.0G
Instructions	14.4G	**720G**

- Intel Core i7 2700K: 56 Ginsn/s peak
  - We need to go 13x faster
  - Make a special-purpose accelerator

Source: Damien Triolet, Hardware.fr

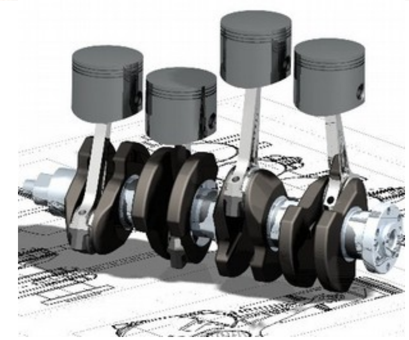
Beginnings of GPGPU

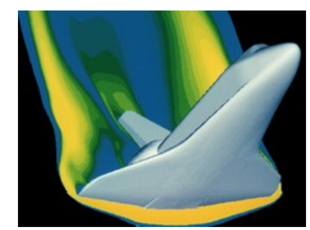


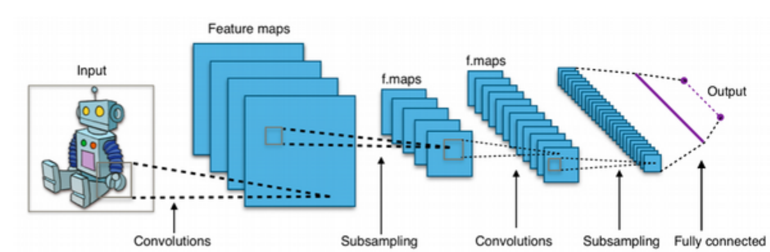
Today: what do we need GPUs for?

1. 3D graphics rendering for games
   - Complex texture mapping, lighting computations…
2. Computer Aided Design workstations
   - Complex geometry
3. High-performance computing
   - Complex synchronization,
     off-chip data movement, high precision
4. Convolutional neural networks
   - Complex scheduling of low-precision linear algebra

- One chip to rule them all
  - Find the common denominator



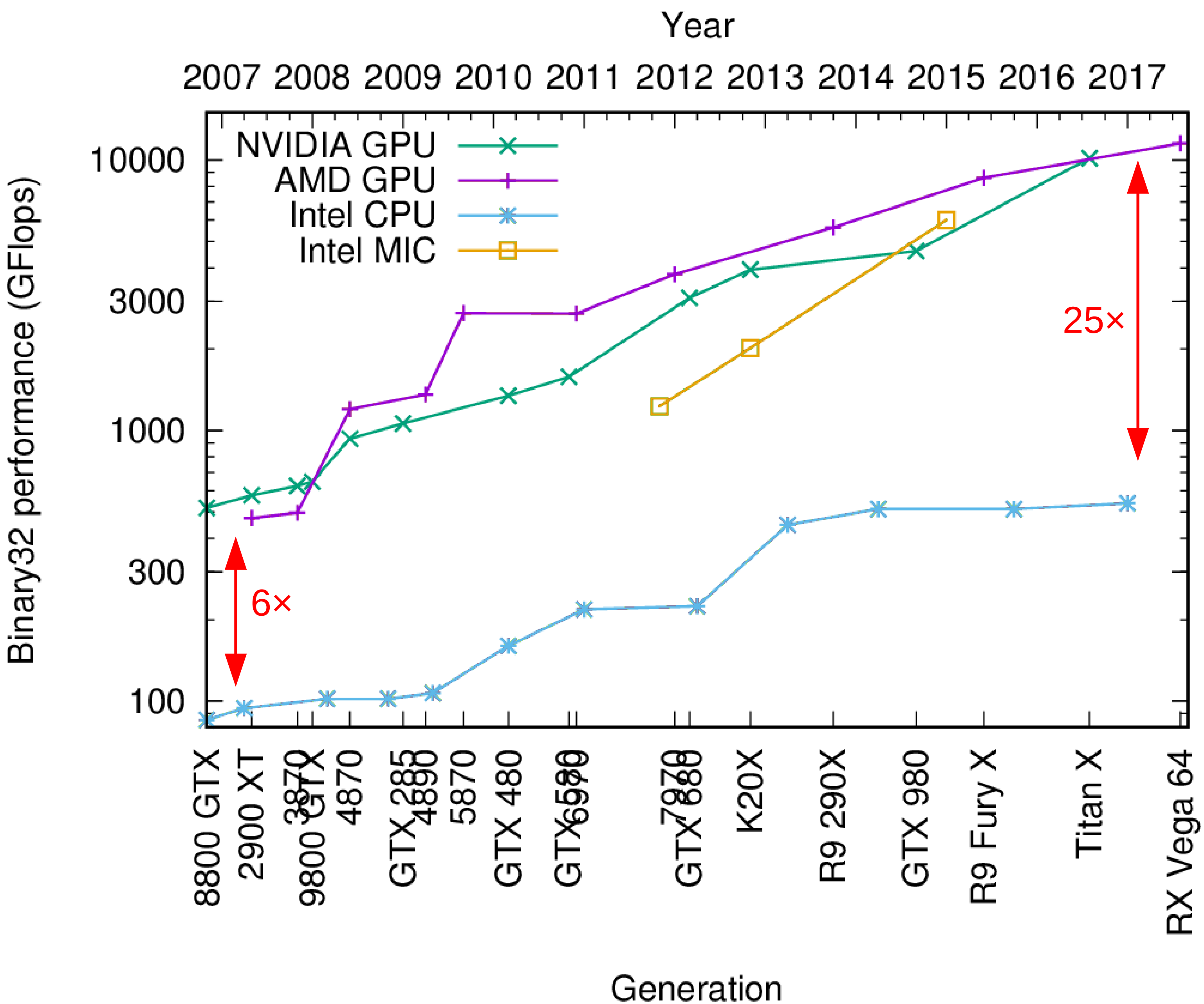




Outline

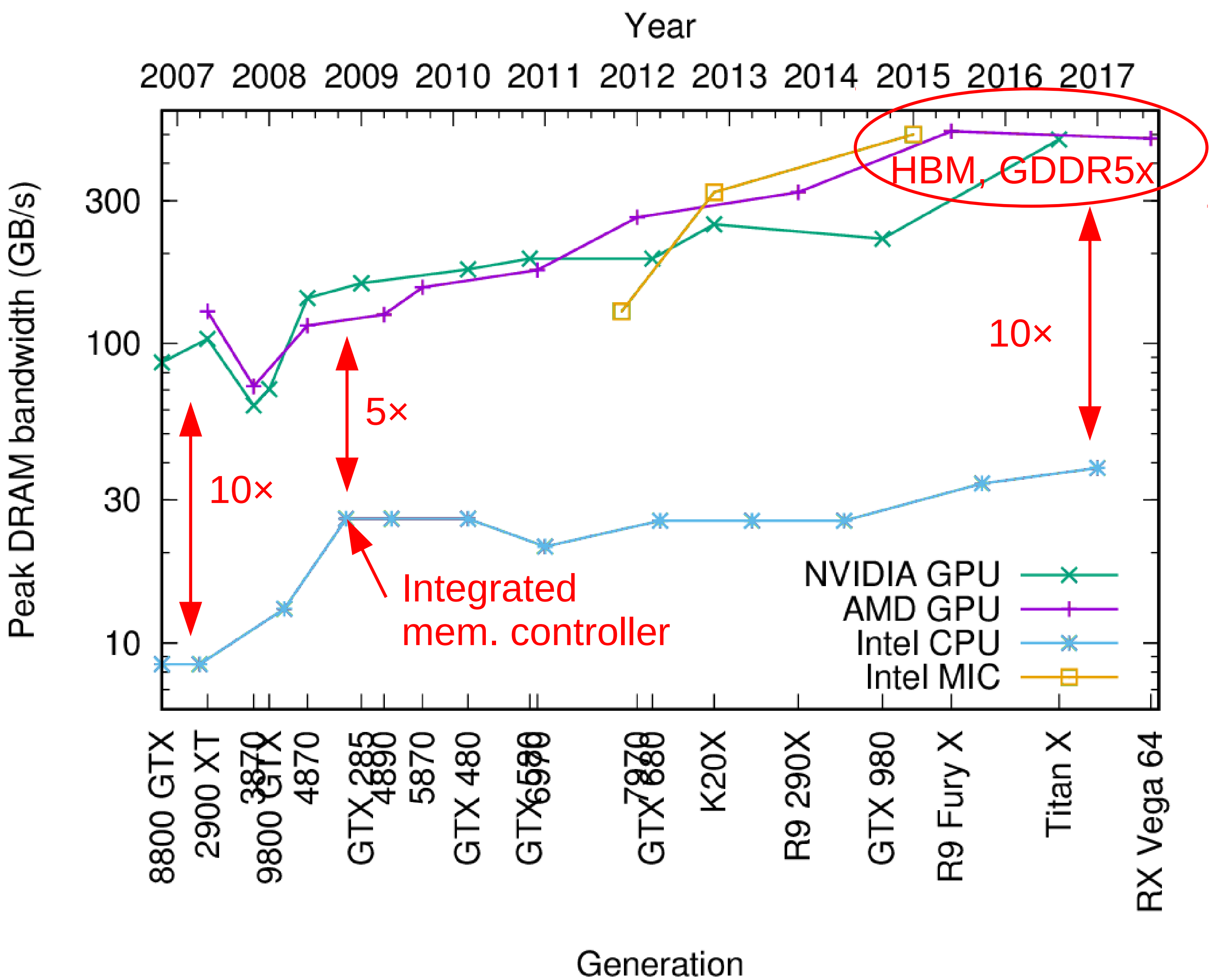
- GPU, many-core: why, what for?
  - Technological trends and constraints
  - From graphics to general purpose
  - Hardware trends
- Forms of parallelism, how to exploit them
  - Why we need (so much) parallelism: latency and throughput
  - Sources of parallelism: ILP, TLP, DLP
  - Uses of parallelism: horizontal, vertical
- Let's design a GPU!
  - Ingredients: Sequential core, Multi-core, Multi-threaded core, SIMD
  - Putting it all together
  - Architecture of current GPUs: cores, memory

Trends: compute performance

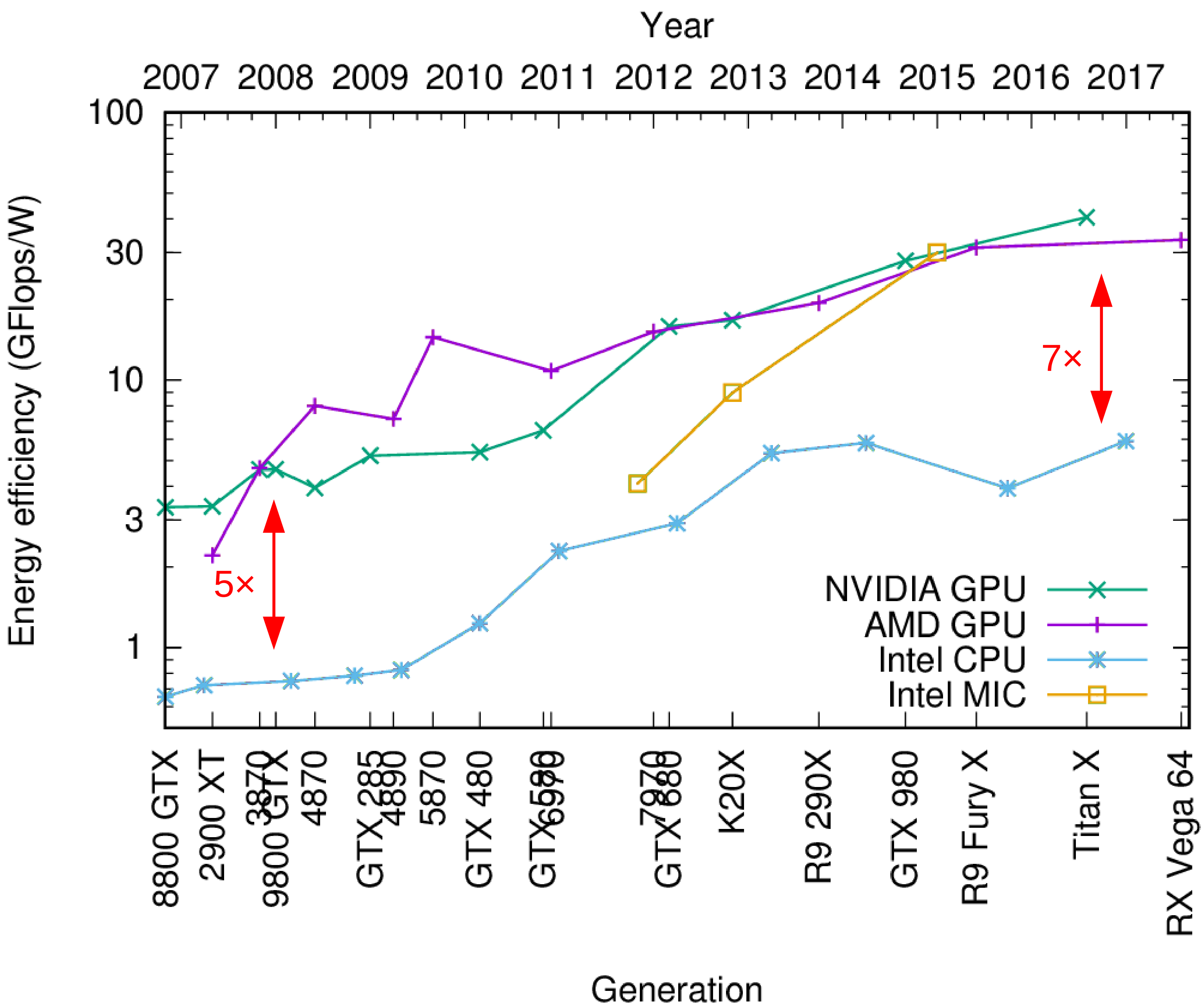


Caveat: only considers **desktop** CPUs. Gap with server CPUs is “only” 4×!

Trends: memory bandwidth



Trends: energy efficiency



Outline

- GPU, many-core: why, what for?
  - Technological trends and constraints
  - From graphics to general purpose
  - Hardware trends
- Forms of parallelism, how to exploit them
  - Why we need (so much) parallelism: latency and throughput
  - Sources of parallelism: ILP, TLP, DLP
  - Uses of parallelism: horizontal, vertical
- Let's design a GPU!
  - Ingredients: Sequential core, Multi-core, Multi-threaded core, SIMD
  - Putting it all together
  - Architecture of current GPUs: cores, memory

What is parallelism?

Parallelism: independent operations which execution can be overlapped

Operations: memory accesses or computations

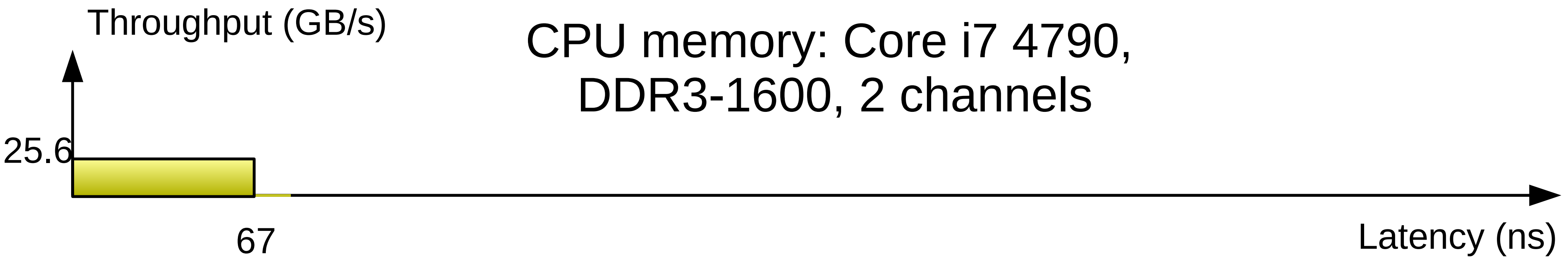
How much parallelism do I need?

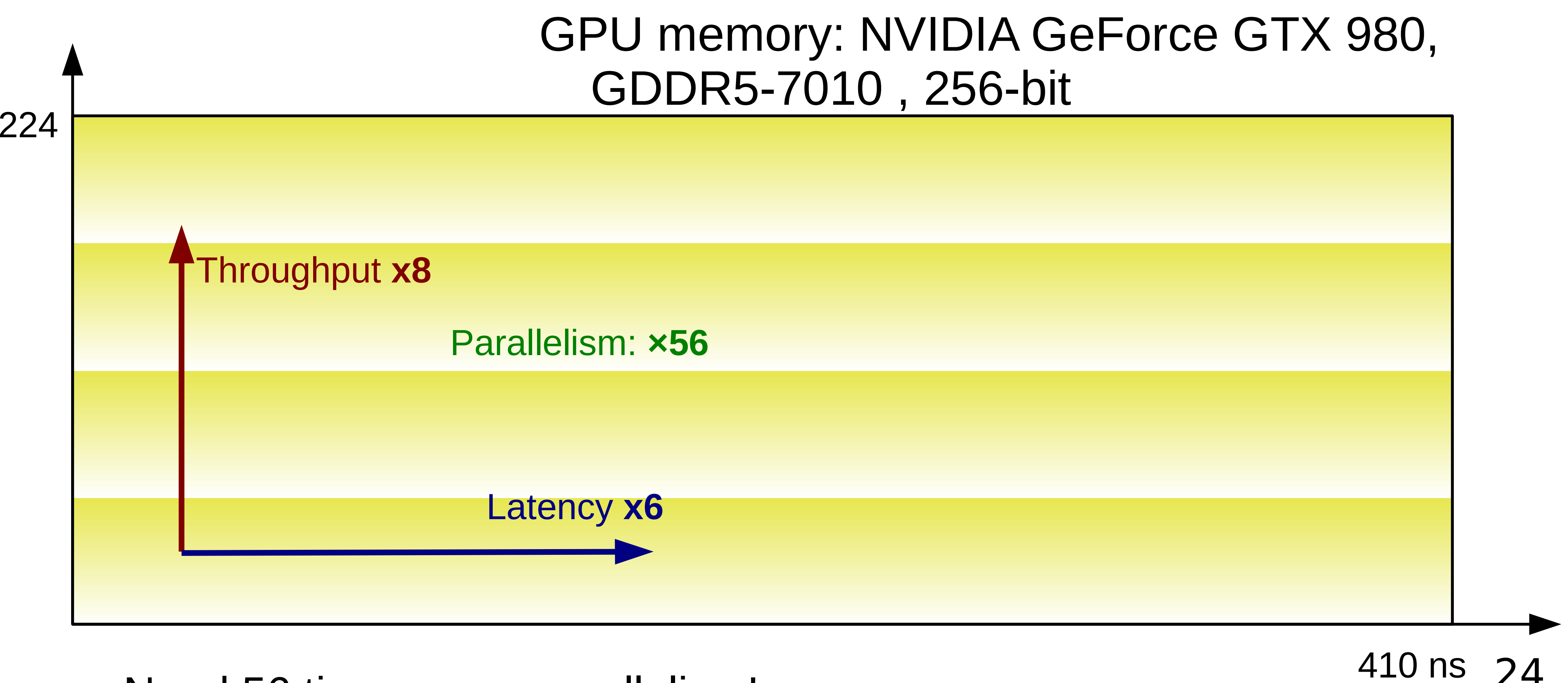
- Little's law in queuing theory
  - Average customer arrival rate **λ** ← throughput
  - Average time spent **W** ← latency
  - Average number of customers **L = λ×W** ← **Parallelism** = **throughput** × **latency**

- Units
  - For memory: **B** = **GB/s** × **ns**
  - For arithmetic: **flops** = **Gflops/s** × **ns**

J. Little. A proof for the queuing formula L= λ W. JSTOR 1961.

Throughput and latency: CPU vs. GPU

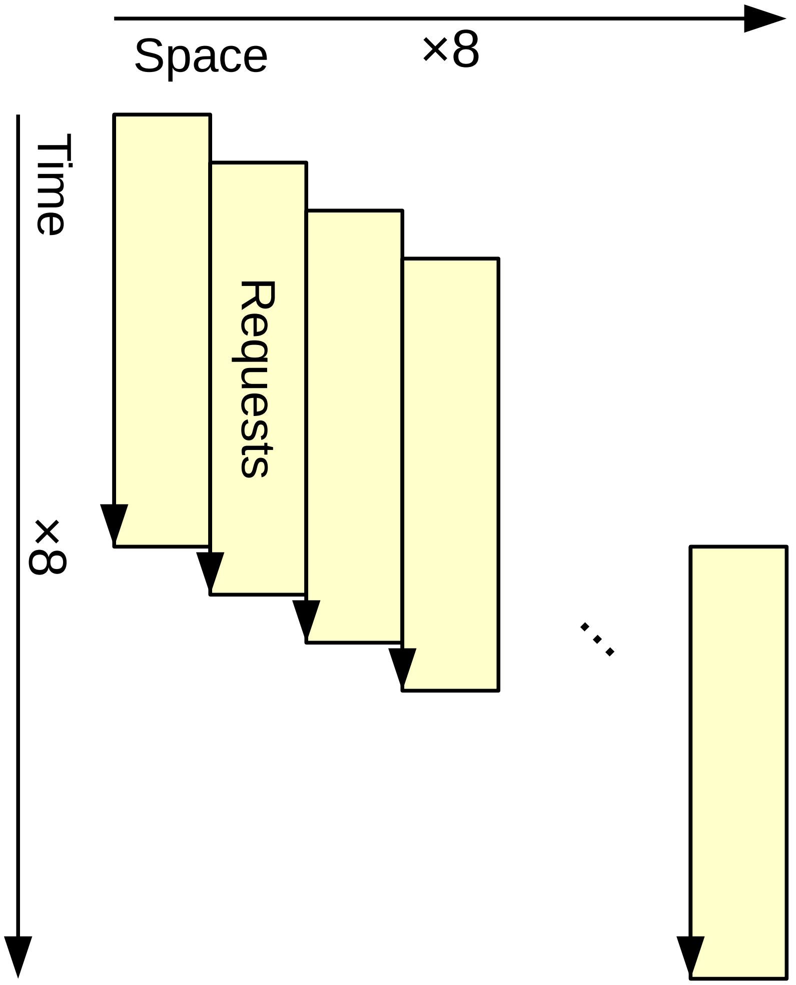




→ Need 56 times more parallelism!

Consequence: more parallelism

- GPU vs. CPU
  - 8× more parallelism to feed more units (throughput)
  - 6× more parallelism to hide longer latency
  - 56× more total parallelism
- How to find this parallelism?

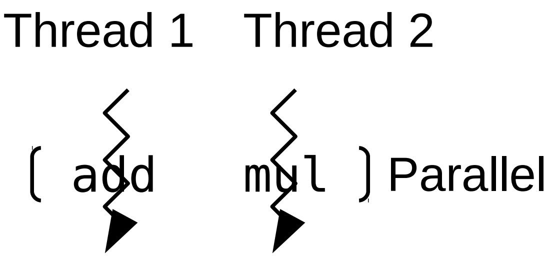


Sources of parallelism

- ILP: Instruction-Level Parallelism
  - Between independent instructions in sequential program

```
add r3 ← r1, r2  ⎫
mul r0 ← r0, r1  ⎭ Parallel
sub r1 ← r3, r0
```

- TLP: Thread-Level Parallelism
  - Between independent execution contexts: threads



- DLP: Data-Level Parallelism
  - Between elements of a vector: same operation on several elements

`vadd r←a,b`

$$\begin{array}{ccc} a_1 & a_2 & a_3 \\ + & + & + \\ b_1 & b_2 & b_3 \\ \hline r_1 & r_2 & r_3 \end{array}$$

Example: $X \leftarrow a \times X$

- In-place scalar-vector product: $X \leftarrow a \times X$

Sequential (ILP)

```
For i = 0 to n-1 do:
    X[i] ← a * X[i]
```

Threads (TLP)

```
Launch n threads:
    X[tid] ← a * X[tid]
```

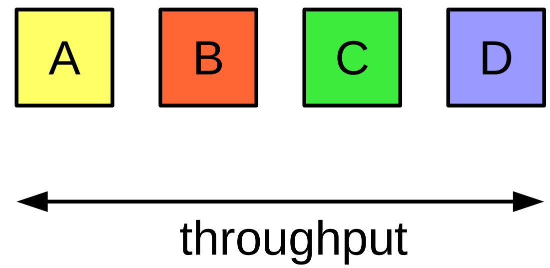
Vector (DLP)

```
    X ← a * X
```

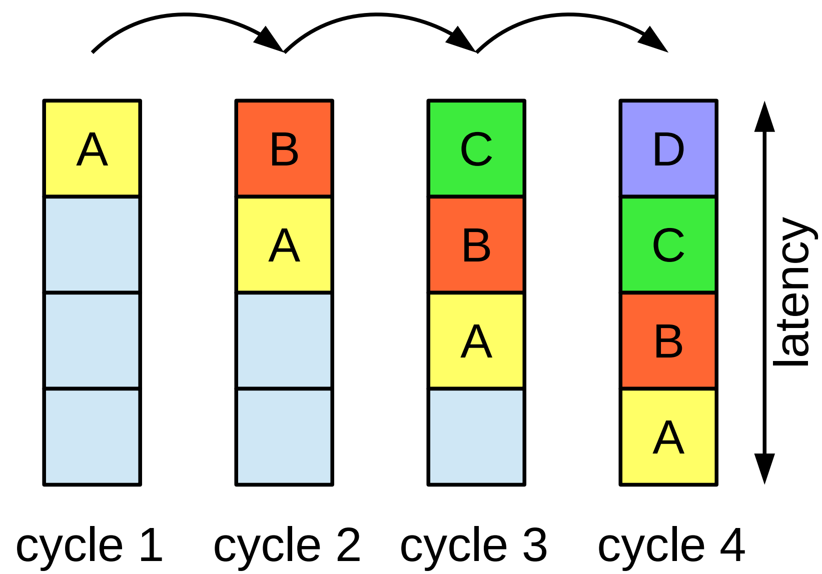
- Or any combination of the above

Uses of parallelism

- “Horizontal” parallelism for throughput
  - More units working in parallel



- “Vertical” parallelism for latency hiding
  - Pipelining: keep units busy when waiting for dependencies, memory



How to extract parallelism?

	Horizontal	Vertical
ILP	Superscalar	Pipelined
TLP	Multi-core SMT	Interleaved / switch-on-event multithreading
DLP	SIMD / SIMT	Vector / temporal SIMT

- We have seen the first row: ILP
- We will now review techniques for the next rows: TLP, DLP

Outline

- GPU, many-core: why, what for?
  - Technological trends and constraints
  - From graphics to general purpose
  - Hardware trends
- Forms of parallelism, how to exploit them
  - Why we need (so much) parallelism: latency and throughput
  - Sources of parallelism: ILP, TLP, DLP
  - Uses of parallelism: horizontal, vertical
- Let's design a GPU!
  - Ingredients: Sequential core, Multi-core, Multi-threaded core, SIMD
  - Putting it all together
  - Architecture of current GPUs: cores, memory

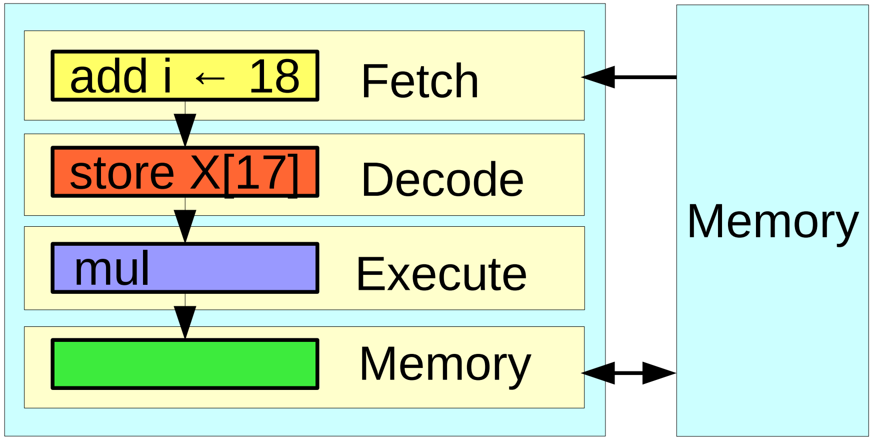
Sequential processor

```
for i = 0 to n-1
   X[i] ← a * X[i]
```

Source code

```
   move   i ← 0
loop:
   load   t ← X[i]
   mul    t ← a×t
   store X[i] ← t
   add    i ← i+1
   branch i<n? loop
```

Machine code

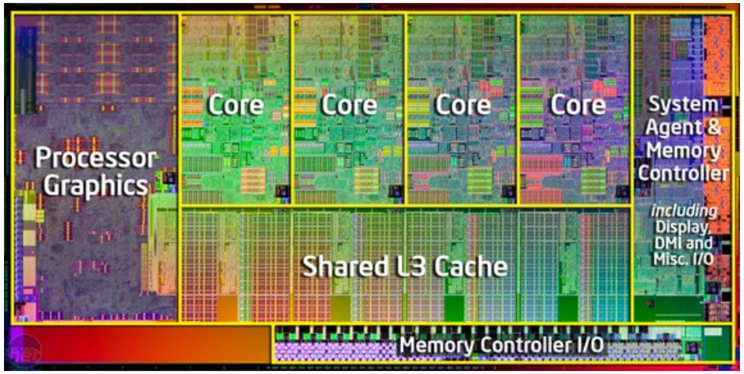


Sequential CPU

- Focuses on instruction-level parallelism
  - Exploits ILP: vertically (pipelining) and horizontally (superscalar)

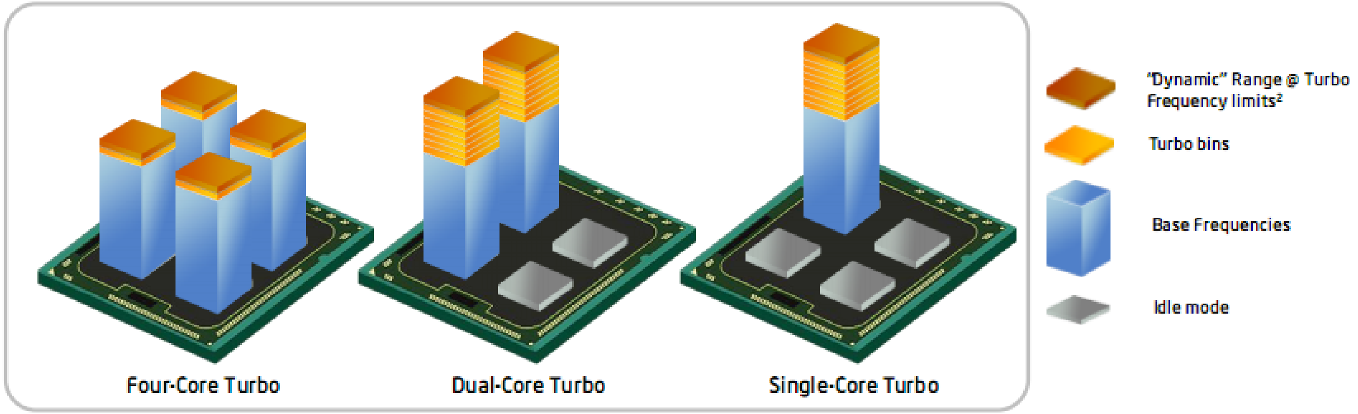
The incremental approach: multi-core

- Several processors
  on a single chip
  sharing one memory space



Intel Sandy Bridge

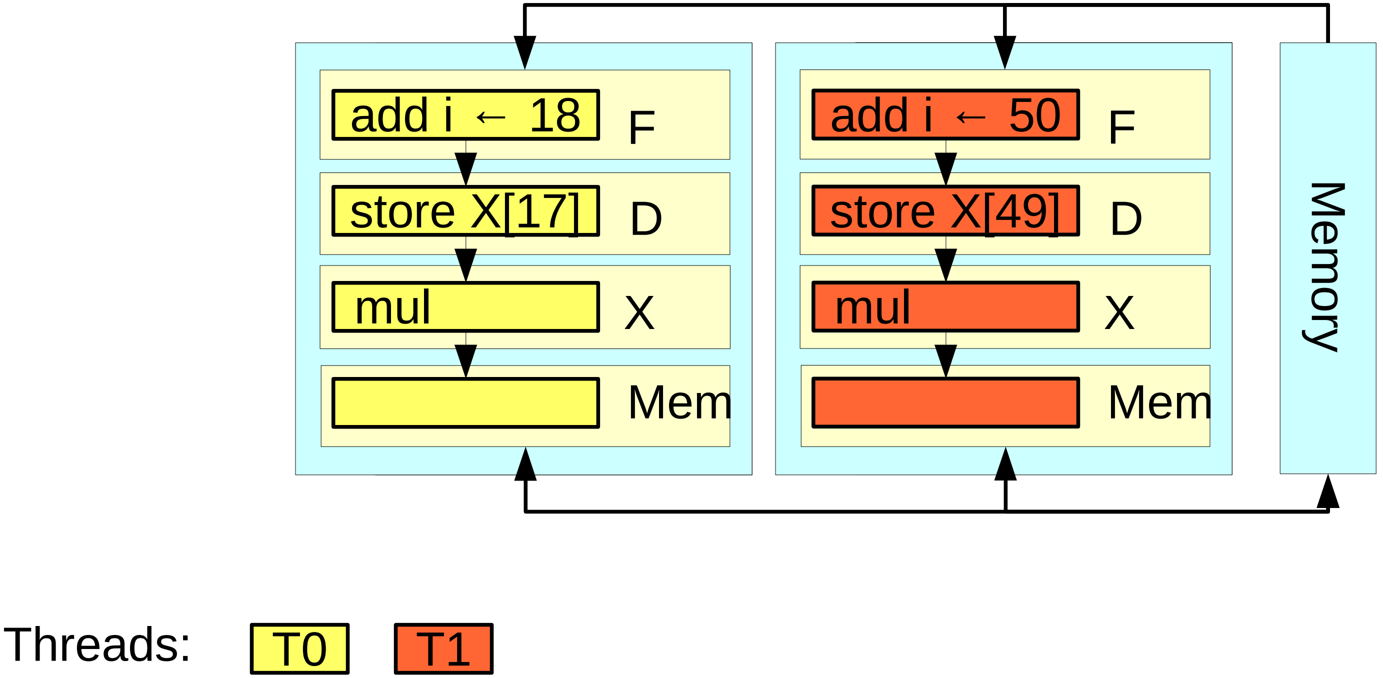
- Area: benefits from Moore's law
- Power: extra cores consume little when not in use
  - e.g. Intel Turbo Boost



Source: Intel

Homogeneous multi-core

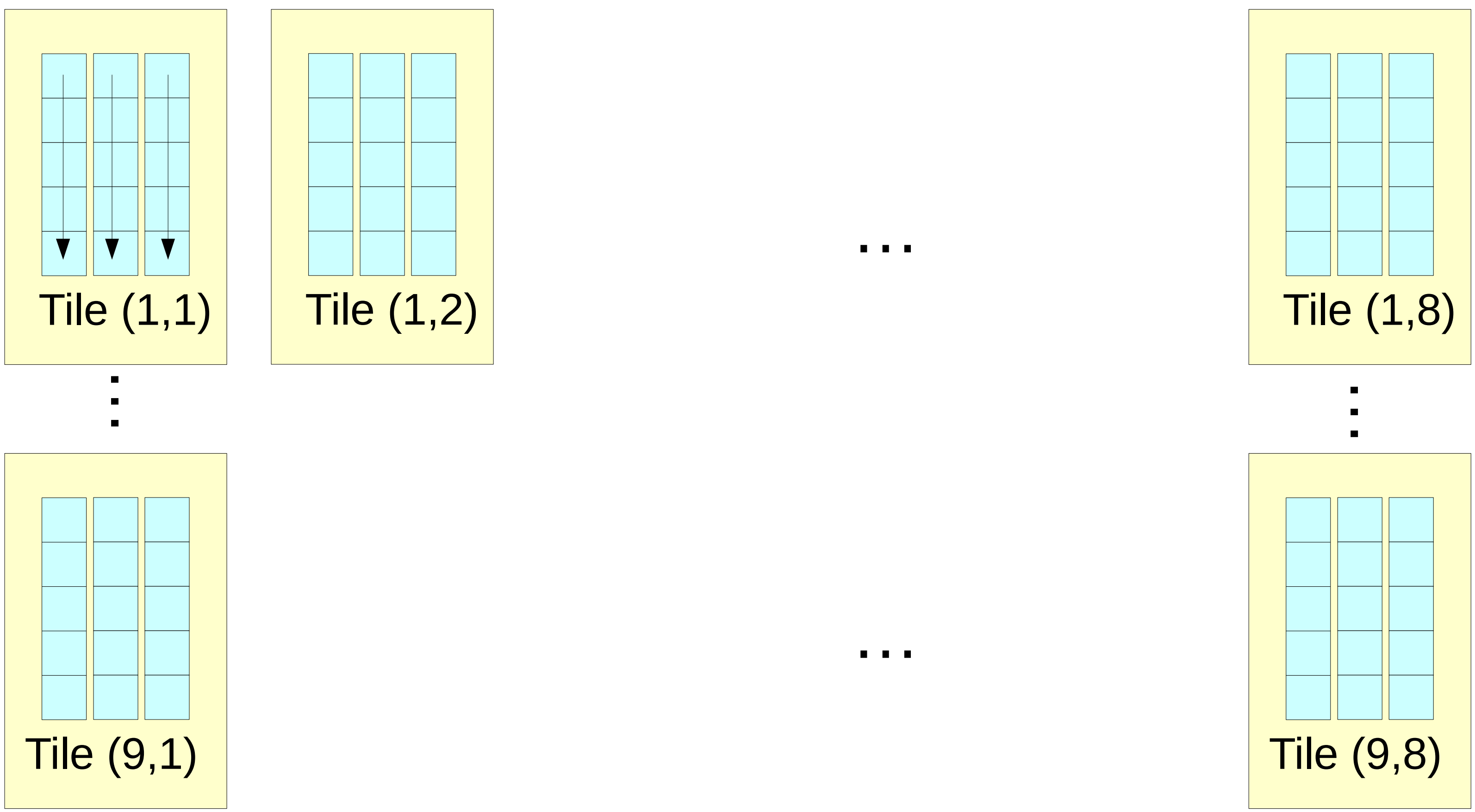
- Horizontal use of thread-level parallelism



- Improves peak throughput

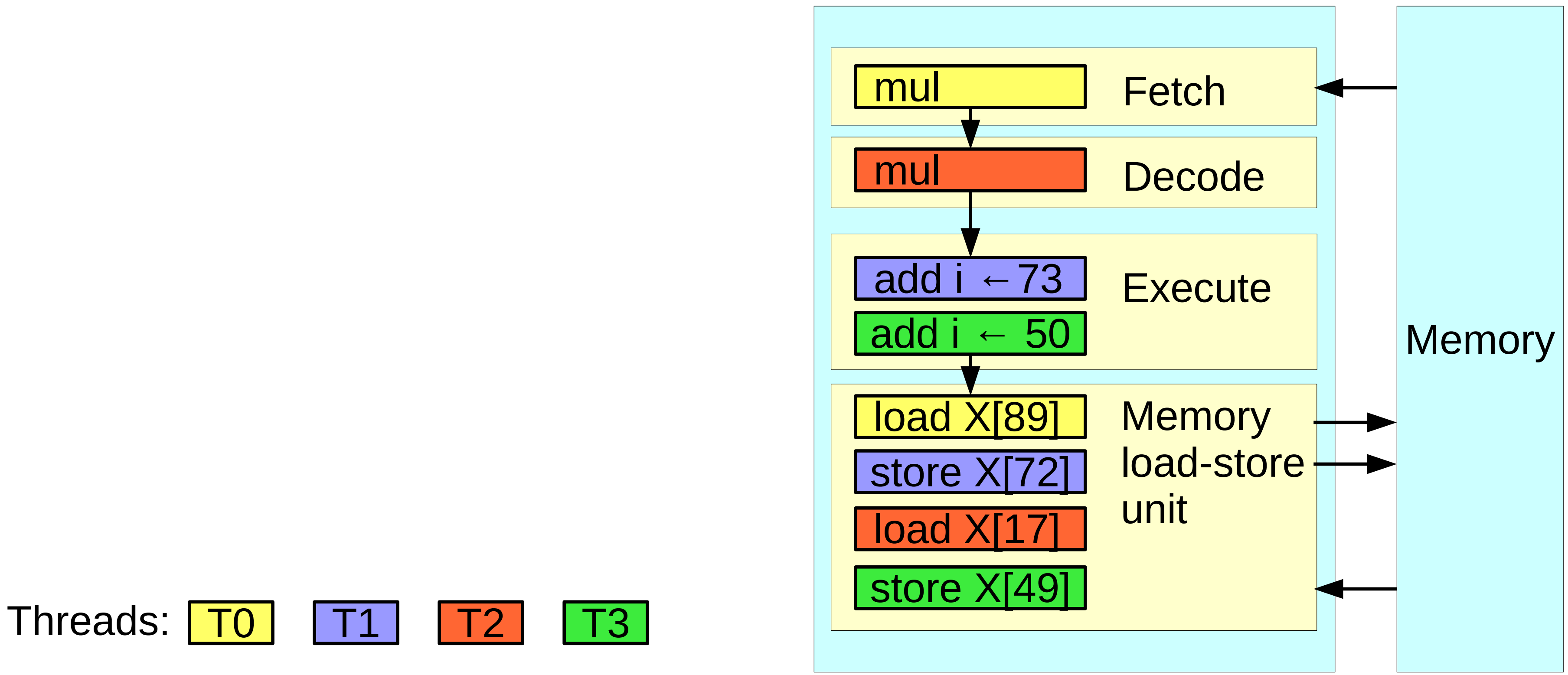
Example: Tilera Tile-GX

- Grid of (up to) 72 tiles
- Each tile: 3-way VLIW processor, 5 pipeline stages, 1.2 GHz



Interleaved multi-threading

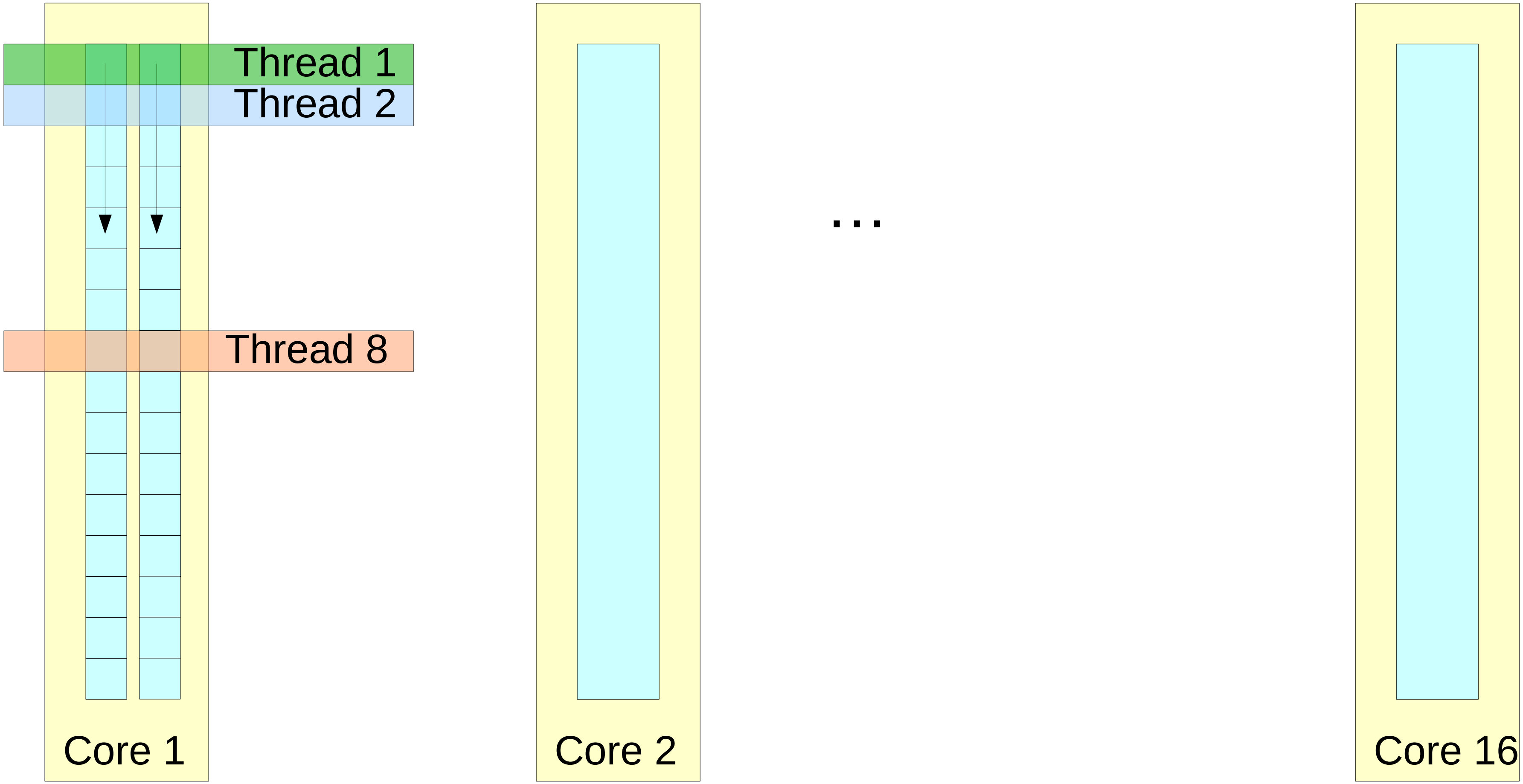
- Vertical use of thread-level parallelism



- Hides latency thanks to explicit parallelism improves achieved throughput

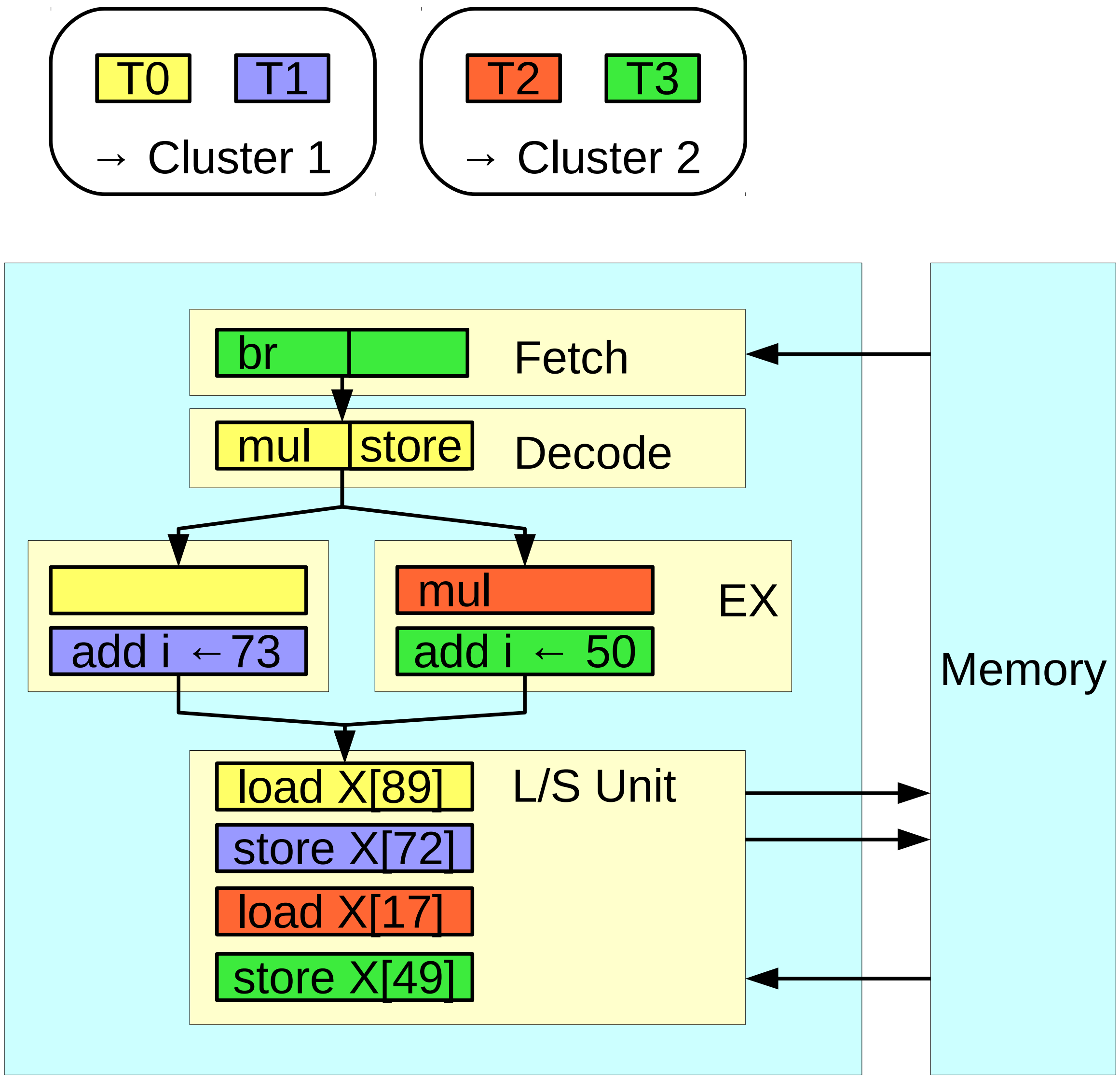
Example: Oracle Sparc T5

- 16 cores / chip
- Core: out-of-order superscalar, 8 threads
- 15 pipeline stages, 3.6 GHz



Clustered multi-core

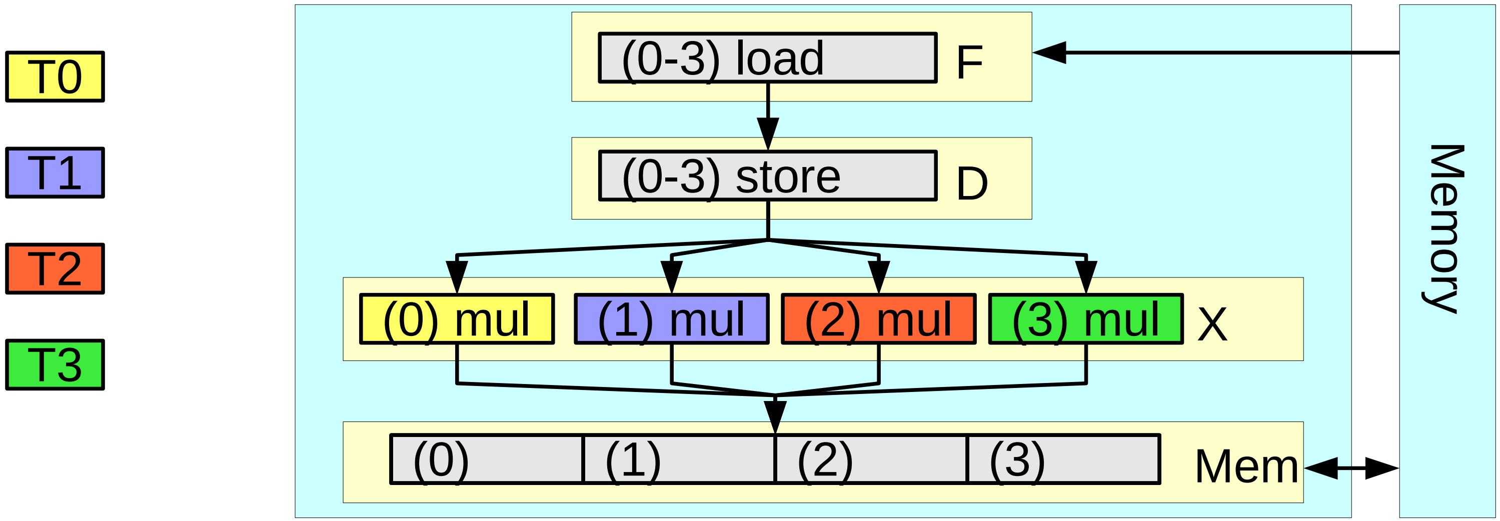
- For each individual unit, select between
  - Horizontal replication
  - Vertical time-multiplexing
- Examples
  - Sun UltraSparc T2, T3
  - AMD Bulldozer
  - IBM Power 7, 8, 9



- Area-efficient tradeoff
- Blurs boundaries between cores

Implicit SIMD

- Factorization of fetch/decode, load-store units
  - Fetch 1 instruction on behalf of several threads
  - Read 1 memory location and broadcast to several registers



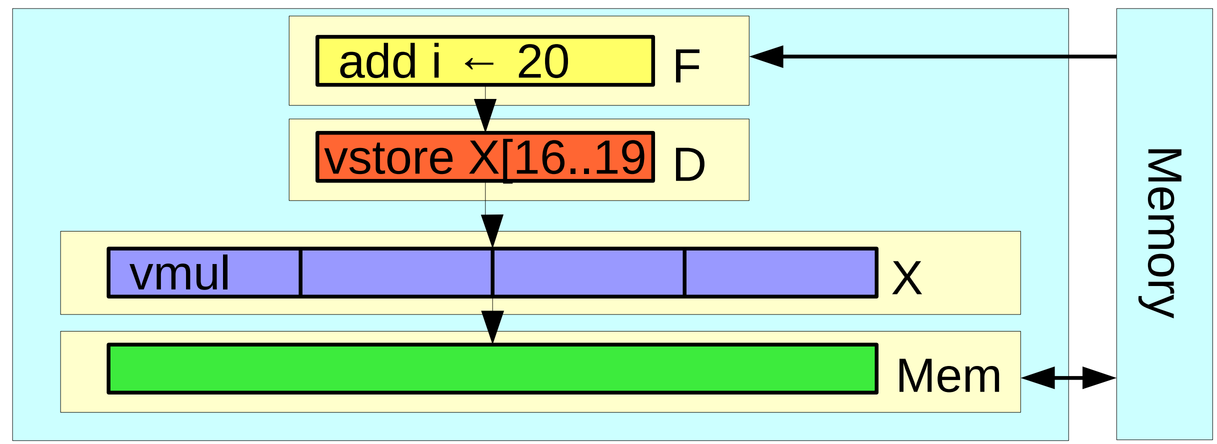
- In NVIDIA-speak
  - SIMT: Single Instruction, Multiple Threads
  - Convoy of synchronized threads: *warp*
- Extracts DLP from multi-thread applications

Explicit SIMD

- Single Instruction Multiple Data
- Horizontal use of data level parallelism

```
loop:
    vload  T ← X[i]
    vmul   T ← a×T
    vstore X[i] ← T
    add    i ← i+4
    branch i<n? loop
```

Machine code



SIMD CPU

- Examples
  - Intel MIC (16-wide)
  - AMD GCN GPU (16-wide×4-deep)
  - Most general purpose CPUs (4-wide to 16-wide)

Outline

- GPU, many-core: why, what for?
  - Technological trends and constraints
  - From graphics to general purpose
- Forms of parallelism, how to exploit them
  - Why we need (so much) parallelism: latency and throughput
  - Sources of parallelism: ILP, TLP, DLP
  - Uses of parallelism: horizontal, vertical
- Let's design a GPU!
  - Ingredients: Sequential core, Multi-core, Multi-threaded core, SIMD
  - Putting it all together
  - Architecture of current GPUs: cores, memory

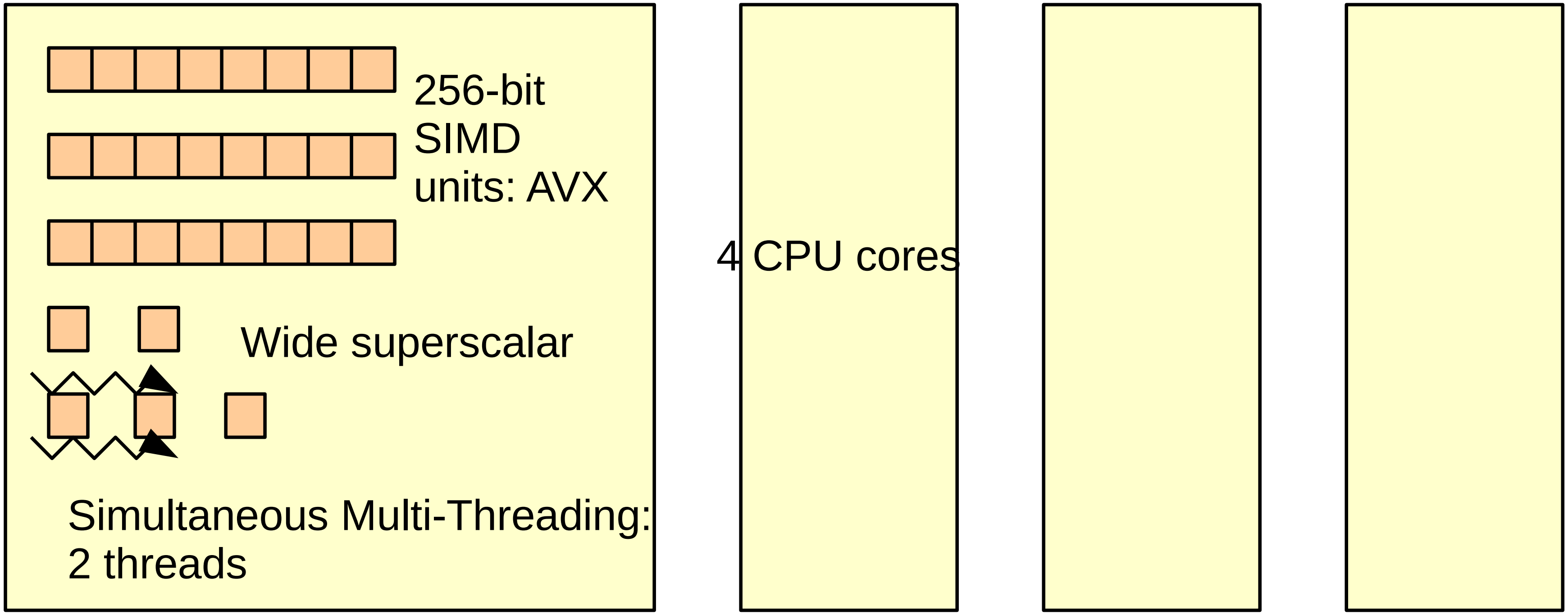
Hierarchical combination

There is no silver bullet!

- All of these techniques face the law of diminishing returns
  - More cores → complex interconnect, hard to maintain cache coherence
  - More threads/core → more register and cache pressure
  - Wider vectors → more performance lost to irregular control/data flow
- Both CPUs and GPUs combine techniques
  - Multiple cores
  - Multiple threads/core
  - SIMD units

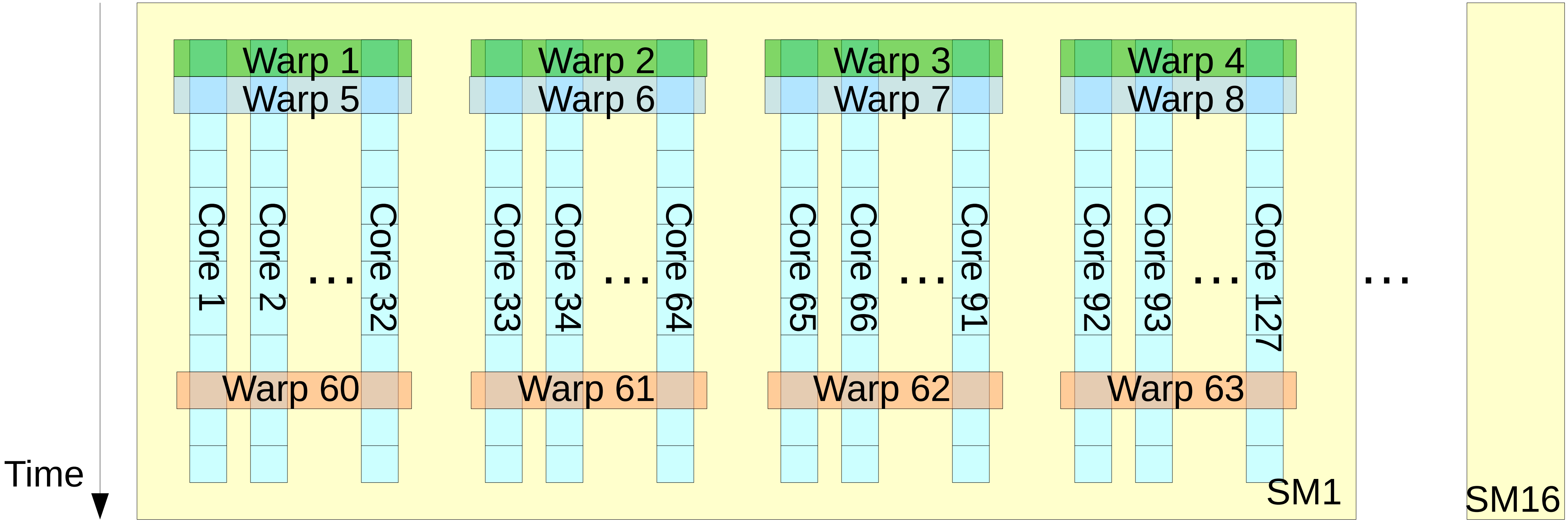
Example CPU: Intel Core i7

- Is a wide superscalar, but has also
  - Multicore
  - Multi-thread / core
  - SIMD units
- Up to 117 operations/cycle from 8 threads



Example GPU: NVIDIA GeForce GTX 980

- SIMT: warps of 32 threads
- 16 SMs / chip
- 4×32 cores / SM, 64 warps / SM



- 4612 Gflop/s
- Up to 32768 threads in flight

Taxonomy of parallel architectures

	Horizontal	Vertical
ILP	Superscalar / VLIW	Pipelined
TLP	Multi-core SMT	Interleaved / switch-on-event multithreading
DLP	SIMD / SIMT	Vector / temporal SIMT

Classification: multi-core

Intel Haswell

	Horizontal	Vertical
ILP	8	
TLP	4	2
DLP	8	

SIMD (AVX)   Cores   Hyperthreading

Fujitsu SPARC64 X

	Horizontal	Vertical
ILP	8	
TLP	16	2
DLP	2	

General-purpose multi-cores:
balance ILP, TLP and DLP

IBM Power 8

	Horizontal	Vertical
ILP	10	
TLP	12	8
DLP	4	

Oracle Sparc T5

	Horizontal	Vertical
ILP	2	
TLP	16	8
DLP		

Cores   Threads

Sparc T:
focus on TLP

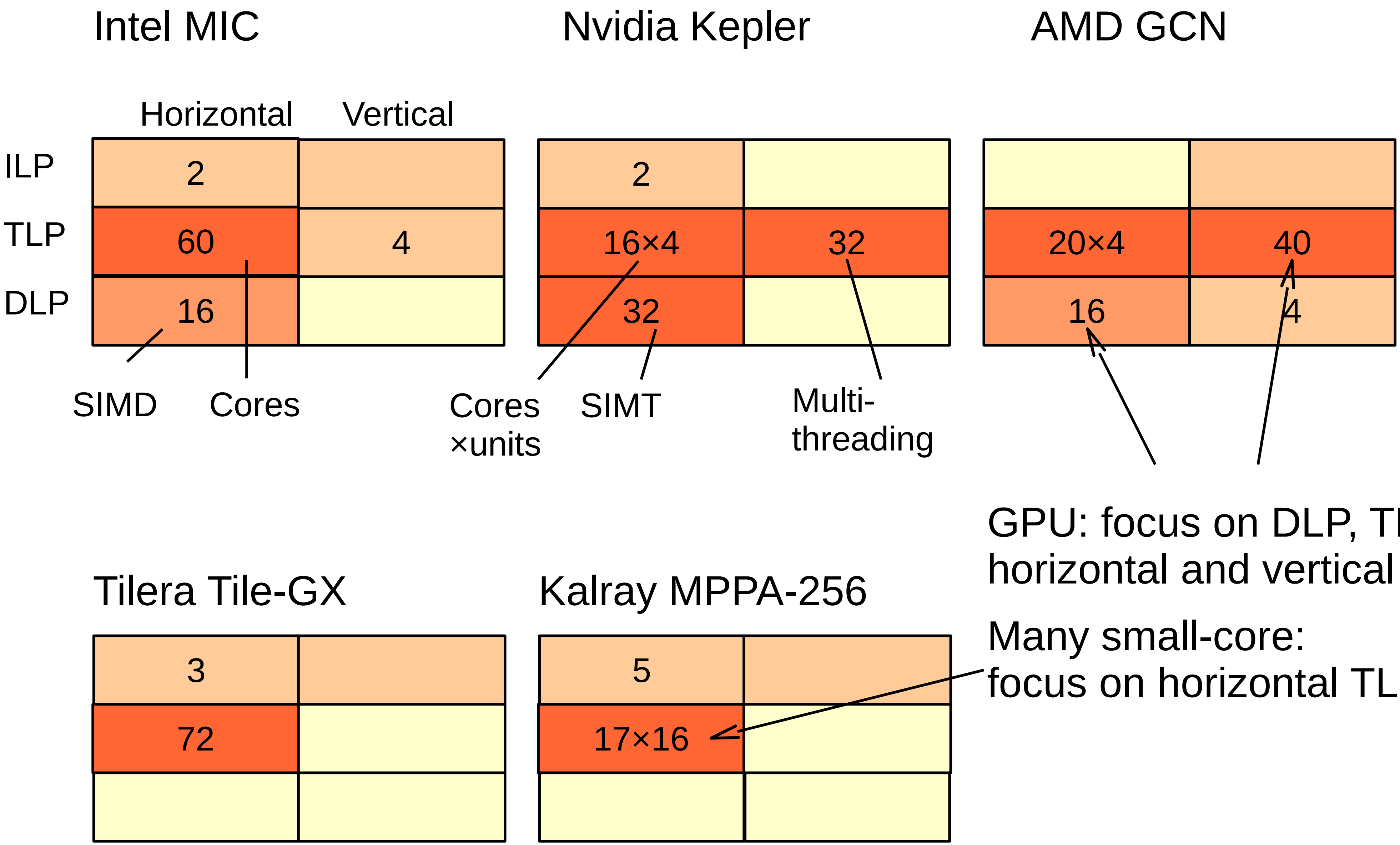
How to read the table

- Given an application with known ILP, TLP, DLP
  how much throughput / latency hiding can I expect?
  - For each cell, take minimum of existing parallelism
    and hardware capability
  - The column-wise product gives throughput / latency hiding

	Sequential code no TLP, no DLP	Horizontal	Vertical
ILP	10	min(8, 10) =8	
TLP	1	min(4, 1)=1	2
DLP	1	min(8, 1)=1	

Max throughput = 8×1×1
for this application
Peak throughput = 8×4×8
that can be achieved
→Can only hope for ~3% of peak performance!

Classification: GPU and many small-core



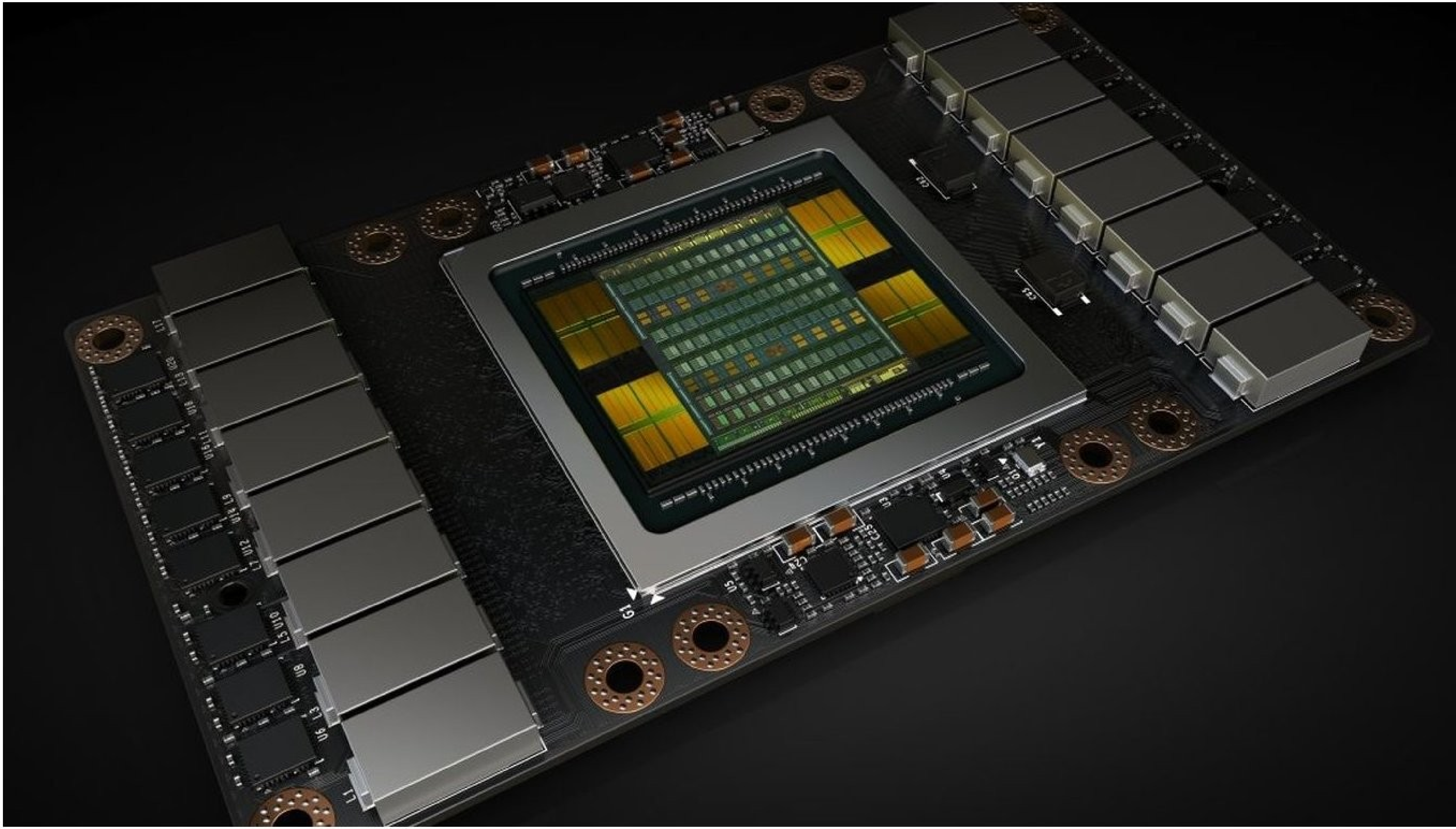
GPU: focus on DLP, TLP horizontal and vertical

Many small-core: focus on horizontal TLP

Outline

- GPU, many-core: why, what for?
  - Technological trends and constraints
  - From graphics to general purpose
- Forms of parallelism, how to exploit them
  - Why we need (so much) parallelism: latency and throughput
  - Sources of parallelism: ILP, TLP, DLP
  - Uses of parallelism: horizontal, vertical
- Let's design a GPU!
  - Ingredients: Sequential core, Multi-core, Multi-threaded core, SIMD
  - Putting it all together
  - Architecture of current GPUs: cores, memory

What is inside a graphics card?

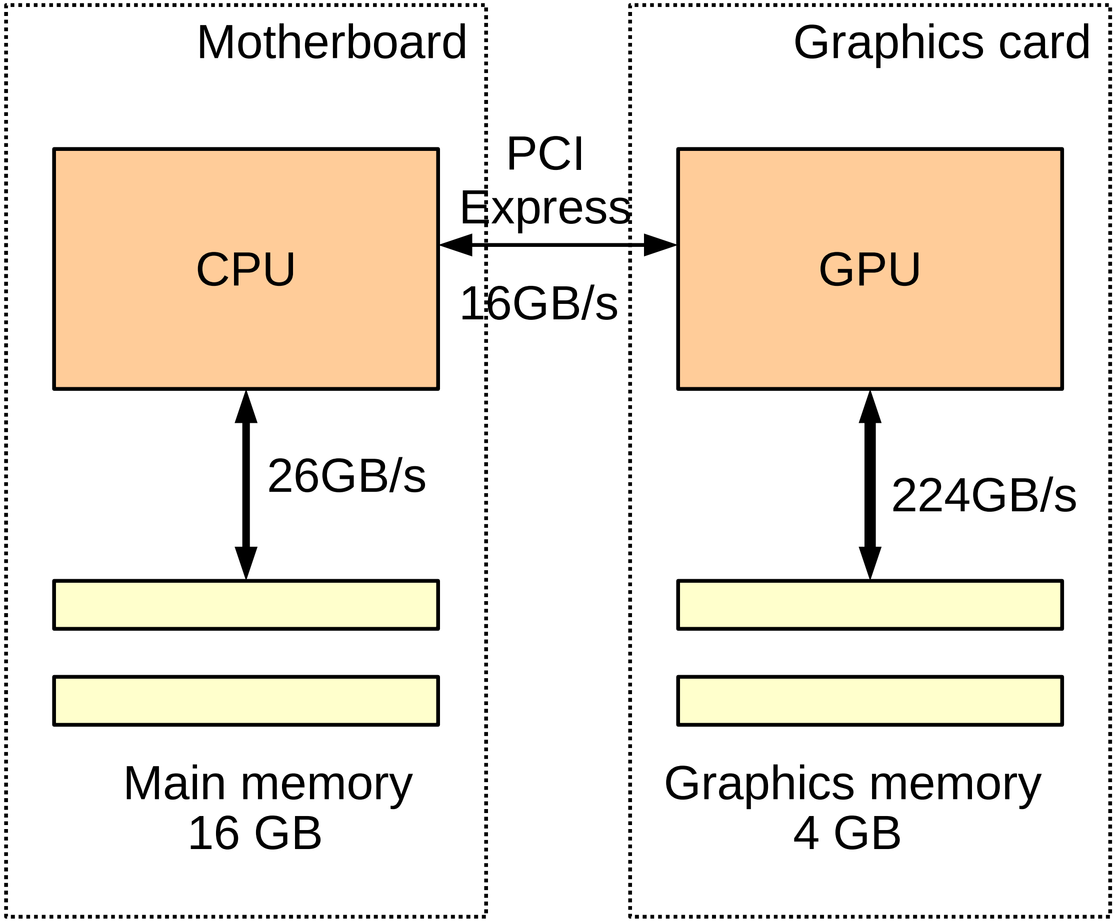


NVIDIA Volta V100 GPU. Artist rendering!

External memory: discrete GPU

Classical CPU-GPU model

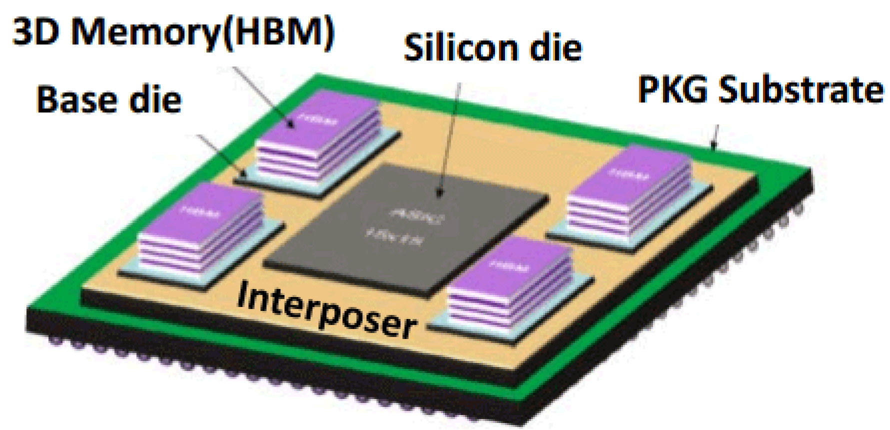
- Split memory spaces
- Need to transfer data explicitly
- Highest bandwidth from GPU memory
- Transfers to main memory are slower



Example configuration:
Intel Core i7 4790, Nvidia GeForce GTX 980

Discrete GPU memory technology

- GDDR5, GDDR5x
  - Qualitatively like regular DDR
  - Optimized for high frequency at the expense of latency and cost
  - e.g. *Nvidia Titan X*: 12 chip pairs x 32-bit bus × 10 GHz → 480 GB/s
- High-Bandwidth Memory (HBM)
  - On-package stacked memory on silicon interposer



  - Shorter traces, wider bus, lower frequency: more energy-efficient
  - Limited capacity and high cost
  - e.g. *AMD R9 Fury X*: 4× 4-high stack × 1024-bit × 1 GHz → 512 GB/s

Maximizing memory bandwidth

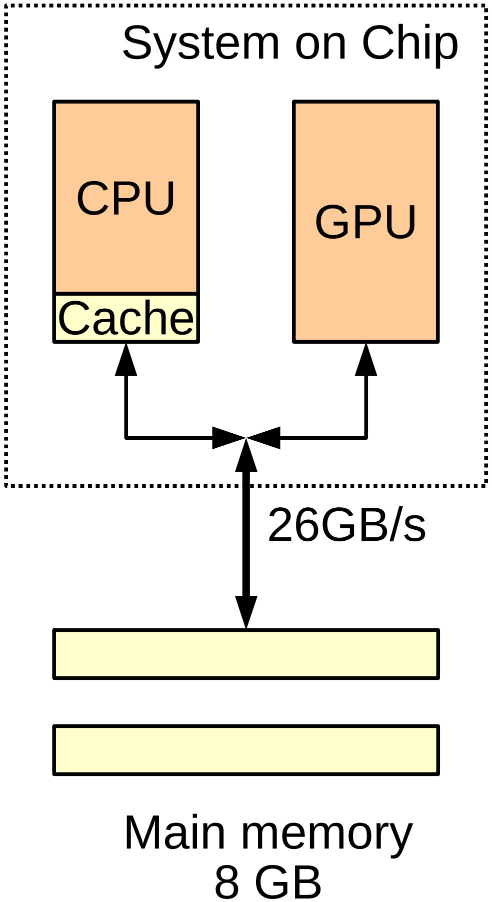
Memory bandwidth is a critical resource

- Cache hierarchy reduces throughput demand on main memory
  - Bandwidth amplification
  - Less energy per access
- Hardware data compression in caches and memory
  - Lossy compression for textures (under programmer control)
  - Lossless compression for framebuffer, z-buffer…

External memory: embedded GPU

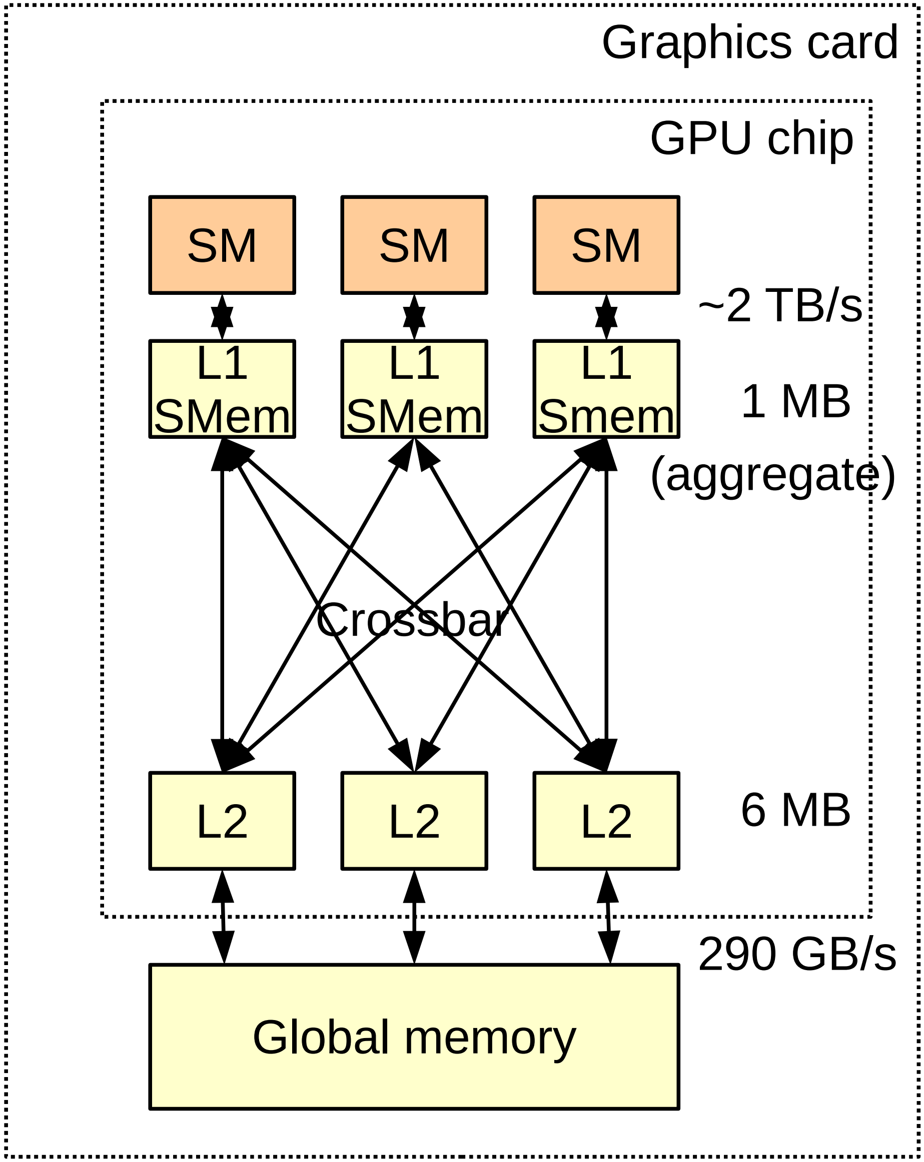
Most GPUs today are integrated

- Same physical memory
- May support memory coherence
  - GPU can read directly from CPU caches
- More contention on external memory



GPU high-level organization

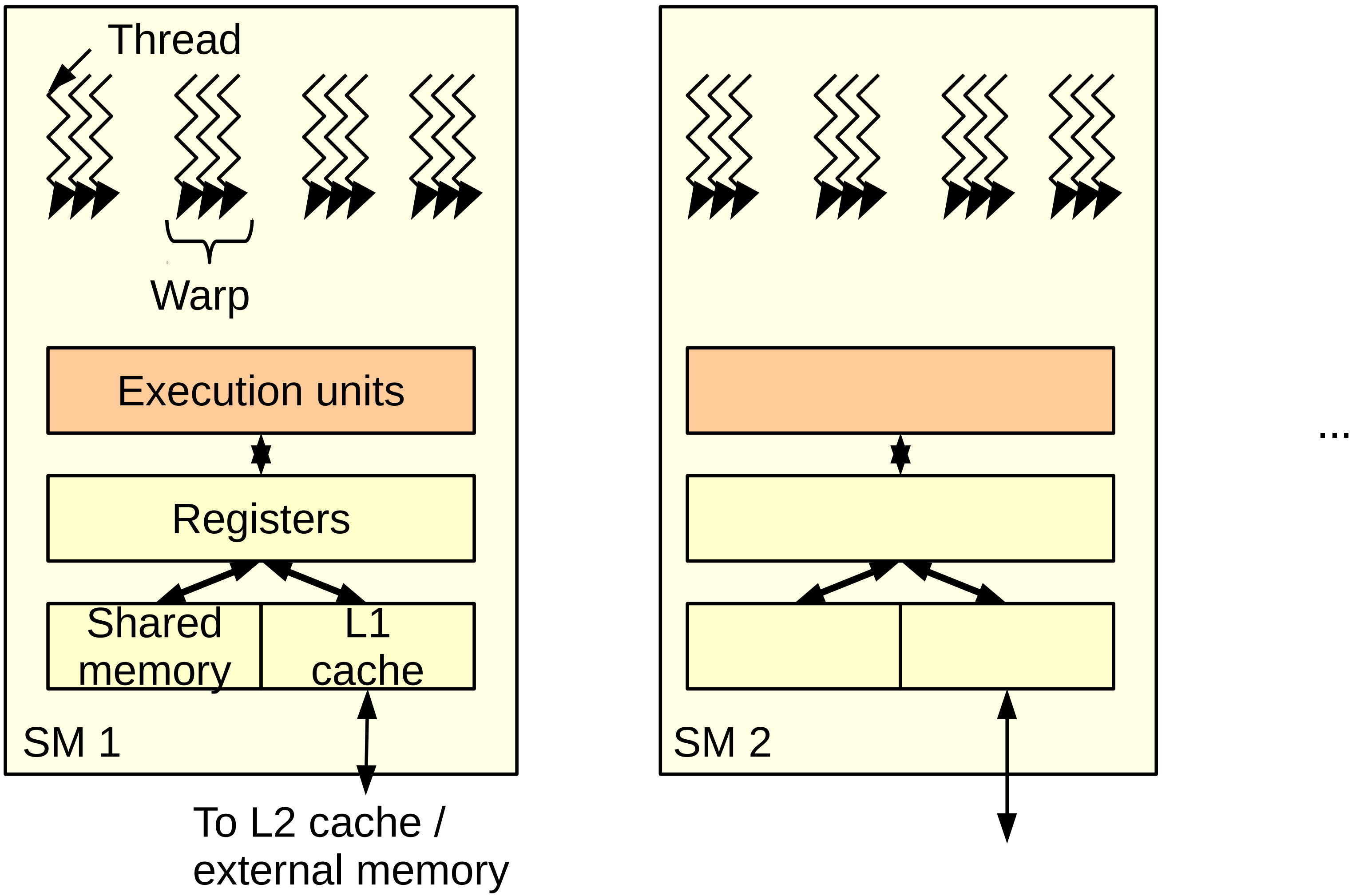
- Processing units
  - Streaming Multiprocessors (SM) in Nvidia jargon
  - Compute Unit (CU) in AMD's
  - Closest equivalent to a CPU core
  - Today: from 1 to 20 SMs in a GPU
- Memory system: caches
  - Keep frequently-accessed data
  - Reduce throughput demand on main memory
  - Managed by hardware (L1, L2) or software (Shared Memory)



GPU processing unit organization

Each SM is a highly-multithreaded processor

- Today: 24 to 48 warps of 32 threads each
  → ~1K threads on each SM, ~10K threads on a GPU



GPU: on-chip memory

- Conventional wisdom
  - Cache area in CPU vs. GPU according to the NVIDIA CUDA Programming Guide:

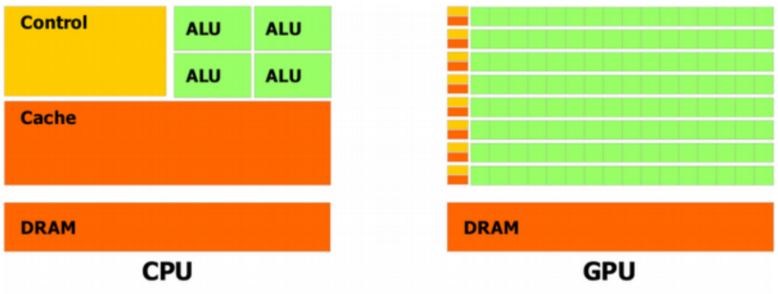
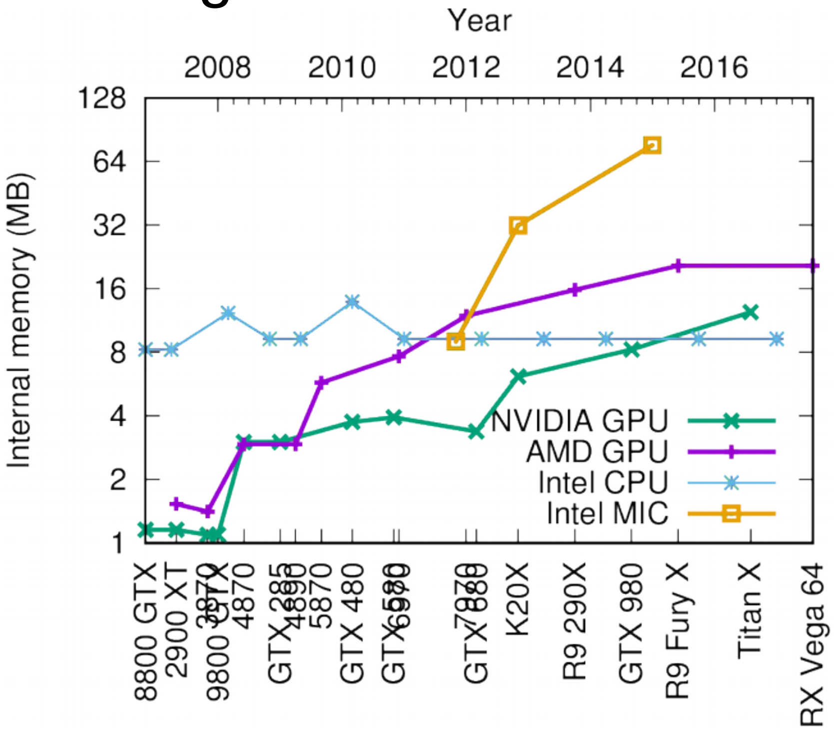


Figure 1-2. The GPU Devotes More Transistors to Data Processing

- But... if we include registers:



- GPUs have more internal memory than desktop CPUs

Registers: CPU vs. GPU

- Registers keep the contents of local variables
- Typical values

	CPU	GPU
Registers/thread	32	32
Registers/core	256	65536
Read / Write ports	10R/5W	2R/1W

- GPU: many more registers, but made of simpler memory

The locality dilemma

- More cores → higher communication latency

  - Solution 1: bigger caches
    (general-purpose multi-cores, Intel MIC)
  - Solution 2: more threads / core
    (GPUs, Sparc T)
    Need extra memory for thread state
    → more registers, bigger caches
  - Solution 3: programmer-managed communication
    (many small cores)

→ Bigger cores

→ More specialized

Where are we heading?

Alternatives for future many-cores

- A **unified** world
  General-purpose multi-cores continue to exploit more ILP, TLP and DLP
  Eventually replace all special-purpose many-cores

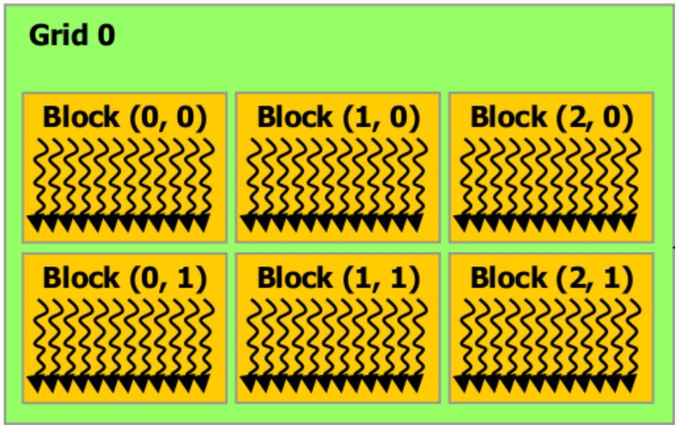
- A **specialized** world
  Varieties of many-cores continue evolving independently
  Co-existance of GPU for graphics, many-core for HPC, many-thread for servers…

- A **heterogeneous** world
  Special-purpose many-cores co-exist within the same chip
  Multi-core CPU + GPU + many-core accelerator…

Next time: SIMT control flow management

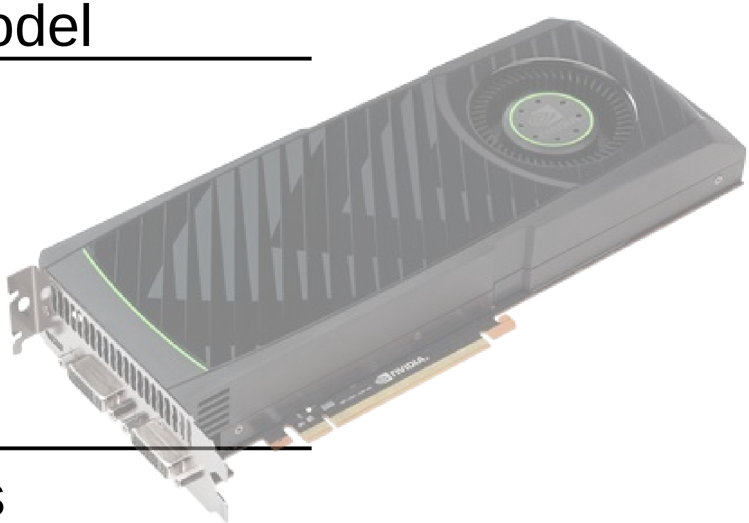
Software

```
__global__ void scale(float a, float * X)
{
    unsigned int tid;
    tid = blockIdx.x * blockDim.x
          + threadIdx.x;
    X[tid] = a * X[tid];
}
```

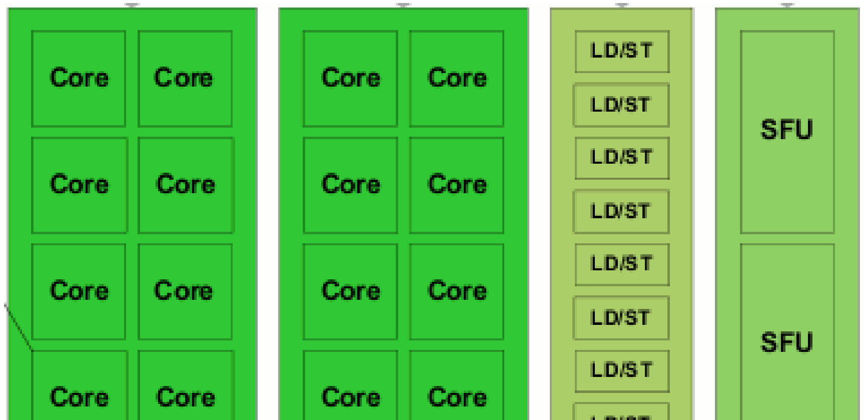


Architecture: **multi-thread** programming model

Dark magic!



Hardware datapaths: **SIMD** execution units



Hardware