Know Your Hardware

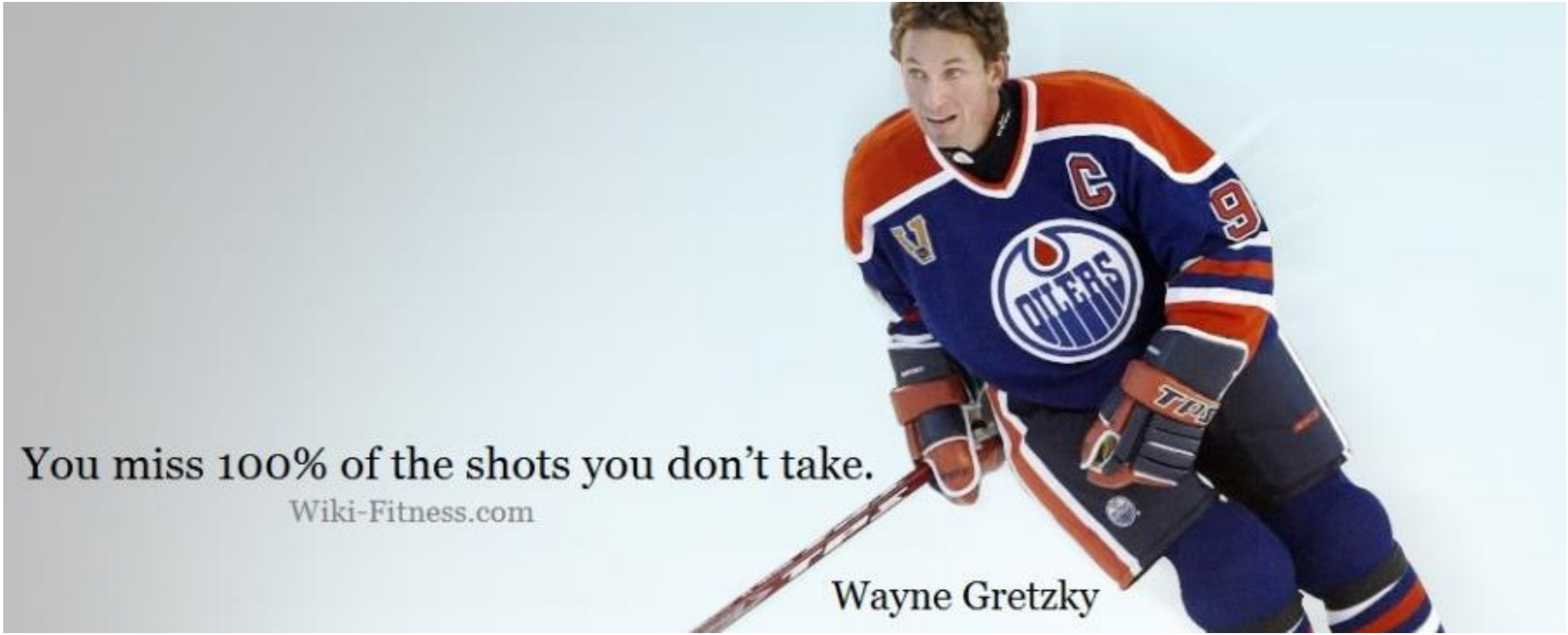
And why you should care

Máté Ferenc Nagy-Egri
Wigner GPU Lab

Table of Contents

- Consumers
  - Consumerism
  - Nomenclatura
  - Features
- Developers
  - What is inside the box?
  - How parallel is parallel?

Ask questions



And all those cryptic numbers

NOMENCLATURE

Consumerism

- Gone are the days when one bought a boiler/washing machine/television and it worked for 10+ years
  - Manufacturers don't have good bussiness models for products that may function for decades
- In Livermore CA, fire dept. #6 there is lightbulb that has been in non-stop operation since 1901
  - How much would such a lightbulb cost today, if they manufactured one?

Consumerism

- When it comes to electronics, there is a strong „hype” built around products
- Consumers are rallied into thinking that the latest significantly outperforms the previous generation of products
- In actual news:

Honest slogans

iT's New.
Buy iT.

„Think different”

- Apple has mastered this technique
- Founder Steve Jobs has converted the tech company into a religion
- Number wars, features, capabilities don’t really matter
- People feel an uncontrollable urge to buy, no matter the added value of the „new”

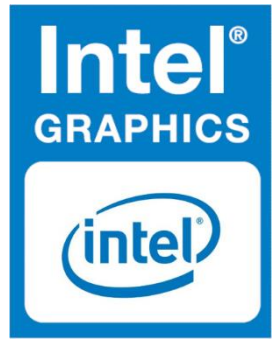
Cogito ergo sum

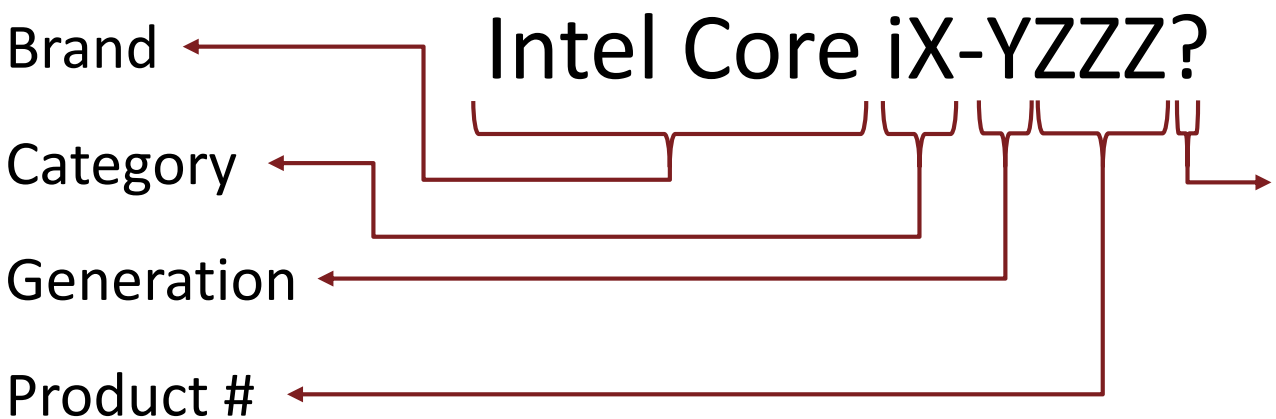
- We work with computers every day!
- Let's try to be a little more educated in what vendors are trying to sell us
  - Don't be mistaken! Affiliates of MediaMarkt/BestByte et al. Not only have their own interest (stock sweep), but are often also uneducated in this regard

„It has an i7 in it, so you can easily run any game in Ultra settings on this." – random tech supermarket expert

Intel product nomenclatura

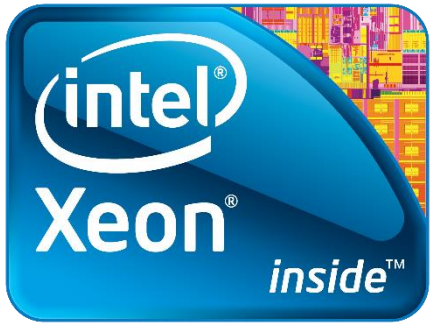
- Intel rates it's consumer product line into 4-5 categories for easier understanding
  - i7 are the best manufactured chips available
  - i5 chips usually have some cores turned off
  - i3 usually lack some key feature (HT, vT)
  - Pentium/Celeron are suited for browsing, office





- Y, Extremely low power
- U, (Ultra-)low power
- K, Unlocked (Overclock ready)
- T, Power optimized lifestyle
- H, High-performance
- Q, Quad-core

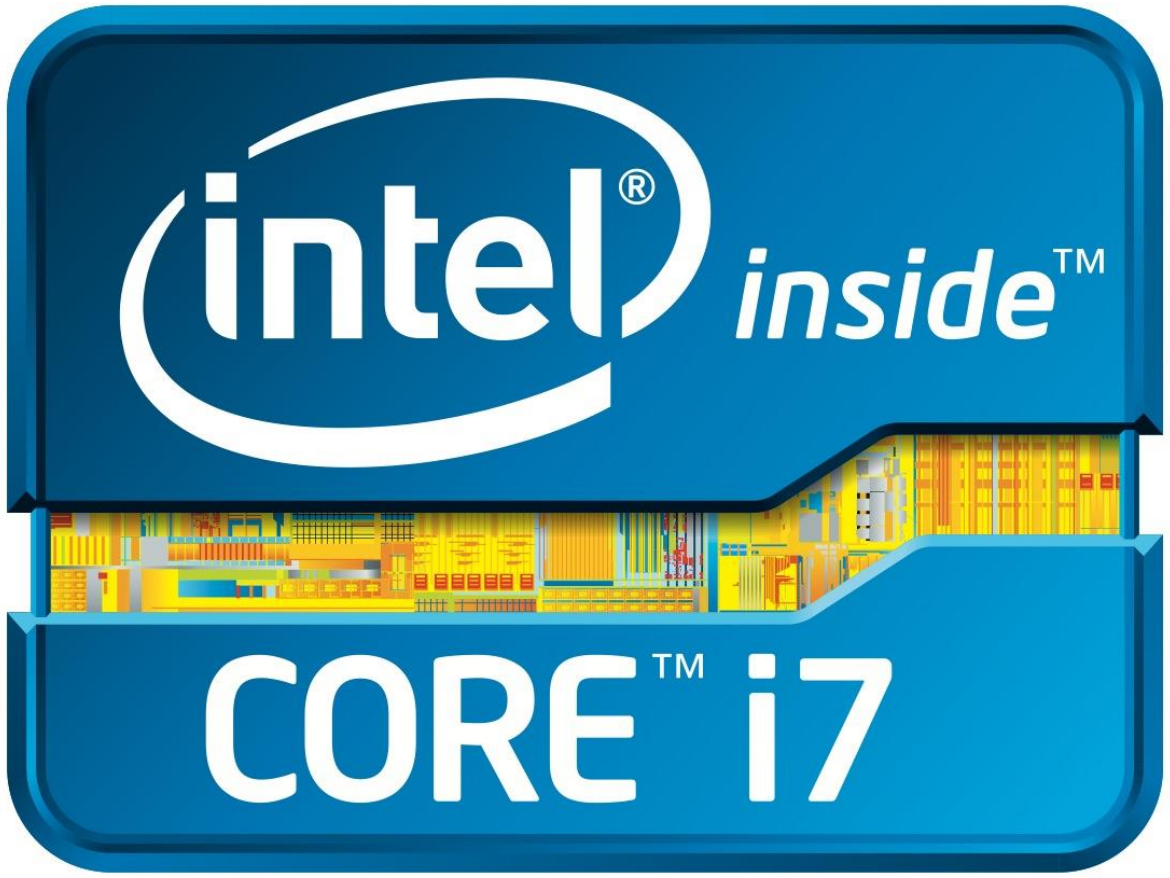
Intel product nomencaltura



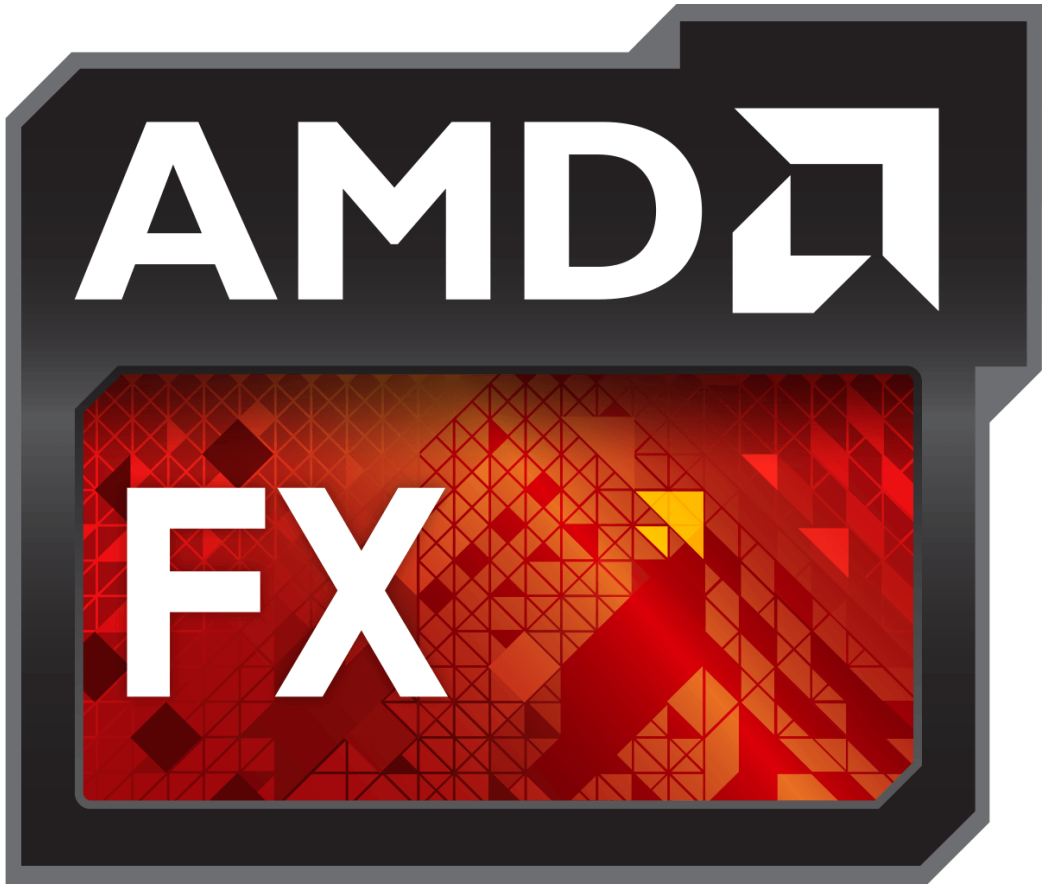
- Intel Xeon is the brand of professional grade products
  - They can be found in both servers and workstations, both desktop and mobile
- Xeon Phi is Intel's approach to many-core computing (see later)
- Intel Iris Pro Graphics is the name of high-performance integrated graphics products found on select high-end chips

Know the diffrence

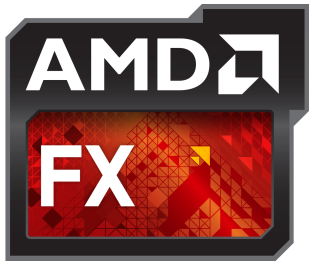
High-end Intel processor brand

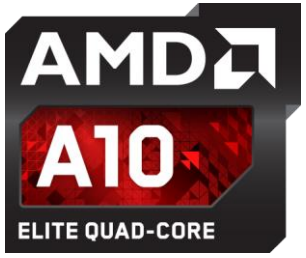


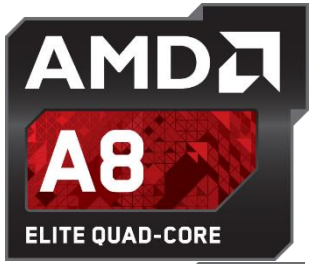
High-end AMD processor brand

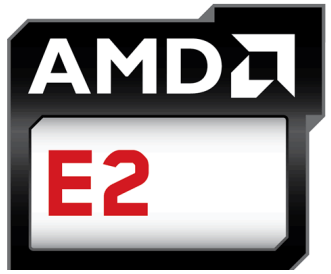


Cost effective alternative

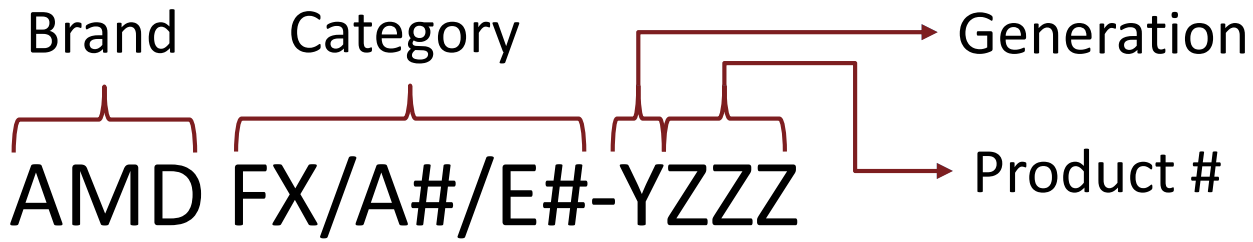




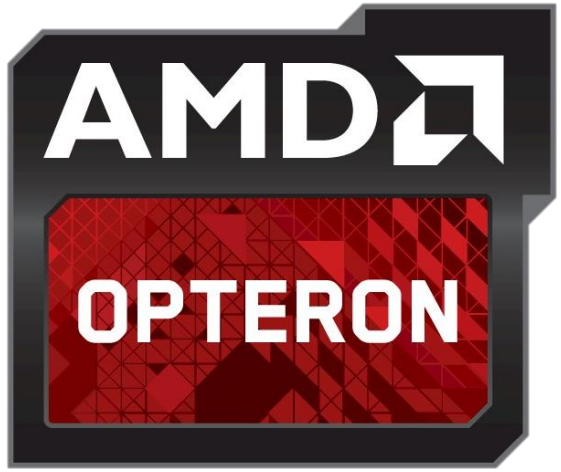


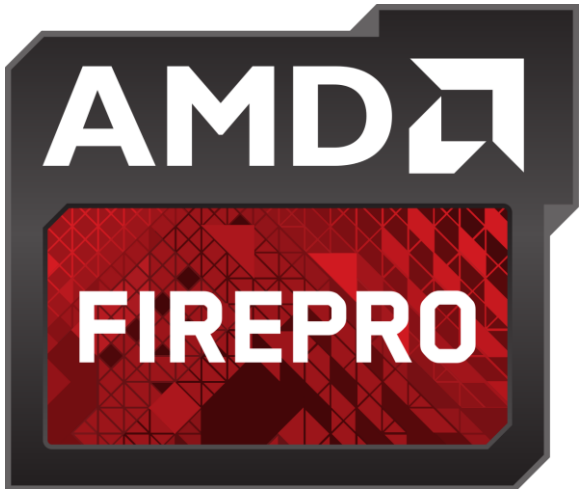


- While AMD used to be head on with Intel processors with their Phenom product line, currently at most it is a capable, cost-effective alternative
- Because of the acquisution of Ati Technologies Inc. in 2006, the integrated graphics system is world leading in both features and performance.



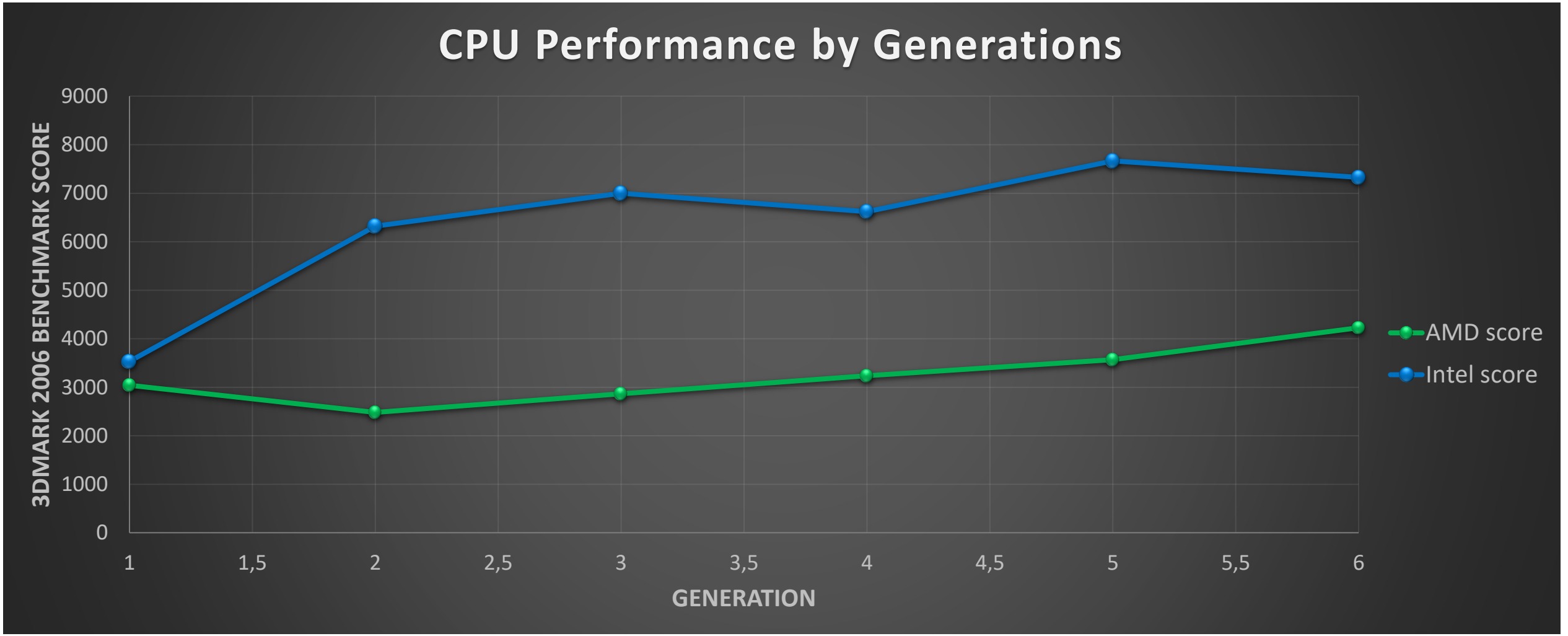
Professional compute





- AMD Opteron is the branding of server and workstation class products
- FirePro is the branding of professional grade graphics
  - Longer warranty
  - ECC VRAM
  - Long term support
  - Priority bug fixing for drivers
- Why does naming conventions matter?

Do generations matter?



Beyond bare numbers

- How strong of a processor do I need if I want to watch FullHD movies? (Both downloaded or streamed)
- It doesn't matter, even a smartphone can do it!
  - With „recent" explosion in display and media resolution, CPU-based decoding is borderline impossible. There is fixed-function hardware to do video decoding

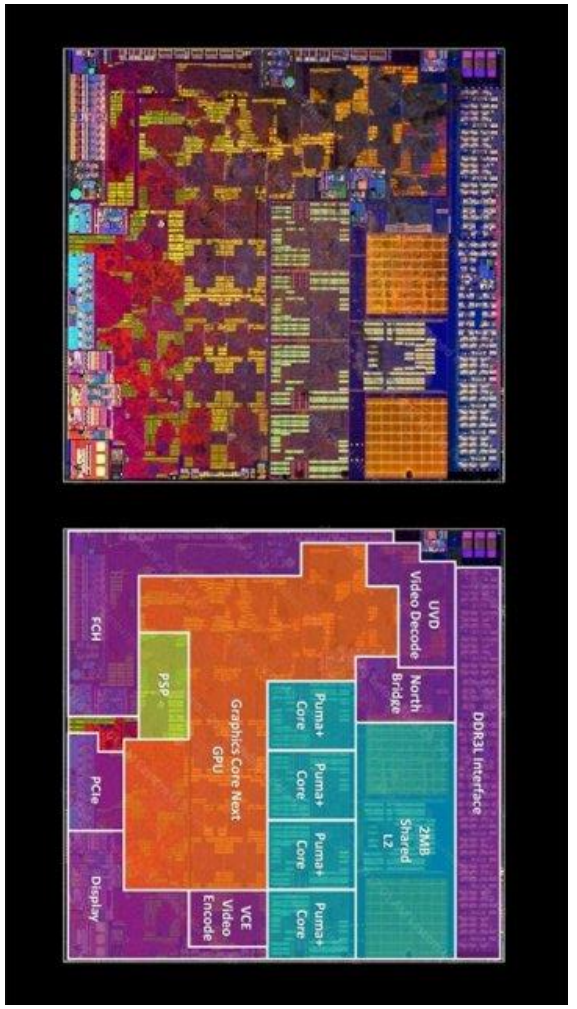
When in doubt, consult your vendor: Snapdragon 210 as found in Microsoft Lumia 550 (rendered)

About video de/encode

- Why does it matter to have a recent processor?
  - Recent processors incorporate recent FFU for de/encoding
  - Media coding formats evolve
  - The „new” standard that is spreading is H.265 (aka. HEVC)
  - It provides same quality as H.264 but in half the bandwidth
  - If your processor cannot decode it, you will need double the WiFi/mobile bandwidth

Beyond bare numbers

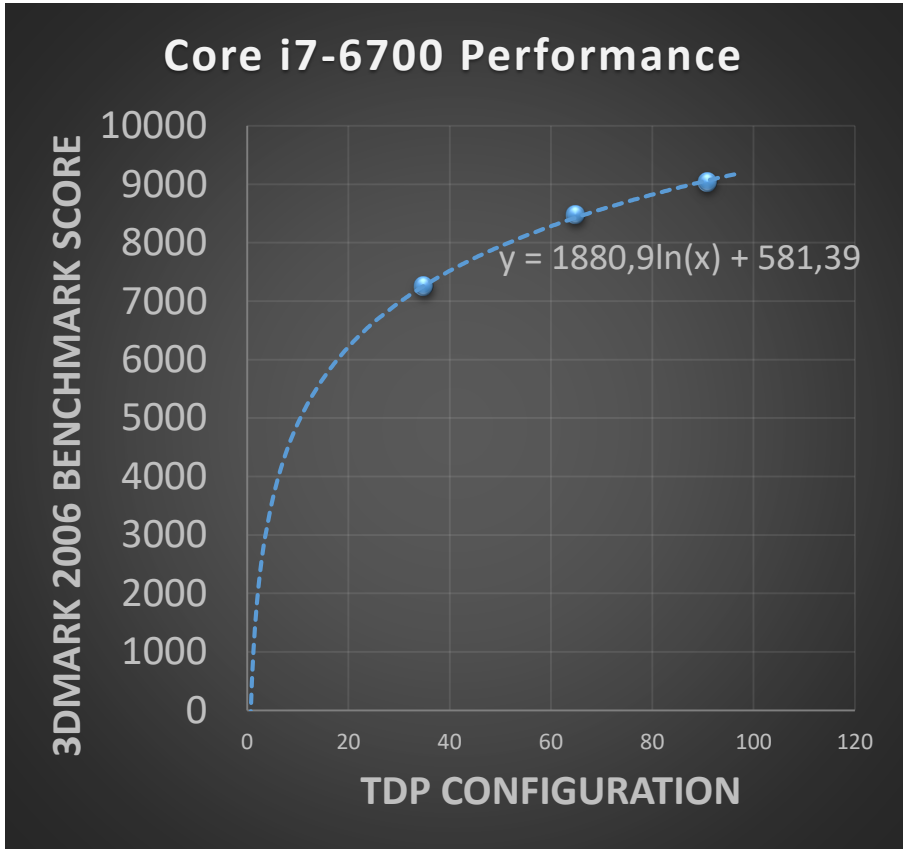
- How strong of a processor do I need if I want all my data encrypted on disk and not sacrifice performance?
- It doesn't matter, so long as you are using professional grade equipment.
  - High-end processors feature fixed-function units for en/decryption (Intel vPRO, AMD PSP)
  - Some HDDs/SSDs also feature such capabilities (boot time password)



AMD Mullins microarchitecture

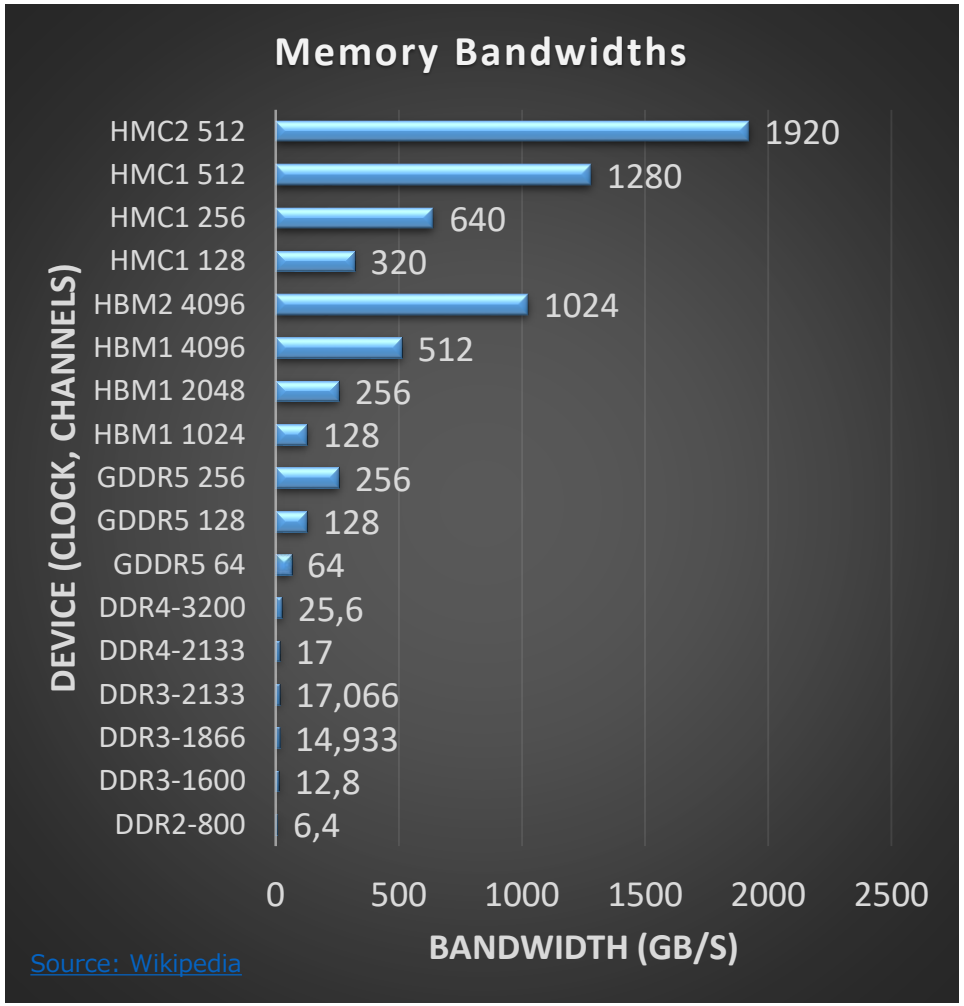
Beyond bare numbers

- How does a desktop Core i7 relate to a mobile one?



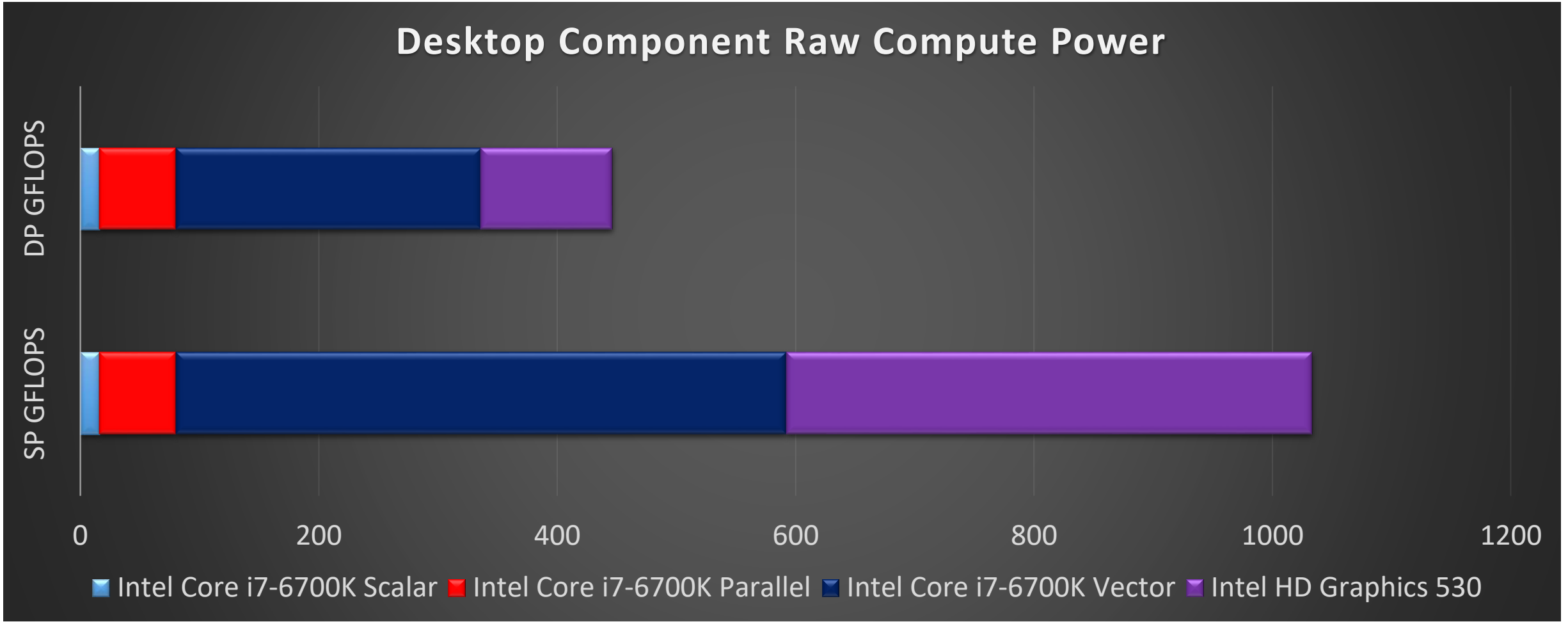
- A desktop Core i7 eats mobile ones for breakfast.
  - Modern processor speed is primarily limited by thermal restrictions (TDP). With better cooling, higher clock rates can be applied.
  - Mobile variant (Core i7-6700HQ) is a different chip. Lower TDP designs provide less cores.

Memory channels

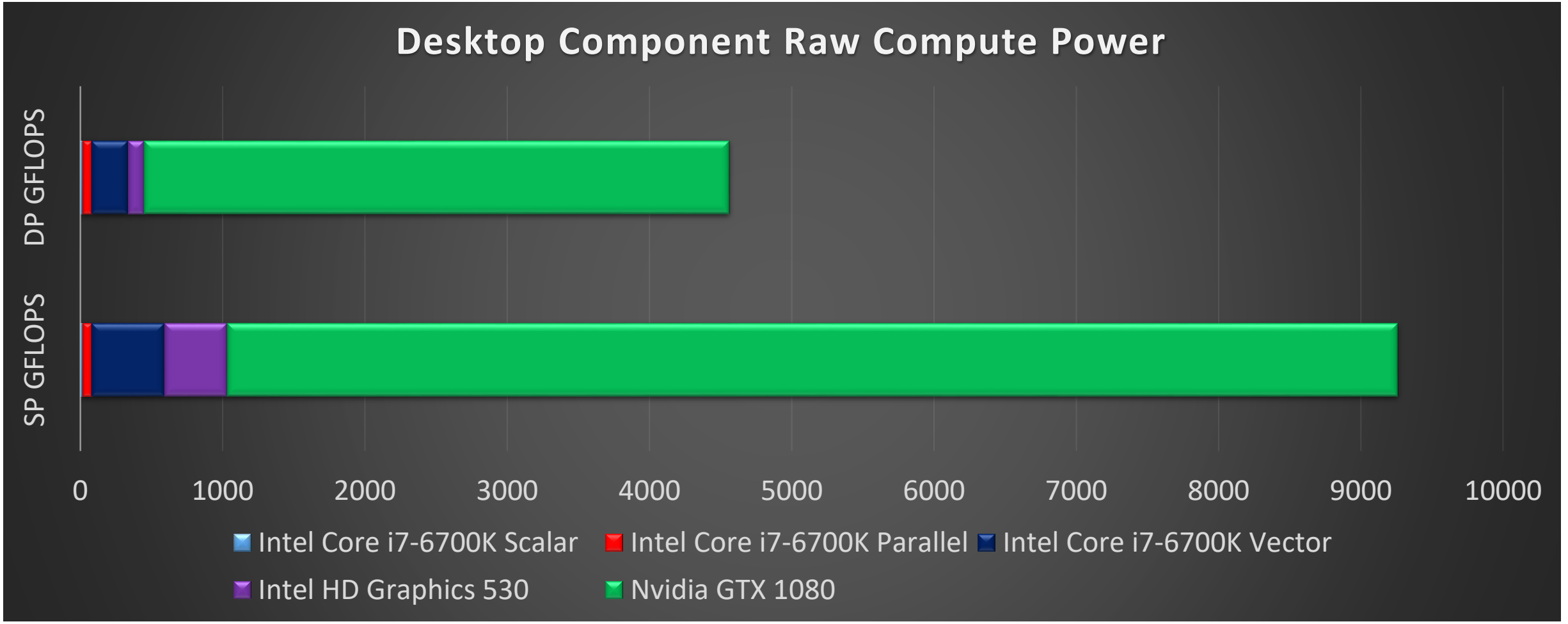


- The number of memory channels may severely impact performance
  - This is the main limiting factor of integrated graphics.
  - The IGP shares main system memory bandwidth with the CPU
- Mobile devices with multiple RAM sockets do not neccessarily provide multiple channels.
- Successor to GDDR5 have not been found
  - HBM (High-Bandwidth Memory) is mainly favored by graphics cards vendors
  - HMC (Hybrid Memory Cube) is favored by supercomputer vendors

The free lunch is over



The free lunch is over



The free lunch is over

„Without addressing vectorization, GPU computing, scalable parallelism, standard C++ is just a scripting system to get the other 99% of the machine through other languages and libraries.”

- Sean Parent

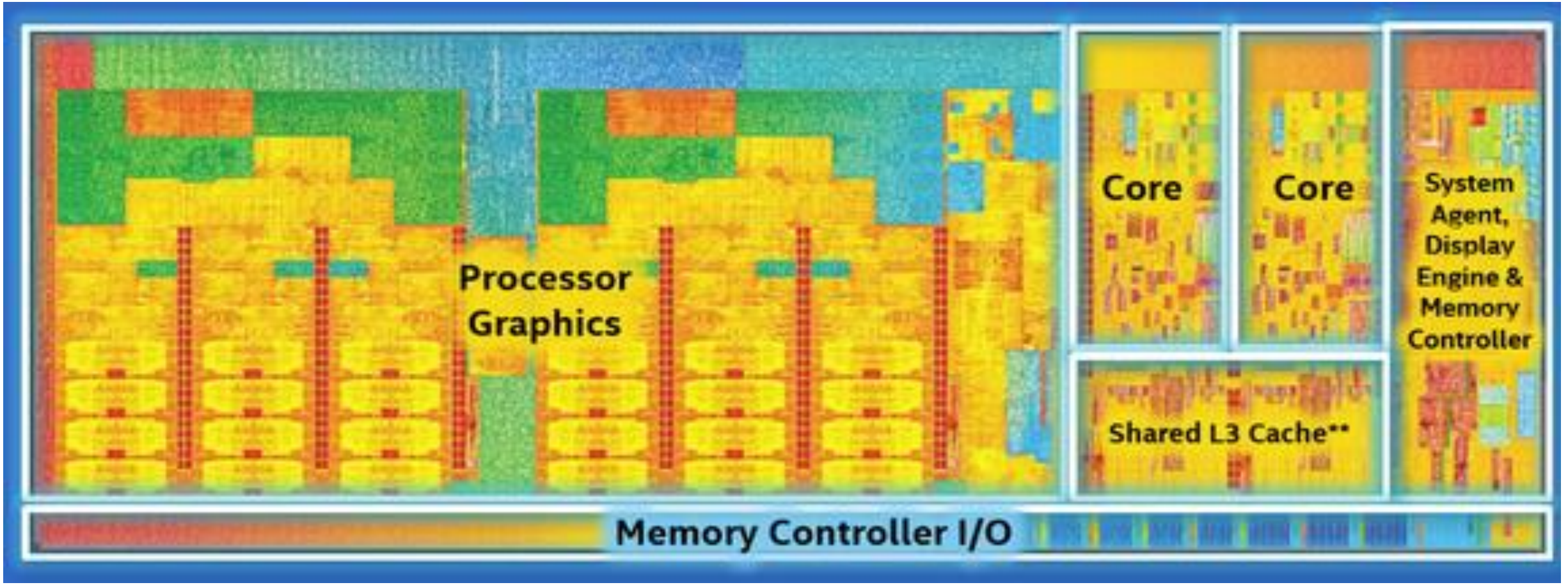
The free lunch is over

- There is ongoing strong effort in getting GPU parallelism to work in **standard** C++.
- All APIs including OpenMP, OpenACC, OpenCL, CUDA, C++AMP and SYCL are just pioneers of this work. Standard solution will learn from all of them and boil it all down to a single, self-consistent entity.
- Best chances are that these efforts will manifest as a TS by C++20
- If someone wishes to catch up on heterogenous computing, a few recent talks worth mentioning:
  - CppCon 2016: Bryce Adelstein Lelbach "The C++17 Parallel Algorithms Library and Beyond„
  - CppCon 2016: Gordon Brown & Michael Wong "Towards Heterogeneous Programming in C++"

Now we are educated consumers. But we are users of HPC and programmers!

TEACH ME MORE SENSEI!

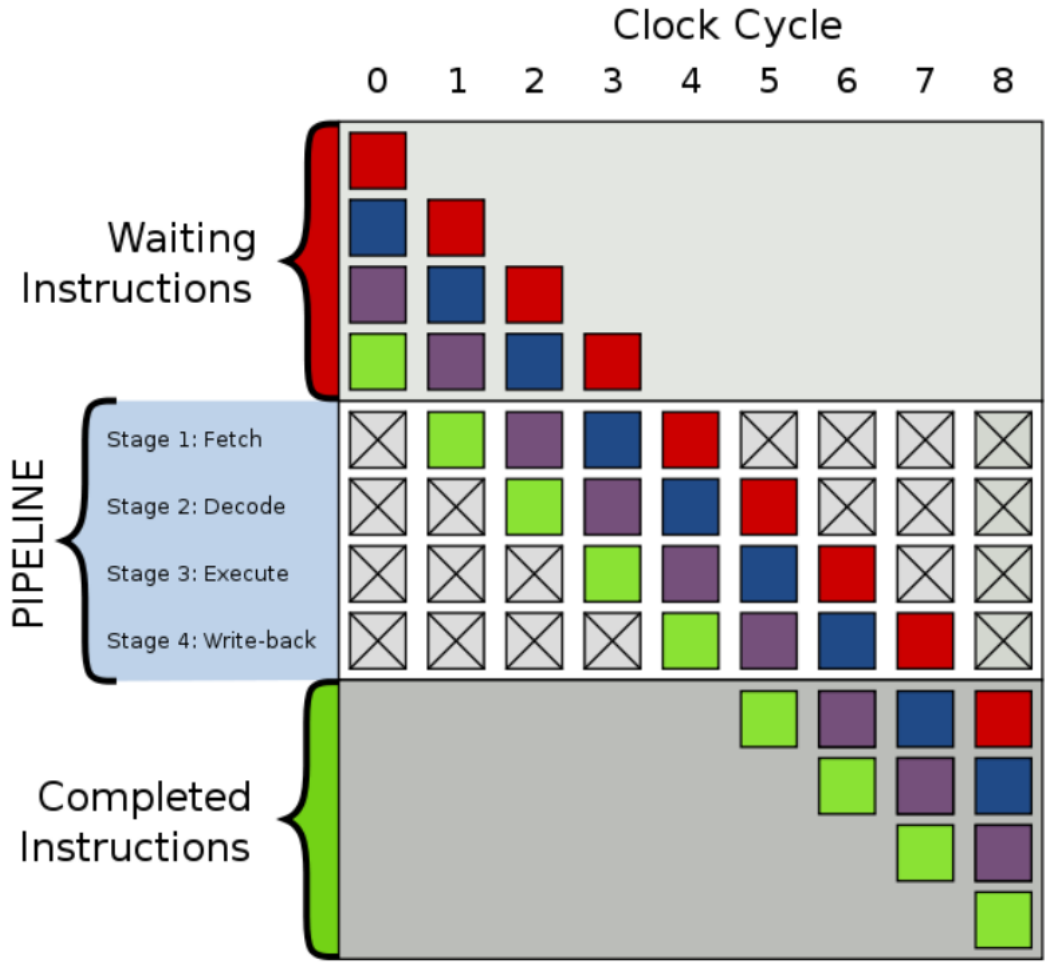
Heterogeneous computing



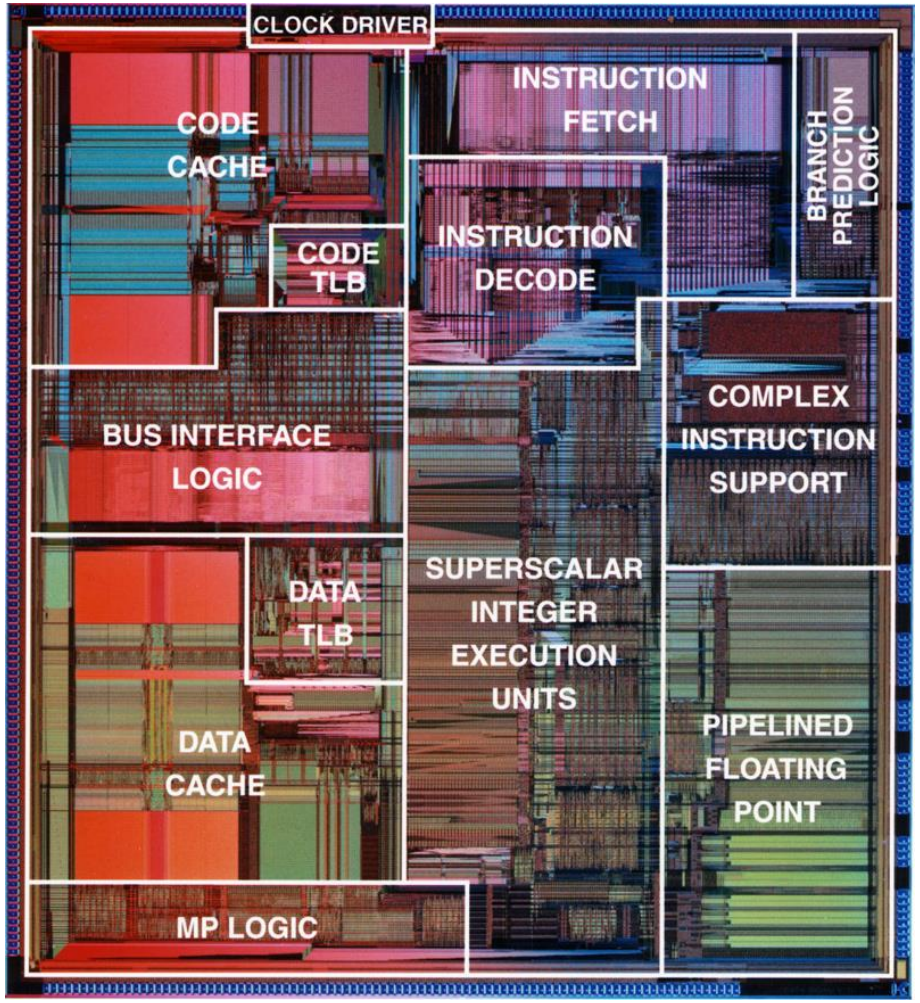
Intel Haswell Y dieshot

CPU internals

- Pipeline
  - Fetch, az utasítás beolvasása
  - Decode, az utasítást megvalósító/emuláló áramkörök kiválasztása
  - Execute, az utasítás végrehajtása
  - Write-back, az eredményt regiszterből memóriába írás
- Latency
  - Mennyi idő, amíg egy adott utasítás átér a szalagon
- Throughput
  - Adott idő alatt hány utasítás megy át (IPC: Instructions Per Cycle)



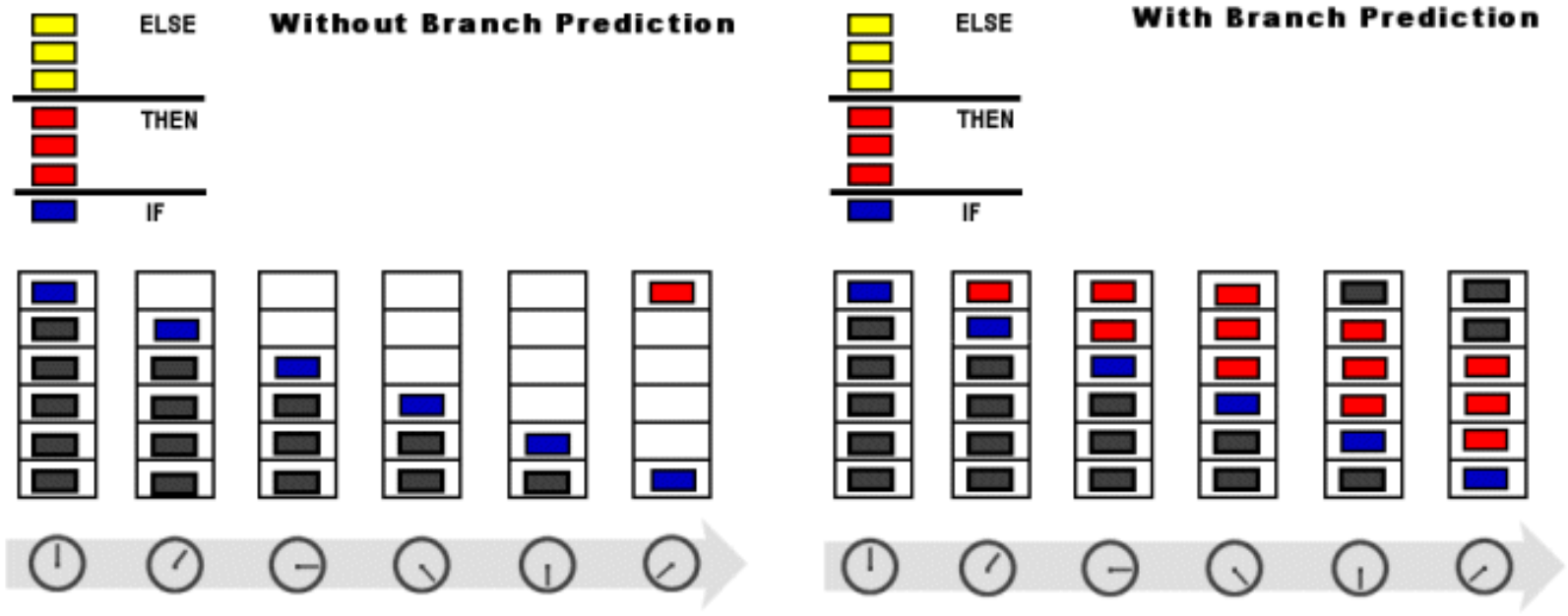
CPU internals



- Code/data cache
- Code/data translation look-aside buffer
  - Helps determining if data of a given address is actually in the cache
- Instruction fetch/decode
- Branch prediction logic (see later)
- ALU units
  - Integer (superscalar)
  - Floating (no vector instructions yet)
  - Complex (transcendental et al.)

CPU internals

Branch prediction is the quintessence of a CPU



- Branch prediction
  - Decides what instruction and data should the fetch unit read
  - If it mispredicts, the pipeline can run dry
- Speculative execution
  - In cases, branch prediction may decide to execute both arms of a branch, so the pipeline does not stall

CPU internals

- Out-of-order execution:
  Causes instructions to execute in a different order, than they arrived in, so long as this does not alter their result (dependency trees)
- Hyper-Threading:
  Advertizes a single physical core as two logical cores, by which multiple tasks use the same pipeline at the same time. By doing so, the superscalar units have a better chance at being utilized
- Translation look aside buffer:
  Due to increasing use of virtualization, memory addresses are increasingly often need to be transformed. The reason of a cache miss might be that the data resides in cache, but it's address does not. TLB mitigates this problem.

CPU internals

- 2006: Intel Core – smaller consumption, macro-op fusion, SSE4
- 2007: Penryn – SSE4.1
- 2008: Nehalem – SSE4.2, integrated PCI-E, DMI controllers, two-level branch prediction
- 2010: Westmere – AES encryption
- 2011: Sandy Bridge – AVX, L3 cache shared with the IGP, hw video decode
  2011: Ivy Bridge – tri-gate transistors,Trusted Execution Technology, hardware random number generation
- 2013: Haswell – AVX2, FMA3, Transactional Synchronization Extensions
- 2014: Broadwell – ADX (arbitrary precision integer arithm.)
- 2015: Skylake – AVX-512, SHA encryption, memory protection

Summa summarum

- Modern CPUs try to execute our program code on a fairly complex pipeline
- They try to predict branching, data and code fetching as well as how is it optimal to execute a series of assembly instructions
- It is no surprise if all if this comes at a cost (transistors, energy consumption and die area), but this is what's needed to run operating systems and many applications in parallel
- All if this in a backward compatible manner a few decades back in time...

OK, but we're the GPU-Lab. What's this gotta do with us?

MANY-CORE ERA, AKA. MASSIVELY PARALLEL PROGRAMMING

Design choices

Latency optimized core

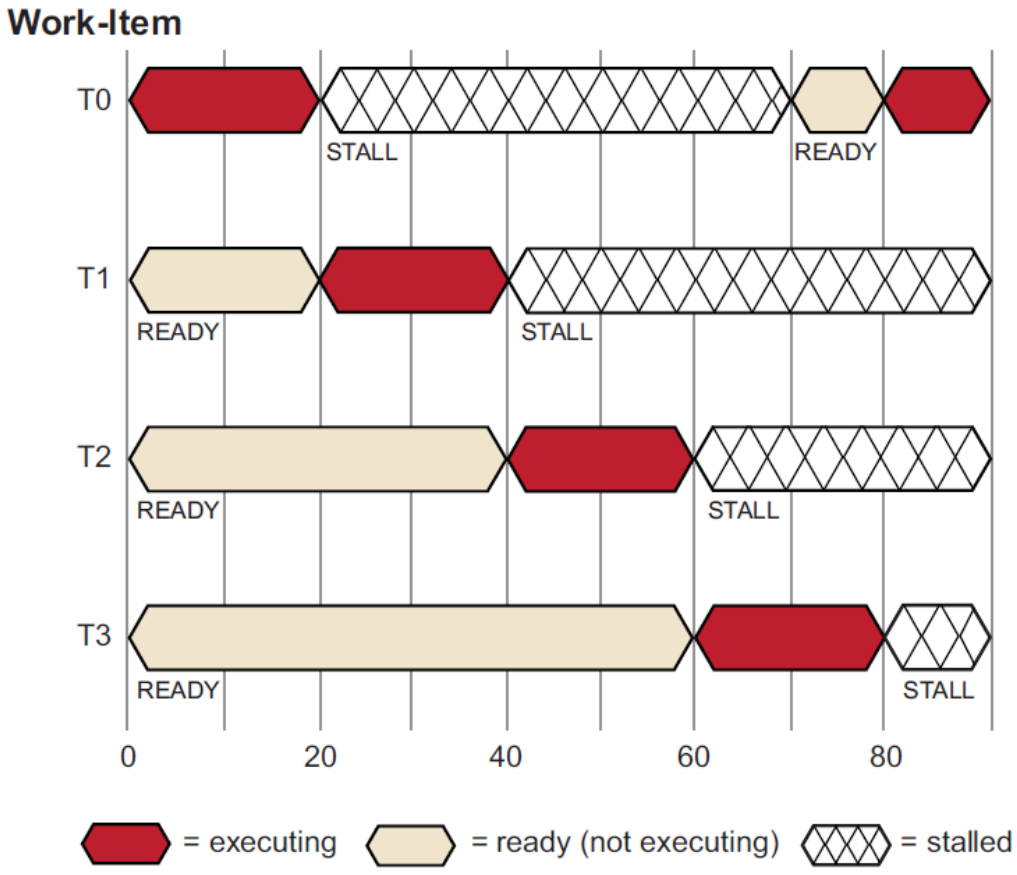
- Large cache sizes
  - Good chance that data is near and not have to be read from memory
- Complex instruction set
  - Special instruction for everything
- High clock speeds
  - HELL YEAH

Throughput optimized code

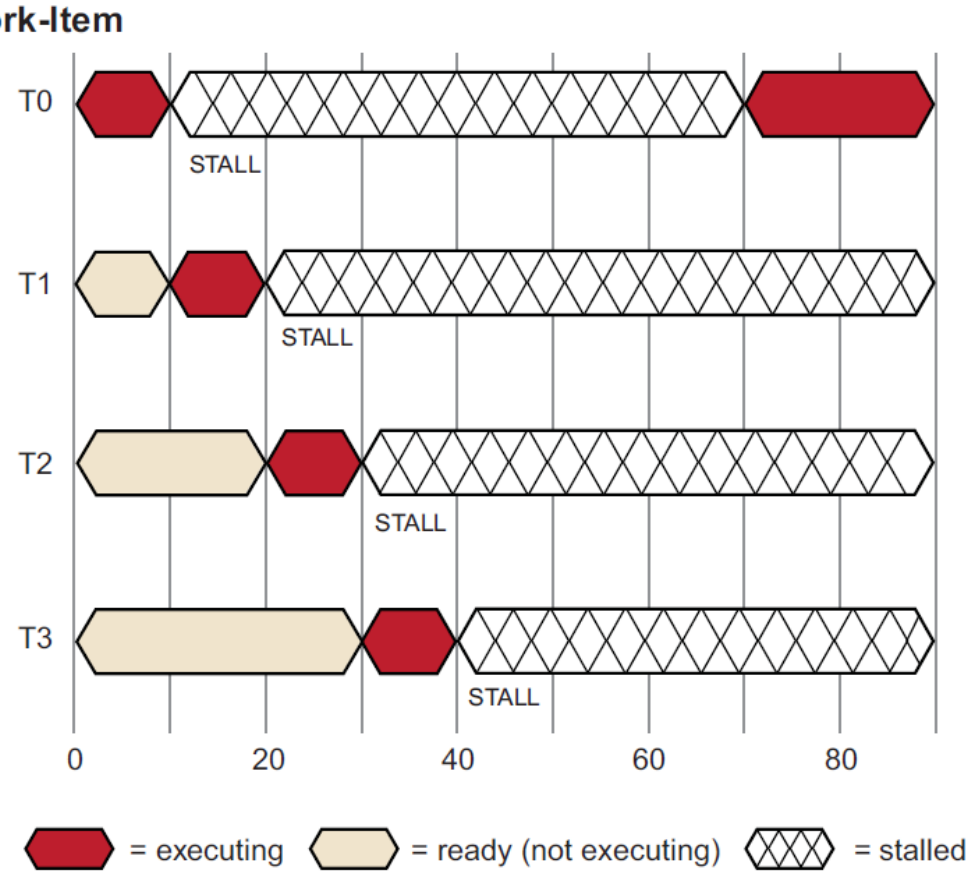
- Small cache sizes
  - If data is not ready, switch threads
- Simple instruction set
  - Less logic = less energy
  - Less energy = less heat
- Low clock speeds
  - Lower clock rate = less heat

Latency hiding

Favorable

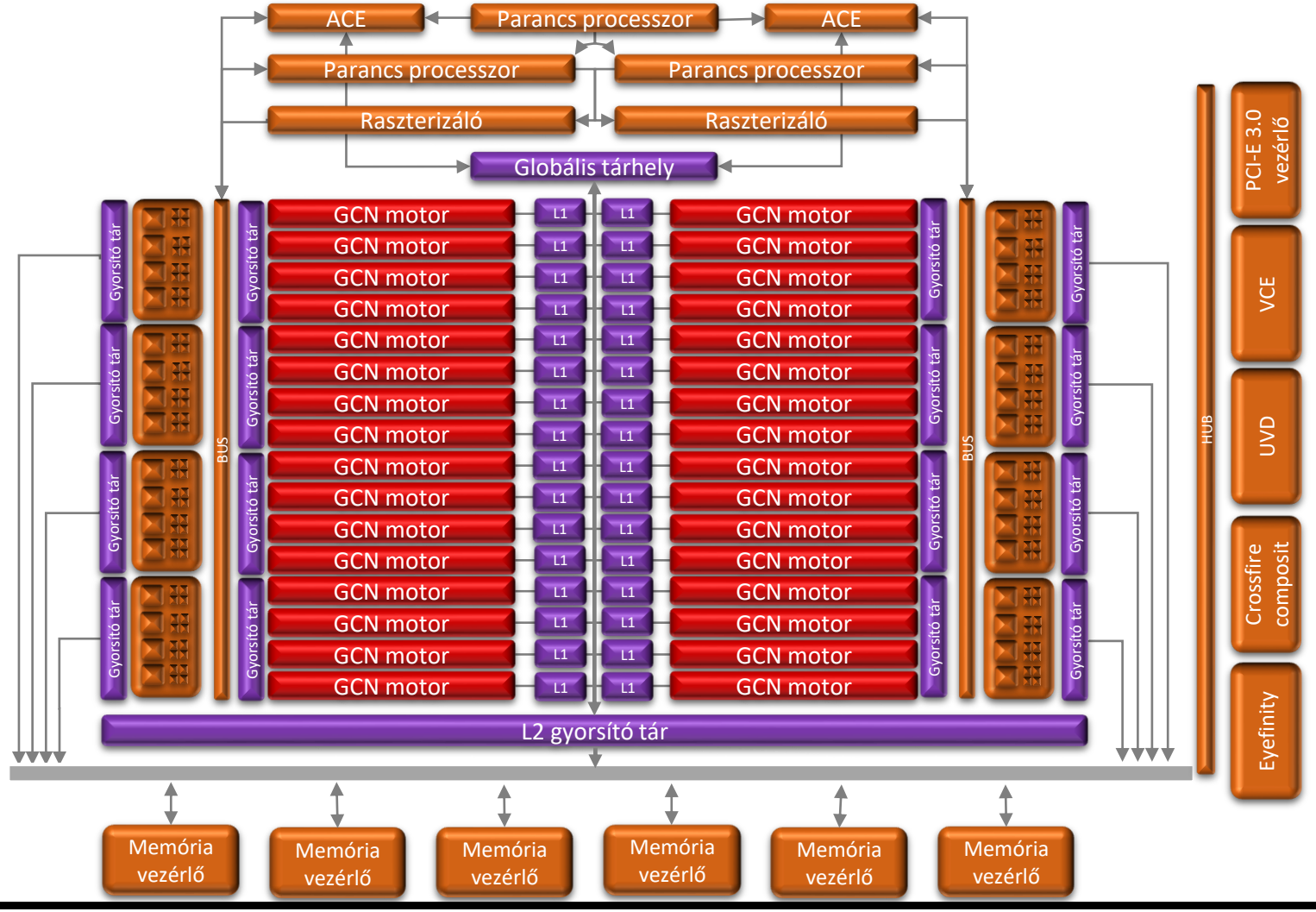


Sub-optimal



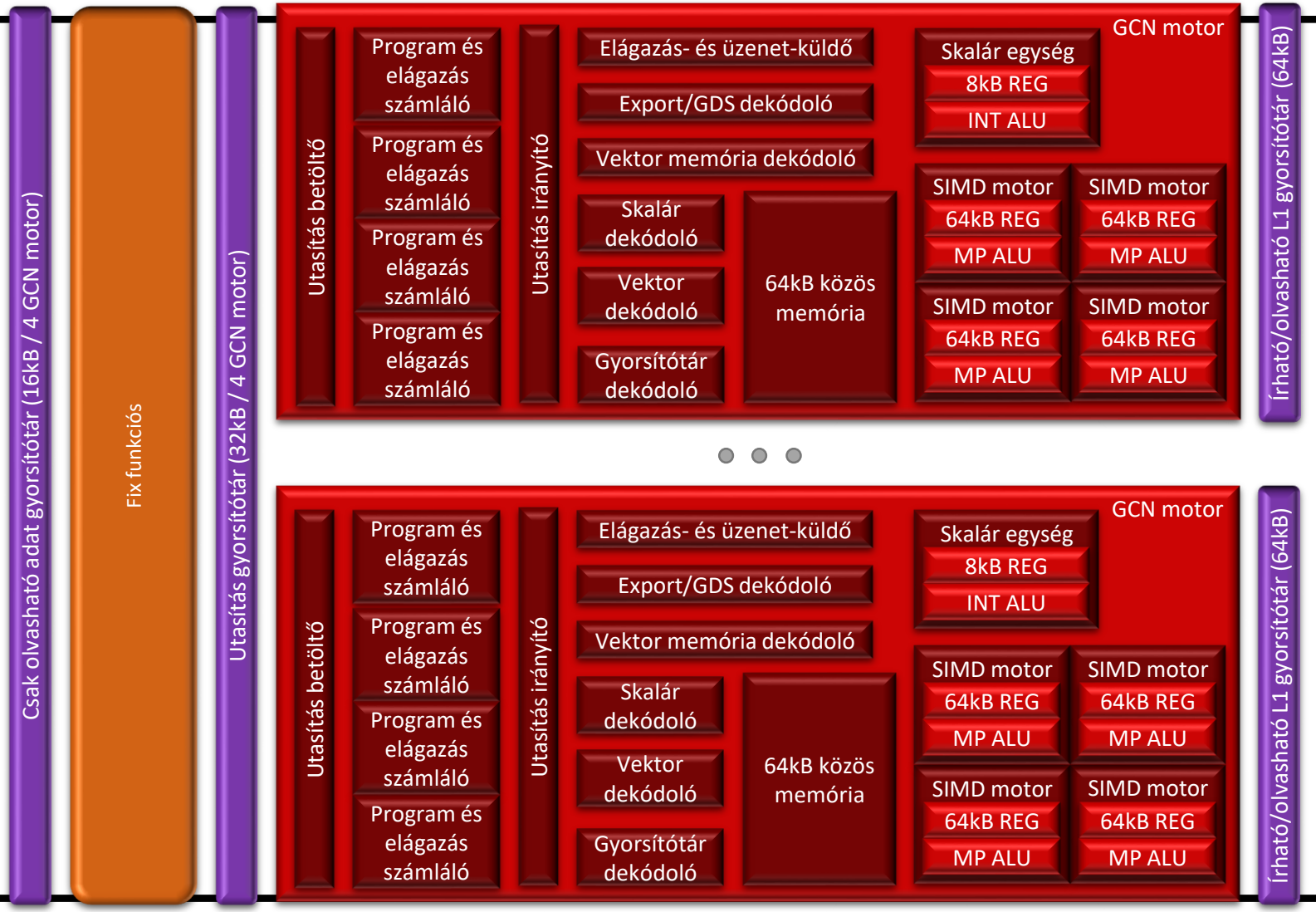
Grahpics Core Next Architecture

- Massively prallel architecture
- Strongly redundand, resistant to manufacturing flaws
  - Monolithic chip vs. Small die
- Fix function units
  - Reduce consumption
- Many small cores for many instances of the same task
- Small cache sizes
  - L3 = 0 MB
  - L2 < 1 MB
  - L1 ~ 32 kB



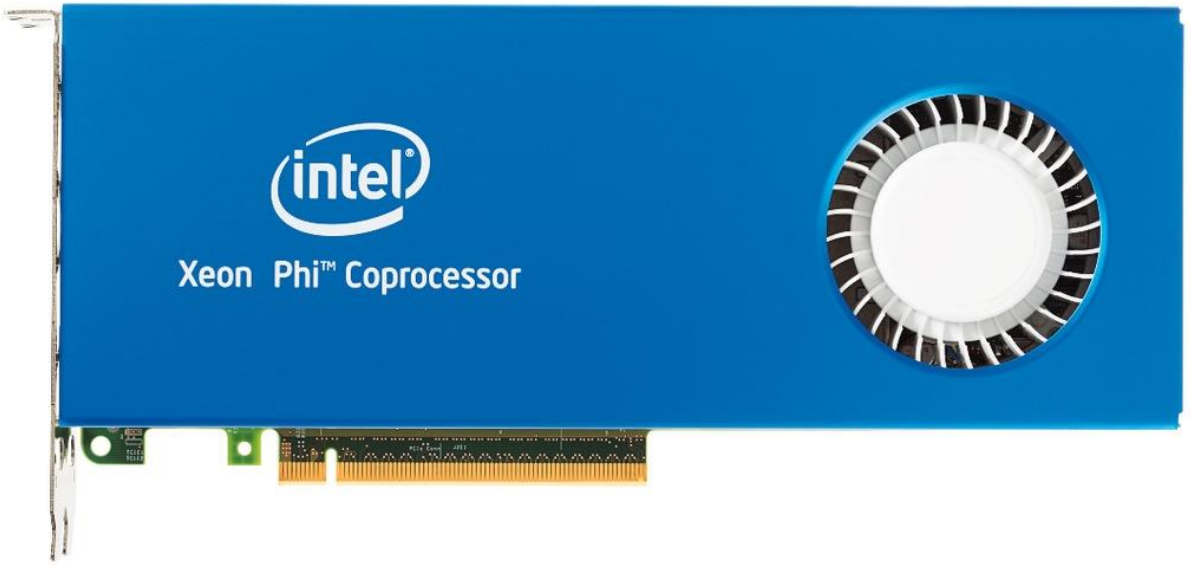
GCN Compute Unit Details

- General purpose compute core
- Traces of legacy architecture can still be found
- Register latency (!)
- Coupled execution
  - While vendor like to advertize each lane of a SIMD engine as a core (luckily, some vendors tend to stop that), it is an ultra-wide SIMD engine
- One GCN CU holds 4*16 wide SIMD engines
- Every 4 CUs share a few fixed function hardware as well as cache



Many Integrated Core

Intel Xeon Phi from the outside



- Not a CPU, not a GPU either
- Low cost of entry to massively parallel programming
- Today it's a separate card, tomorrow it will be a regular socket server CPU

Intel Xeon Phi from the inside