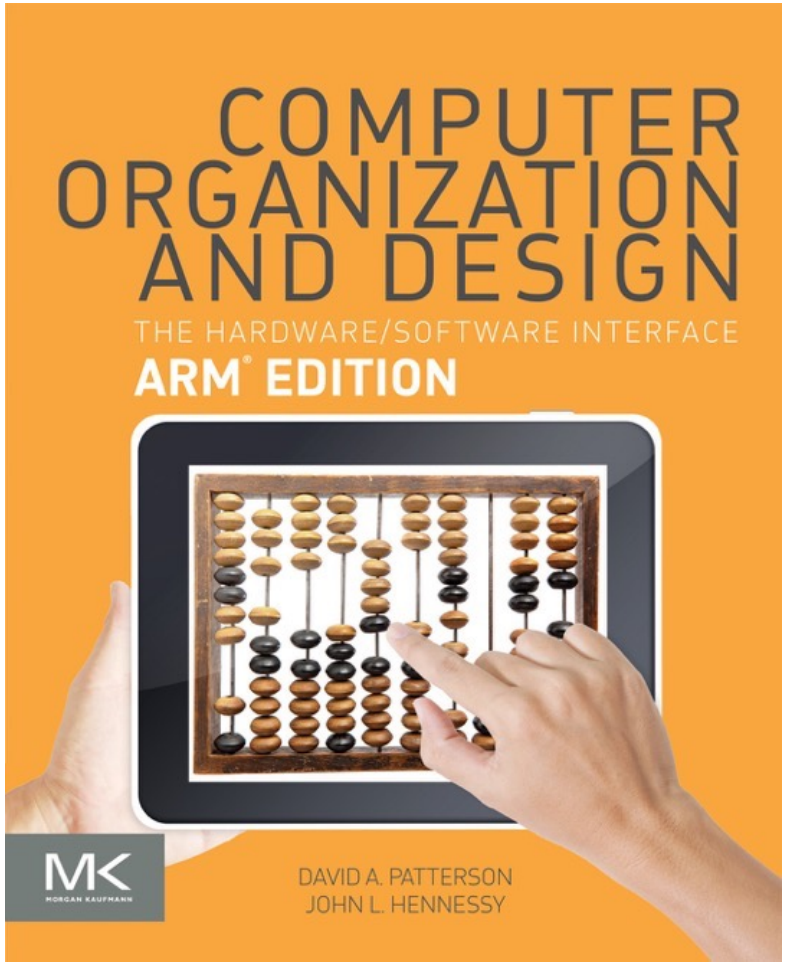
EEEN301 Embedded systems

Lecture 7 2023

Need for Speed!

Reference book

- David A. Patterson and John L. Hennessy, Computer Organization and Design, ARM edition, Morgan Kaufmann Publishers.



Is Moore's law over?

17 Sep 2018 | 20:09 GMT

David Patterson Says It's Time for New Computer Architectures and Software Languages

Moore's Law is over, ushering in a golden age for computer architecture, says RISC pioneer

By **Tekla S. Perry (/author/perry-tekla-s)**

Photo: Peg Skorpinski/UC Berkeley

David Patterson

David Patterson—University of California professor, Google engineer, and RISC pioneer (http://news.berkeley.edu/2018/03/21/david-patterson-pioneer-of-modern-computer-architecture-receives-turing-award/)—says there's no better time than now to be a computer architect.

That's because Moore's Law really is over, he says: "We are now a factor of 15 behind where we should be if Moore's Law were still operative. We are in the post–Moore's Law era."

This means, Patterson told engineers attending the 2018 @Scale Conference (https://atscaleconference.com/events/the-2018-scale-conference/) held in San Jose, that "we're at the end of the performance scaling that we are used to. When performance doubled every 18 months, people would throw out their desktop computers that were working fine because a friend's new computer was so much faster." But last year, he said, "single program performance only grew 3 percent, so it's doubling every 20 years. If you are just sitting there waiting for chips to get faster, you are going to have to wait a long time."

For a computer architect like Patterson, this is actually good news. It's also good news for innovative software engineers, he pointed out. "Revolutionary new hardware architectures and new software languages, tailored to dealing with specific kinds of computing problems, are just waiting to be developed," he said. "There are Turing Awards waiting to be picked up if people would just work on these things."

As an example on the software side, Patterson indicated that rewriting Python into C gets you a 50x speedup in performance. Add in various optimization techniques and the speedup increases dramatically. It wouldn't be too much of a stretch, he indicated, "to make an improvement of a factor of 1,000 in Python."
On the hardware front, Patterson thinks domain-specific architectures just run better, saying, "It's not magic— there are just things we can do." For example, applications don't all require that computing be done at the same level of accuracy. For some, he said, you could use lower-precision floating-point arithmetic instead of the commonly used IEEE 754 (https://en.wikipedia.org/wiki/IEEE_754) standard.

The biggest area of opportunity right now for applying such new architectures and languages is machine learning, Patterson said. “If you are a hardware person,” he said, “you want friends who desperately need more computers.” And machine learning is “ravenous for computing, which we just love.”

Today, he said, there’s a vigorous debate surrounding which type of computer architecture is best for machine learning, with many companies placing their bets. Google has its Tensor Processing Unit (TPU) (https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu), with one core per chip and software-controlled memory instead of caches; Nvidia (https://www.nvidia.com/en-us/)’s GPU has 80-plus cores; and Microsoft is taking an FPGA approach. And Intel (http://www.intel.com/), he said, “is trying to make all the bets,” marketing traditional CPUs for machine learning, purchasing Altera (http://fortune.com/2015/08/27/why-intel-altera/) (the company that provides FPGAs to Microsoft), and buying Nervana (https://www.engadget.com/2017/10/17/intel-ai-deep- learning-nervana-npp/), with its specialized neural-network processor (similar in approach to Google’s TPU).

Along with these major companies offering different architectures for machine learning, Patterson says there are at least 45 hardware startups tackling the problem. Ultimately, he said, the market will decide.

“This,” he says, “is a golden age for computer architecture.”

Software is hitting a wall when it comes to performance and energy efficiency.

So we use hardware acceleration.

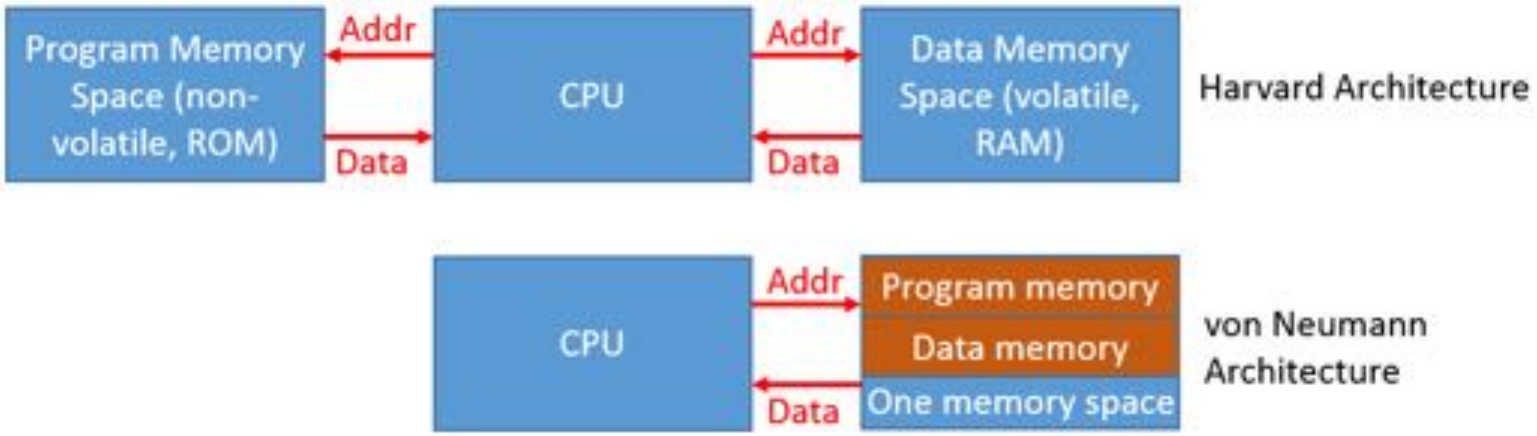
Graphics cards for gaming!

Consider the Iphone processor https://en.wikipedia.org/wiki/Apple_A11

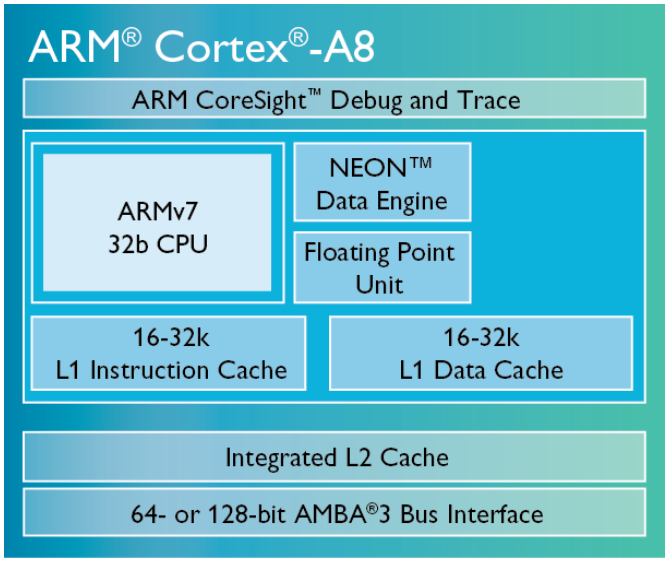
The A11 features an Apple-designed 64-bit ARMv8-A six-core CPU. The A11 also integrates an Apple-designed three-core graphics processing unit (GPU), the M11 motion coprocessor, a new image processor and a dedicated neural network hardware engine.

So much more than just a multicore processor!!

Architecture is key to performance!



Harvard architecture with it's separate memory buses allows for the simultaneous fetching of data and instructions.



Having separate Instruction and Data caches also allows simultaneous fetching.
A von Neumann architecture with the speed advantage of the Harvard architecture!

The caches themselves provide an additional speed advantage.

Plus, we have dedicated hardware accelerator blocks – NEON, FPU

Digital Signal Processors (DSP)

A DSP is used to perform real time processing. Usually the digital filtering of discrete time data coming from an ADC. An FIR filter performs a convolution between the input data stream and a set of filter coefficients. This is performed at the input data rate which is usually several times faster than the Nyquist rate. The DSP architecture is optimized for this task.

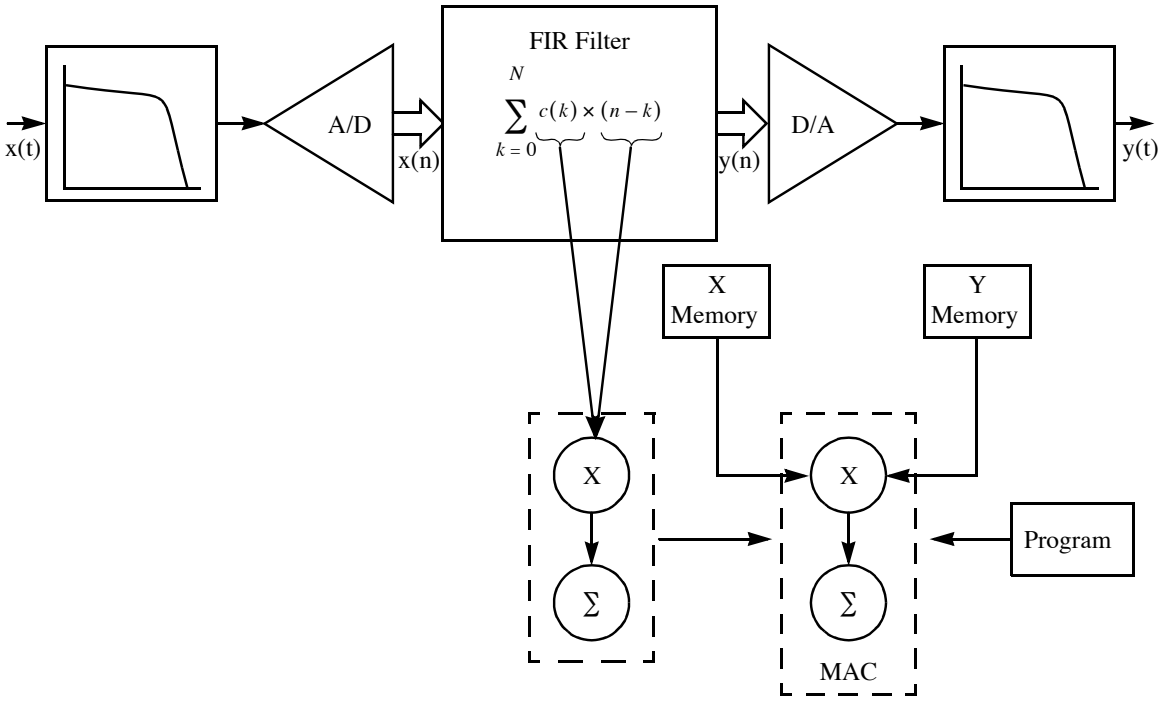


Figure 1-4. Mapping DSP Algorithms Into Hardware

Most DSP devices have separate Program and X, Y data memories and data buses. Often called a ”Dual Harvard Architecture”. This allows the simultaneous fetching of an Instruction and X (data), Y (filter coefficient) data.

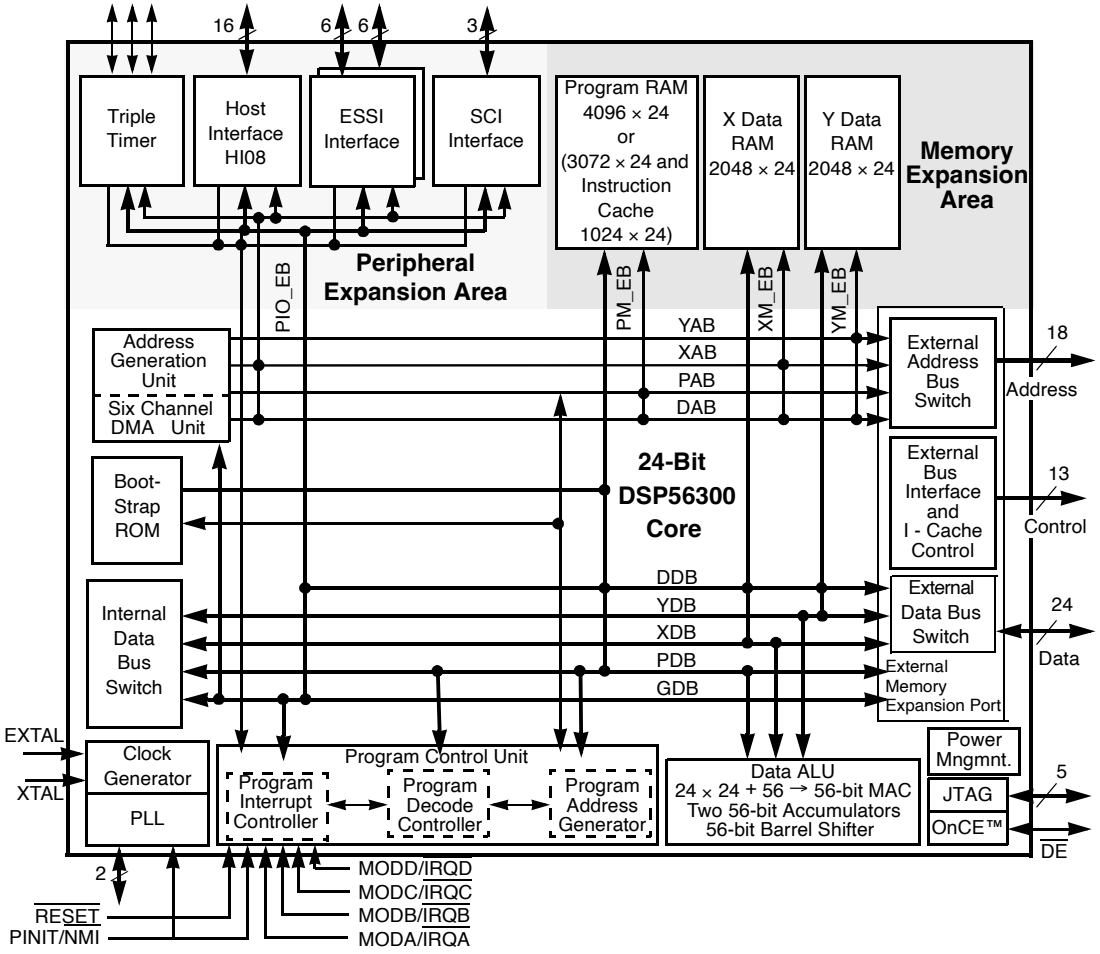
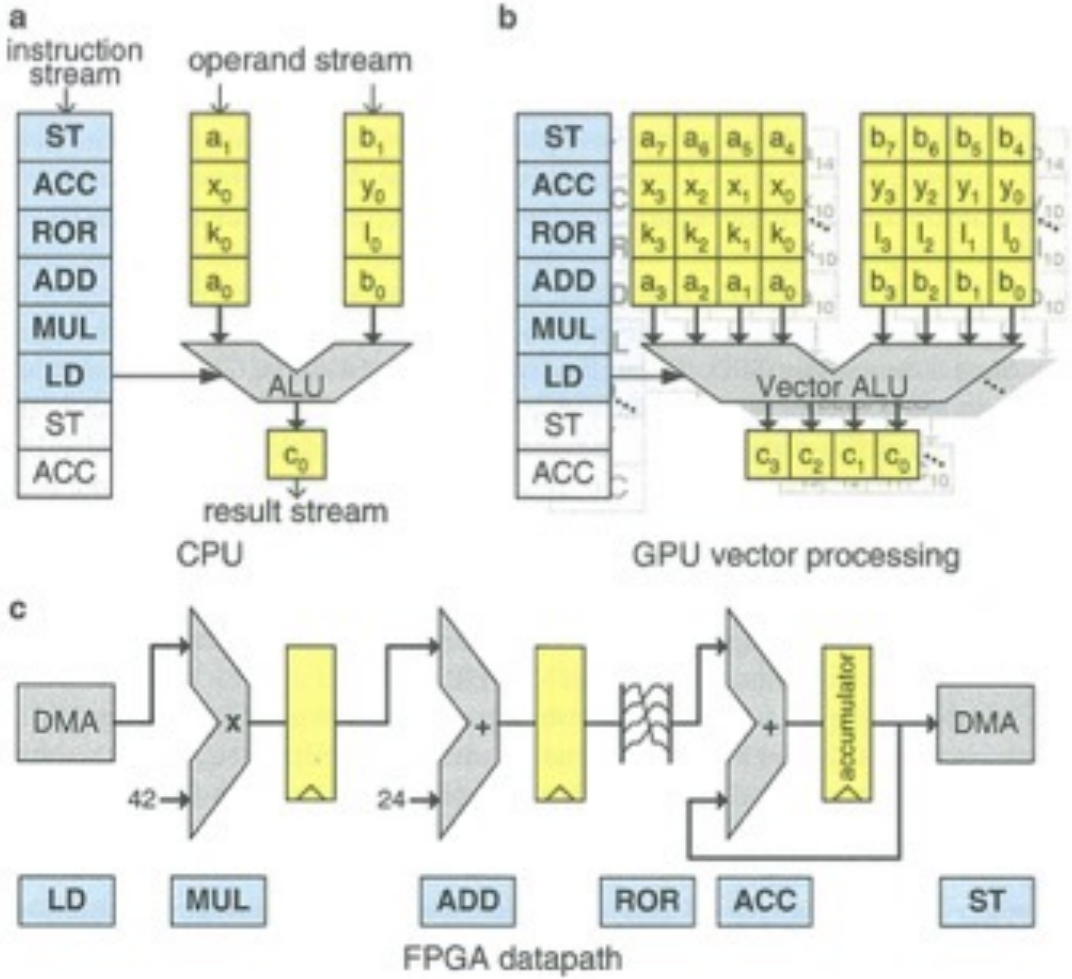


Figure 2-1. DSP56303 Block Diagram

Dedicated looping registers as well as special “DSP” instructions ensure maximum performance.

GPU and FPGA

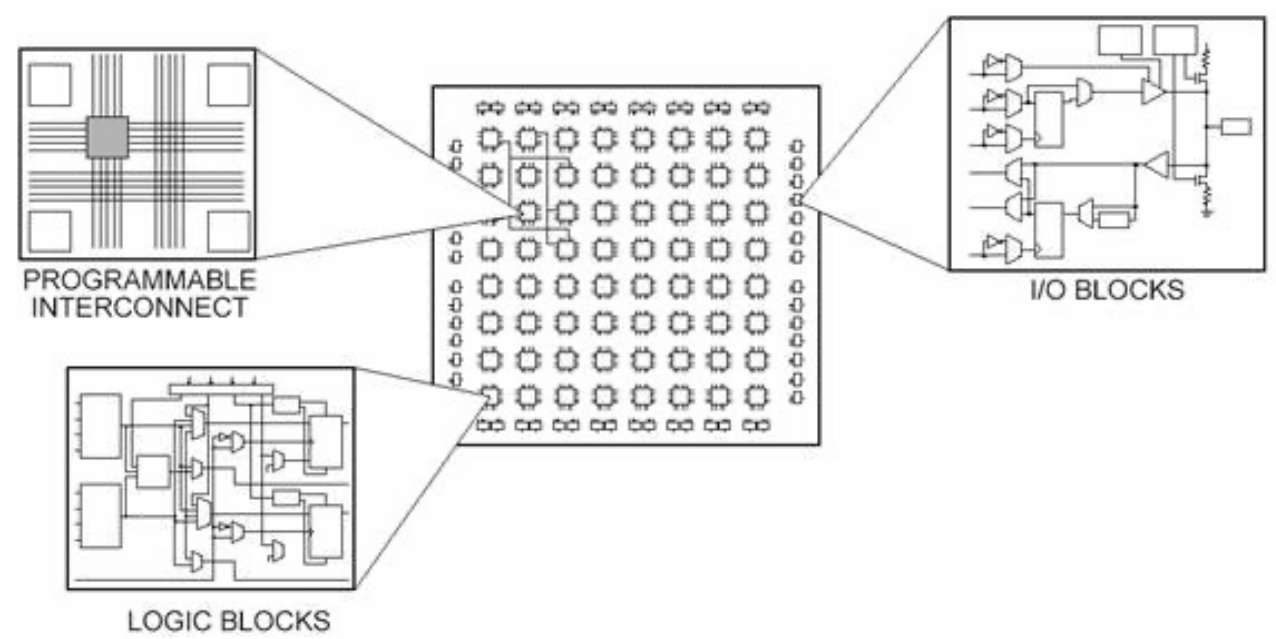
Good when operations can be parallelized or pipelined such as in Image/Video processing.



a) Traditional processor, one instruction at a time.
b) GPU vector processor, multiple operations in parallel using same instructions (SIMD). Often requires low level assembly language programming to get max performance.
c) Reconfigurable data path. A chain of dedicated(custom ALU) processing stages in parallel. Potentially one ALU per pixel!

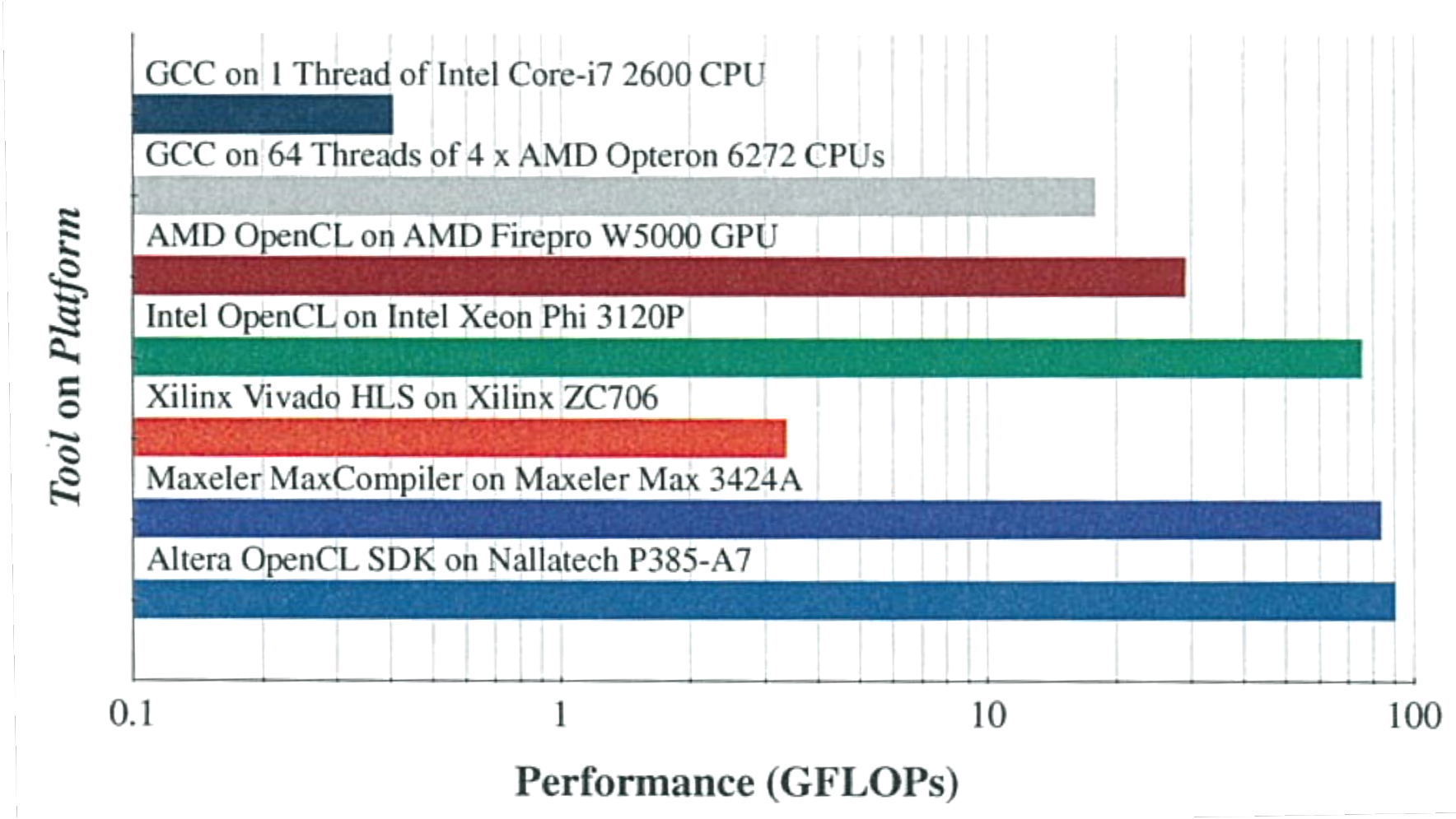
Inside an FPGA

- A regular array of configurable logic blocks (CLBs), switching matrices and I/O blocks – all programmable.



- Each CLB is made up of gates, flipflops and multiplexers
- Also other resources: DSP blocks, multipliers, RAM, Clocking etc.
- EEEN402 is a course dedicated to FPGAs

Relative performance of High Level approaches. Xeon Phi has 57 cores and like GPUs it consumes a lot of power. Maxeler provide IP cores and large FPGAs. Nallatech provide PCIe boards with large FPGA devices.



Xilinx, Alveo. High performance FPGA based hardware acceleration for your desktop PC.
33 Terra Operations Per second!!

12,288 DSP slices!

	Product Name	Alveo U250
Dimensions	Width	Dual Slot
	Form Factor, Passive Form Factor, Active	Full Height, ¾ Length Full Height, Full Length
Logic Resources [1]	Look-Up Tables	1,728K
	Registers	3,456K
	DSP Slices	12,288
DRAM Memory	DDR Format	4x 16GB 72b DIMM DDR4
	DDR Total Capacity	64GB
	DDR Max Data Rate	2400MT/s
	DDR Total Bandwidth	77GB/s
	HBM2 Total Capacity	–
	HBM2 Total Bandwidth	–
Internal SRAM	Total Capacity	57MB
	Total Bandwidth	47TB/s
Interfaces	PCI Express®	Gen3 x16
	Network Interface	2x QSFP28
Power and Thermal	Thermal Cooling	Passive, Active
	Typical Power	110W
	Maximum Power	225W
Time Stamp	Clock Precision	–
Compute Performance	INT8 TOPs	33.3
	Machine Learning	Machine Learning Solution Brief
	Acceleration Applications	Acceleration Application Solutions

Consider Amazon Web Services (AWS).

https://aws.amazon.com/ec2/instance-types/f1/

Amazon EC2 F1 is a cloud based compute instance with field programmable gate arrays (FPGAs) that you can program to create custom hardware accelerations for your application.