An Introduction to FPGAs

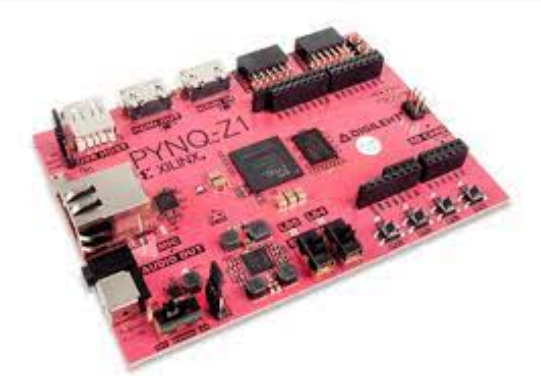
Lecture 2 for Information Processing

Aaron Zhao, Imperial College London, a.zhao@imperial.ac.uk

Why an FPGA is an interesting device to consider

My name is Aaron, and my research looks at the intersections between algorithm, hardware and security in Deep Learning (DL) Systems.

(a) Virtex ultrascale



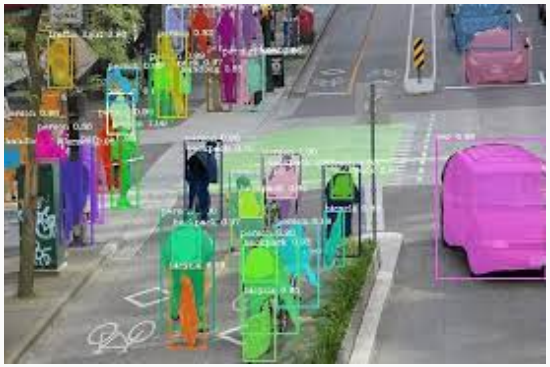
(b) PYNQ Z1

(c) Alveo U250

There is now an incresing need for high-performance systems at low power edge systems

(a) Robotics

(b) UAVs



(c) AI-based Vision systems

High throughput and low latency computation on the cloud

(a) Genetics/genomics

(b) Large-scale simulations

(c) Large AI models

The need for speed and low power

Do we have the 'Tesla"?

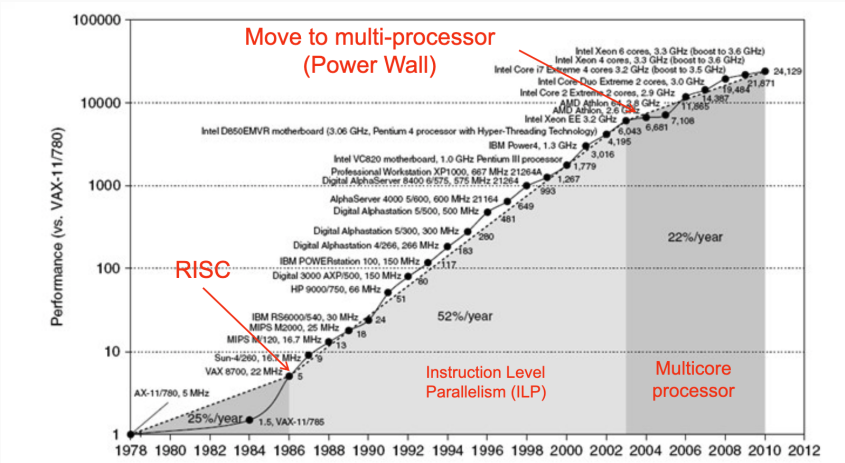
A better metric might be energy efficiency Ops/Second/Watt/Area

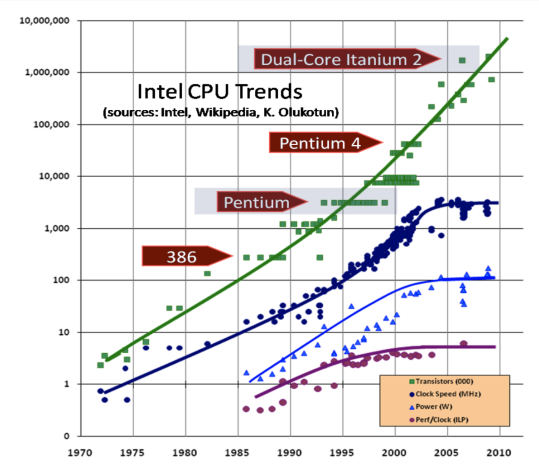
Approach 1: Use existing hardware platforms, but deisng better algorithms or software

- Design or write more “efficient” algorithms
- Use approximations, for instance, subsampling
- Consider the architecture of the system and optimize your program

Approach 2: Buy better hardware

Do nothing. Just wait for the next generation processor.



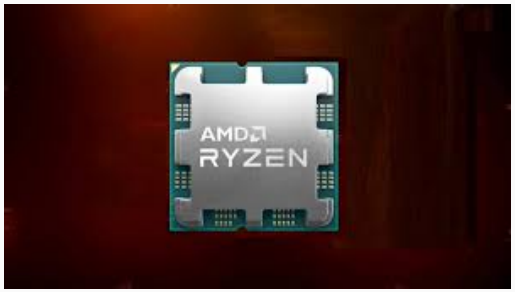


No more free-lunch

Approach 3: Use heterogeneous systems

CPUs combined with other hardware (eg. parallel architectures) on the same SoC.

- Multi- or even Many-core CPUs (AMD Ryzen Threadripper PRO 7995WX, 96 Cores, 192 Threads 2.5GHz and max 5.1GHz)
- Graphic Processing Units (Nvidia H100, 640 Tensor Cores, 128 RT Cores, 80 Streaming Multiprocessors (SMs) and 18,432 CUDA cores)
- Field Programmable Gate Arrays (around 2M Logic Elements/ LUTs, around 5K DSPs)



(a) AMD Ryzen Zen4

(b) Nvidia H100

(c) Intel Stratix 10

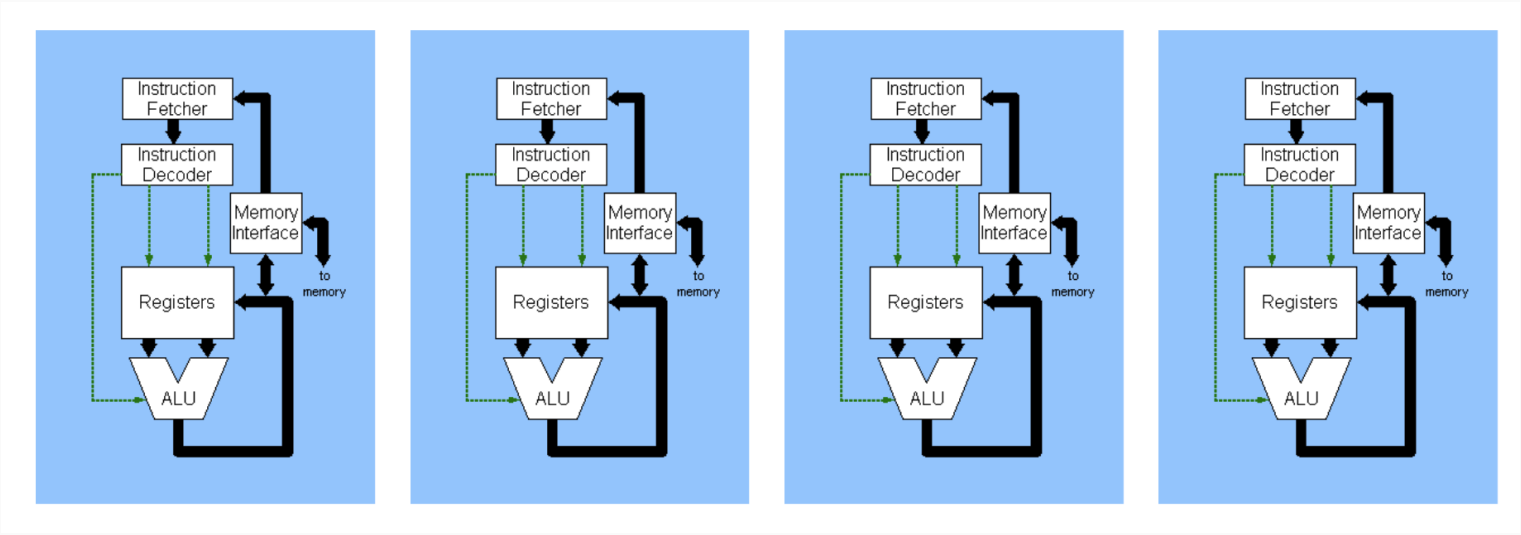
Flexibity and Performance trade-off

- The left-hand side may have better power efficiency on the specific tasks.
- The right-hand side is more flexible and requires shorter dev cycles and efforts.

Benefits come from customising the hardware to the application, and also by tuning your application for the hardware, this is also known as software-hardware co-optimization.

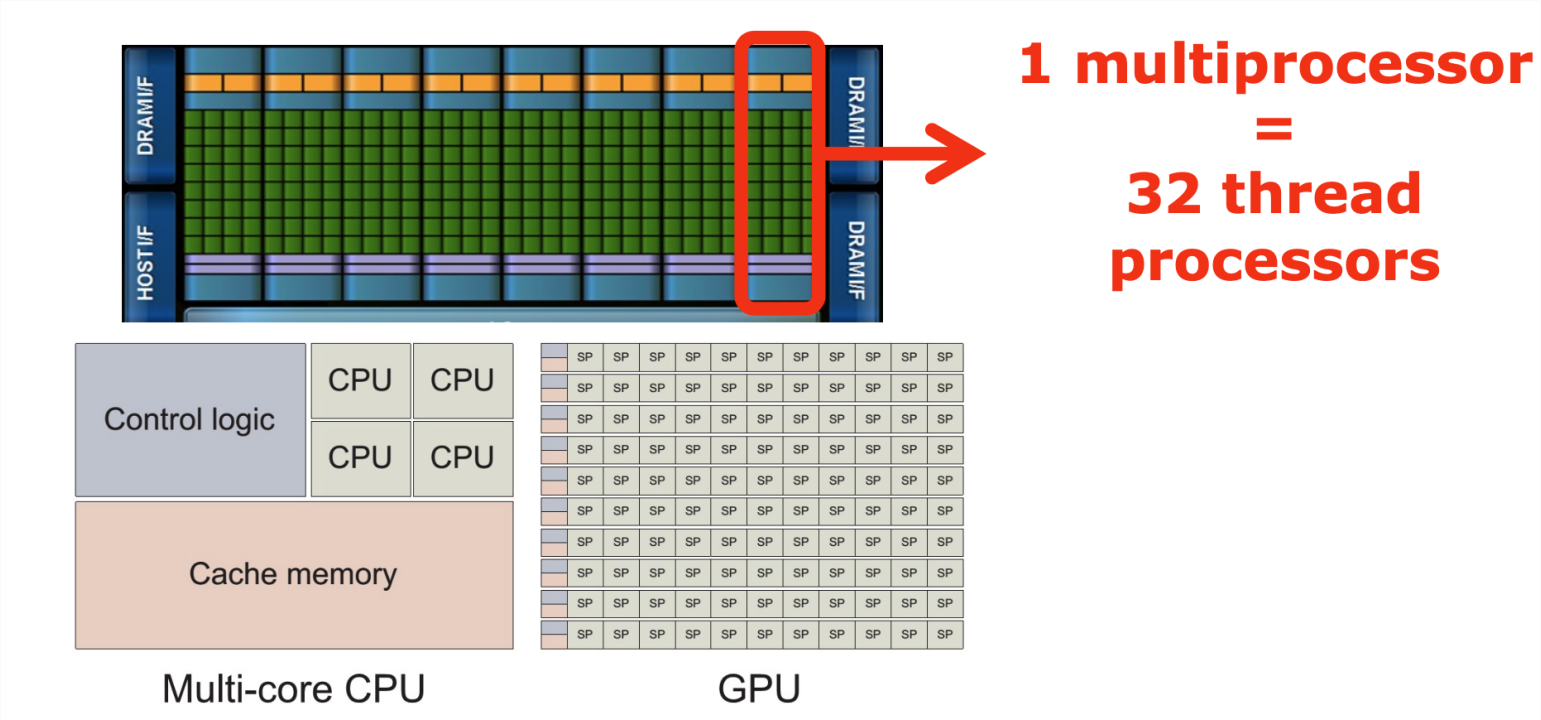
Device Comparison: Multi-core CPUs

- Each core is fairly powerful and runs at a very high frequency.
- Complex memory hierarchy.
- Up to linear speed up (extremely optimistic!).



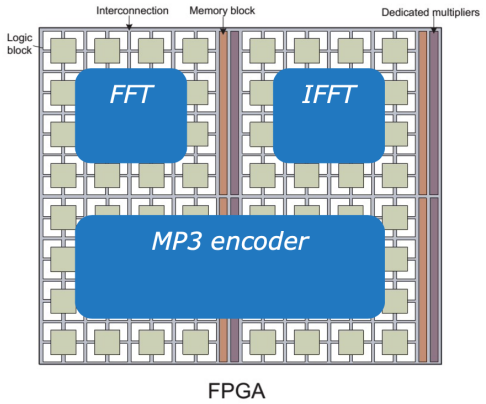
Device Comparison: GPUs

- Many light-weight thread processors (SMs, in Nvidia world) =¿ Hide memory latency.
- All thread processors execute the same sequential code.
- SIMD architecture, massive data parallelism.



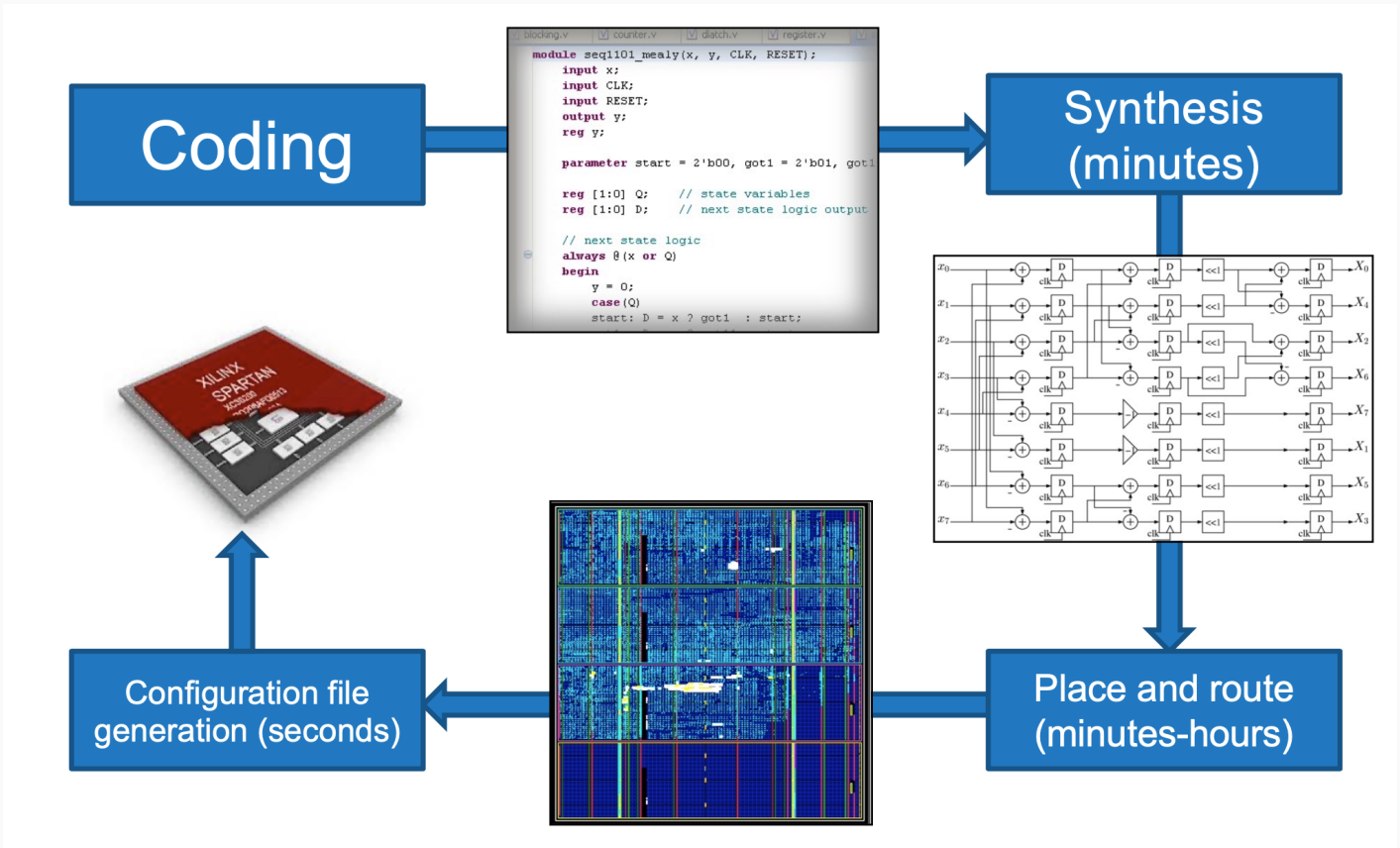
Device Comparison - FPGAs

- (Re-)programmable digital hardware – can implement any digital circuit.
- Can exploit low-level pipeline parallelism
- Logic blocks evaluate simple Boolean functions.
- Interconnection resources connect blocks to implement complex systems.



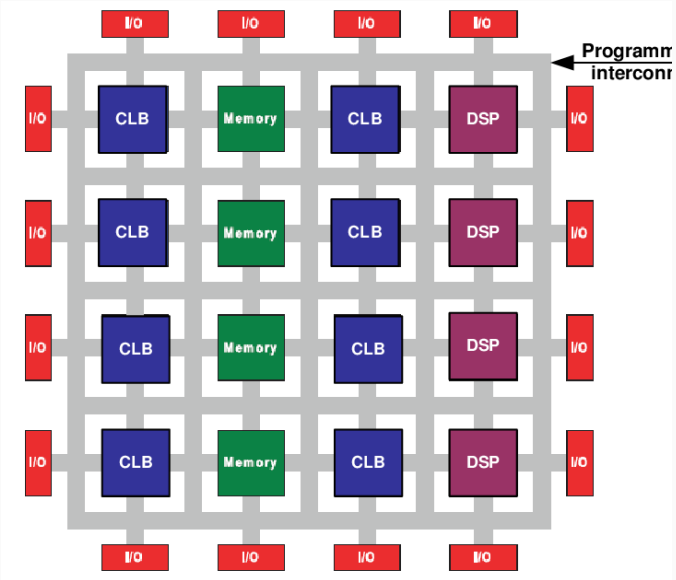
Computation in space

FPGA design flow



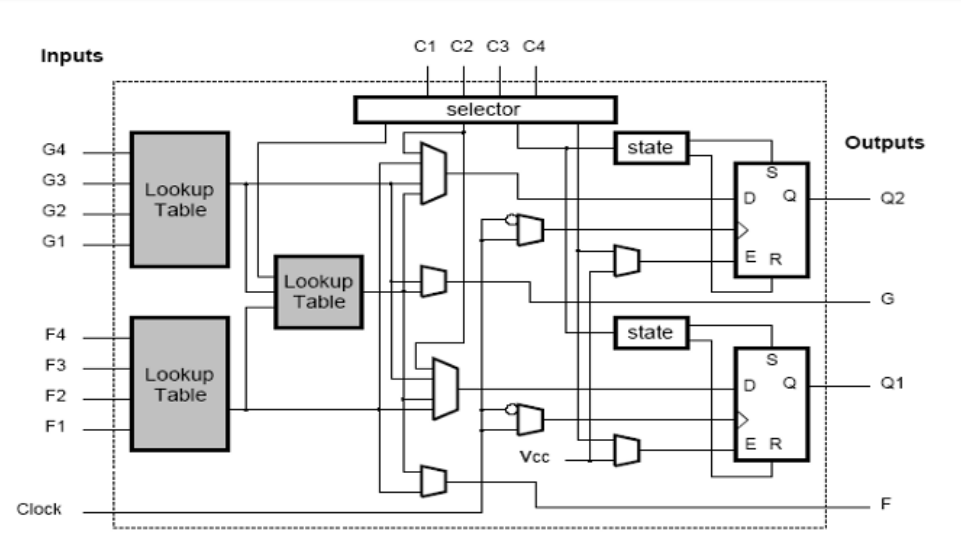
Field Programmable Gate Arrays (FPGAs)

- Xilinx (now part of AMD) is the first to introduce SRAM based FPGA using Lookup Tables (LUTs)
- Components
  - Configurable Logic Block (CLB)
  - IO Block
  - Programmable Interconnects
  - DSPs
  - Memory



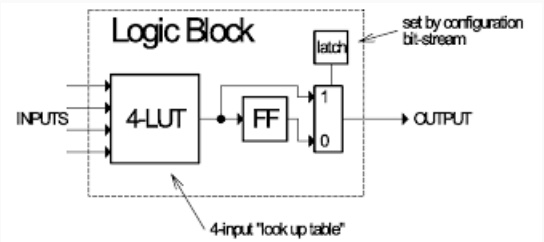
CLB

- Each Configurable Logic Block (CLB) has 2 main Look-up Tables (LUTs) and 2 registers.
- The two LUTs implement two independent logic functions F and G.
- Shown here is the CLB for Xilinx XC4000 devices.

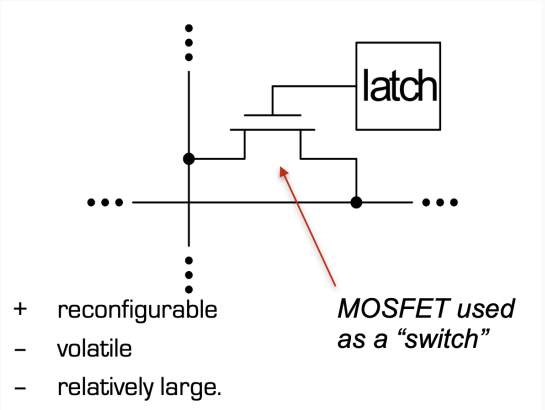


LUT

- LookUp Table (LUT) is implemented using latches:
  - 4-LUT (i.e. 4-input LUT) implements any truth table with 4 inputs, this constructs combinatorial logic functions
  - Requires 24 storage elements, each implemented with a latch (similar to a flip-flop, but half the size roughly, 1-bit memory)
  - Multiplexer select one latch to output
  - "Configuration bit stream" is loaded under user control



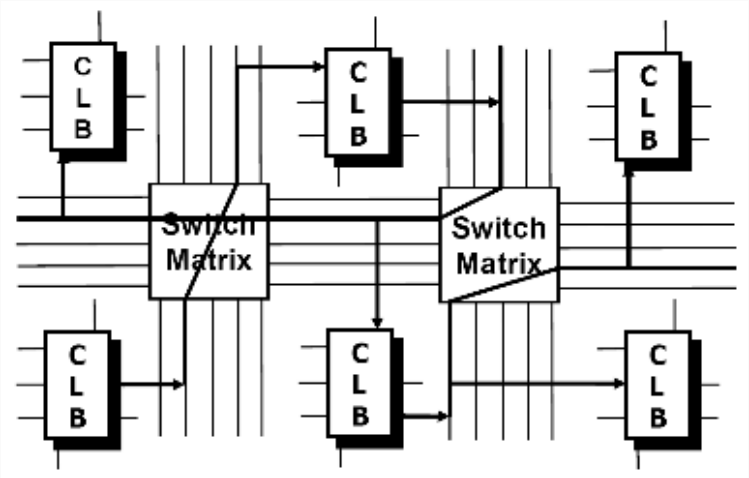
(a) 4-input LUT



(b) Latch

Programmable Interconnect

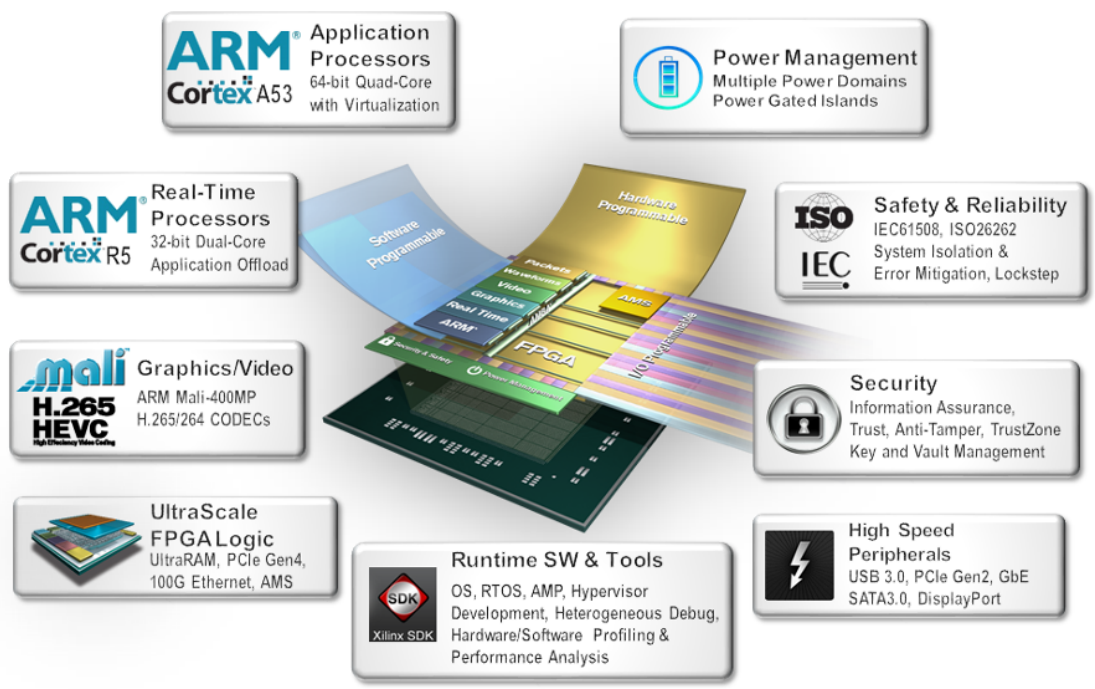
- Switch-box provides programmable interconnect
  - Local interconnects are fast and short
  - Horizontal and vertical interconnects are of various lengths



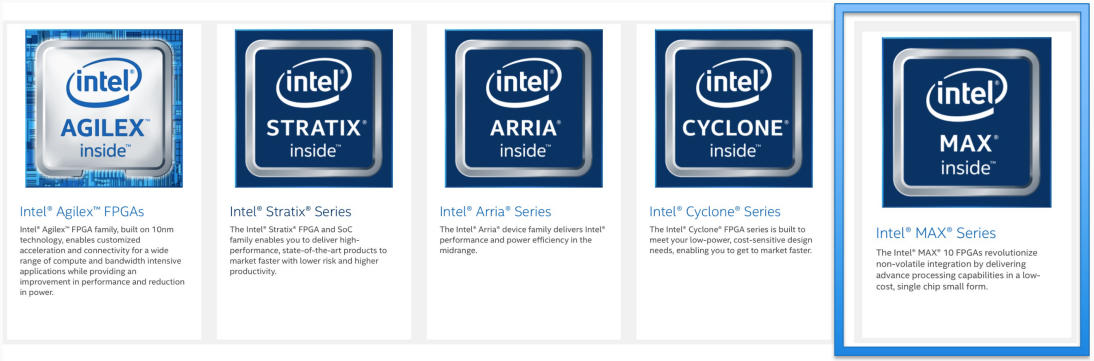
End of the Dinosaurs Age

Modern FPGA devices – Heterogeneous

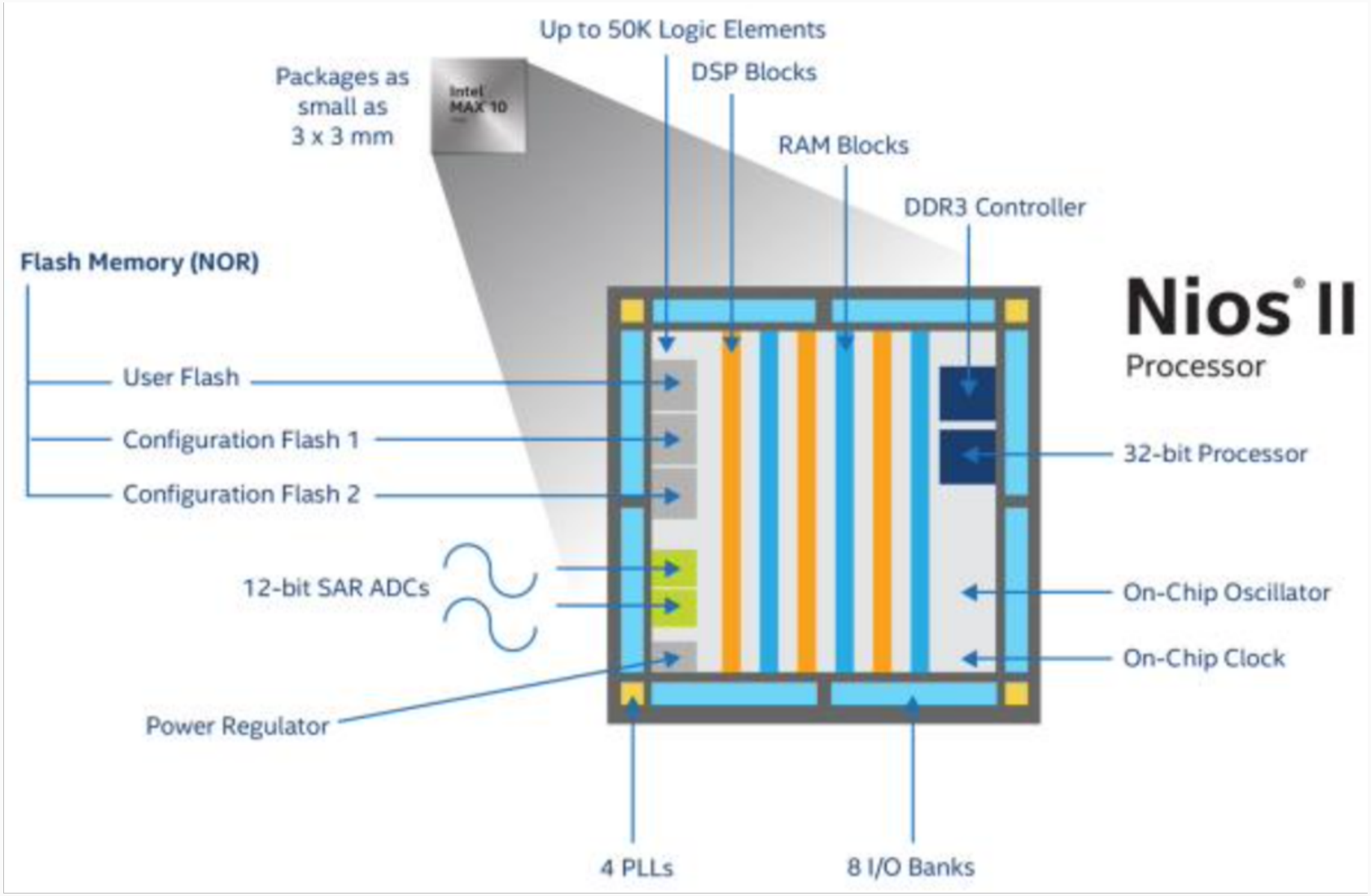
Pre-built ASIC components are already integrated



Current available FPGAs from Intel



NIOSII: the soft core at MAX10



Questions?

Questions?