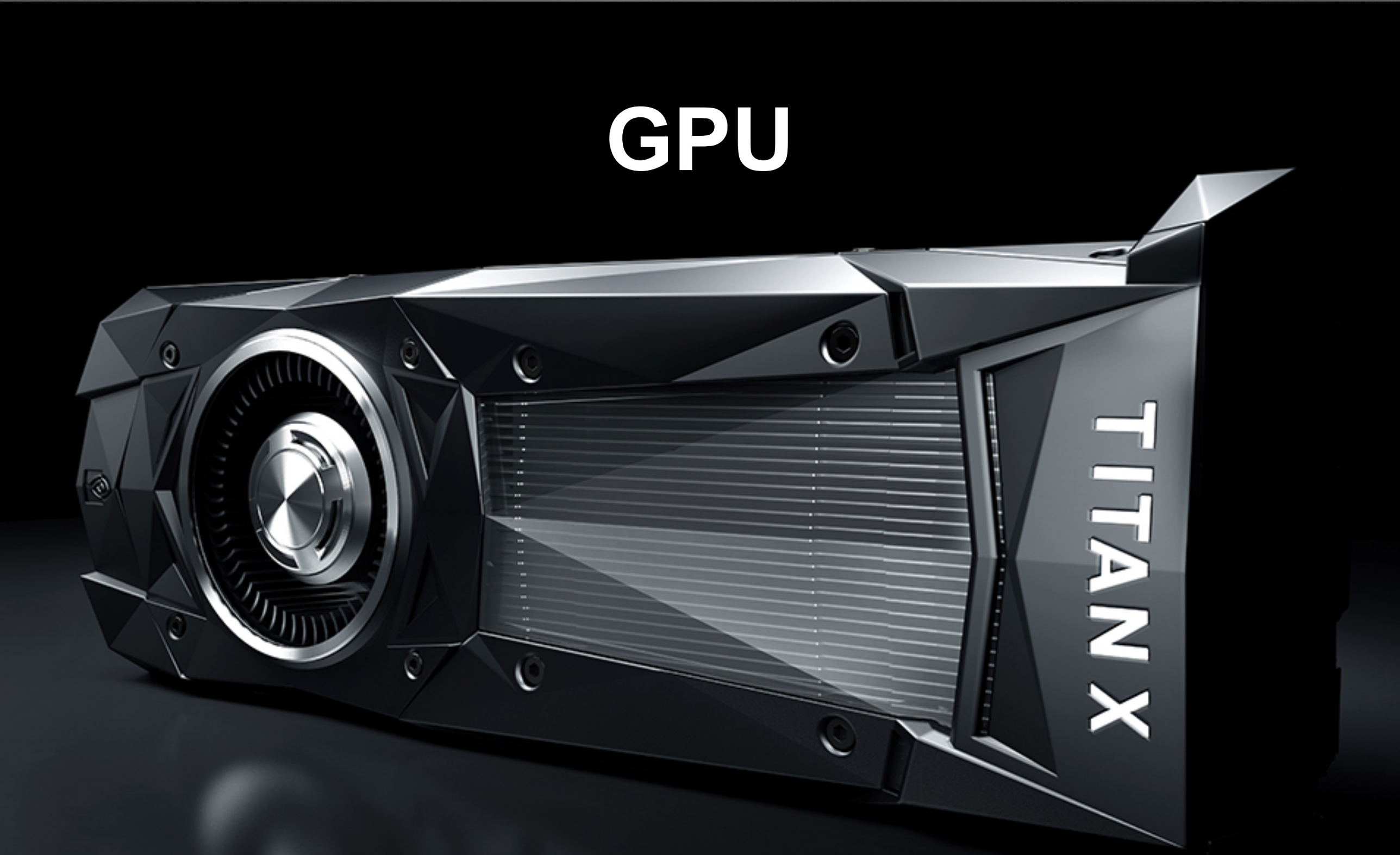
A Brief Introduction to GPU, TPU and FPGA in Deep Learning Field

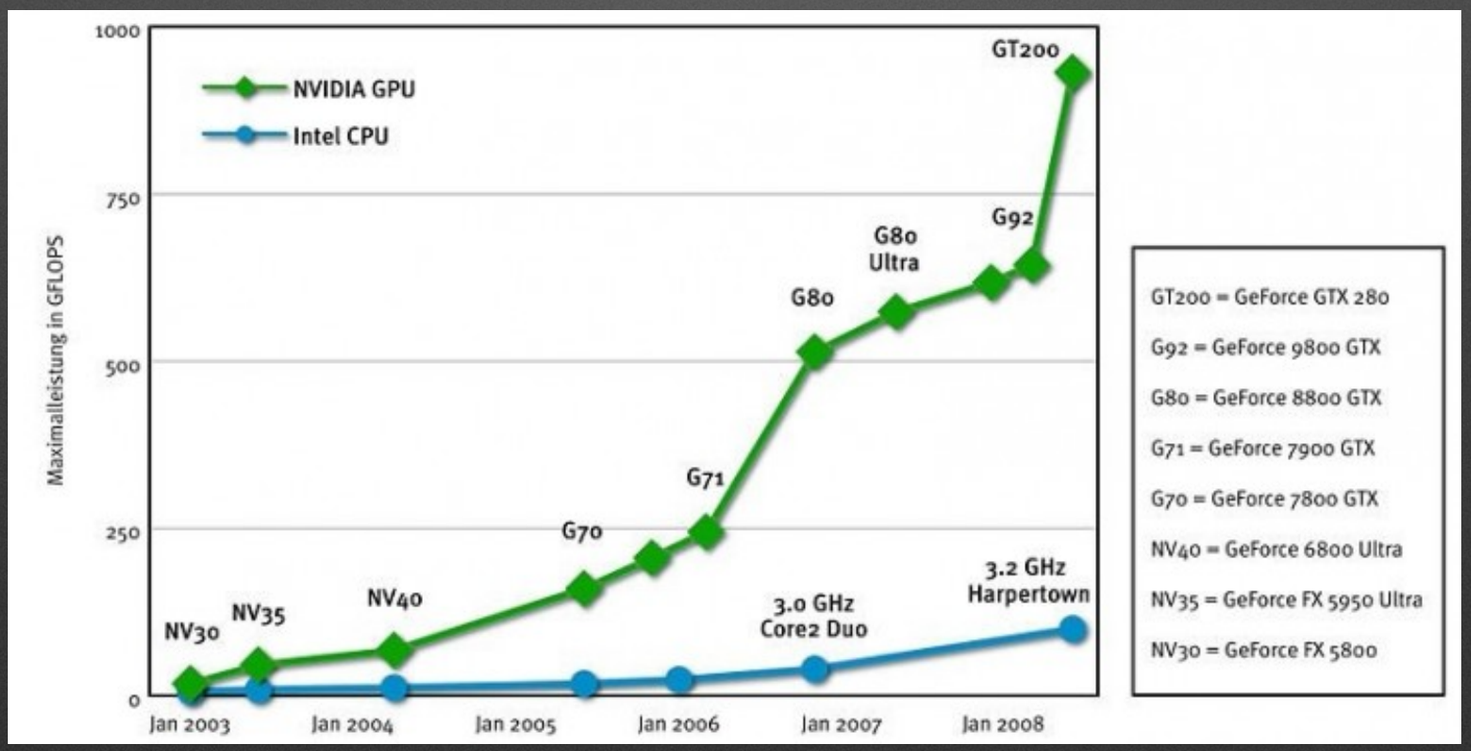
GPU



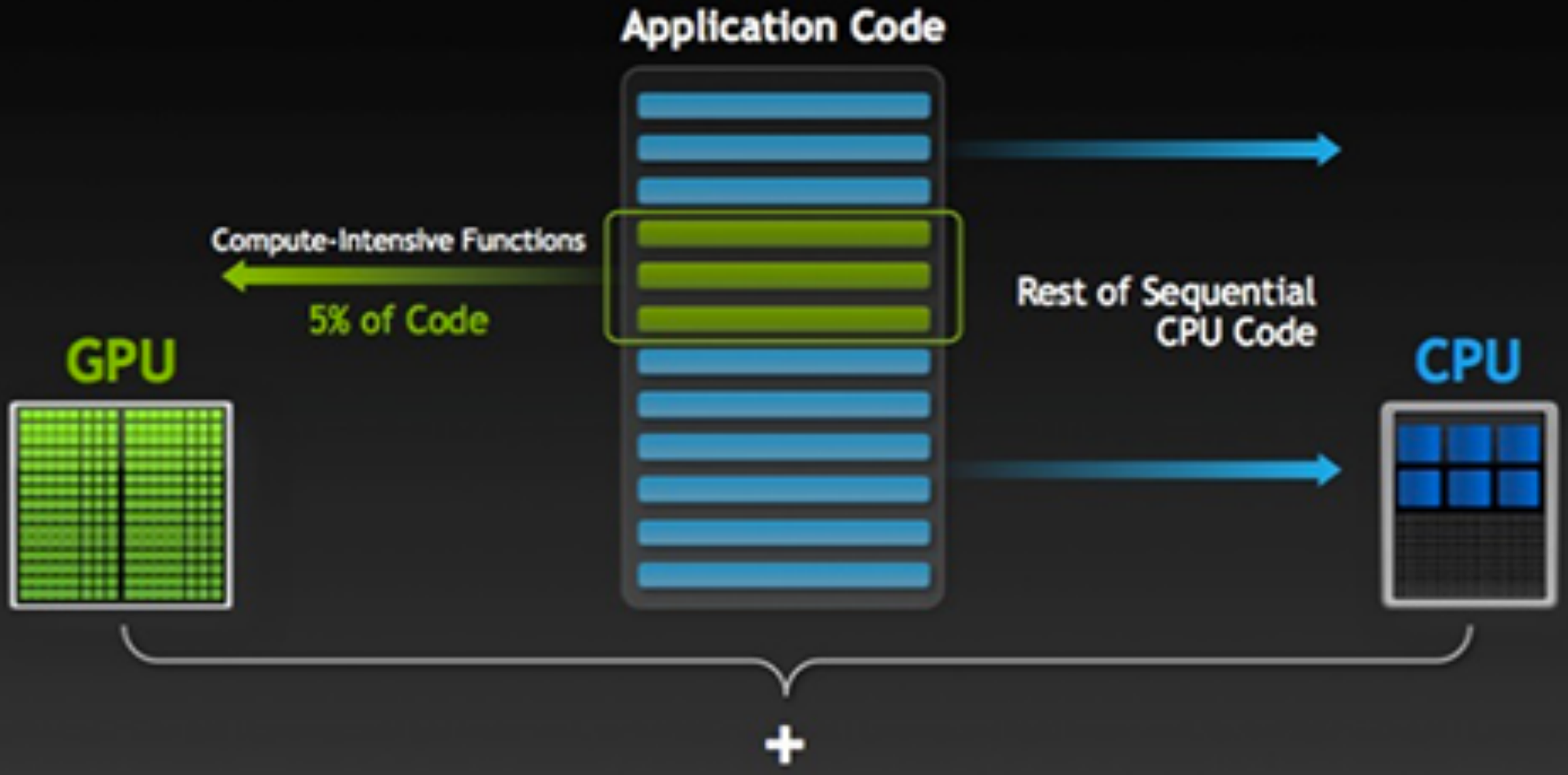
CPU的微架构示意图

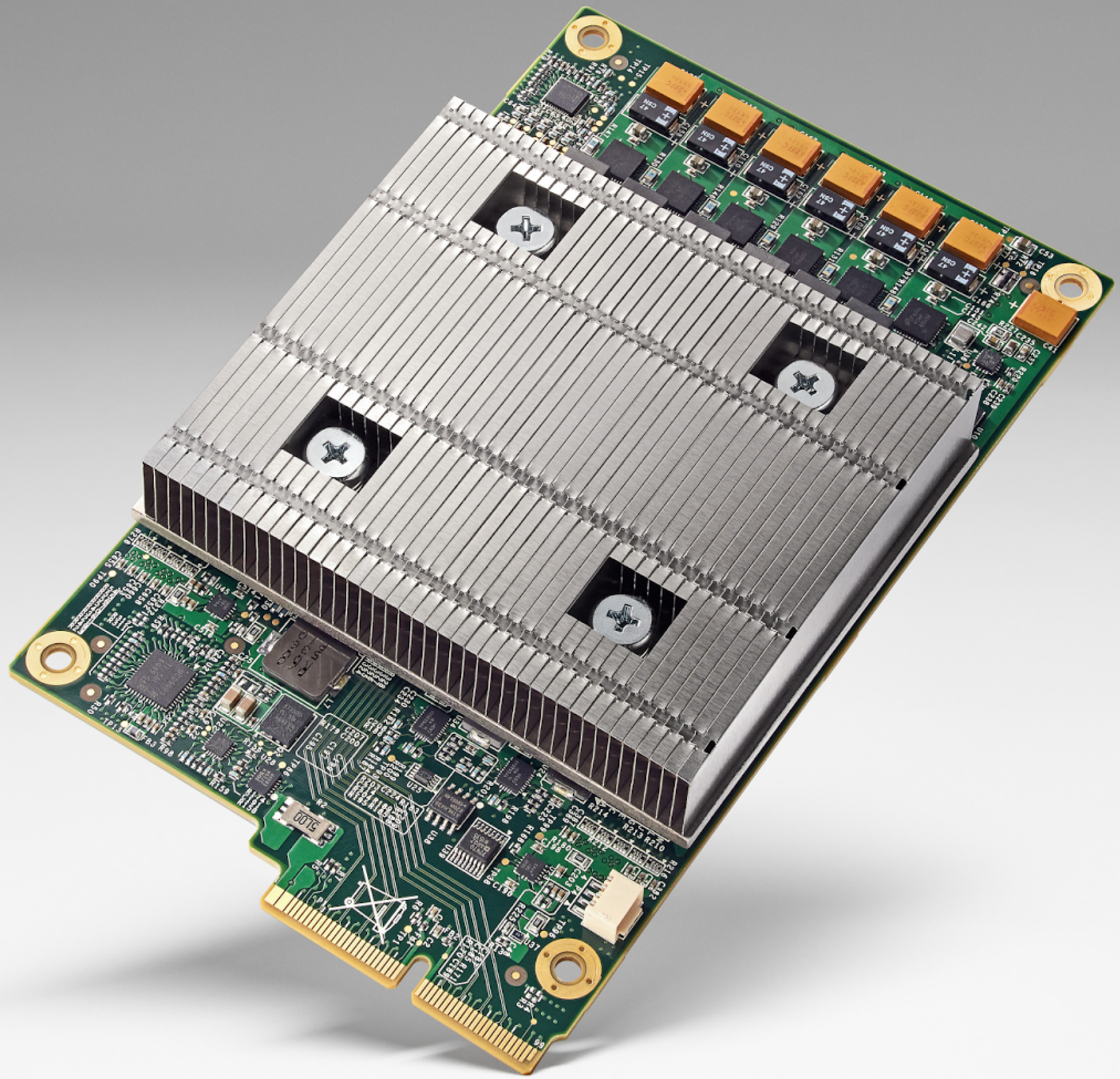
（ALU-用于计算的晶体管）

GPU的微架构示意图



How GPU Acceleration Works





TPU

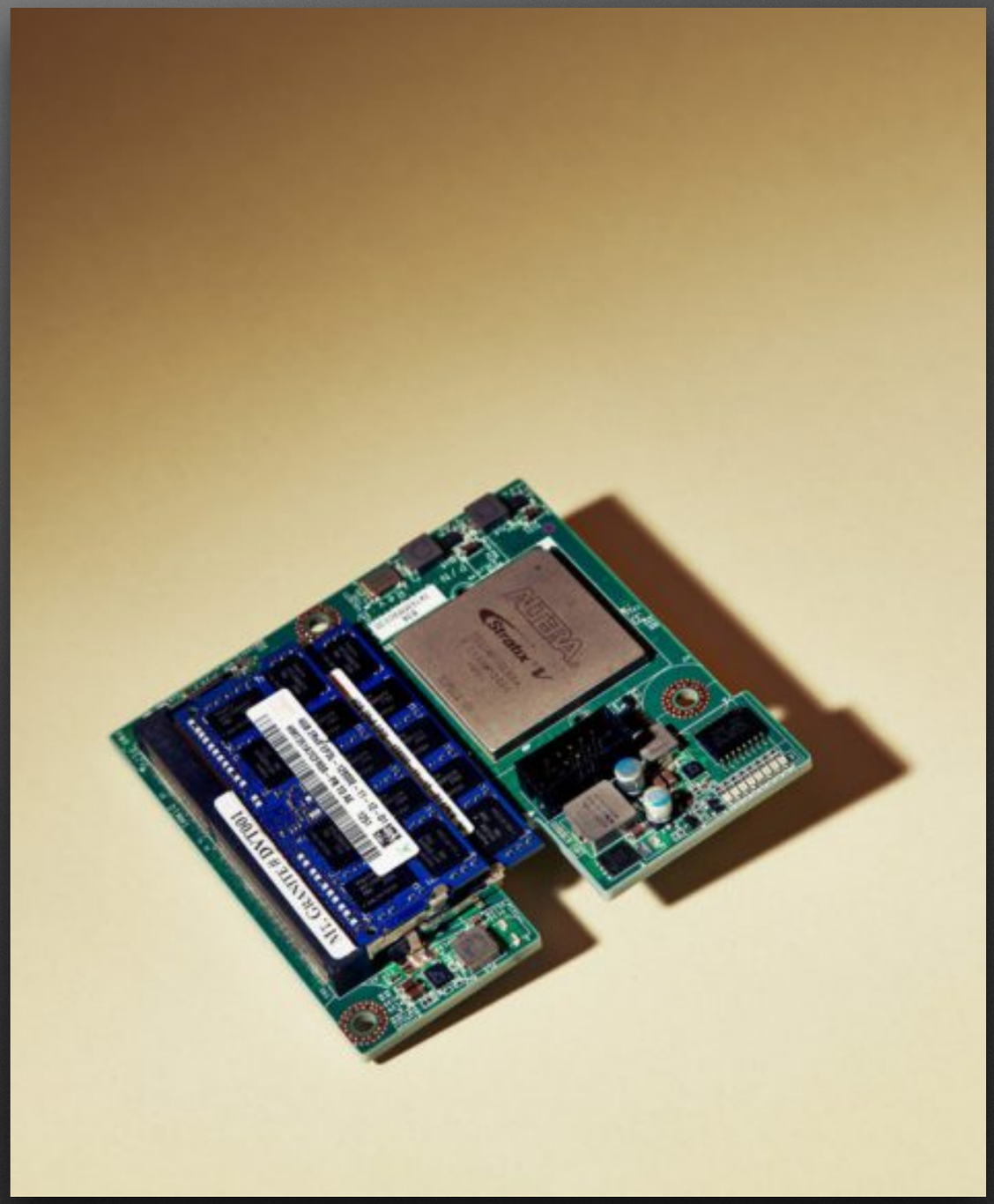
Tensor Processing Unit is custom ASIC google built specifically for machine learning — and tailored for TensorFlow. Google have been running TPUs inside data centers for more than a year, and have found them to deliver an order of magnitude better-optimized performance per watt for machine learning.

	AlphaGo Vs. FAN Hui (without TPU)	AlphaGo Vs. LEE Sedol (with TPU)
CPU	1202	48
GPU	174	8

FPGA

Project Catapult

- Microsoft has revealed that Altera FPGAs have been installed across every Azure cloud server, creating what the company is calling “the world’s first AI supercomputer.”
- The Catapult hardware costs less than 30 percent of everything else in the server, consumes less than 10 percent of the power, and processes data twice as fast as the Microsoft could without it.



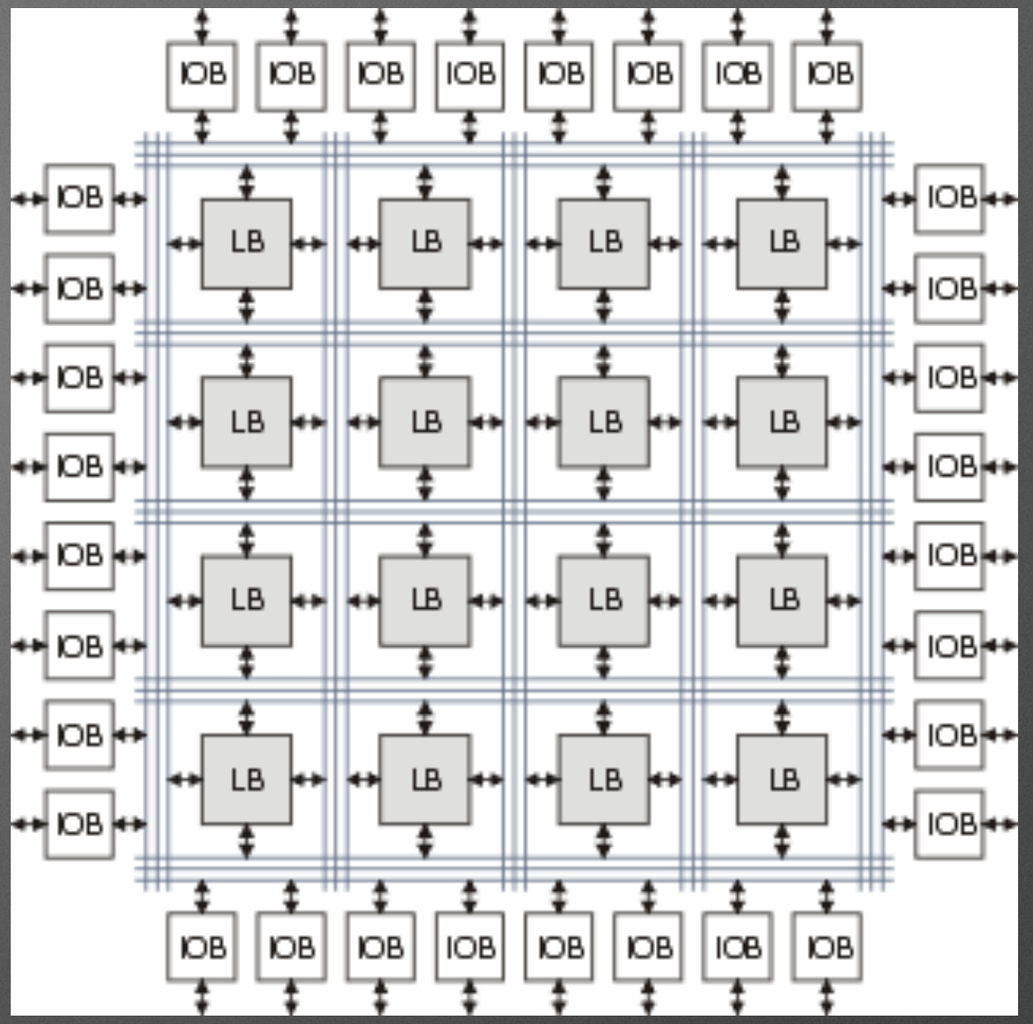
A server equipped with four FPGAs translated all 1,440 pages of the famous Russian novel War and Peace into English in just 2.6 seconds; a single 24-core CPU server took 19.9 seconds and required 60 more watts of power to do so.

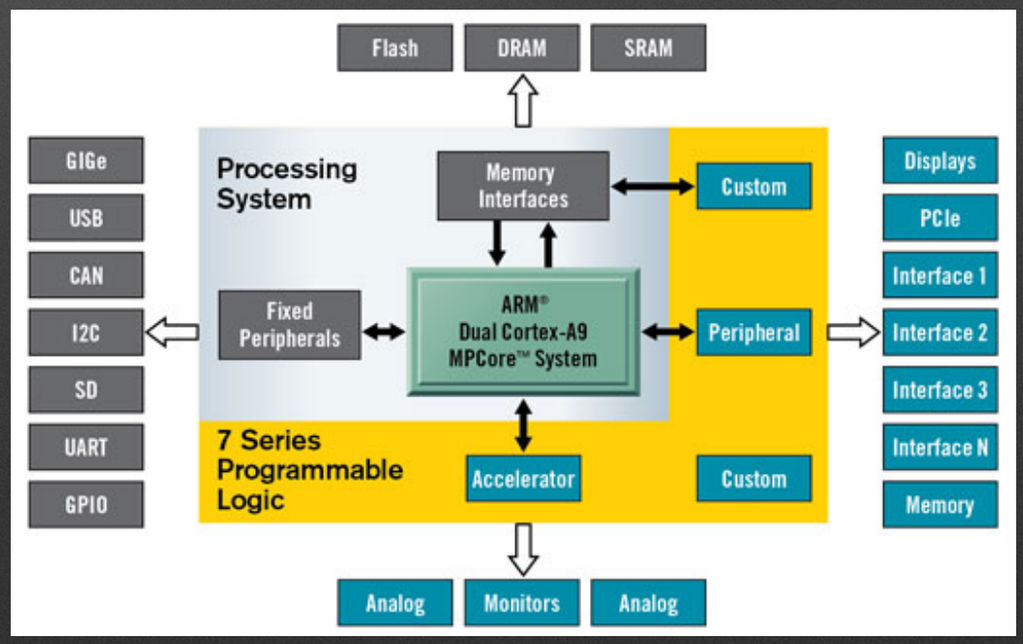
FPGA

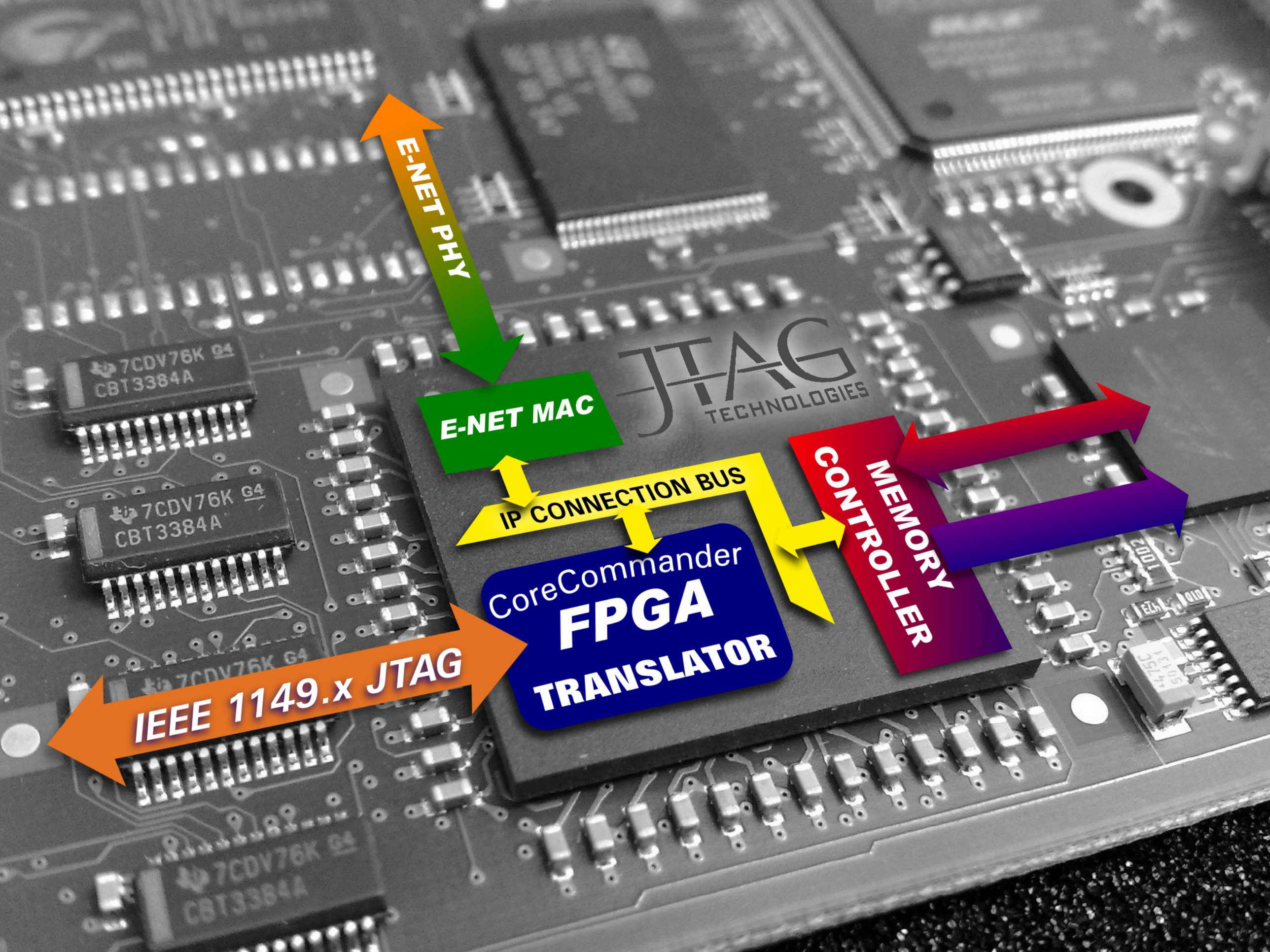
A field-programmable gate array (FPGA) is an integrated circuit designed to be configured by a customer or a designer after manufacturing – hence "field-programmable".

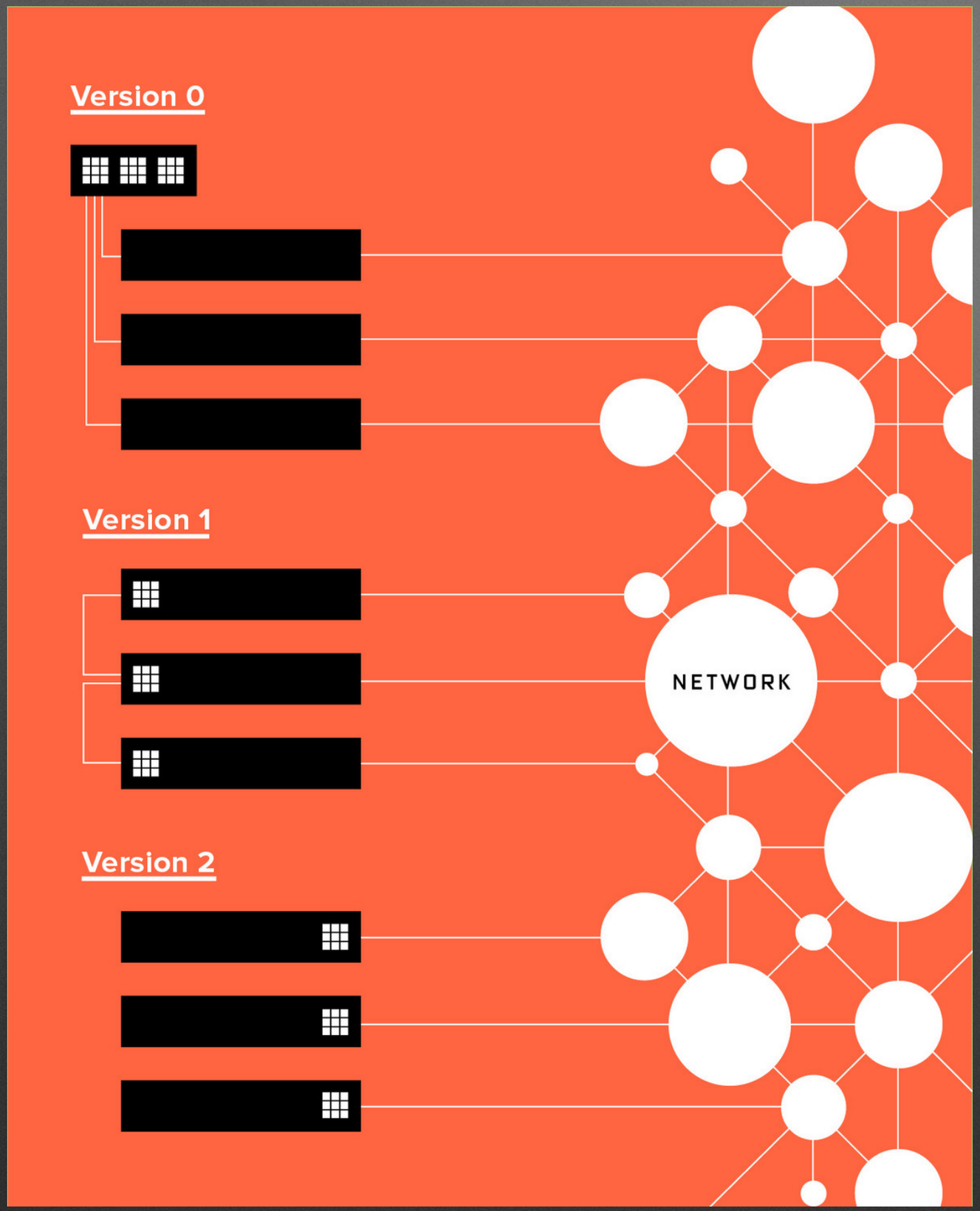
FPGA with IP Cores

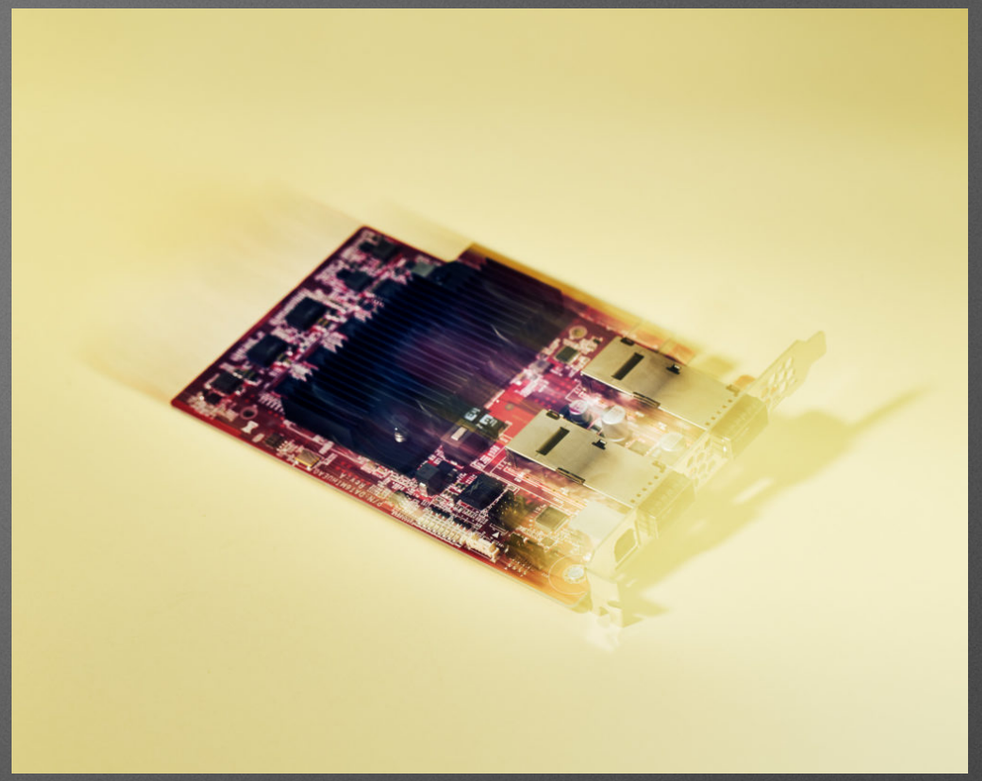
- Hard Cores
- Soft Cores
- Firm Cores

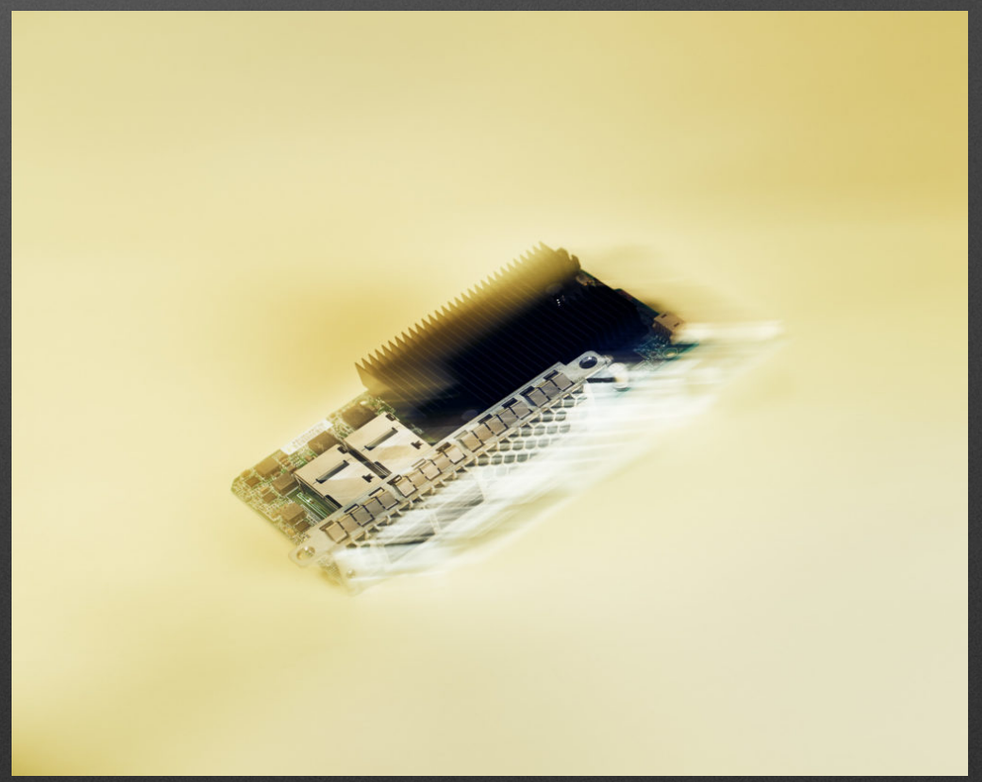





E-NET PHY
E-NET MAC
IP CONNECTION BUS
CoreCommander
FPGA
TRANSLATOR
IEEE 1149.x JTAG
MEMORY CONTROLLER
JTAG
TECHNOLOGIES


Version 0
Version 1
NETWORK
Version 2





	CIFAR-10 [4]	ImageNet 1K [1]	ImageNet 22K [2]	Max Device Power
Catapult Server + Stratix V D5 **[3]**	2318 images/s	134 images/sec	91 images/sec	25W
Catapult Server + Arria 10 GX1150 **[8]**	-	~233 images/sec (projected)	~158 images/sec (projected)	~25W (projected)
Best prior CNN on Virtex 7 485T **[5]**	-	46 images/sec[3]	-	-
Caffe+cuDNN on Tesla K20 **[6]**	-	376 images/sec	-	235W
Caffe+cuDNN on Tesla K40 **[6]**	-	500-824 images/sec[4]	-	235W

Q&A