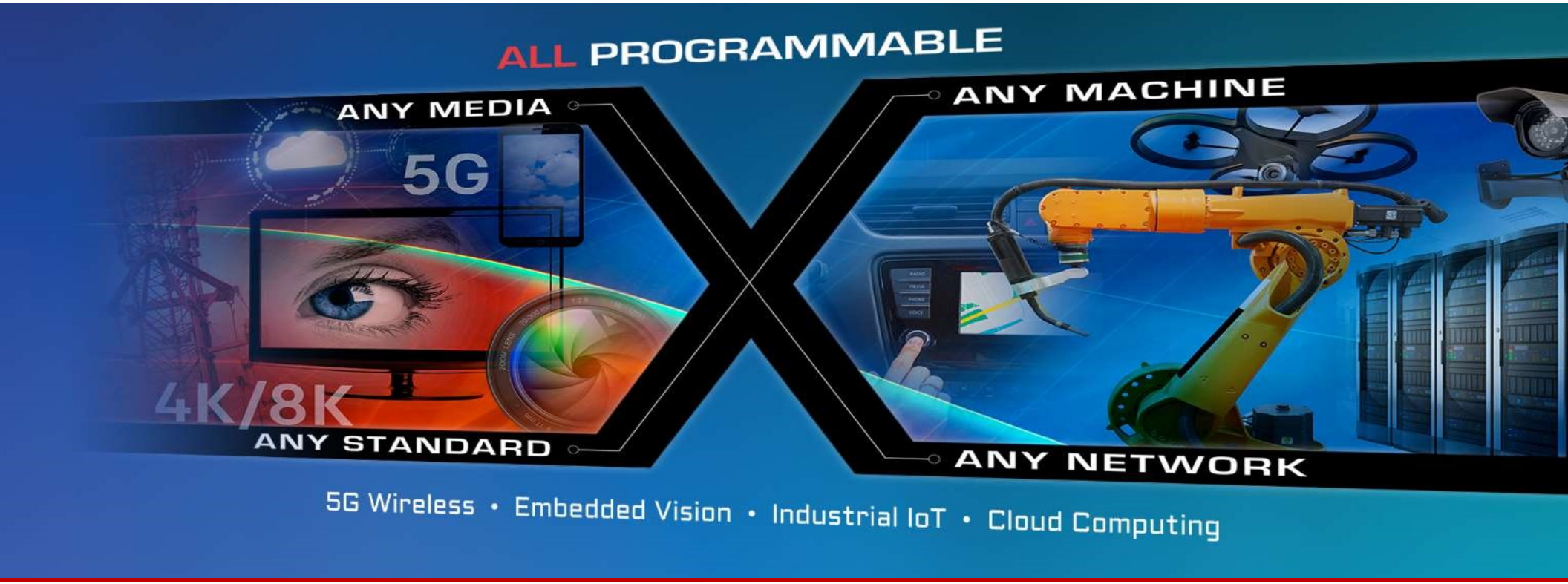


XILINX
ALL PROGRAMMABLE™

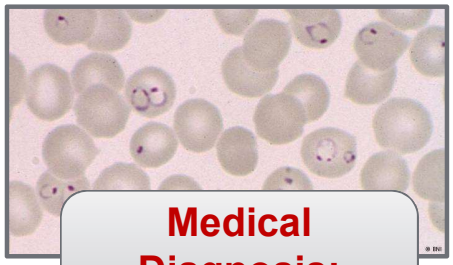
The computational battle for deep learning

Kees Vissers

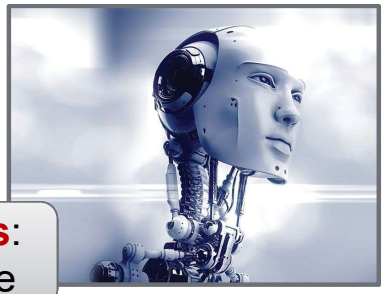
Challenge : Diverse Applications with Diverse Design Targets

Translate & AlphaGo:
Huge networks

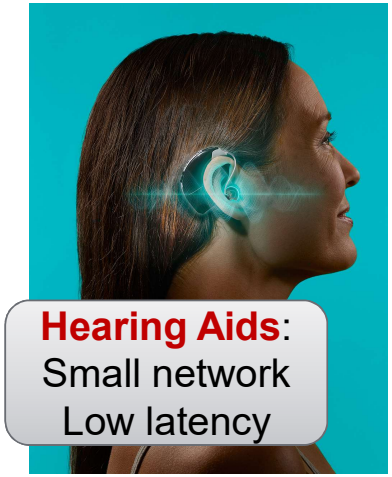
AlphaGo



Medical Diagnosis:
Small networks

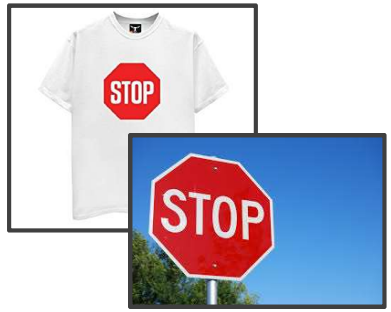


Robotics:
Real-time



Hearing Aids:
Small network
Low latency

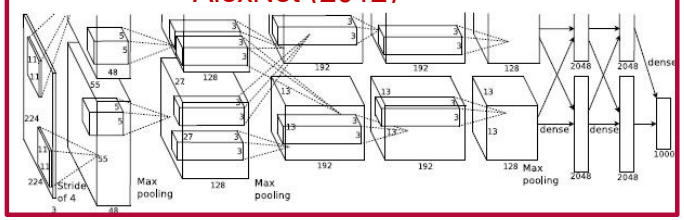
ADAS
High accuracy
Low latency



Challenge:
Different use cases require different networks & different figures of merits (speed, latency, energy, accuracy)

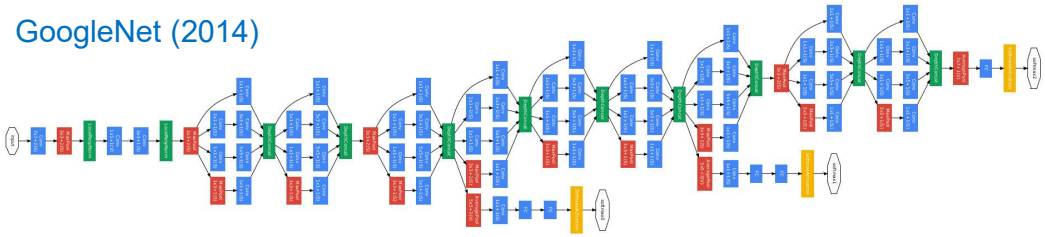
Challenge: Neural Networks Will Continue to Change

AlexNet (2012)

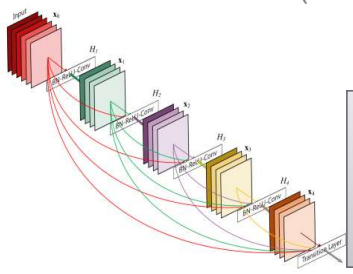


Number and **types** of layers are changing

GoogleNet (2014)



DenseNet (2016)



Data representations and **quantization methods** are changing

Graph Connectivity is changing

Challenge 2:
Continuous stream of new algorithms

What is the opportunity and what matters

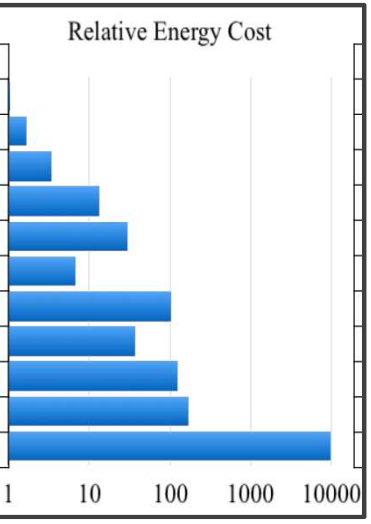
- What is deep learning: Neural Networks
  - Convolutional Neural Nets
  - LSTMs
  - Reinforcement learning
  - Training or interference

- What are figures of merit for a solution:
  - $$
  - Performance/Watt/$
  - Absolute Power bound, e.g. 5W, 50W, 500W?
  - Latency
  - Hours/days to program a solution

The title of the panel is already wrong?

- The ~~computational~~ battle for deep learning: the **energy** battle!,
- so it is about minimizing the data movement: programmed memory transfers versus caches
- Historically: supercomputing, today DATAcenters.

Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FP Mult	1.1
32b FP Mult	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640



What do we have?

- CPU: optimized for SpecInt and SpecFp, and for Databases, OS
  - Some cores, large caches, cache coherency, given Memory, Disk, Network abstractions
- GPU: optimized for Graphics, re-purposed for High Performance Compute
  - Large number of small cores, vector processing
  - Multi-threading to hide memory latency
  - External memory synchronization model, very high external memory bandwidth (power!!)
  - Historically Floating Point, now tuned to the problem (16bit, 8bit)
- FPGAs, a range of devices:
  - Do it yourself anything, ASIC programming flow
  - Started as bit-oriented logic, now includes word oriented DSPs (MAC), Memory (BRAM, URAM), processors (ARM, Microblaze), High-Speed I/O.
  - Very good at networking and networking interfaces (bit oriented, high speed)
- Programmers that want to use Python, Caffe, Tensorflow, MxNet…

What do we want

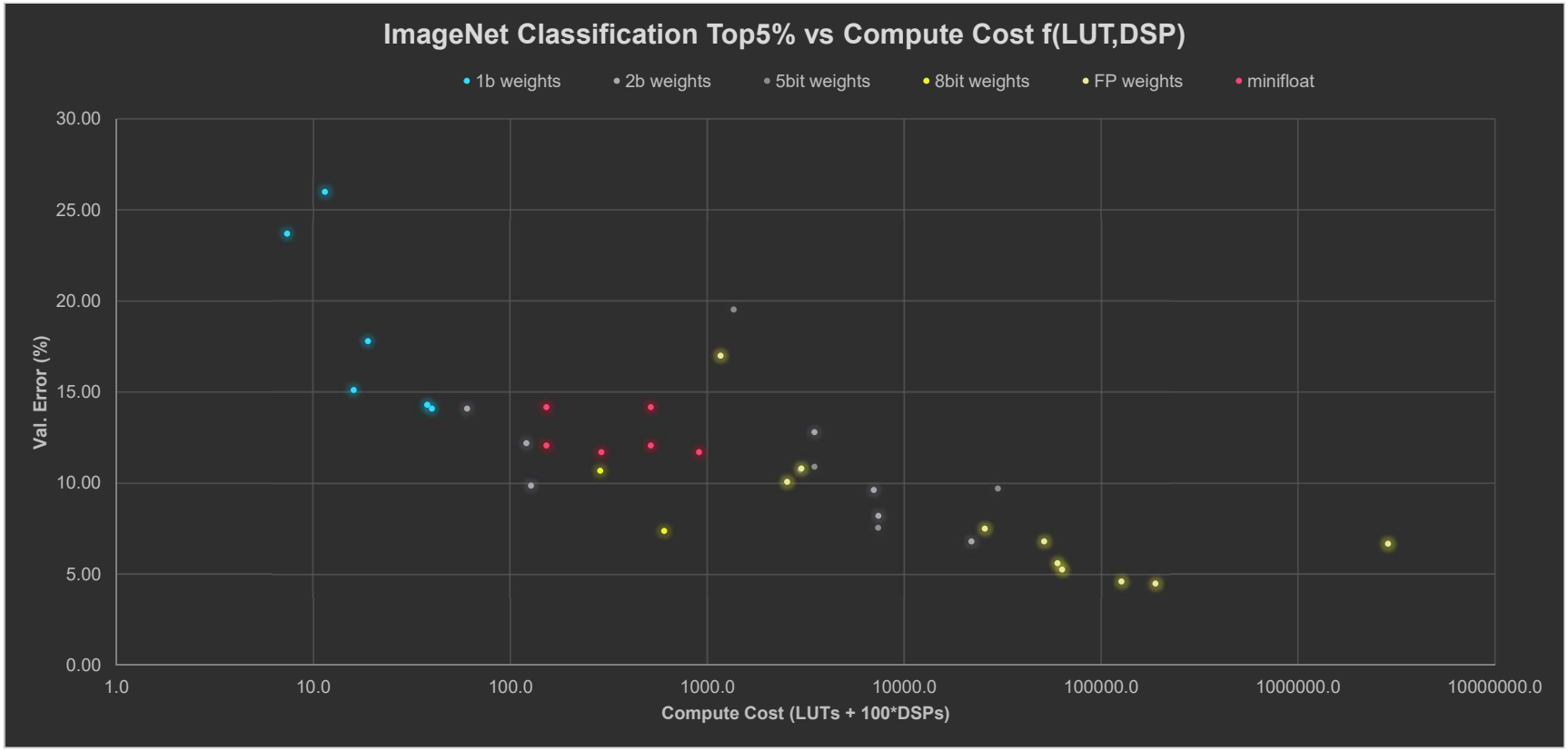
- 1-10 TOPS/Watt, effective (not Peak Performance that nobody can reach)
- Exploiting massive concurrency (that is coded in 7 nested loops???)
- A tool that takes my Network in a given framework, E.g. Caffe, tensorflow, and maps that efficiently to the desired hardware (embedded and datacenter multi-server)
  - And change the algorithm all the time
  - And no bugs in the tools please.

What is the opportunity for FPGAs (and next gen)

- Programming tools (datacenter level and embedded level)
  - Overlay architectures that are efficient and tuned per application sub-domain
  - Tuned neural network to bitstream, dataflow style, architectures
  - Tool that shows retraining for reduced precision implementations
  - Interoperability (Open Neural Network Exchange ONNX, NNEF)
- tools that tell me the POWER consumption up front, for a program!
- Proposal for new abstractions that leverage benefits: e.g. streaming compute, dataflow, flow control, micro-services, (this is NOT C)

- FPGA today: low latency, any (including reduced) precision, distributed memory and amazing integration with other functions.

what accuracy: chose your NETWORK and Precision



What can YOU do

- Network designers: Novel Networks that are optimal given the hardware concepts
- Academia: new networks, concepts, abstractions, prototype tools, conceptual architectures
- Software Startups: Take above mentioned innovation and turn publications in usable products!
- Hardware startups: maybe (need to include a software startup)
- Cloud (Provider) industry: amazing services that ‘just work’. Apps on top of that.
- (FPGA) Chip Industry: Tools and tune silicon to address the opportunity given the new application domain.

Conclusion

- We are just at the start
- The opportunity is in front of us (is one team ahead?)
- Innovation in this field will go very rapidly

Follow Xilinx

XILINX
ALL PROGRAMMABLE™


facebook.com/XilinxInc


twitter.com/XilinxInc

youtube.com/XilinxInc


linkedin.com/company/Xilinx


plus.google.com/+Xilinx