XILINX
ALL PROGRAMMABLE™

Scalable Machine Learning with Reconfigurable Devices

Michaela Blott, Principal Engineer, Xilinx Research

Agenda

- **Background – Xilinx Research**
- **Machine Learning**
- **Research Efforts**
- **Summary & Outlook**

Xilinx Research - Ireland

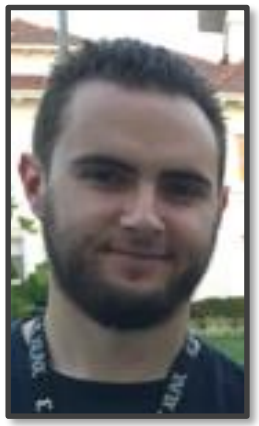
- **Since 13 years**
- **Part of the worldwide CTO organization (7 out of 36)**
- **AI Lab expansion part-financed through IDA Ireland**
- **Increasingly external funding (H202))**
- **Visiting professors on sabbatical**

Ivo Bolsens
CTO

Kees Vissers
Dist. Eng.

The Current Xlabs Dublin-Team

- **Yaman Umuroglu, Ken O'Brien, Nick Fraser, Giulio Gambardella, Alessandro Pappalardo, Peter Ogden (from left to right)**
  - More faces to be added soon



- **Plus 2 in Xilinx University Program (Cathal McCabe, Katy Hurley)**

Plus a Very Active Internship Program

- **On average 4-6 interns at any given time**
  - From top universities all over the world
- **Overall**
  - 67 interns since 2007
  - Many collaborations have come from this
  - Many found employment

Mission: Application-driven technology development

- **Identify strategic applications for Xilinx**
  - Data centers and machine learning
- **Derisk emerging technologies**
  - HLS, HBM
- **Build technology leadership through vision, communication and partnership with customers, partners, and universities**
- **Quantifying value proposition within the competitive landscape**
  - Testdriving, benchmarking, etc.
  - **Currently: for FPGAs in Machine Learning**

Agenda

- **Background – Xilinx Research**
- **Machine Learning**
- **Research Efforts**
- **Summary & Outlook**

New York Times: "The Great A.I. Awakening"

(Dec 2016)

Elon Musk's Billion-Dollar AI Plan Is About Far More Than Saving the World

The Race For AI: Google, Twitter, Intel, Apple In A Rush To Grab Artificial Intelligence Startups

World's Largest Hedge Fund to Replace Managers with an AI System

Drones Can Defeat Humans Using Artificial Intelligence

ELON MUSK'S BILLION-DOLLAR CRUSADE TO STOP THE A.I. APOCALYPSE

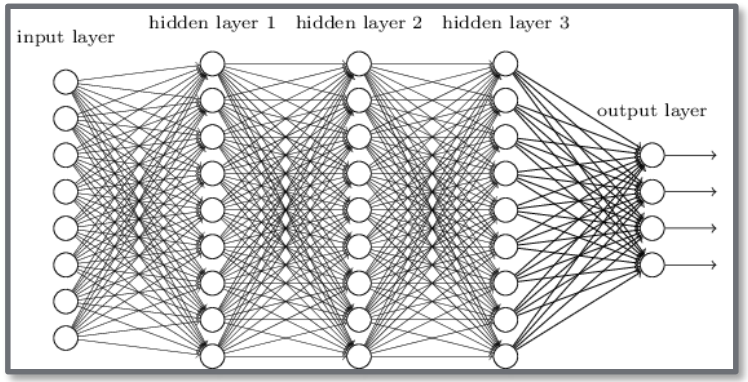
Elon Musk is famous for his futuristic gambles, but Silicon Valley's latest rush to embrace artificial intelligence scares him. And he thinks you should be frightened too. Inside his efforts to influence th rapidly advancing field and its proponents, and to save humanity from machine-learning overlords.

BY MAUREEN DOWD
APRIL 2017

- **Demonstrated to work well for numerous use cases**

CNNs – The Universal Approximator

- **Convolutional neural networks are the predominant ML algorithm used**
  - Mimics the human brain
  - Outperforming humans and traditional CV algorithms for image classification



- **NNs are the "universal approximation function"**
  - If you make it big enough and train it with enough data
  - Requires zero domain expertise
  - Will increasingly replace other algorithms (unless for example simple rules can describe the problem)

- **And solve previously unsolved problems**

Machine Learning will help address the **Grand Engineering Challenges of the 21st Century (NAE)**

- **Make solar energy economical**
- **Reverse-engineer the human brain**
- **Secure cyper space**
- **Restore & improve urban infrastructure**
- **Engineering better medicine**
- **Advance health informatics**
- **…**

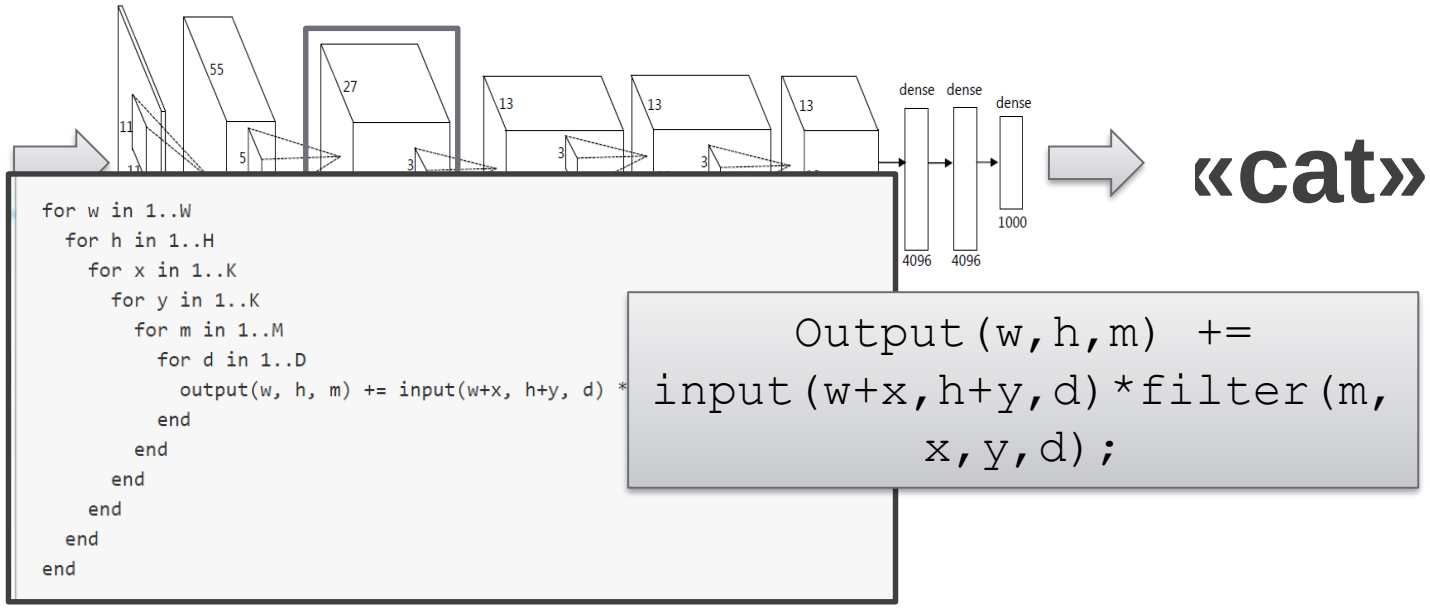
Jeff Dean, Google @ Strata Data Conference, 2018

> "I actually think machine learning is going to help with all of these," the legendary computer scientist said. "I think there are actually going to be significant breakthroughs in some of these Grand Challenges that are at least in part fueled by the fact that we now have machine learning at scale with many of these techniques that can really push us forward in the areas of commuter vision, language understanding, speech recognition, and automating and solving engineering problems."

What is the challenge?

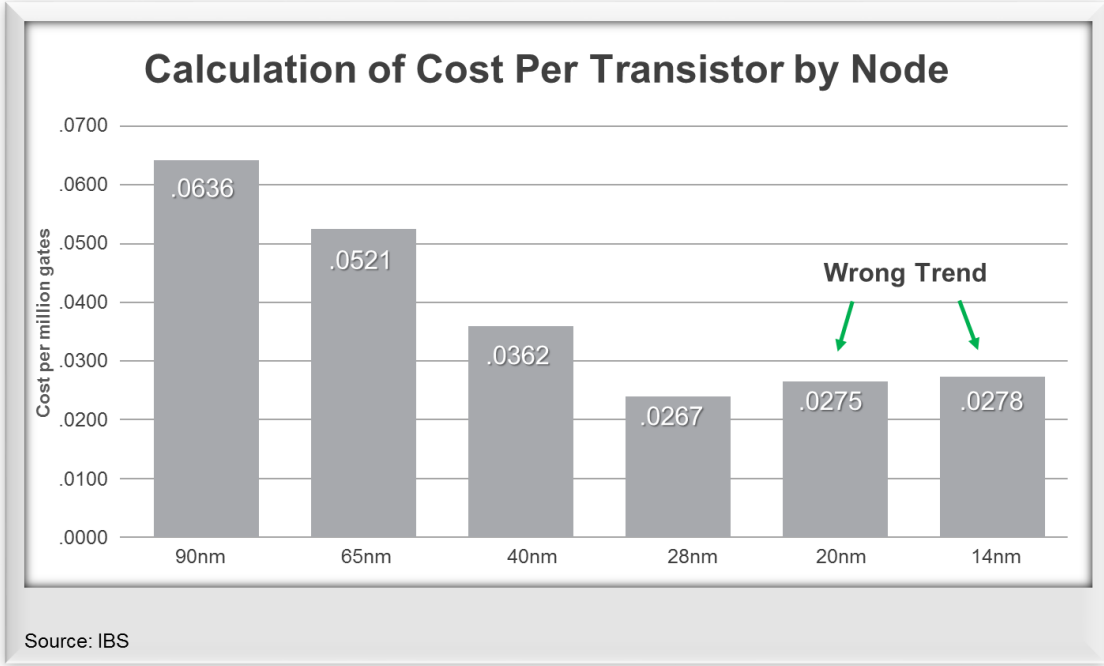
Highly Compute and Memory Intensive

- **The predominant CNN computation is linear algebra**
  - Demands lots of (simple) computation and lots of parameters (memory)
    - AlexNet: 244MB & 1.5GOPS, VGG16: 552MB & 30.8GOPS; GoogleNet: 41.9MB & 3.0GOPS for ImageNet

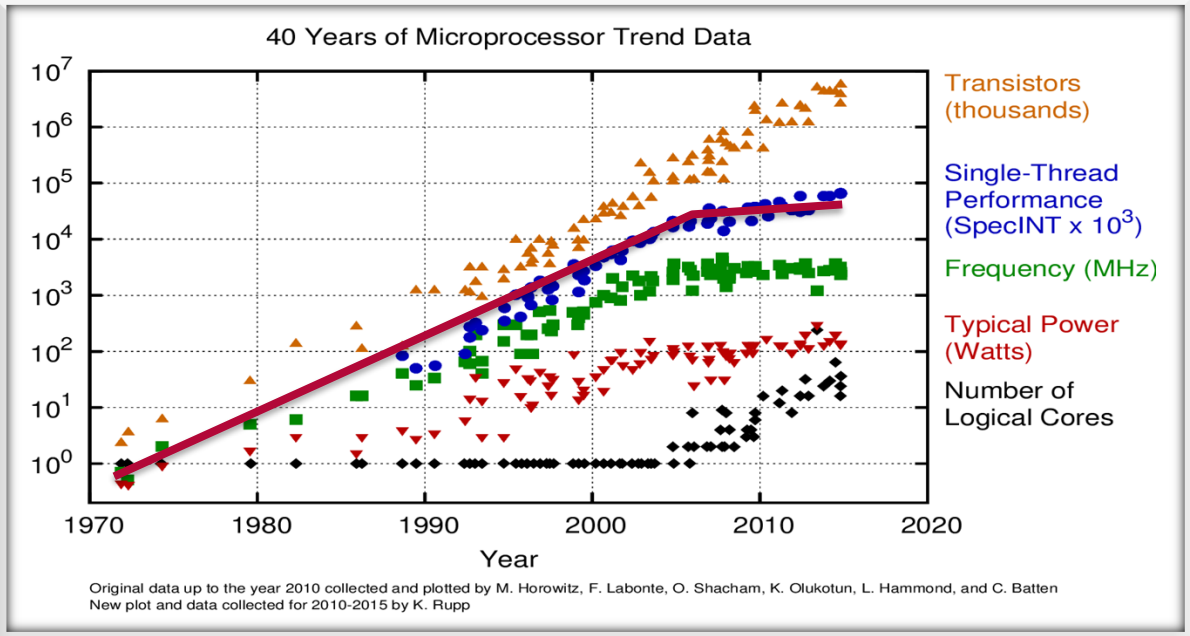


- **Billions of multiply-accumulate ops & tens of megabytes of parameter data**

On Crash course with End of Moore's Law



Source: IBS



Requires Heterogeneous Architectures & Acceleration

- **Compute performance is no-longer scaling and becomes more expensive**

Many Applications Require Different Networks

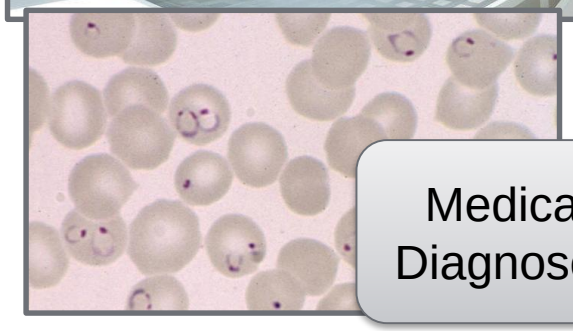
Translation Service

AlphaGo

Gaming strategy

3D reconstruction from drone images

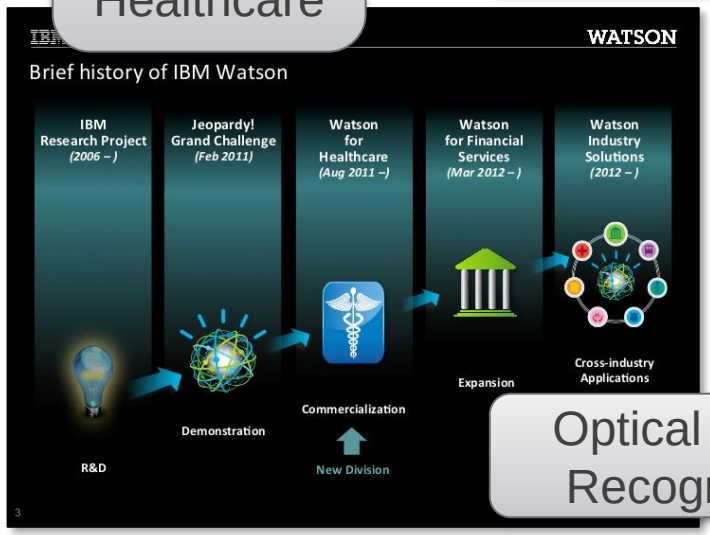
ADAS



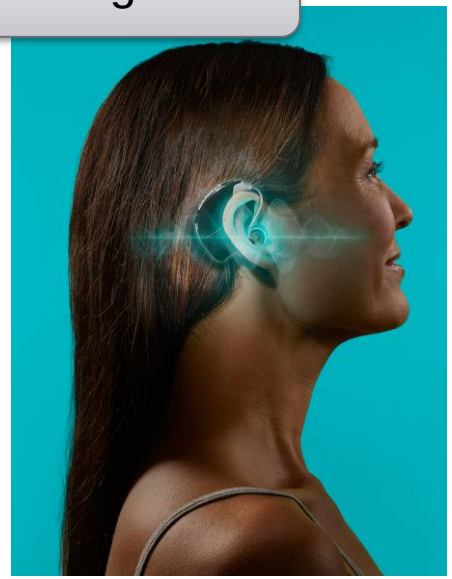
Medical Diagnoses

Real-time, sensor-based-control

Data Analysis for Healthcare



Hearing Aids



Optical Char. Recognition

Recommender Systems

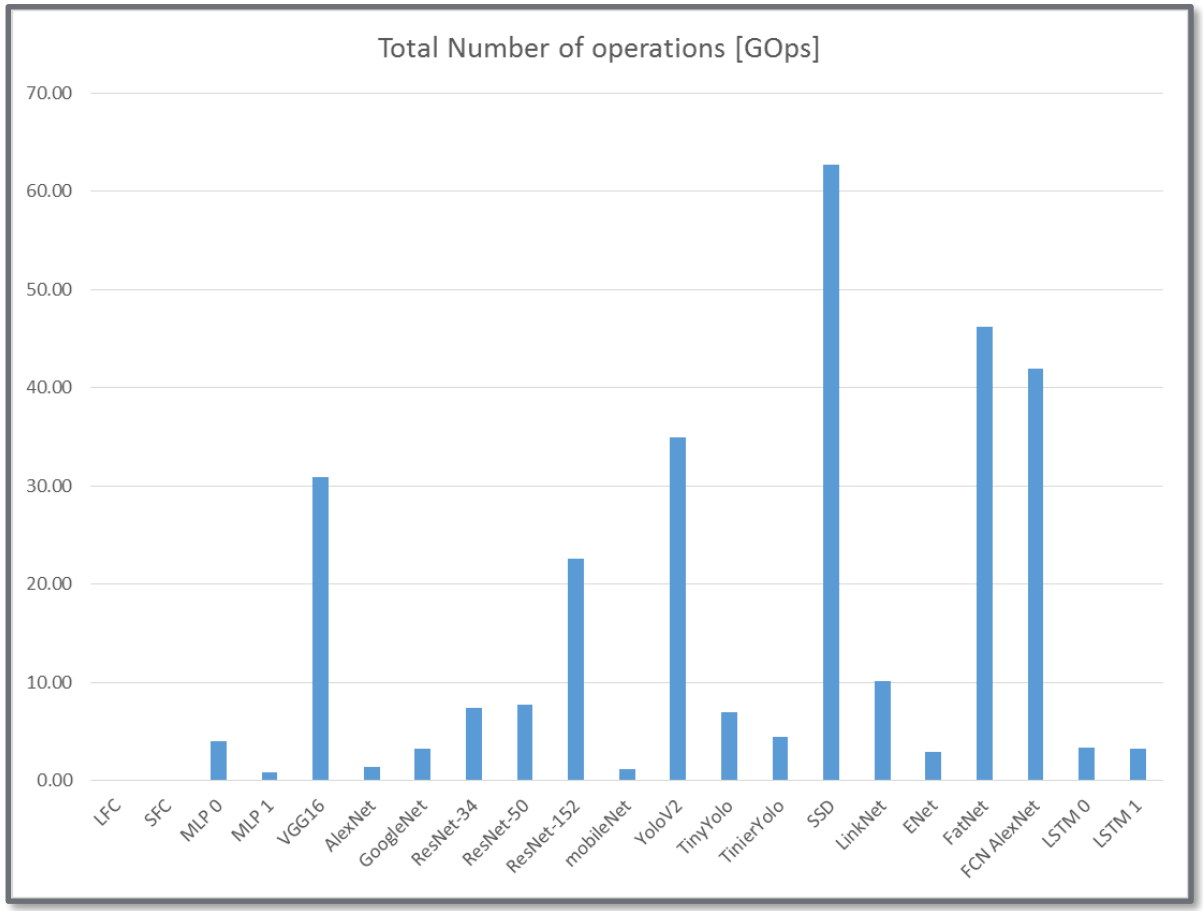
Huge Variation in Compute & Memory Requirements

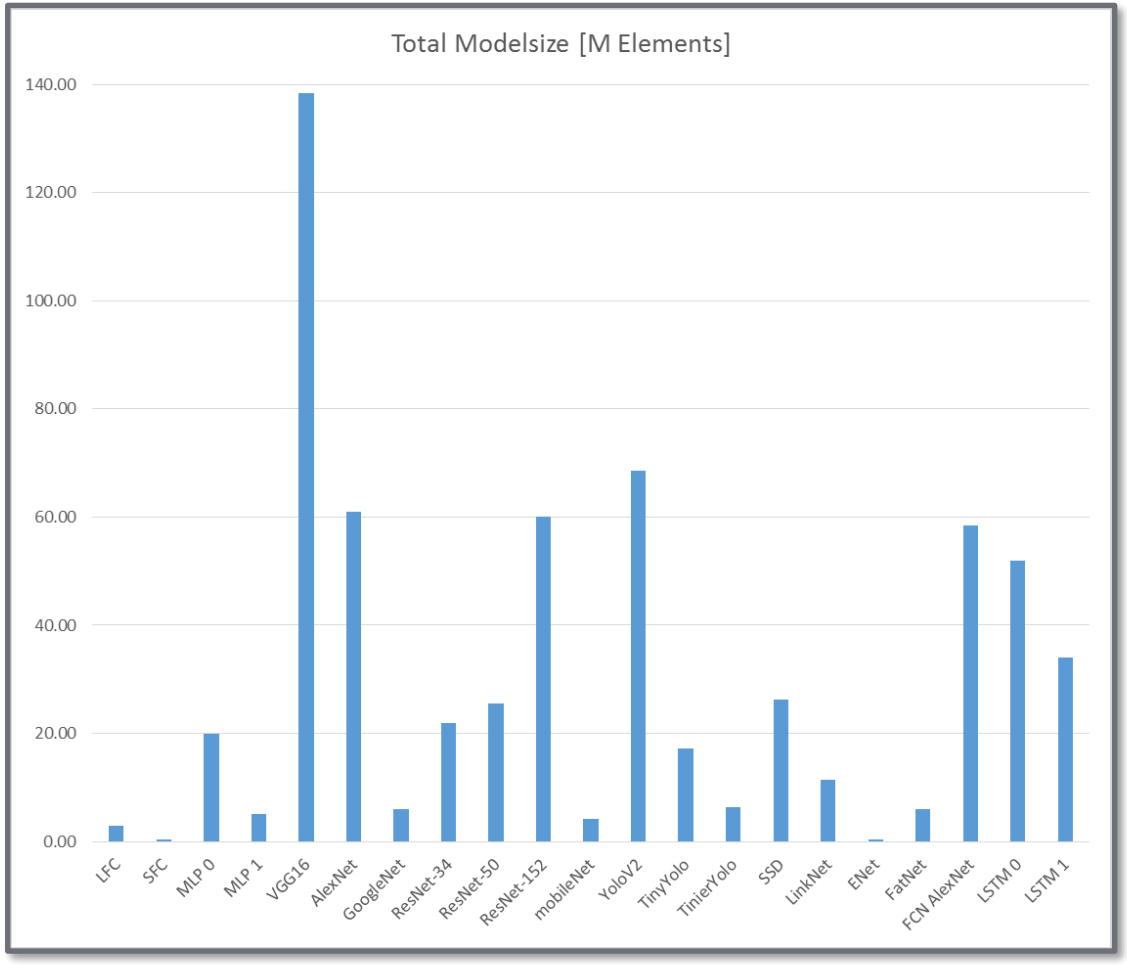
Hyperparameters	Device Performan [TOps/s]	Weights Precision [Bits]	Weights Precision [Bits]	Activation Tensors [Bits]	BRAM/URAM Bandwidth [Gbit/s]	Batch Size [Elements]	Sequence Length [Elements]	BRAM Size [Bits]	URAM Size [Bits]	URAM/BRAM Cost Factor [Ratio]
	233	8	2	8	36	16	500	36864	294912	1.7

Model			Architecture Independent					Dataflow							Activation Tensors		w/ On-Chip Weights		
Task	Neural Network Topology	Link/Ref	Total Ops w/ Batch Size = 1 [GOps]	Total Activation Tensors w/ Batch Size = [M Elements]	Total Model Size [M Elements]	Input Width [Elements]	Input Height [Elements]	TOTAL Activation Tensors w/ window buffer [M Elements]	Total Memory [M Elements]	Number of BRAMs for Activation Tensors 8-bit [BRAM 36K]	Number of URAMs for Activation Tensors 8-bit [URAM 288K]	Number of BRAMs for On-Chip Weights 2-bit [BRAM 36K]	Number of URAMs for On-Chip Weights 2- [URAM 288K]	Total Number of BRAMs [BRAM 36K]	Total Number of URAMs [URAM 288K]	Number of BRAMs for Activation Tensors 8-bit [BRAM 36K]	Number of URAMs for Activation Tensors 8-bit [URAM 288K]	Number of BRAMs for On-Chip Weights 2-bit [BRAM 36K]	Number of URAMs for On-Chip Weights 2-bit [URAM 288K]	Total Number BRAMs [BRAM 36K]
MLP	LFC	link	0.	0.	2.11			0.	2.11	3	0	16431	0	16434	0	76	0	412	0	
	SFC	internal_lir	0.	0.	0.13			0.	0.13	3	0	45729	0	45732	0	298	0	429	0	
	MLP 0	link	4.	0.	20.			0.	20.											
	MLP 1	link	0.84	0.	5.			0.	5.											
Image Classification	VGG16	link	30.94	9.65	138.34	224.	224.	0.55	138.89	3	18	6339	0	6342	18	33	2834	4	1023	
	AlexNet	link	1.38	0.42	60.95	227.	227.	0.06	61.02	2	5	9477	452	9479	457	32	132	2	498	
	GoogleNet	link	3.22	4.83	5.98	227.	227.	0.13	6.11	25	13	4828	0	4853	13	156	175	36	6	
	Inception v3	link	19.91	14.14	23.8	299.	299.	0.64	24.44	26	37	1410	0	1436	37	74	608	40	27	
	Resnet-18	link	3.65	2.38	11.68	224.	224.	0.25	11.93	2	17	3359	0	3361	17	68	167	12	75	
	ResNet-34	link	7.35	3.64	21.78	224.	224.	1.84	23.63	2	33	5366	0	5368	33	91	164	16	110	
	ResNet-50	link	7.72	10.14	25.5	244.	244.	5.55	31.06	2	17	3433	0	3435	17	279	663	22	150	
	ResNet-101	link	15.14	15.26	44.44	224.	224.	0.48	44.92	2	34	1900	0	1902	34	105	684	22	154	
	ResNet-152	link	22.56	21.58	60.04	224.	224.	13.87	73.91	2	51	1339	0	1341	51	220	670	22	210	
	mobileNet	link	1.14	5.14	3.19	224.	224.	0.39	3.58	1	13	6855	0	6856	13	470	639	20	7	
Image Detection	YoloV2	link	34.9	8.91	67.12	416.	416.	0.56	67.68	2	21	6668	0	6670	21	2	1202	5	461	
	TinyYolo	link	6.97	2.31	18.5	416.	416.	0.15	18.65	2	9	6895	72	6897	81	19	599	3	180	
	TinierYolo	InternalLir	4.45	2.61	6.38	416.	416.	0.15	6.53	2	8	6735	0	6737	8	16	624	4	46	
	R-FCN ResNet50	link	17.21	12.59	36.47	244.	244.	6.61	43.09	3	20	1541	0	1544	20	125	682	28	165	
	SSD	link	62.75	19.78	26.28	300.	300.	1.14	27.42	30	33	6584	12	6614	45	37	4396	10	153	
Semantic Segmentation	LinkNet	link	17.18	9.3	11.35	640.	360.	0.3	11.65	4	20	3141	10	3145	30	57	1700	23	54	
	ENet	InternalLir	1.33	21.41	0.32	40.	70.	0.22	0.54	45	22	6006	0	6051	22	1673	414	85	0	
	FatNet	InternalLir	22.43	71.14	4.89	320.	560.	0.78	5.67	31	33	5918	0	5949	33	205	3086	41	3	
	FCN AlexNet	InternalLir	41.95	5.56	56.94	500.	500.	0.29	57.23	3	10	7463	0	7466	10	2	911	1	390	
Seq2Seq	OCR LSTM	link	0.16	0.01	0.16	500.	25.													
Full Model	DeepSpeech2 3xBiRNN	link	18.73	2.12	38.18	500.	161.													
	DeepSpeech2 3xBiGRU	link	55.86	2.12	113.95	500.	161.													
	DeepSpeech2 5xBiRNN	link	21.27	2.59	43.38	500.	161.													
	DeepSpeech2 5xBiGRU	link	63.51	2.59	129.55	500.	161.													
	ePhi DeepSpeech2 5xBiLS	link	24.86	1.72	50.68	500.	161.													
	LSTM 0	link	3.33	NA	52.	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	LSTM 1	link	3.26	NA	34.	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Recurrent Layers Only	DeepSpeech2 3xBiRNN	link	18.56	1.09	37.88	500.	161.													
	DeepSpeech2 3xBiGRU	link	55.7	1.09	113.65	500.	161.													
	DeepSpeech2 5xBiRNN	link	21.11	1.64	43.09	500.	161.													
	DeepSpeech2 5xBiGRU	link	63.35	1.64	129.26	500.	161.													
	ePhi DeepSpeech2 5xBiLS	link	24.71	0.95	50.41	500.	161.													

* Recurrent models: assumes 1 output element per 1 input element of the input sequence, summation to reduce a bidirectional layer
* counting FM buffers, full dimensions once between layers
* for Weight Staging Bandwith: only convolutional layers considered
* for total memory for systolic array factor 2 for weights included

Huge Variation in Compute & Memory Requirements

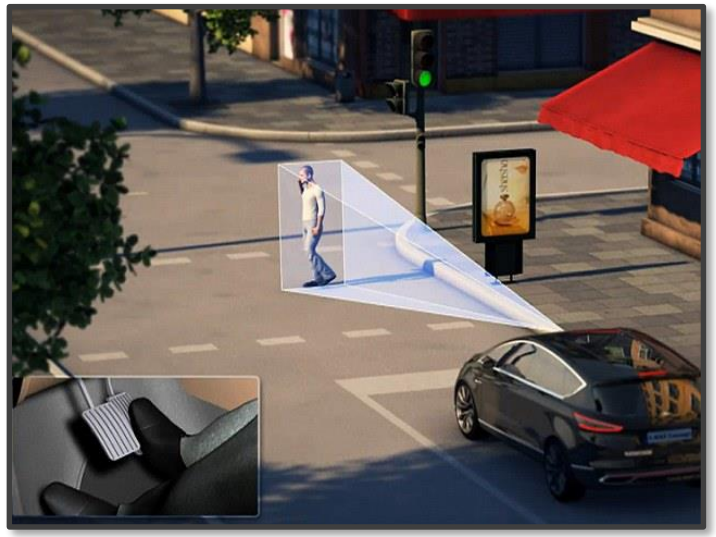


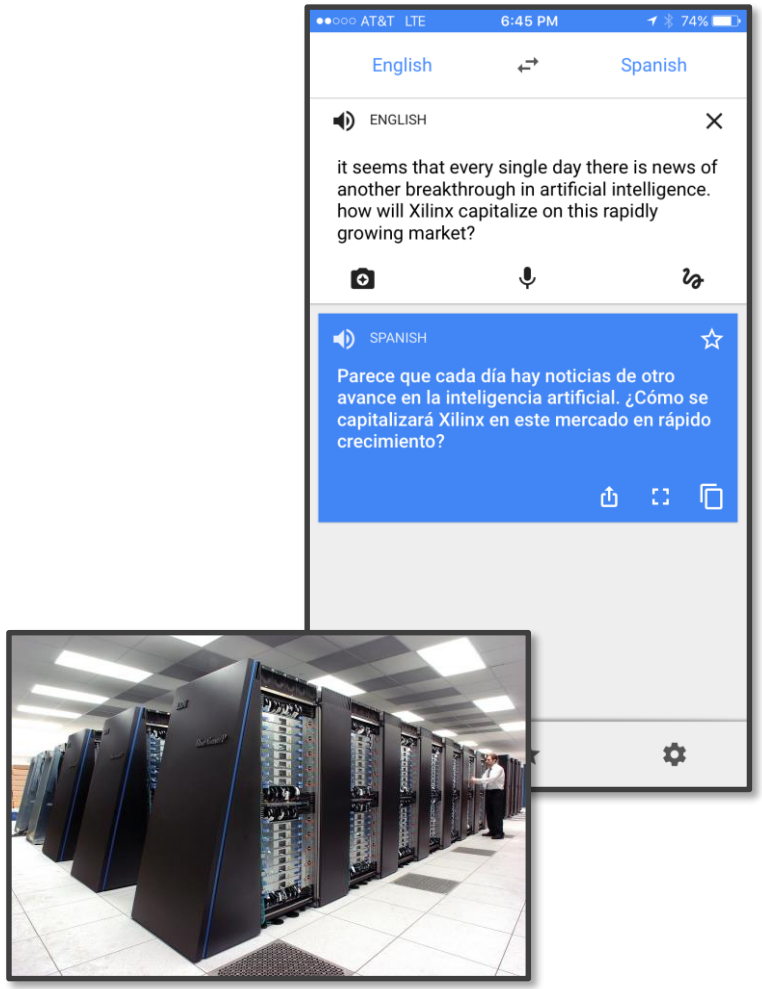


Different Use Cases, Different Figures of Merits

Accuracy, speed, power, latency, cost

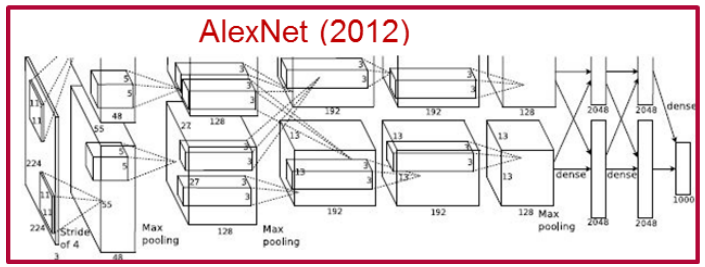
- **Accuracy** is most important for ADAS
- Hearing aids require low **power** and very low **latency**
- Robotics & online services require < 7ms response time
- Augmented reality requires high **throughput**…

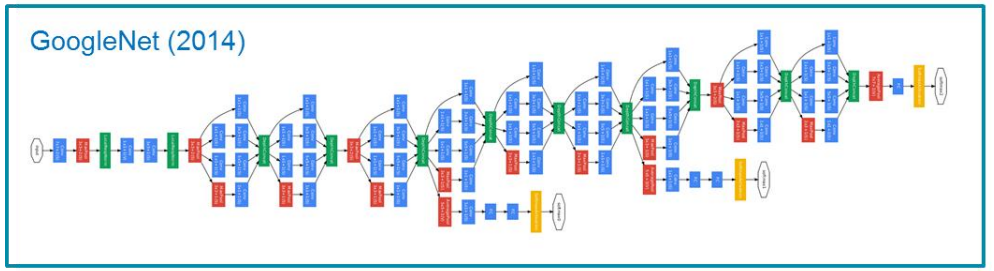


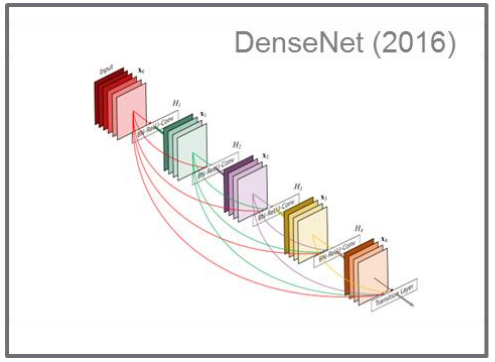


Neural Networks Change @ Increasing Rate

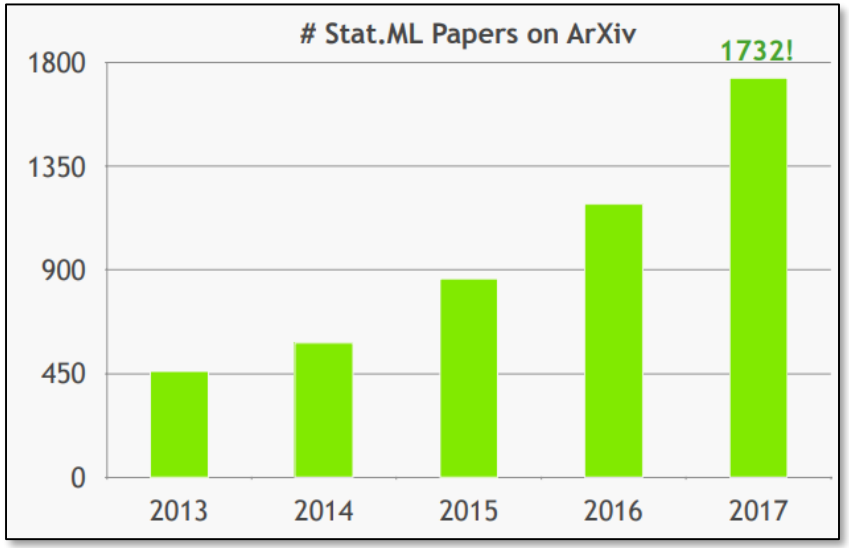
- **Graph connectivity, number and types of layers are changing**







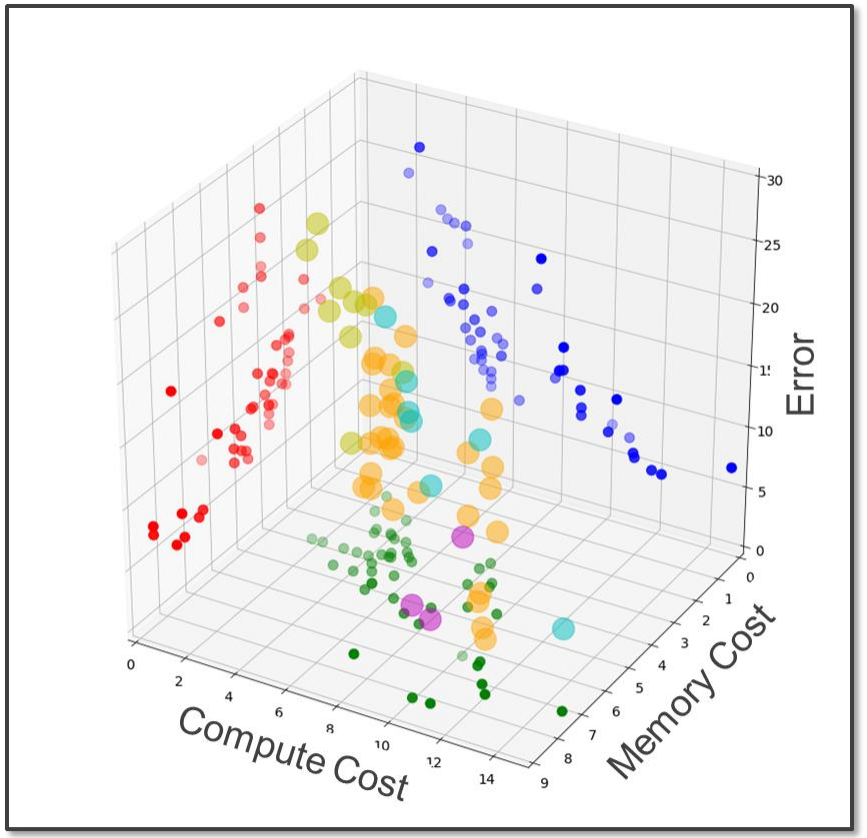
- **Increasing stream of research**



Ce Zhang, ETH Zurich, Systems Retreat 2018

Machine Learning Challenges

- **Machine Learning is a very demanding use case**
  - High compute and memory demands
  - With a high variation
- **Complicated design space**
  - Different applications
  - Different and changing algorithms
  - Different figures of merits
- **Need to be addressed through architectural and algorithmic innovation**
  - As semiconductor industry is facing the end of Moore's Law



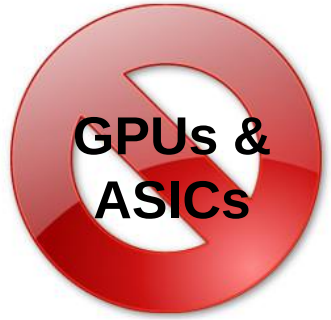
ImageNet Classification: Error, compute cost, memory requirements, topology

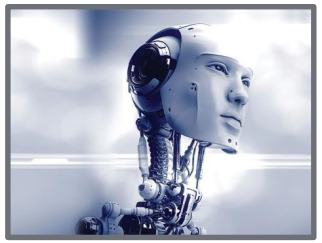
FPGA Opportunity

Flexibility to adjust to use cases

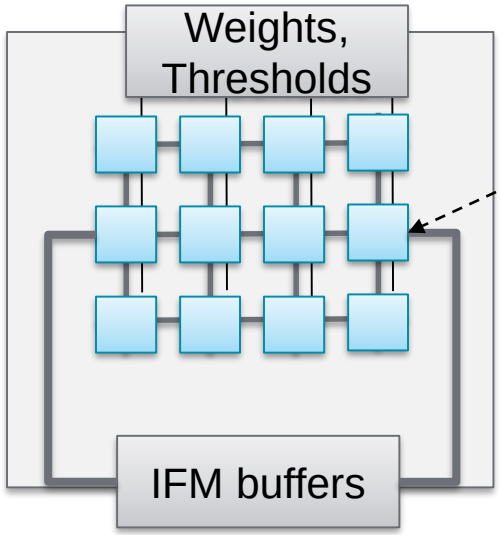
Combined with a broad range of devices

Performance scalability through custom arithmetic

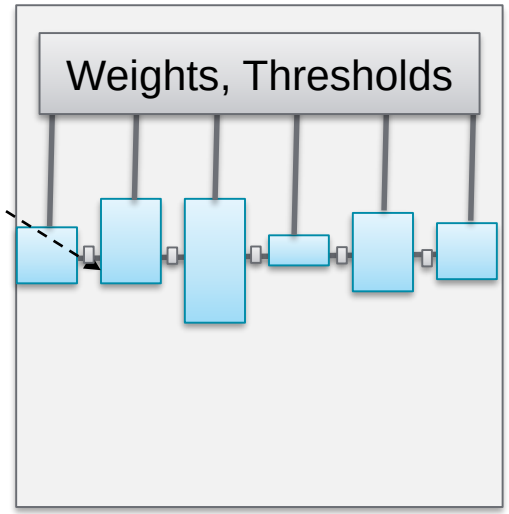




Less resources for very large networks



Customized Operators with PL, DSPs



Dataflow for lower latency

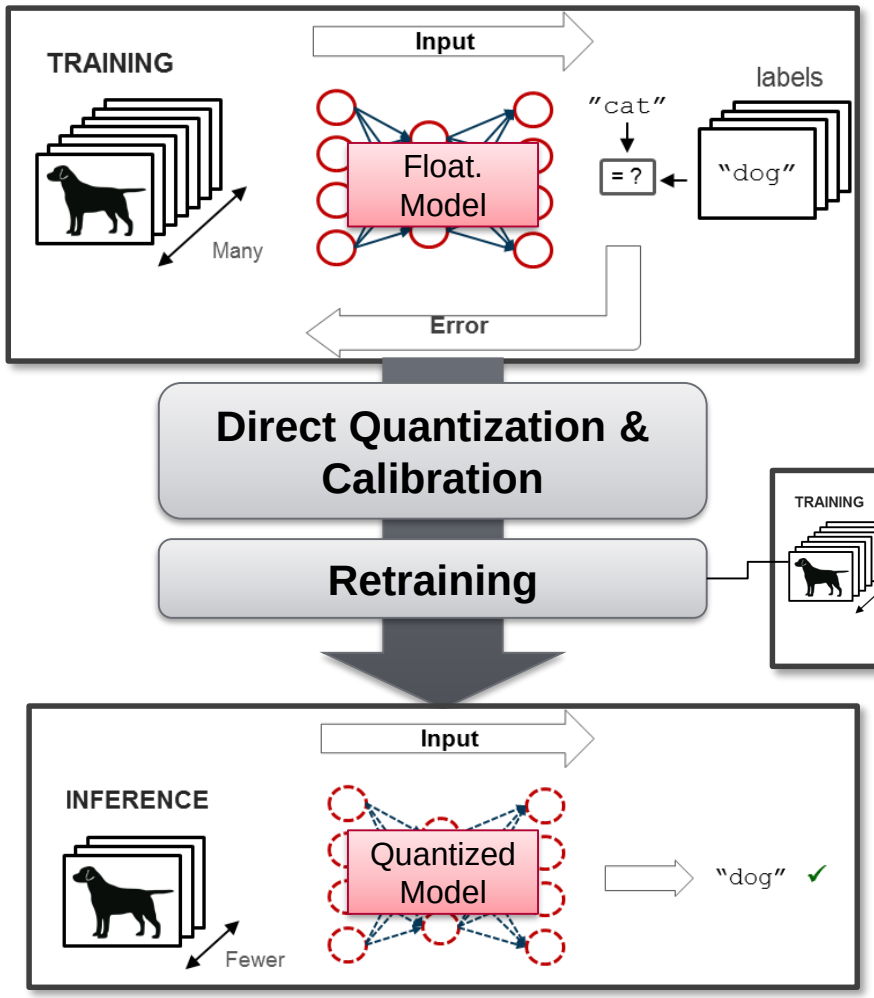
Our Research Effort:

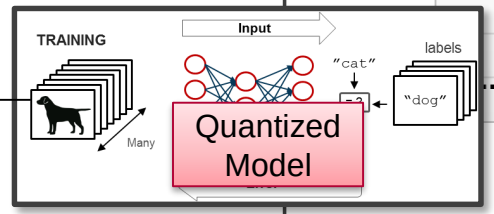
- **Changing neural network algorithm by reducing precision in data types to provide performance scalability, compute efficiency**
  - Numerical representations, precision, quantization
- **Customizing architecture to hit specific design targets**
- **Through automated tool flow (FINN) and open source platforms (PYNQ and AWS) to provide ease of use**

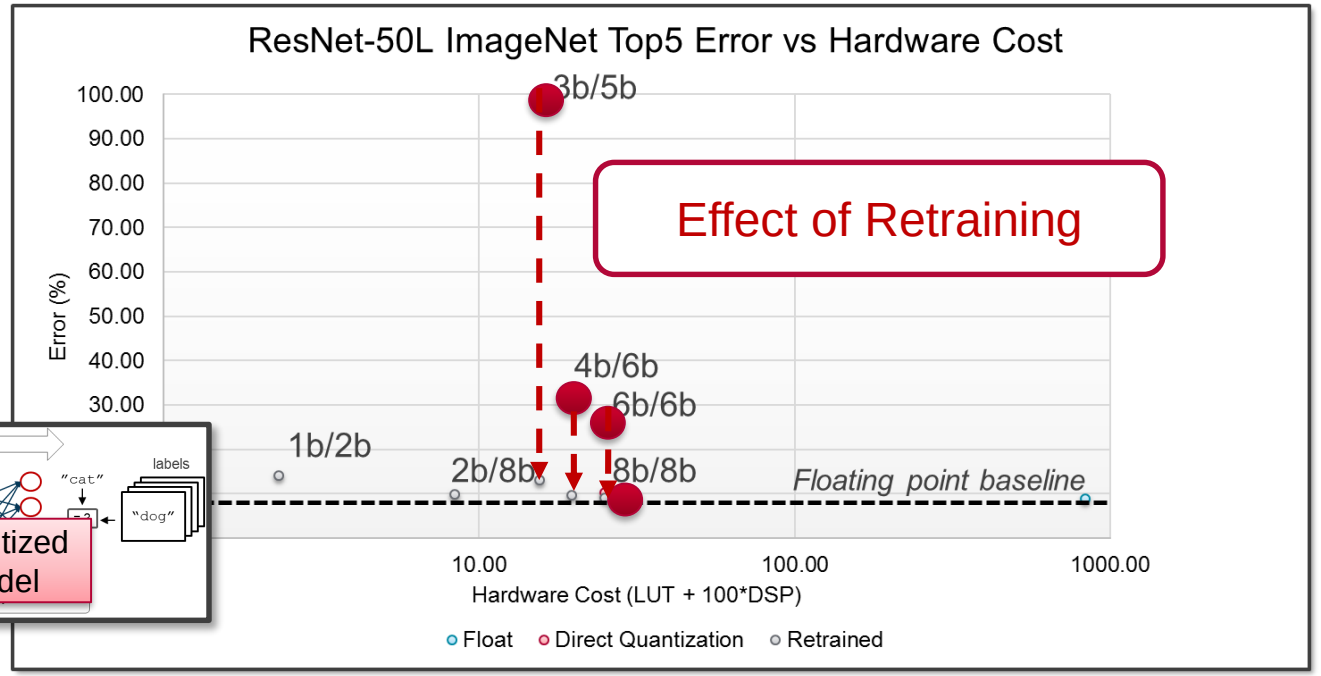
Agenda

- **Background – Xilinx Research**
- **Machine Learning**
- **Research Efforts: Reduced Precision Neural Networks**
- **Summary & Outlook**

From Floating Point to Reduced Precision NNs







- **Direct quantization & calibration**
  - **Deploying a different model to the one we trained**
  - **Works surprisingly well for 8b**
- **<8bit: retraining helps a lot**

Reducing Precision

Scales Performance & Reduces Memory

- **Reducing precision from 32b to 1b shrinks LUT cost**
  - Instantiate **100x** more compute within the same fabric

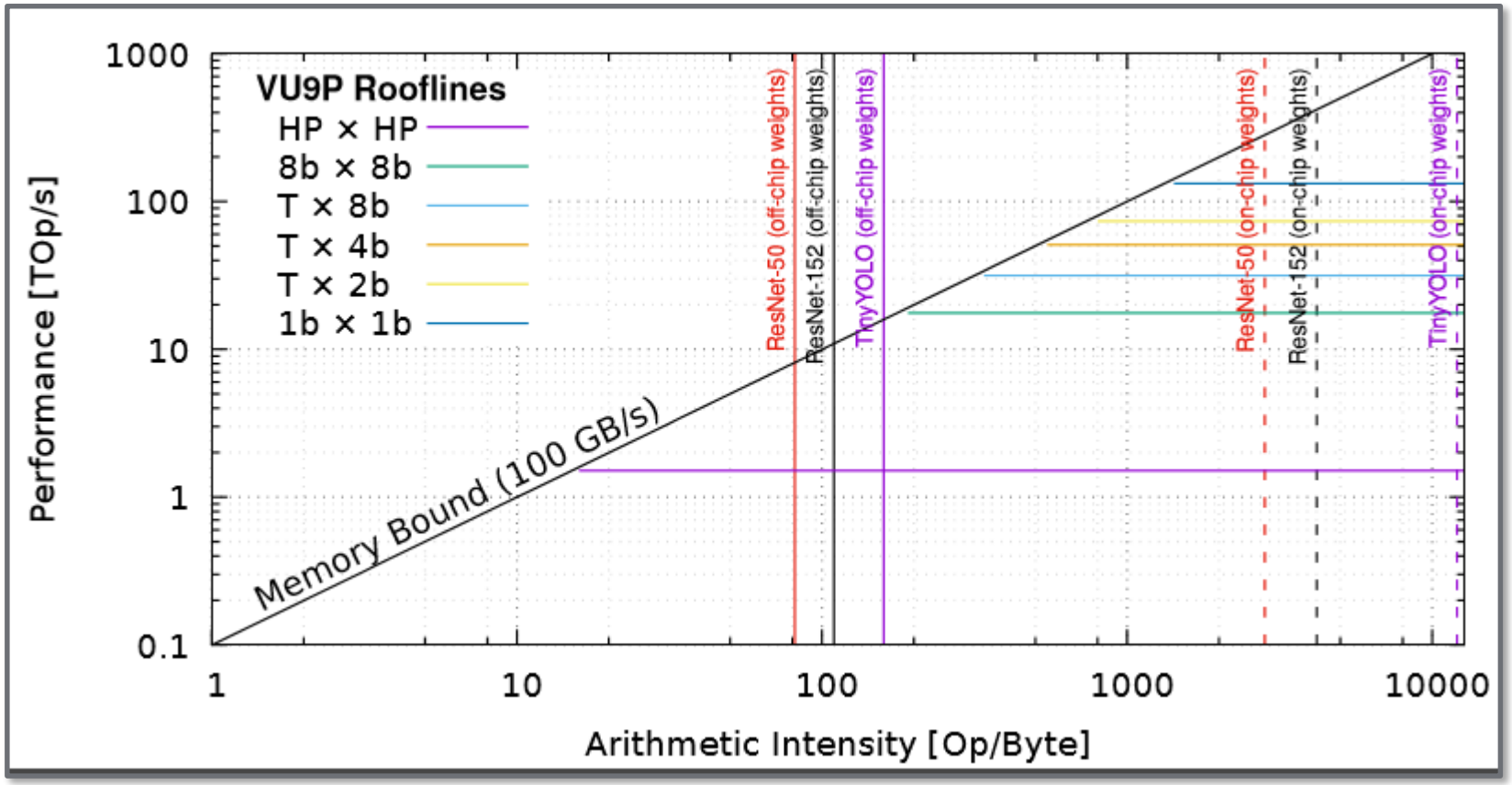
- **Potential to reduce memory footprint**
  - NN model can stay on-chip => no memory bottlenecks

Precision	Cost per Op LUT / DSP	Modelsize [MB] (ResNet50)	TOps/s (VU9P)
1b	2.5 / 0	3.2	~100
8b	45 / 0	25.5	~6
32b	178 / 2	102.5	~1

Assumptions: Application can fill device to 70% (fully parallelizable) 300MHZ
HLS overhead included

Reducing Precision provides Performance Scalability

Example: ResNet50, ResNet152 and TinyYolo



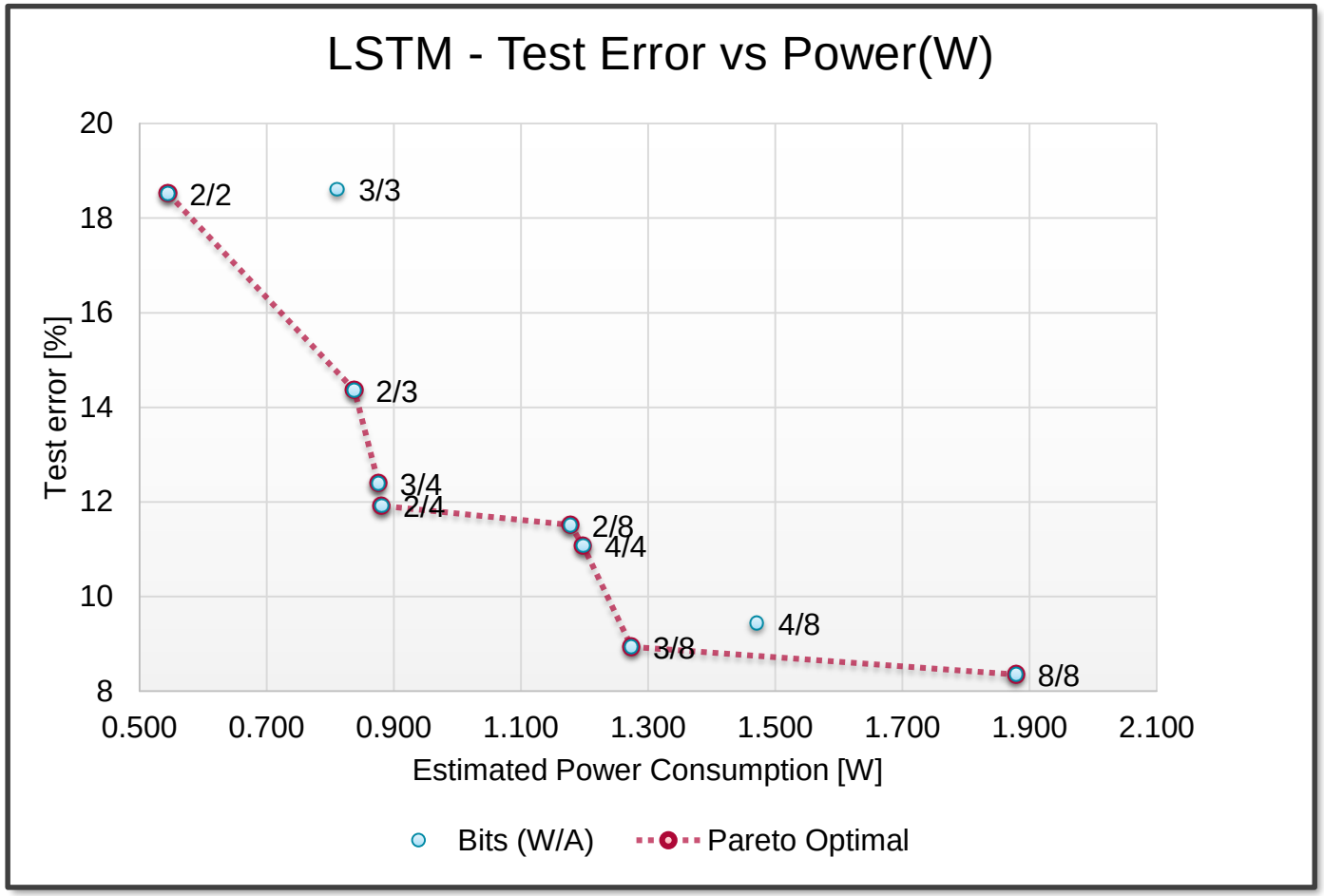
RP scales compute performance

Theoretical Peak Performance for a VU9P with different Precision Operations

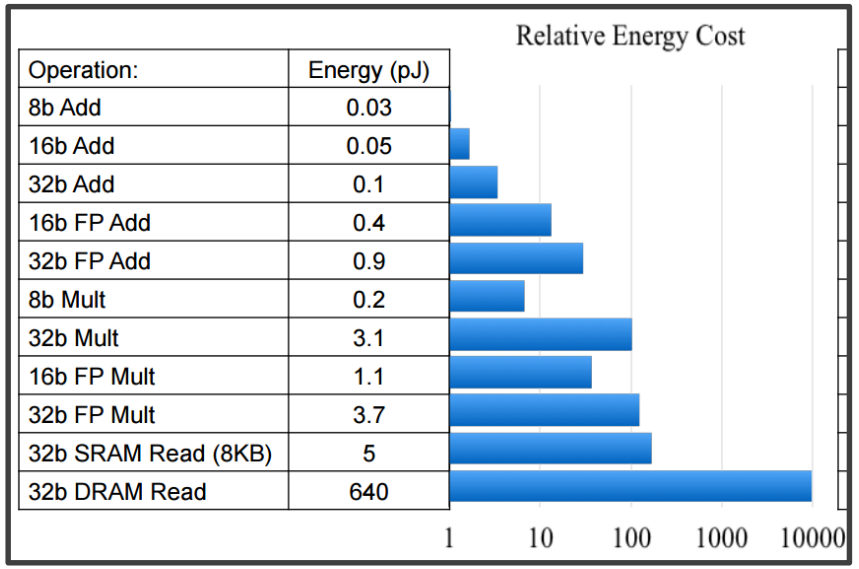
RP reduces model size=> to stay on-chip

Up to 100x

Reduced Precision Inherently Saves Power



Target Device ZU7EV ● Ambient temperature: 25 °C ● 12.5% of toggle rate ● 0.5 of Static Probability ● Power reported for PL accelerated block only

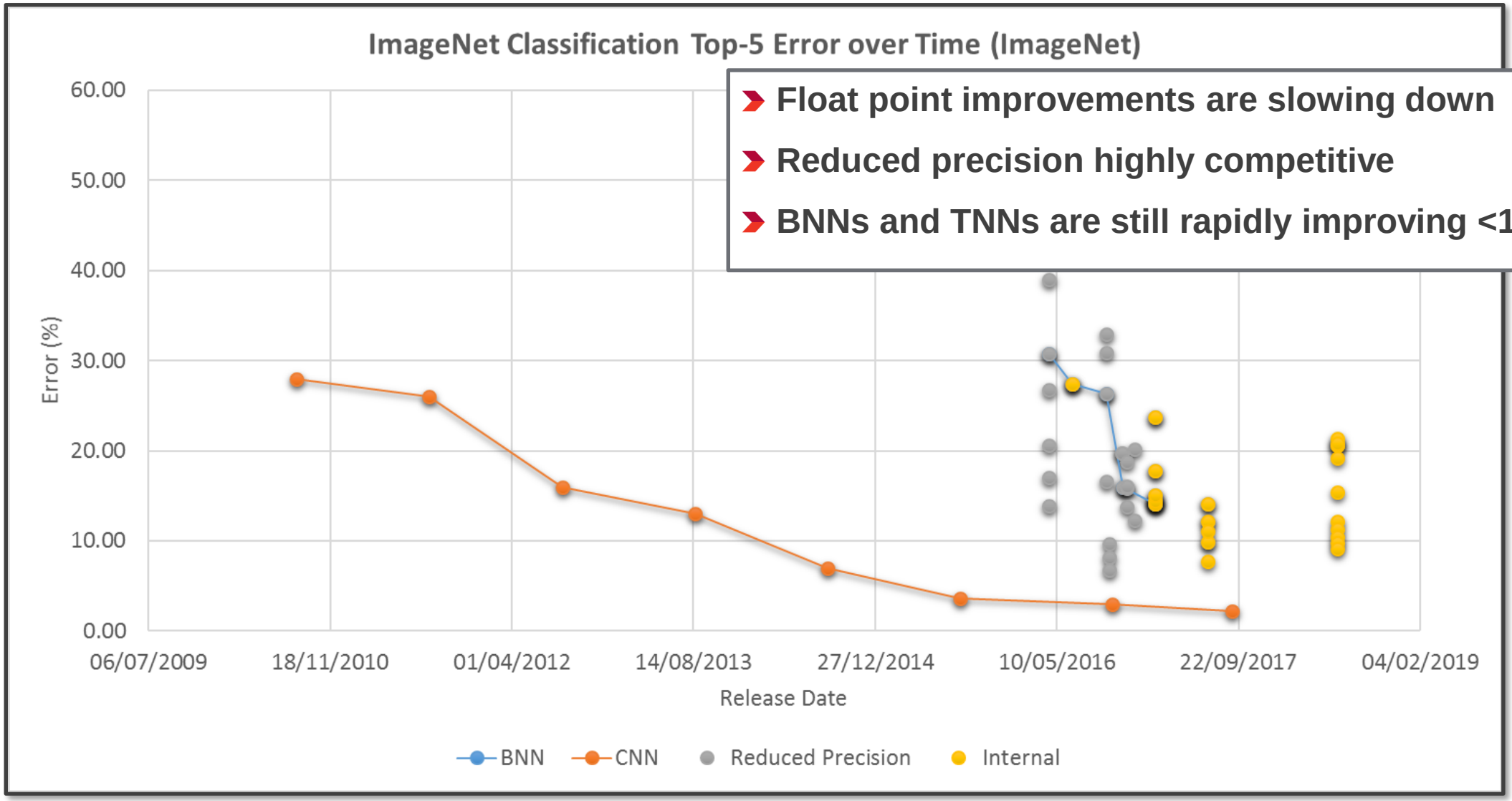


Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FP Mult	1.1
32b FP Mult	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640

*Source: Bill Dally (Stanford), Cadence **Embedded Neural Network Summit, February 1, 2017***

What are the downsides of reduced precision?

RPNNs: Closing the Accuracy Gap

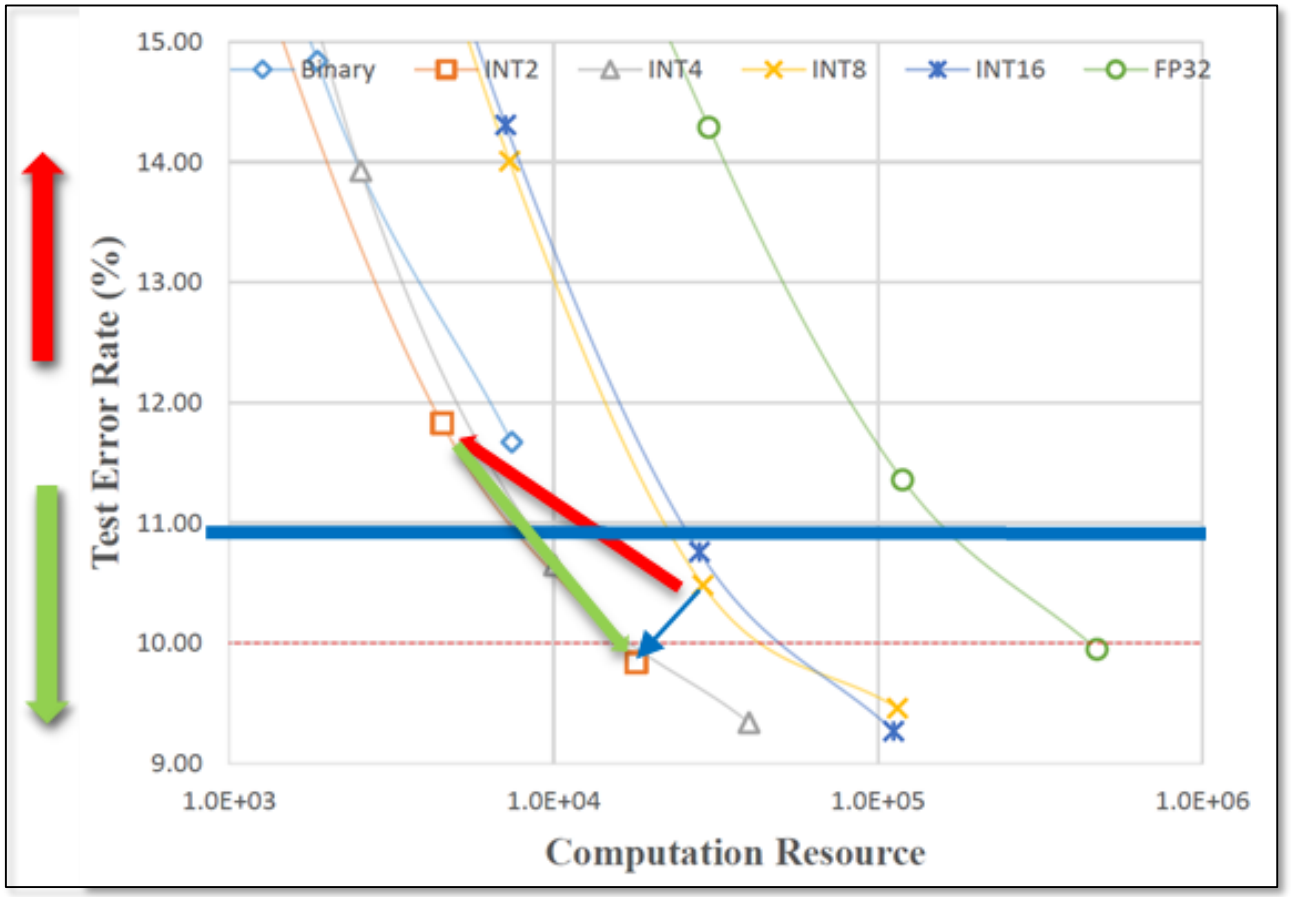


- Float point improvements are slowing down
- Reduced precision highly competitive
- BNNs and TNNs are still rapidly improving <10% top5

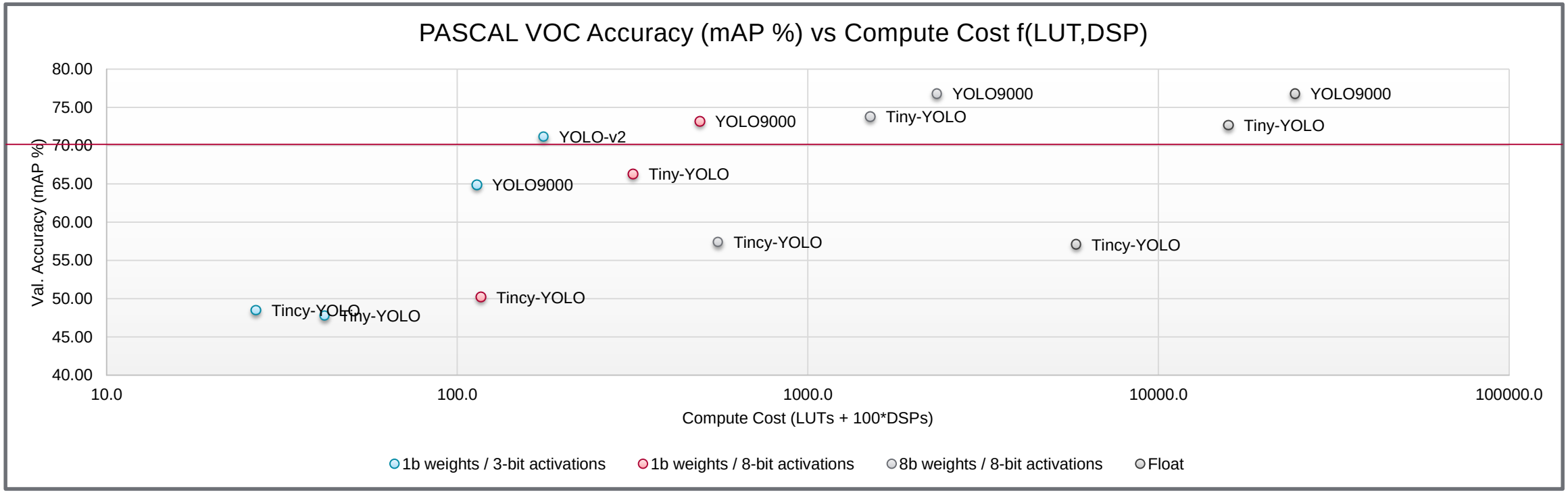
How to recuperate accuracy

Imperial College London

- **We can recuperate accuracy through increasing layer size is possible**
  - Reducing precision, reduces hardware cost & increases error
  - Recuperate accuracy by retraining & increasing network size
- **Check out Alex's talk in Session 1:**
  - "Accuracy to Throughput Trade-offs for Reduced Precision Neural Networks on Reconfigurable Logic"



Topological Changes

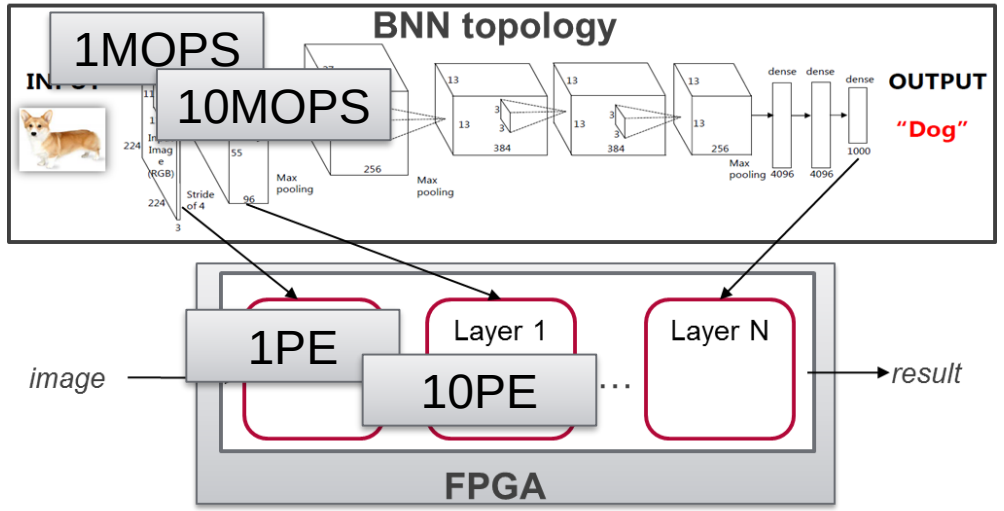


- **RPNNs can achieve floating point accuracy through larger networks at lower hardware cost**

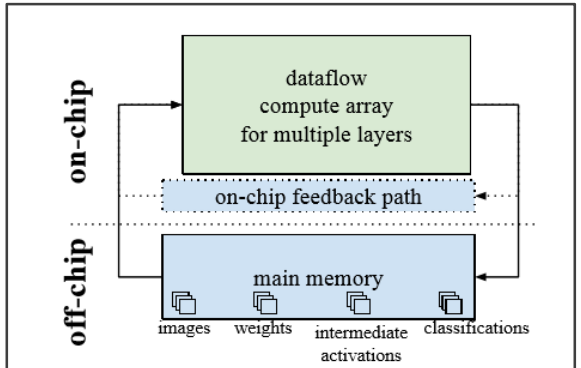
Automating & Customization

FINN: Customized Architectures

- **Custom-tailored hardware**
  - Customized feed-forward dataflow architecture to match network topology
  - Customized to meet design requirements
  - Customized data types (n-bit)



- **If dataflow exceeds target resource**
  - Folding into multi-Layer offload



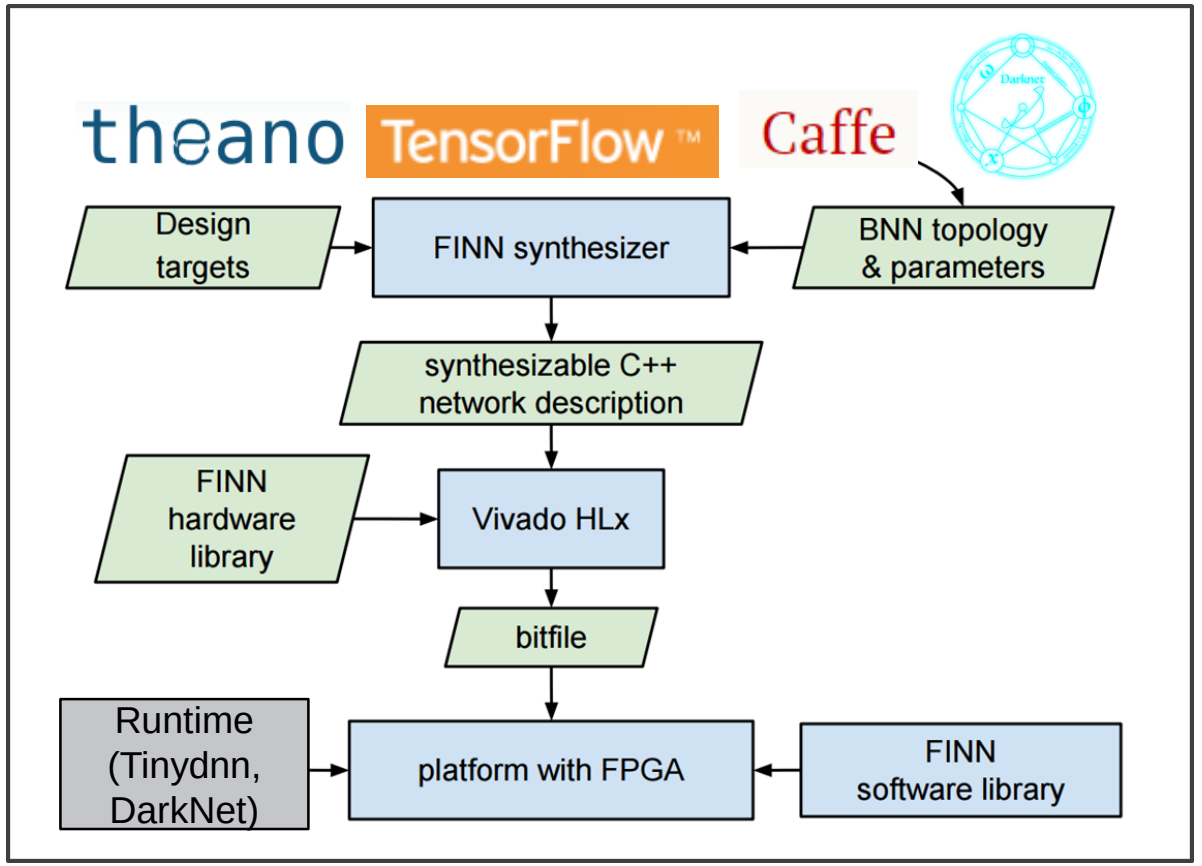
(b) Multilayer Offload Architecture with Maximally-Sized Homogeneous Compute Arrays for Different Precisions

Automatically generated from CNN description

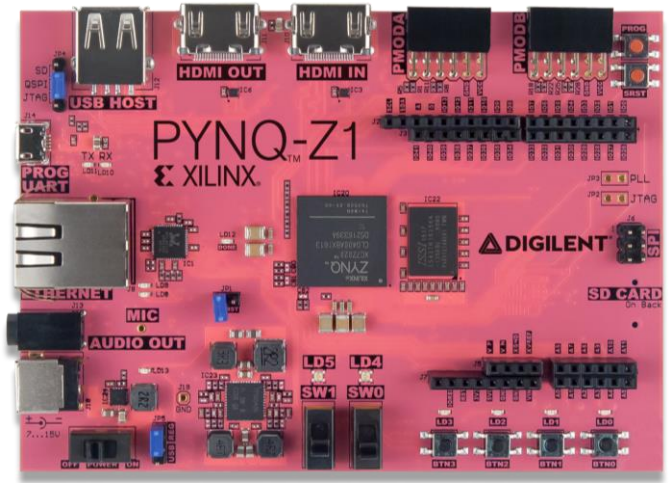
- **Uses a synthesizable C++ NN description**
- **Enables flexibility & scalability and supports portability, rapid exploration**

Synthesizable CNN Description

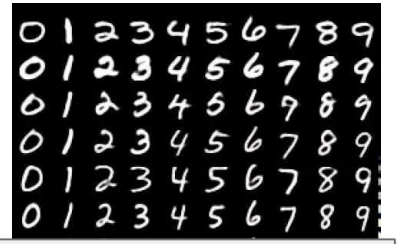
```
void DoCompute(ap_uint<64> * in, ap_uint<64> * out) {
#pragma HLS DATAFLOW
   stream<ap_uint<64> > memInStrm("memInStrm");
   stream<ap_uint<64> > InStrm("InStrm");
          .
          .
          .
   stream<ap_uint<64> > memOutStrm("memOutStrm");

   Mem2Stream<64, inBytesPadded>(in, memInStrm);
   StreamingMatrixVector<L0_SIMD, L0_PE, 16, L0_MW, L0_MH, L0_WMEM, L0_TMEM>
          (InStrm, inter0, weightMem0, thresMem0);
   StreamingMatrixVector<L1_SIMD, L1_PE, 16, L1_MW, L1_MH, L1_WMEM, L1_TMEM>
          (inter0, inter1, weightMem1, thresMem1);
   StreamingMatrixVector<L2_SIMD, L2_PE, 16, L2_MW, L2_MH, L2_WMEM, L2_TMEM>
          (inter1, inter2, weightMem2, thresMem2);
   StreamingMatrixVector<L3_SIMD, L3_PE, 16, L3_MW, L3_MH, L3_WMEM, L3_TMEM>
          (inter2, outstream, weightMem3, thresMem3);
    StreamingCast<ap_uint<16>, ap_uint<64> >(outstream, memOutStrm);
    Stream2Mem<64, outBytesPadded>(memOutStrm, out);
}
```



Numerous Platforms – From Embedded to Cloud

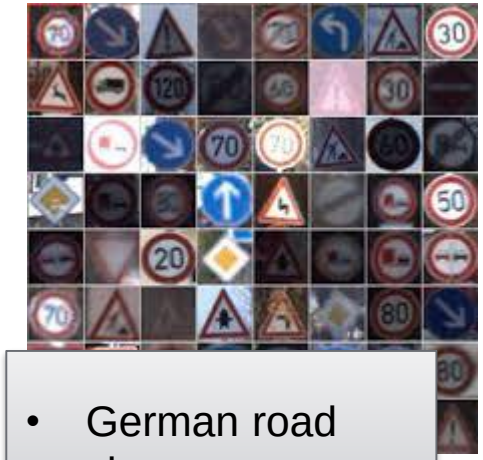


Numerous Datasets

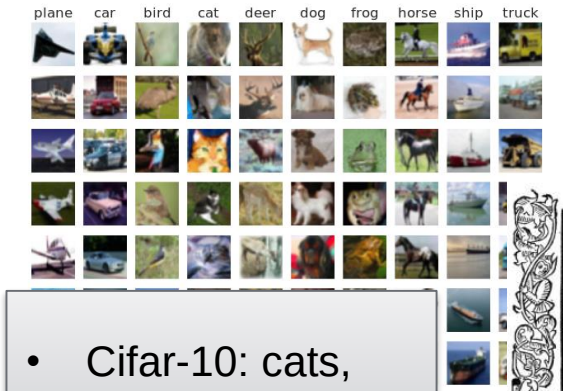


- MNIST handwritten digits

- Streetview house numbers

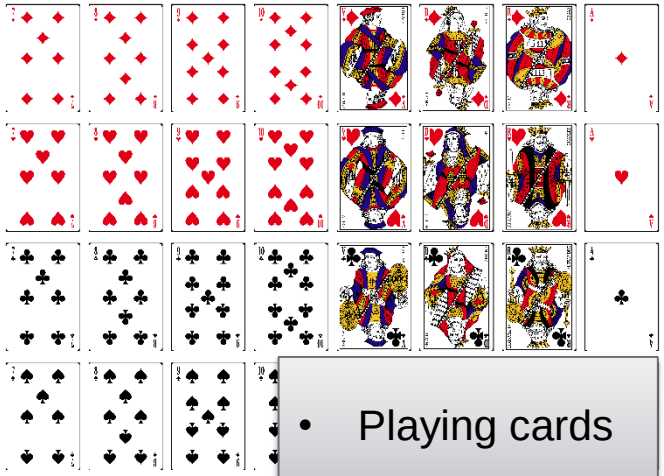


- German road signs

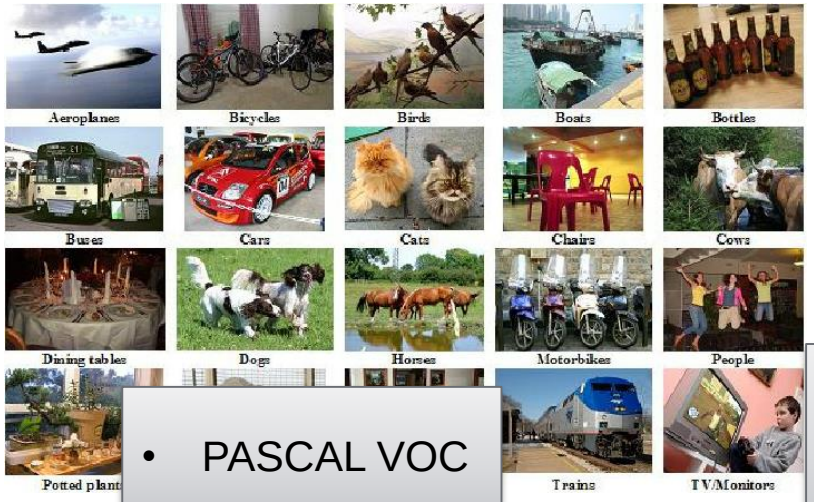


- Cifar-10: cats, dogs, etc

- Fraktur



- Playing cards

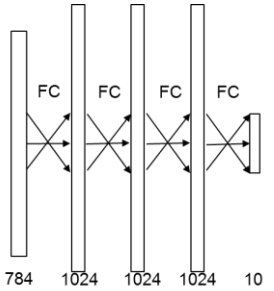


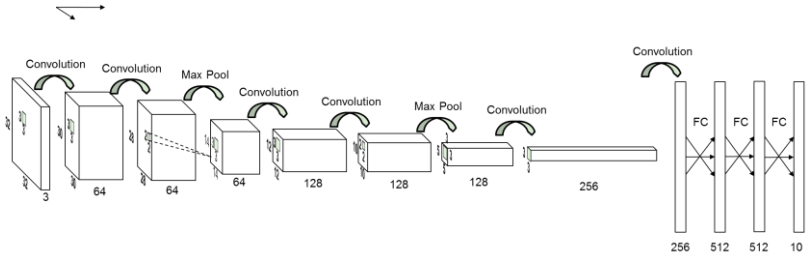
- PASCAL VOC

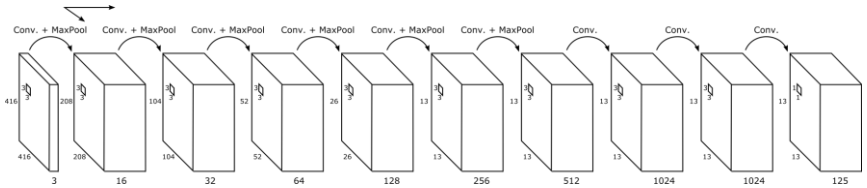
- Imagenet

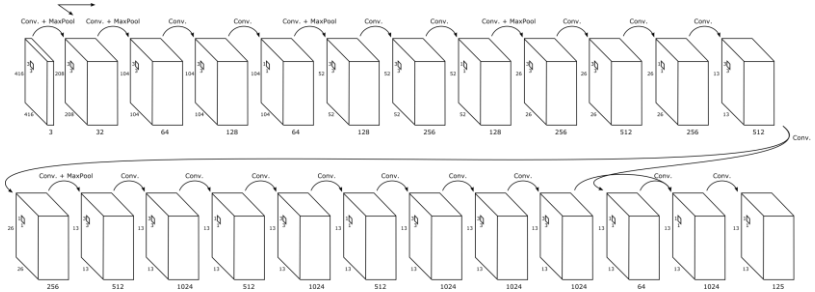
Numerous Test Networks

- **Multilayer Perceptron (1b weights, 1b act), MNIST**
  - Up to 5.8MOPS/frame
- **VGG-16 derivative (1b weights, 1b act), SVHN, CIFAR-10, traffic signs, playing cards)**
  - Up to 1.2GOPS/frame
- **DorefaNet – AlexNet derivative (mostly 1b weights, 2b act) (ImageNet)**
  - Up to 3.9GOPS/frame
- **YoloV2, Yolo9000, TinyYolo (1b weights, 8b act) (VOC, COCO)**
  - 34.9, 19 and 7.0GOPS/frame
- **LSTM, for OCR on Fraktur**









FINN Results

- **Performance**
  - VOC Object recognition: Quantized TinyYolo @ **55fps @ 7Watt** (batch=1) for embedded (ZU3EG)
  - ImageNet Classification: Dorefanet @ **11 TOPS on AWS F1** instance
  - Scaled binary operations to **51TOPS on AWS F1** and **5.2 TOPS on ZU3EG & 1000x over Raspberry Pi**

- **Energy efficiency: measured 433GOPS/Watt**

- **Flexibility & Scalability**
  - Different platforms can easily be targeted from embedded to cloud
  - Different use cases, networks & training data sets

- **While being sufficiently accurate**
  - <10% top5 for ImageNet classification

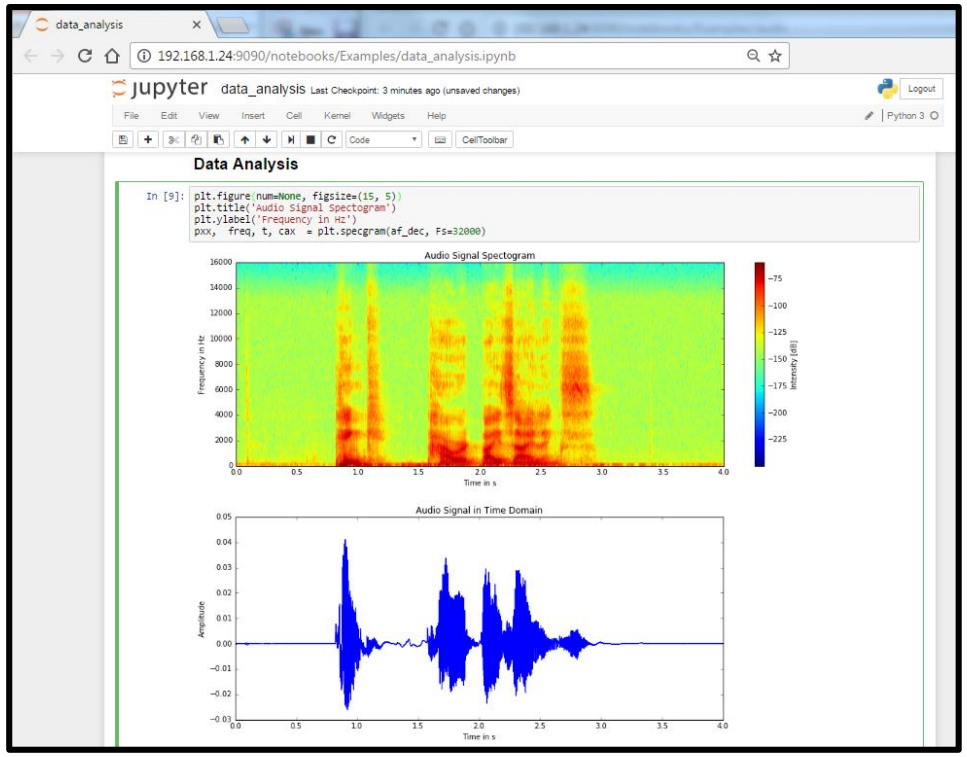
Bridging the Gap to the ML Community…

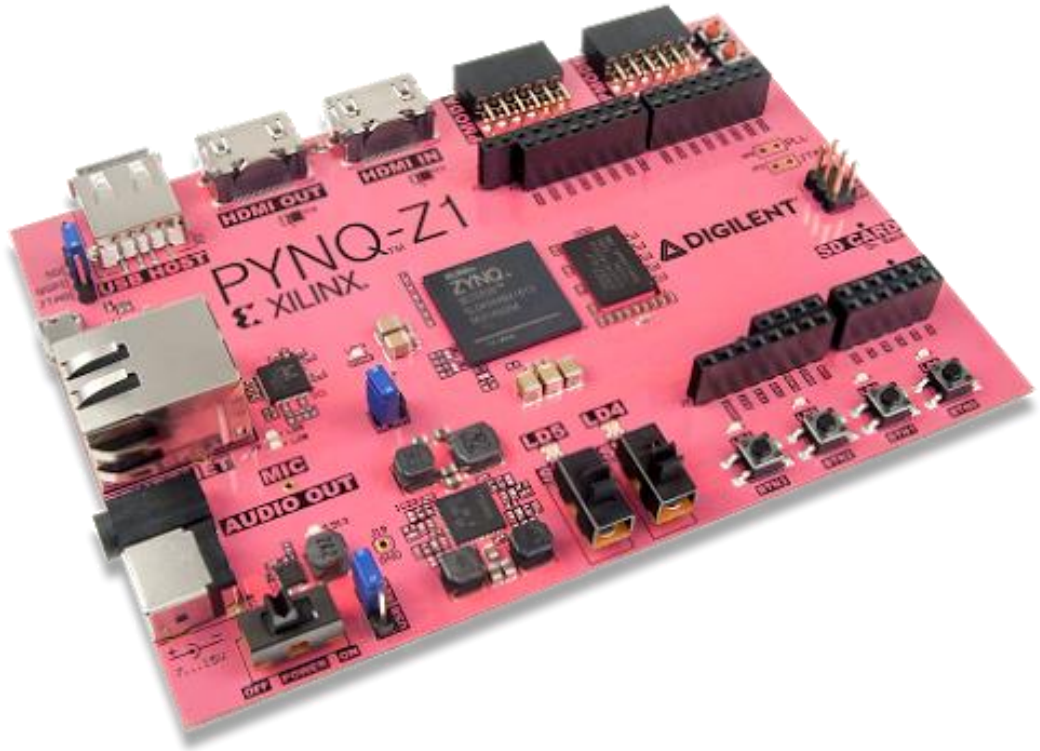
… or how do we address Ease of Use

Python Productivity Kit for Zynq

To enable ML community

PYNQ™





python™

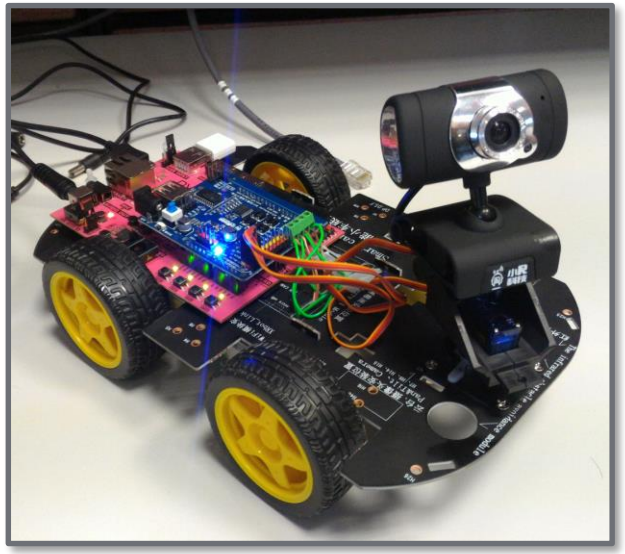
jupyter

Linux

XILINX ALL PROGRAMMABLE™

Build-out of ML Libraries on PYNQ

- **Pynq-bot**
  - Pynq-Z1 robotics kit: https://github.com/Xilinx/PYNQ-BOT



- **Expansion of machine learning library**
  - ZipML: linear regression
  - FINN new networks TincyYolo and DorefaNet
    - https://github.com/Xilinx/BNN-PYNQ
    - https://github.com/Xilinx/QNN-MO-PYNQ
    - https://github.com/Xilinx/FINN
  - LSTM for optical character recognition
    - https://github.com/Xilinx/LSTM-PYNQ

- **Repositories: linked through http://www.pynq.io/**

Agenda

- **Background – Xilinx Research**
- **Machine Learning**
- **Research Efforts**
- **Summary & Outlook**

Summary

- **ML has the potential to address many of the grand engineering challenges of this century**
- **However, compute & memory requirements are huge and flexibility and scalability are key**
- **FPGAs can play an important role here, in particular in conjunction with reduced precision and customized architectures**
  - Orders of magnitude improvement in performance, resources and power consumption
- **Many challenges remain**

Outlook

Ease of Use – bridging the gap to ML community

- **Help build out libraries!**

Accuracy

- Novel training techniques to improve accuracy
- Automating topological changes
- Reduction in training time

Architecture Exploration

- **Help understand the choices!**

Thank You.

Be in touch regarding

FINN, PYNQ and AWS

Publications

- **FPGA 2017: FINN: A Framework for Fast, Scalable Binarized Neural Network Inference**
  - https://arxiv.org/abs/1612.07119
- **PARMA-DITAM 2017: Scaling Binarized Neural Networks on Reconfigurable Logic**
  - https://arxiv.org/abs/1701.03400
- **ICCD 2017: Scaling Neural Network Performance through Customized Hardware Architectures on Reconfigurable Logic**
  - https://ieeexplore.ieee.org/abstract/document/8119246/
- **H2RC 2016: A C++ Library for Rapid Exploration of Binary Neural Networks on Reconfigurable Logic**
  - https://h2rc.cse.sc.edu/2016/papers/paper_25.pdf
- **ICONIP'2017: Compressing Low Precision Deep Neural Networks Using Sparsity-Induced Regularization in Ternary Networks**
  - https://arxiv.org/abs/1709.06262
- **CVPR'2018: SYQ: Learning Symmetric Quantization For Efficient Deep Neural Networks**
- **DATE 2018: Inference of quantized neural networks on heterogeneous all-programmable devices https://ieeexplore.ieee.org/abstract/document/8342121/**
- **ARC'2018: Accuracy Throughput Tradeoffs for Reduced Precision Neural Networks**