CustoNN2: Customizing Neural Networks on FPGAs

High-Performance IT Systems group

Dr. Tobias Kenter
Prof. Dr. Christian Plessl

16 July 2018

Paderborn Center for Parallel Computing

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Neural Network Success Stories

POST MAGAZINE

Why Baidu's breakthrough on speech recognition may be a game changer

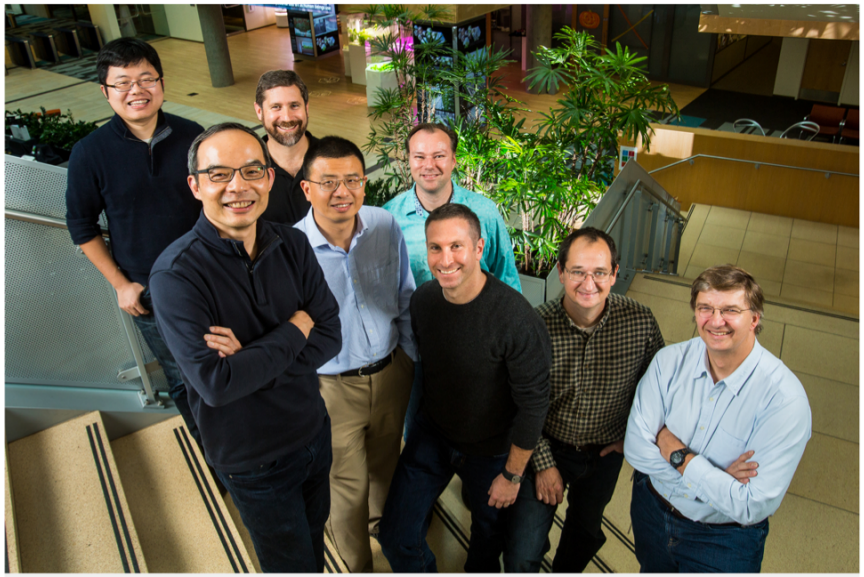
Deep Speech 2, a speech recognition network developed by China's answer to Google, is so stunningly accurate it can transcribe Chinese better than a person, writes Will Knight

BY **MIT TECHNOLOGY REVIEW**
19 MAR 2016

Microsoft Store Products Support

Next The Official Microsoft Blog The Fire Hose Microsoft On the Issues Transform

Historic Achievement: Microsoft researchers reach human parity in conversational speech recognition



Microsoft researchers from the Speech & Dialogue research group include, from back left, Wayne Xiong, Geoffrey Zweig, Xuedong Huang, Dong Yu, Frank Seide, Mike Seltzer, Jasha Droppo and Andreas Stolcke. (Photo by Dan DeLong)

Posted October 18, 2016

By **Allison Linn**

Microsoft has made a major breakthrough in speech recognition, creating a technology that recognizes the words in a conversation as well as a person does.

Neural Network Success Stories

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For

Archive > Volume 529 > Issue 7587 > Articles > Article

ARTICLE PREVIEW

view full access options ▸

NATURE | ARTICLE

日本語要約

Mastering the game of Go with deep neural networks and tree search

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George va
Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc
Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever,
Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel & Demis Hassabis

Affiliations | Contributions | Corresponding authors

Nature **529**, 484–489 (28 January 2016) | doi:10.1038/nature16961
Received 11 November 2015 | Accepted 05 January 2016 | Published online 27 Janu

tastehit

Home Demo Pricing FAQ Blog

Neural Network Success Stories

MIT Technology Review

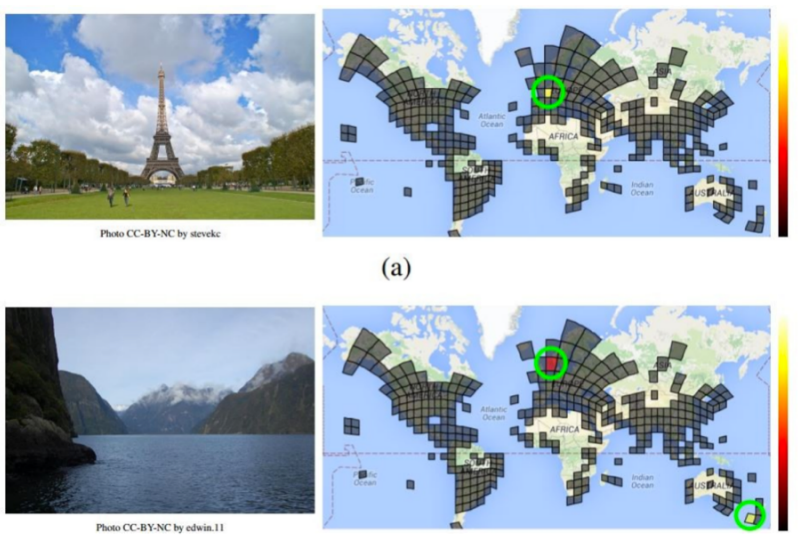
Topics+ Top Stories

Computing

Google Unveils Neural Network with "Superhuman" Ability to Determine the Location of Almost Any Image

Guessing the location of a randomly chosen Street View image is hard, even for well-traveled humans. But Googl latest artificial-intelligence machine manages it with relati ease.

by Emerging Technology from the arXiv February 24, 2016



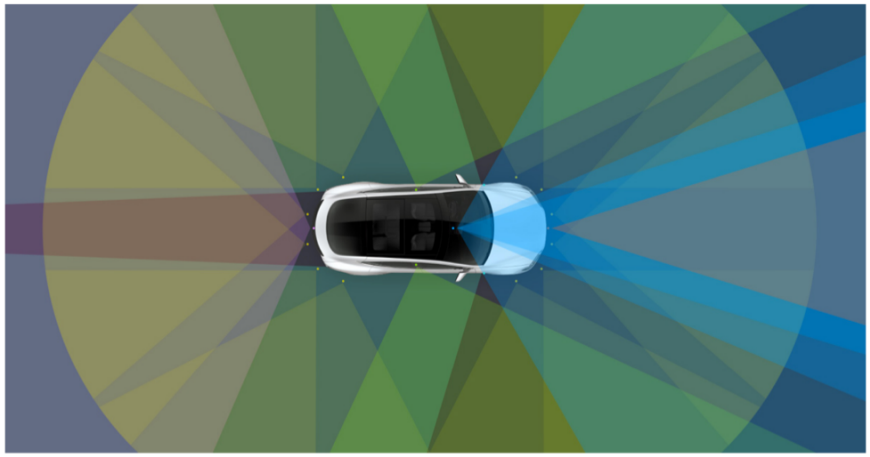
Neural Network Success Stories

TESLA MODEL S MODEL X MODEL 3 ENERGY CHARGING

Blog Videos Press

All Tesla Cars Being Produced Now Have Full Self-Driving Hardware

The Tesla Team • October 19, 2016



Self-driving vehicles will play a crucial role in improving transportation safety and accelerating the world's transition to a sustainable future. Full autonomy will enable a Tesla to be substantially safer than a human driver, lower the financial cost of transportation for those who own a car and provide low-cost on-demand mobility for those who do not.

NVIDIA

DRIVERS PRODUCTS DEEP LEARNING AND AI COMMUNITIES SUPPORT SHOP ABOUT NVIDIA

HOME DEEP LEARNING VIRTUAL REALITY DRIVING PRO GRAPHICS GAMING

Posted on OCTOBER 20, 2016 by DANNY SHAPIRO

Tesla Motors' Self-Driving Car "Supercomputer" Powered by NVIDIA DRIVE PX 2 Technology

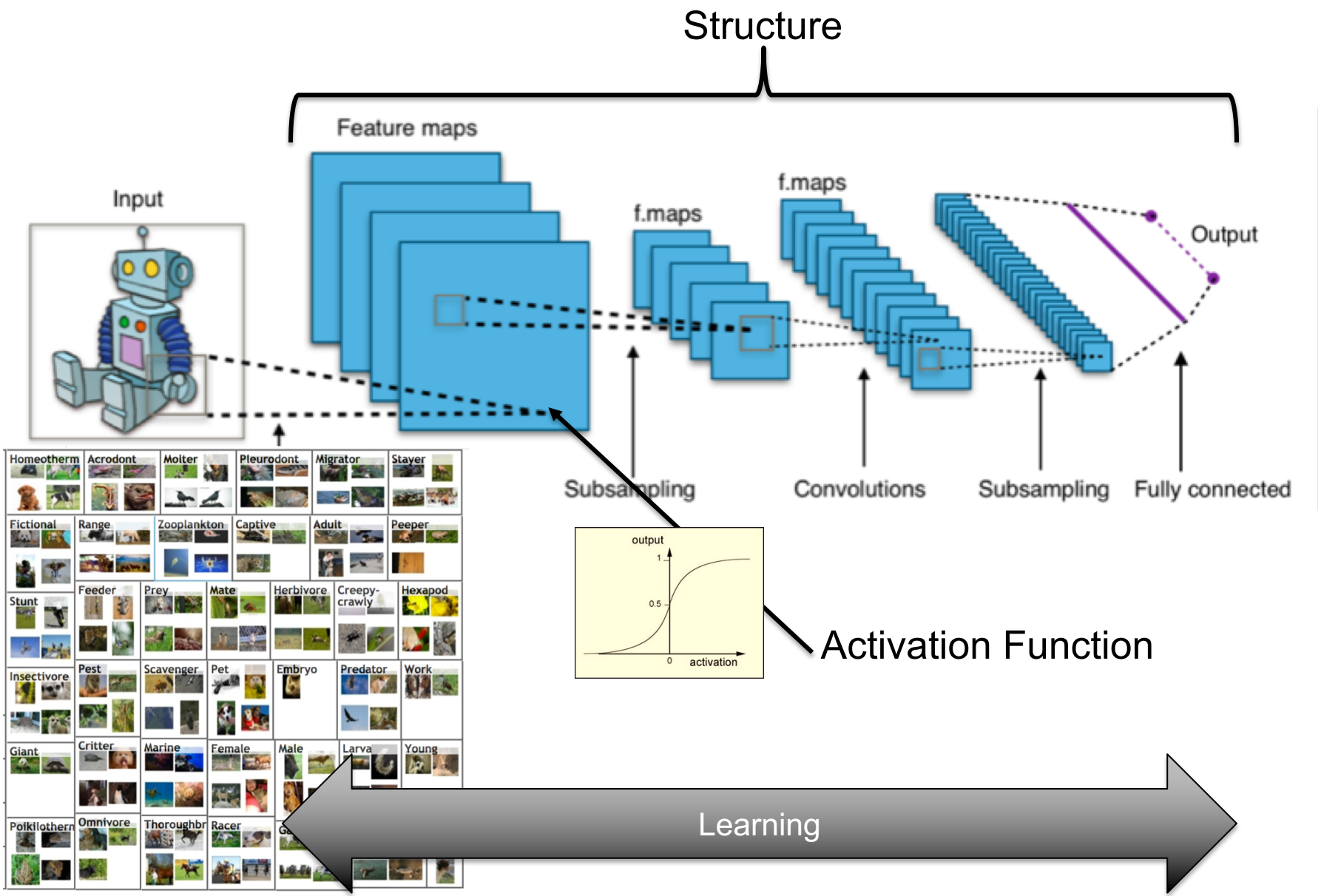
Tesla Motors has announced that all Tesla vehicles — Model S, Model X, and the upcoming Model 3 — will now be equipped with an on-board "supercomputer" that can provide full self-driving capability.

The computer delivers more than 40 times the processing power of the previous system. It runs a Tesla-developed neural net for vision, sonar, and radar processing.

This in-vehicle supercomputer is powered by the NVIDIA DRIVE PX 2 AI computing platform.

NVIDIA DRIVE PX 2 is an end-to-end AI computing system that uses groundbreaking approaches in deep learning to perceive and understand the car's surroundings.

Designing CNNs

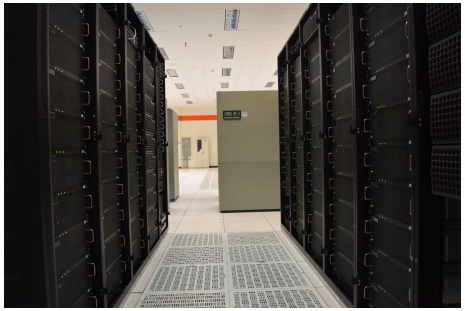


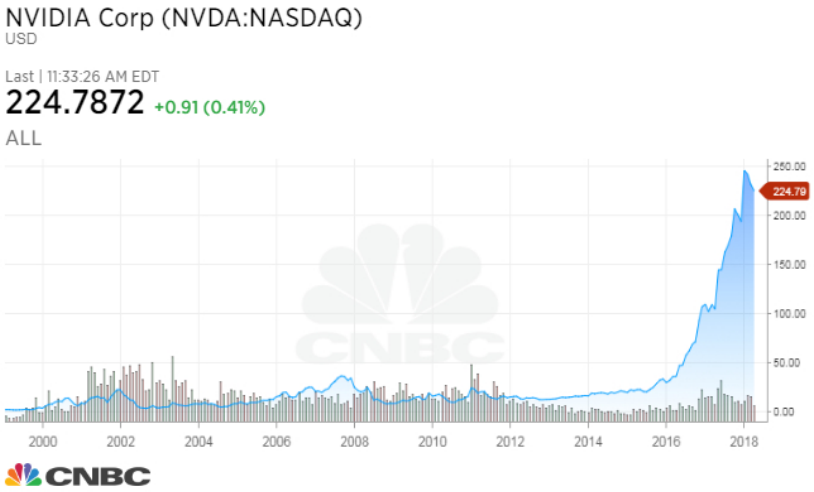
[Sources: ImageNet Database (http://www.image-net.org/), Wikimedia Commons]

CNN's Hardware Demands

- AlexNet [2012]
  - 650,000 neurons
  - 60 million parameters (249 MB)
  - 1.5 billion floating point operations to classify one image
  - Training: 5-6 days on 2 GTX 580 GPUs
- AlphaGo
  - Training: > 4 weeks on 50 GPUs

➢ Massive GPU clusters

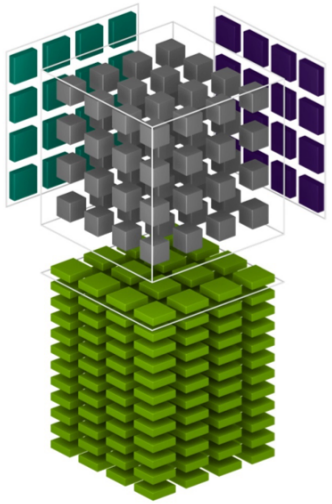




[Krizhevsky, A., Sutskever, I., Hinton, G.E.: **Imagenet classification with deep convolutional neural networks**. In: *Advances in neural information processing systems*, pp. 1097–1105 (2012)]

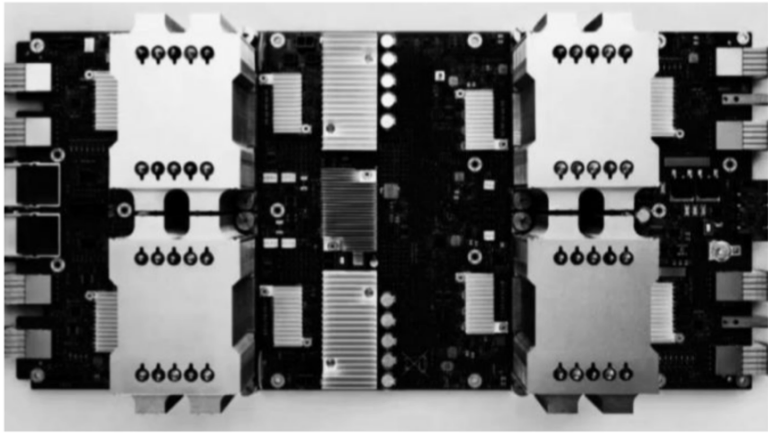
Special hardware for CNNs

- Specialized hardware for CNNs
  - Nvidia Tensor Cores
  - Google TPUs (3rd generation 2018)

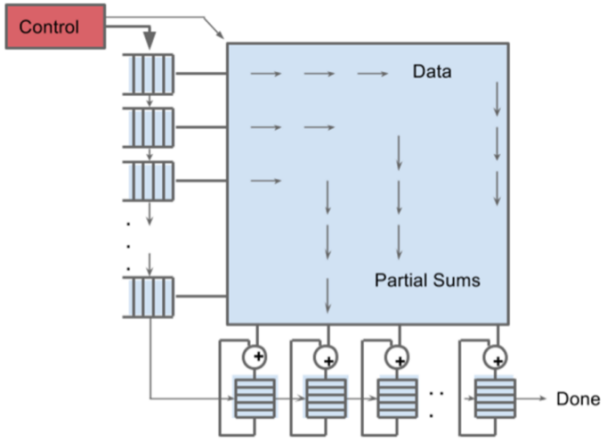


TEARING APART GOOGLE'S TPU 3.0 AI COPROCESSOR

May 10, 2018 Paul Teich

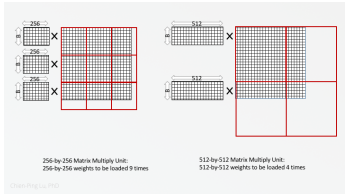


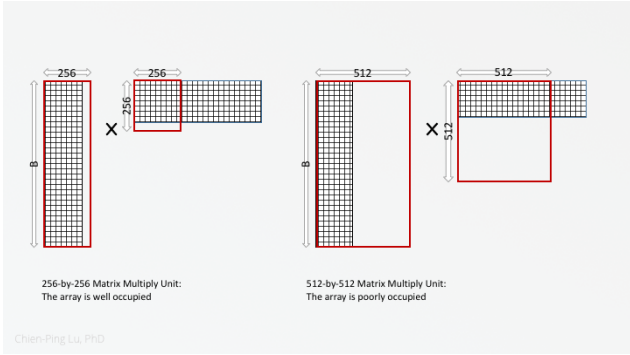
Google did its best to impress this week at its annual IO conference. While Google rolled out a bunch of benchmarks that were run on its current Cloud TPU instances, based on TPUv2 chips, the company divulged a few skimpy details about its next generation TPU



FPGAs for CNNs

- Configurable Hardware
  - Customize operations, connections, data reuse
  - One tensor format does not fit all topologies





  - There is active debate on best data formats for CNNs

Binarized Neural Networks: Training Neural Networks with Weights and Activations Constrained to $+1$ or -1

Matthieu Courbariaux[*1]
Itay Hubara[*2]
Daniel Soudry[3]
Ran El-Yaniv[2]
Yoshua Bengio[1,4]
[1]Université de Montréal
[2]Technion - Israel Institute of Technology
[3]Columbia University
[4]CIFAR Senior Fellow
*Indicates equal contribution. Ordering determined by coin

Abstract

We introduce a method to train Binarized Neural Networks (BNNs) - neural networks with binary weights and activations at run-time. At training-time the binary weights and activations are used for computing the parameters gradients. During the forward pass, BNNs drastically

2018 IEEE 26th Annual International Symposium on Field-Programmable Custom Computing Machines

ReBNet: Residual Binarized Neural Network

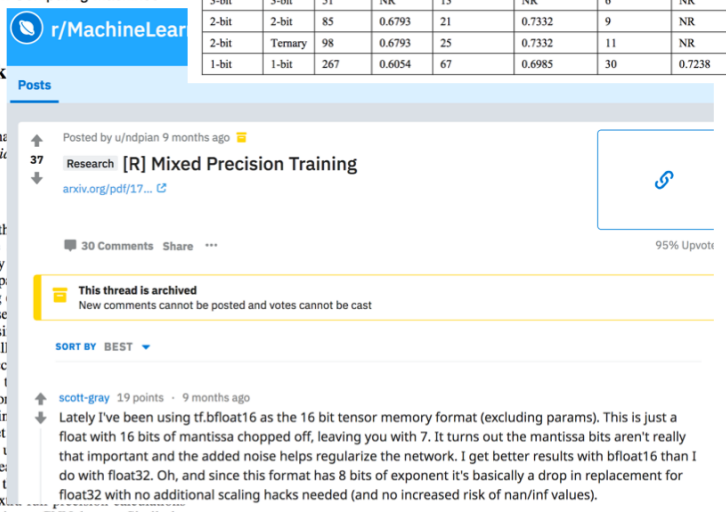
Mohammad Ghasemzadeh, Mohammad Samragh, and Farinaz Kousha
Department of Electrical and Computer Engineering, University of California
{*mghasemzadeh, msamragh, farinaz*}*@ucsd.edu*

***Abstract*—This paper proposes ReBNet, an end-to-end framework for training reconfigurable binary neural networks on software and developing efficient accelerators for execution on FPGA. Binary neural networks offer an intriguing opportunity for deploying large-scale deep learning models on resource-constrained devices. Binarization reduces the memory footprint and replaces the power-hungry matrix-multiplication with light-weight XnorPopcount operations. However, binary networks suffer from a degraded accuracy compared to their fixed-point counterparts. We show that the state-of-the-art methods for optimizing binary networks accuracy, significantly increase the implementation cost and complexity. To compensate for the degraded accuracy while adhering to the simplicity of binary networks, we devise the first reconfigurable scheme that can adjust the classification accuracy based on the application. Our proposition improves the classification accuracy by representing features with *multiple* levels of residual binarization. Unlike previous methods, our approach does not exacerbate the area cost of the hardware accelerator. Instead, it provides a tradeoff between throughput and accuracy while the area overhead of multi-level binarization is negligible.**

[17], significantly reducing th
sumption. Consequent to the
dataflow architecture of binary
than their fixed-point counterp
There are several remaining
CNNs. First, the training phase
often slow and the final classi
lower than the model with full
been shown that the loss of acc
by training binary networks
Nevertheless, this method fo
minishes the performance gain
effort, authors of XNOR-net
accuracy of the binary CNN
computed by averaging the fe
method, however, sacrifices
CNN accelerator by adding ex
to the computation flow of binary CNN layers. Similarly,

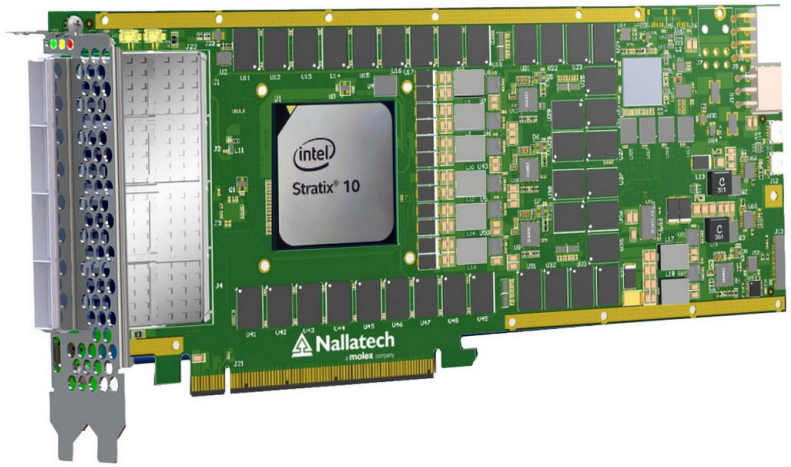
TABLE IV. THROUGHPUT AND ACCURACY FOR VARIOUS PE CONFIGURATIONS ON RESNET TOPOLOGIES

		ResNet-34 1x Wide		ResNet-34 2x Wide		ResNet-34 3x Wide		ResNet-50	
Activation	*Weight*	*Eq TOPS*	*Top-1 Acc*	*Eq TOPS*	*Top-1 Acc*	*Eq TOPs*	*Top-1 Acc*	*Eq TOPS*	*Top-1 Acc*
FP32	FP32	7	0.7359	NR	NR	NR	NR	7	0.7622
8-bit	8-bit	8	0.7093	2	NR	1	NR	8	0.7243
8-bit	Ternary	43	0.6919	11	NR	5	NR	43	0.7038
8-bit	Binary	52	NR	13	NR	6	NR	52	NR
4-bit	4-bit	18	0.7033	5	0.7453	2	NR	18	0.7188
3-bit	3-bit	51	NR	13	NR	6	NR	51	NR
2-bit	2-bit	85	0.6793	21	0.7332	9	NR	85	NR
2-bit	Ternary	98	0.6793	25	0.7332	11	NR	98	NR
1-bit	1-bit	267	0.6054	67	0.6985	30	0.7238	267	0.6263



Our Infrastructure

- Noctua cluster
  - Right now build up by Cray in PC²
  - Includes 32 latest generation Intel Stratix 10 FPGAs
    - Direct FPGA-to-FPGA connections
  - Plus two smaller experimental clusters
    - Xilinx FPGAs
    - Intel CPU+FPGA prototypes

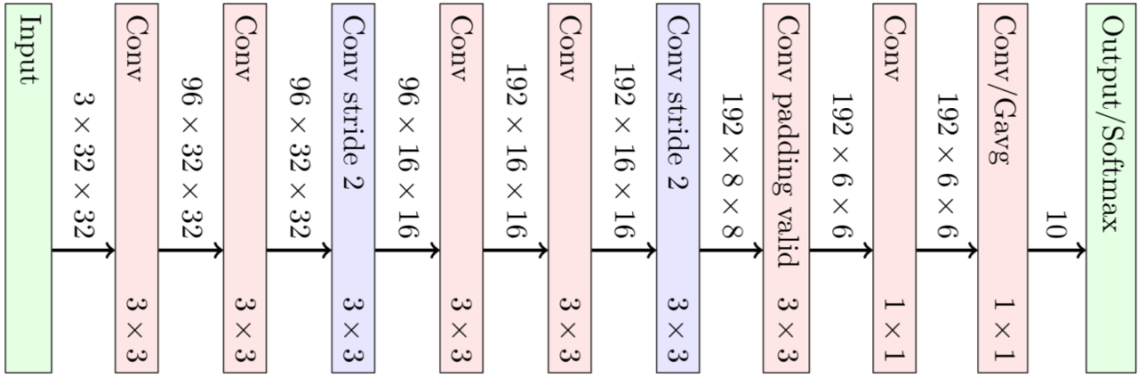


- Programming FPGAs with OpenCL

```
22  // AOC kernel demonstrating device-side printf call
23
24  __kernel void hello_world(int thread_id_from_which_to_print_message) {
25    // Get index of the work item
26    unsigned thread_id = get_global_id(0);
27
28    if(thread_id == thread_id_from_which_to_print_message) {
29      printf("Thread #%u: Hello from Altera's OpenCL Compiler!\n", thread_id);
30    }
31  }
32
```

Results of first PG CustoNN

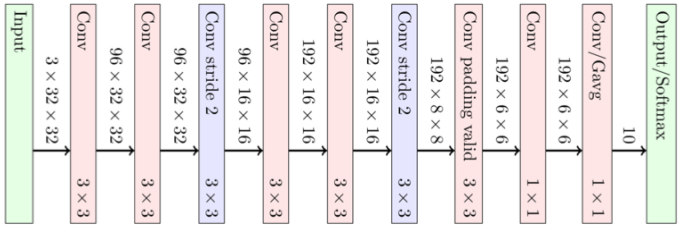
- OpenCL-based FPGA design for Xilinx and Intel FPGAs
- Custom topology keeping all weights in local FPGA memory

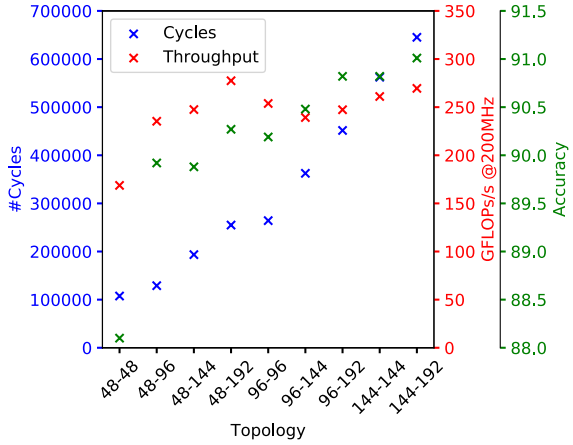


- Trained with Tensorflow for image recognition
- Up to 200 GFLOPs performance (Xilinx FPGA)
- Research paper submitted, in preparation for resubmission
- Functional, but inefficient fixed-point implementation

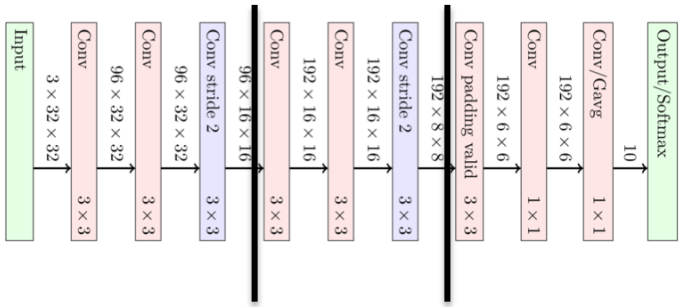
Goals PG CustoNN2

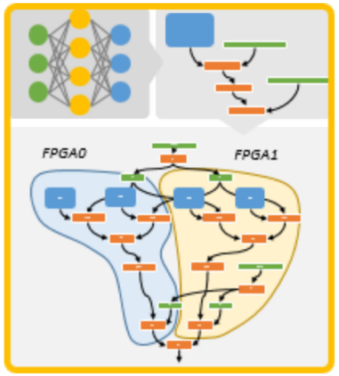
- Codesign of topology and hardware
  - Weights in DDR memory – support for more state-of-the-arte CNNs

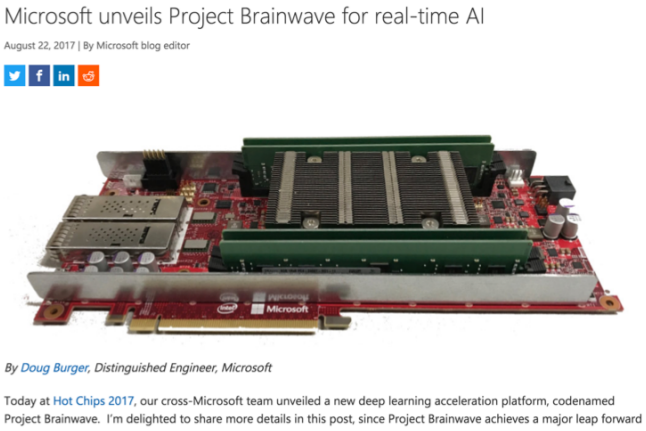




- Catch up on custom data types
  - e.g. bfloat16, libraries for efficient fixed point
- Exploit capabilities of new FPGAs
  - 3-4x more resources per FPGA
  - Stretch calculations over several FPGAs






Microsoft unveils Project Brainwave for real-time AI

August 22, 2017 | By Microsoft blog editor

By Doug Burger, Distinguished Engineer, Microsoft

Today at Hot Chips 2017, our cross-Microsoft team unveiled a new deep learning acceleration platform, codenamed Project Brainwave. I'm delighted to share more details in this post, since Project Brainwave achieves a major leap forward

CustoNN - Summary

- **Project Group for CS and CE students**

- Goals
  - ➢ Codesign of CNN topology and FPGA implementation
  - ➢ Realize custom data types via OpenCL or HLS
  - ➢ Brainwave-like CNN stretching over 32 FPGAs

- Fields of interest
  - Neural networks / deep learning
  - OpenCL or other accelerator languages
  - Accelerator architectures

- Supervisors
  - Christian Plessl, christian.plessl (at) uni-paderborn.de
  - Tobias Kenter, kenter (at) uni-paderborn.de, ☎ 05251/60-4340