FPGA-based Computing in the Era of Artificial Intelligence and Big Data

Speaker: Eriko Nurvitadhi (Intel Labs)

Acknowledgements: Mahesh Iyer, Aravind Dasu (Intel PSG)

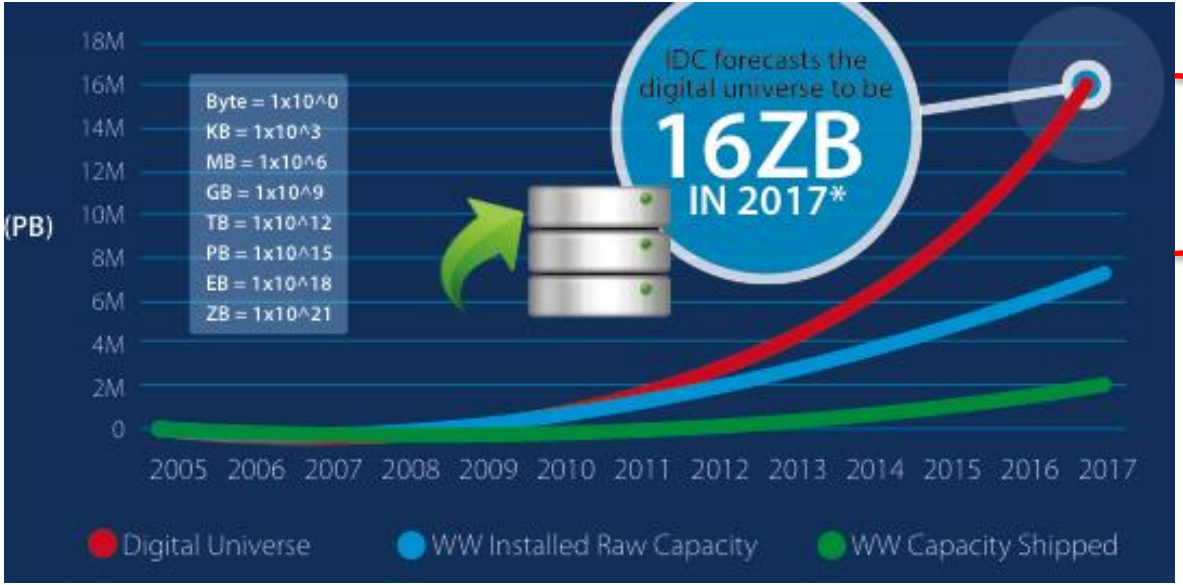
Talk Outline

Part 1: Trends in AI and Big Data

Part 2: Trends and opportunities for FPGA

Part 3: Research highlights

Amount of Available Data Grows Rapidly



No choice but to throw data away!?

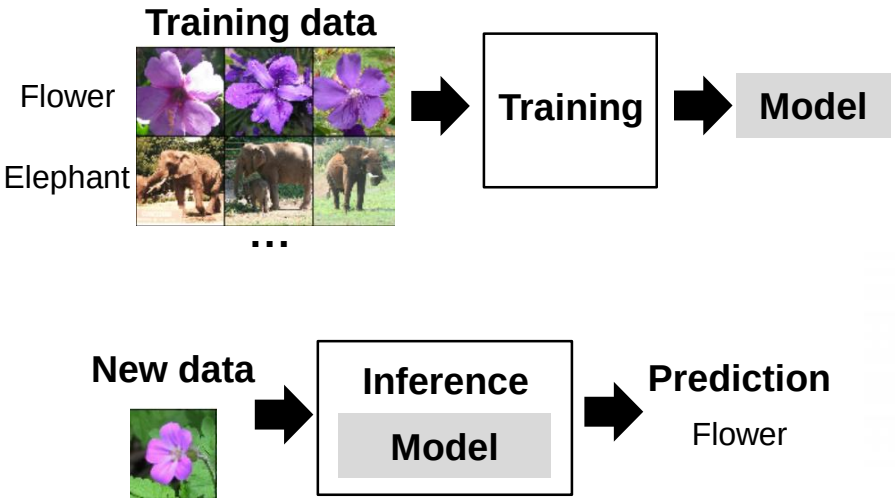
(Imagine how many pictures of cats and babies we would miss…)

Source: L. Ceze, ISAT Summer'15

How to take advantage of such abundance of data?

AI can Extract Insights from Abundance of Data

Let's look at deep neural networks (DNNs), a very popular branch of AI



DESIGNLINES | INDUSTRIAL CONTROL DESIGNLINE

Microsoft, Google Beat Humans at Image Recognition

Deep learning algorithms compete at ImageNet challenge

By R. Colin Johnson, 02.18.15 14

Share Post | Share on Facebook | Share on Twitter | G+ | in

PORTLAND, Ore. -- First computers beat the best of us at chess, then poker, and finally Jeopardy. The next

ACHIEVING HUMAN PARITY IN CONVERSATIONAL SPEECH RECOGNITION

W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu and G. Zweig

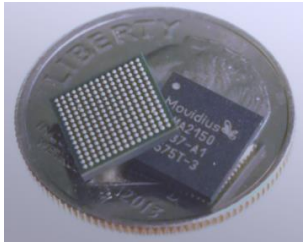
Microsoft Research
Technical Report MSR-TR-2016-71

AI revolution is happening - let's look at how it impacts computing

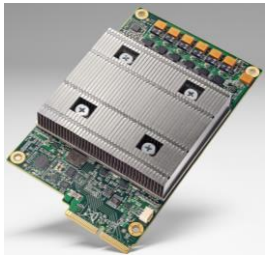
AI is Changing Hardware and Software Services

Hardware for AI **now available**

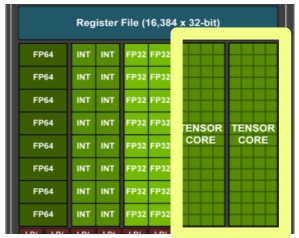
Intel Movidius



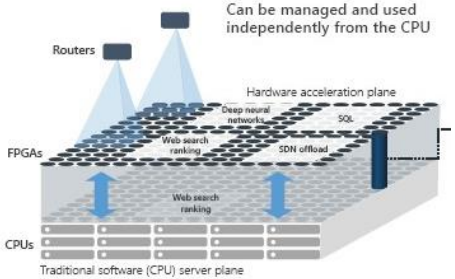
Google TPUs



Nvidia Volta GPU+TensorCore

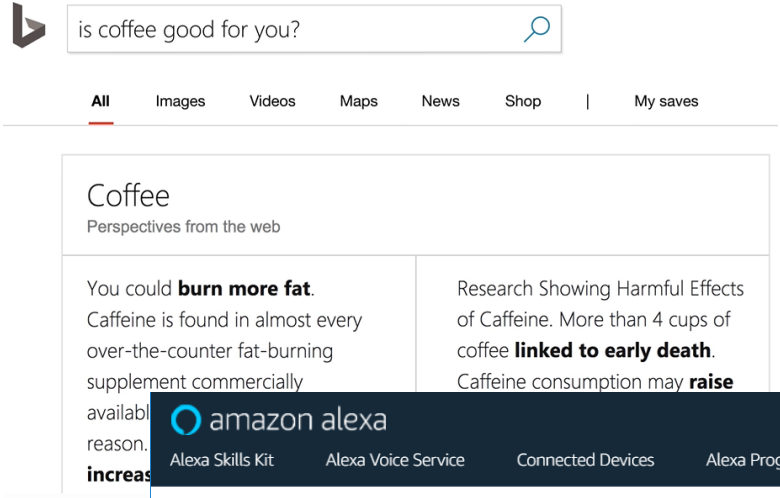


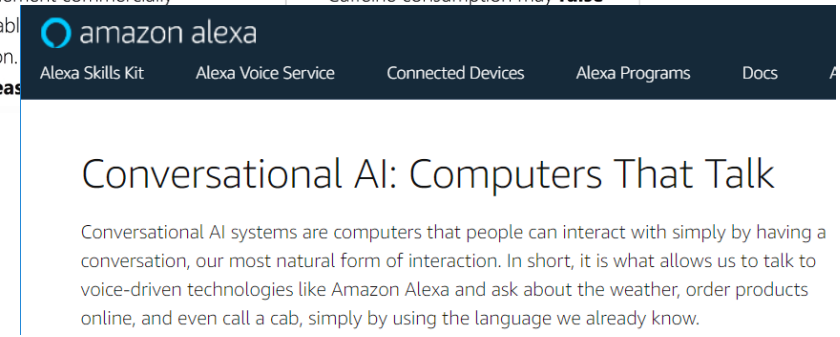
Microsoft Brainwave (AI cloud using Intel FPGAs)



… many more not shown and in the pipeline

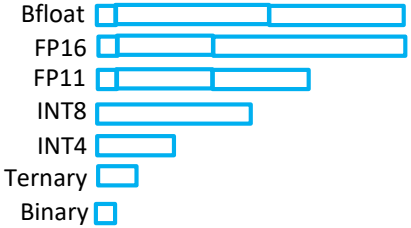
Real-time intelligent services **on the rise**



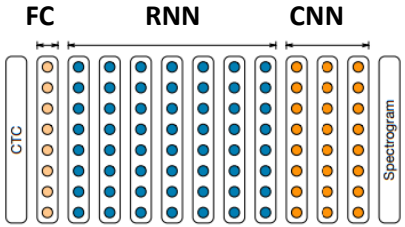


Evolving AI Needs Programmable & Customizable HW

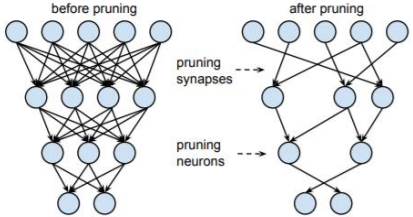
Mix of precisions
(just a few examples below)



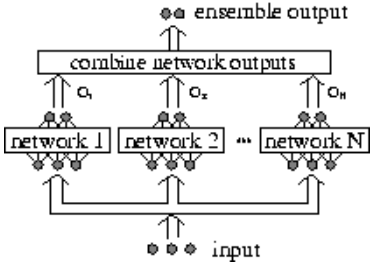
Mix of NN layer types
(e.g., DeepSpeech2)



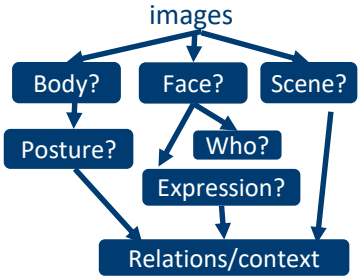
Mix of sparse layers
(e.g., NIPS'15)



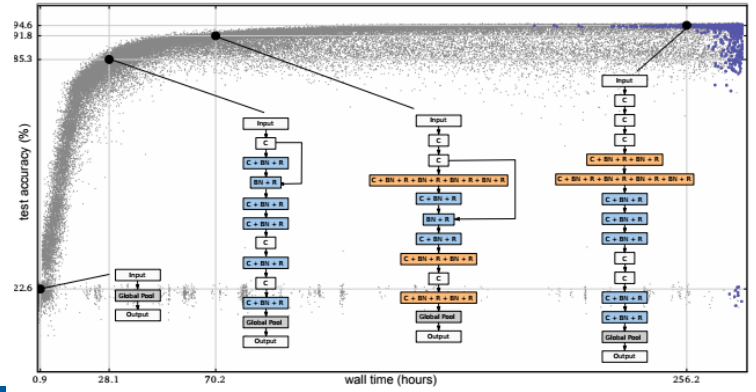
Ensemble of NNs



Mix of uses
(e.g., context analyses)



AutoML: Computer-generated custom NNs (e.g., available in Google Cloud)



Implications to Computing

- Deluges of data → process near data, minimize movement
- Myriad AI algorithm variations → programmability + customizability
- Interactive, real-time AI services → latency optimized architecture
- Larger data and AI models → need scalable solution

FPGAs are strong in all these areas

Talk Outline

Part 1: Trends in AI and Big Data

Part 2: Trends and opportunities for FPGA

Part 3: Research highlights

FPGA Overview

Fine-grained general-purpose spatial arch.
(bit-level, cycle-level, dataflow-level programmable)

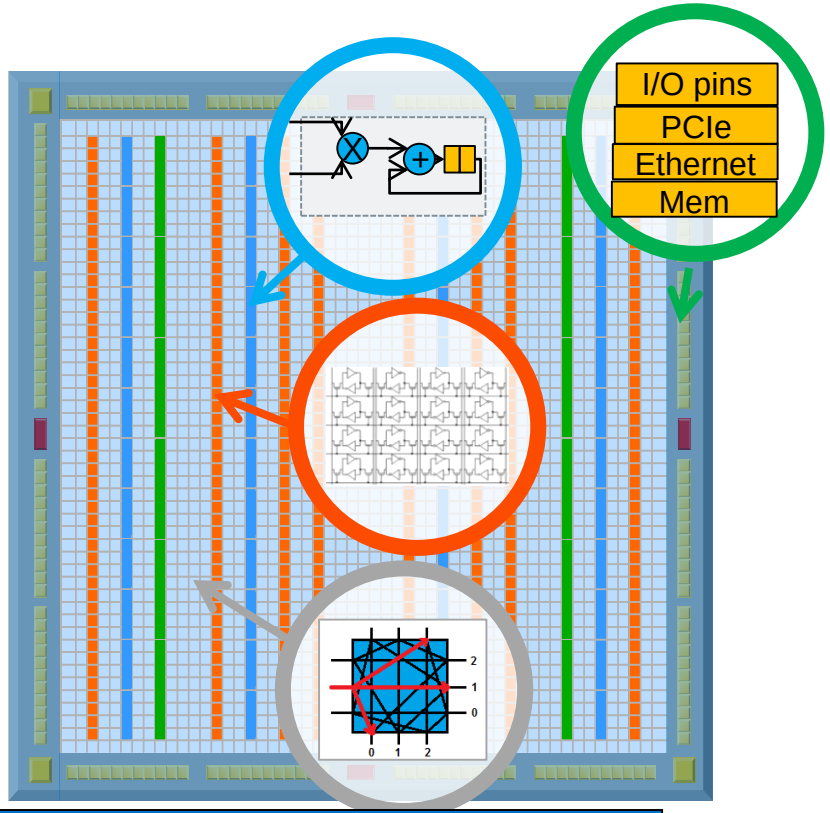
Sea of Programmable Logic and Routing

1000s of hard compute units (“DSP”)
(e.g., 10 TFLOP/s FP32 in Stratix 10 2800)

1000s of Hard Scratchpads (“M20Ks”)
(e.g., ~30MBs total size, ~10s TB/s in Stratix 10 2800)

Many I/Os options

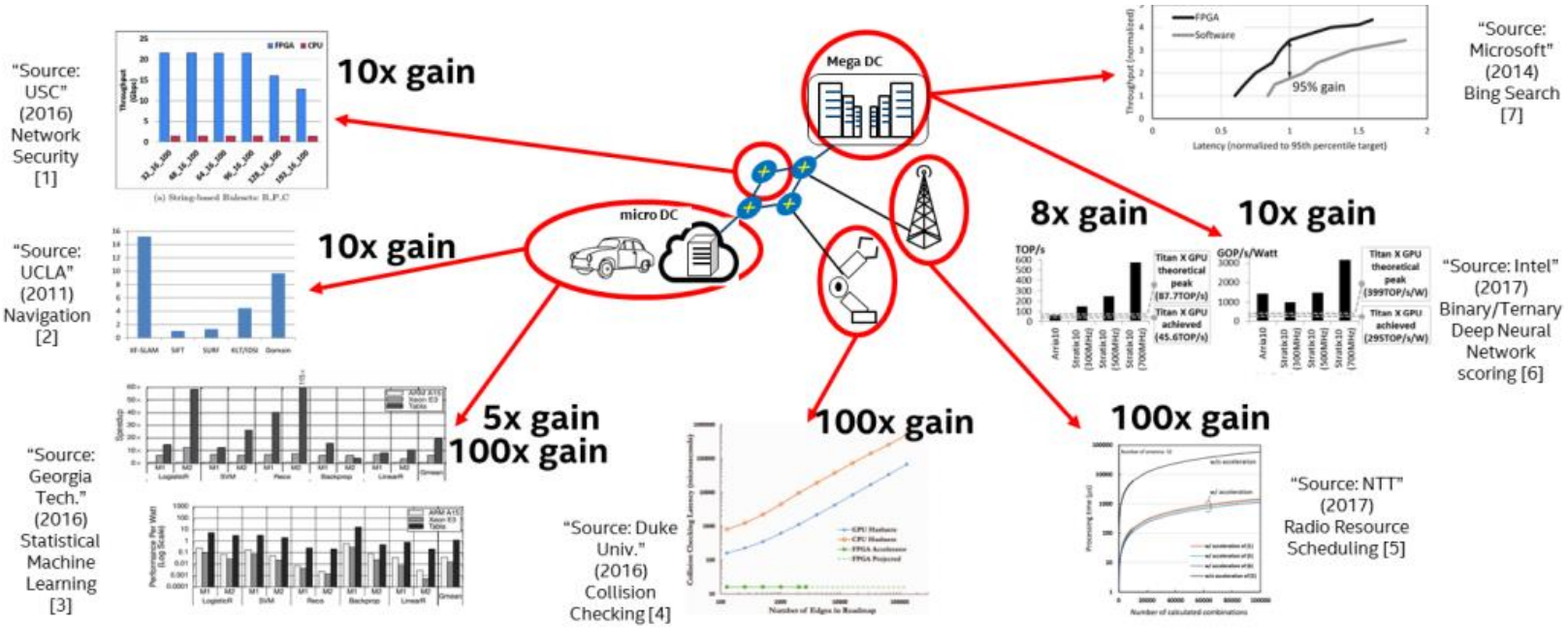
"Bare-metal" RTL “program”



Great for near-data latency-sensitive fine-grained apps

Complementary to other general-purpose architectures (CPU, GPU)

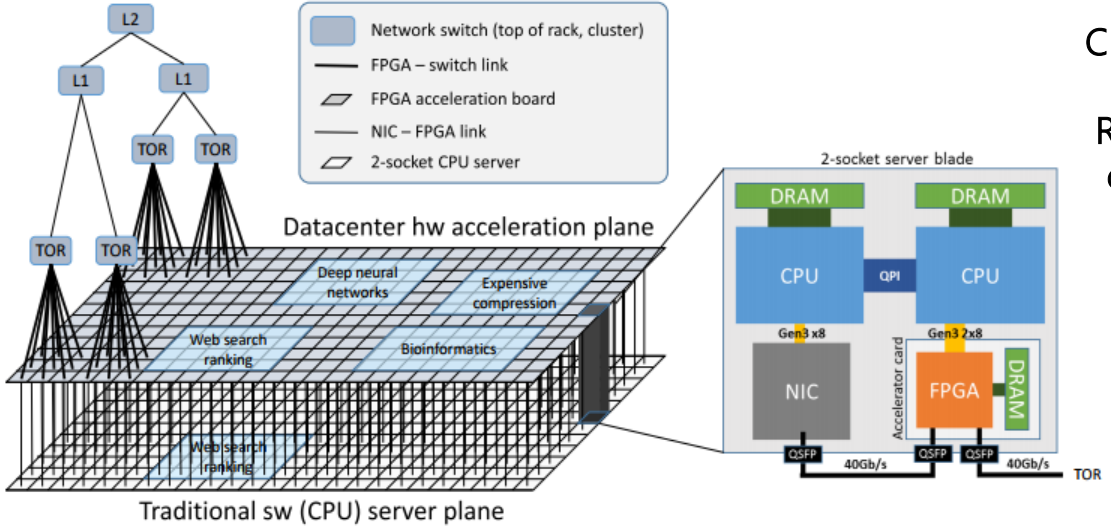
FPGAs are in Many Places, from Edge to Cloud

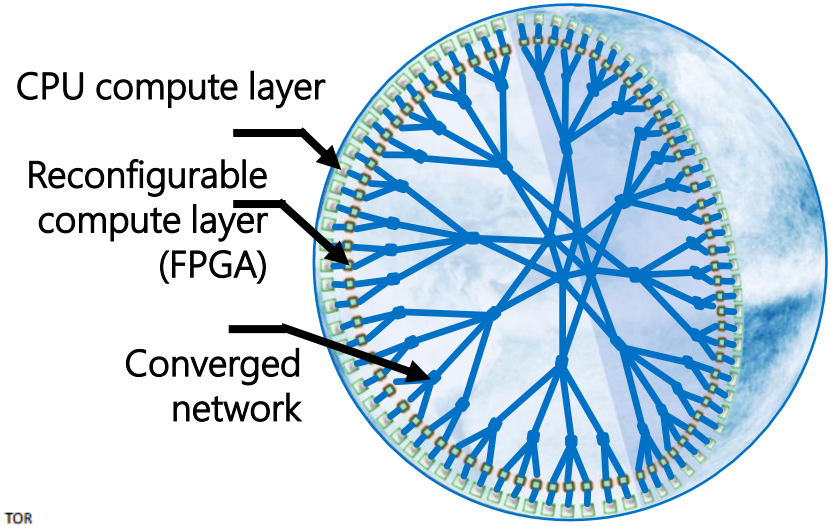


FPGAs are already in strategic points, where data moves across edge/cloud

[3] D. Mahajan et al., "TABLA: A unified template-based framework for accelerating statistical machine learning," High Performance Computer Architecture (HPCA), 2016.
[4]. S. Murray, W. Floyd-Jones, Y. Qi, G. Konidaris and D. J. Sorin, "The microarchitecture of a real-time robot motion planning accelerator," MICRO, 2016.
[5]. Yuki Arikawa, et al., "High-speed radio-resource scheduler with hardware accelerator for fifth generation mobile communications systems," IEICE Communications Express, Advance Publication, 2017.
[6]. E. Nurvitadhi, G. Venkatesh, et. al., "Can FPGAs Beat GPUs in Accelerating Next-Generation Deep Neural Networks?" FieldProgrammable Gate Arrays (FPGA), 2017.
[7]. A. Putnam, A. M. Caulfield, E. S. Chung, et. al., "A reconfigurable fabric for accelerating large-scale datacenter services," international symposium on Computer architecuture (ISCA), 2014.

FPGAs are in Data Center and Scalable



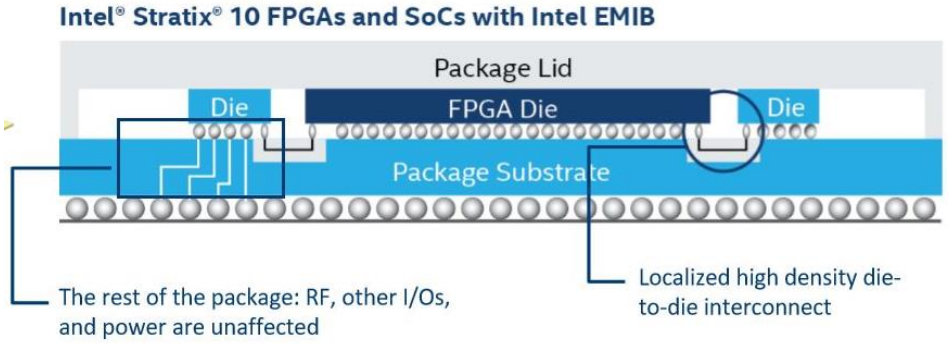


Figures from: Microsoft Configurable Cloud talk/paper

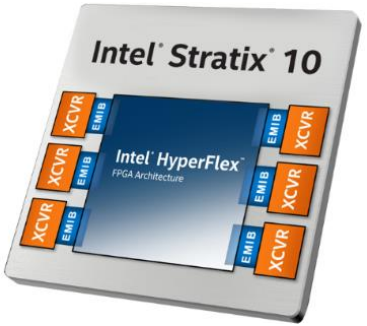
E.g. MSFT reconfigurable cloud: "planet scale" apps on many networked FPGAs

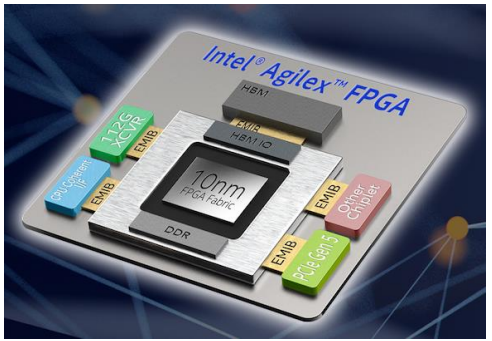
Others are also integrating FPGAs in their cloud (e.g., Amazon, Baidu)

FPGAs can Take Advantage of Heterogenous Integration



Source: S. Shumarayev (Hotchips 2017)

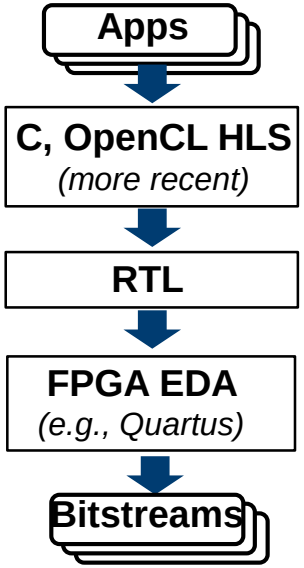




Intel FPGAs already heterogeneously integrated

FPGAs are Programmable in Different Ways

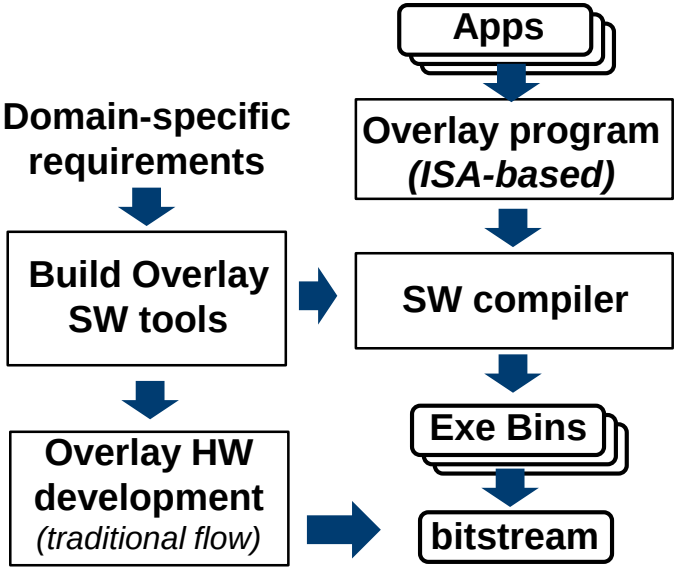
Raising abstraction for HW-oriented users



SW-oriented flow with Overlays

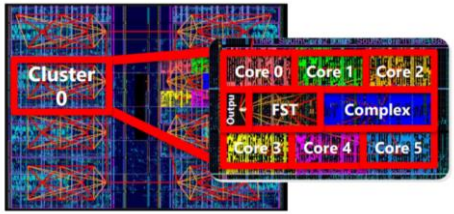
SW abstraction, compile speed

Domain customizations

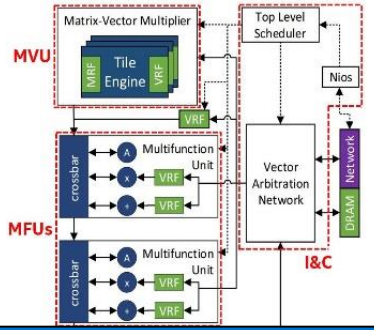


Examples of commercial overlays

Bing's FFE processor *[ISCA'14]*



Brainwave AI overlay *[ISCA'18]*



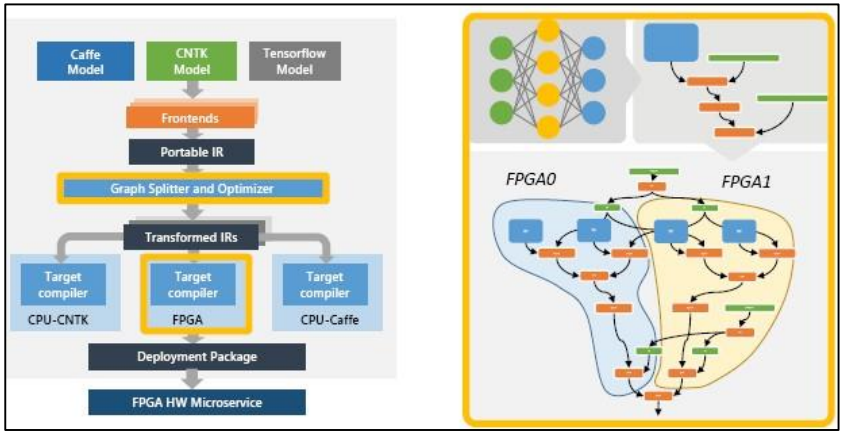
[ISCA
[ISCA

Overlay goes beyond HW-oriented FPGA EDA → SW flexibility + programming speed

FPGAs Shine in Low-latency Scalable DNN Inference

Inference goals	FPGA strengths
Interactive, real-time	**Near data compute** (High on-chip BW, next to arith)
On various AI algos	**Customizable** (e.g., numerics, dataflows)
On big data and model	**Great I/Os for scaling** (many xcvrs for multi-node)
With high Efficiency	**Energy efficient, fine grained, spatial fabric**

Commercial example:
Microsoft Brainwave AI Cloud based on Intel FPGAs [HC'17]



Tremendous opportunities in combined AI requirements and FPGA capabilities

[HC'17] E. Chung, J. Fowers, K. Ovtcharov, et. al., "Accelerating Persistent Neural Networks at Datacenter Scale," Hotchips 2017.

Talk Outline

Part 1: Trends in AI and Big Data

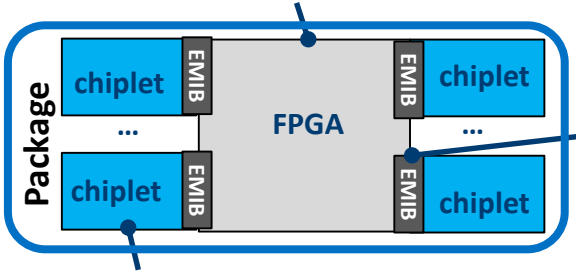
Part 2: Trends and opportunities for FPGA

Part 3: Research highlights

- **Enhancing FPGAs with AI Chiplets**
- Software-level programmability using FPGA overlays

Enhancing FPGAs with Chiplets for AI [FPL'18][FCCM'19]

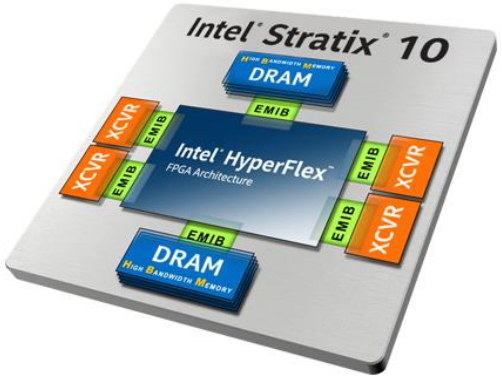
FPGA: flexibility for custom application-specific ops



Chiplets: efficiency for domain-shared ops

Intel's 2.5D EMIB for integration

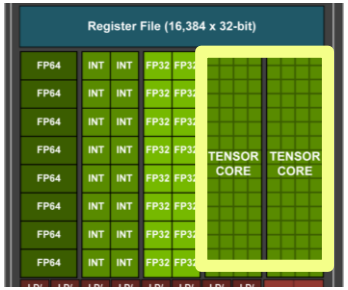
Stratix 10 FPGAs already System-in-Package: use 2.5D EMIBs for xcvr and mem chiplets



We proposed chiplets for AI → scalable, customizable, shared ecosystem

[FPL'18] E. Nurvitadhi, J. Cook, A. Mishra, D. Marr, et. al., "In-Package Domain-Specific ASICs for Intel® Stratix® 10 FPGAs: A Case Study of Accelerating Deep Learning Using TensorTile ASIC," FPL 2018.
[FCCM'19] E. Nurvitadhi, D. Kwon, A. Jafari, A. Boutros, et. al., "Why Compete When You Can Work Together: FPGA-ASIC Integration for Persistent RNNs," to appear at FCCM 2019.

Very Scalable: Can Mix & Match FPGAs + Chiplets



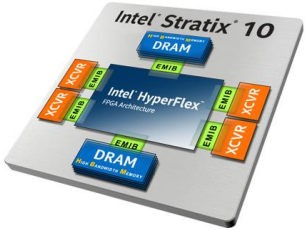
Volta: GPU + Tensor Cores
125 TOPs (FP16)
26 MB on-chip RAMs

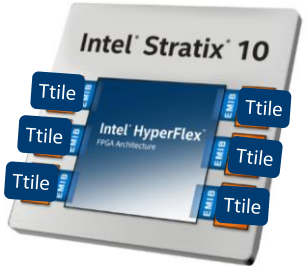
1.5x and 2.7x better peak TOPs and RAMs

Est. 69 TOPs (INT8) in a T-tile

Ttile

Stratix 10 400 (378K LEs)



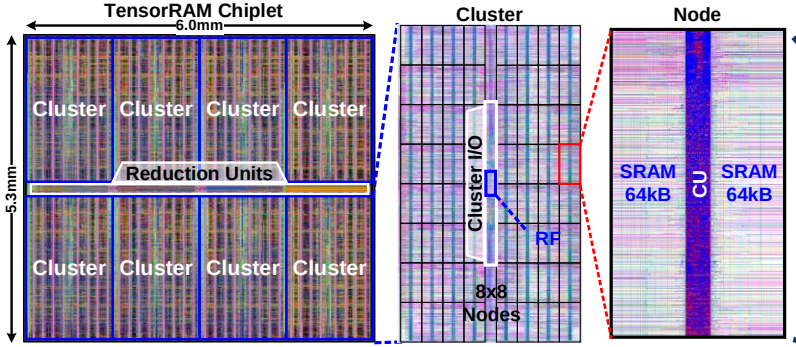


S10 2800 + 6x compute-intensive chiplets on 14nm
Est. 194 TOPs (FP16)
72MB on-chip RAMs

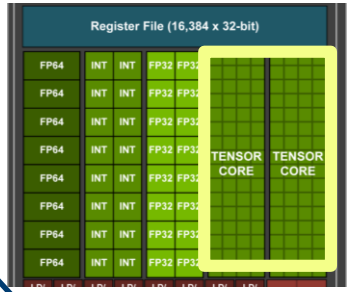
Small → **Large**

Small to large FPGAs in Stratix 10 family, with links to 1 or more AI Chiplets.
Example of compute-intensive 14nm "TensorTile" chiplet shown

Very Scalable: Can Mix & Match FPGAs + Chiplets



TensorRAM 32 mm² 10nm layout [IL/CRL]



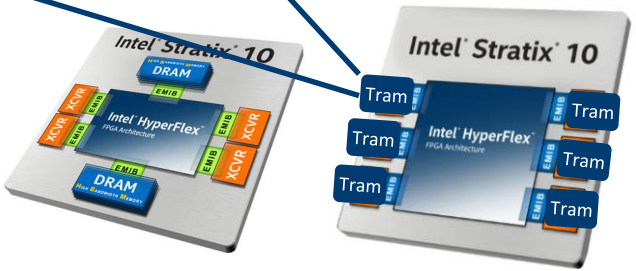
Volta: GPU + Tensor Cores
125 TOPs (FP16)
26 MB on-chip RAMs

3.1x and 15.8x better peak TOPs and RAMs

Est. 64 TOPs (INT8) in a T-ram

Tram

Stratix 10 400 (378K LEs)



S10 2800 + 6x data-intensive chiplets on 10nm
Est. 393 TOPs (INT8)
412MB on-chip RAMs

Small

Large

Can also build multiple variants of chiplets. Example of data-intensive 10nm chiplet "TensorRAM" for persistent AI shown here.

Talk Outline

Part 1: Trends in AI and Big Data

Part 2: Trends and opportunities for FPGA

Part 3: Research highlights

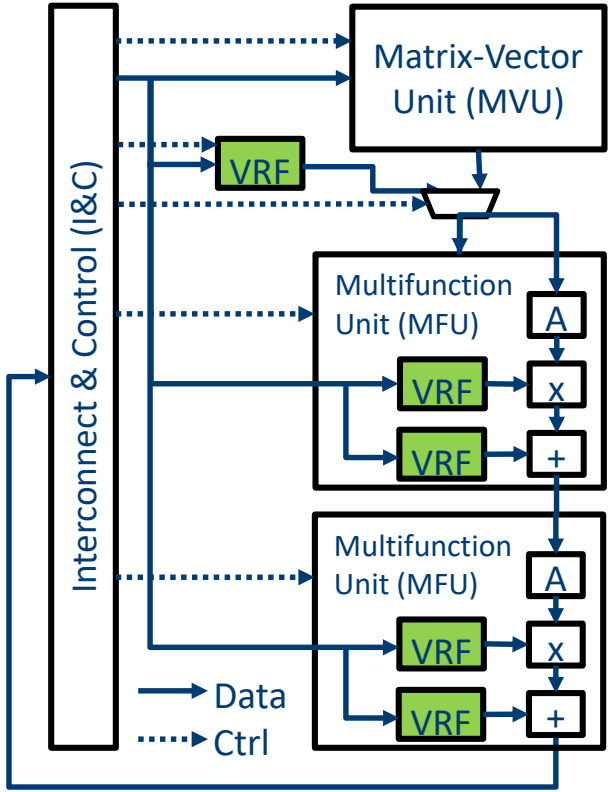
- Enhancing FPGAs with AI Chiplets
- **Software-level programmability using FPGA overlays**

Case Study on Brainwave-like Overlay [FCCM'19]

// Overlay Program example
// in tensor Instruction Set Architecture (ISA)

```
v wr( InitialVrf , ivrf xt ) ;
// xWf = xt * Wf + bf
v rd( InitialVrf , ivrf xt ) ;
mv mul(mrf Wf);
vv add(asvrf bf) ;
v wr(AddSubVrf, asvrf xWf);
// xWi = xt * Wi + bi ...
// xWf = xt * Wo + bo ...
// xWc = xt * Wc + bc ...
```

Compile



Key questions: How fast can we write software? Does it perform well?

Overlay is Easy to Program, Runs Fast, & Customizable

Study:

- **Batch-1 RNN/GRU/LSTM (Deepbench)**
- **Implemented programs in tensor ISA**
- **Compiled to a single-bitstream overlay**

Ver	Description	Inst Count	Est. Engr. Hours	RAM Footprint
V1	Baseline functionality	20	4 hrs	Base
V2	Loop unroll, SW pipeline	20	1 hr	1x V1
V3	Efficient graph mapping	19	1 hr	1.1x V2

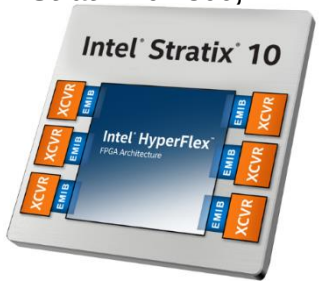
Created & optimized programs in ~hours

Compile time in ~ms

Load program to FPGA overlay in ~us

Overlay on a large FPGA

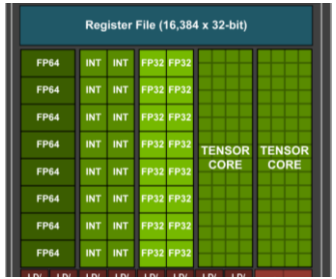
Stratix 10 2800, 14nm



vs.

cuDNN for persistent AI on a large GPU

Titan V (Volta), 12nm



FPGA offers better latency than GPU, by 3x (FP32) & 10x (INT8) on average

Why? GPU underutilized it's available TOP/s (even with latest cuDNN software from Nvidia with persistent AI support)

Overlay Challenge

- Overlay targets entire domain, so it has overheads of general architecture
  - e.g., Instruction Set Arch (ISA), instruction processing (fetch, decode, etc)
- Need highly optimized implementation to mitigate generality overhead
  - One or few bitstream(s) per domain, but highly tuned (e.g., best frequency, packing)
- Currently, high effort human experts to guide EDA tool to best implementation
  - E.g., floor planning based on overlay architecture knowledge
  - E.g., guiding the scaling to larger FPGAs → need scalable floor planning methods

Need EDA innovations for overlays to achieve better results with less effort

In Closing

- It’s exciting time for FPGAs!
- FPGAs are well-positioned for the era of AI and Big Data
  - AI algorithm optimizations and evolution → needs programmability + configurability
  - Interactive real-time AI services → needs latency-optimized solutions
  - Too much data from various sources → needs near-data and scalable processing
- Research opportunities
  - Beyond traditional FPGA EDA → making FPGAs software-programmable (overlay)
  - Heterogeneous FPGA-based systems → FPGA + chiplets


intel®
experience
what's inside™