AI/ML HIGH PERFORMANCE COMPUTING WITH FPGAS

Jose Roberto Alvarez

Senior Director, CTO Office

Programmable Solutions Group

Legal Disclaimers

† Tests measure performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

Diverse Applications for FPGAs

Advanced, multi-function accelerators

Flexibility for highly-differentiated products

H/W re-programmability for evolving market requirements & changing standards

EDGE/EMBEDDED

NEC NeoFace Facial Recognition Accelerator

NETWORK

Rakuten Cloud-Native Mobile Network

DATA CENTER

Microsoft AI Acceleration at Cloud Scale for Azure and Bing

Industry Has Reached an Inflection Point

"…It may prove to be more economical to build large systems out of smaller functions, which are separately packaged and interconnected."

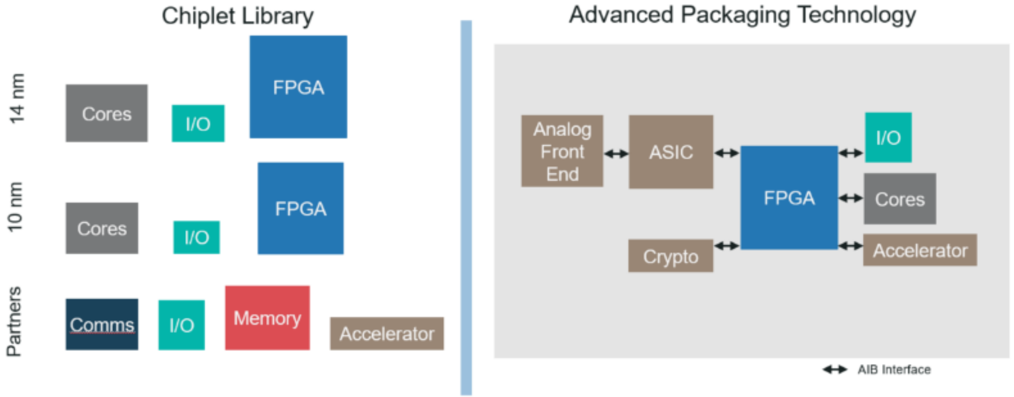
Gordon E. Moore

Industry Has Reached an Inflection Point

Rapidly Emerging Workloads Demands Flexibility and Interoperability

Breakthrough Packaging Technologies Approach On-Die Capabilities

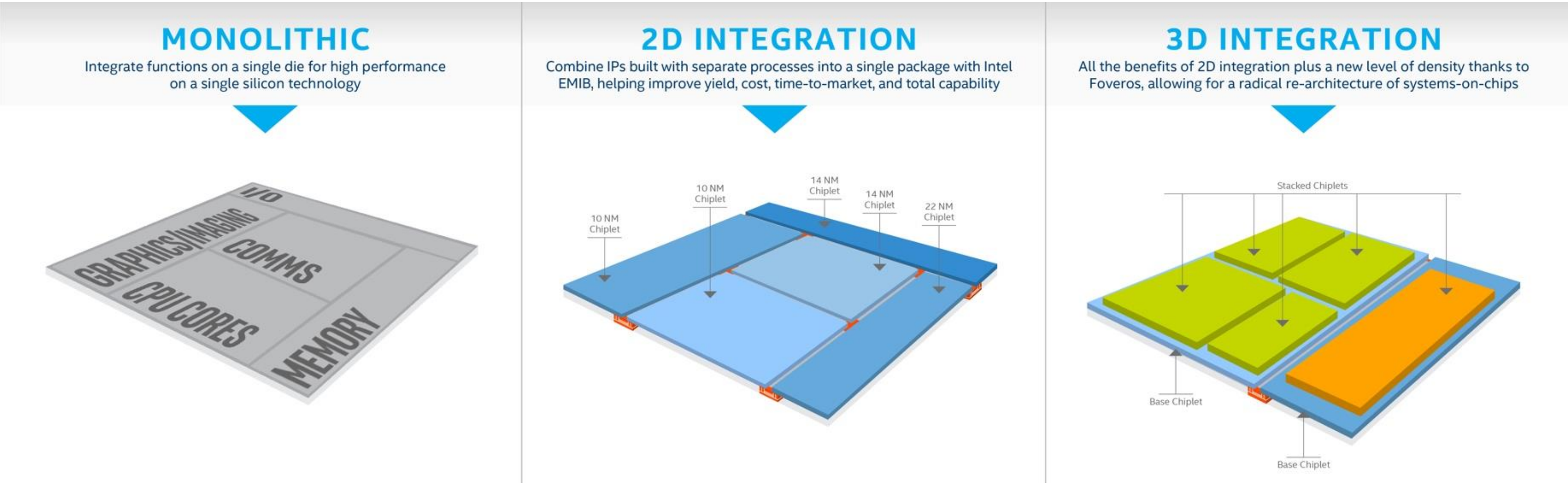
Heterogeneous Integration – Innovation Through Chiplet Ecosystem!



HETEROGENEOUS INTEGRATION

2D and 3D Packaging Drive New Design Flexibility

The combination of advanced 2D and 3D packaging technologies allows Intel to flexibly combine smaller chiplets of IP to meet the demands of a huge range of applications, power envelopes, and form factors. Intel® embedded multi-die interconnect bridge (EMIB) and Foveros are advanced 2D and 3D packaging technologies, delivering high performance at low cost.



MONOLITHIC

Integrate functions on a single die for high performance on a single silicon technology

2D INTEGRATION

Combine IPs built with separate processes into a single package with Intel EMIB, helping improve yield, cost, time-to-market, and total capability

3D INTEGRATION

All the benefits of 2D integration plus a new level of density thanks to Foveros, allowing for a radical re-architecture of systems-on-chips

Any-to-Any Heterogenous 3D Packaging

EMBEDDED MULTI-DIE INTERCONNECT BRIDGE

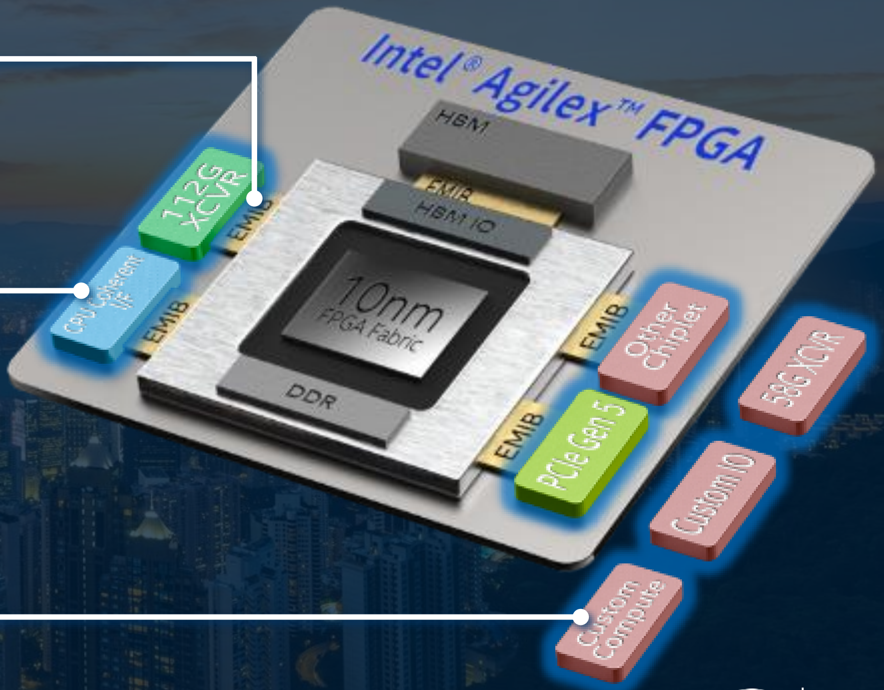
No compromise die-to-die packaging interconnect with high performance, low cost, and high density

CHIPLET

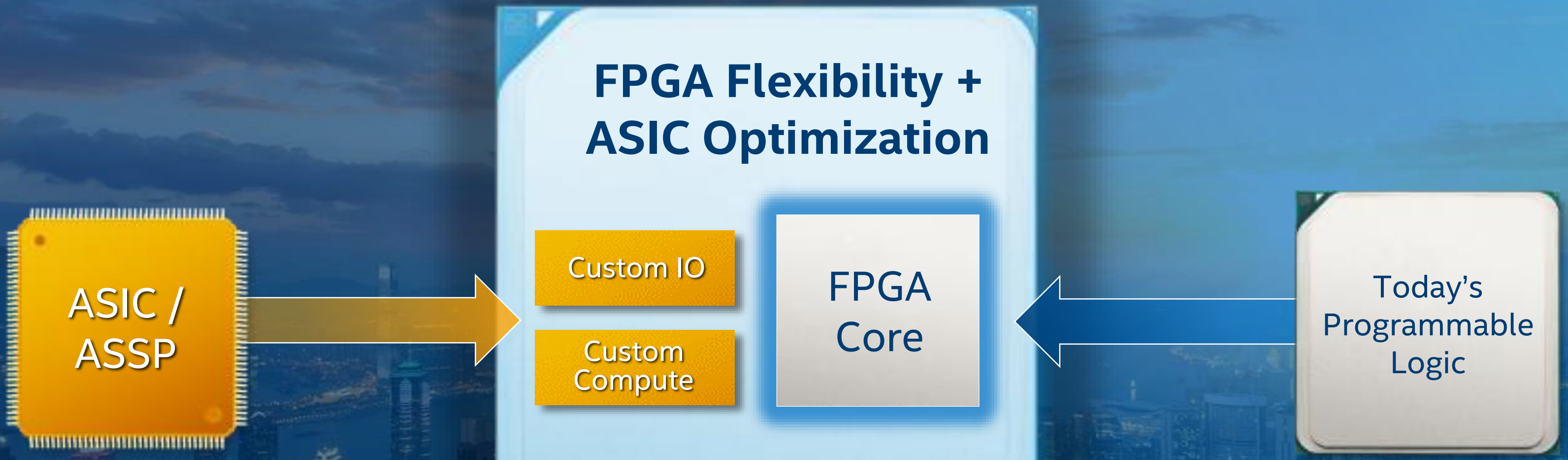
Library of chiplets including transceivers, custom I/O & custom compute tiles

Intel® eASIC™

Ultimate customization with integrated Intel® eASIC™ chiplet



Delivering Best of Both Worlds



ASIC / ASSP

FPGA Flexibility + ASIC Optimization

Custom IO

Custom Compute

FPGA Core

Today's Programmable Logic

Targeted Power & Performance

Maximum Hardware Flexibility

Agile and Flexible Artificial Intelligence

UP TO

40 TFLOPS[1]

DSP PERFORMANCE

Configurable DSP

- FP32
- BFLOAT16
- FP16
- INT8

intel AGILEX inside

Flexibility for evolving AI workloads and integrating AI with other functions

Supports low-precision INT7 through INT2 configurations for high performance

Only FPGA supporting hardened BFLOAT16 & FP16

Complementary to:

intel XEON PLATINUM inside

intel CORE i7 8th Gen

intel ATOM inside

intel NERVANA inside

intel MOVIDIUS inside

oneAPI Optimized Applications, Middleware, Frameworks

FPGA Summary

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”)

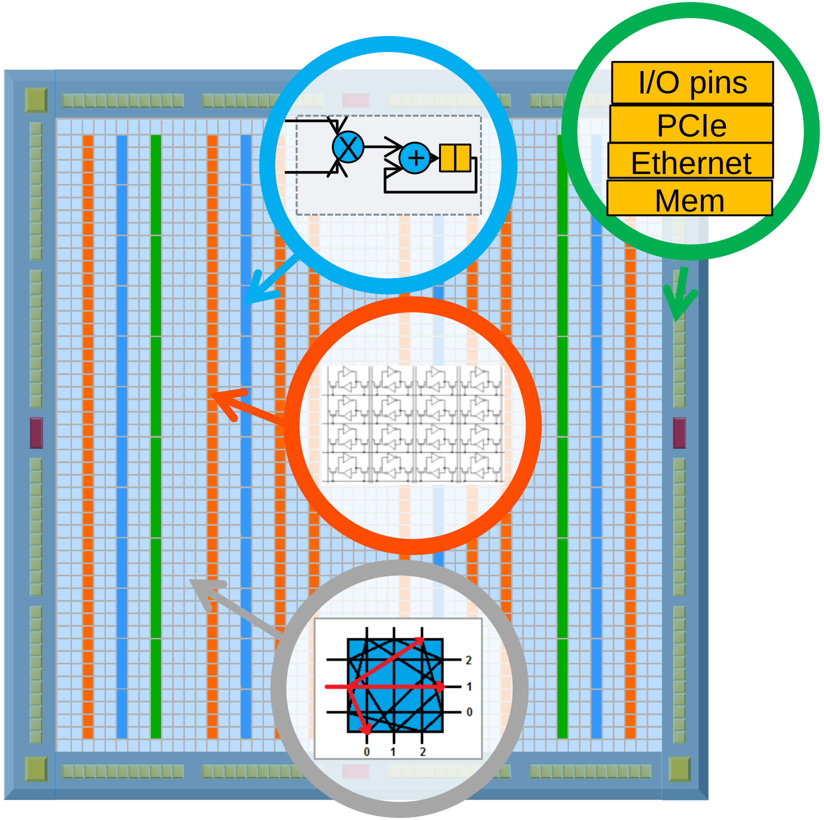
(e.g., ~10 TFLOP/s FP32 in Stratix 10 2800)

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

Commercial DNN Platforms Continue to Evolve

Few years ago

DNNs on general purpose platforms

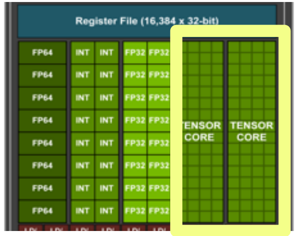
CPU

GPGPU

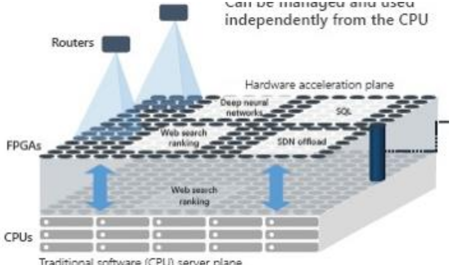
Some Examples Now

Customized DNNs platforms are available (i.e., HW accelerators in cloud/embedded/IoT)

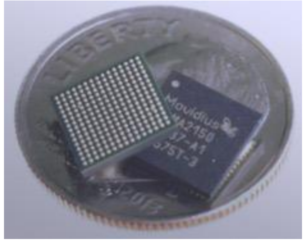
Nvidia Volta GPU+TensorCore



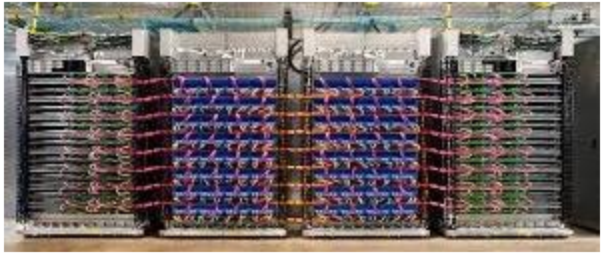
Microsoft Brainwave (AI cloud using Intel FPGAs)


Can be managed and used independently from the CPU
Routers
Hardware acceleration plane
Deep neural networks
SQL
Web search ranking
SDN offload
FPGAs
Web search ranking
CPUs
Traditional software (CPU) server plane

Intel Movidius



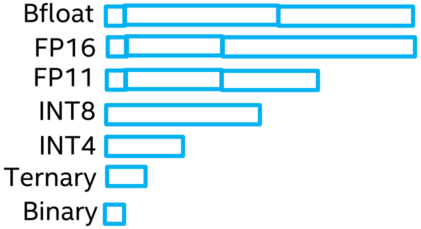
Google Cloud TPU



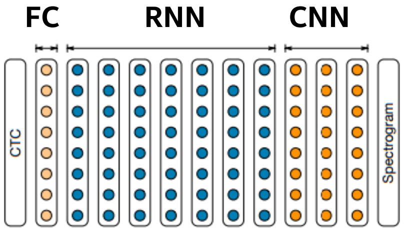
Source: E. Nurvithadi, et.al. Intel Labs

DNN Trends Going Forward

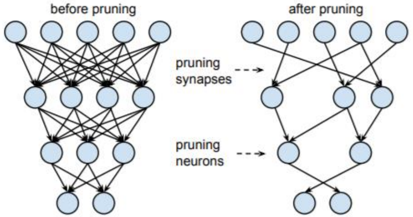
Mix of precisions
(just a few examples below)



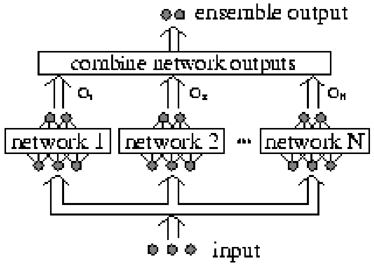
Mix of NN layer types
(e.g., DeepSpeech2)



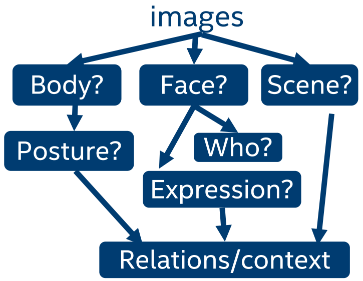
Mix of sparse layers
(e.g., NIPS'15)



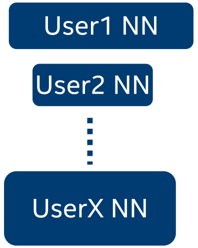
Ensemble of NNs



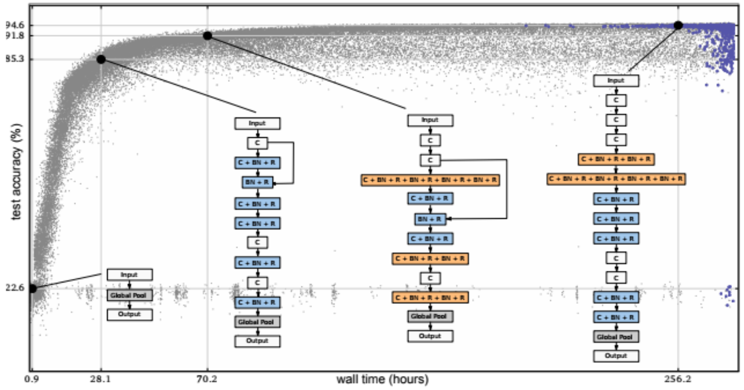
Mix of uses
(e.g., context analyses)



Multi Tenants



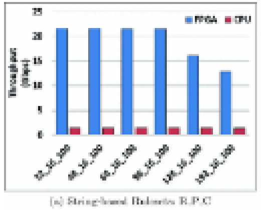
AutoML: Computer-generated custom NNs
(e.g., available in Google Cloud)



Needs programmability + customizability

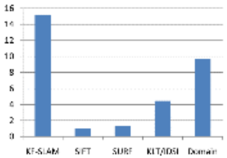
FPGAs Are Excellent Accelerators for Latency-Sensitive Irregular Computations

“Source: USC” (2016) Network Security [1]



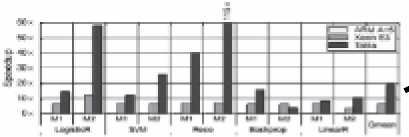
10X gain†

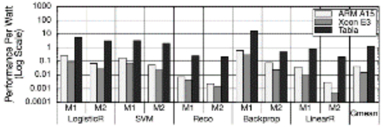
“Source: UCLA” (2011) Navigation [2]



10X gain †

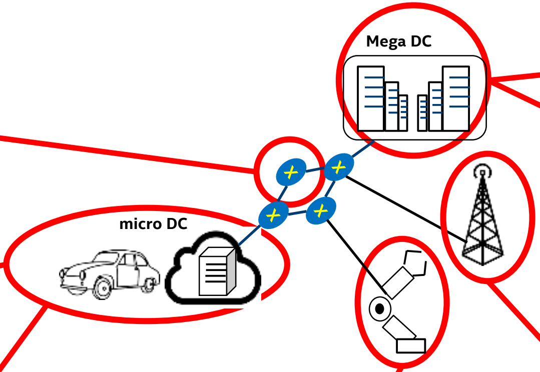
“Source: Georgia Tech.” (2016) Statistical Machine Learning [3]



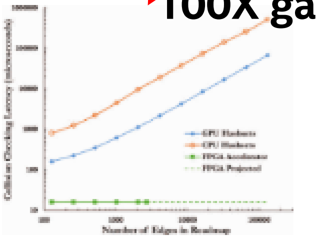


5X gain †

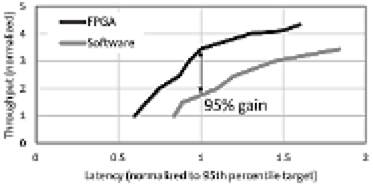
100X gain †



“Source: Duke Univ.” (2016) Collision Checking [4]



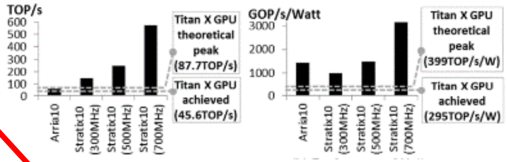
100X gain †



“Source: Microsoft” (2014) Bing Search [7]

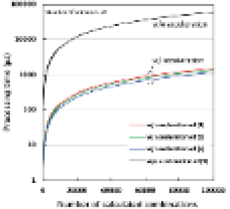
8X gain †

10X gain †



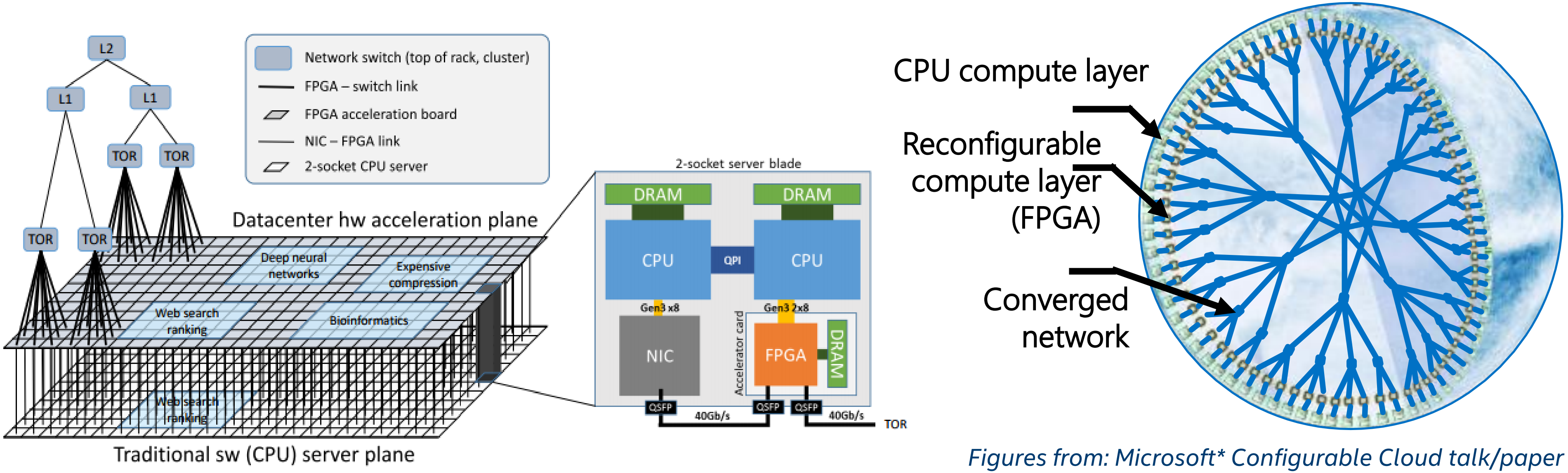
“Source: Intel” (2017) Binary/Ternary Deep Neural Network scoring [6]

100X gain †



“Source: NTT” (2017) Radio Resource Scheduling [5]

FPGAs Are Scalable



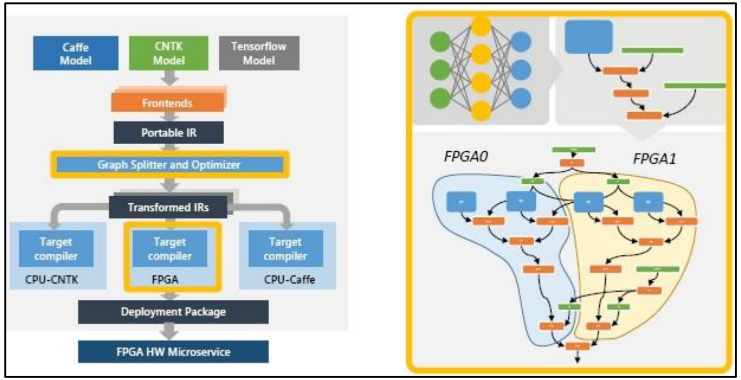
Figures from: Microsoft Configurable Cloud talk/paper*

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

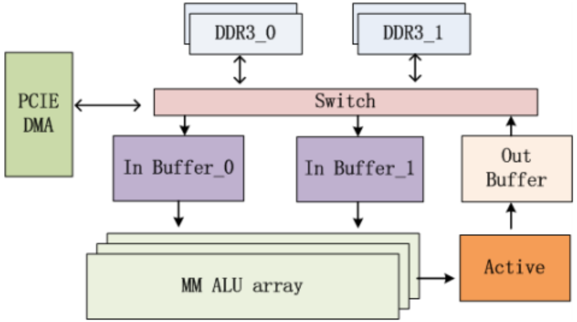
Others are also integrating FPGAs in their cloud (e.g., Amazon, Baidu)

FPGAs already used for Deep Learning Today

E.g., MSFT BrainWave (2017)



E.g., Baidu's SDA for DNNs (2015)



ISFPGA'18 Panel on DL: value of FPGA's flexibility

- MSFT: 1 week to add a new op in BrainWave
- MSFT: an intern repurposed BrainWave for another workload in 12 weeks
- FB Research: FPGAs for irregular DLs (e.g., FFTs)

Implications to Computing + Opportunities for FPGAs

- Too much data → process near data, minimize movement
  - *FPGAs already in key places where data flows (network, storage, sensors)*
- Myriad AI algorithm variations → programmability + customizability
  - *FPGAs spatial fabric are highly configurable + general purpose programmable*
- Interactive, real-time AI services → latency optimized architecture
  - *FPGAs are known for their deterministic and latency-optimized characteristics*
- Larger data and AI models → need scalable solution
  - *FPGAs are tightly coupled with transceivers to facilitate multi-node connections*

Tremendous opportunities in combined AI requirements and FPGA capabilities

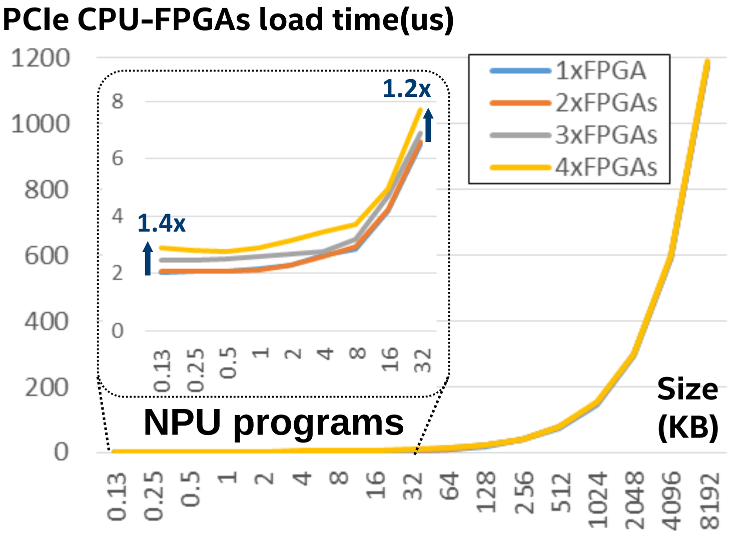
Case Study of programming RNNs on NPU Overlay

How fast can we write + optimize programs?

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

Can optimize RNN programs within ~hours
Compile time in seconds

How fast can we program our overlay?

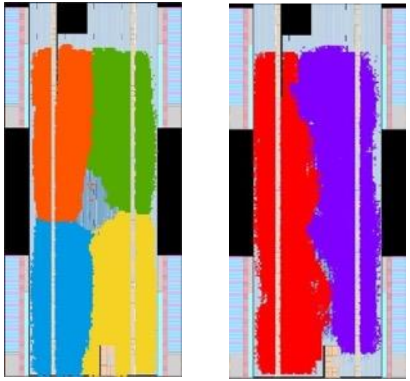


Typical programming time <8us,
even when targeting multiple FPGAs

Does it perform well?

2 bitstreams, tuned to Stratix 10 2800

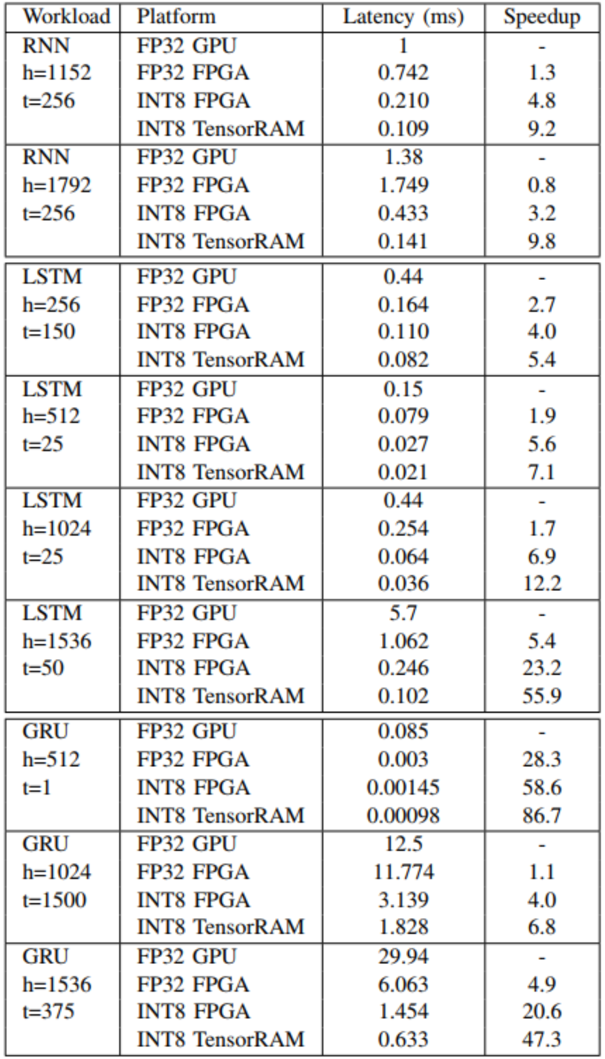
NPU	ALMs	M20Ks	DSPs	Freq. (MHz)	Peak TOPS
INT8 4T-120D-40L	567,982 (61%)	9,018 (77%)	4,880 (85%)	275	10.6
FP32 2T-64D-32L	286,024 (31%)	4,441 (38%)	4,768 (83%)	200	2.5



(a) 4-tile INT8

(b) 2-tile FP32

Ran various RNNs, GRUs, LSTMs from Deep Bench (batch1)



Workload	Platform	Latency (ms)	Speedup

RNN
h=1152
t=256
	FP32 GPU	1	-

	FP32 FPGA	0.742	1.3
	INT8 FPGA	0.210	4.8
	INT8 TensorRAM	0.109	9.2

RNN
h=1792
t=256
	FP32 GPU	1.38	-

	FP32 FPGA	1.749	0.8
	INT8 FPGA	0.433	3.2
	INT8 TensorRAM	0.141	9.8

LSTM
h=256
t=150
	FP32 GPU	0.44	-

	FP32 FPGA	0.164	2.7
	INT8 FPGA	0.110	4.0
	INT8 TensorRAM	0.082	5.4

LSTM
h=512
t=25
	FP32 GPU	0.15	-

	FP32 FPGA	0.079	1.9
	INT8 FPGA	0.027	5.6
	INT8 TensorRAM	0.021	7.1

LSTM
h=1024
t=25
	FP32 GPU	0.44	-

	FP32 FPGA	0.254	1.7
	INT8 FPGA	0.064	6.9
	INT8 TensorRAM	0.036	12.2

LSTM
h=1536
t=50
	FP32 GPU	5.7	-

	FP32 FPGA	1.062	5.4
	INT8 FPGA	0.246	23.2
	INT8 TensorRAM	0.102	55.9

GRU
h=512
t=1
	FP32 GPU	0.085	-

	FP32 FPGA	0.003	28.3
	INT8 FPGA	0.00145	58.6
	INT8 TensorRAM	0.00098	86.7

GRU
h=1024
t=1500
	FP32 GPU	12.5	-

	FP32 FPGA	11.774	1.1
	INT8 FPGA	3.139	4.0
	INT8 TensorRAM	1.828	6.8

GRU
h=1536
t=375
	FP32 GPU	29.94	-

	FP32 FPGA	6.063	4.9
	INT8 FPGA	1.454	20.6
	INT8 TensorRAM	0.633	47.3

Our S10 FPGA NPU has 6x lower peak FP32 TOP/s than a large Titan V GPU

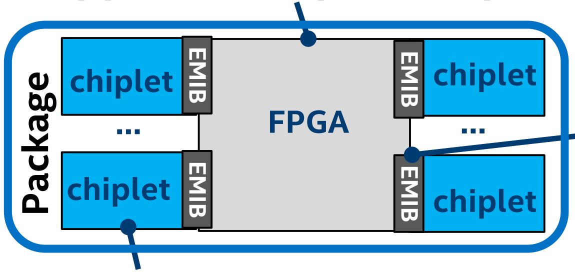
However, S10 performs better by 3x (FP32) & 10x (INT8) on average

Why? GPU underutilize its available TOP/s, even with latest cuDNN with persistent AI support

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

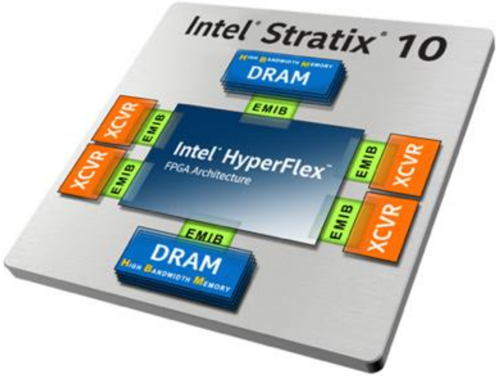
Stratix10 FPGAs already System-in-Package: use 2.5D EMIBs to offer xcvr and memory chiplets next to FPGA

FPGA: flexibility for custom application-specific ops



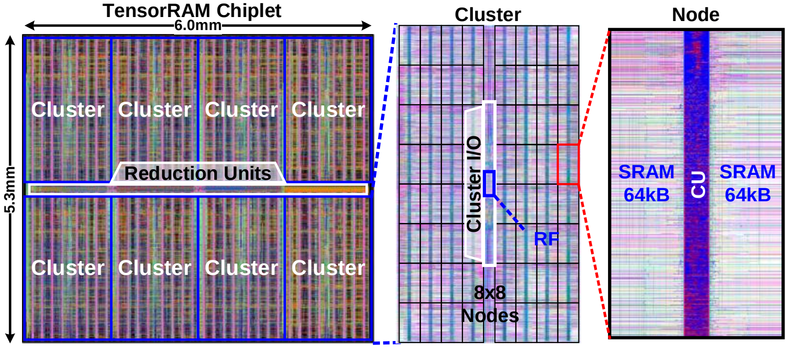
Intel’s 2.5D EMIB for integration

Chiplets: efficiency for domain-shared ops

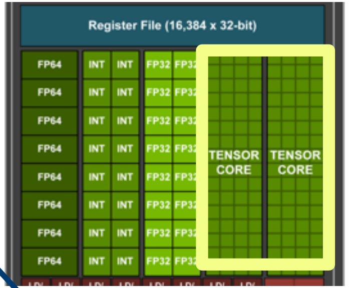


[FPL’18] E. Nurvitadhi, J. Cook, A. Mishra, D. Marr, et. al., “In-Package Domain-Specific ASICs for Intel® Stratix® 10 FPGAs: A Case Study of Accelerating Deep Learning Using TensorTile ASIC,” FPL 2018.
[FCCM’19] E. Nurvitadhi, D. Kwon, A. Jafari, A. Boutros, et. al., “Why Compete When You Can Work Together: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

Very Scalable: Can Mix & Match FPGAs + Tiles



TensorRAM 32 mm^2 10nm layout [IL/CRL]



Volta: GPU + Tensor Cores

125 TOPs (FP16)

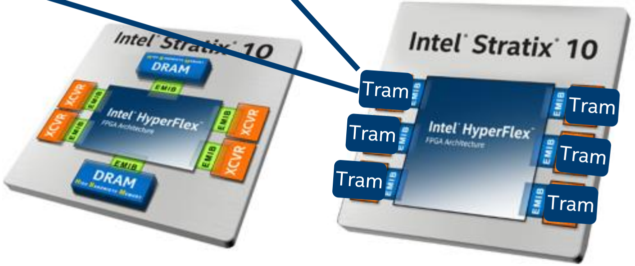
26 MB on-chip RAMs

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)

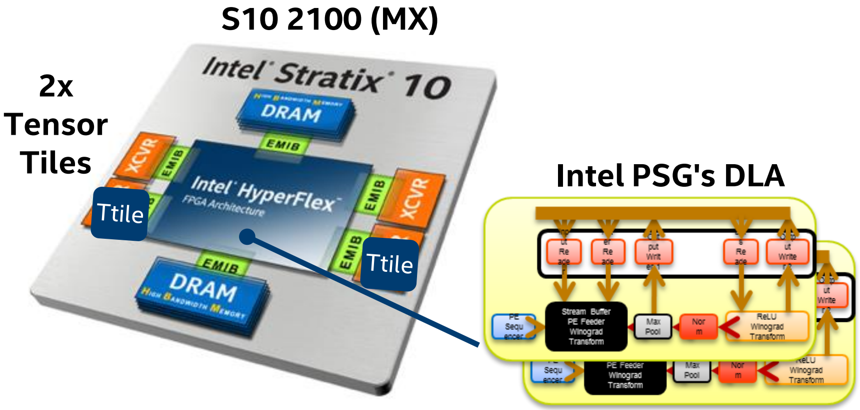
412MB on-chip RAMs

Small → Large

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

Case Studies

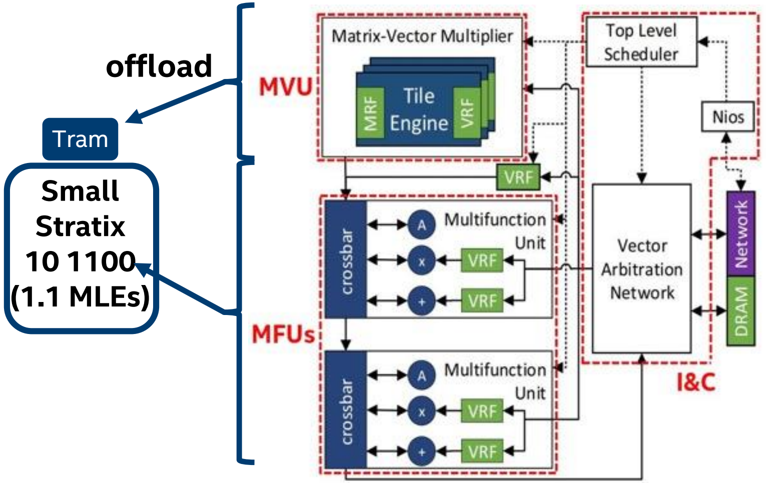
(1) AlexNet (FP16) Inference, low batch



The 2x tiles improved performance and performance/Watt by 4x and 3.3x vs. FPGA only for CNN workload

(2) Persistent RNN/GRU/LSTM (INT8)

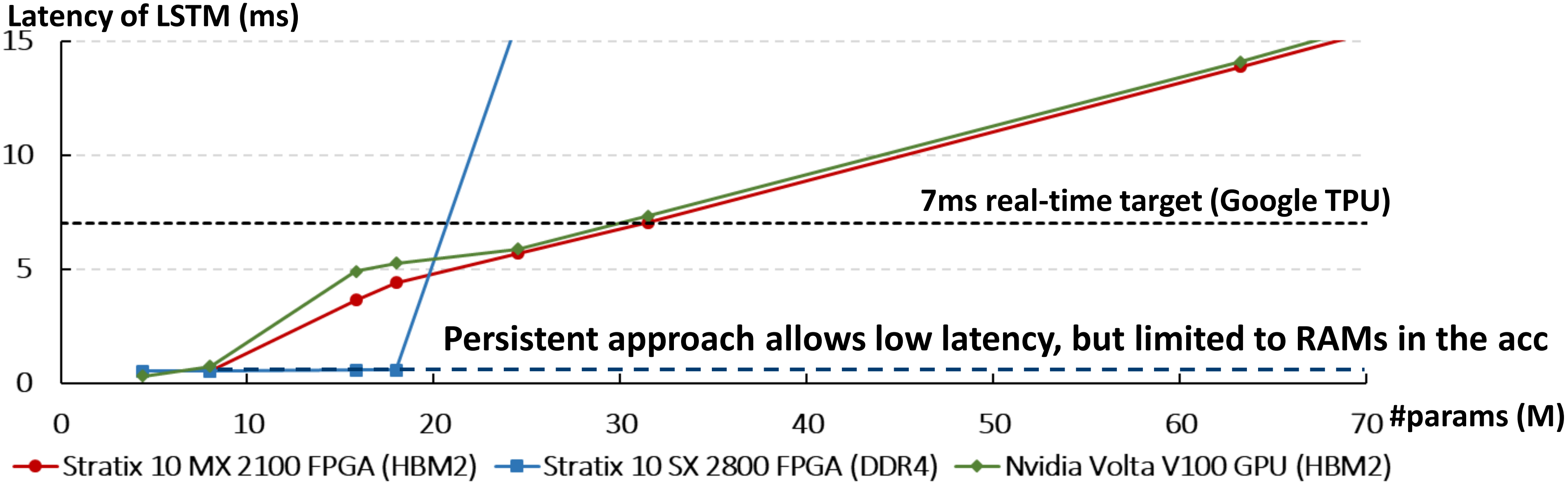
Brainwave-like NPU on S10 1100



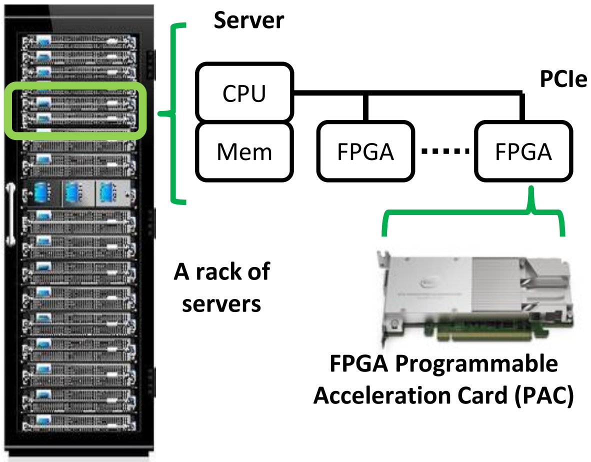
A small S10 1100 with 1 T-RAM (INT8) offers 16x better latency than GPU (FP32) and 34x energy efficiency accross RNN, GRU, LSTM workloads

How to keep low latency inference as AI models gets bigger?

- AI models gets better and **larger**
  - E.g., NMT use ~200M parameters, BERT use ~300M
- Single accelerator solution does not scale well



Persistent Model Across Multiple FPGAs in a Server

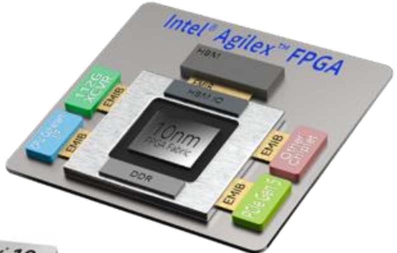


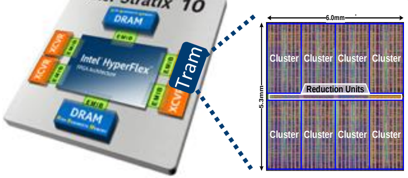
A rack of servers

FPGA Programmable Acceleration Card (PAC)

A10 1150 (~4MB BRAMs)

S10 2800 (~28MB BRAMs)





S10 110 + TensorRAM research chiplet [FCCM'19] (~80MB on FPGA+chiplet RAMs)

Dell Server R740 (Xeon CPU + multiple FPGA PCIe PACs)

Multiple generations of FPGA, with 1x to 8x FPGAs in a server

The FPGA for the Data-Centric World

PROCESS DATA

2ND GENERATION INTEL® HYPERFLEX™ FPGA ARCHITECTURE

UP TO 40% HIGHER PERFORMANCE[1,3]

UP TO 40% LOWER POWER[1,3]

UP TO 40 TFLOPS DSP PERFORMANCE[2,3]

STORE DATA

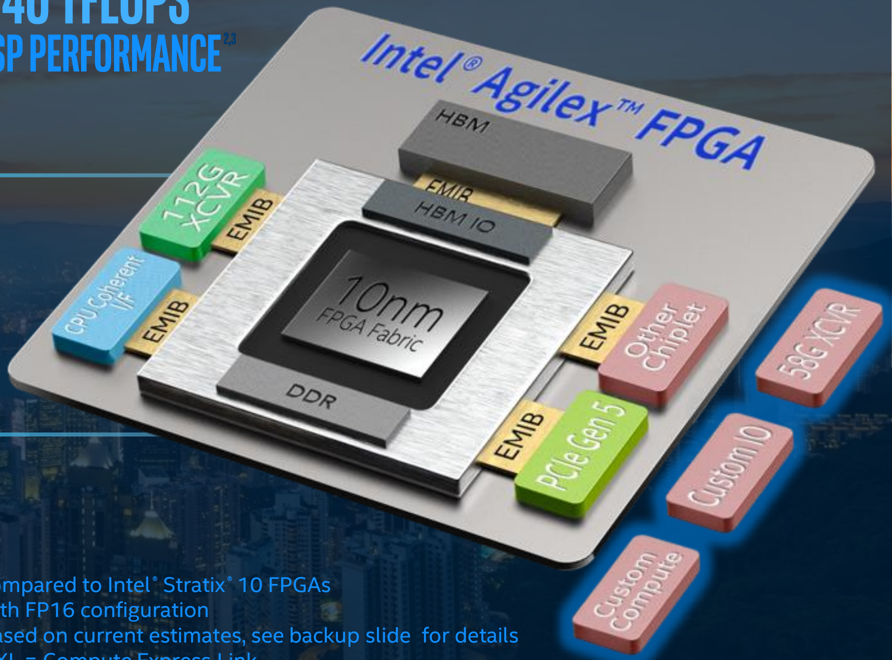
DDR5 & HBM

INTEL® OPTANE™ DC PERSISTENT MEMORY SUPPORT

MOVE DATA

INTEL® XEON® PROCESSOR COHERENT CONNECTIVITY (CXL*) & PCIE* GEN5

112G TRANSCEIVER DATA RATES



[1] compared to Intel® Stratix® 10 FPGAs
[2] with FP16 configuration
[3] Based on current estimates, see backup slide for details
CXL = Compute Express Link

Intel® Agilex™ FPGAs for the Data-Centric World

Intel® Agilex™ FPGAs enable transformative applications in edge computing, embedded, networking (5G/NFV), and data centers

Any-to-Any integration enables Intel to offer FPGAs with application-specific optimization and customization, delivering new levels of flexibility and agility

First FPGAs leveraging Intel's unmatched innovation:
10 nm process, 3D packaging, Intel Xeon® Scalable Processor cache coherency via CXL*, 112G transceivers, PCI Express* (PCIe*) Gen5, OneAPI, Intel eASIC™ devices, Intel Optane™ DC persistent memory support

intel®