XILINX

Advances and Trends in Computing Architectures for Deep Learning

Michaela Blott
Distinguished Engineer
Nov 2020

Background

- Xilinx
  - Fabless semiconductor company, founded in Silicon Valley in 1984
  - Today: ~4000 employees, $2.8B revenue
  - Invented the FPGA

- Xilinx Research - Dublin
  - Established almost 15 years ago
    - ~10 researchers plus university program
    - Highly active internship program, 80+ interns over the last 10years
  - Focus: FPGAs in Machine Learning
    - Building systems, architectural exploration, algorithmic optimizations, benchmarking
    - Quantifying the value of our devices in this space
  - In collaboration with partners, customers and universities

XILINX

What are FPGAs?

Customizable, Programmable Hardware Architectures

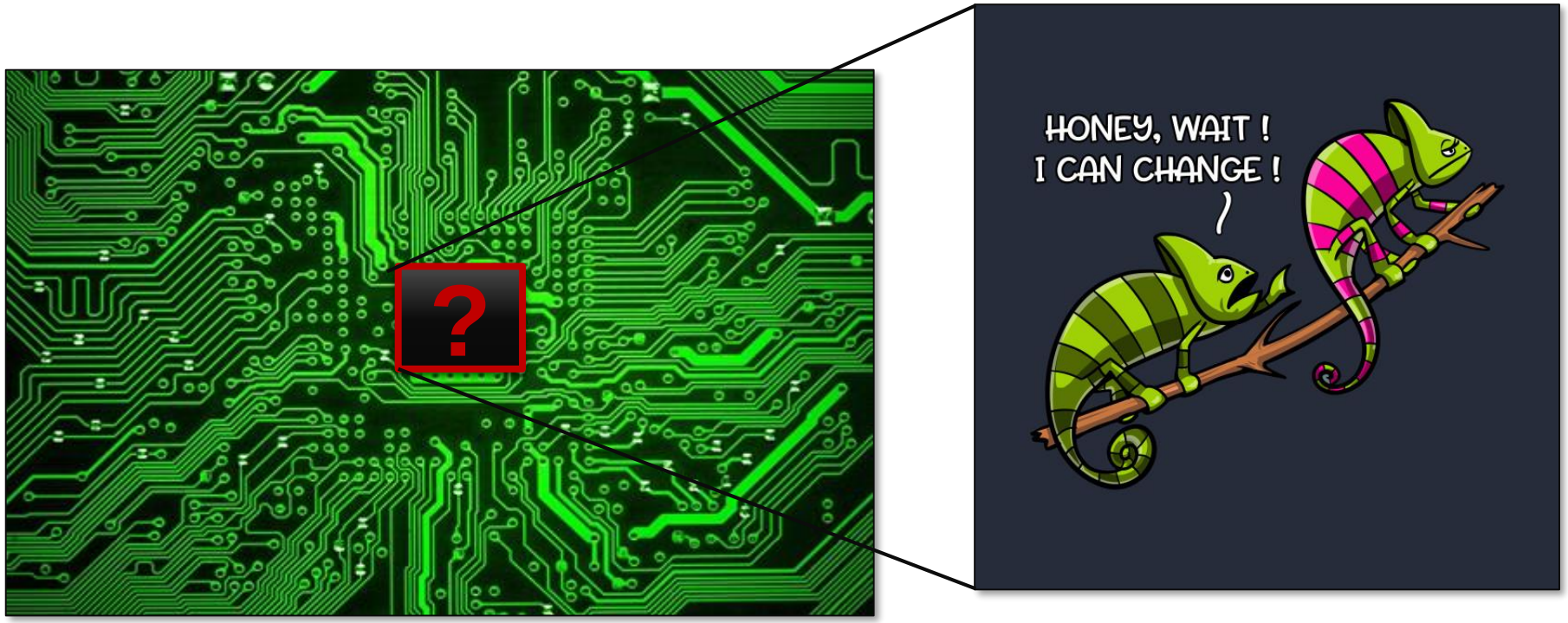
- The chameleon amongst the semiconductors…

  - Customizes IO interfaces, compute architectures, memory subsystems to meet the application
- Use case: Nothing else works, and you want to avoid ASIC implementation; or ASIC emulation

Non-standard IOs

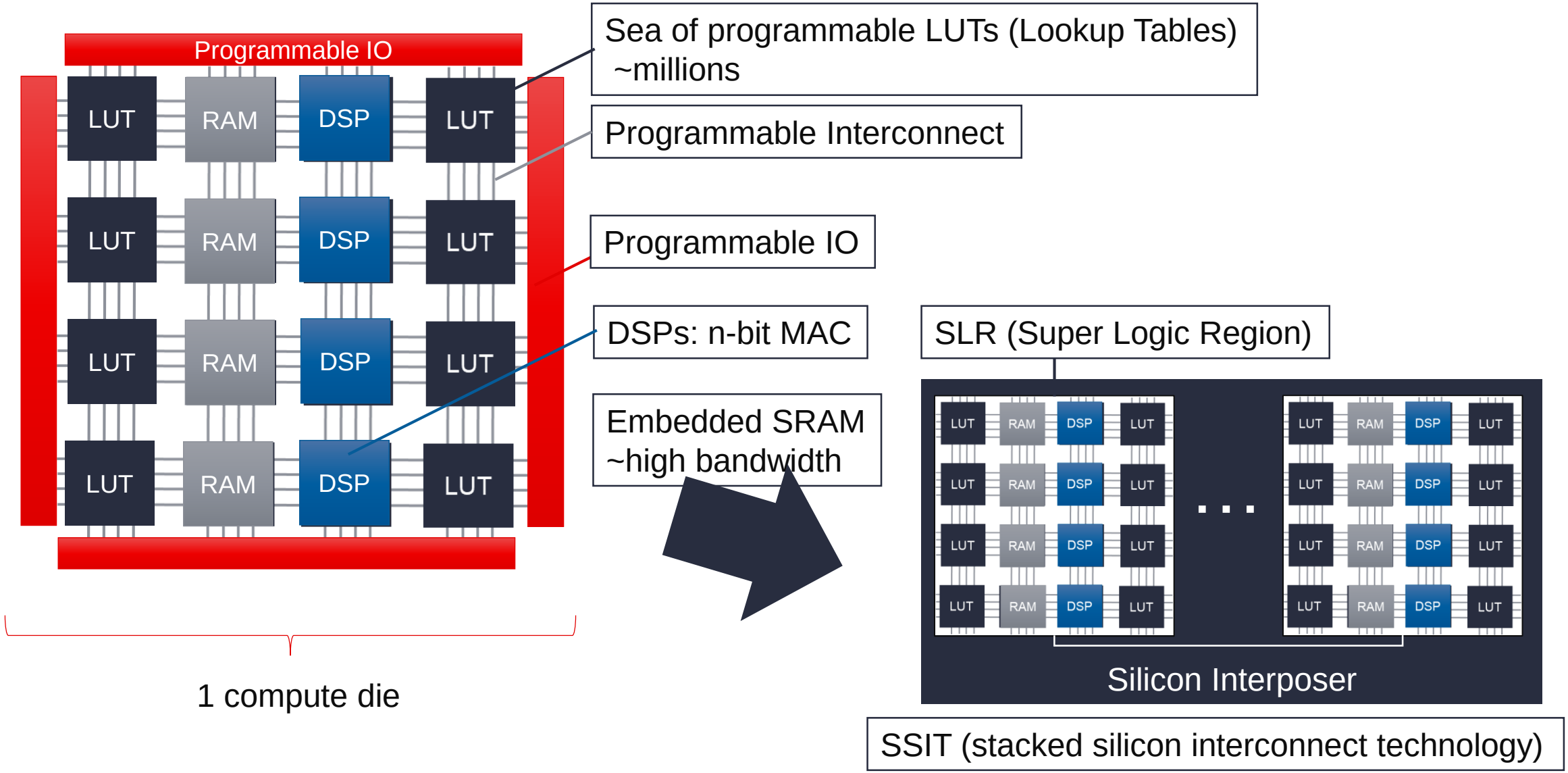
Different functionality?

Higher performance or efficiency metrics?



XILINX

What are FPGAs?

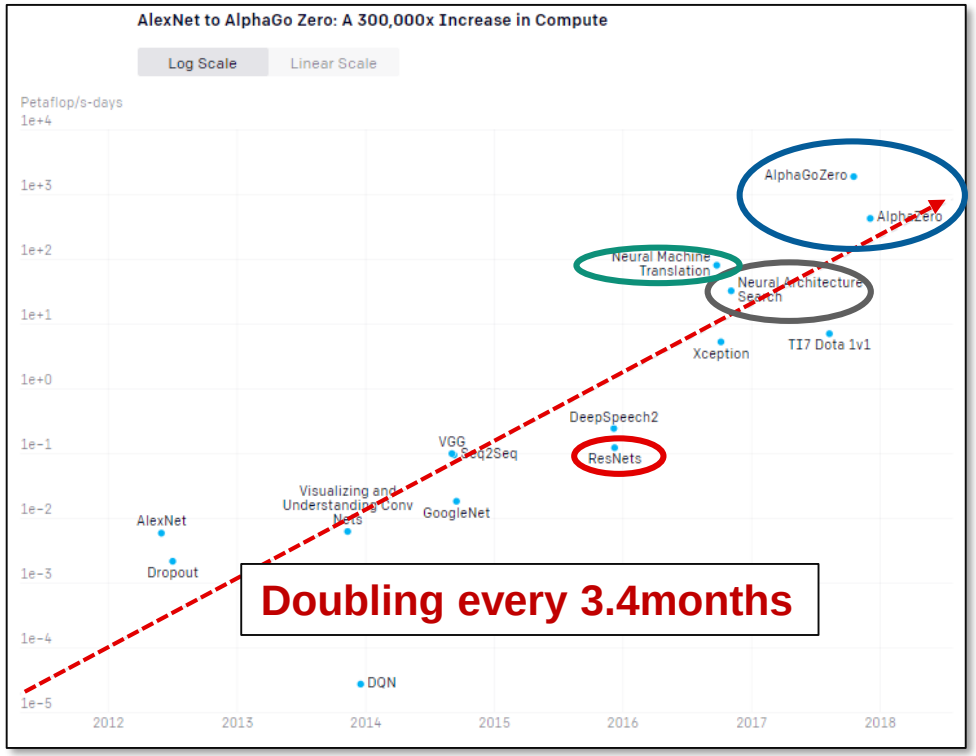


XILINX

Challenges in Deployment of Deep Learning

Growing Compute & Memory Requirements

- Compute & memory requirements are large and growing
  - Post-2012, compute has been doubling every 3.4 months
  - Rise of **language processing** (Alexa, Siri,..)
    - Turing-NLG: A 17-billion-parameter language model by Microsoft (Feb,2020)*
    - GPT-3: 175 billion-parameters by OpenAI
  - Fuelled by **Reinforcement Learning (RL)** and **Network Architecture Search (NAS)**



Source: https://blog.openai.com/ai-and-compute/

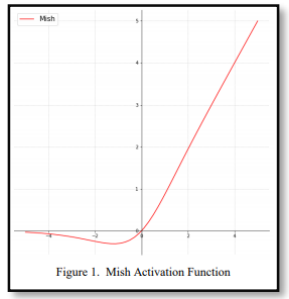
*Source: https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

XILINX

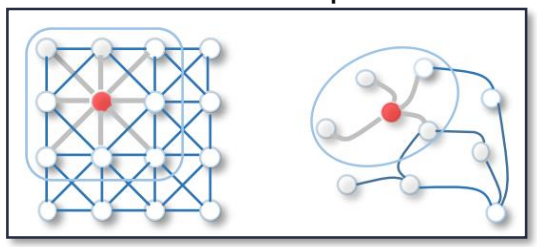
Algorithms are Changing

- New operators & layer types
  - Mish: A Self Regularized Non-Monotonic Neural Activation Function: https://arxiv.org/abs/1908.08681
- Runtime changes
  - Wang, X. et al. "Skipnet: Learning dynamic routing in convolutional networks." ECCV'2018.
- New (non-Euclidean) input data
  - For example point clouds, social network data, netlists, protein interaction networks
- Neural connectivity
  - Discovering neural wirings: https://arxiv.org/abs/1906.00586

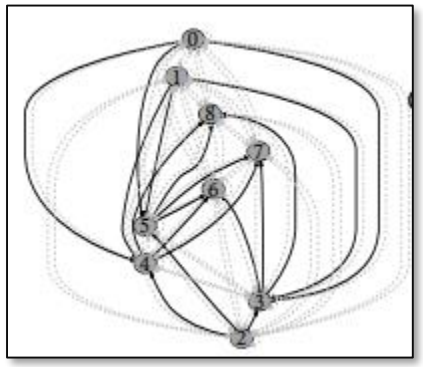
New activation functions



Non-Euclidean input data



Discover neural connectivity

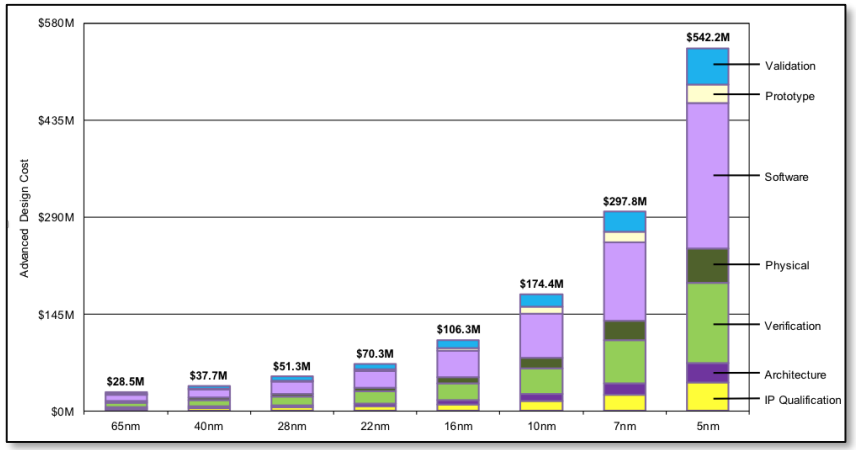


- Speed of change is increasing
  - Between 1998 and 2018, the volume of peer-reviewed AI papers has grown by more than 300%
  - Source: Artificial Intelligence Index Report 2019, Stanford

XILINX

Challenges in the Semiconductor Landscape

- Manufacturing difficulties of shrinking transistor sizes beyond 5nm
  - FINFET doesn't scale to 3nm
- Design costs are exploding
- Limited performance & power benefits with smaller technology nodes



Source: IBS

Hitting the physical limits of silicon-based computing

Moving away from standard van Neumann architectures

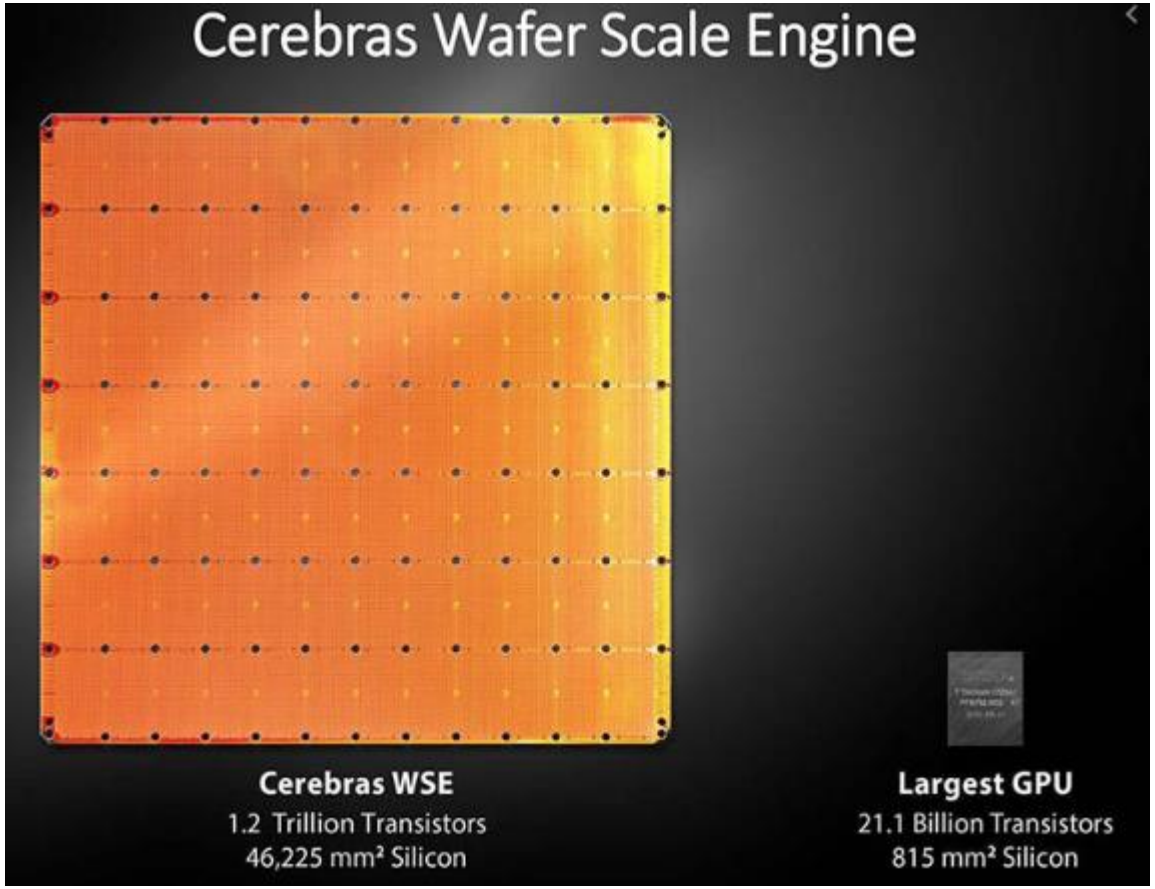
Architectural innovation becomes paramount

XILINX

Innovation is needed to provide the necessary performance scalability

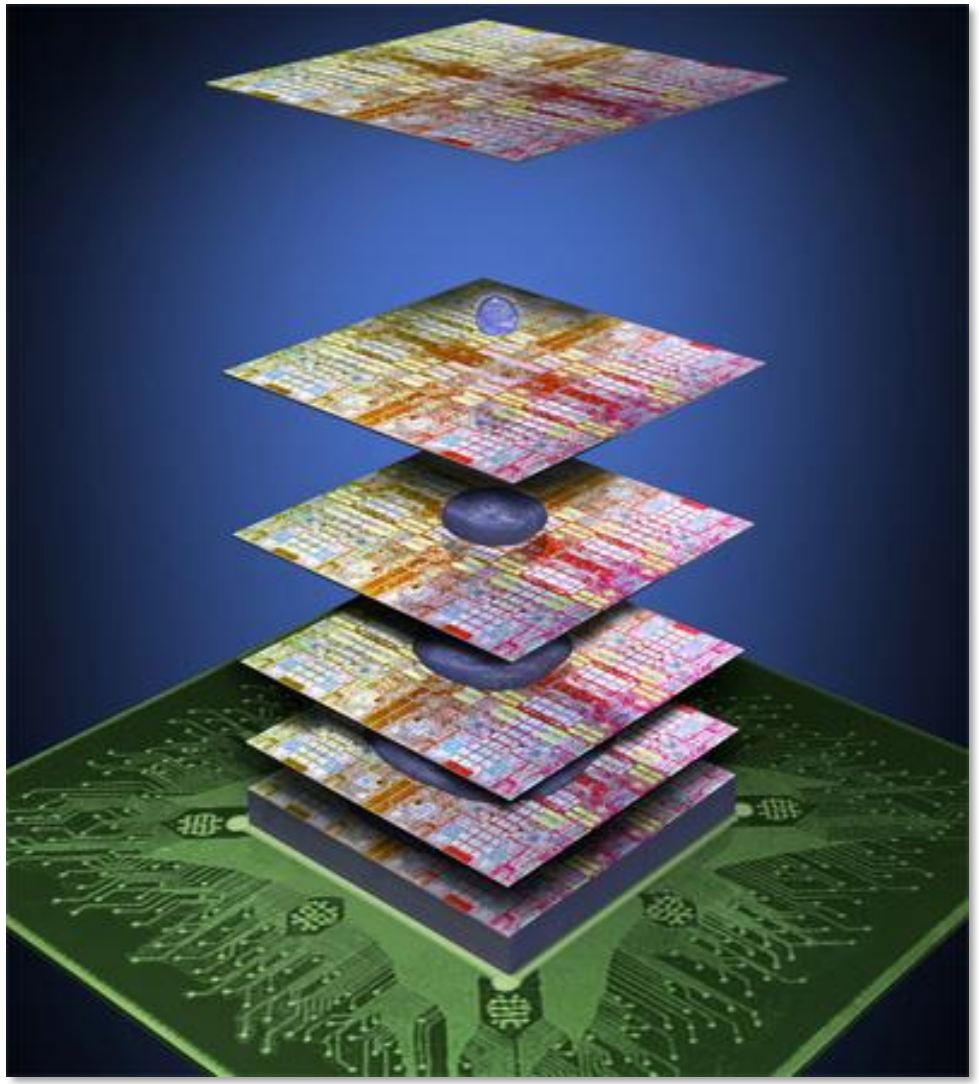
Innovative Approaches – Going Wide

- Cerebras: Waver-Scale Computing
  - Targeting ML training



Source: HotChips2019

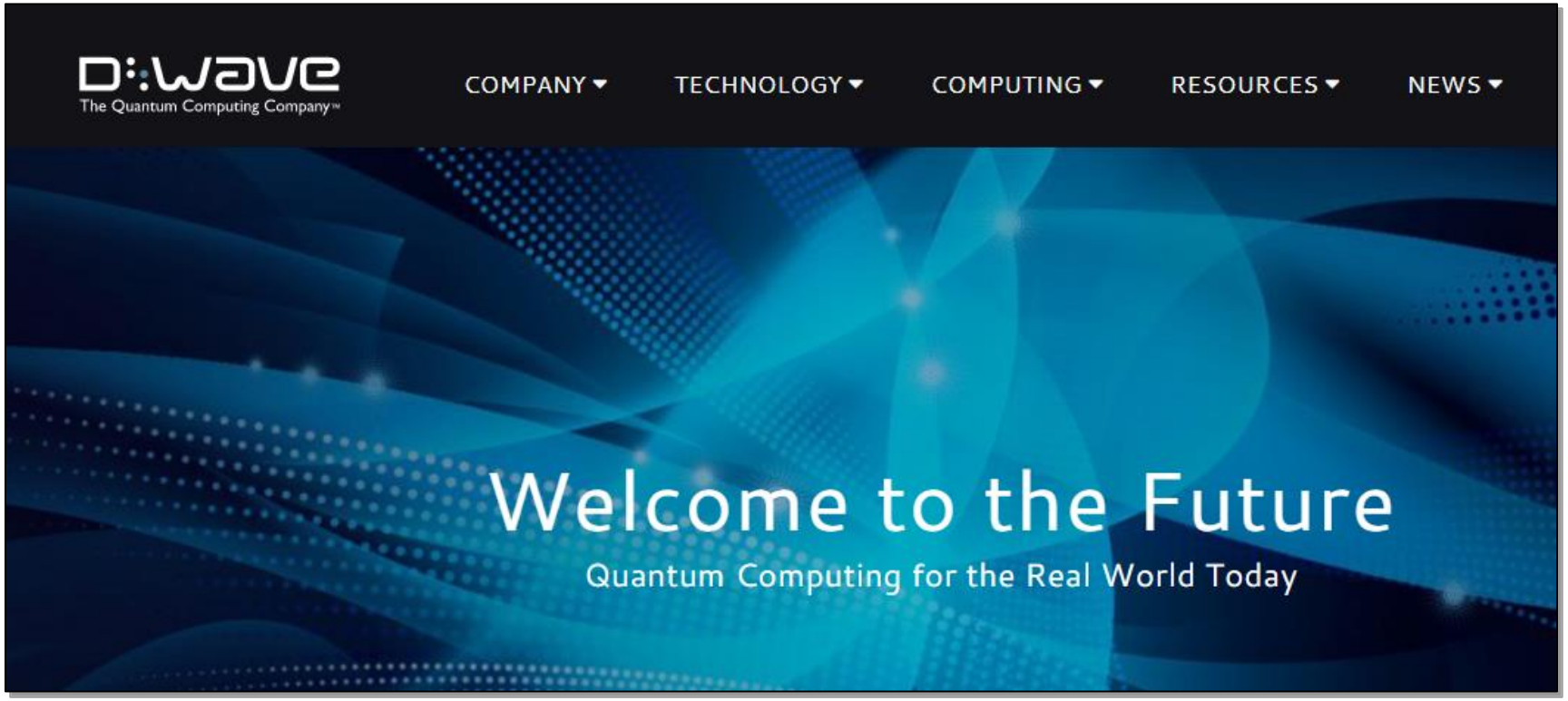
Innovative Approaches – Going High with 3D Die Stacking



XILINX

Innovative Approaches – Quantum Computing

- Dwave: Quantum Computing
  - For HPC and ML applications



XILINX

Innovative Approaches – Analog Neuromorphic Computing

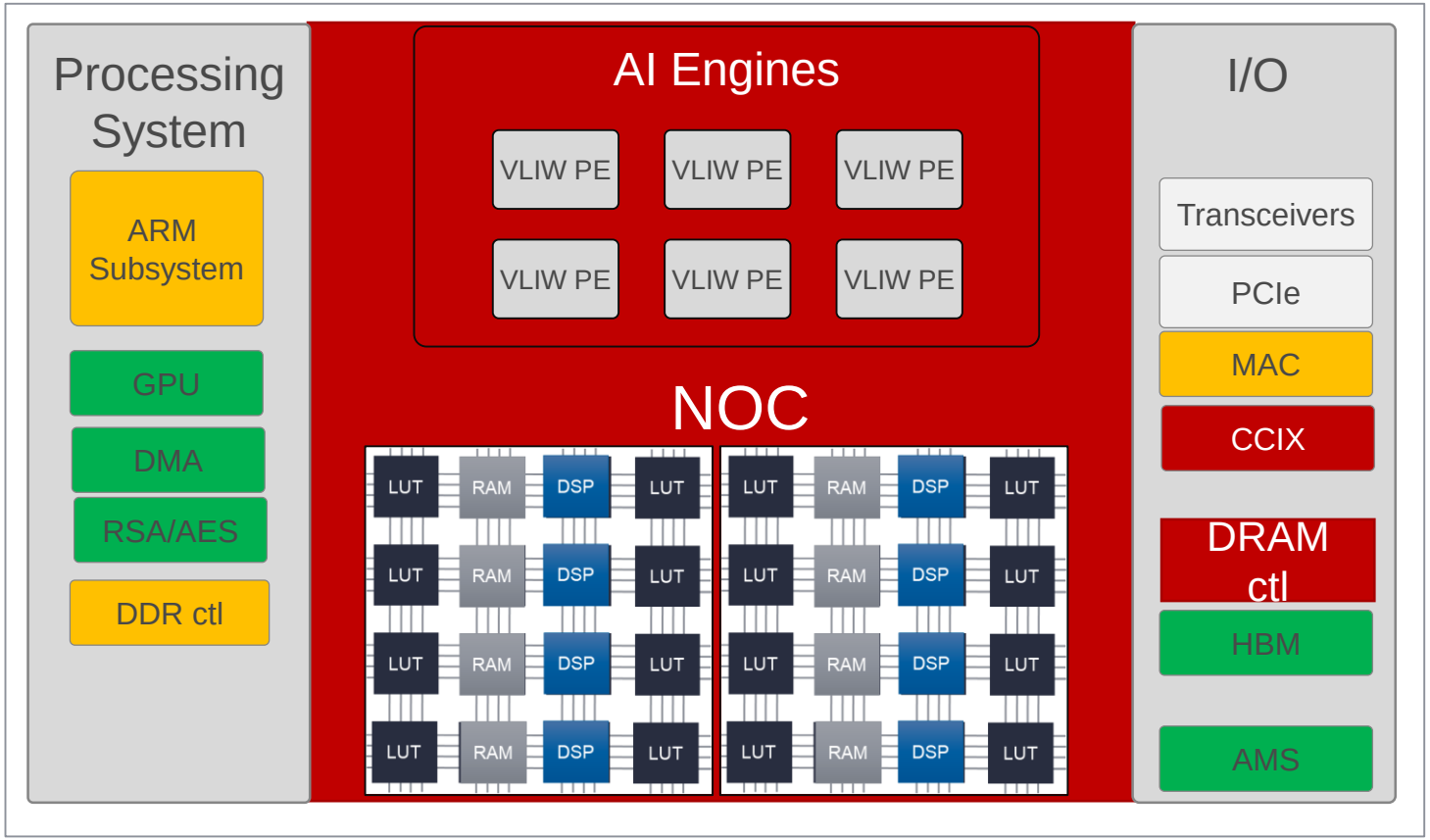
ASPINITY

XILINX

Performance Scalability through Specialization

Specialization = > Increasingly Heterogeneous

5 Series
7 Series
U+ Series
Versal

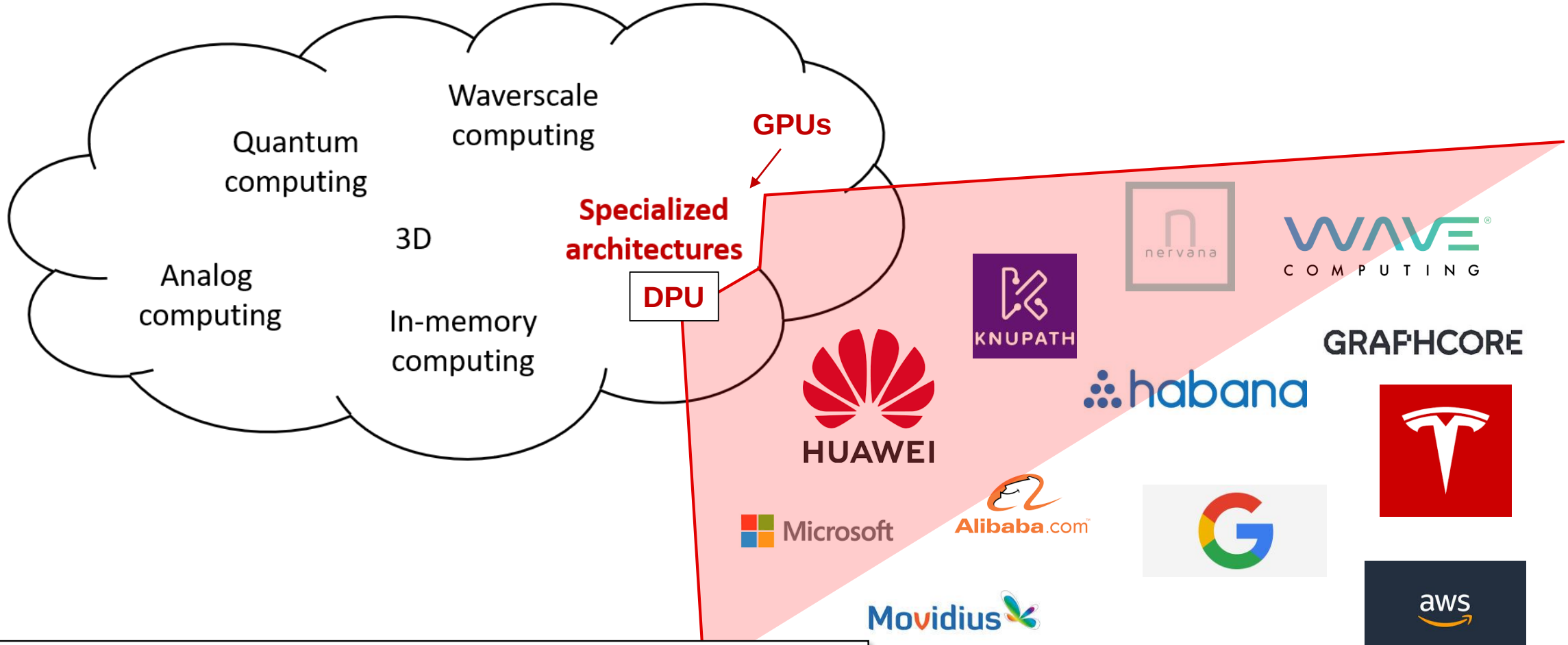


Xilinx Example:
FPGA -> ACAP

More hardened functionality (=> heterogeneous)
to improve compute density and save power

XILINX

Innovative Approaches



Specialized hardware architectures for ML workloads

XILINX

DPU Compute Architecture
Specialization, Performance & Flexibility

slower
flexible
generic

10k+ fps

10k-100m fps

100m+ fps

faster
more efficient
co-designed
specialized

Customize for DNN in general

Matrix of Processing Engines (MPE)

Customized for topologies

Spatial Processors (SP)

Topologies fully co-designed for hardware architecture

Circuit is the CNN

ASICs

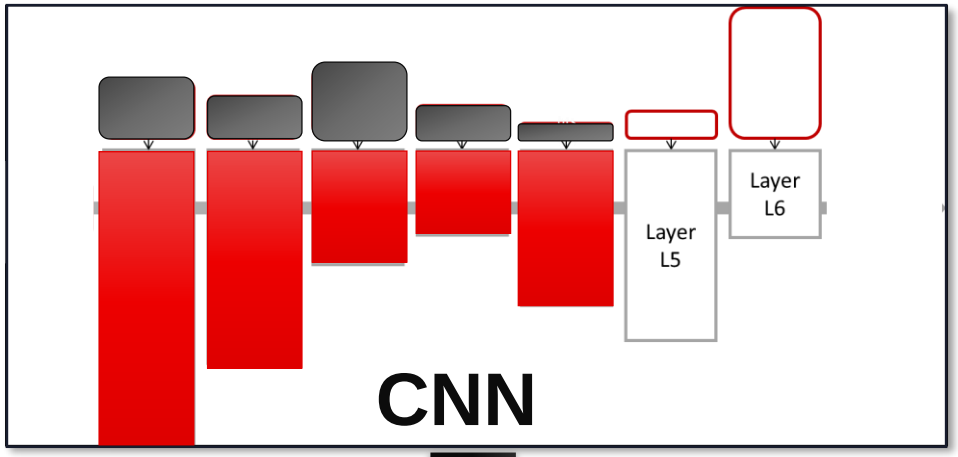
FPGAs

XILINX

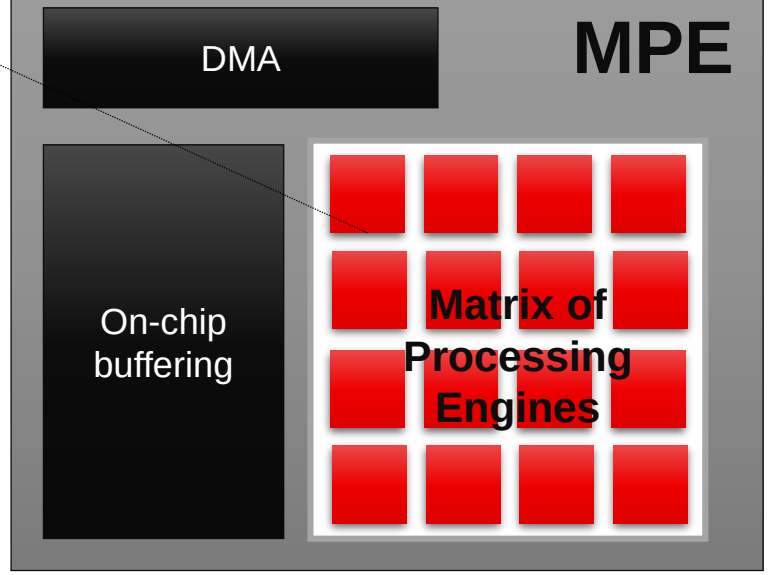
Matrix of Processing Engines Customizing for DNN in General

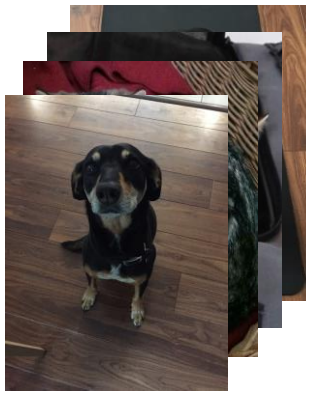
- Popular layer-by-layer compute
- Batching to achieve high compute efficiency

- Specialized operators
- ALU types
  - tensor-, matrix- or vector-based



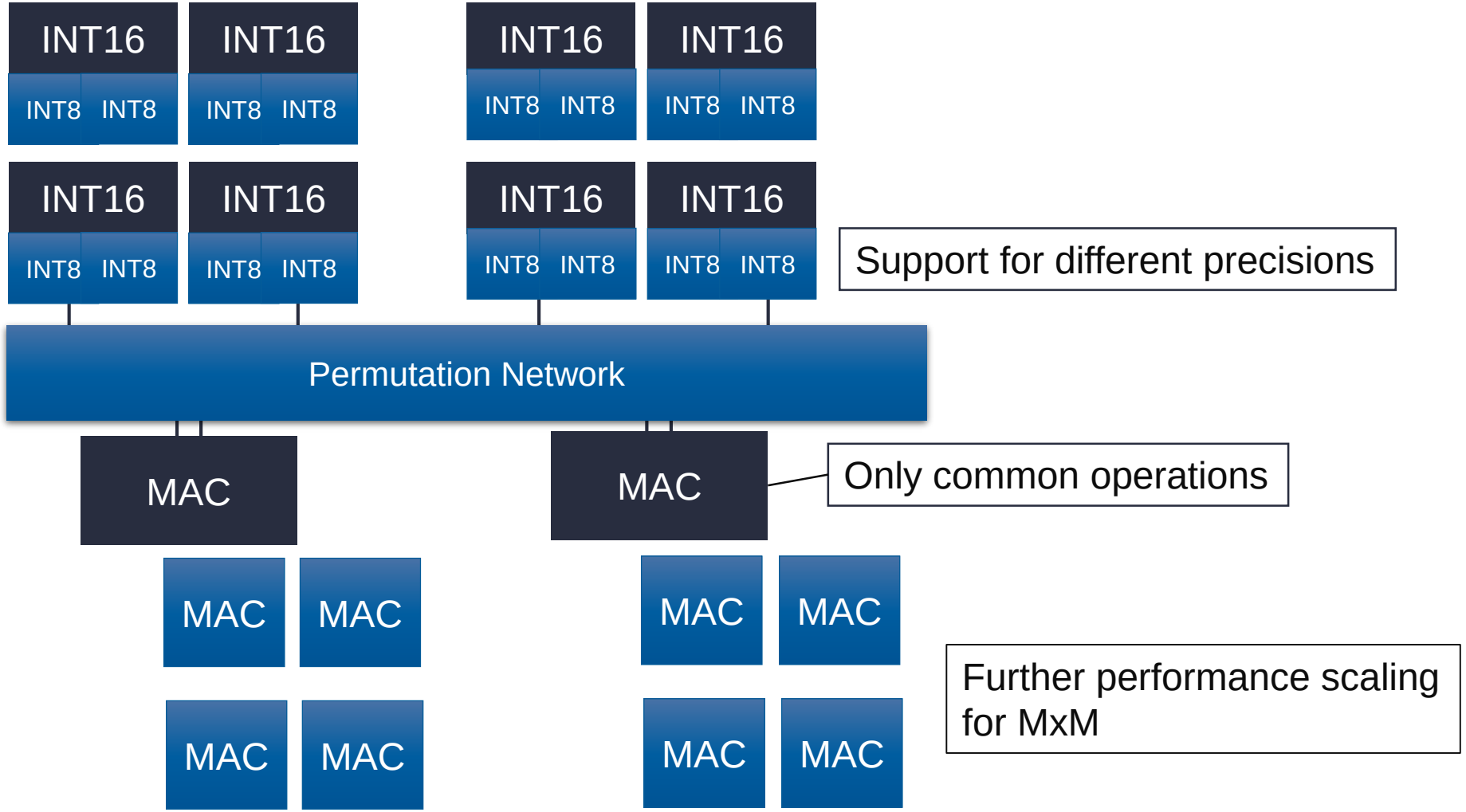
MAC, VLIW,
Vector Processor





XILINX

MPE: Specialization of Processing Engines



XILINX

DPU Compute Architecture
Specialization, Performance & Flexibility

slower
flexible
generic

10k+ fps

10k-100m fps

100m+ fps

faster
more efficient
co-designed
specialized

Customize for DNN in general

Matrix of Processing Engines (MPE)

Customized for topologies

Spatial Processors (SP)

Topologies fully co-designed for hardware architecture

Circuit is the CNN

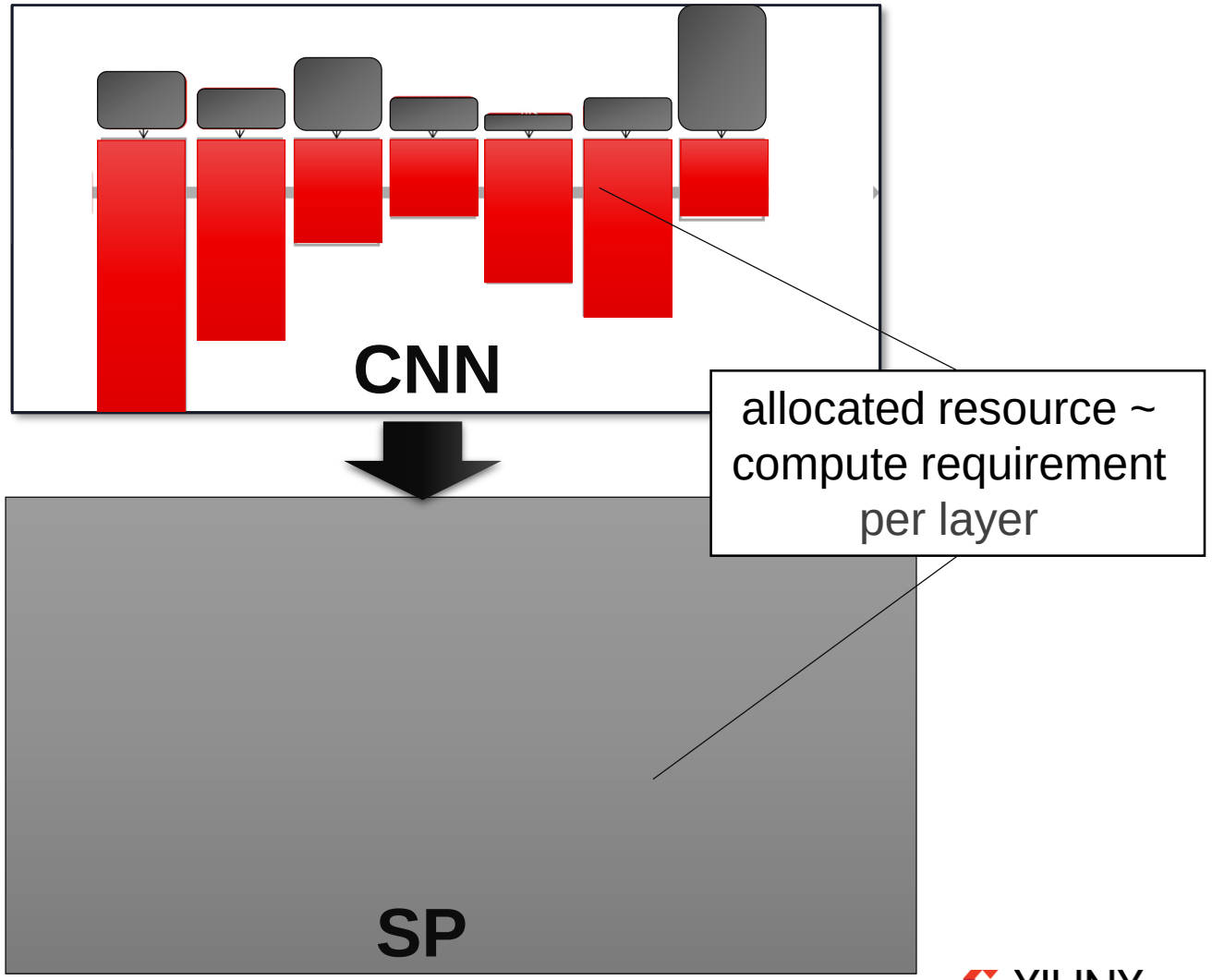
ASICs

FPGAs

XILINX

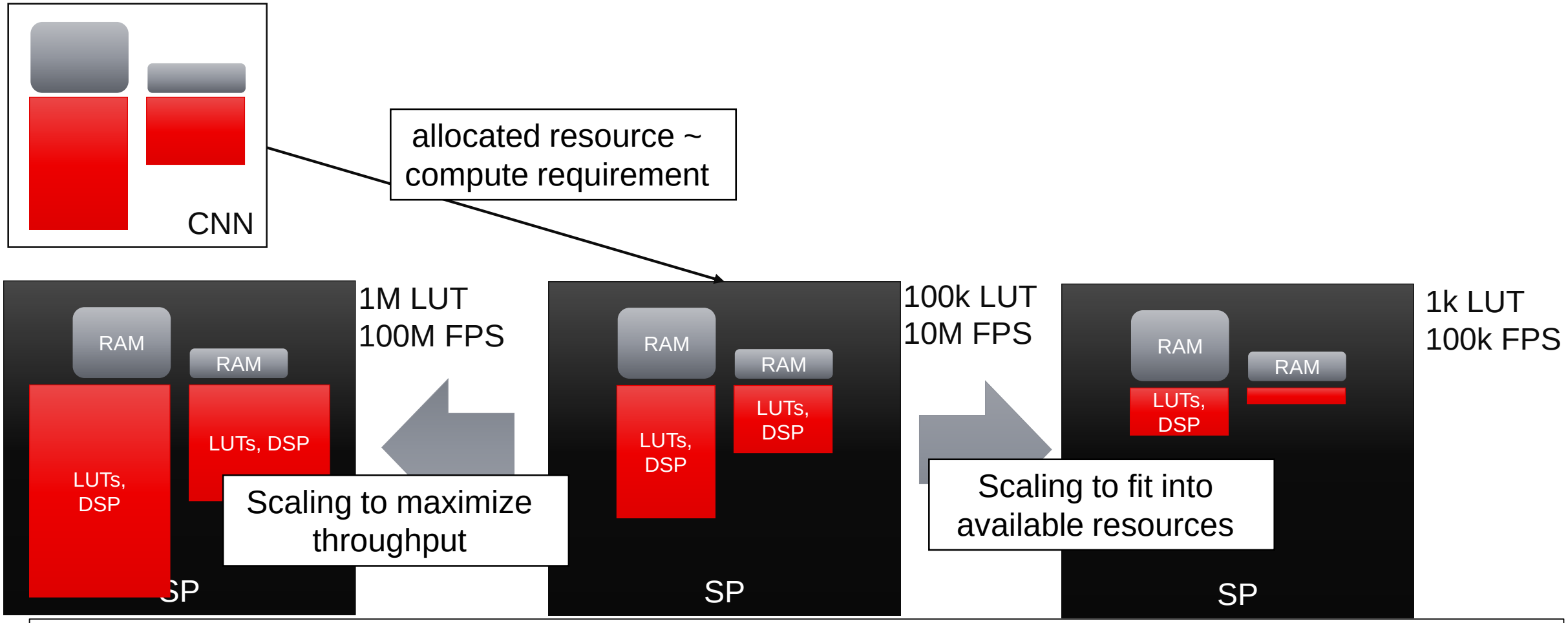
Spatial Processors: Customizing for Specific Topologies

- Hardware architecture mimics the topology
- Customize everything to the specifics of the CNN
- Benefits:
  - Improved efficiency
  - Lower latency
  - Higher throughput
- FPGAs rather than ASICs



XILINX

Spatial Architectures:
Scaling to Meet Performance & Resource Requirements



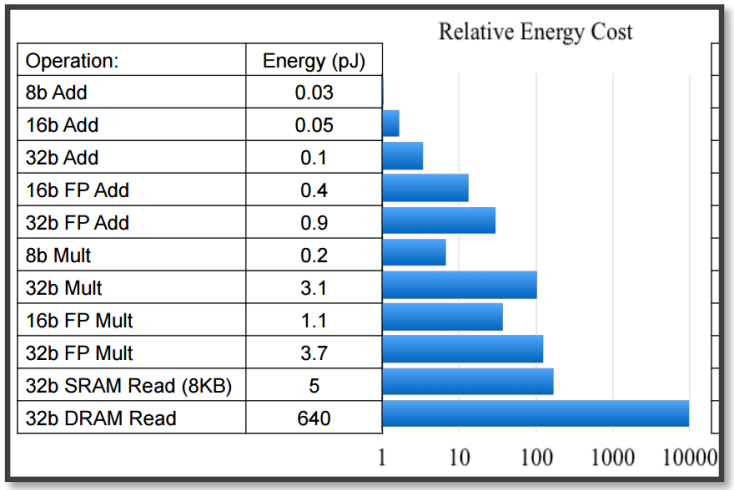
1. Scale performance & resources to meet the application requirements
2. If resources allow, we can completely unfold to create a circuit that inferences at clock speed

XILINX

Customizing Arithmetic

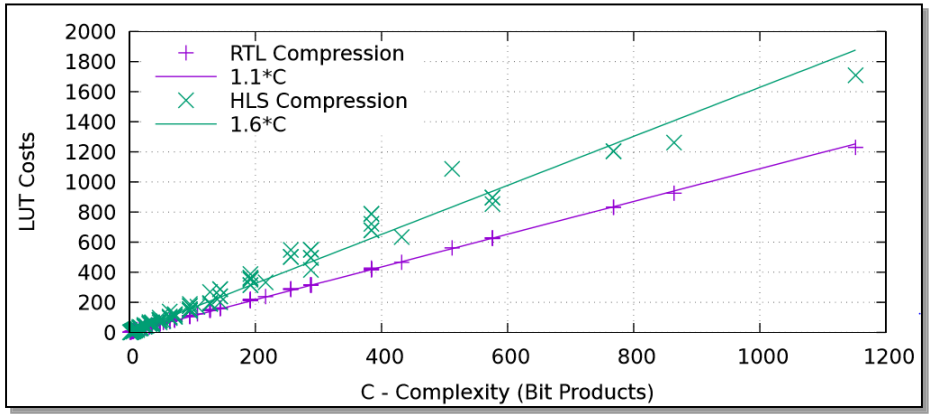
Customizing Arithmetic to Minimum Precision Required

- Shrinks hardware cost & scales performance
  - Instantiate ~**100x** more compute within the same fabric, thereby scale performance **100x**
- Reduces memory footprint
  - NN model can stay on-chip => no memory bottlenecks
- Inherently saves power

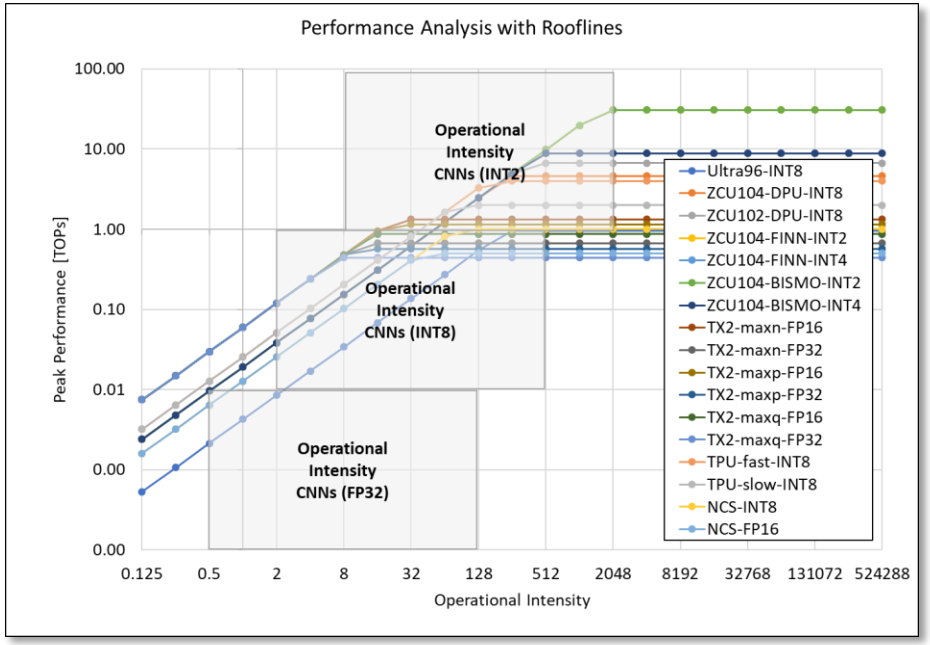


Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FP Mult	1.1
32b FP Mult	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640

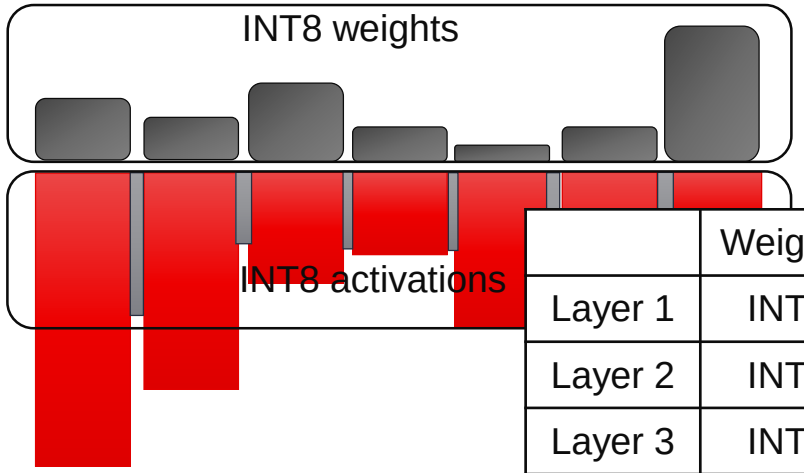
[Adapted from Horowitz. Computing's Energy Problem (and what we can do about it), ISSCC'14]



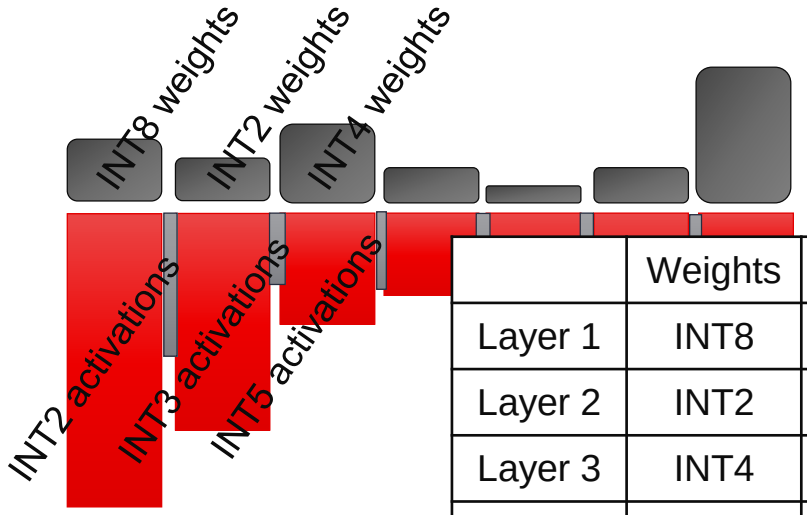
C= size of accumulator * size of weight * size of activation



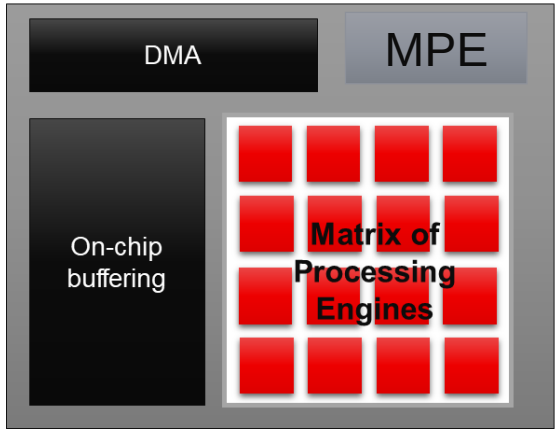
Granularity of Customizing Arithmetic

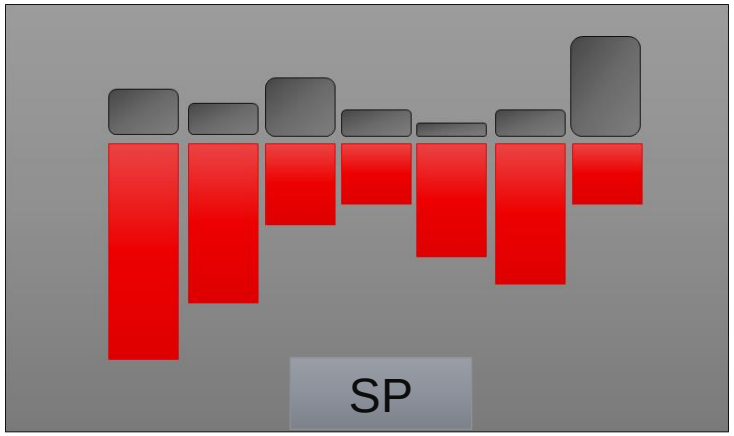


	Weights	Act.
Layer 1	INT8	INT8
Layer 2	INT8	INT8
Layer 3	INT8	INT8
…		



	Weights	Act.
Layer 1	INT8	INT2
Layer 2	INT2	INT3
Layer 3	INT4	INT5
…		

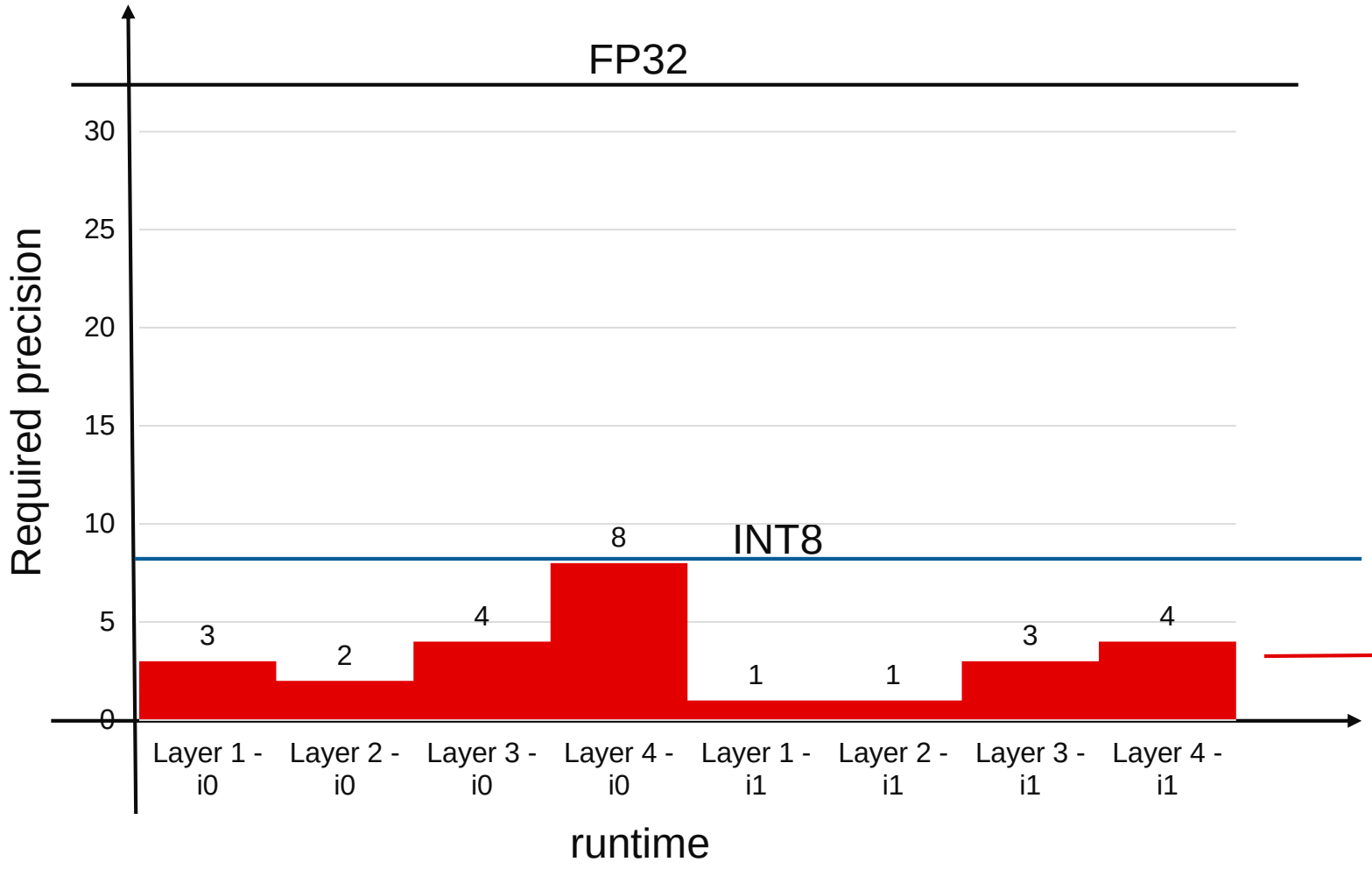




Spatial architectures enable fine granular customization of arithmetic

XILINX

Run-Time Programmable Precision



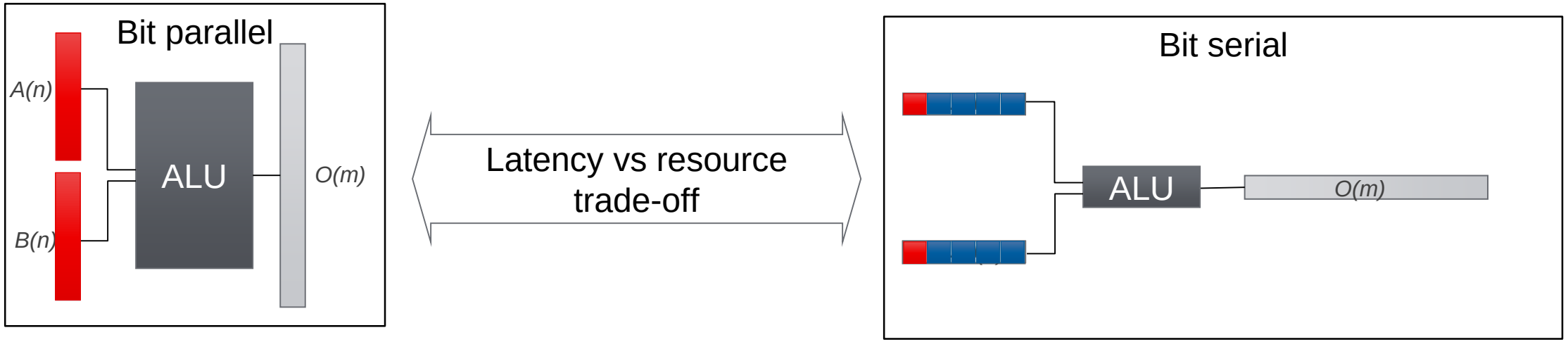
hardened customized arithmetic

runtime programmable precision (transprecision computing)

XILINX

Run-time Programmable Precision
with Bit-serial Architectures

Comparison to Traditional Bit-Parallel:



- Bit-serial: At run-time decide the precision at runtime – providing specialization without loosing generality
- Cost of Flexibility:
  - FPGA Evaluation for Matrix Multiply: almost no cost and provides equivalent bit-level performance at chip-level for low precision
  - System-level complexity

https://arxiv.org/pdf/1806.08862.pdf

XILINX

Extreme Specialization

DPU Compute Architecture
Specialization, Performance & Flexibility

slower
flexible
generic

10k+ fps

10k-100m fps

100m+ fps

faster
more efficient
co-designed
specialized

Customize for DNN in general

Customized for topologies

Topologies fully co-designed for hardware architecture

Matrix of Processing Engines (MPE)

Spatial Processors (SP)

Circuit is the CNN

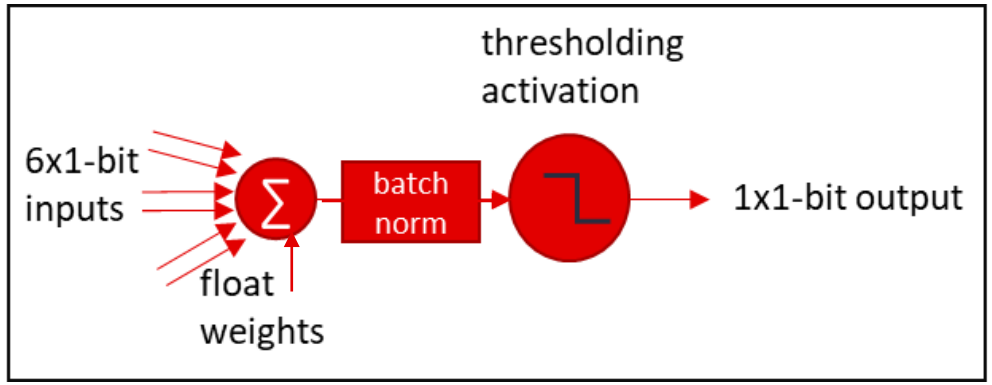
LogicNet

ASICs

FPGAs

XILINX

LogicNets with FPGAs



Traditional

Pick CNN topology → Train → Optimize hardware → Deploy

Adjust the parameters of CNN while iterating on training dataset until accuracy

LogicNets

Pick a circuit (=unrolled CNN) → Train → Deploy

Adjust the parameters of CNN (=LUT contents) while iterating on training dataset until accuracy

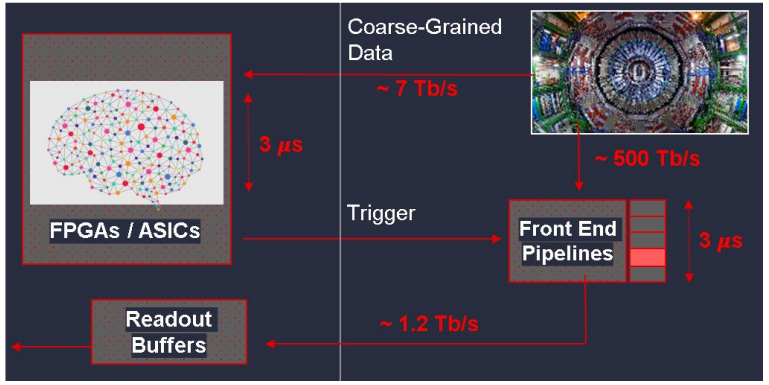
Source: https://arxiv.org/abs/2004.03021

- Maximum performance by design (classification at clock rate)

XILINX

LogicNets Key Results

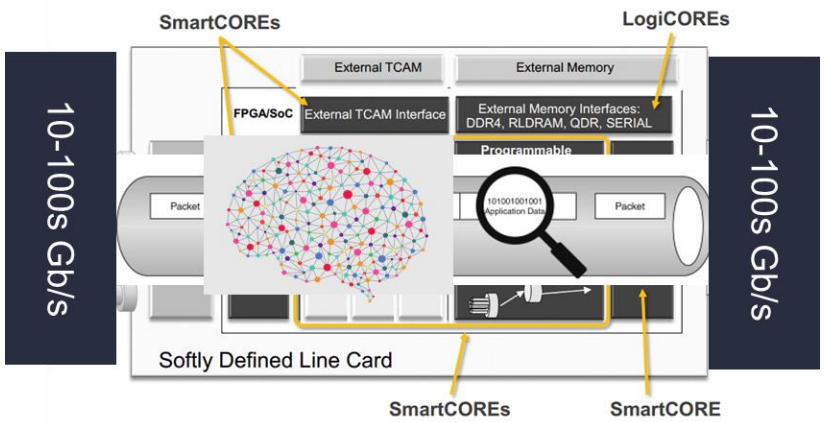
hls4ml JSC dataset [Duarte et al.]



Jet Tagging (CERN LHC)

~**72%** accuracy
using ~**38k** LUTs
at **427 M** samples / second
with **13 ns** latency

UNSW-NB15 Network Intrusion Detection dataset [Moustafa et al.]



Network Intrusion Detection

~**91%** accuracy
using ~**16k** LUTs
at **471 M** samples / second
with **9 ns** latency

Open source release upcoming as part of FINN (Q1 2021)

Preprint: https://arxiv.org/pdf/2004.03021
Video: https://youtu.be/qCyK5v84jpI

XILINX

Challenge

How can we enable a broader spectrum of end-users to be able to specialize hardware architectures and co-design solutions?

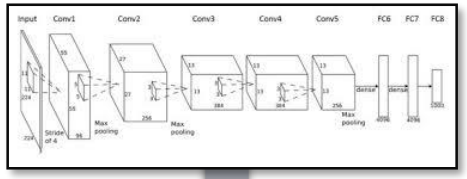
Project Mission

- Providing tools and platforms for exploration of CNN compute architectures

- End-to-end flow
  - ML engineers can create specialized hardware architectures on an FPGA
    - with spatial architectures and custom precision

- Open source
  - Transparency and flexibility for the fast changing landscape of algorithms
    - if not supported, you can add your own

From CNN to FPGA Deployment



Brevitas
Training in pytorch
Algorithmic optimizations

- Train or even learn reduced precision CNNs
- Library of standard layers
- Pretrained examples

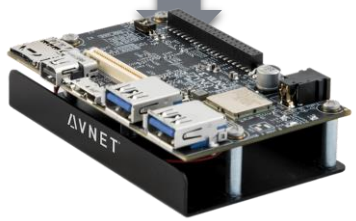
ONNX Intermediate Representation

FINN compiler
Specializations of hardware architecture

- Perform optimizations
- Map to Vivado HLS
- Create CNN hardware IP

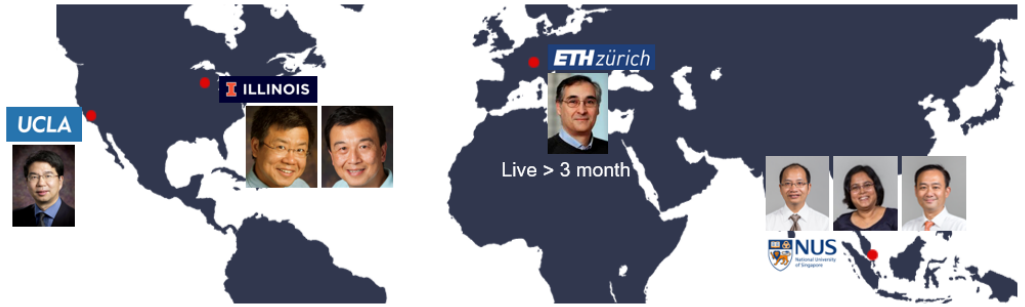
Deployment with PYNQ

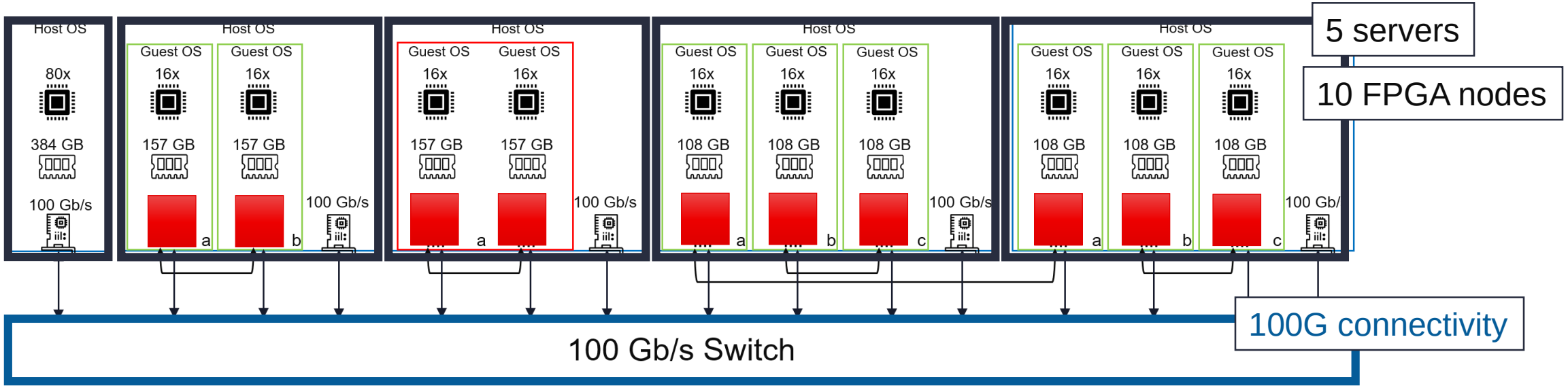
- Embeds the CNN IP into an infrastructure design
- Generates Python run-time (based on PYNQ)
- Enables integration with your application
- Works on embedded and Alveo platforms



Infrastructure for Experimentation

- Xilinx academic compute clusters
  - 4 centres world-wide
  - Free to use
  - Enabling research community
- Not only for FINN

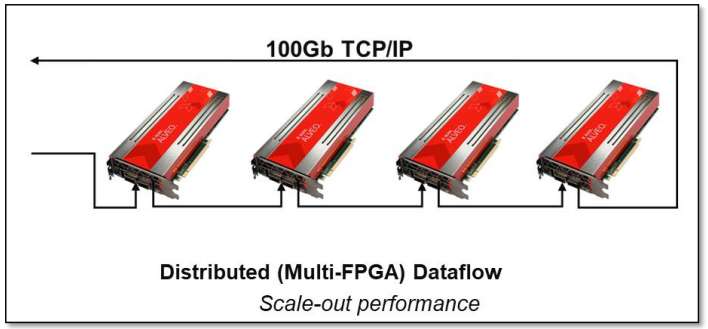




XILINX

FINN Status

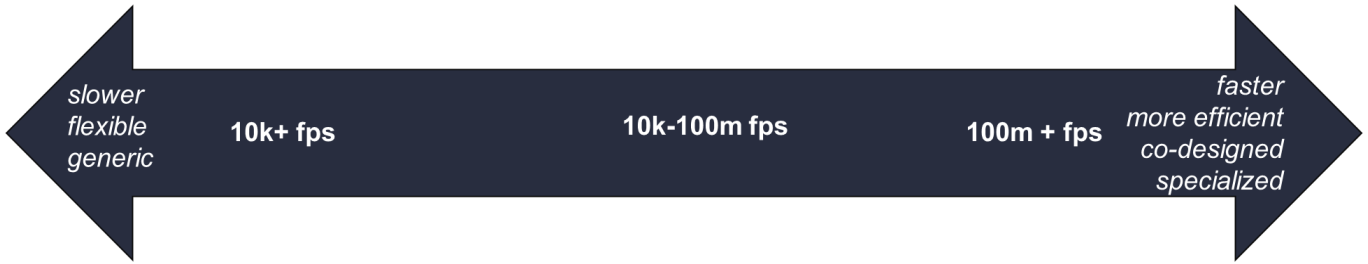
- Ongoing development
  - Support for residual topologies, depthwise convolutions
  - ResNet-50, MobileNet-v1 imminent
  - Multinode deployment on XACC
- Looking to build-up a community
  - Many student, hobbyist, and school projects
  - University classes with FINN @ Stanford, Charlotte, NTNU
    - Online material in preparation
  - Industrial applications
- Looking to create differentiating application portfolio
  - Extreme throughput (100M+ fps) ultra-low latency



If you're interested, we'd love to hear from you ☺

Summary

Summary – Future Work



- Spectrum of innovative architectures emerge to address compute and memory requirements in CNNs
- Specialization of hardware architecture are critical to making this a reality
- With more flexibility, more opportunity to customization
  - FPGAs allow to specialize to the specifics of individual use cases while providing generality
  - Tools such as FINN are needed to overcome complexity in the design entry and make technology accessible
- Please be in touch, if you're interested in collaborating ☺

XILINX

Thank You

More information can be found at:
https://xilinx.github.io/finn