



Scaling DNN Inference for Extreme Throughput

Michaela Blott
Distinguished Engineer, Xilinx Research
Dec 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



What are FPGAs?

Customizable, Programmable Hardware Architectures

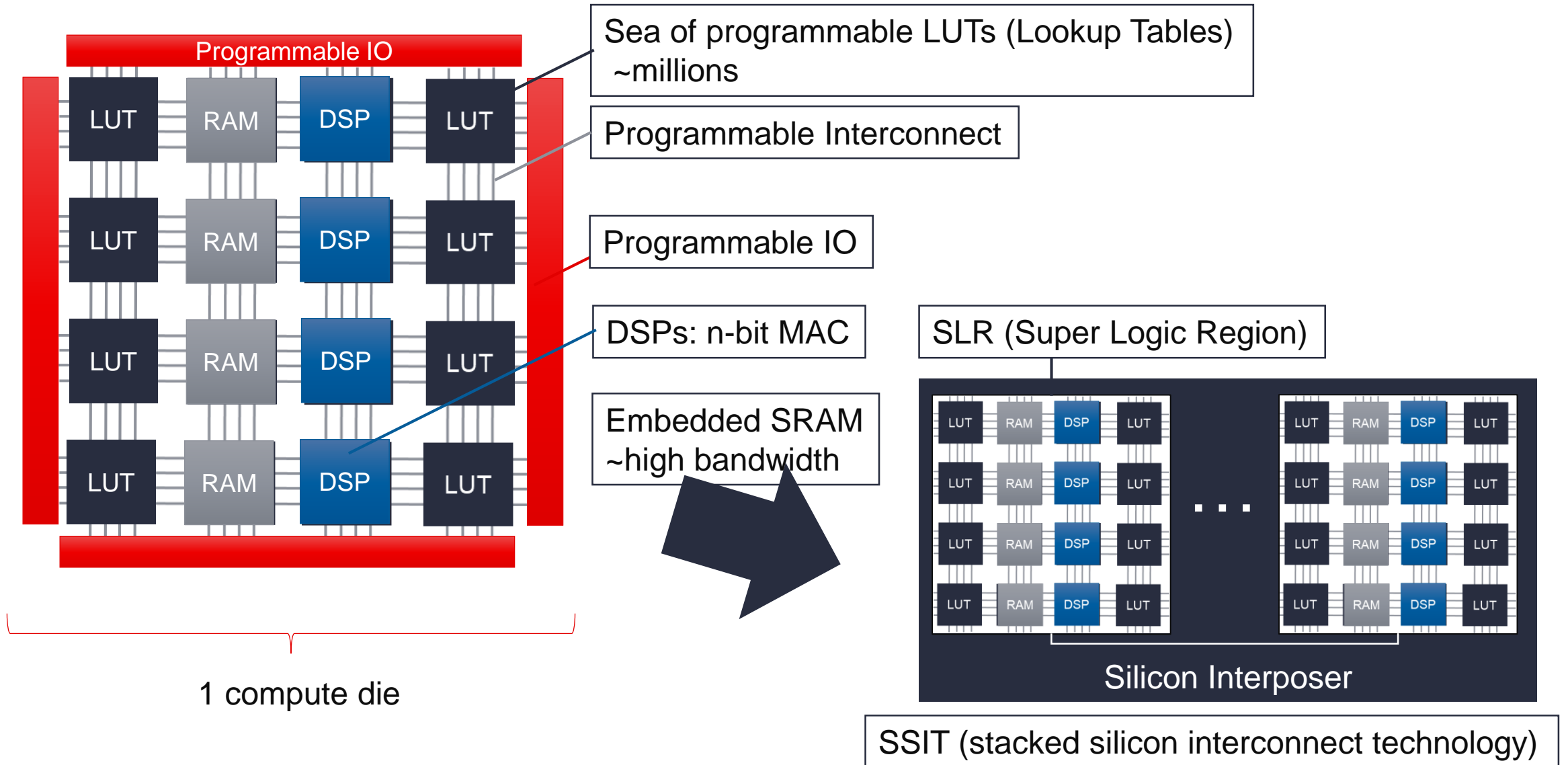
- ▶ The **chameleon** amongst the semiconductors...
 - Customizes IO interfaces, compute architectures, memory subsystems to meet the application
- ▶ **Use case:** Nothing else works, and you want to avoid ASIC implementation; or ASIC emulation

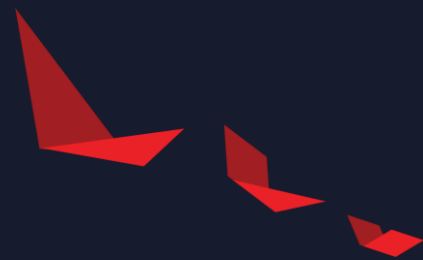


- Non-standard IOs →
- Different functionality? →
- Higher performance or efficiency metrics? →





What are FPGAs?

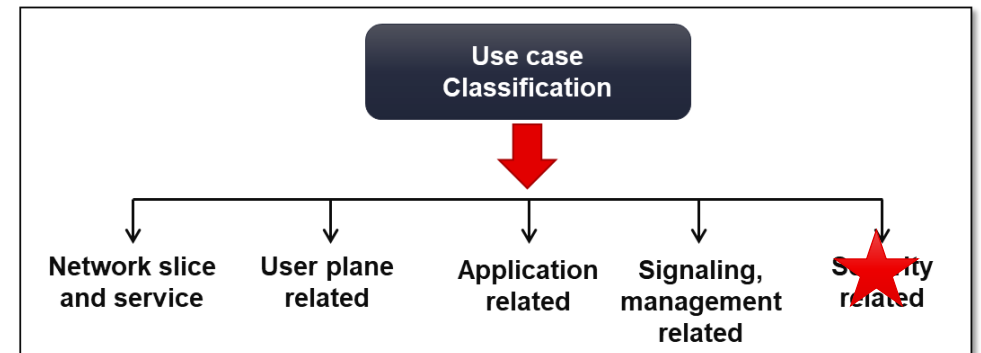




Challenges in Deploying DNNs in Communications

DNNs in Communications

- ▶ Many emerging use cases
 - Traffic classification
 - Traffic monitoring and statistics
 - Traffic prediction
 - Network intrusion detection  
 - Physical layers
- ▶ Implementation of individual basic components
 - Hashing/ indexing
 - Sorting
- ▶ ITU has identified and classified 30 use cases
 - ITU-T Y.3170-series Supp 55



[1] https://www.cl.cam.ac.uk/~ey204/teaching/ACS/R244_2018_2019/papers/Kraska_SIGMOD_2018.pdf

[2] http://learningsys.org/sosp19/assets/papers/22_CameraReadySubmission_Abstract_SOSP_19_ML_Sys_workshop-4.pdf

[3] <https://aip.scitation.org/doi/full/10.1063/1.5140609>

[4] <https://hal.archives-ouvertes.fr/tel-01206266/document>

[5] <https://tel.archives-ouvertes.fr/tel-01876701/document>

[6] <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8054694>

Specific Challenges for DNNs in Communication: Throughput

- ▶ Extreme high throughput requirements
- ▶ Highest reported inference performance
 - 55kfps ResNet50 (423TOP/sec) MLPerf [1]
- ▶ Even a 1MOP/inference model would require
 - **600TOP/s** for 400Gbps
 - 30TOP/s for 20Gbps

Throughput requirements are extremely high

- Beyond the limit of latest AI silicon
- Limiting complexity of DNNs

	Throughput
5G (20Gbps)	30MRps
100Gbps	150MRps
400Gbps	600MRps

MRps: Million Requests per Second
Assuming 64B / packet

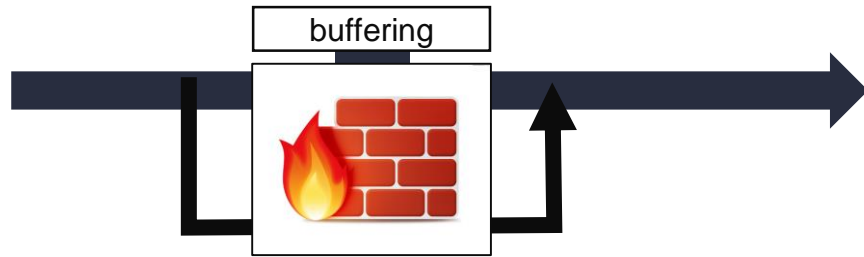
Datasheet performance of
State of the Art AI accelerators:

	Performance
Ascend 910	256TOP/s *
Colossus (Graphcore)	250TOP/s *
A100 (Nvidia)	312TOP/s * 1248TOP/s **

*BF/FP16
**INT8

Specific Challenges for DNNs in Communication: Latency

- ▶ Ultra low latency requirements in any form of cognition cycle:
 - Translates to buffering requirements



	Buffer [10ns]	Buffer [1us]	Buffer [1msec]
5G (20Gbps)	0.2Kb	20Kb	20Mb
100Gbps	1Kb	100Kb	100Mb
400Gbps	4Kb	400Kb	400Mb

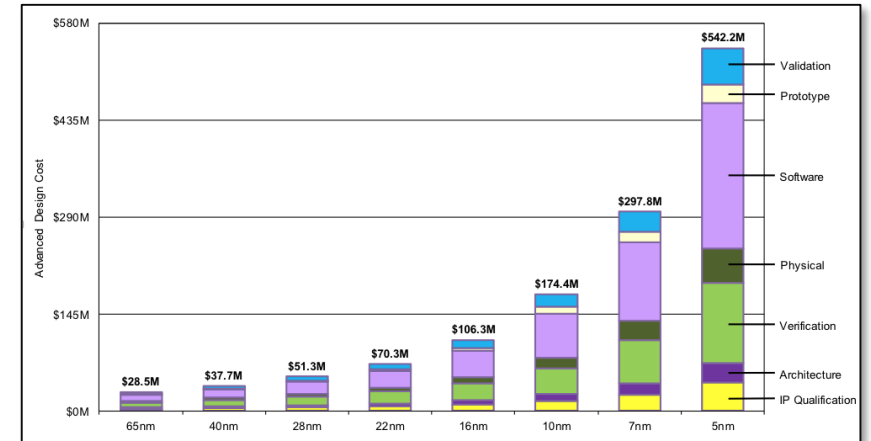
- ▶ Typical latency in MLPerf
 - Closed data center, single stream latency: 2-7msec [1]



[1] <https://mlperf.org/inference-results-0-7/>

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

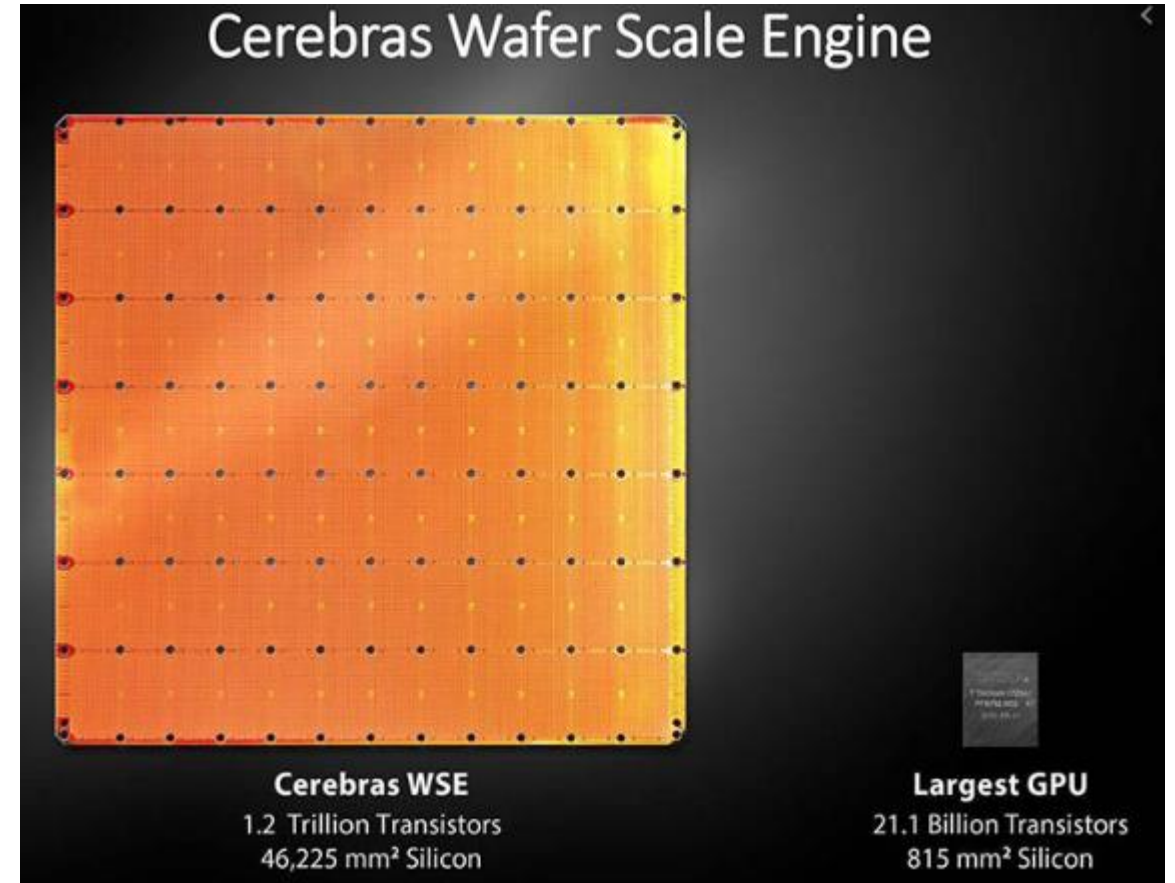
Architectural innovation becomes paramount



**Innovation is needed to provide the
necessary performance scalability**

Innovative Approaches – Going Wide

- ▶ Cerebras: Wafer-Scale Computing
 - Targeting ML training



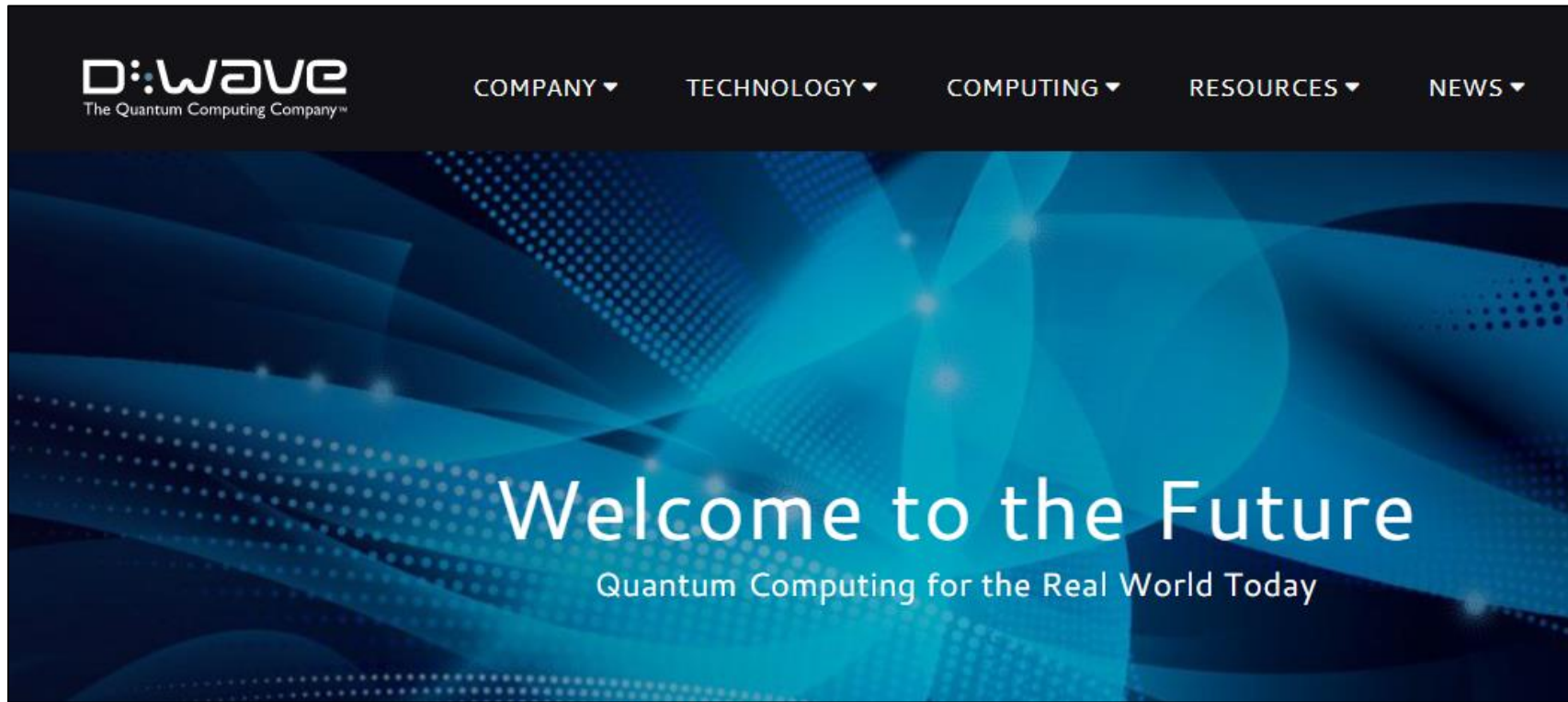
Source: HotChips2019

Innovative Approaches – Going High with 3D Die Stacking



Innovative Approaches – Quantum Computing

- ▶ Dwave: Quantum Computing
 - For HPC and ML applications



Innovative Approaches – Analog Neuromorphic Computing

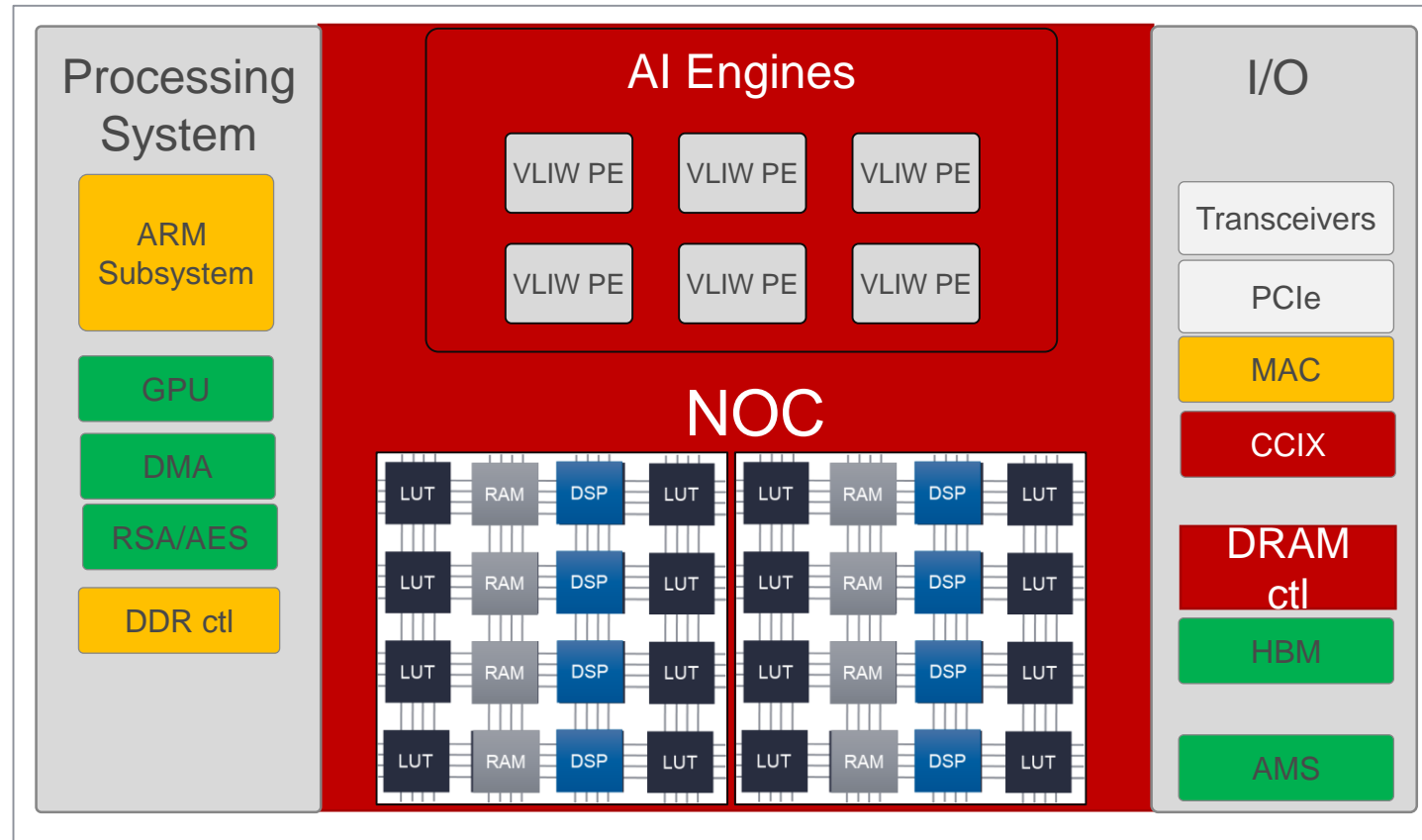


Performance Scalability through Specialization



Specialization for Performance Scalability

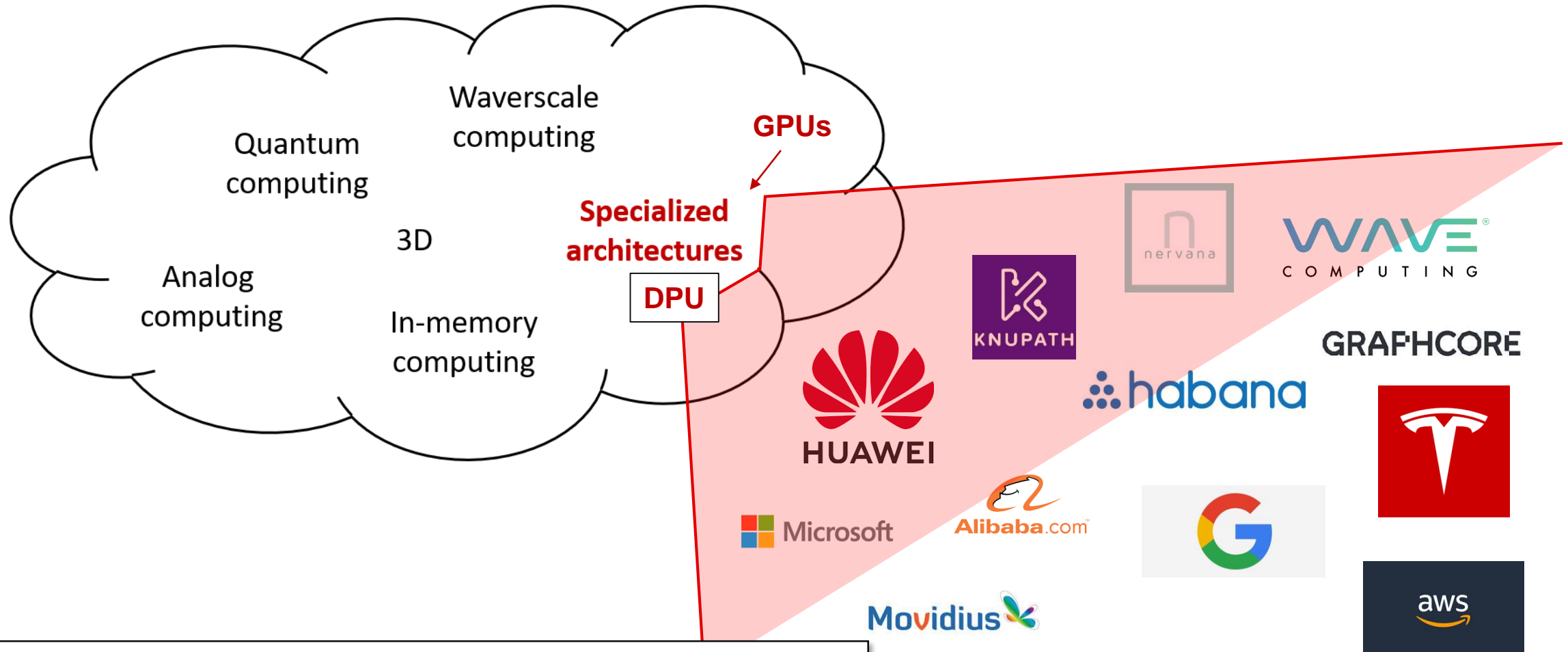
5 Series
7 Series
U+ Series
Versal



Xilinx Example:
FPGA -> ACAP

More hardened specialized functionality
to improve compute density and save power

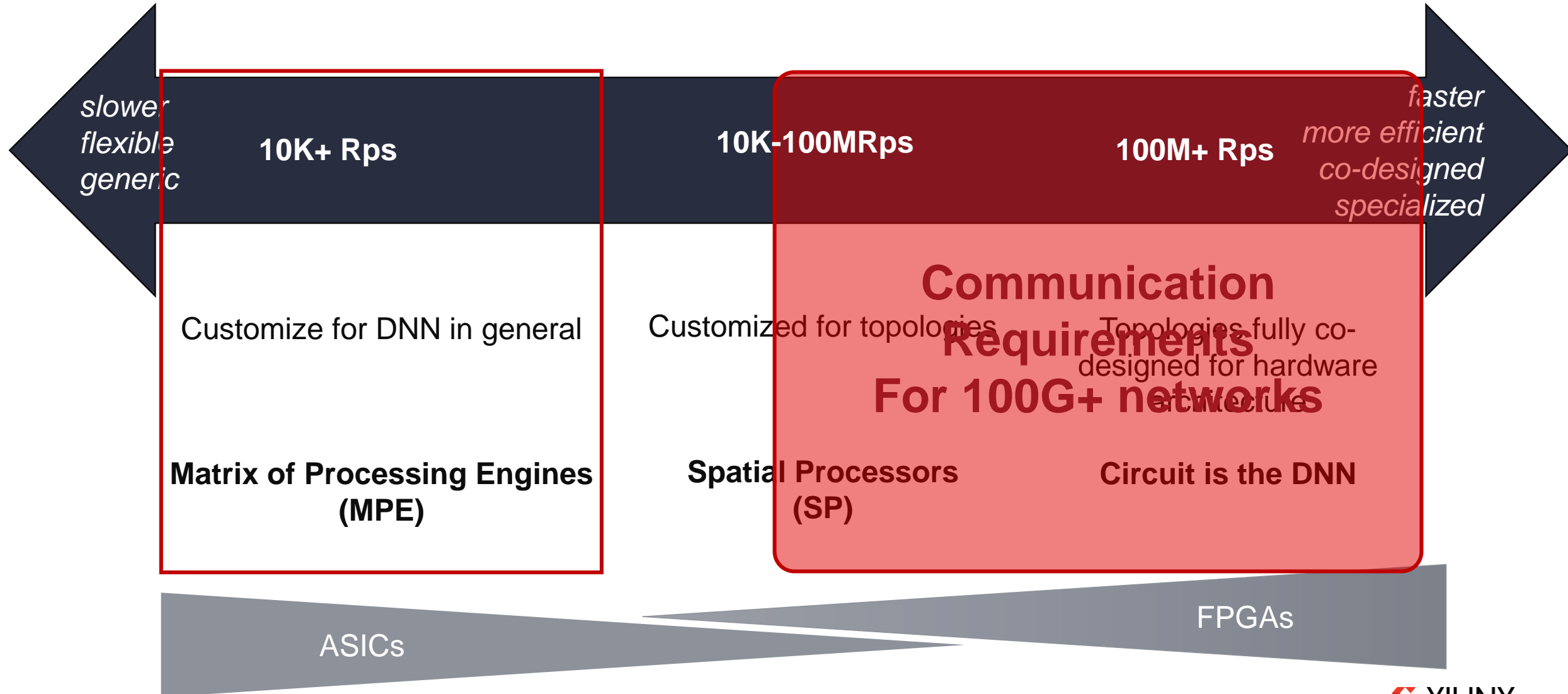
Innovative Approaches



Specialized hardware architectures for ML workloads

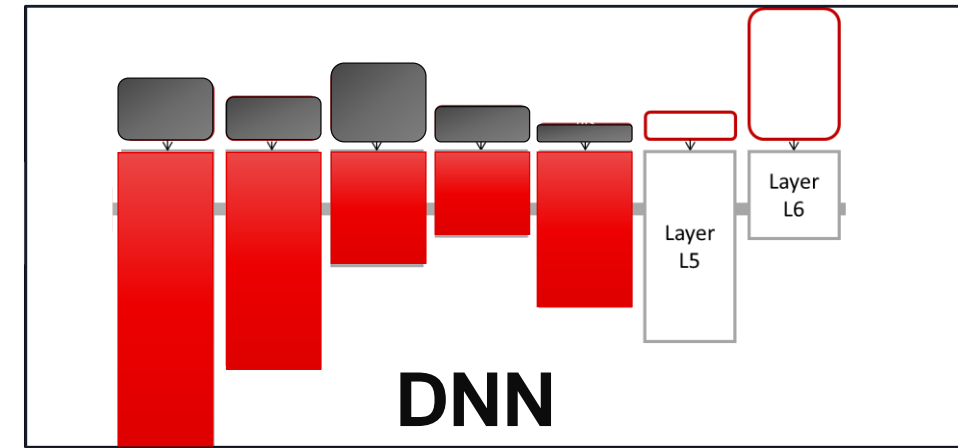
DPU Compute Architecture

Specialization, Performance & Flexibility

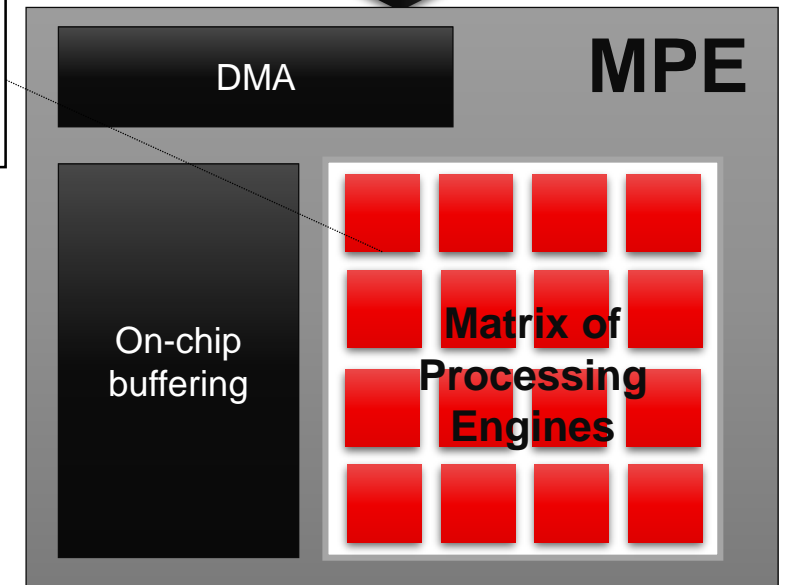
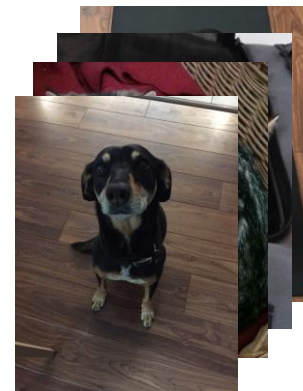


Matrix of Processing Engines Customizing for DNN in General

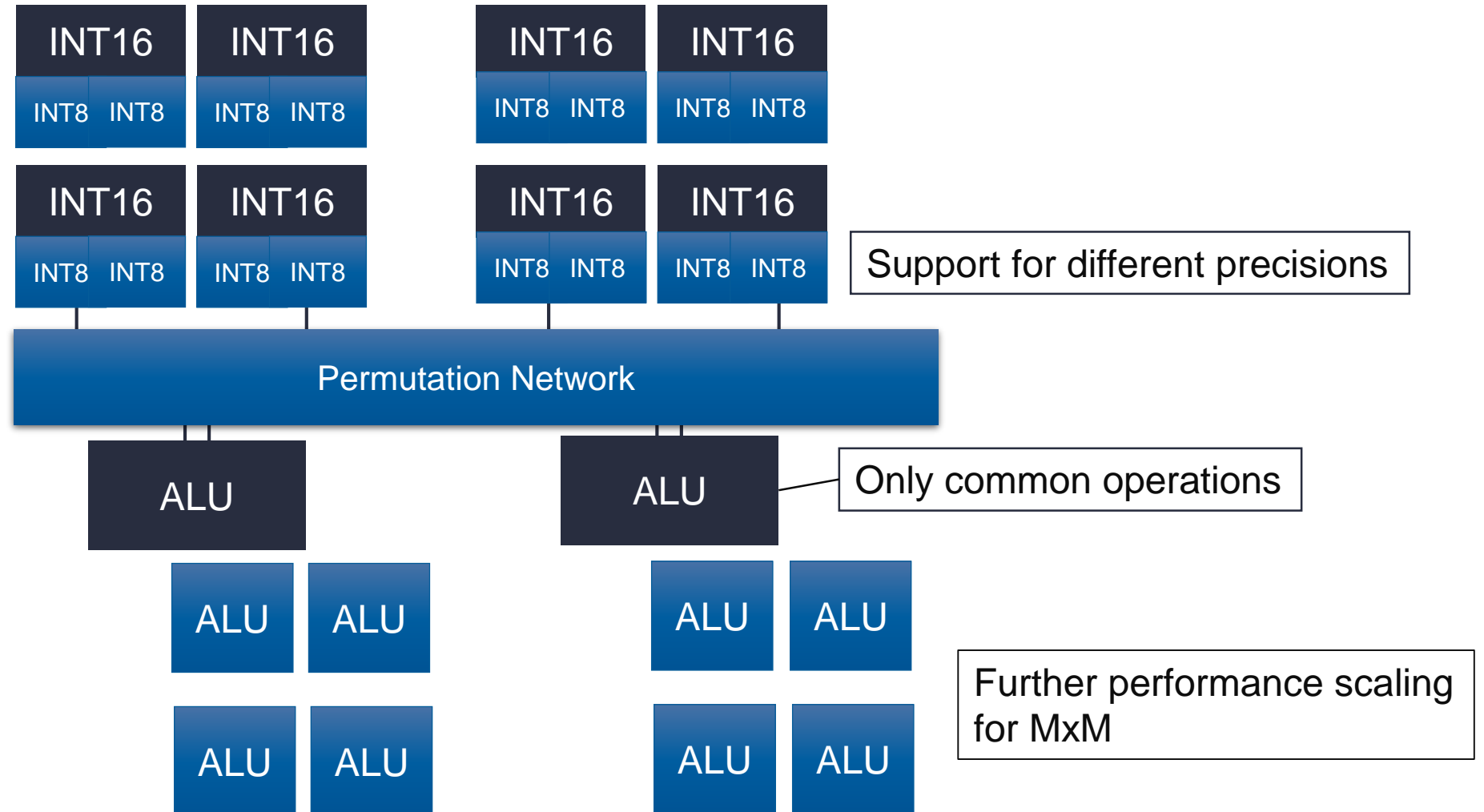
- ▶ Popular layer-by-layer compute
- ▶ Batching to achieve high compute efficiency
- ▶ Specialized processing engines
 - Operators
 - ALU types
 - tensor-, matrix- or vector-based



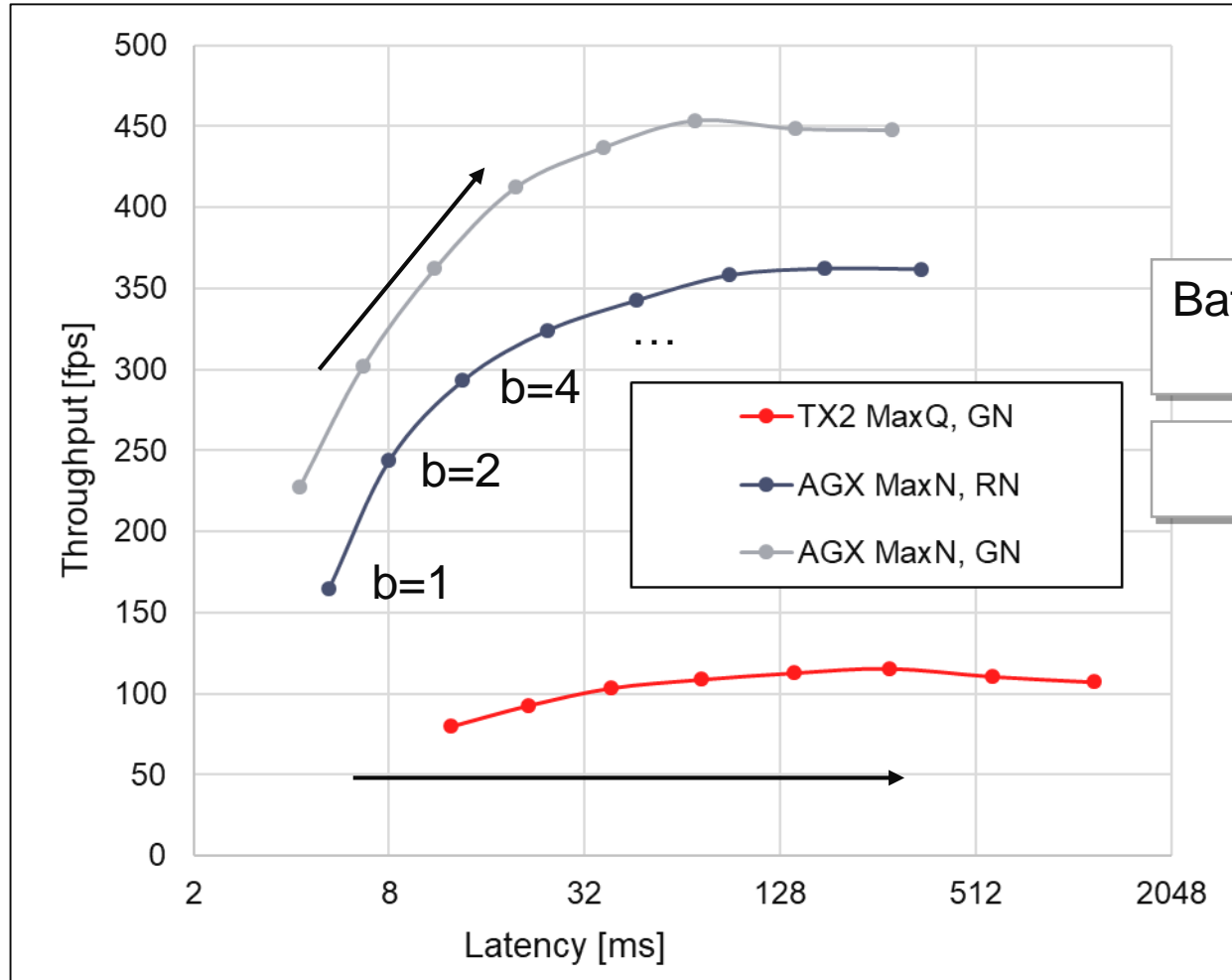
MAC, VLIW,
Vector Processor



MPE: Specialization of Processing Engines



MPE: Latency Implications of Batching



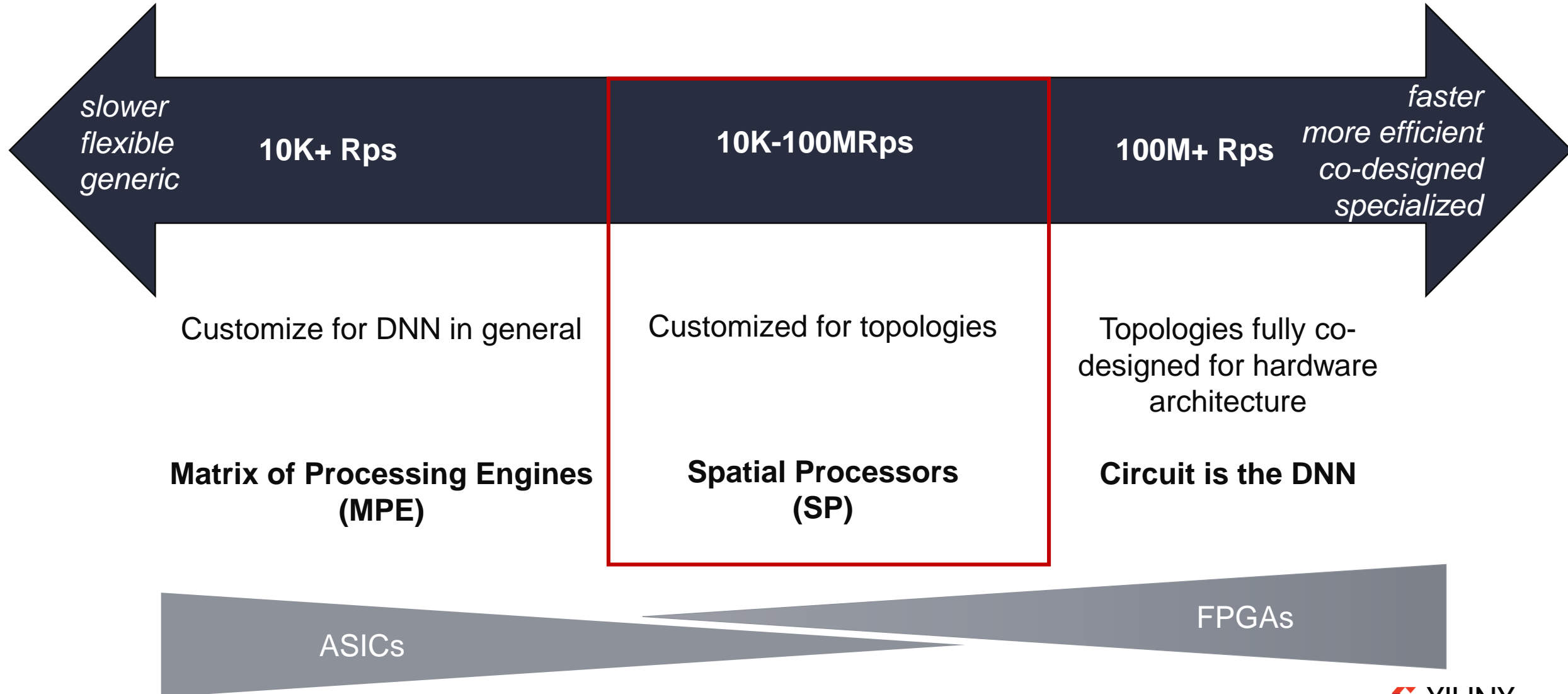
Batching improves throughput until high utilization is reached

At cost of increased latency

Embedded measurement of system-level latency, FP16
<https://rcl-lab.github.io/QtibenchWeb/>

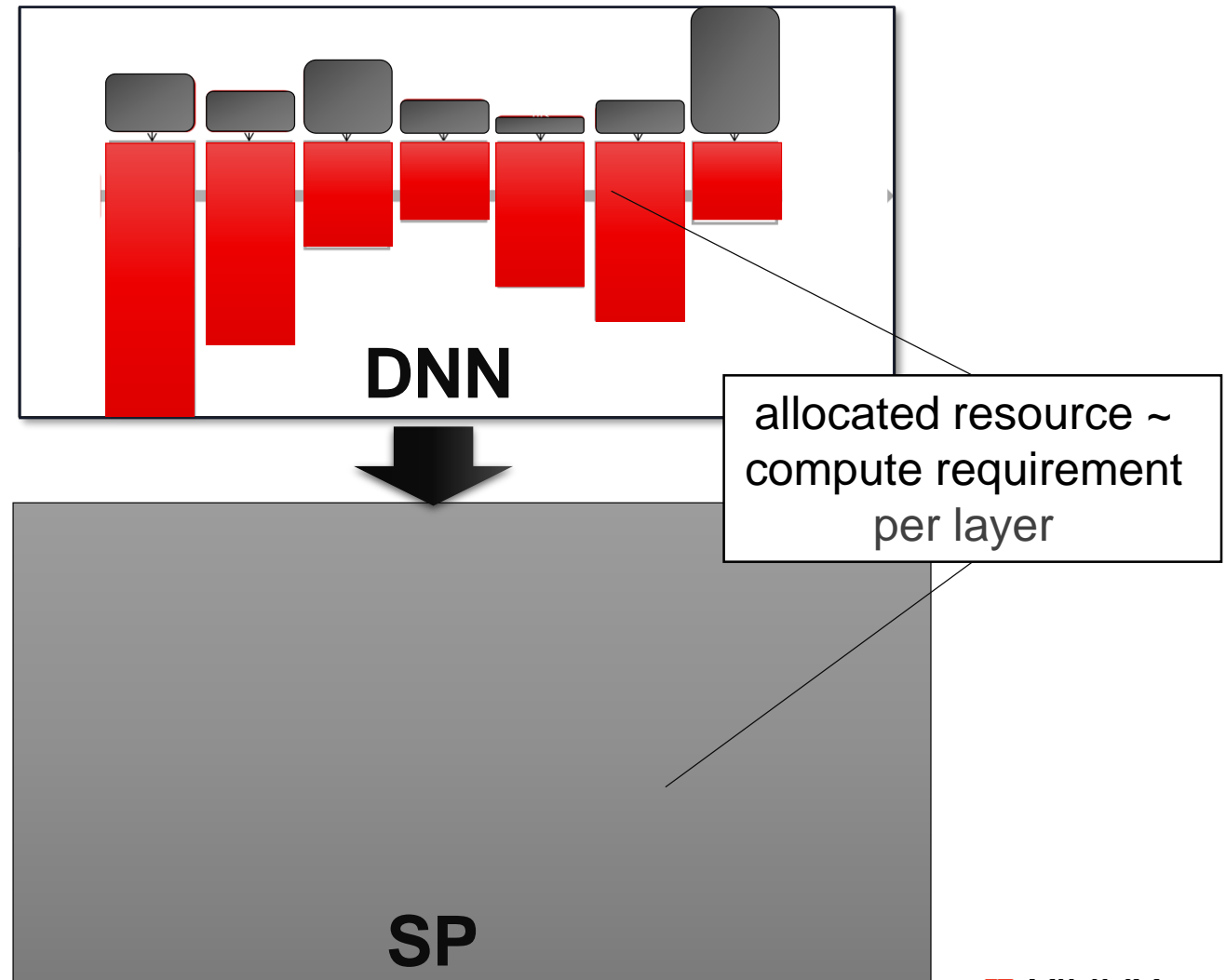
DPU Compute Architecture

Specialization, Performance & Flexibility

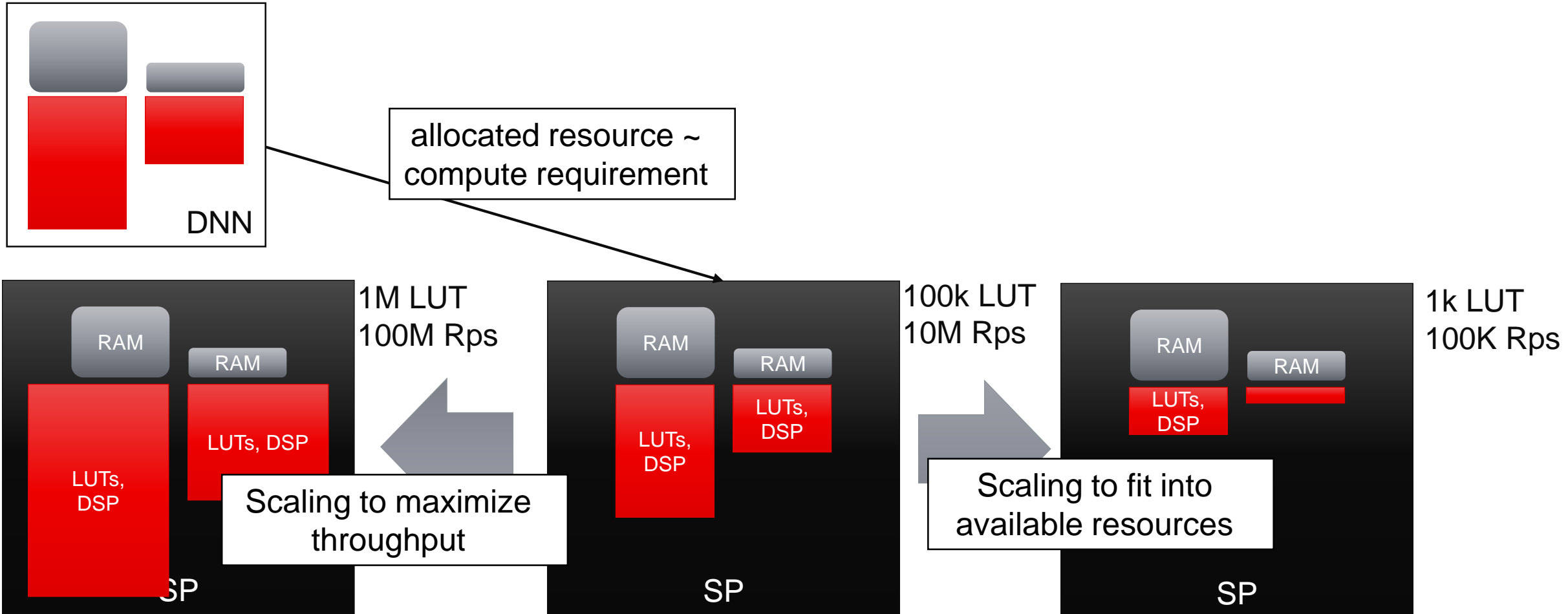


Spatial Processors: Customizing for Specific Topologies

- ▶ Hardware architecture mimics the topology
- ▶ Customize everything to the specifics of the DNN
- ▶ Benefits:
 - Improved efficiency
 - Low fixed latency
 - Higher throughput
- ▶ FPGAs rather than ASICs



Spatial Architectures: Scaling to Meet Performance & Resource Requirements

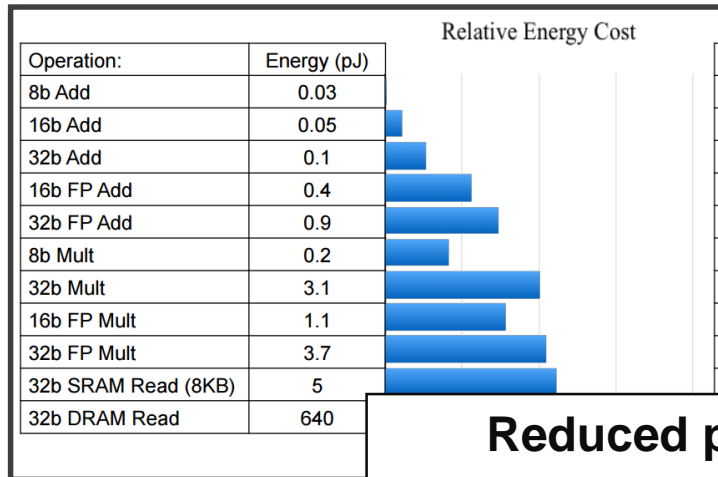


1. Scale performance & resources to meet the application requirements
2. If resources allow, we can completely unfold to create a circuit that inferences at clock speed (communications!)

Customizing Arithmetic

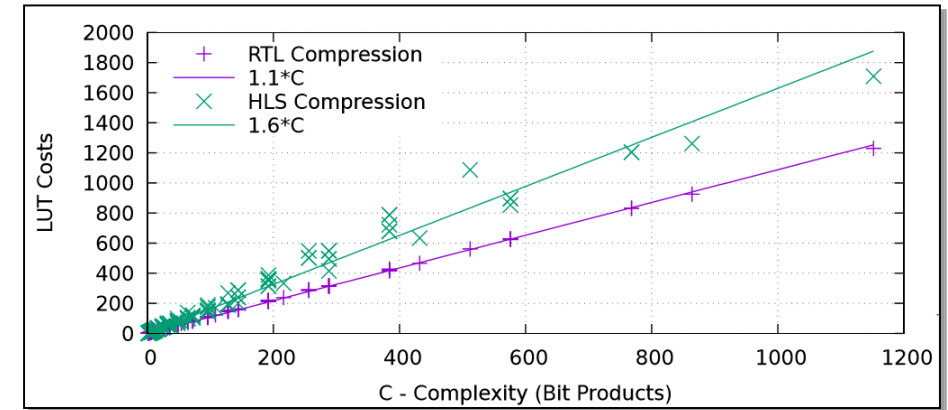
Customizing Arithmetic to Minimum Precision Required

- ▶ 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

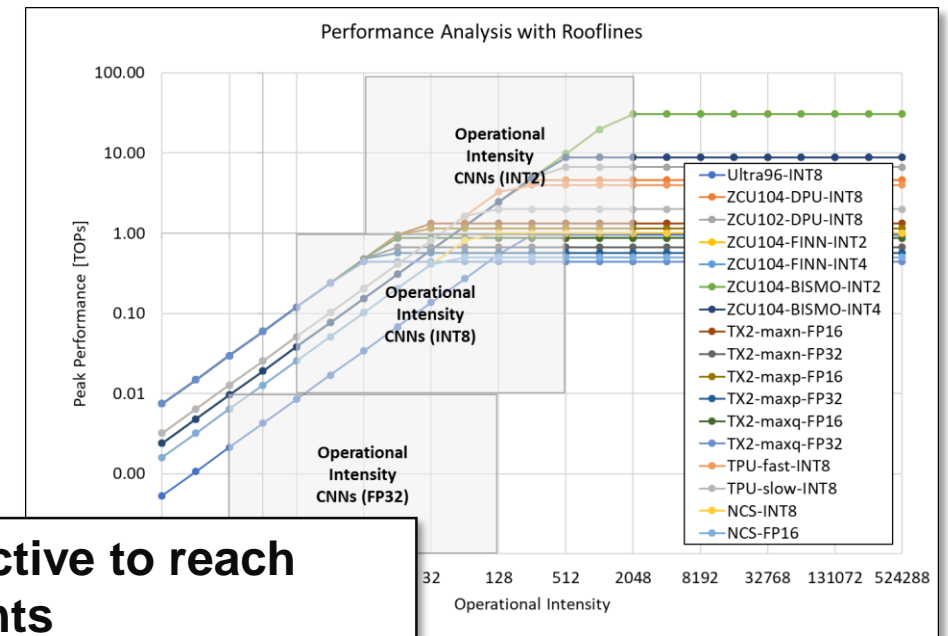


[Adapted from Horowitz, "The Green Swine Gun: Solving the Problem (and what we can do about it), ISSCC'14]

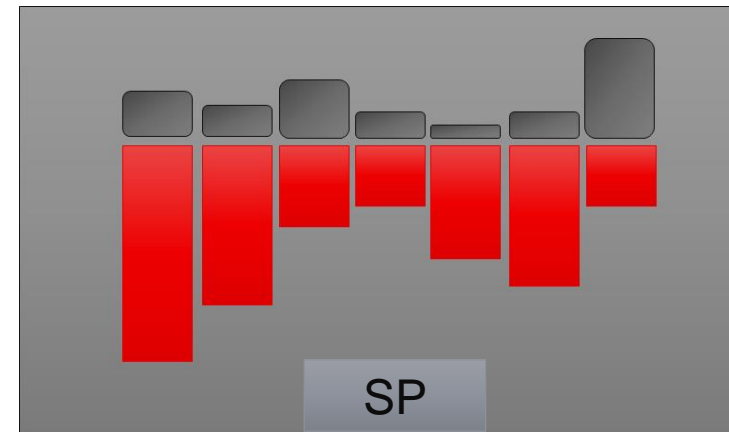
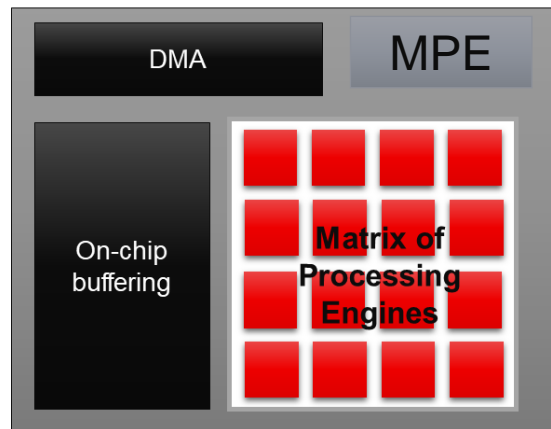
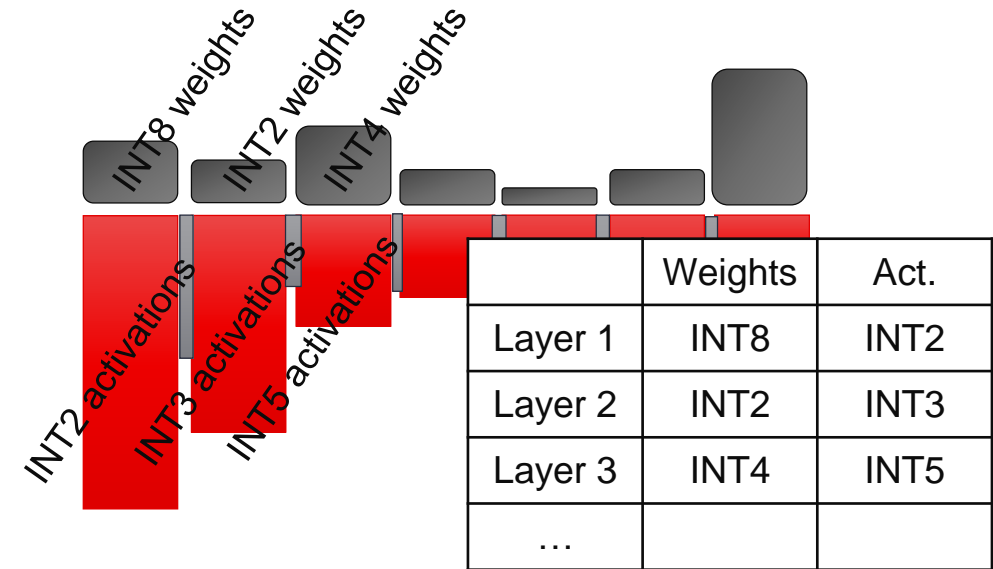
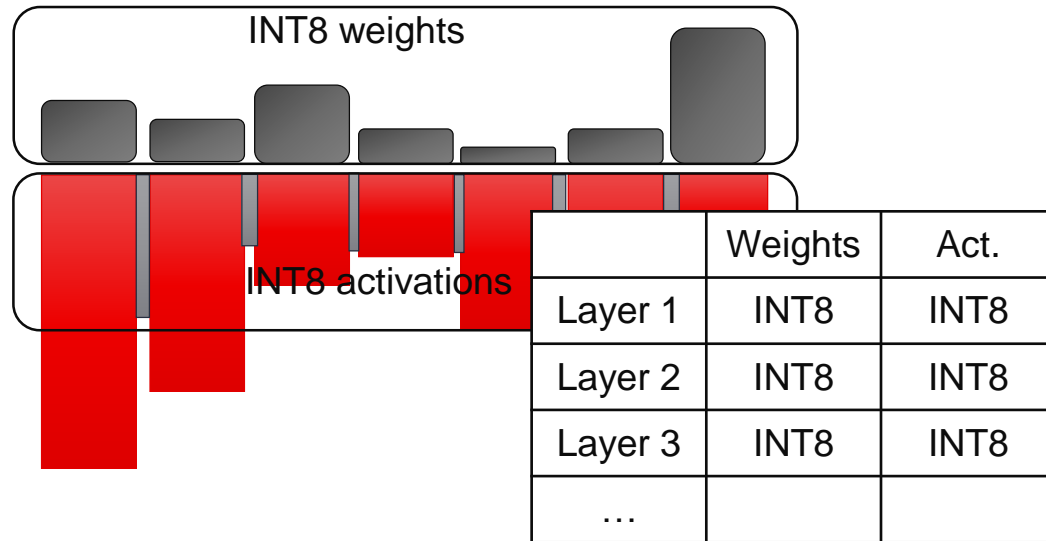
Reduced precision can be highly effective to reach communication requirements



$C = \text{size of accumulator} * \text{size of weight} * \text{size of activation}$



Granularity of Customizing Arithmetic

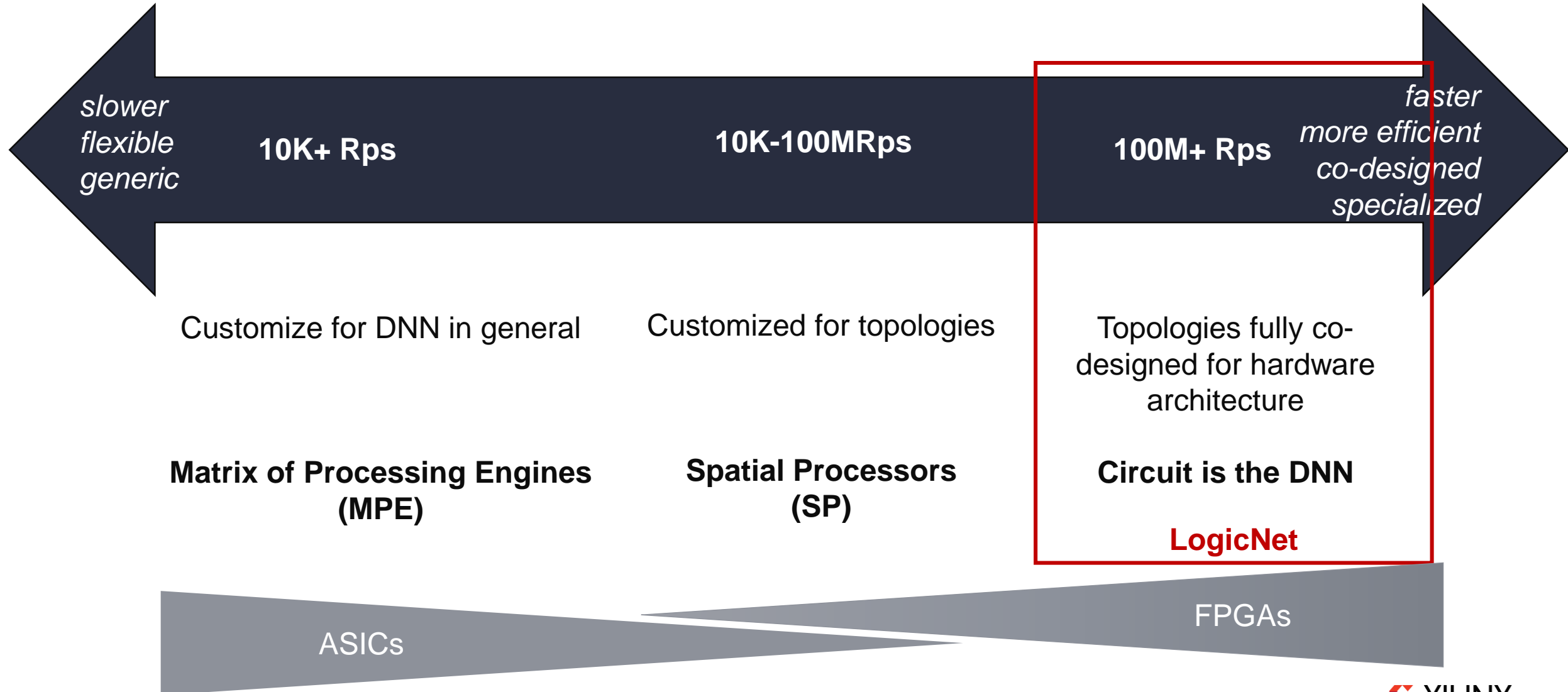


Spatial architectures enable fine granular customization of arithmetic

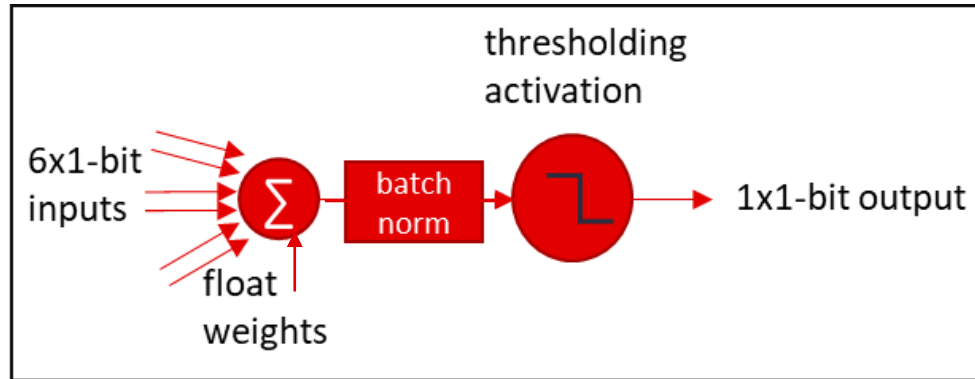
Extreme Specialization

DPU Compute Architecture

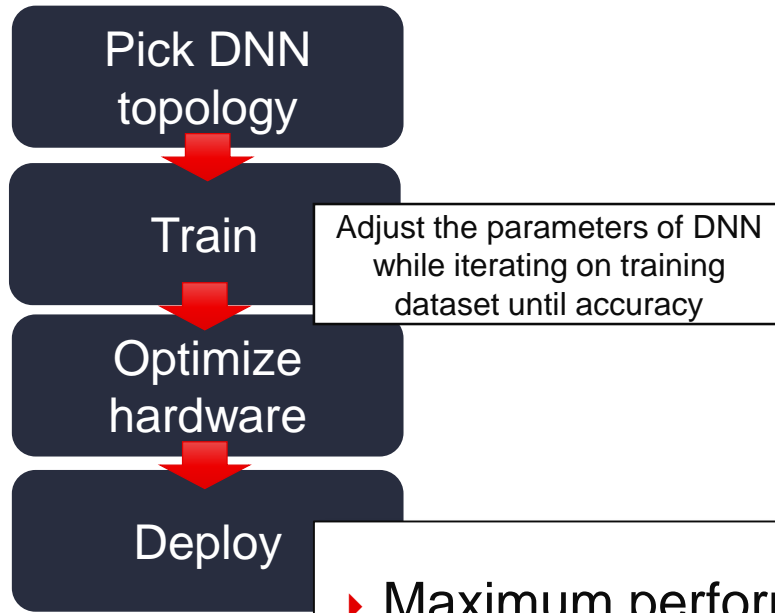
Specialization, Performance & Flexibility



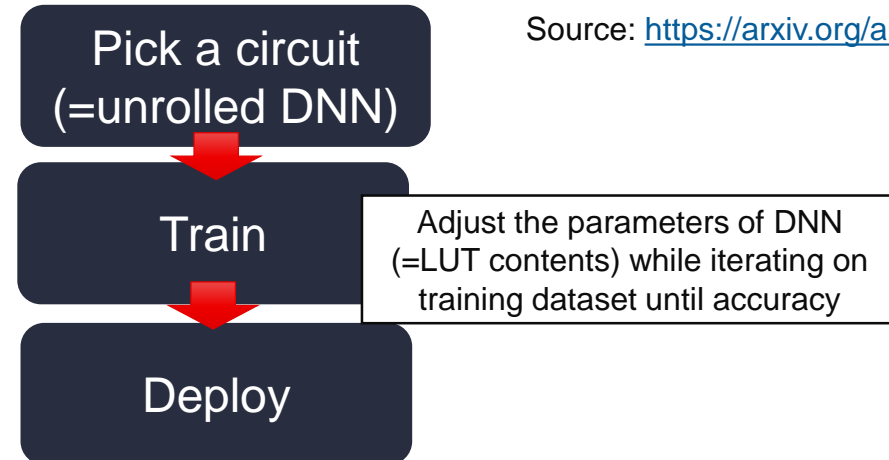
LogicNets with FPGAs



Traditional



LogicNets

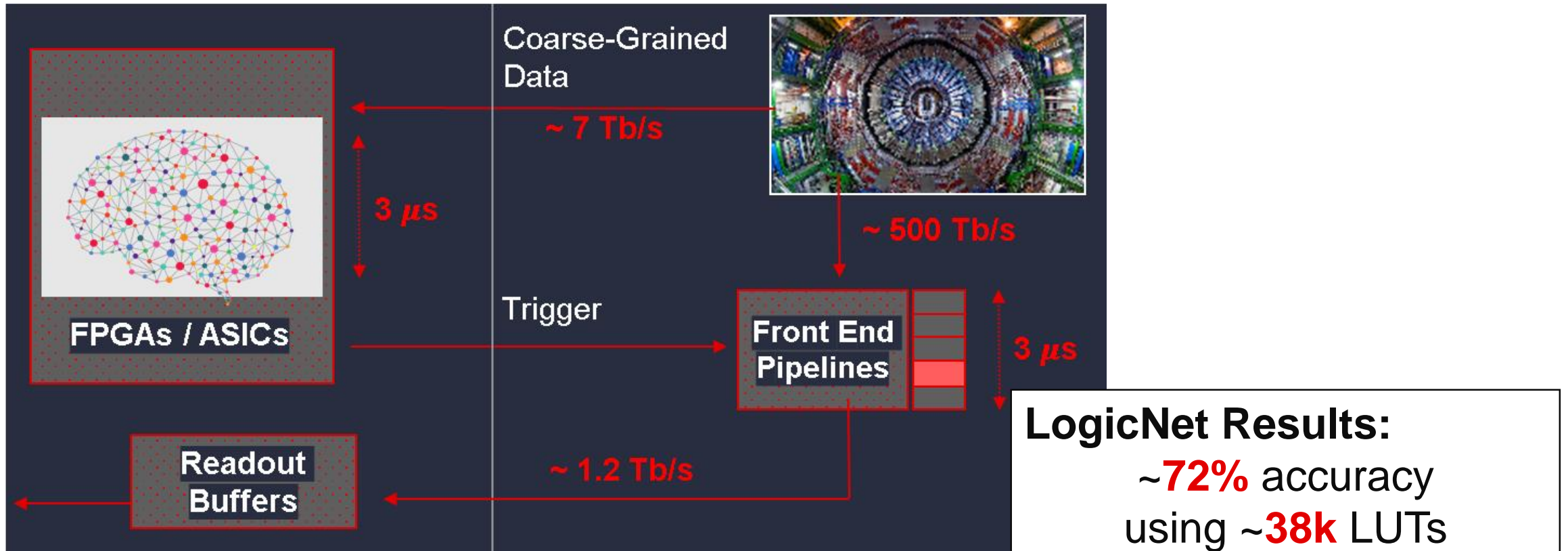


Source: <https://arxiv.org/abs/2004.03021>

▶ Maximum performance by design (classification at clock rate)

LogicNets Results

Jet Tagging (CERN LHC)



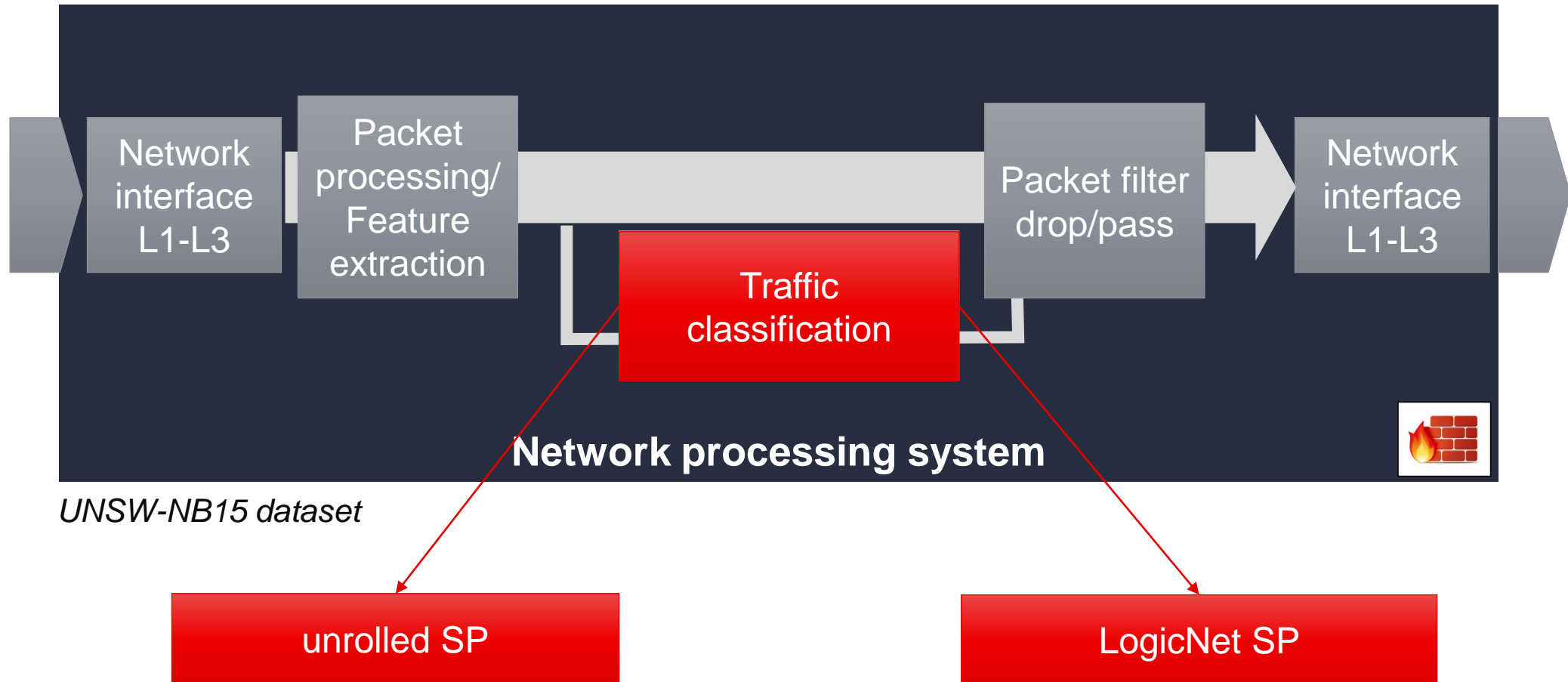
hls4ml JSC dataset [Duarte et al.]

Promising results for processing Tbps
Open source release upcoming as part of FINN (Q1 2021)



Scaling to Extreme Throughput in Network Intrusion Detection

Deep Network Intrusion Detection System



Results

DNN	Unrolled SP	LogicNet SP
topology	MLP	Circuit is the topology
#layers	3	4
neurons / layer	64	10s - 100s
#bits / weight & activation	2b	2b
#bits / inputs & output	binary	binary
Inputs / neuron	64	7
accuracy	91.9%	91.3%

If we can change the topology

Sparsity to suit to fabric

Optimization	spatially unrolled, customized arithmetic	Learned circuit
Throughput	Expected* 208MRps	471MRps
Latency	1.2usec	9nsec
Clock	208MHz	471MHz

100Gbps throughput requirements are met

Extreme low latency

low clock

400Gbps throughput requirements are met

High clock rate

UNSW-NB15 Network Intrusion Detection

Spatial processing, customized arithmetic and learned circuits can help scale to communication throughput and latency requirements

Challenge

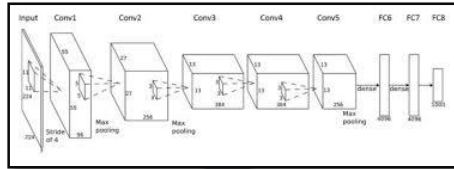
How can we enable a broader spectrum of end-users to be able to specialize hardware architectures and co-design solutions?

Project Mission



- ▶ Providing tools and platforms for exploration of DNN 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 DNN to FPGA Deployment




Brevitas
Training in pytorch
Algorithmic optimizations

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

ONNX Intermediate Representation

FINN compiler
Specializations of
hardware architecture

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

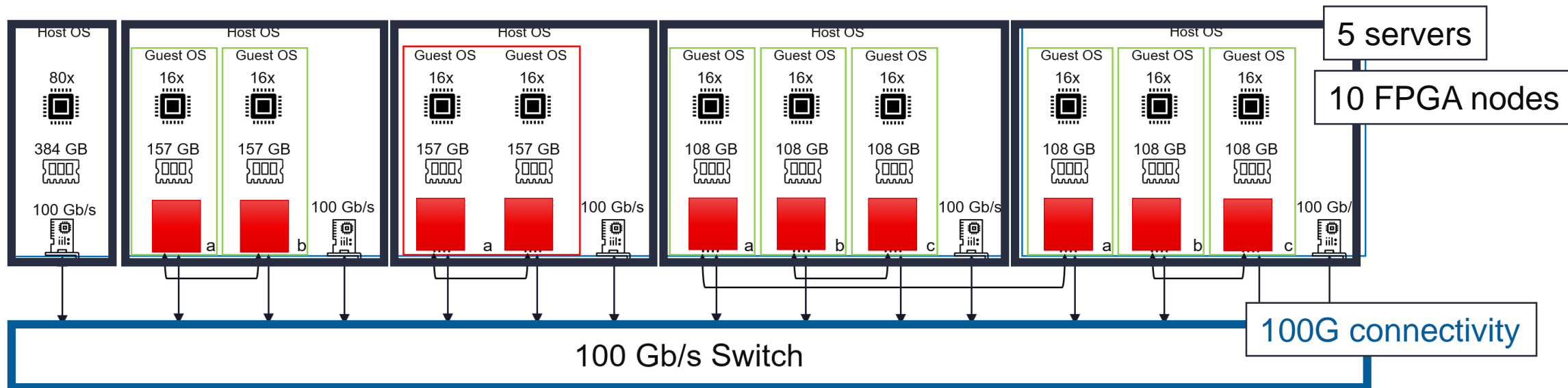
Deployment
with 

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



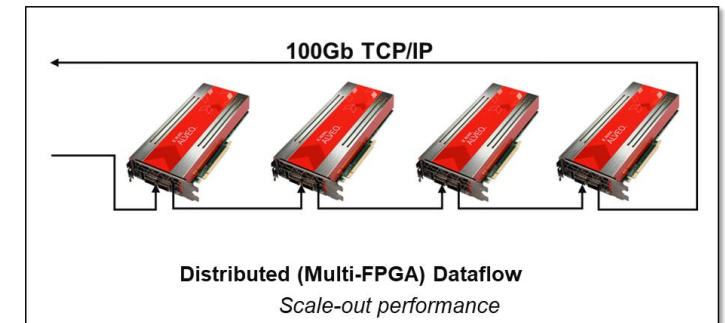
Infrastructure for Experimentation

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



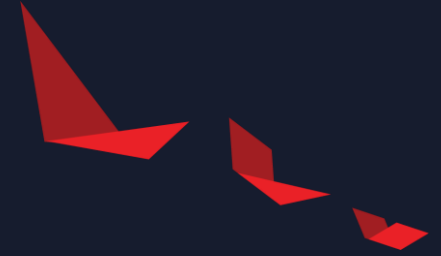
FINN Status

- ▶ Ongoing development
 - Support for residual topologies, depthwise convolutions
 - LogicNet
 - 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

- ▶ Specialization in hardware architectures is key to scaling performance to meet requirements of DNNs in communications
- ▶ With more flexibility, more opportunity to customization
 - FPGAs allow to specialize to the specifics of individual use cases without losing generality
- ▶ SPs with customized arithmetic and LogicNets are shown to meet 100Gbps – 400Gbps requirements in NIDS (as well as high energy physics)
- ▶ 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 😊**



Thank You

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