

Pervasive and Sustainable AI with Adaptive Computing Architectures

Michaela Blott

Senior Fellow

AMD Research & Advanced Development

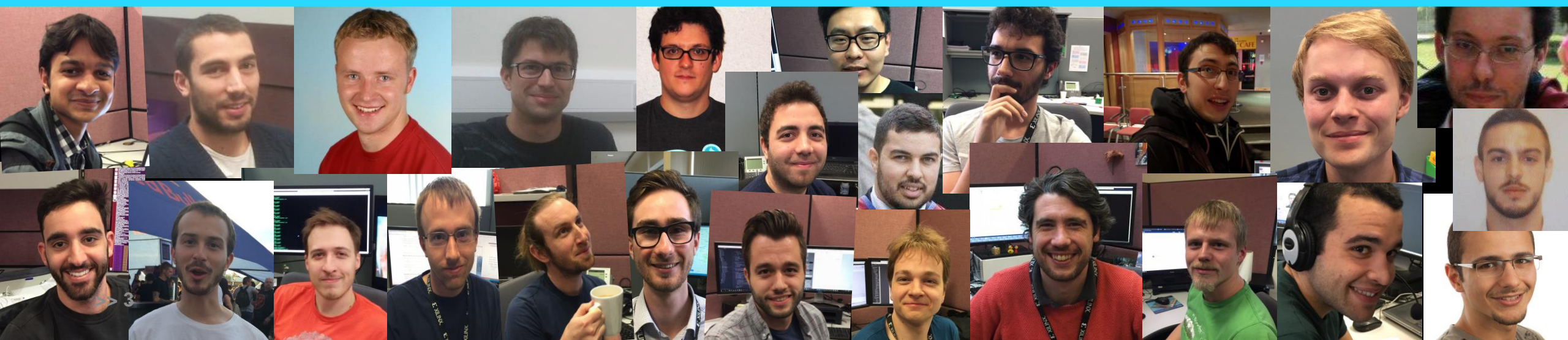
AMD Research and Advanced Development (RAD)

- **Integrated Comms and AI Lab**
 - ~20 researchers plus university program
 - 5 different locations
 - Established as Xilinx Research Labs 18 years ago
- **Focus: AI and Communications**
 - Building systems, architectural exploration, algorithmic optimizations, benchmarking
 - In collaboration with partners, customers, and universities
 - ETH Zuerich, Paderborn University, Imperial College, KIT, NTNU, Politecnico di Milano, NUS, University of Sydney



Active Internship Program

- On average 10 interns at any given time
 - From top universities all over the world
- Overall
 - 100+ interns since 2007
 - Many collaborations have come from this
 - Many found employment



CONTEXT

DNNs and Their Potential

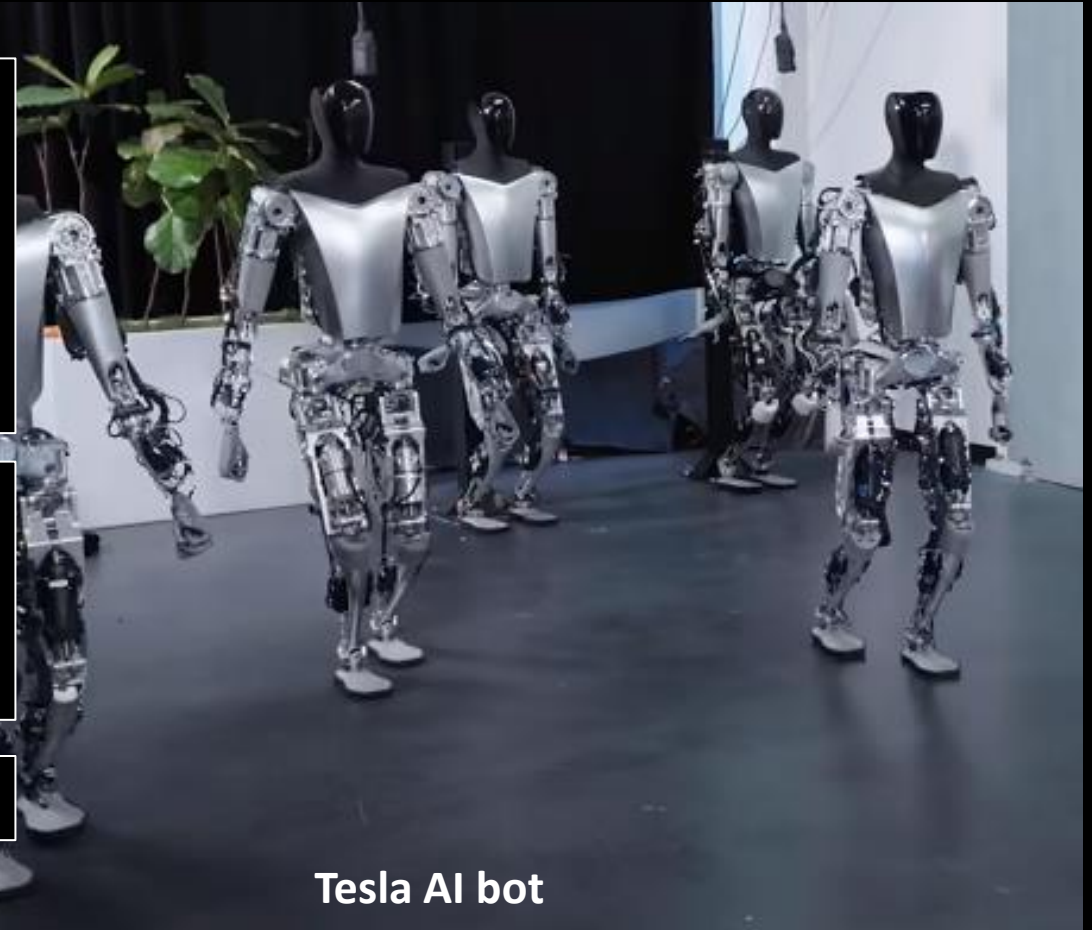
Huge potential

- Requires little domain expertise
- NNs are a “universal approximation function”
- If you make it big enough and train it long enough
 - Can outperform humans and existing algorithms on specific tasks

Solves previously unsolved problems

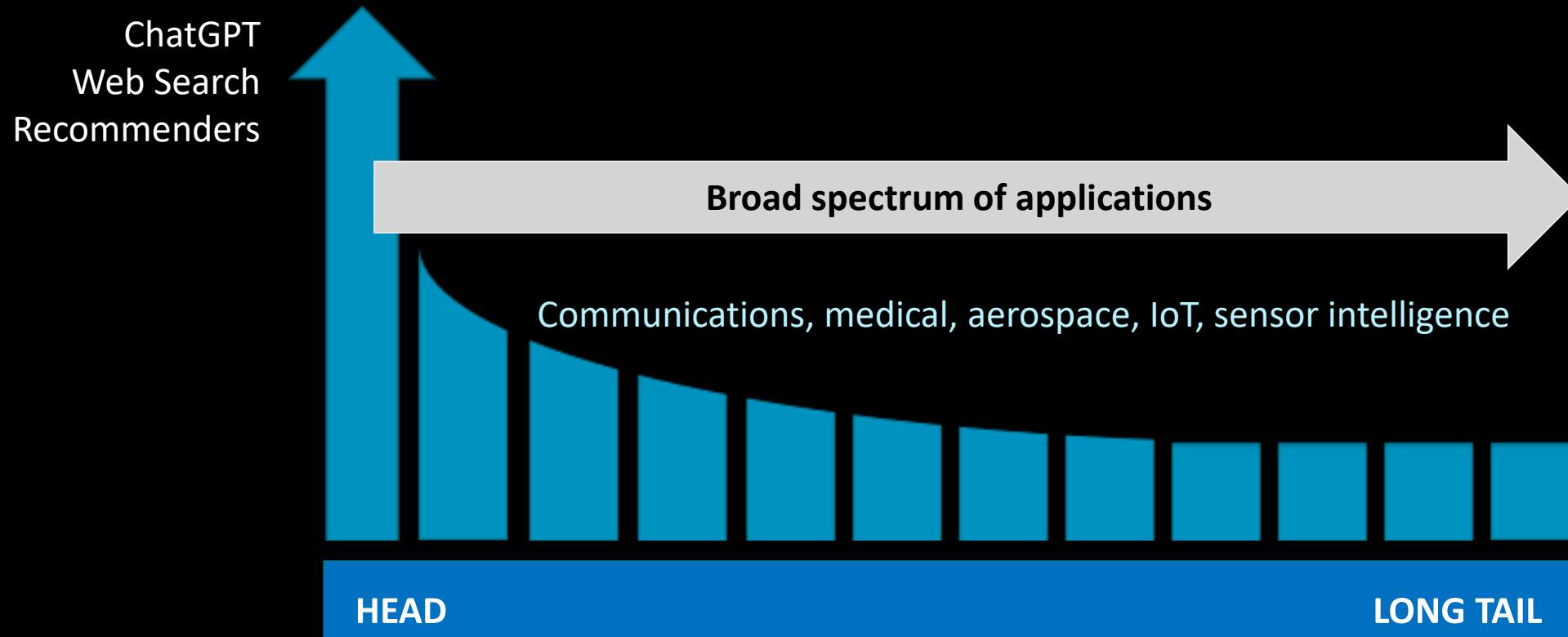
- Code, text and image generation, and GPT-4 even passed the bar exam in the 90th percentile
- Protein folding

Increasing adoption in many different applications



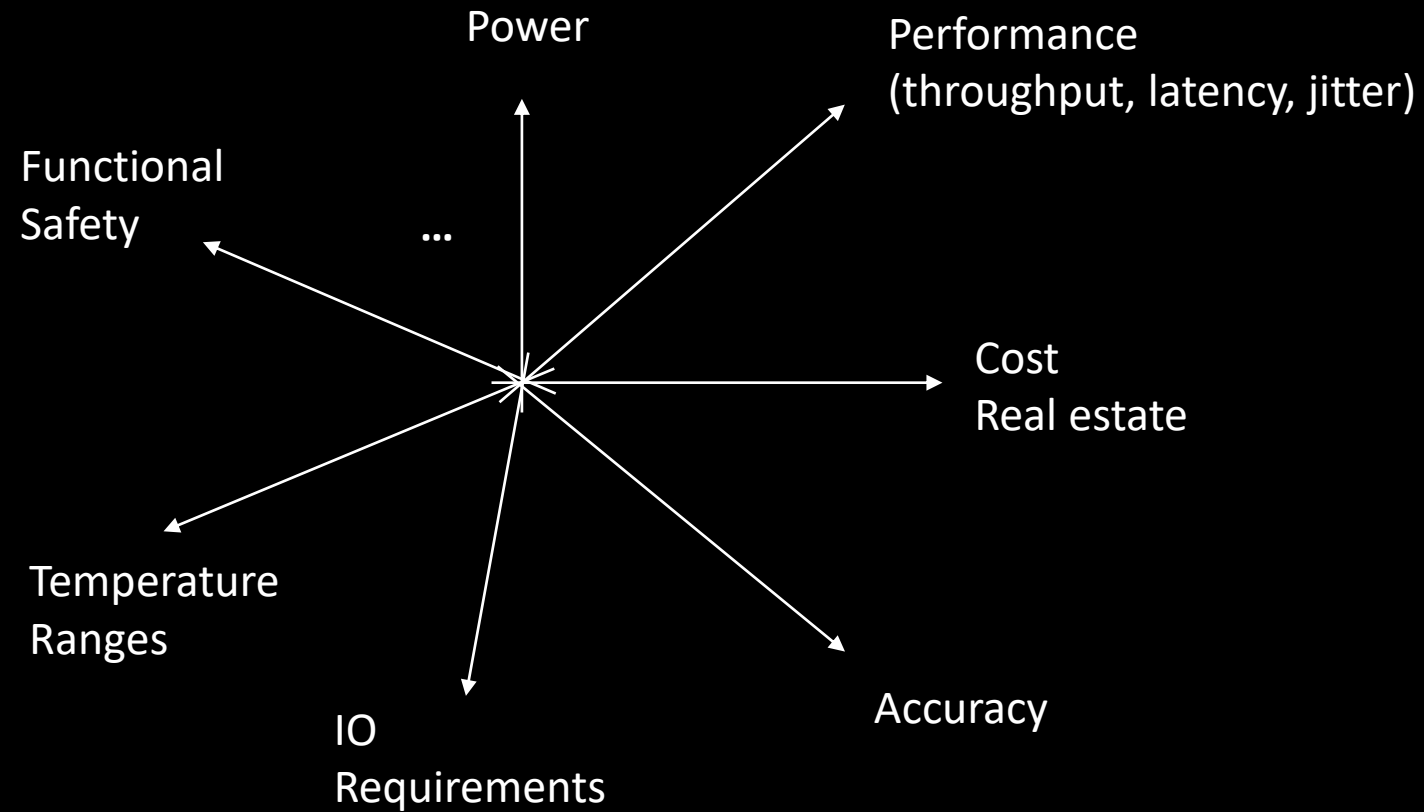
Tesla AI bot

Pervasive AI



Adapted from TED Talk: Andrew Ng "How AI could empower any business"

Pervasive AI Comes with Diverse Requirements



Examples of Diverse Requirements

- **IoT/Embedded**
 - Small resource footprint, low power (<10W), low latency (msec) and zero jitter
- **High-Frequency Trading**
 - High-frequency trading (HFT) is an arms race of acquiring data and executing trading decisions fastest
 - Multimillion-dollar advantages through nanosecond differences
 - Extreme low latency requirements (nsec) as DNNs are being adopted for better trading decisions
- **High-Energy Particle Physics**
 - CERN CMS Experiment needs nsec latency for setting recording trigger
 - Incoming data needs to be processed at 7 Tbps
 - Extreme latency requirements (nsec)

Examples of Diverse Requirements - Communications

- Extreme throughput (100s Minferences/sec)
 - Line-rate processing for $n \cdot 100\text{G}$ Ethernet
- Low latency (<msec)
 - Real-time communications (5G and 6G)
 - Reduce buffering demands
 - No execution run-time with batching but streaming integration
- Fusing with signal processing on lower protocol layers*

- DNNs are increasingly penetrating both wireless and wired telecommunications
 - monitoring, prediction, optimizing, learned physical interfaces
- Extreme throughput and low latency requirements

Dynamic Workloads

AI is a highly active research area

- Algorithms are still changing, science is not mature yet
 - Next data type? FP32 -> INT8 -> BF16 -> FP8 => Logarithmic?
 - Next operator that changes the compute paradigm? Transformers have arrived in 2017 and are now everywhere->?
 - Next generative paradigm? VAE -> GAN -> Denoising Diffusion
- Fundamentally disruptive ideas
 - Hinton's NeurIPS 2022 keynote speech on Forward-Forward learning – backpropagation not be needed in the future?*
- Customer workloads are changing during the development cycle
 - Models are in flux (optimization)
 - First 3GPP 6G specification expected in 2028

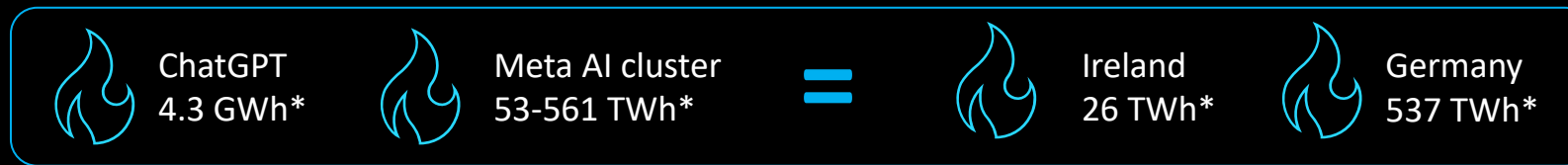
Discover neural connectivity

*<https://syncedreview.com/2022/12/08/geoffrey-hintons-forward-forward-algorithm-charts-a-new-path-for-neural-networks/>

**Audibert, Rafael & Lemos, Henrique & Avelar, Pedro & Tavares, Anderson & Lamb, Luís. (2022). On the Evolution of A.I. and Machine Learning: Towards Measuring and Understanding Impact, Influence, and Leadership at Premier A.I. Conferences. 10.48550/arXiv.2205.13131.

Sustainability & Energy Consumption

- Energy footprint on par with whole industrial nations



- Current DNN algorithms represent a **sledgehammer approach**
 - Extremely inefficient

100s kilo Watts
matrix multiply



Scope for Improvement:
Estimated 10^5



20Watts

The carbon footprint of ChatGPT. An estimate of the carbon emissions... | by Chris Pointon | Dec, 2022 | Medium

<https://www.semianalysis.com/p/meta-discusses-ai-hardware-and-co>

Germany - Energy consumption in Germany (worlddata.info)

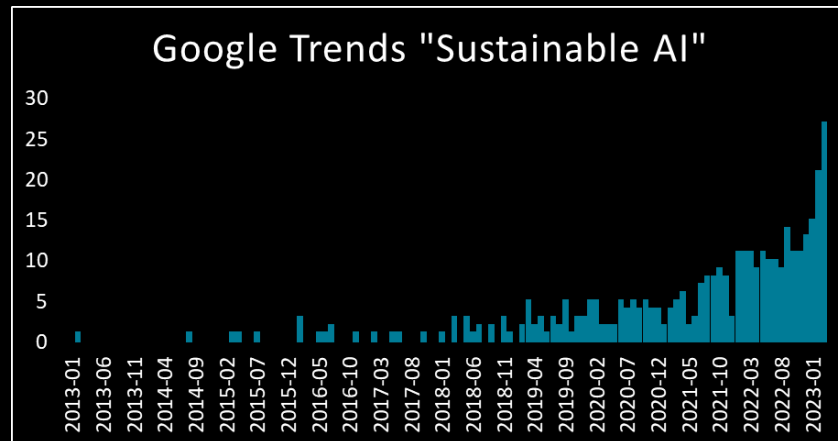
Ireland - Energy consumption in Ireland (worlddata.info)

**Yu Wang, Tsinghua University, Feb 2016 <https://www.numenta.com/blog/2022/05/24/ai-is-harming-our-planet/>

*TWh = Tera Watt hours

Paradigm Will Shift towards Energy Efficient AI

- Energy will become the limiting factor to scaling NNs

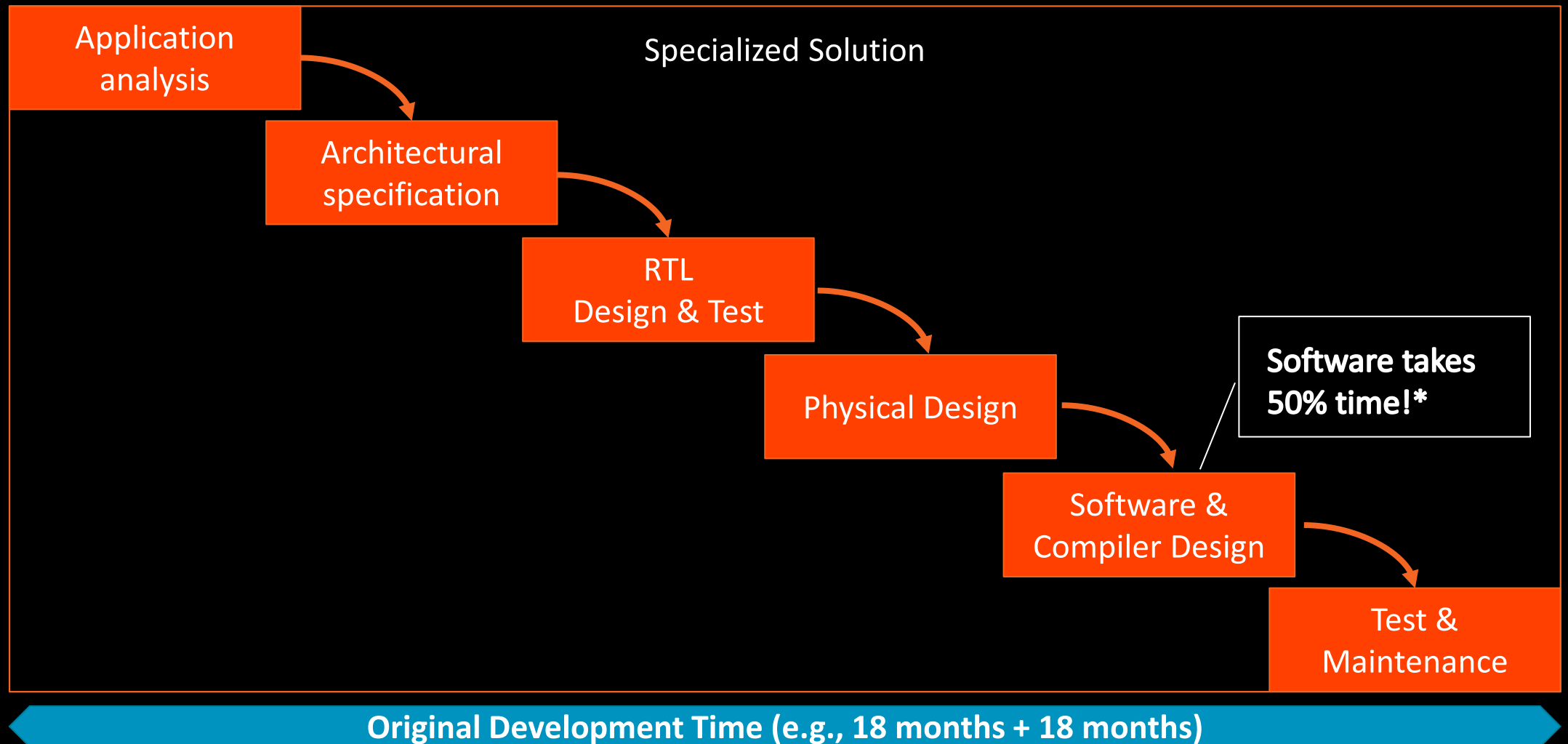


Specialization Is #1 Industry Approach to Energy Efficiency



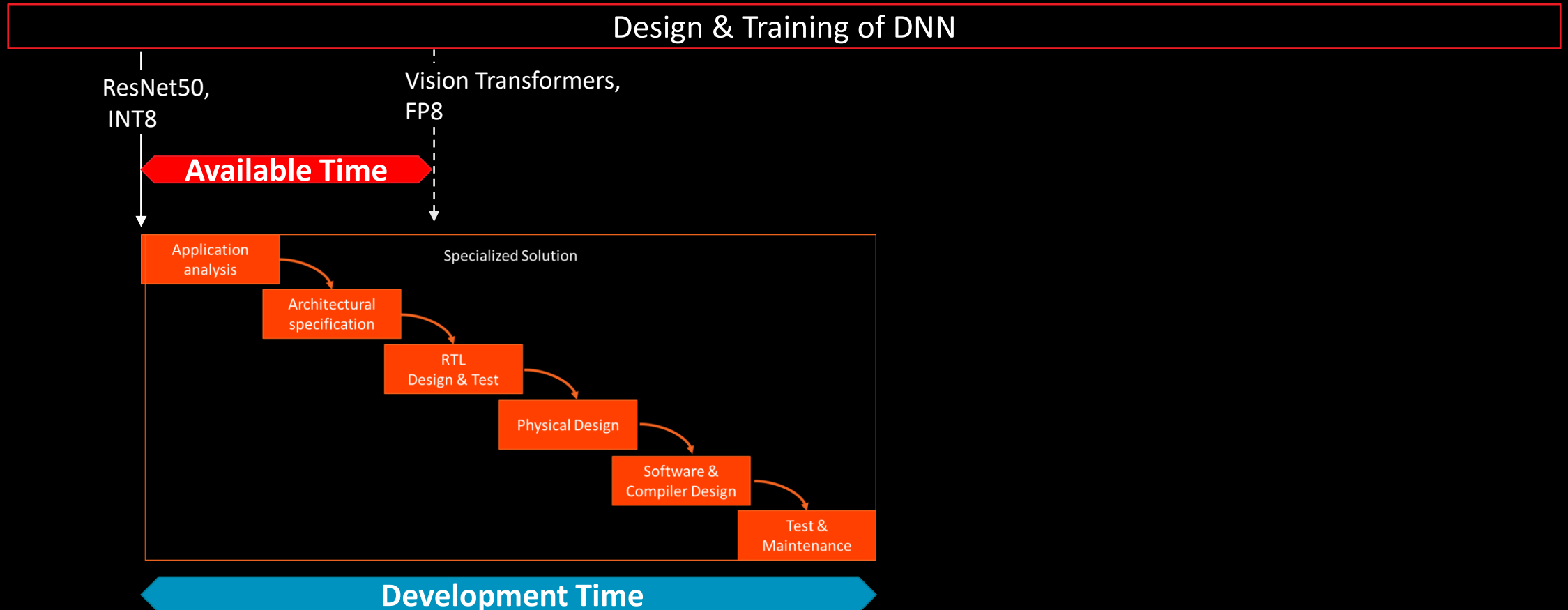
Solution Specialization

Classical Hardware Accelerator Design Process (Waterfall)



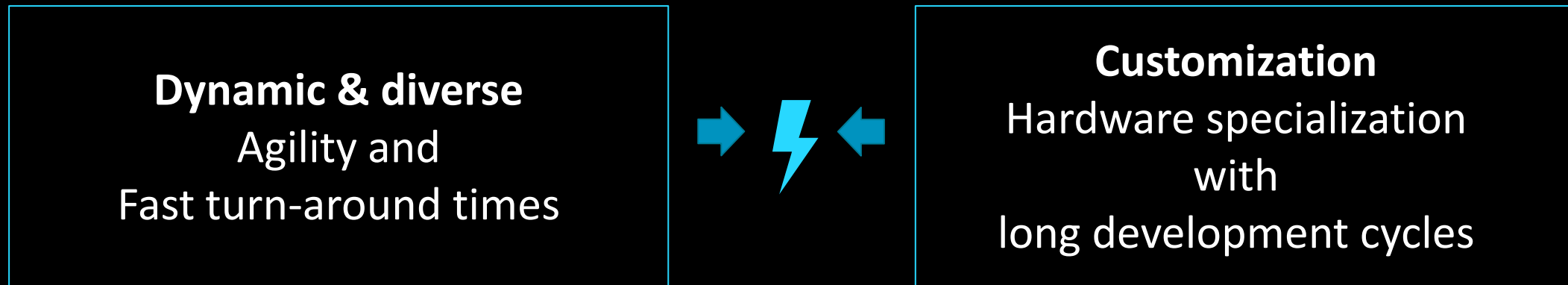
*Source: [Chip Design and Manufacturing Cost under Different Process Nodes: Data...](#) | Download Scientific Diagram (researchgate.net)

Dynamic and Diverse Workloads vs. Solution Specialization



Challenges in a Nutshell

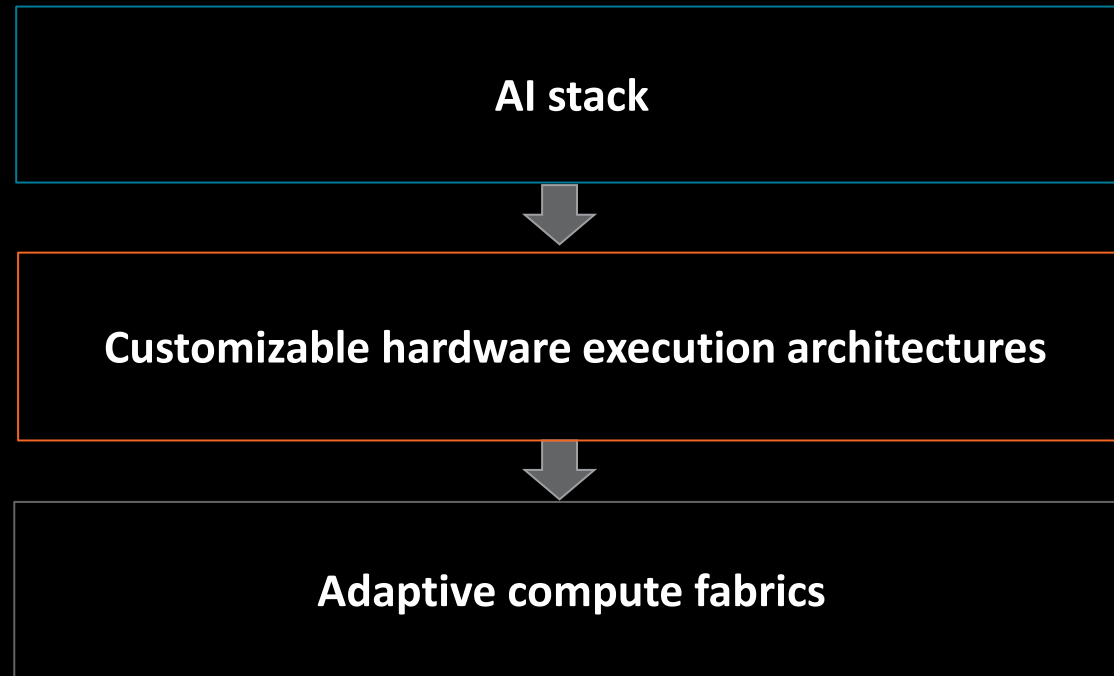
Dynamic, Diverse & Highly Customized



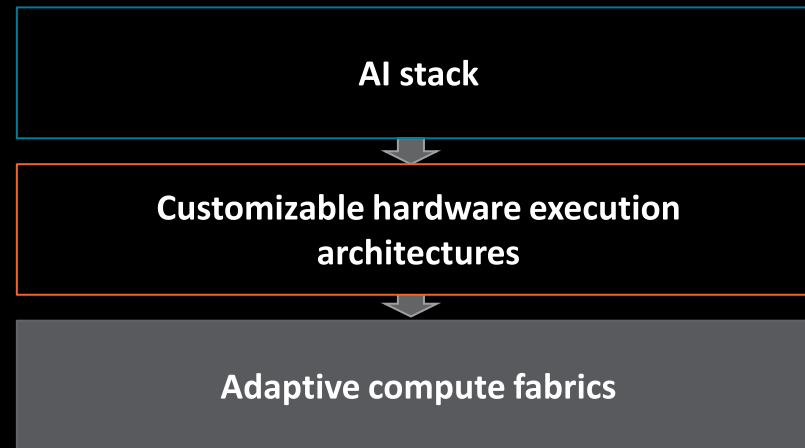
Agility in Customization is King

Enabling Rapid Specialization with Adaptive Compute Fabrics and Agile AI Stacks

Enabling Rapid Specialization with Adaptive Compute Fabrics and AI Stacks

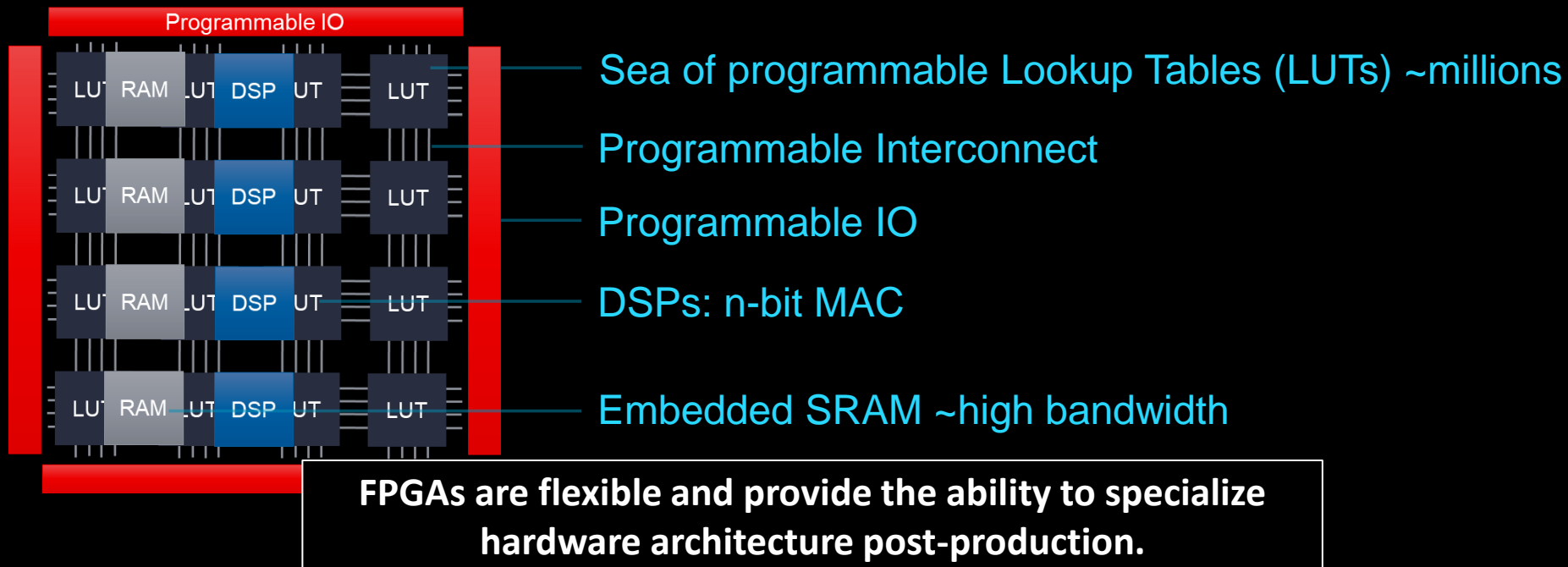


What are adaptive compute fabrics? FPGAs and AIEs



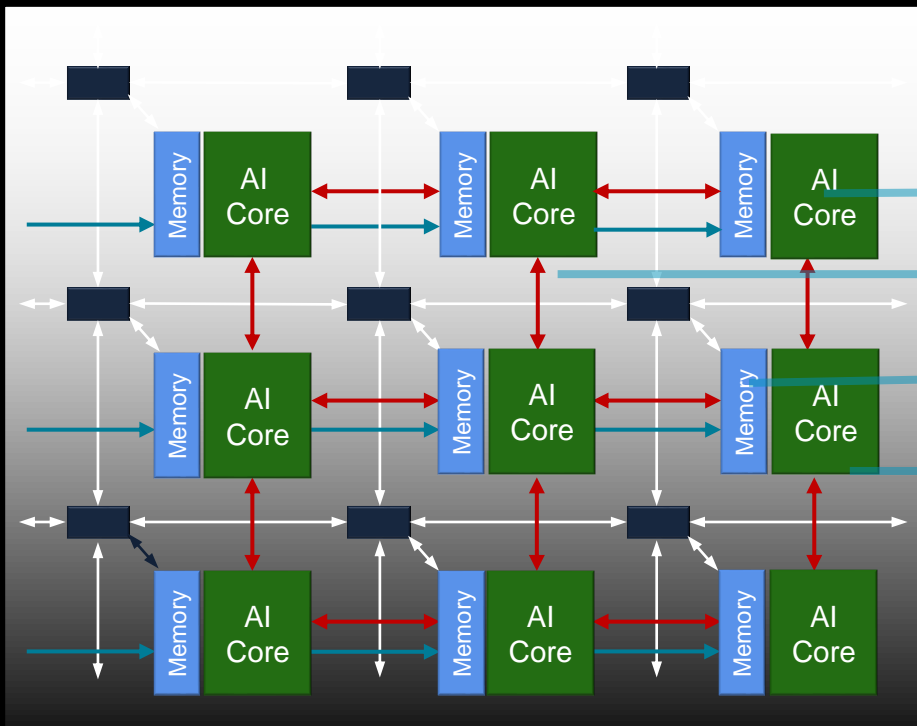
Primer: Adaptive Computing – FPGAs

- FPGAs are the **chameleon** amongst the semiconductors: flexible, adaptive mostly homogeneous hardware architectures that enable **post-production customization at the architectural level**
- Customize
 - IO interfaces
 - **Functionality post-silicon** (compression, encryption, NN accelerator, key value store,...)
 - **Compute architectures & memory subsystems** to meet specific use case's performance or energy targets



Primer: Adaptive Computing – AIEs

- AI Engines (AIEs): new form of higher performant, adaptive compute fabric
 - Higher performance through hardened vector processing in VLIW cores, just word-based (instead of bit-based) with native support for ML-optimized data types (e.g., INT8, block float,...)
 - Great flexibility because of interconnectivity and separate control flow
 - => **adapt the execution architecture to different workloads**



Matrix of VLIW/SIMD vector processors (10s...100x)

Flexible interconnect

Tightly coupled, embedded memory (1..10s MB)

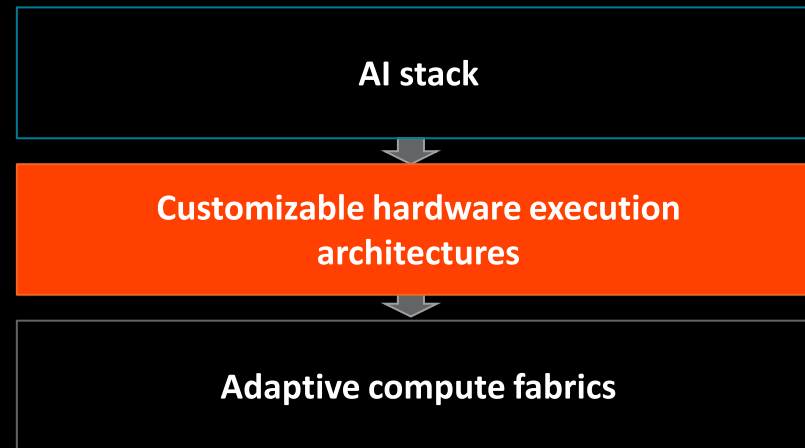
AIE are software compiled and don't require synthesis

FPGAs Are Diverse and Widely Deployed

- **~100 Product Families**
 - Spanning Si nodes from 350 nm to 7 nm
 - Devices ship for +20s year (for example, Coolrunner XPLA3)
- **500+ Base Parts**
 - Different fabric sizes and mixtures between DSPs and LUTs
 - Combinations with high-speed serial IO, ADC, HBM, ARM cores, and other hard IP
- **Three basic temperature grades & three speed grades**
- **Other variants**
 - Radiation hardened and custom variants

FPGAs available in a broad spectrum of parts to cater to the diverse requirements in pervasive AI

Extreme Specialization of the Hardware Architecture (post-silicon)

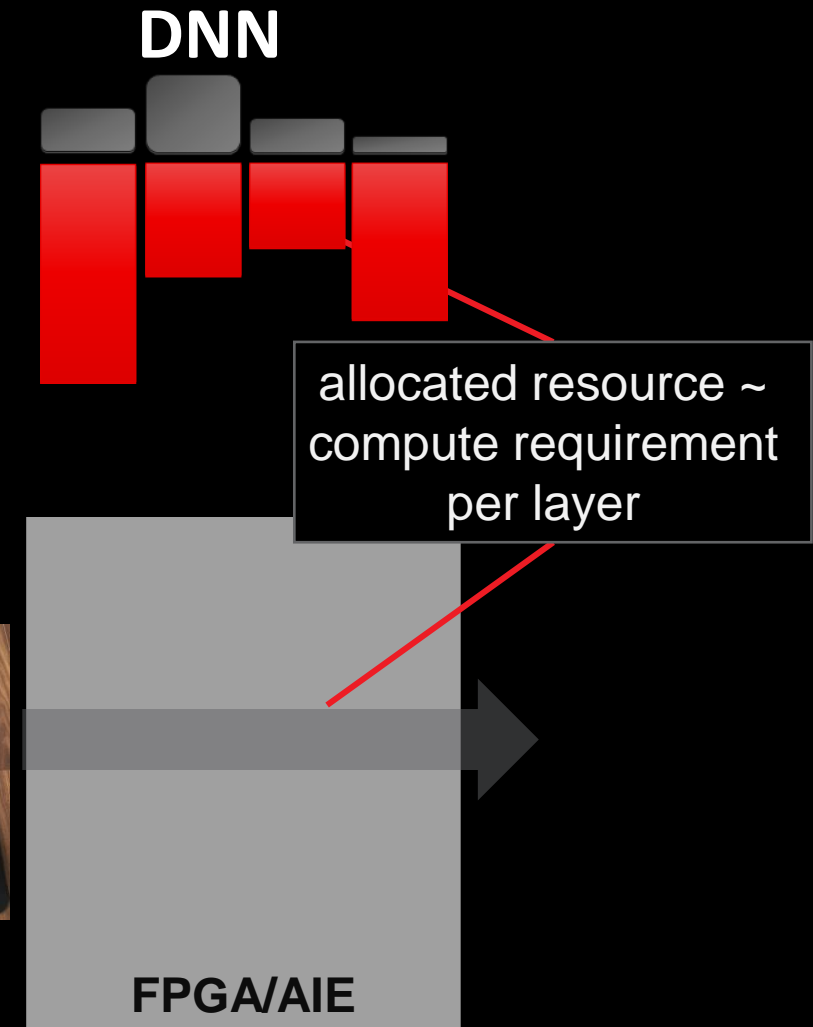


Key Concepts



Dataflow - Specializing for Individual Topologies

- Hardware instantiates the topology as a dataflow architecture
- Customize everything to the specifics of the given DNN, its operations and connectivity
- Benefits: energy efficiency, latency and throughput scalability

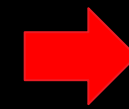


Dataflow - Energy Efficiency

- Architecture only computes and stores what's needed in the specific use case
 - Customized memory and compute subsystem
- Minimizes movement & storing of data
 - Activations are not buffered externally; they are in SRAM and registers moved directly from one layer to next
- High efficiency through concurrent communication and compute
 - Each layer starts computing as soon as first inputs are available
 - Shortens execution time => energy saving ($E = P * time$)

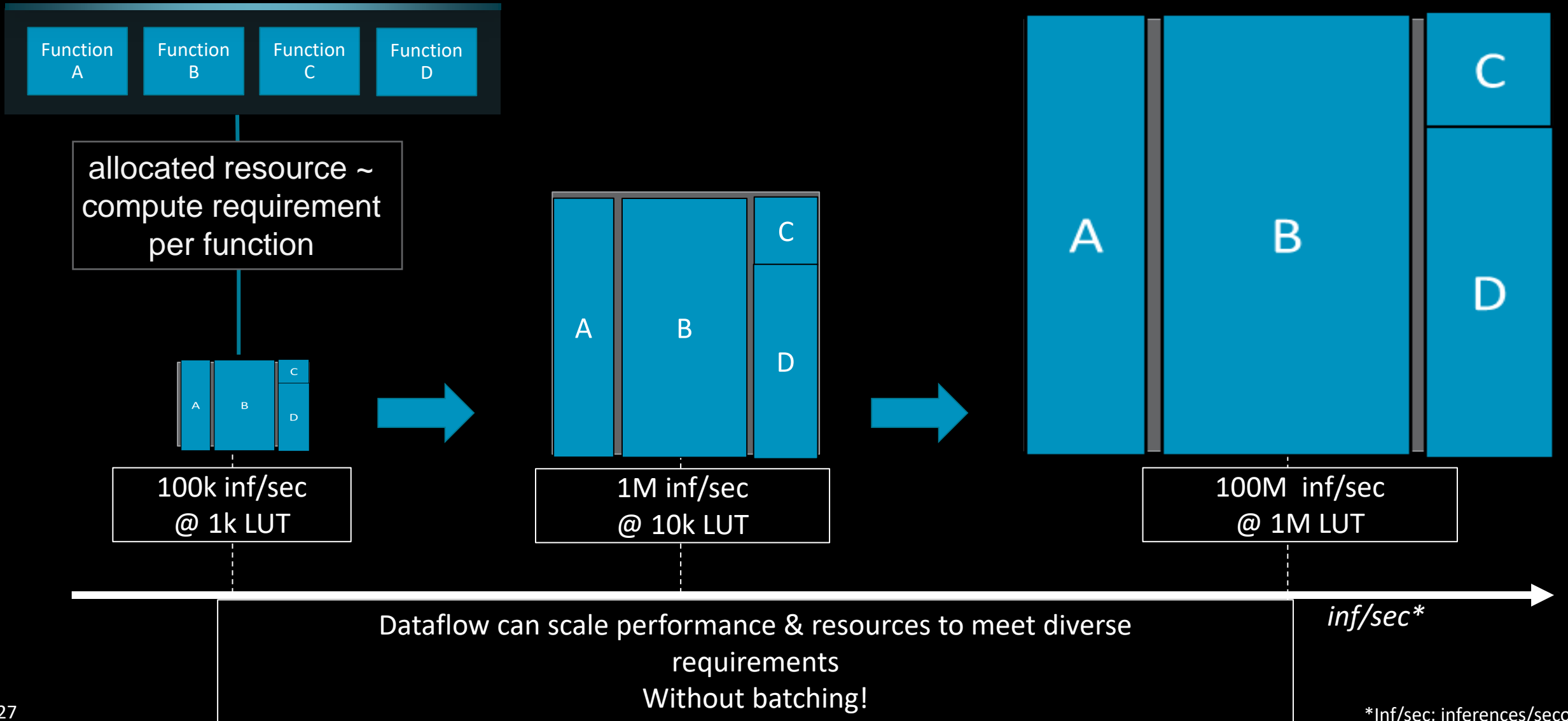
Operation		Picojoules per Operation		
		45 nm	7 nm	45 / 7
+	Int 8	0.03	0.007	4.3
	Int 32	0.1	0.03	3.3
	BFloat 16	--	0.11	--
	IEEE FP 16	0.4	0.16	2.5
	IEEE FP 32	0.9	0.38	2.4
×	Int 8	0.2	0.07	2.9
	Int 32	3.1	1.48	2.1
	BFloat 16	--	0.21	--
	IEEE FP 16	1.1	0.34	3.2
	IEEE FP 32	3.7	1.31	2.8
SRAM	8 KB SRAM	10	7.5	1.3
	32 KB SRAM	20	8.5	2.4
	1 MB SRAM ¹	100	14	7.1
GeoMean ¹		--	--	2.6
DRAM		Circa 45 nm	Circa 7 nm	
	DDR3/4	1300 ²	1300 ²	1.0
	HBM2	--	250-450 ²	--
	GDDR6	--	350-480 ²	--

Table 2. Energy per Operation: 45 nm [16] vs 7 nm. Memory is pJ per 64-bit access.



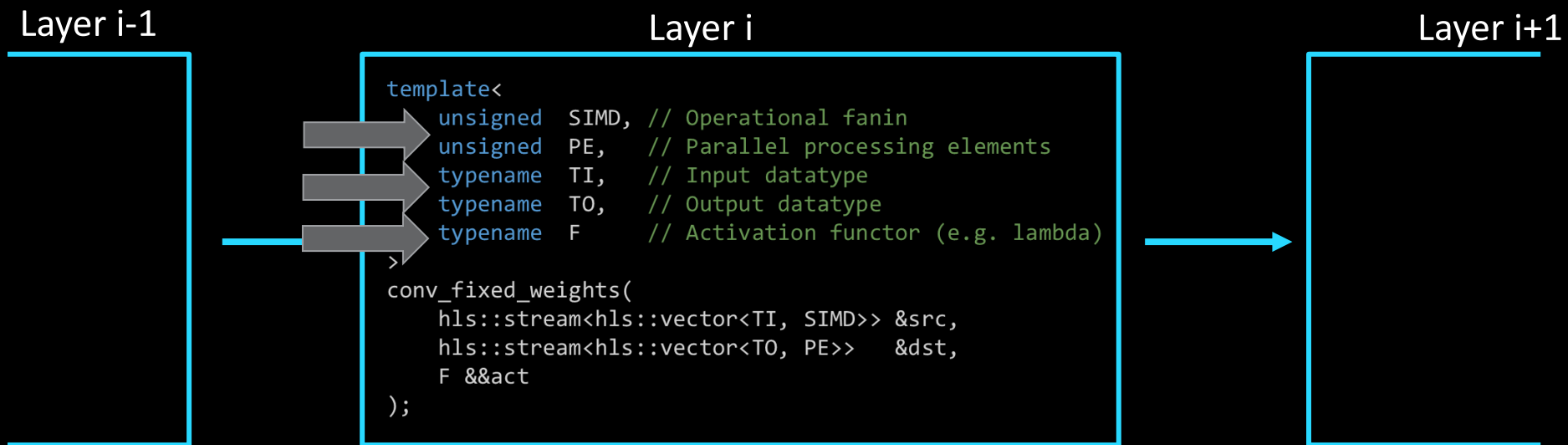
Jouppi, Norman P., et al. "Ten lessons from three generations shaped Google's TPUv4i: /ISCA'2021.

Dataflow - Adapt and Scale to Diverse Workloads

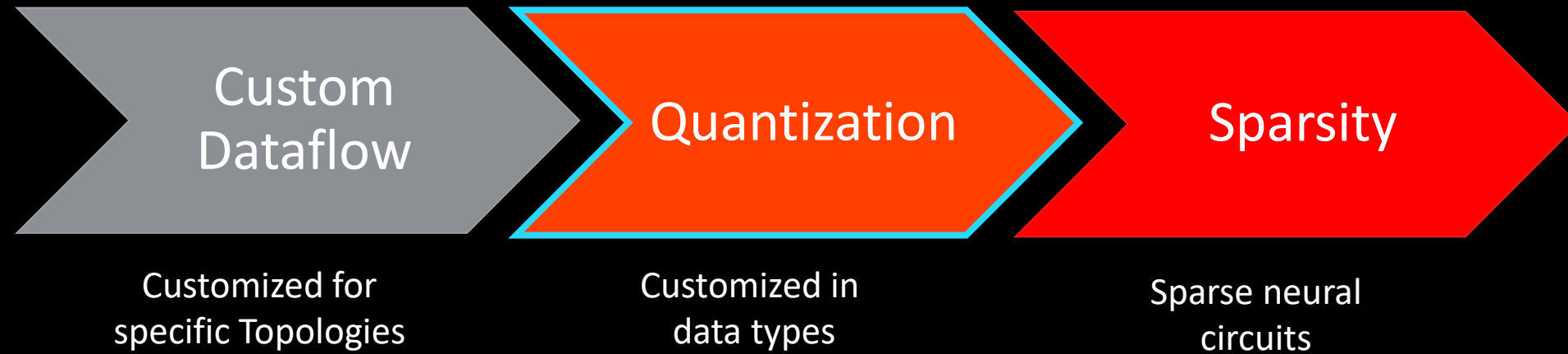


Dataflow - Parameterizable Kernel Library

- Kernels representing the individual layers, which can be parameterized
 - Degree of parallelism (output channels, input channels, kernel dimensions, ...) for different performance/resource trade-offs
 - Data types (INT8, ternary, INT2, ...)
 - Behavior (activation function)
- Composable through streaming I/O
- Programmed in synthesizable C++ (Vitis HLS)



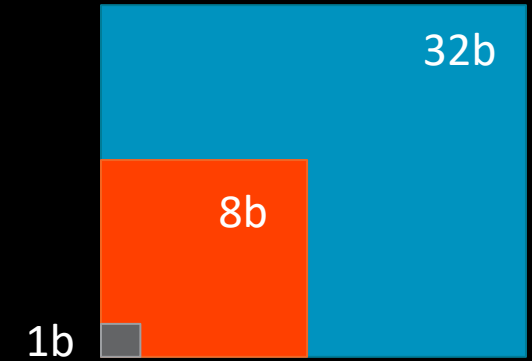
Key Concepts



Customizing Arithmetic to Minimum Precision

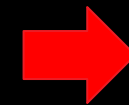
Quantization

- Reducing precision shrinks hardware cost/scales performance
 - For integer datatypes, LUT cost proportional to bitwidths in weight and activations (e.g., INT1 : INT8: 70x)
 - Instantiate n-times more compute within the same fabric, thereby scale performance n-times or shrinks hardware cost
- Energy
 - Faster execution => less energy ($E = P * time$)
 - Using reduced precision operators saves energy
 - Reduces memory footprint
 - ResNet50 @ 32b: 102.5 MB, ResNet50 @ 2: 6.4 MB
 - NN model can stay on-chip => no external memory access => saves energy



Operation		Picojoules per Operation		
		45 nm	7 nm	45 / 7
+	Int 8	0.03	0.007	4.3
	Int 32	0.1	0.03	3.3
	BFloat 16	--	0.11	--
	IEEE FP 16	0.4	0.16	2.5
	IEEE FP 32	0.9	0.38	2.4
×	Int 8	0.2	0.07	2.9
	Int 32	1.48	1.48	2.1
	BFloat 16	--	0.21	--
	IEEE FP 16	1.1	0.34	3.2
	IEEE FP 32	3.7	1.31	2.8
SRAM	8 KB SRAM	10	7.5	1.3
	32 KB SRAM	20	8.5	2.4
	1 MB SRAM ¹	100	14	7.1
GeoMean ¹				2.6
		Circa 45 nm	Circa 7 nm	
DRAM	DDR3/4	1300 ²	1300 ²	1.0
	HBM2	--	250-450 ²	--
	GDDR6	--	350-480 ²	--

¹ is pJ per 64-bit access.



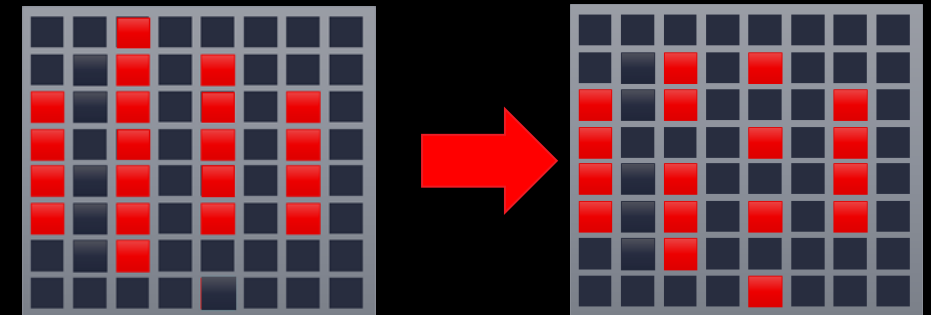
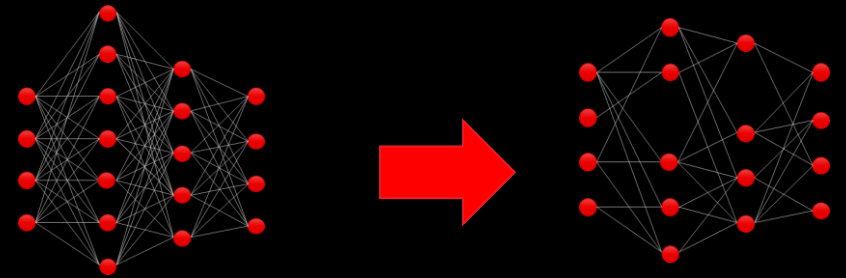
Jouppi, Norman P., et al. "Ten lessons from three generations shaped google's tpuv4i: /SCA'2021.

Key Concepts



Sparsity – Energy Efficiency

- DNNs are naturally sparse
- Massive scope to improve ML efficiency through sparsity
 - The human brain is highly sparse (98%) & operates on the power of a light bulb (~20W)*
- Sparse topologies result in irregular compute patterns which are difficult to accelerate on vector- or matrix-based execution units
 - Poor efficiency
- With streaming dataflow architectures, where every neuron and synapse is represented in the hardware, we can maximize efficiency



FPGA

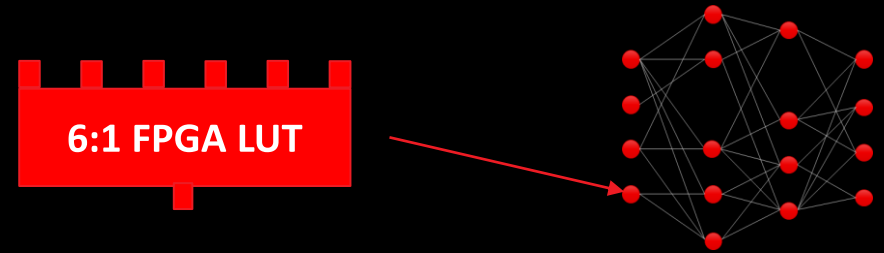
**Optimized
Dataflow
on FPGA**

Sparsity – Extreme Codesign with LogicNets

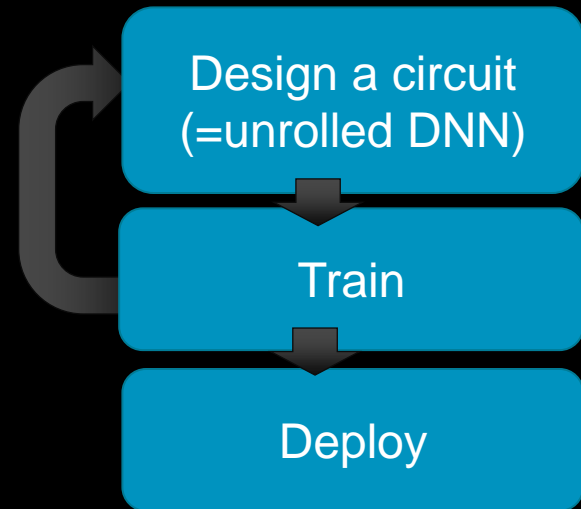
- **Idea**

- A LUT in an FPGA can represent a neuron
- Design a highly sparse circuit in an FPGA
- Represent this as a DNN to the training framework
- Learn the LUT contents

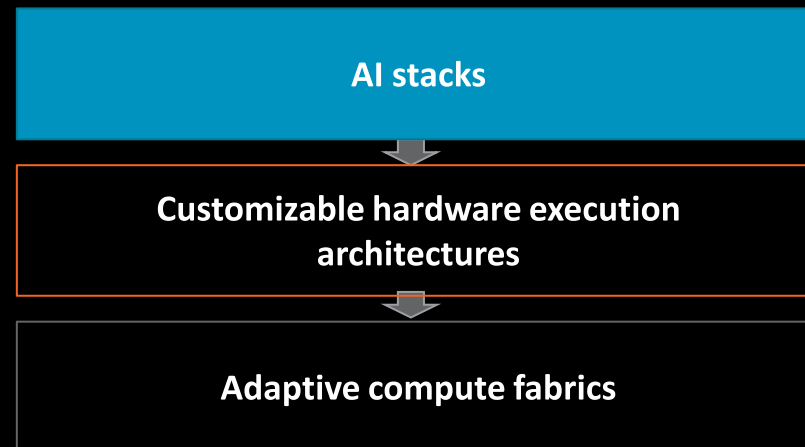
High-efficiency and maximum performance by design (classification at clock rate)



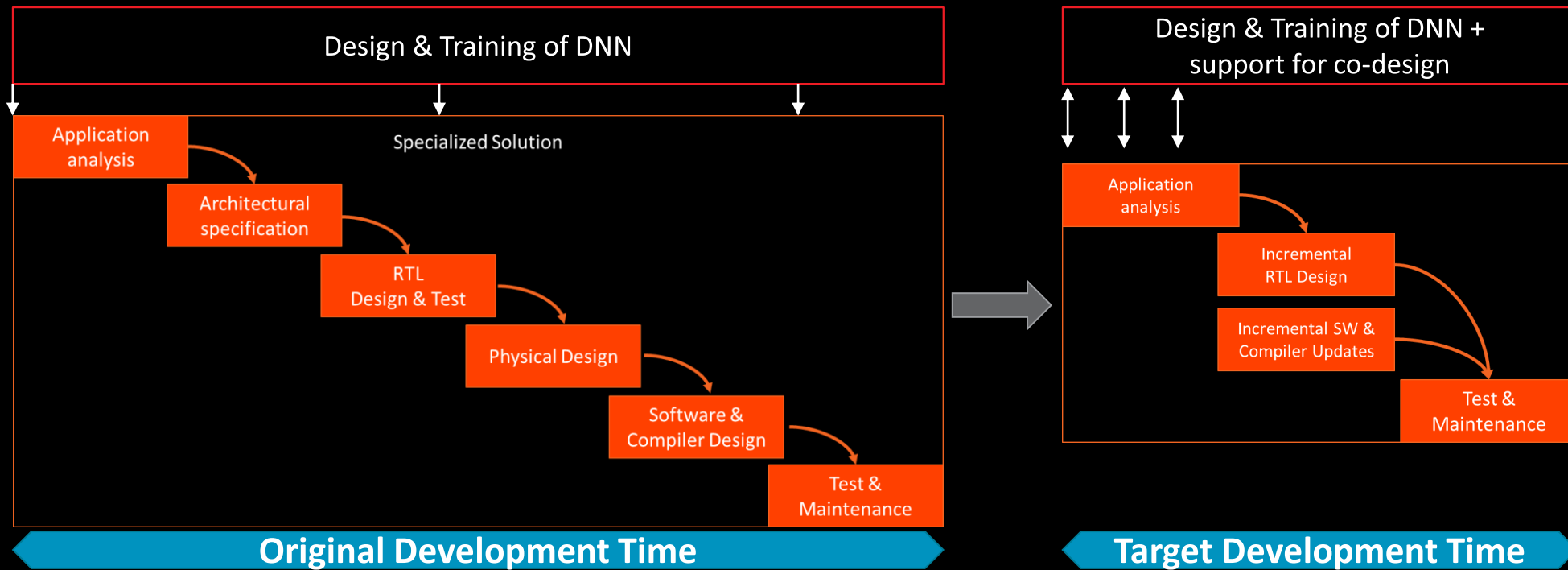
Adjust the parameters of DNN (=LUT contents) while iterating on training dataset until accuracy



How can we support this specialization through agile AI stacks? (FINN with Brevitas)



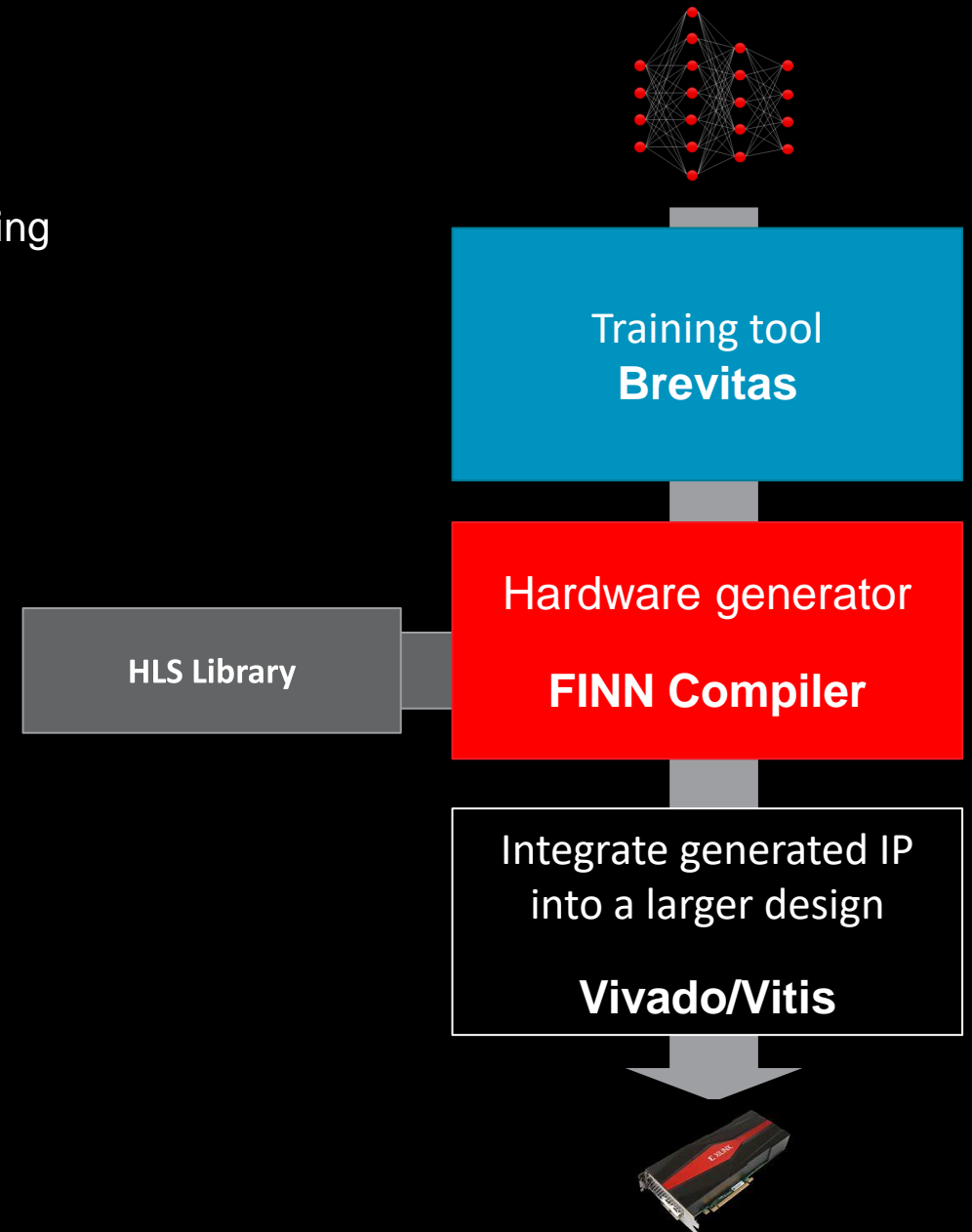
Faster Iterations with Shortened Development Cycles



- Adaptive Computing eliminates the need for physical design
- Generalizable architectures which can incrementally adapt to new requirements
- Paired with graph compiler which automates the specialization
- Agile quantization support in training library

Example: & Brevitas

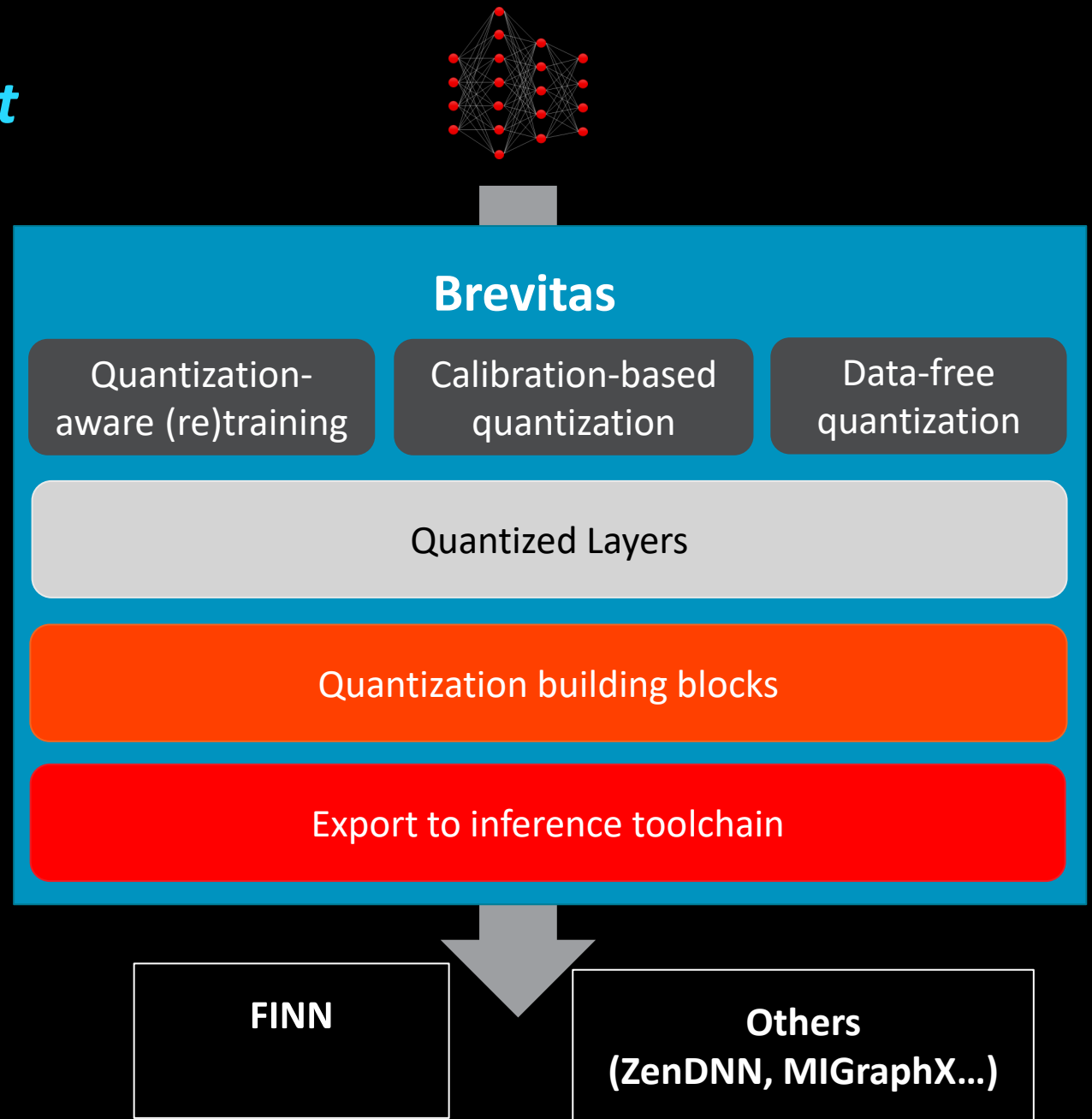
- ▶ End-to-end flow – from DNN to bitstream
 - Enables generation of highly customized hardware architectures using **quantization** and **dataflow** and **fine-granular sparsity**
- ▶ Components
 - Training tool: Brevitas
 - Hardware generator (FINN)
 - Kernel library (HLS)
- ▶ Open-source
 - Easy collaboration with customers
 - Flexibility to adapt to fast-moving application space
 - Third-party contributions



Brevitas - PyTorch Library

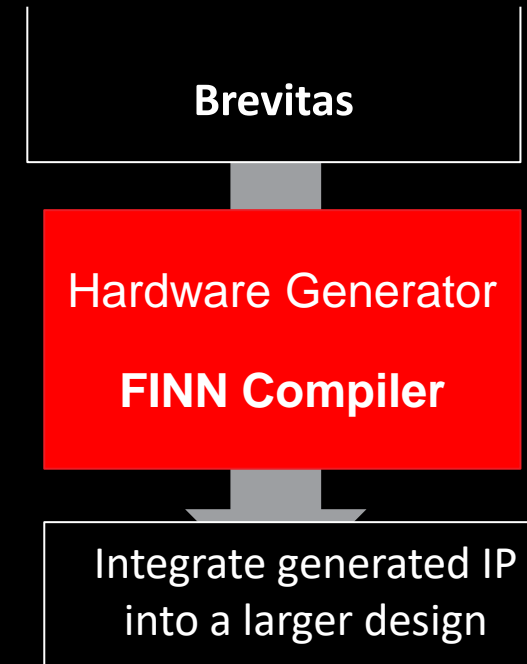
Offering Agile Quantization Support

- First class support for custom data types and operators at ML framework level
 - Arbitrary precision integer, float, block-style quantization
 - Extendible to user-defined datatypes and operators and support for any hardware-specific datatype at training
- Composable building blocks at multiple abstraction levels that can be arbitrarily combined
- Integration with different compiler stacks
 - Exports commonly used representation format (for example ONNX)

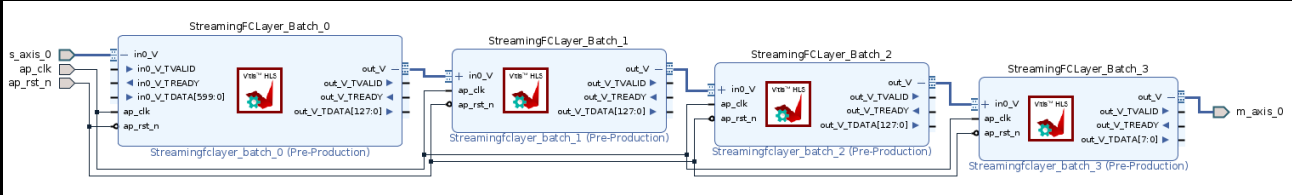


FINN Compiler

- Modular **graph compiler** with well-defined abstraction levels
- Incrementally lowers ONNX graph to a hardware description through **transformations**
- Performs **optimizations**
 - Layer fusion
- Explores the **design space**
 - Calculates the degrees of parallelism for each kernel using resource cost and performance models
- **Code-generates** a dataflow C++ description using the parameterizable kernel library
- Creates **DNN hardware IP**



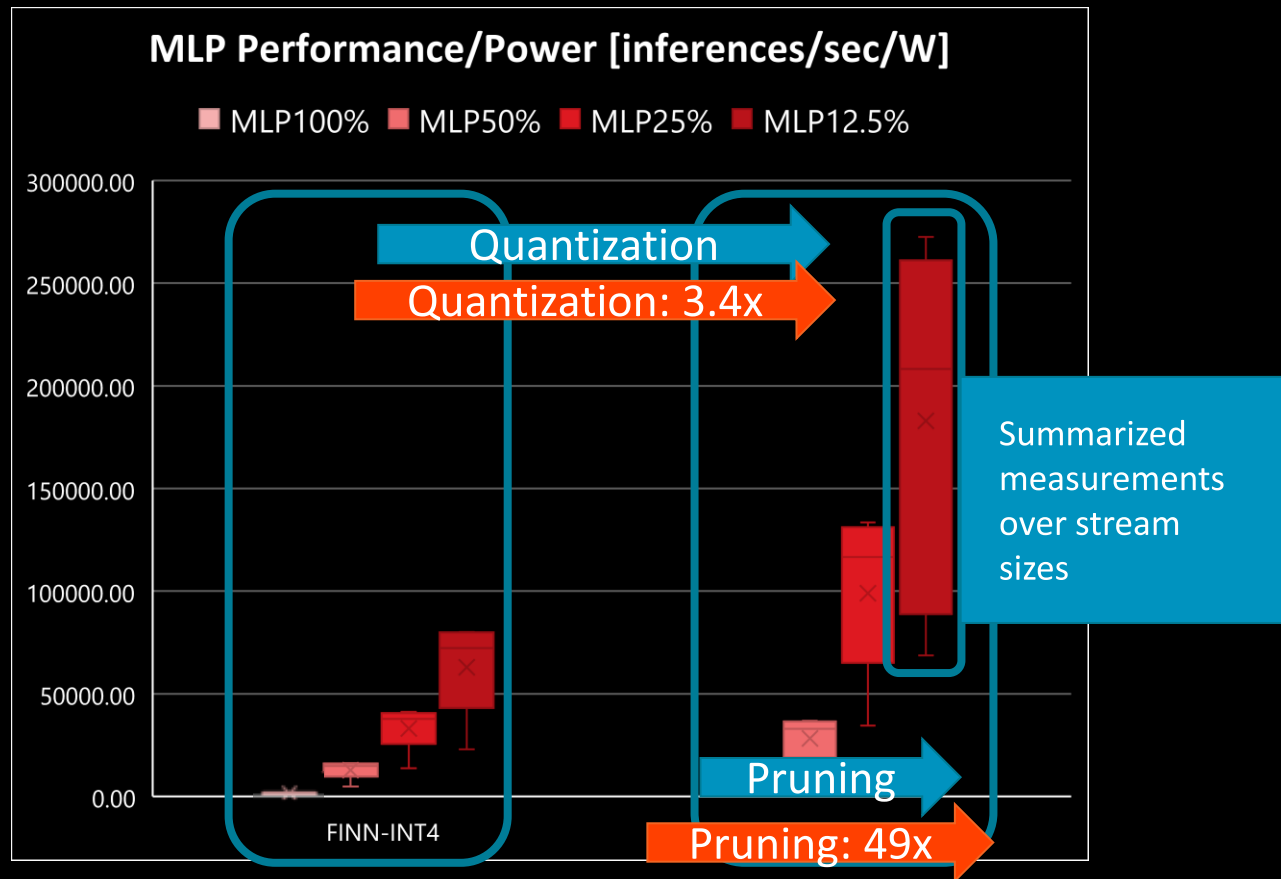
```
hls::stream<ap_int<185>> in
hls::stream<ap_int<100>> inter0, inter1, ...
...
StreamingFCLayer<BINARY, BINARY, ..>(in, inter0, ...)
StreamingFCLayer<BINARY, BINARY, ..>(inter0, inter1, ..)
...
```



Some Example Results

Energy Efficiency through Quantization and Sparsity

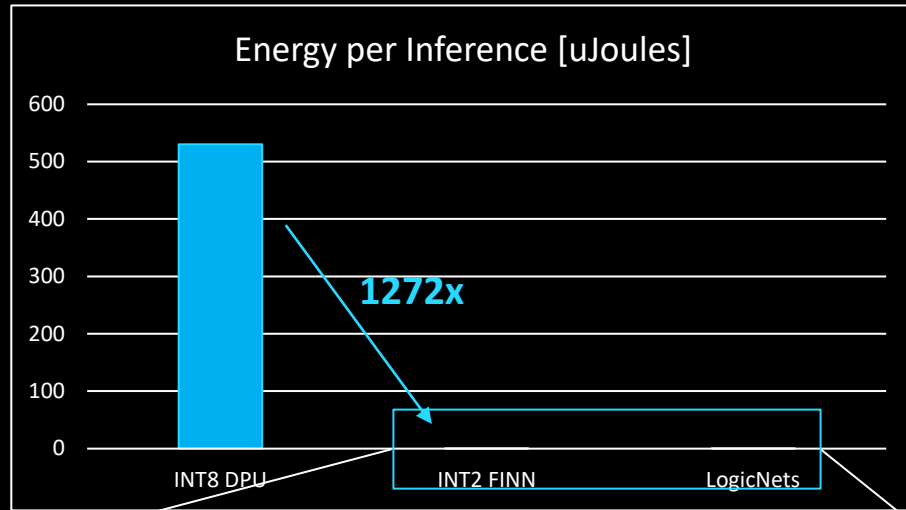
- Benchmarking activity* across topologies, devices, and optimization schemes
- Example representing typical behavior: one MLP and one CNV, using quantization & pruning on an FPGA (FINN)



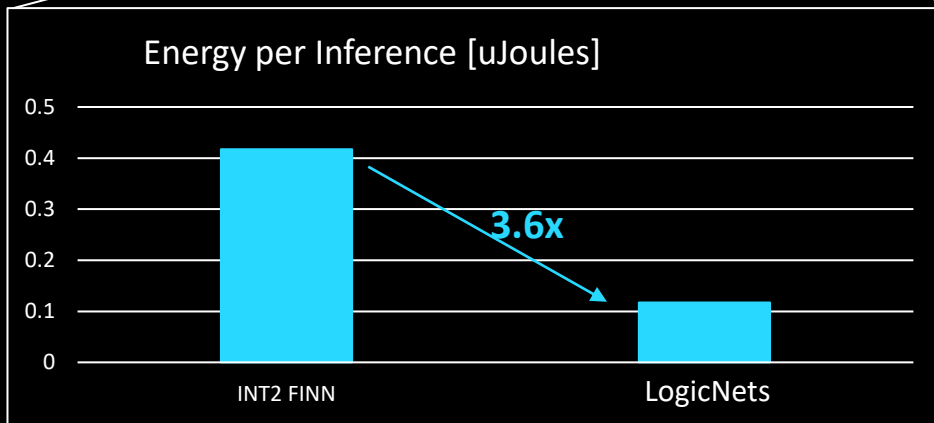
Significant energy efficiency through pruning and quantization on FPGAs possible

Energy Efficiency: FINN & LogicNets

Results Demonstrate the Potential



Reducing precision & Dataflow =>
1272x improvement



LogicNets: 3.6x over FINN

Energy calculated
LogicNets assume

Total: ~4500x Energy Improvement through Post-Silicon Hardware Specialization
Much more work coming...

Details:
Network Security Application
Malware Classifier
UNSW dataset
MLP 92k Ops/inference
INT8 with VitisAI,
INT2 with Brevitas and FINN
Board power ZCU104

Cyber Security – Line-rate Classification with Nanosecond Latency



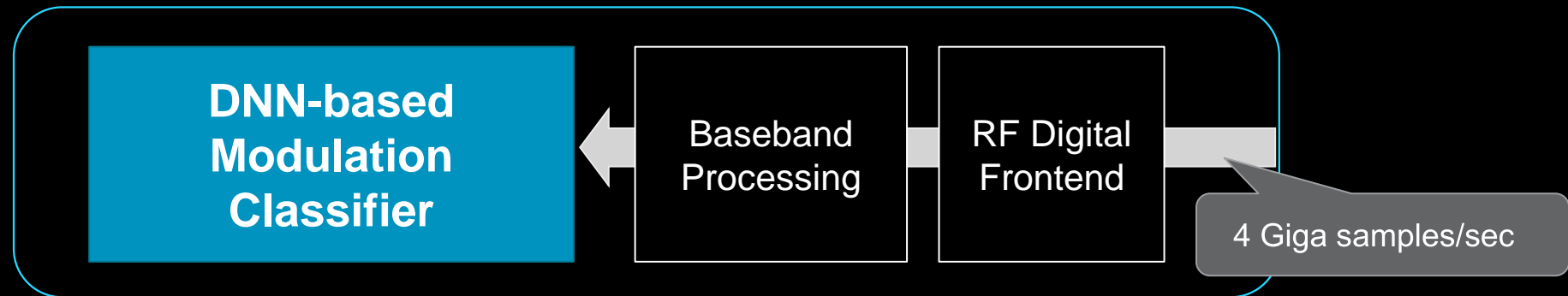
- FINN implementation of UNSW-NB15 malware classifier
 - 2b weights & activations
 - 91.9% accuracy
 - 300M inferences/sec with 18 nsec latency
 - 8k LUT
- FINN implementation of DDoS classifier trained on CIC-IDS2017 dataset
 - 2b weights & activations
 - 85% F1-score (binary classification using flow-based per-packet features)
 - 19.2M inferences/sec, 52nsec latency
 - 18.6K LUTs

Work in progress:
Expected to scale to 300M
inferences/sec too ...

Diversity

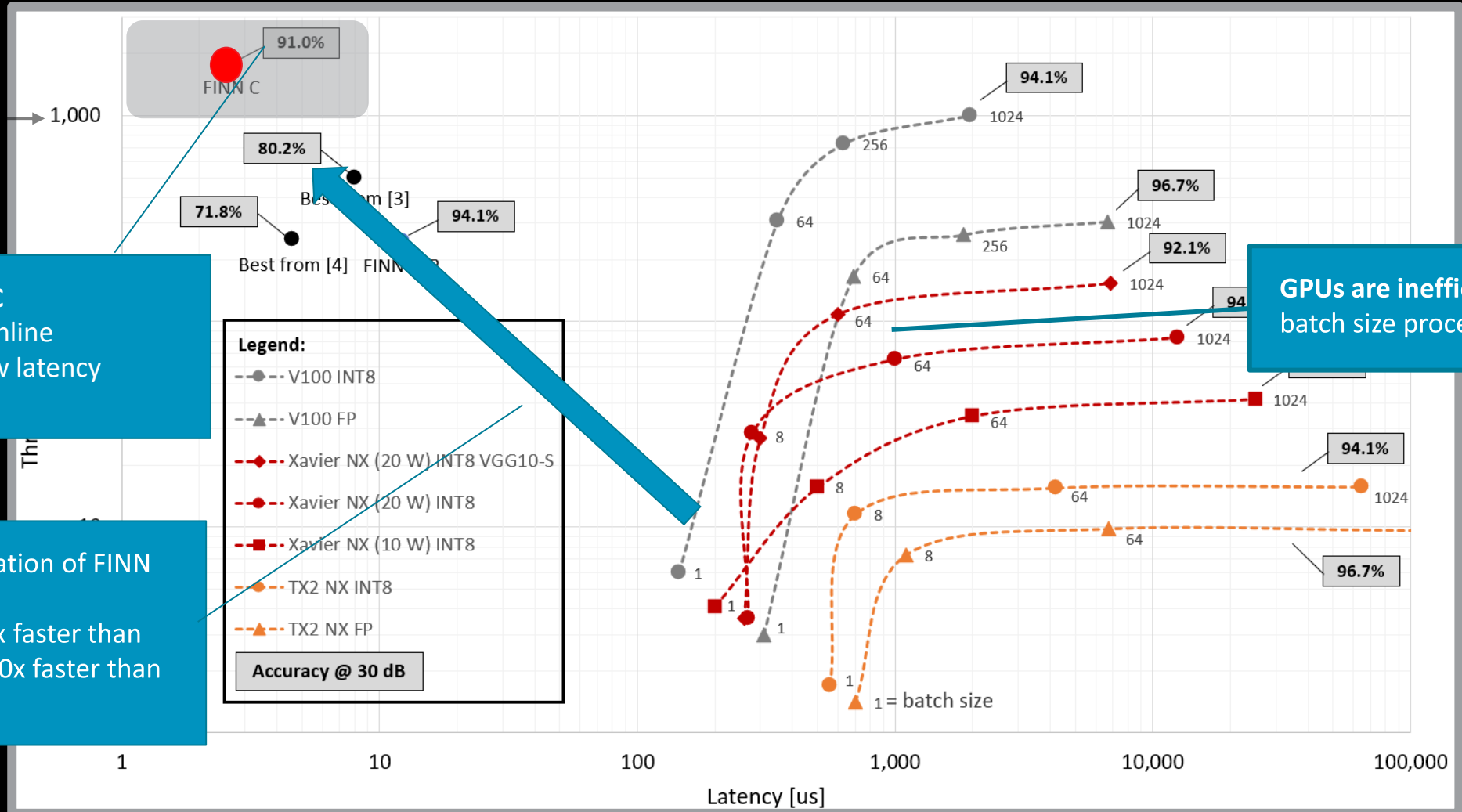
Modulation Classification: GHz sampling rate & usec latency

- What's in my RF spectrum? Rapidly label + understand RF spectrum
 - What modulations are used?
- Key enabler for many applications and key component of an AI-enabled (cognitive) software-defined radio
 - For example, spectrum interference monitoring, dynamic spectrum access
- DNNs promising for modulation classification



Challenge: At GHz sampling, we need Minfps inference throughput

DNN-Based Modulation Classification (RadioML)



> GHz

Dataflow on RFSoc
Enables real-time inline processing with low latency

Customer Evaluation of FINN confirms:
measured >100x faster than Xeon CPU and 10x faster than GPU

GPUs are inefficient at low batch size processing

Low latency

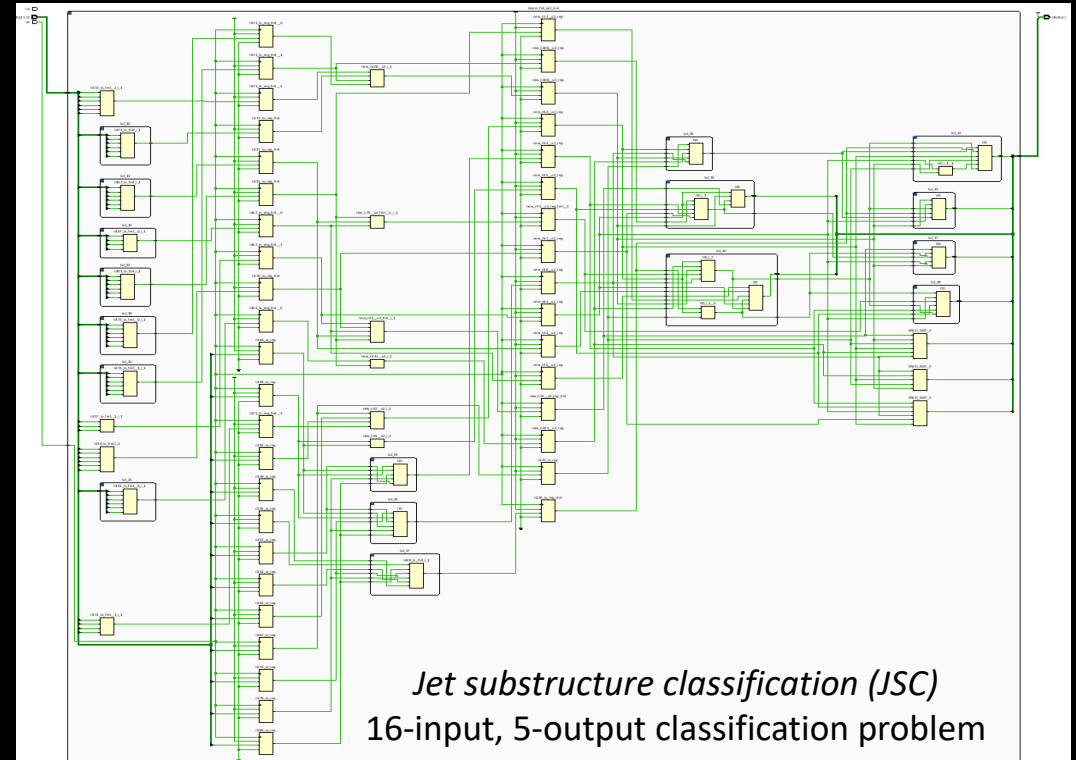
DF on a ZCU111: 1.75 GSamples/sec, 2.6 usec latency

Diversity

LogicNets Results – Tiny (!!!) and Fast

- **DNN in similar area compared to an FPGA 32b adder**
- **High-energy particle physics CERN L1 trigger experiment**
 - Inference rate: 666 Minferences/sec*
 - Latency: 3 nsec
 - Resources: 30 LUTs

A Complete Neural Network @ 70% Accuracy!

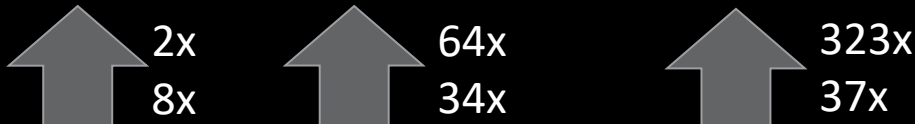


Diversity

LogicNets Results

- Quotation from Petersen et al., Dec 2022 @ NeurIPS:
 - “FINN [...] the **fastest method** for classifying MNIST at an accuracy of 98.4%,”*

	Acc. [%]	LUT	Latency [nsec]	Inferences/sec
✓ FINN	98.4	83k	2,440	1.6M
	95.8	91k	310	12.4M



	Acc. [%]	LUT	Latency [nsec]	Inferences/sec
LogicNets-M	97.7	45k	38	517M
LogicNets-S	95.8	12k	9	458M

“World’s fastest MNIST classifier”* - now even faster

Synthesized with Vivado 2019.2; F_{Max} equals inference rate

*Petersen et al. "Deep Differentiable Logic Gate Networks." NeurIPS, 2022.

FINN: Diverse Engagements and Open-Source Adoption

- **Communications**
- **Medical**
- **Sensor Intelligence**
- **Automotive**
- **High-energy particle physics**
- **Aerospace & Defense**
- **High-frequency Trading**

- **Open-source Adoption**
 - **~2000 stars, 230k+ Brevitas downloads, 72k+ QONNX, 17k+ FINN compiler downloads**
- **Three best paper awards**
- **> 1000 citations**

Available: Customer support through AMD CSE organization

<https://xilinx.github.io/finn>

<https://github.com/Xilinx/brevitas>

Summary

Pervasive AI: dynamic and diverse long tail of AI applications

Paradigm shift towards energy efficiency

Enabling Rapid Specialization with Adaptive Compute Fabrics, Customized Execution Architectures and Agile AI Stacks

Adaptive computing available in great diversity and can help by customization of hardware execution architectures

- Dataflow, shrinking precision, fine granular sparsity

Speed-up and automate specialization through graph compilers such as FINN and training libraries Brevitas

Proof points from FINN, Brevitas and LogicNets demonstrate the potential for energy savings, and addressing truly diverse requirements

Heterogeneous Accelerated Compute Clusters (HACCs)

Focus on heterogenous and adaptive computing

- Supporting high-end compute research
- Bare metal access to adaptive compute hardware
- HACC community
- Growing community of over 100 institutions



www.amd-haccs.io

AMD HPC Fund

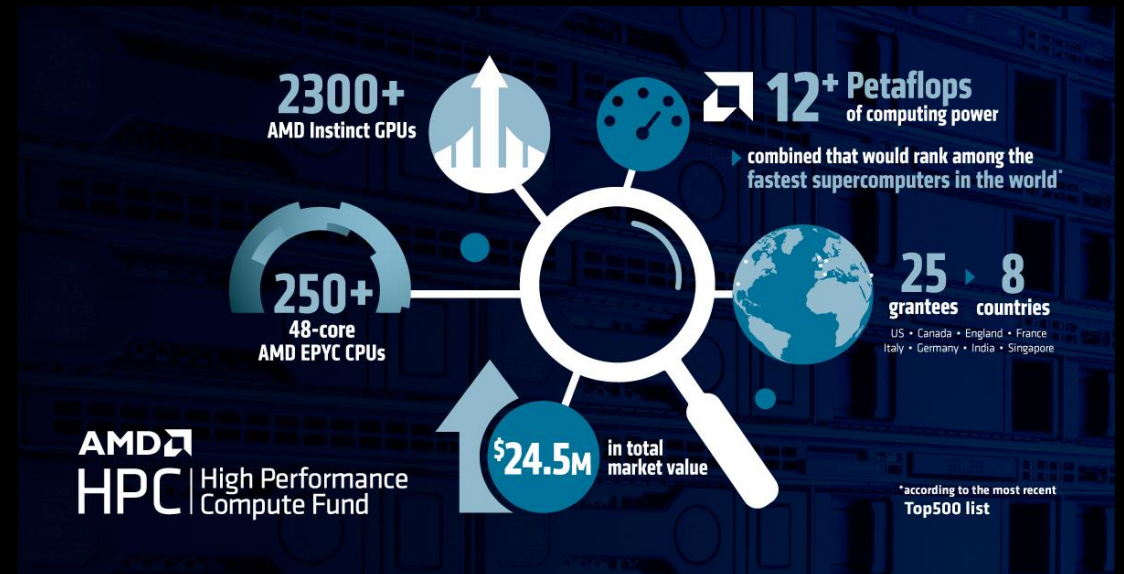
Accelerating Science in the Public Interest

- Cloud access to AMD HPC CPU & GPU technologies
- Customized technical training
- E-learning sessions
- Networking opportunities with peers around the world

Get involved!

<https://www.amd.com/en/corporate/hpc-fund>

<https://www.amd-haccs.io/>



AMD
EPYC

AMD
INSTINCT

XILINX
ALVEO™

XILINX
VERSAL™

AMD 

Abstract

- In the context of AI, we face a plethora of challenges that extend beyond the widely discussed performance scalability required to meet the growing demands of compute and storage in the latest models. These challenges encompass sustainability, pervasiveness, agility, and diversity, all of which are needed to cater to a constantly evolving range of applications and algorithms from endpoint to edge and cloud. In this talk, we explore how AMD adaptive devices and agile compiler stacks can provide solutions by delivering post-production hardware specialization and co-designed algorithms. This results in highly optimized AI systems which not only provide the necessary performance scalability but also bring a reduction in carbon footprint while addressing the needs of a broad range of diverse applications with the necessary agility.