Ubiquitous AI by the Agile Generation of Inference Solutions

Dr. Thomas Preusser
AMD Research and Advanced Development

AMD
together we advance_

Itinerary

1. RAD Integrated Communications & AI Lab
2. Broader AI Context
3. Long-Tail Challenges
4. Technological Levers
5. Brevitas & FINN for Enabling Agility
6. Conclusions

AMD · RAD Integrated Communications & AI Lab

Team
- ~25 top-class researchers/PhDs
- Europe, Singapore, US
- With broad expertise in AI, compilers and compute architectures
- Highly active internship program, ~10 across the different sites

Subjects
- Model optimization & quantization
- Customized compute architectures for inference acceleration
- Networking for distributed AI

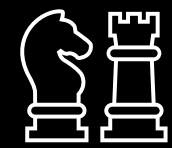
Close Collaboration
- In close collaboration with universities, partners and customers

Broader AI Context

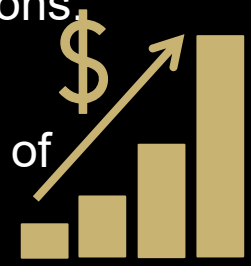
Broader AI Context

- **Many Views and Beliefs (bulls & bears)**
  - **Transcending the Human Scale**: Master information volume beyond the bandwidth of an individual.
  - **Clever Hans Effect**: User clues and observer bias wanting to recognize intelligence.

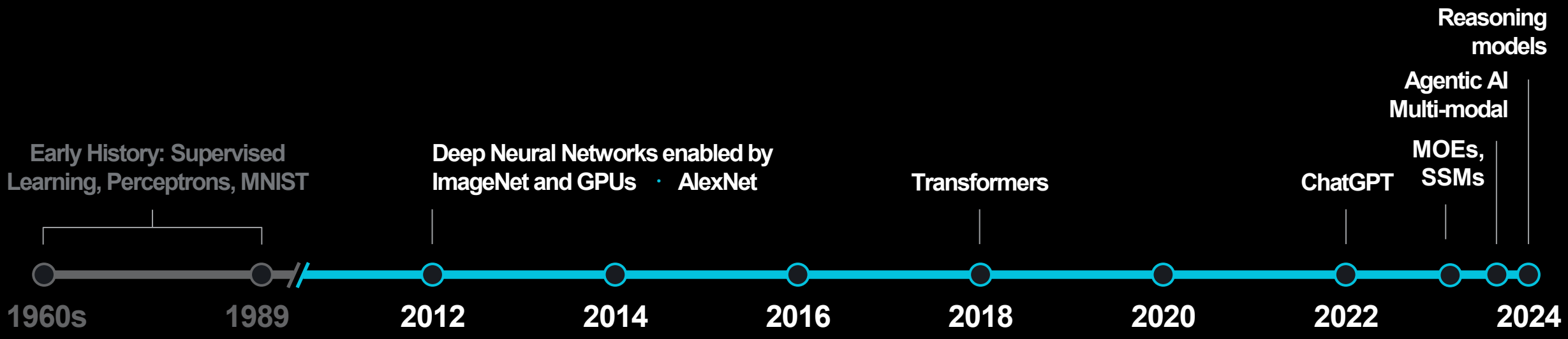
- **Artificial General Intelligence (AGI) Debate**
  - AGI will be attained through scaling models, more compute & data.
  - The huge VC investments demonstrate the broad consensus on a splendid future.
  - Scaling AI by growing compute & data exponentially is not making the models more ***intelligent***.
  - An imagined superpower overrides all risks and fuels a hype running on the fear of missing out.



- **Heavy Investment** is happening and will create opportunities across a wide spectrum of applications.
- State actors are joining the race on sovereignty grounds.
- Tremendous activity that will ultimately reveal where we're able to land between the dumb scaling of an imposter emulation and an intelligently reasoning agent.

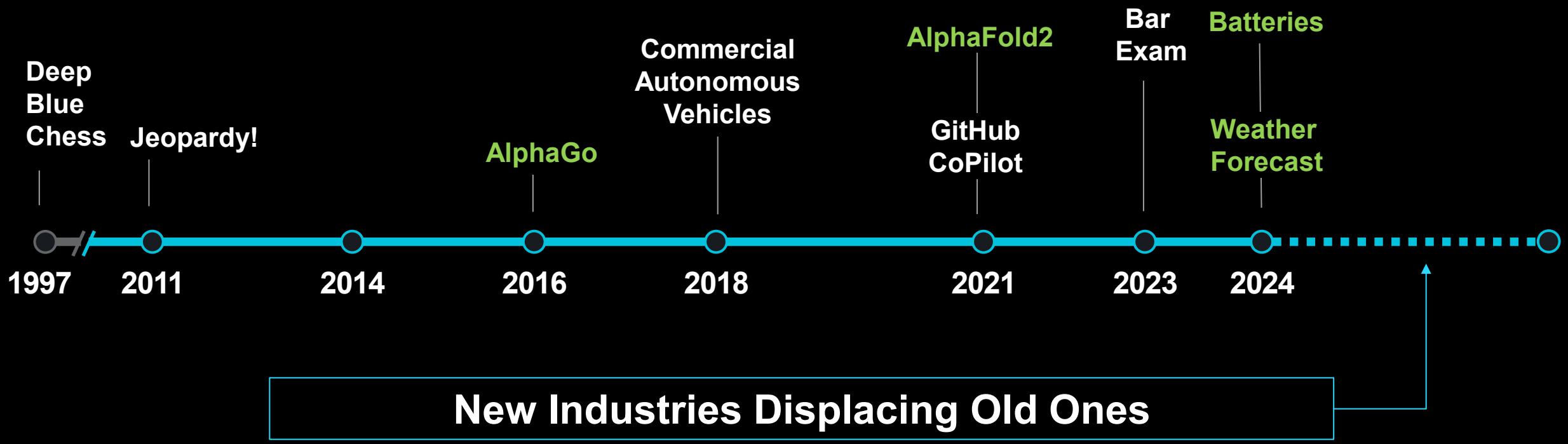


Evolution of AI is Accelerating



Charlie Giattino, Edouard Mathieu, Veronika Samborska and Max Roser (2023) - "Artificial Intelligence" Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/artificial-intelligence' [Online Resource]

Inflection Points

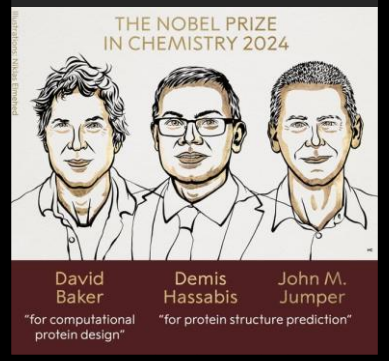


Eric Schmidt: “The AI Revolution is underhyped”

https://www.ted.com/talks/eric_schmidt_the_ai_revolution_is_underhyped

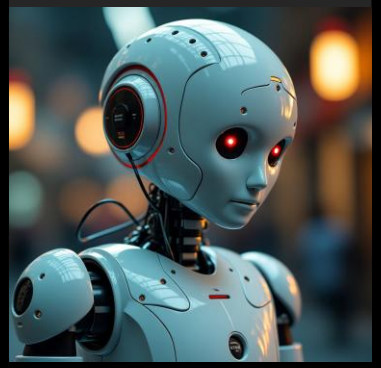
Inflection Points

Healthcare & Life Sciences



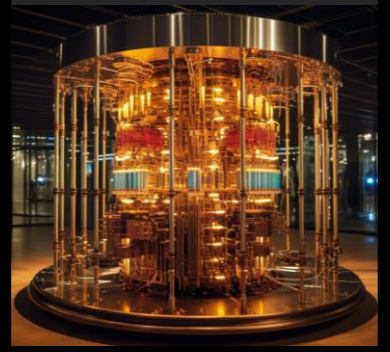
Code Productivity Tooling

Robotics



Computer Graphics

Next Gen HPC with AI & Quantum



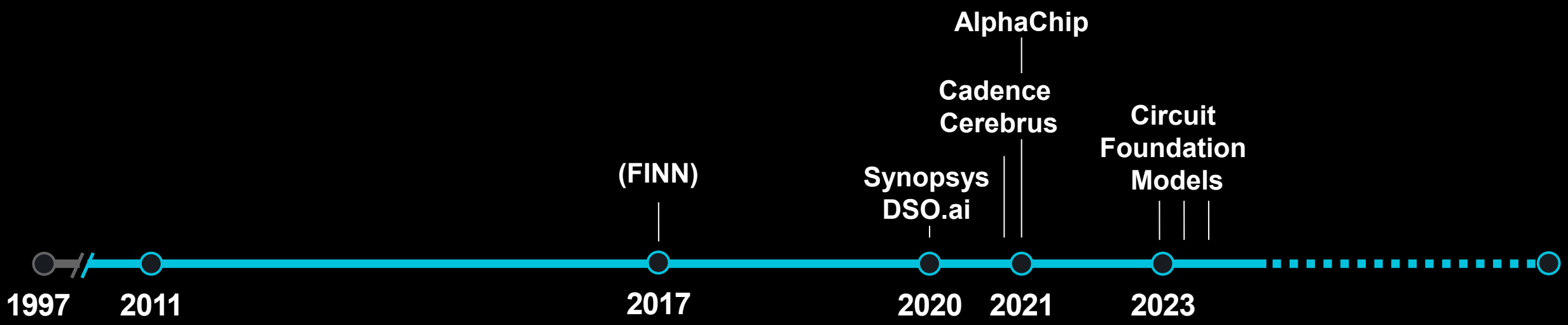
...

Today's AI

AI Revolution

AI Pervasiveness

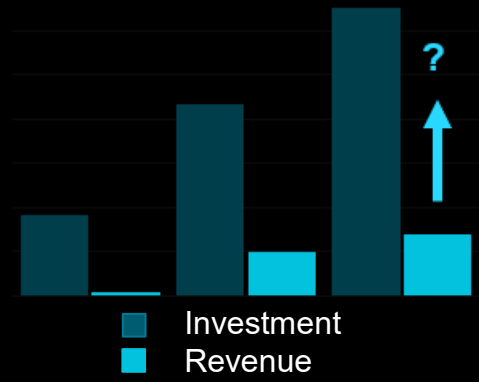
First Inflection Points in EDA!

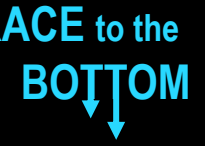


[2504.03711] A Survey of Circuit Foundation Model: Foundation AI Models for VLSI Circuit Design and EDA

Enabling the AI Revolution
Economic Challenges

- **Monetization challenge**
- **Key contributor: High inference cost**
  - Expensive cloud service due to compute and memory intensive workloads
- **Race-to-bottom dynamic for inference pricing**
  - Large Language Models (LLMs) similar in their capabilities and easy to switch
  - Short time window to amortize R&D cost
  - Open-source models closing in (DeepSeek, Qwen)





Inference Cost – How can we lower the computational cost of running inference?

Today's AI — AI Revolution — AI Pervasiveness

Reported revenue for 2023, internally estimated numbers based on https://www.sequoiacap.com/article/ais-600b-question/ for visualization

There Must Be Headroom

AI requires MegaWatts

est. 10^5 x

Human brain ~20Watts

- Three Mile Island is reopening and selling its power to Microsoft (CNN Sep. 2024)
- Transforming power delivery with 1MW per rack from 48 Volts direct current (VDC) to +/- 400 VDC [1]
- “A nuclear fusion breakthrough is needed to meet the vast energy requirements of future artificial intelligence.” — Sam Altman, CEO of OpenAI [2]

Images generated with Stable Diffusion

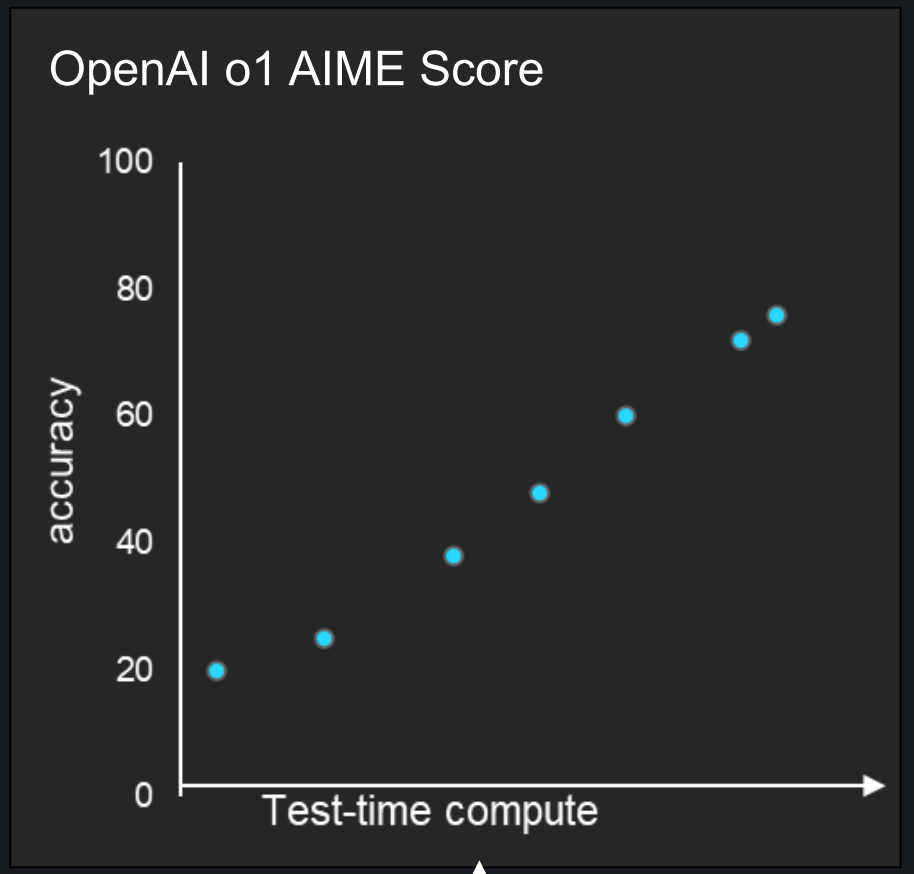
[1] 2025 OCP EMEA Summit, AI infrastructure is hot. New power distribution and liquid cooling infrastructure can help
[2] https://www.independent.co.uk/tech/openai-ai-nuclear-fusion-energy-b2479985.html

Can we get more efficient?

- **Linear improvements with exponentially more compute?**

And is this really what we call intelligence?

- **“Prompted DeepSeek R1 to choose a number between 1 and 100” (LLMDevs)**
  - Took 96seconds

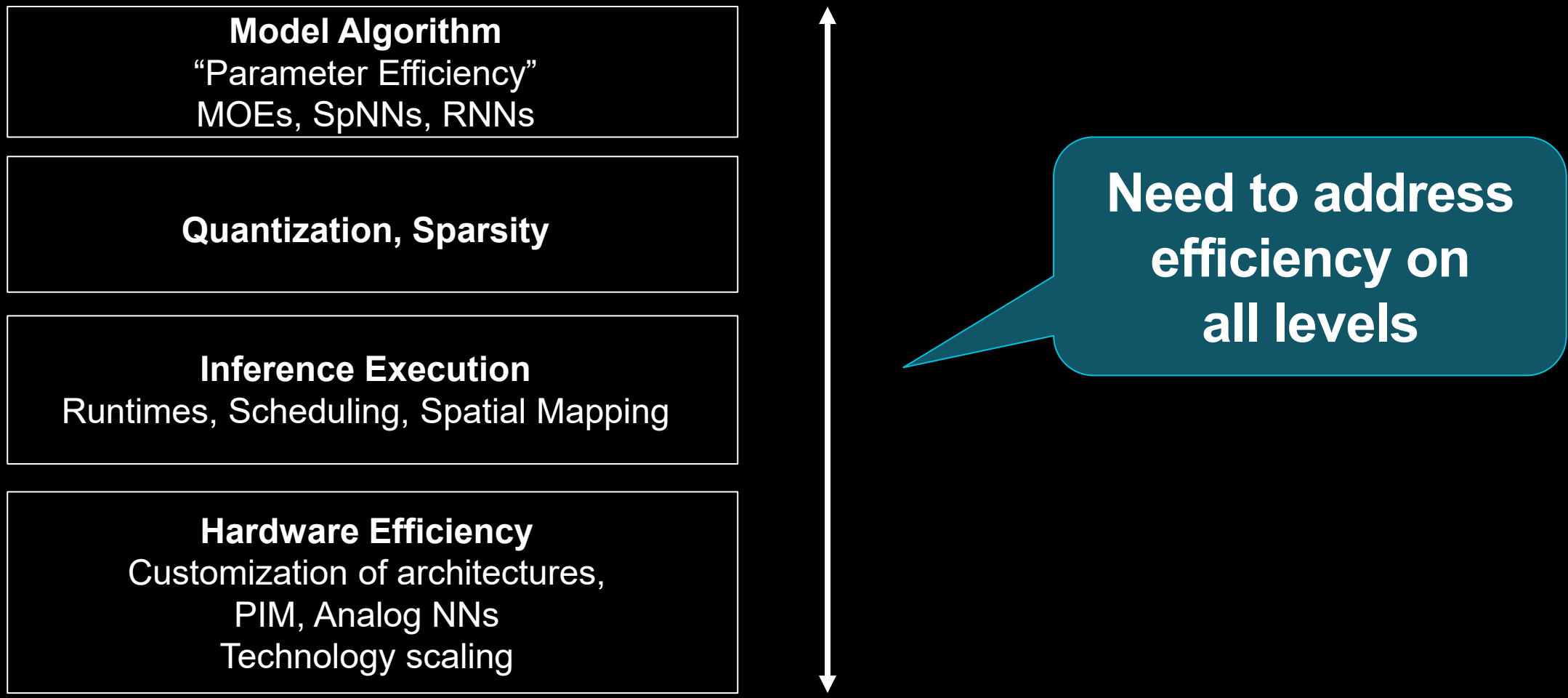


Log scale x-axis!

Data from https://openai.com/index/learning-to-reason-with-llms

https://www.reddit.com/r/LLMDevs/comments/1ieyflc/prompted_deepseek_r1_to_choose_a_number_between_1/

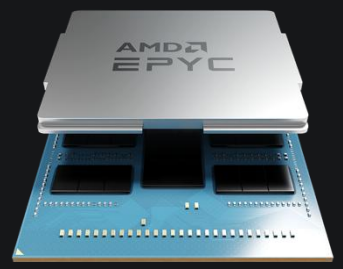
Inference Efficiency Enablers



Efficiency through Customization

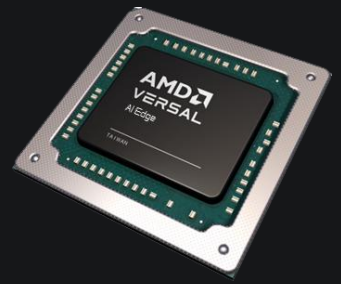
Energy Efficiency Focus across Product Range

at all levels



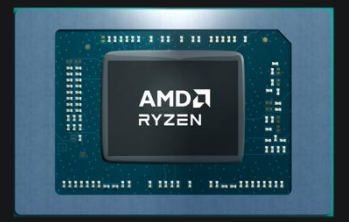
4th Gen
AMD EPYC™
Processors

AMD Instinct™
AI Accelerators



AMD Versal™
FPGAs and SoCs

AMD Radeon™
GPUs

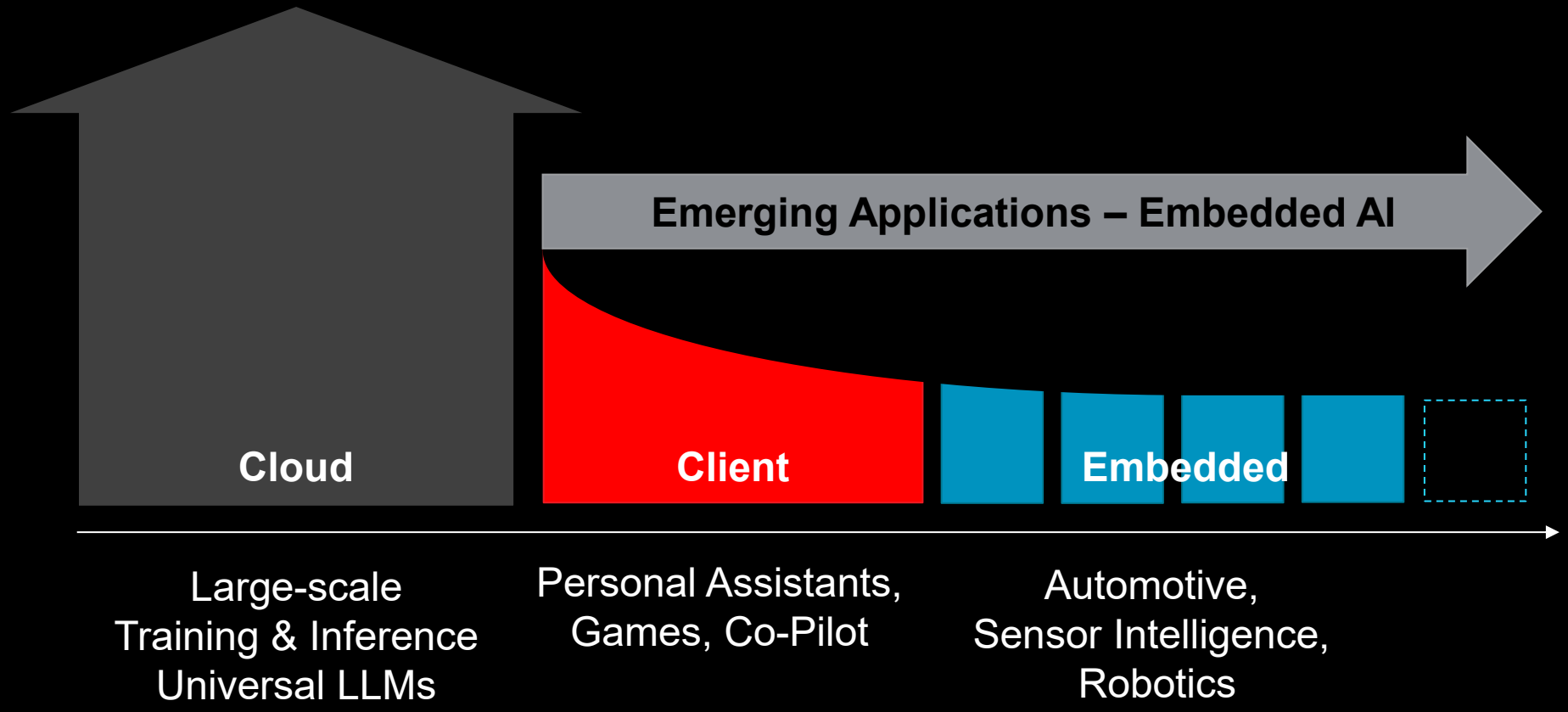


AMD Ryzen™
Mobile Processors

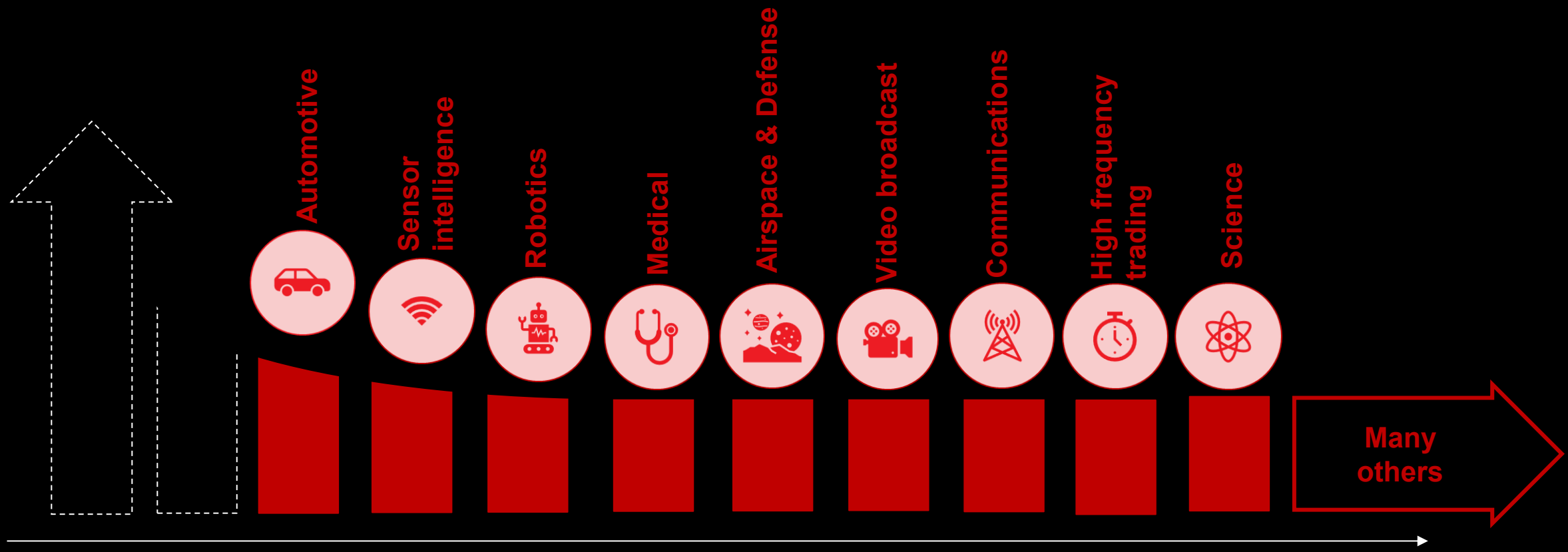
From Cloud to Embedded

Long-Tail Challenges

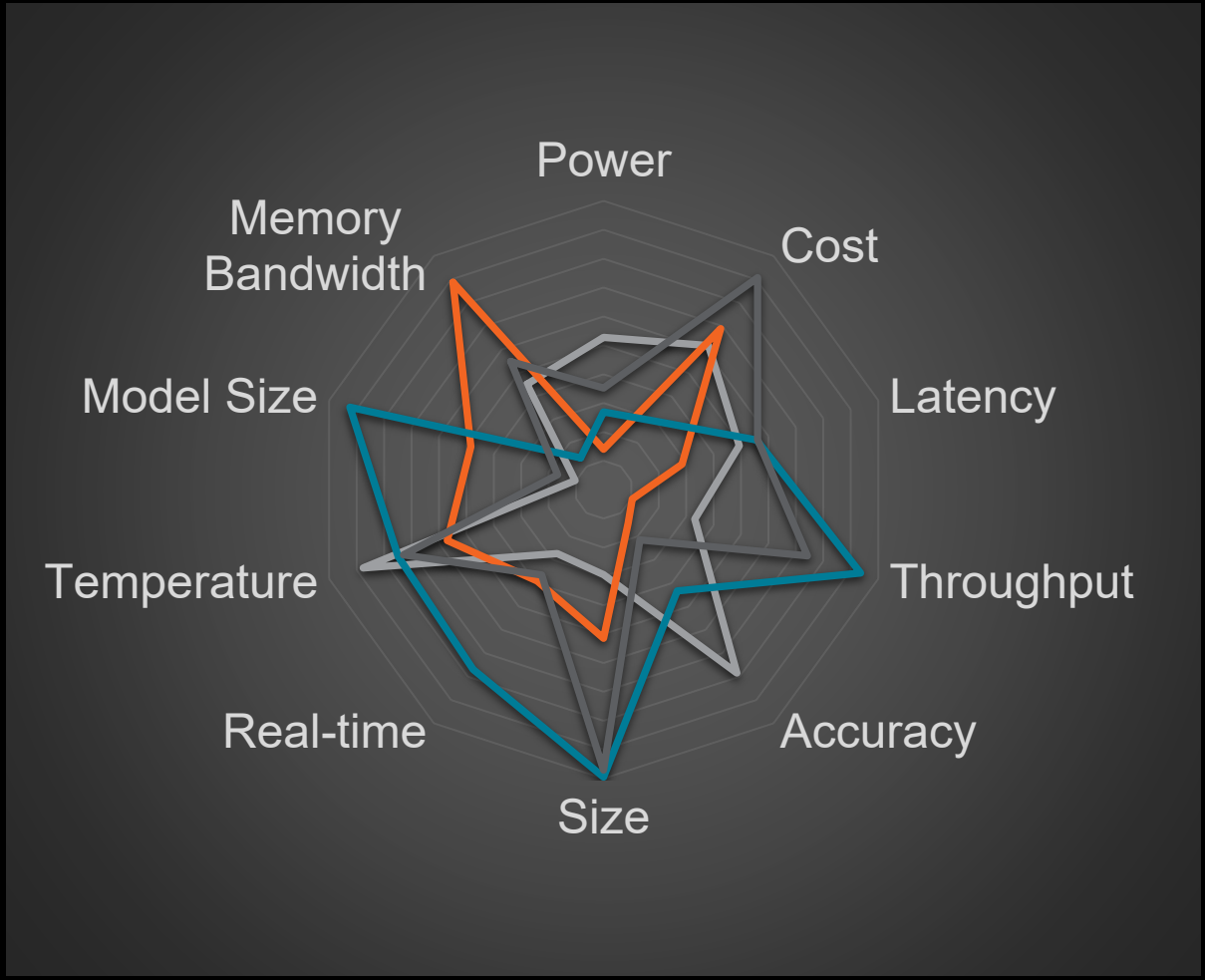
Ubiquitous AI



Many Application Domains in the Embedded Space



Diverse Requirements & High Degree of Specialization

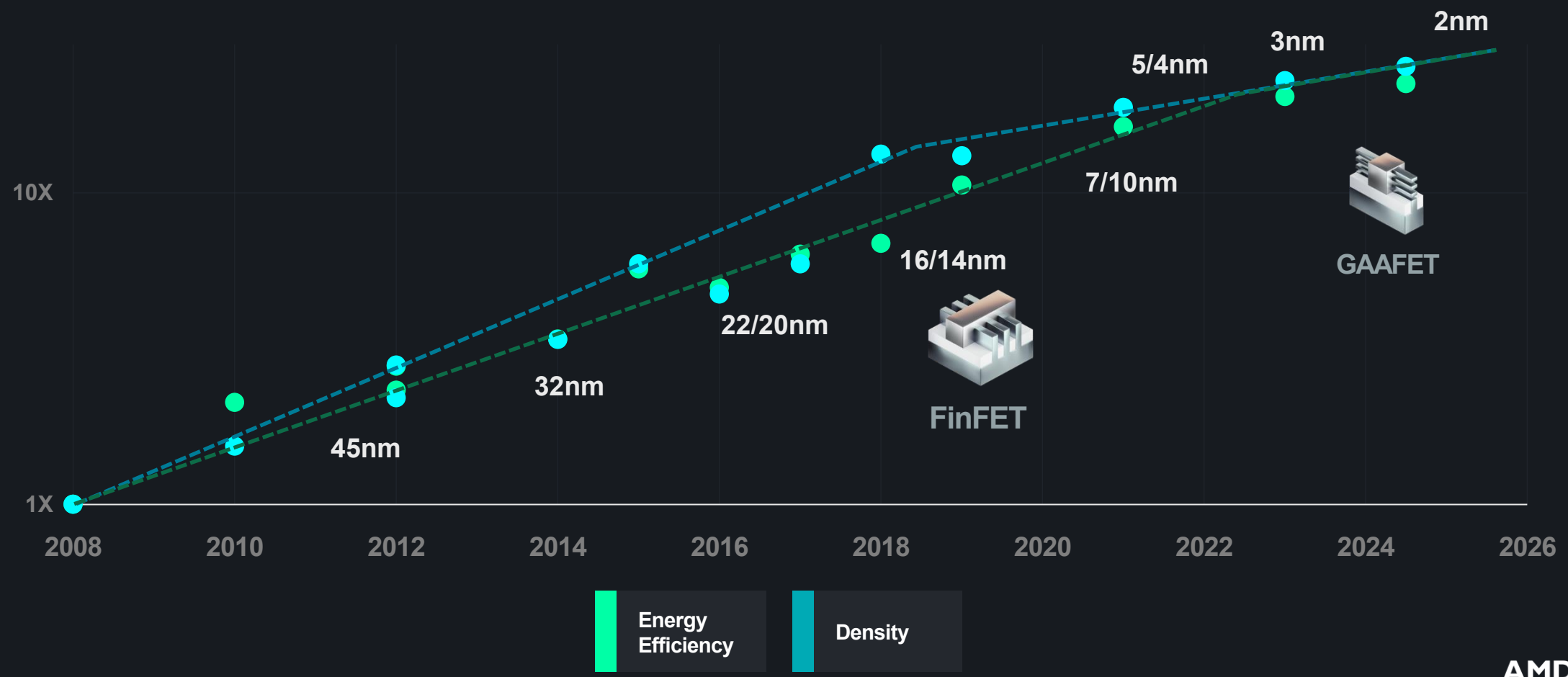


- **System specialization serves the diversity of requirements.**
- **Demanding combinations of constraints must be accounted for in used building blocks.**
- **Benefit: Well-scoped problems escape the AGI question. It's all about scaling – albeit "dumb".**

Technological Levers

The Silicon Technology Dividend

Gains are Slowing but Essential



Based on AMD internal estimates

The Advanced Packaging Dividend

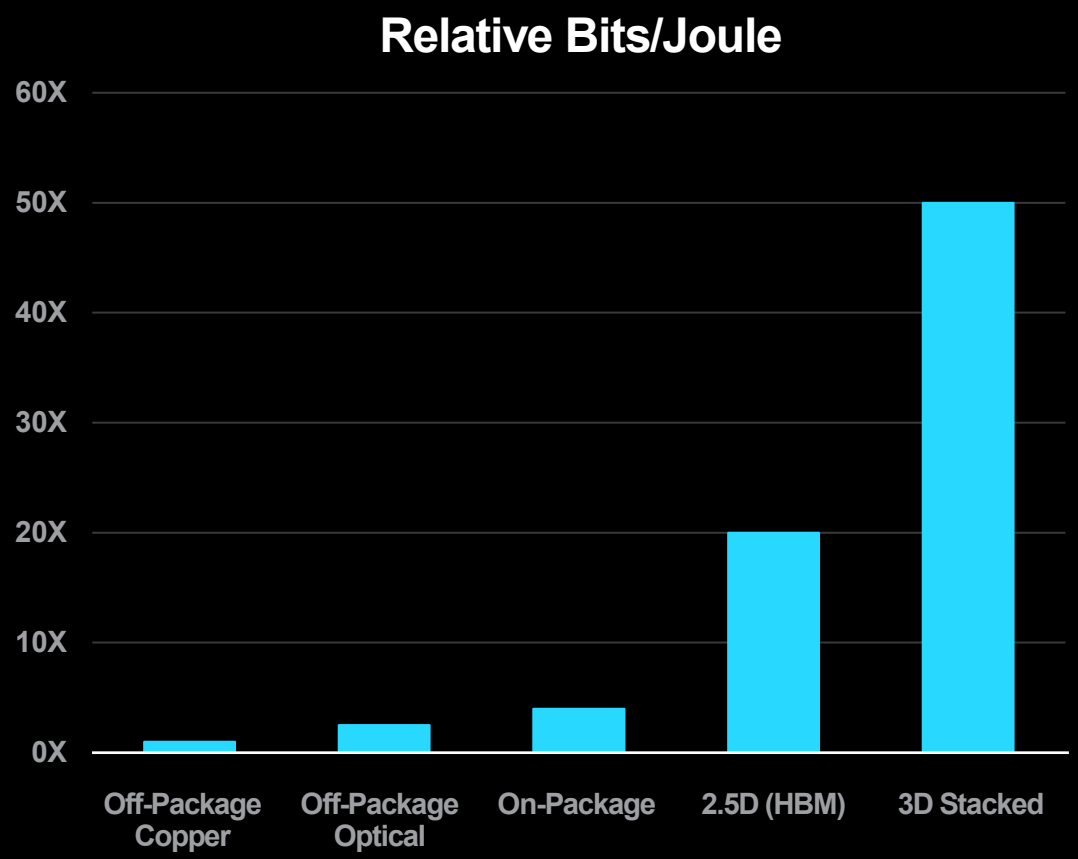
Gains in Energy-Efficient Performance

Energy-efficient performance needs tight integration.

2.5D enables co-packed compute with HBM.

AMD 3D V-Cache™ technology drives energy efficiency leadership.

Advanced 3D hybrid bonding provides by orders of magnitude the densest, most power-efficient chiplet interconnect through higher bandwidth and lower latency.



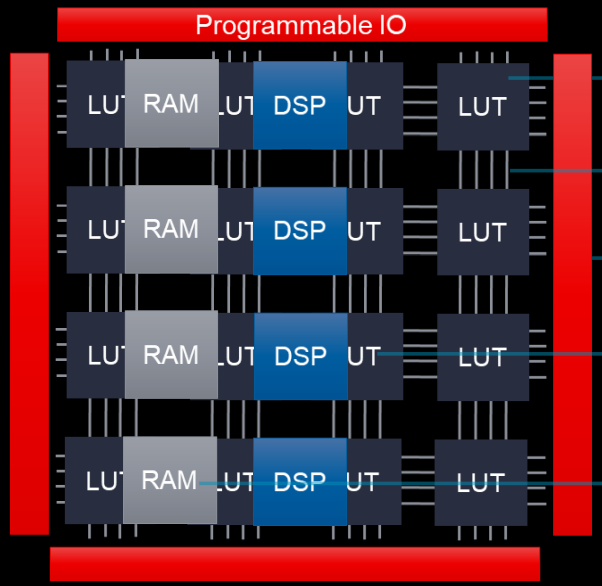
L. Su and S. Naffziger, "1.1 Innovation For the Next Decade of Compute Efficiency," 2023 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2023, pp. 8-12, doi: 10.1109/ISSCC42615.2023.10067810

Source: Internal AMD Data

The FPGA Dividend

Making Extreme Customization Affordable

- **Chameleon** amongst the semiconductors – ready silicon on attractive technology node:
  - shortens time to deployment, and
  - lowers non-recurring engineering cost.
- **Customize**:
  - IO Interfaces: DAC/ADC, Transceivers, GPIO
  - Functionality: signal conditioning, compression, encryption, NN acceleration, …



Sea of programmable Lookup Tables (LUTs) ~millions

Programmable Interconnect

Programmable IO

DSPs: n-bit MAC

Embedded SRAM ~high bandwidth

FPGAs offer post-production HW specialization.

The Quantization Dividend

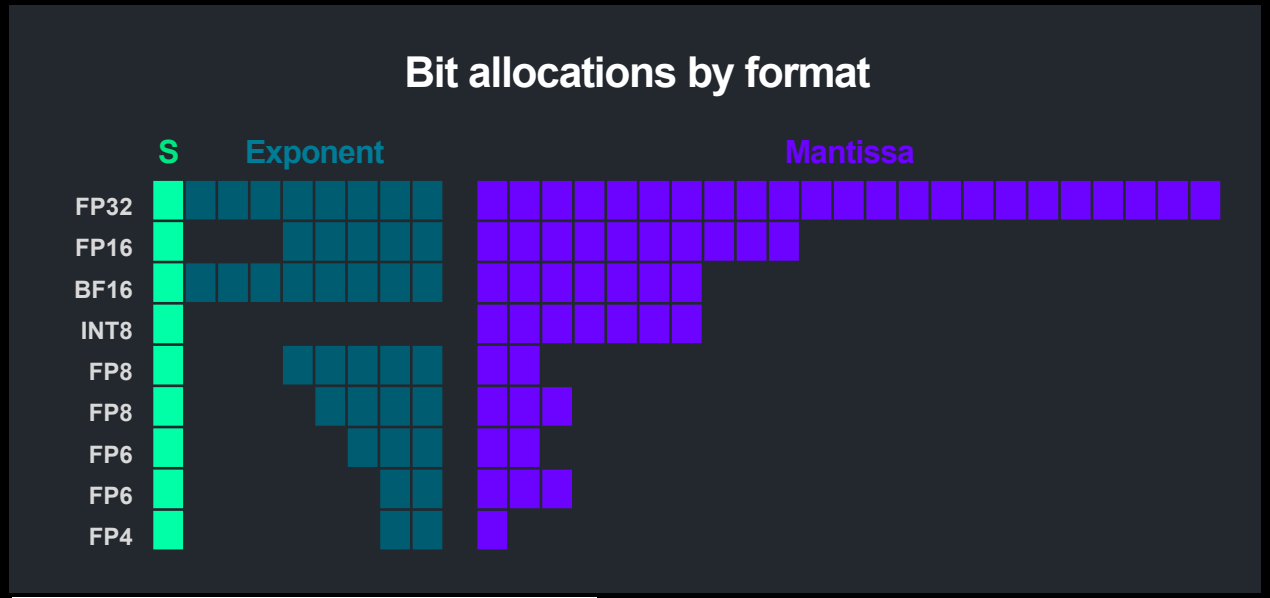
Customization that Makes a Difference

Adapt algorithms to lower-precision arithmetic for space and energy benefits:

- Smaller storage and buffers (linear),
- More efficient arithmetic (quadratic for MULs).
- Tradable for proportional upscaling headroom.

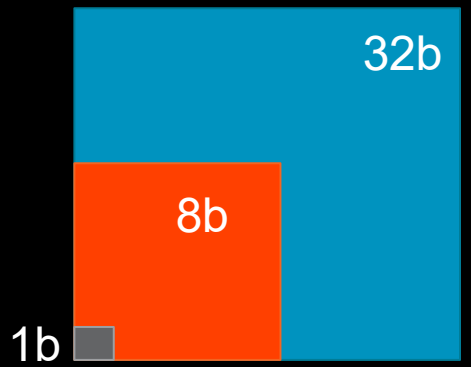
Techniques:

- PTQ exploiting algorithmic resilience.
- QAT for advancing quality of results.
- A2Q (accumulator-aware quantization) for optimizing underlying hardware.



	Operation	Picojoules per Operation 45 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		0.07	2.9
	Int 32		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.

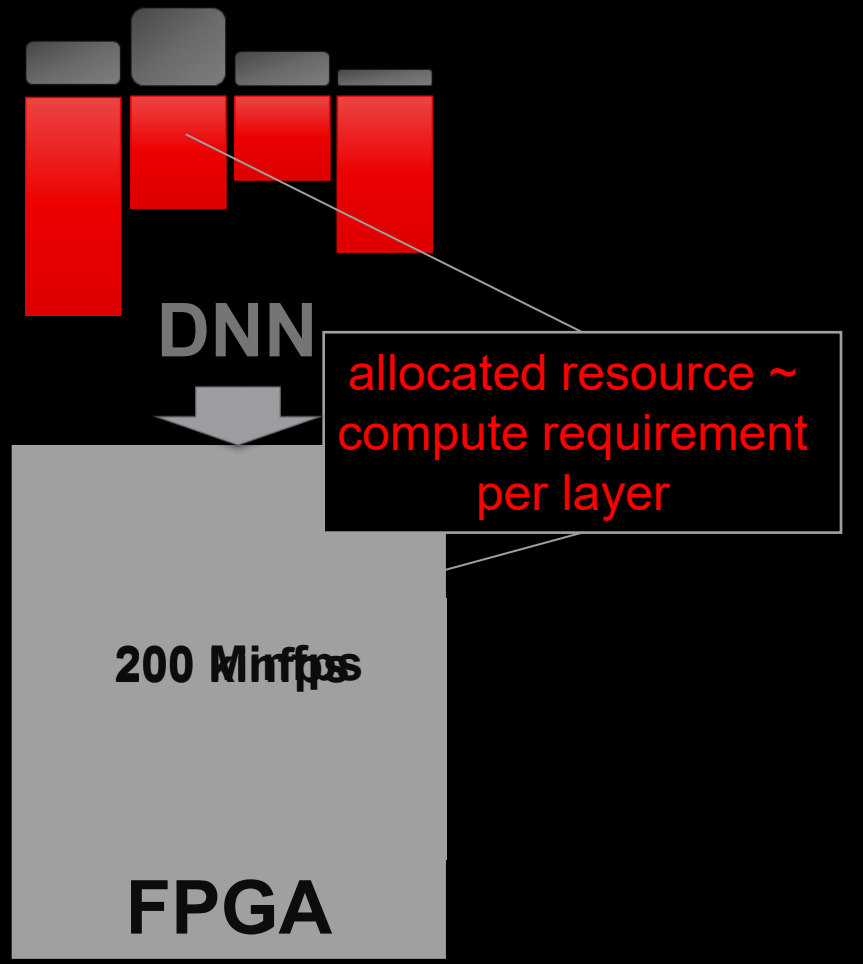


Source: Internal AMD Data

The Dataflow Dividend

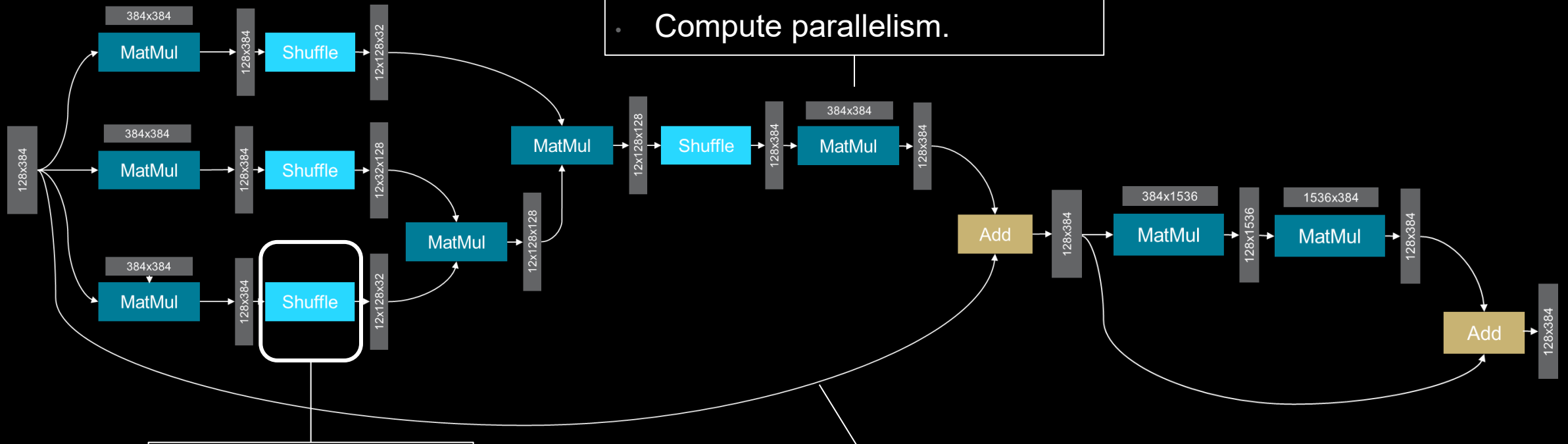
Keep Everyone Moving

- Structural dataflow representation of a topology in hardware:
  - Improved efficiency:
    - Fine-grained (per-layer) customization (precision, parallelism, …).
    - Partial, on-chip feature map buffering typically suffices.
  - Low fixed latency:
    - No need for batching to attain performance,
    - Predictable, guaranteed execution latency for all inputs.
- Scale performance & resources right according to needs:
  - Spectrum from **full parallel unroll** of bottlenecks for maximum performance
  - To compute **folded in time** for minimum resource footprint.



Further Benefits of Customization

Example: Attention Block



Finely tuned resource investment:

- Buffer sizes &
- Compute parallelism.

Fused operations:

- Less buffering,
- Lower latency &
- Higher efficiency.

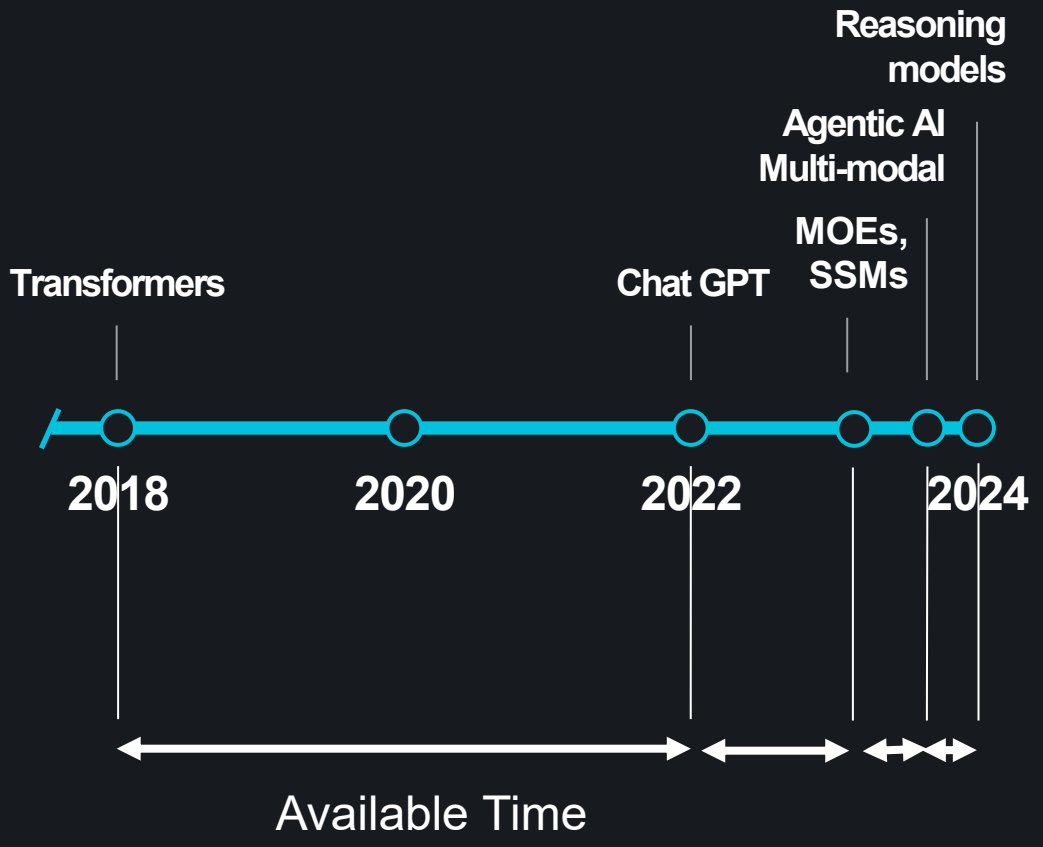
On-chip buffering & data movement:

- Low latency &
- Energy efficiency.

FINN & Brevitas for Enabling Agility

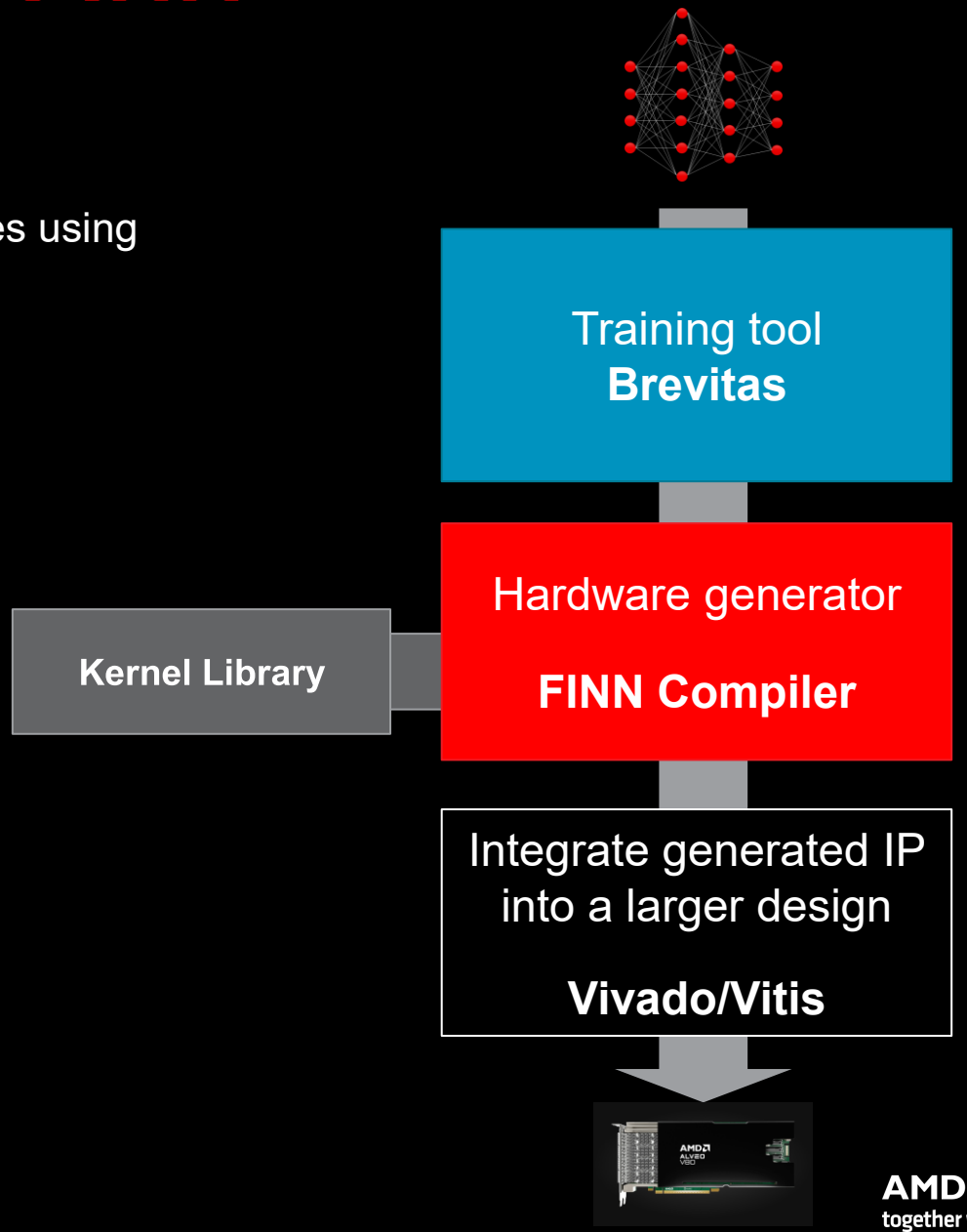
Agility is Key

- Time window for customization is getting shorter & shorter as the rate of innovation accelerates.



Tooling for Agility: Brevitas & FINN

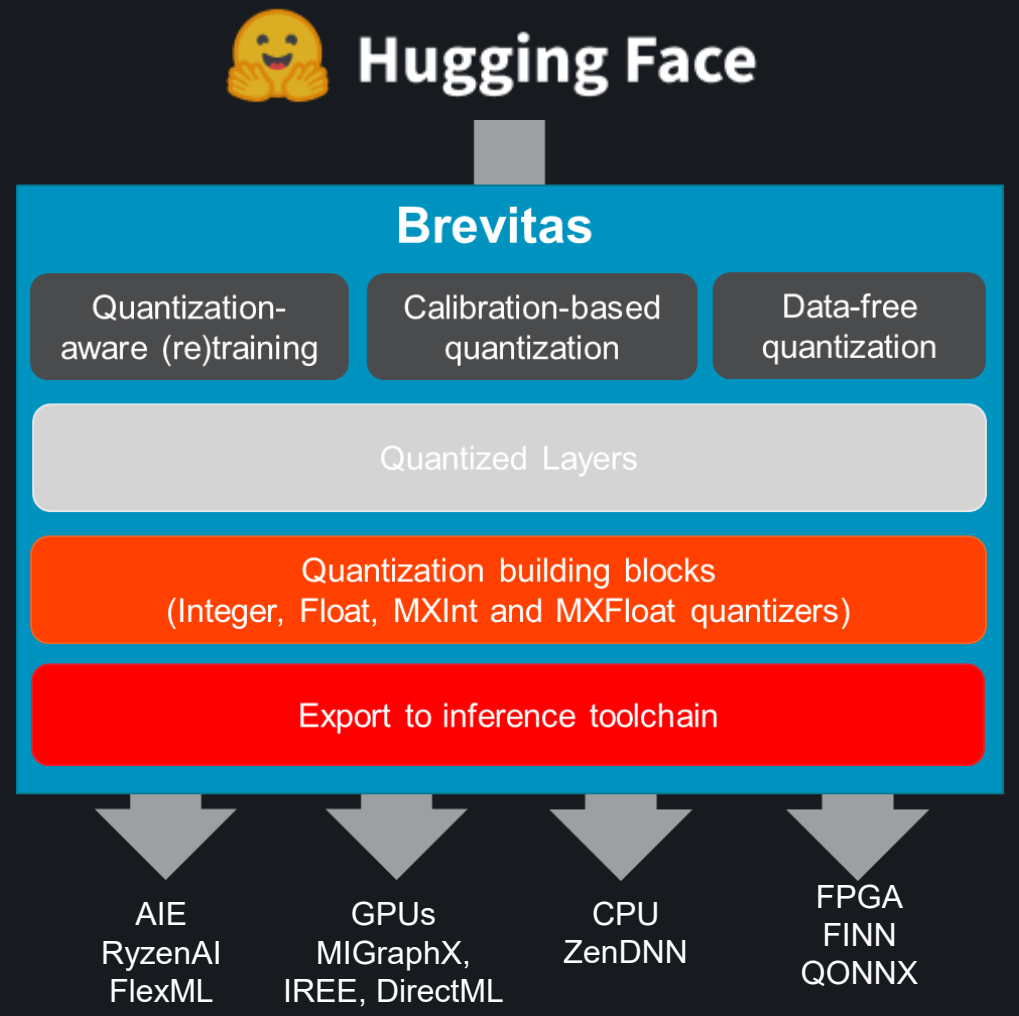
- End-to-end flow – from DNN to bitstream
  - Enables generation of highly customized hardware architectures using **quantization** and **dataflow**.
- Components
  - Brevitas: Training tool,
  - FINN: Hardware generator, and
  - Kernel library (HLS/RTL).
- Open-source
  - Easy collaboration with partners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



Brevitas - PyTorch Library "Agile Quantization Support"

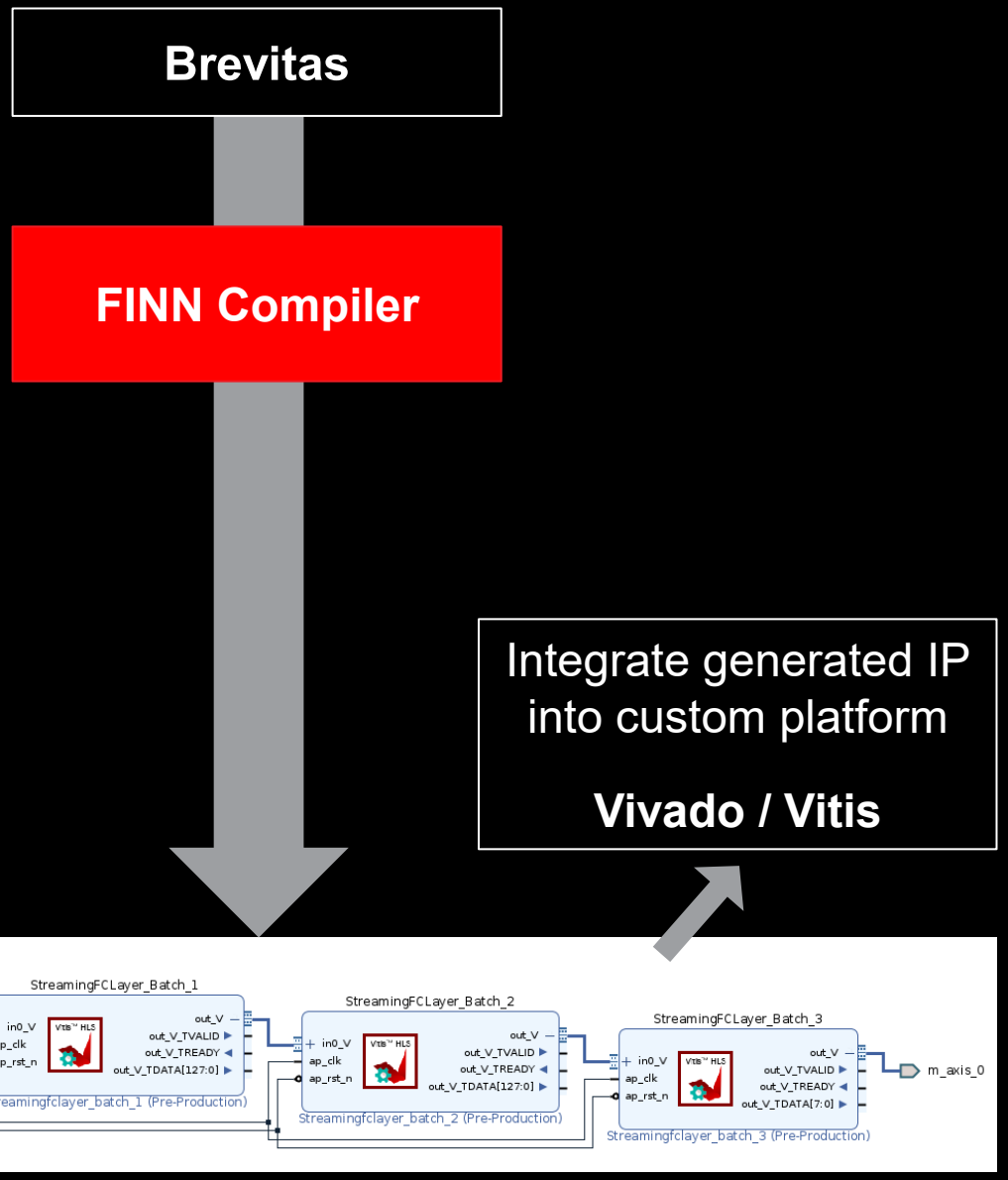
- **First class support for custom datatypes and operators at ML framework level**
  - Arbitrary precision integer, float, incl. FP8, block-style (MX)
  - Extendible to user-defined datatypes, operators and support at training time
- **Composable building blocks at multiple abstraction levels that can be arbitrarily combined**
  - Easy to broaden/scale model scope
  - Supports QAT & PTQ
    - GPxQ, SmoothQuant, Bias Correction, Weight Equalization,…
- **Enabling quantization research**
- **Hardware-independent through diverse export flows**

https://github.com/Xilinx/brevitas



FINN Compiler

- Modular **graph compiler** performing transformations to:
  - Incrementally lower ONNX graph to a hardware description.
  - Perform **optimization**: Layer fusion, streamlining, …
- Explores the **design space**:
  - Calculates the degrees of parallelism for each kernel using resource cost and performance models.
- Creates **DNN hardware IP** by stitching together parametrized kernels from the **kernel library**.



FINN‘s Agility Journey

Departing from extreme low precision

- FINN started with building *binarized* neural network inference solutions (XNOR-Net style).
- Support for 4-bit and 8-bit solutions required robust and innovative kernel generalizations.
- Using higher-precision layers exclusively for IO was bearable.
- Floating-point fallback layers are coming.

Growing robustness for more complex topologies

- Started with serial CNN topologies.
- Support for residual (ResNet) topologies.
- Going for transformers and the time dimension.

Reducing complexity for deploying custom transforms & operators

- New interfaces coming in FINN to be exploited by tooling being built with Microsoft Research.

Classic Thresholding in FINN

Assumption

very low precison → *very few* quantization levels → *very few* thresholds.

State of Affairs

FINN unrolls full breadth of thresholds $(T_i)_{i=0,\dots,n-1}$ with $n = 2^k - 1$ for a k-bit output:

$$y = \sum_{i=0}^{n-1} (T_i \leq x)$$

- Already optimal for binary outputs.
- Acceptable for up to about 4-bit outputs.
- Exponentially increasing pain beyond.

Benefit: Memory access pattern independent from input.

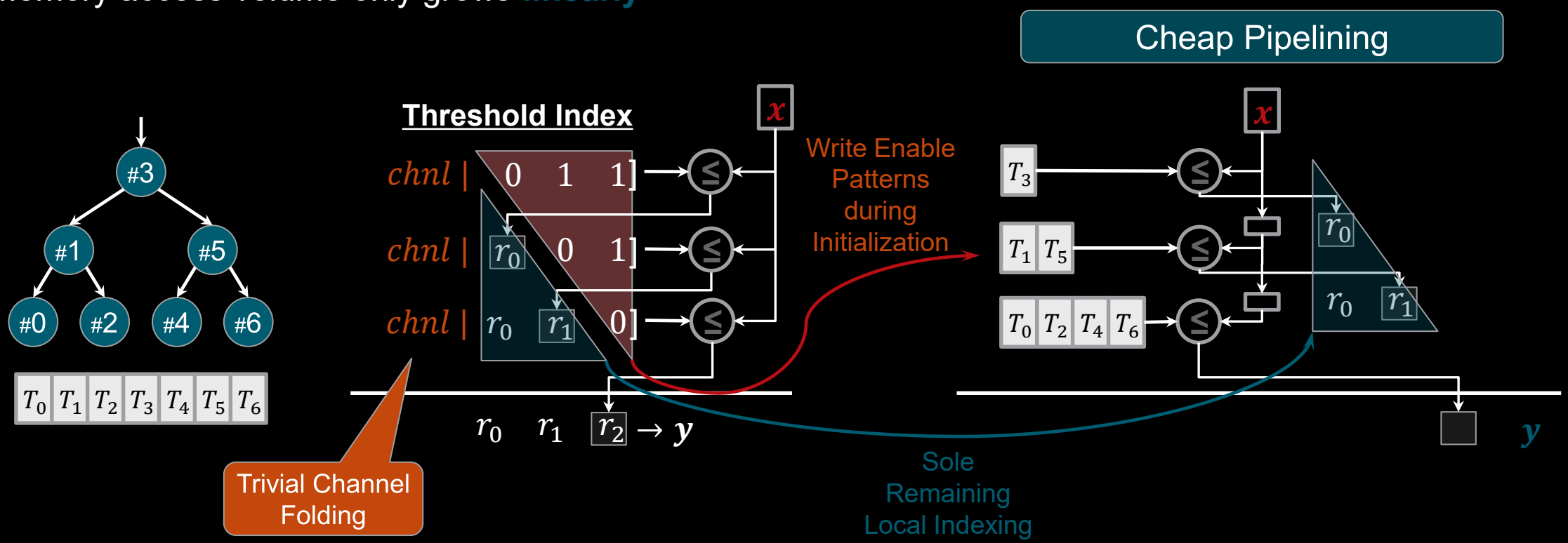
However, **exponentially** growing costs *per op*

- Threshold memory traffic, and
- Parallel comparator demand for throughput

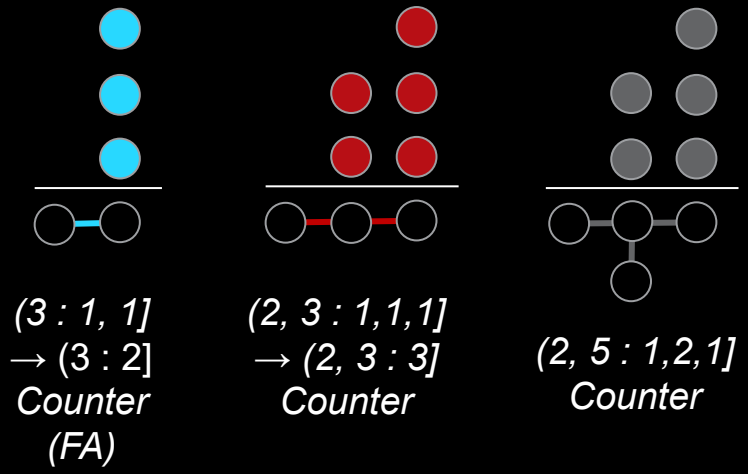
Exploiting Binary Search

The Deal

- Thresholds must be **sorted**.
- Memory access becomes **input-dependent**
- Memory access volume only grows **linearly**



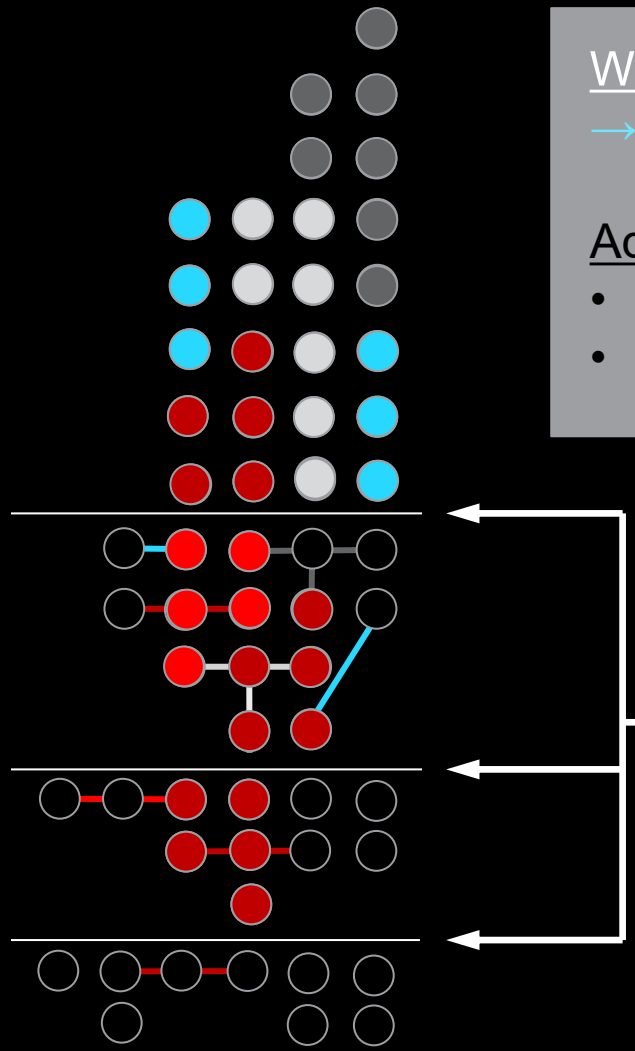
Compressor Tree Construction for Dot-Product Reduction



(3 : 1, 1] → (3 : 2] Counter (FA)

(2, 3 : 1,1,1] → (2, 3 : 3] Counter

(2, 5 : 1,2,1] Counter



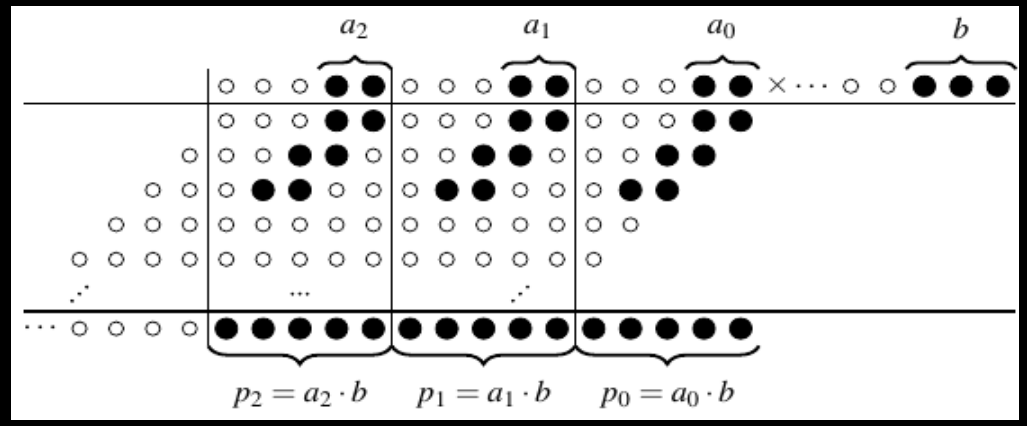
Natural Pipelining Opportunities

Within Stage:
→ LUT → CC* →

Admissable Chaining:
- CC of full slices before Versal.
- Also LUT cascade on Versal.

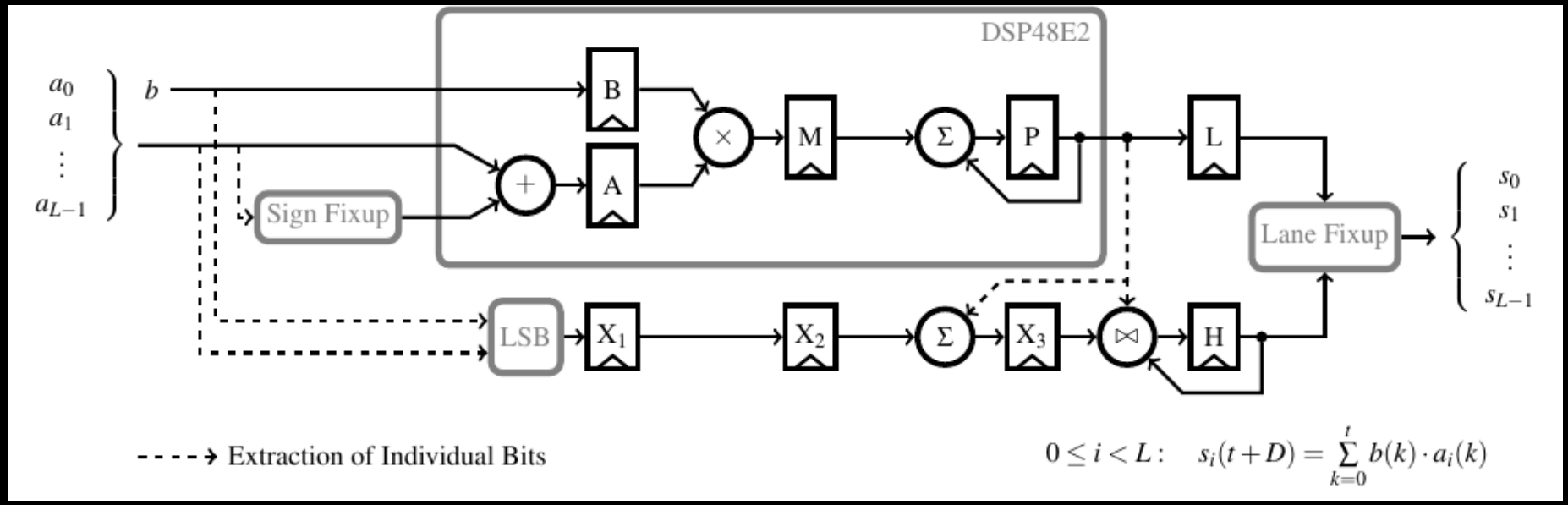
Konstantin J. Hoßfeld, Hans Jakob Damsgaard, Jari Nurmi, Michaela Blott, and Thomas B. Preußer: ***High-Efficiency Compressor Trees for Latest AMD FPGAs***, ACM TRETS, February 2024. https://doi.org/10.1145/3645097

Low-Precision DSP Packing

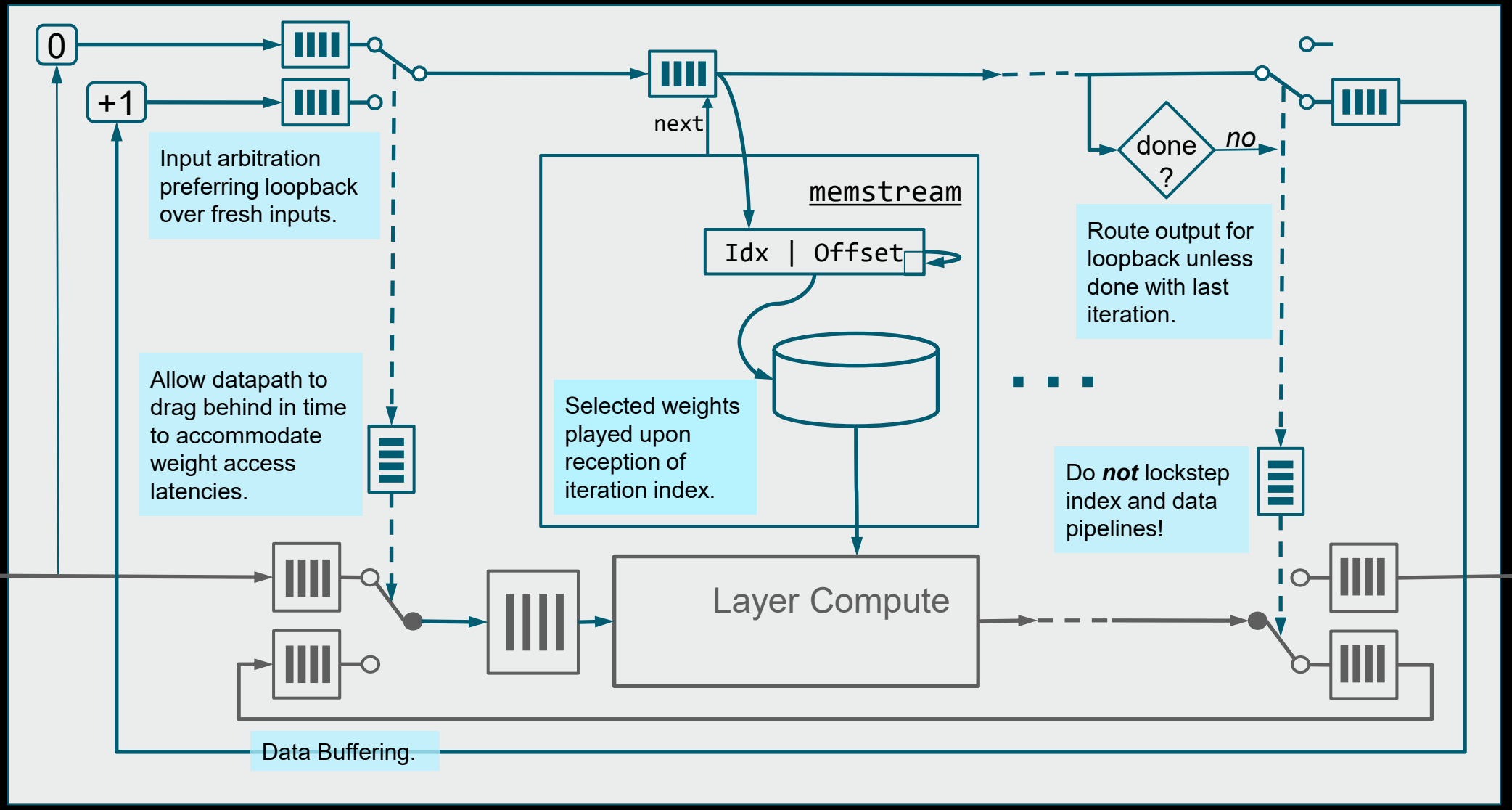


Performance per DSP48E2
- HLS: 1 MAC/cycle
- **This Technique:**
  - **8×8 bits: 2 MAC/cycle**
  - **4×4 bits: 4 MAC/cycle**

- Multiplier shared across vectorized multiplicands
- Vectorization radix to allow sufficient lane spacing
- Accumulation spill-over monitored and corrected in external fabric

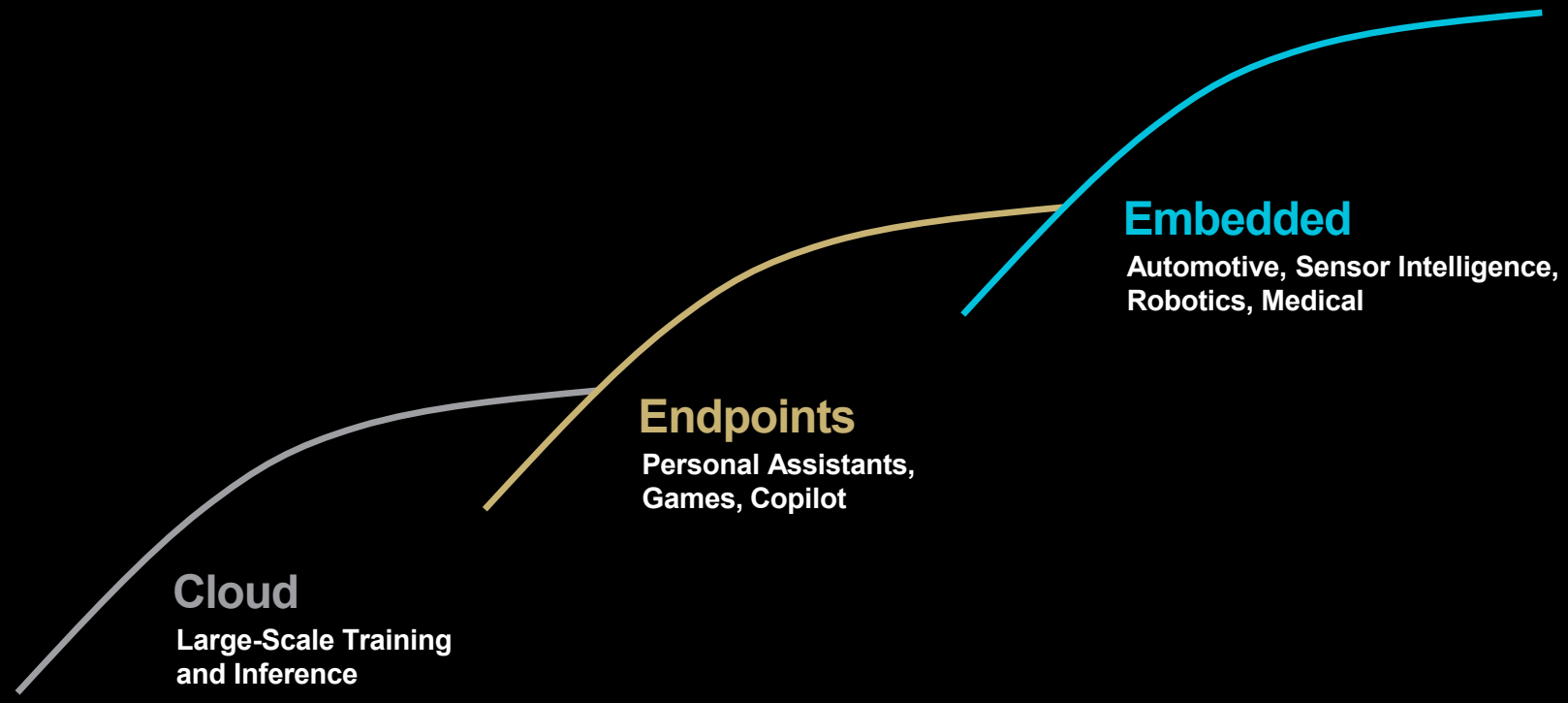


Loop Operator Tapping into the Time Dimension



Conclusions

The AI Revolution in Supercycles



Today's AI

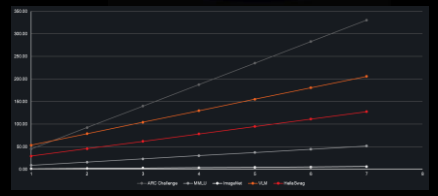
Virtual › Personalized › Physical

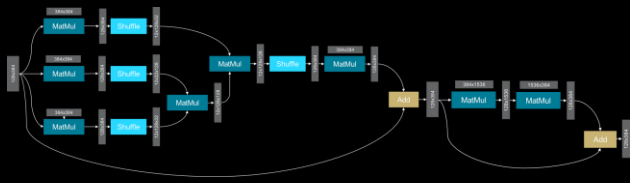
With Increasing Unit Volume

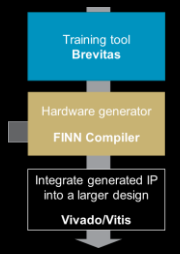
AI Pervasiveness

Summary & Conclusion

- AI is disrupting many industries today and many more to come.
- Inference efficiency is the #1 challenge int the AI revolution on all levels.
- Specialization of accelerators with dataflow, quantization and sparsity promises orders of magnitude of improvements.
- Agile end-to-end flows essential to keep up with rate of innovation.
- Pressing need for agility in tools like FINN, Brevitas & ***Brainsmith***.







Thank You.

Disclaimer

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION


AMD
together we advance_