SYNOPSYS®

Navigating the Future: Processor IP Trends Driving Advanced SoCs

Dr. Yankin Tanurhan
Sr. Vice President of Engineering, Synopsys

37 Years of Advancing Chip Design

Leading electronic design automation tools and services

Broadest portfolio of interface, foundation, processor and security IP

Pioneer in electronics systems solutions and AI-powered EDA

#12 global software company by revenue

$5.79B* Revenue (TTM)

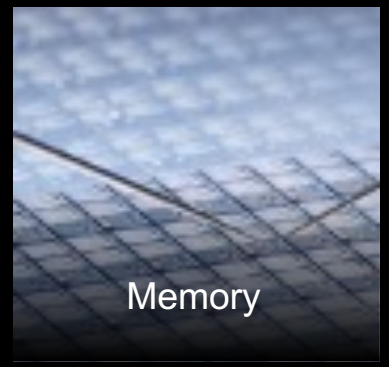
~19K* Employees

3,395 Patents

28% R&D Investment

* Excluding SIG

Change Is Occurring Across Many Industries

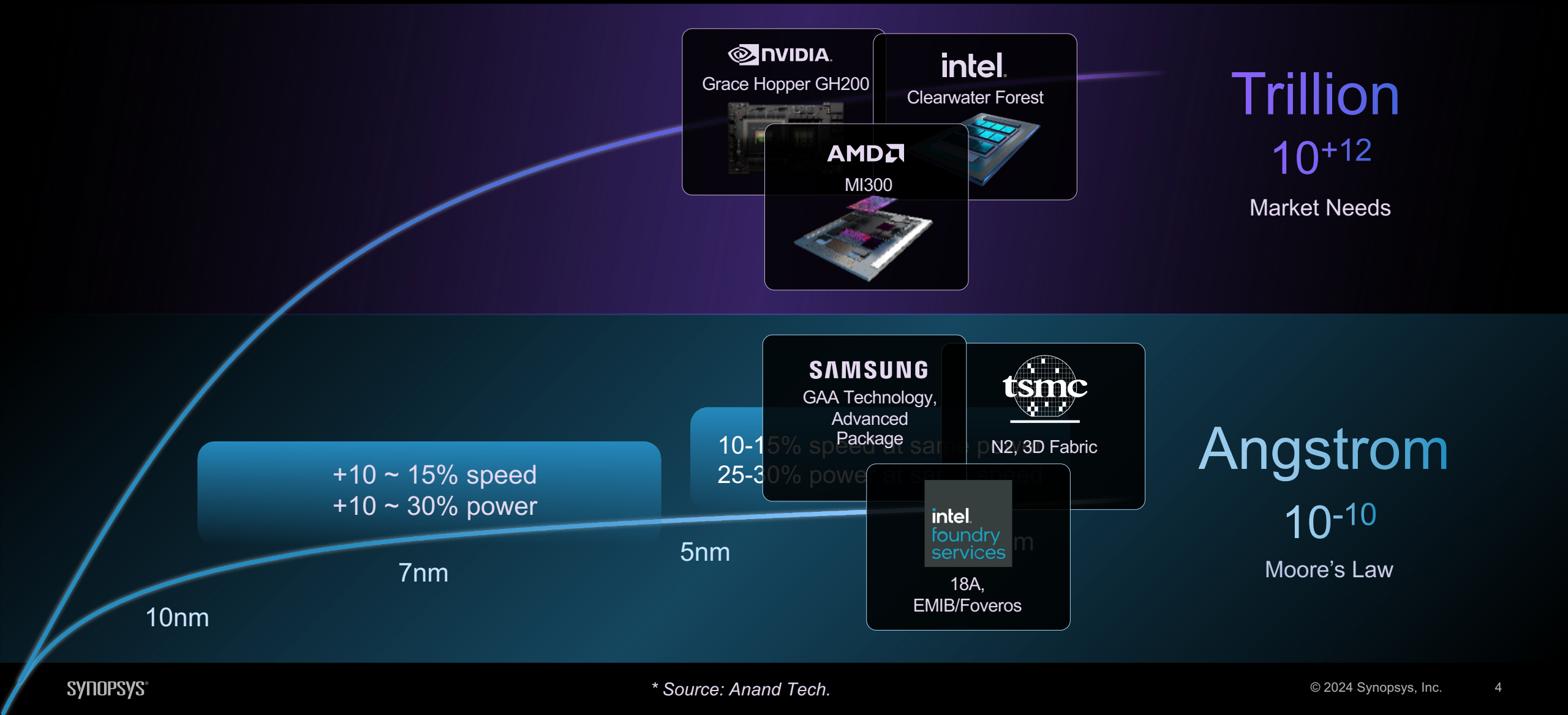


Demand for Smart Everything seems limitless

Innovation is fueled by semiconductor and software advances

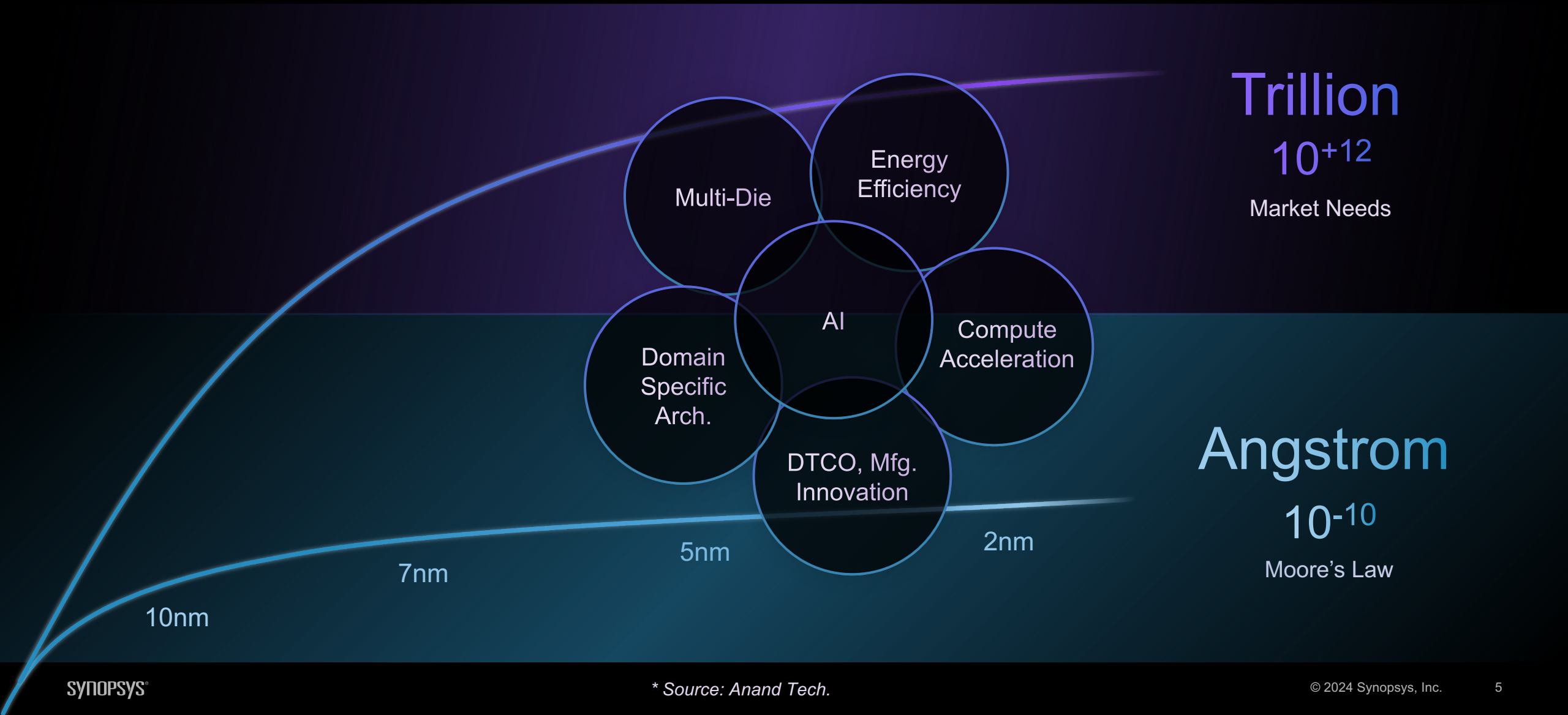
Driven by fusion of big data, massive compute, and machine learning

Process Scaling Gains Lagging Compute Growth Needs



** Source: Anand Tech.*

Multiple Innovations Closing the Gap



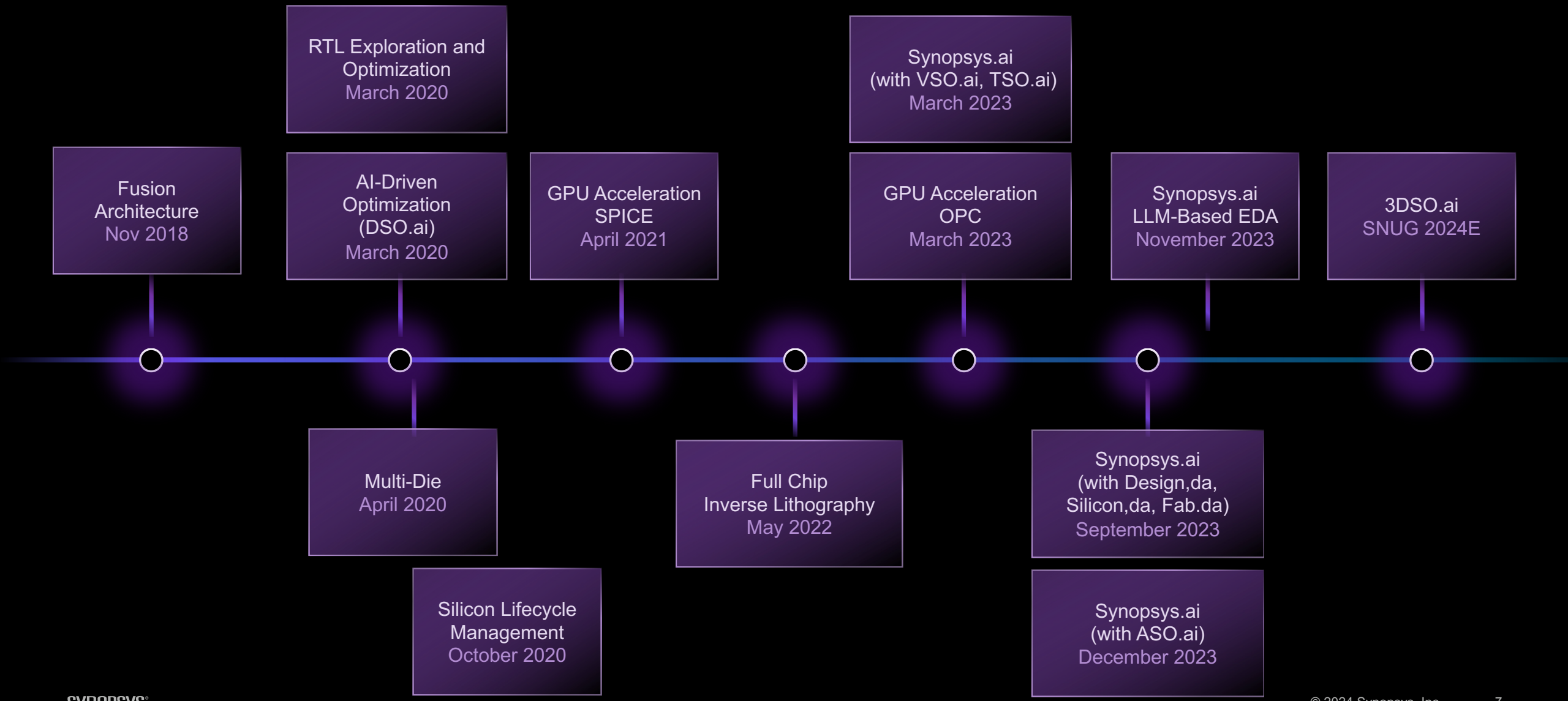
** Source: Anand Tech.*

Semiconductor Chips and Software are
at the center of these monumental transformations

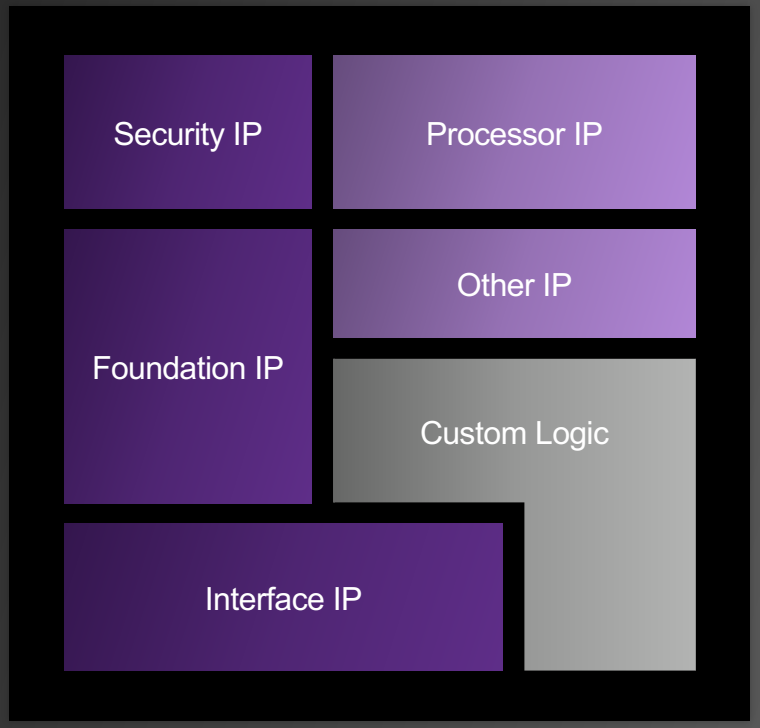
Software Drives Differentiation

SoC Designers & Software Developers Embracing Open
Standard Solutions for **Collaboration and Flexibility**

Pioneering Innovations Enable Semiconductor Progress



Synopsys has the Broadest IP Portfolio for Modern SoCs



25 years of investment & commitment

#2 IP provider worldwide

Leader in Interface IP

Leader in Foundation IP

#2 in Processor CPU IP

Reduce design risk and speed time-to-market with high-quality IP

Growth in Designers Adopting RISC-V

3600+ RISC-V Int'l Members
Across 70 countries; 30% increase in 2022

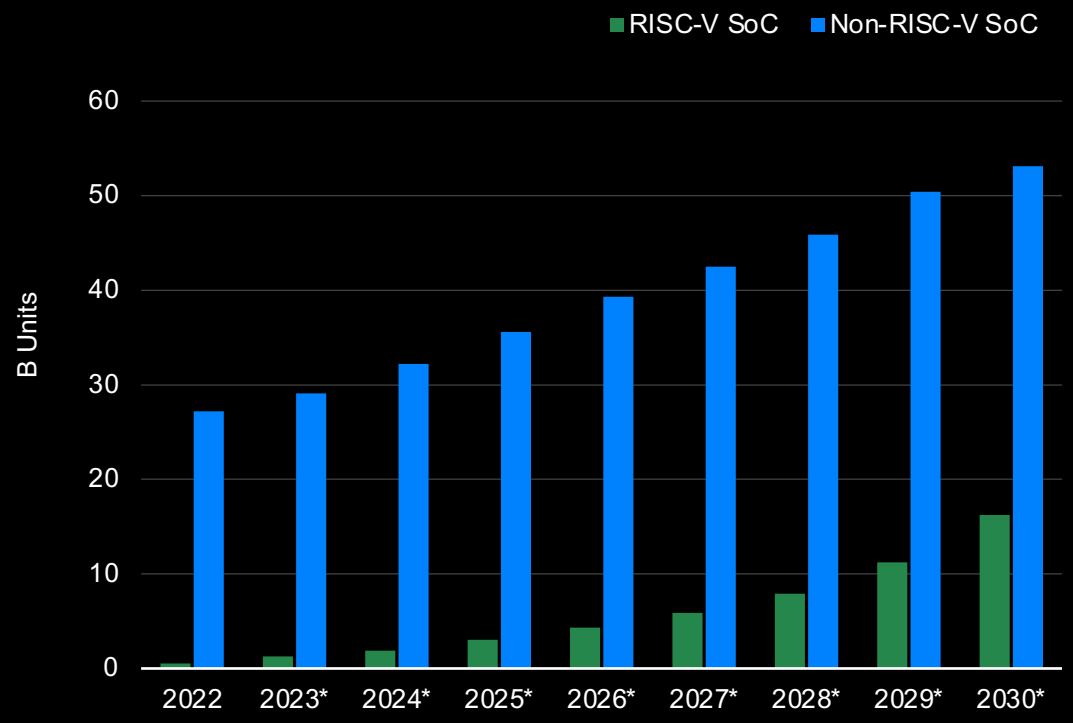
108 Chip Companies
SoC, IP, FPGA

54 Software Companies
Dev tools, firmware, OS

13 Industries Represented
Cloud, mobile, HPC, ML, automotive

Source: RISC-V International, May 2023

>40% CAGR Forecasted for SoCs with RISC-V Processors



Source: The SHD Group

* Forecast

Recent RISC-V Collaborations & Milestones

March 23

RISC-V IP/chip supplier StarFive receives strategic investment from search giant Baidu

May 18

Meta reveals details of in-house RISC-V AI chip

June 5 – 9

1st RISC-V Summit Europe

Aug 3

Tenstorrent raises $100M in funding

Led by Hyundai Motor Group and Samsung Catalyst Fund

Oct 17

Qualcomm to bring RISC-V wearable platform to Wear OS by Google

May 31

Linux Foundation Europe launches RISE

Board includes Andes, Google, Intel, Imagination, MediaTek, NVIDIA, Qualcomm, Red Hat, Rivos, Samsung, SiFive, T-Head, and Ventana

Aug 4

German JV established by Bosch, Infineon, Nordic, NXP and Qualcomm

Initial focus automotive, eventually IoT

Nov 7

Synopsys announces ARC-V portfolio of processors

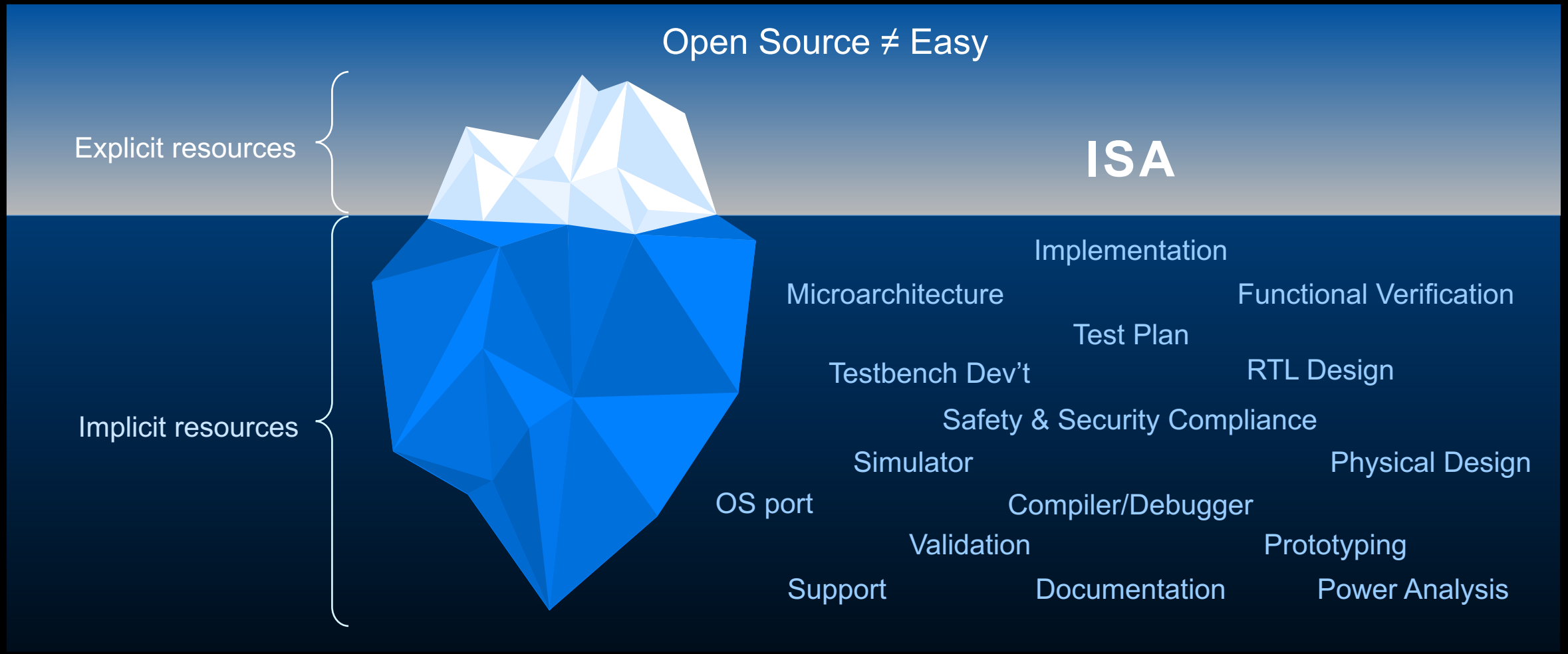
RISC-V Chip Design Industry Challenges and Opportunities

Ecosystem Maturity & Toolchain Support	Designers contend with less optimized tools, slowing development, debugging, & face limited resources due ecosystem's growth.
Performance Optimization	Complexity in achieving optimal performance & need for thorough understanding of the architecture.
Software Compatibility & Porting	Difficulty in ensuring software compatibility & significant effort required for porting existing software.
Security Considerations	Ensuring robust security features & vigilance against potential vulnerabilities.

Continued investment, industry collaboration, and advancements in SW/HW are crucial for RISC-V's success

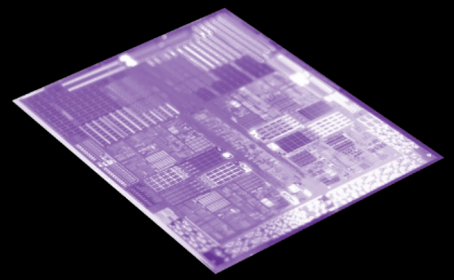
It’s All About the Implementation

Quality & Expertise Required for Implementation Success



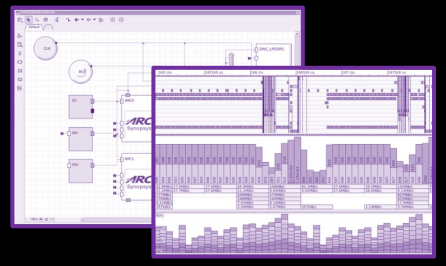
Synopsys Solutions for RISC-V

Optimize QoR and Productivity



- Fusion QuickStart Implementation Kit (QIK), reference flows & guides for superior PPA
- Tunable to target application & process to 2nm

Early Architecture Exploration & SW Development

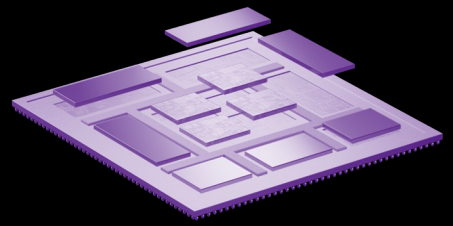


- Early performance/ power analysis & optimization of multicore SoCs
- Optimize HW/SW partitioning
- Flexible ARC-V virtual prototypes for SW dev

Accelerate SW Development w/HW

- Start SW development months before HW availability
- Avoid costly re-spins
- Actionable power profiling on full design & software workload

Broad Portfolio of IP

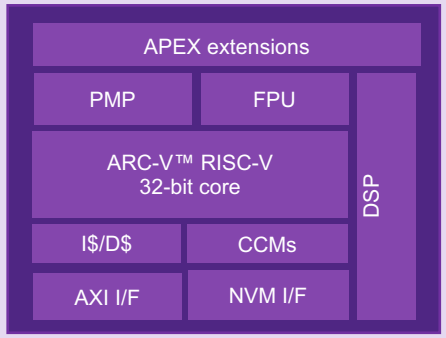


- ARC-V Processor IP with robust dev toolkit
- Silicon-proven Interface IP minimizes risk
- High-speed, low-leakage & low-power Foundation IP optimizes PPA
- Efficient Security IP helps protect data

Addressing Market Demand & Challenges with ARC-V IP

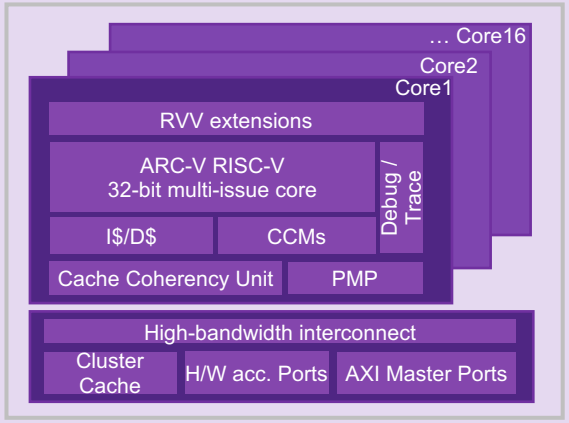
Builds on 25 Years of Customer Success & Leading PPA Efficiency

RMX Series



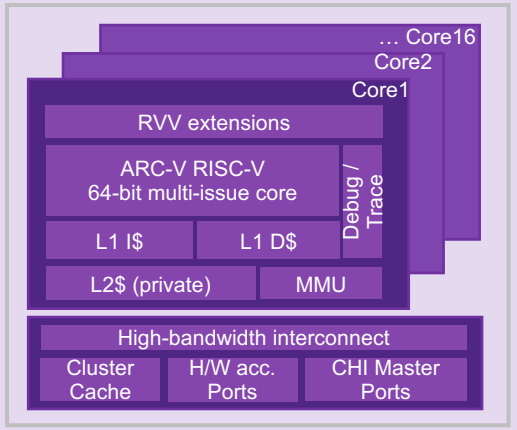
- 32-bit embedded processor, optimized for ultra-low power 3- and 5-stage pipeline configurations
- Optional DSP extensions
- FuSa hybrid mode

RTX Series



- 32-bit processor, optimized for real-time apps w/dual-issue 10-stage pipeline
- H/W virtualization support
- Up to 16x cores, plus H/W accelerators

RPX Series



- 64-bit multi-core host processor supporting user/supervisor profiles
- Multi-cluster cache coherency with AMBA-CHI interfaces

Functional Safety Features Accelerate SoC Safety Certification

For Processor IP, It's All About the Ecosystem

Broad Ecosystem Support for ARC-V Software Development

Synopsys Tools & Software

MetaWare Development Toolkit

- Visual Studio Code-Based IDE
- Simulink® Model-Based Design Support
- Highly-Tuned Runtime & Compute Libraries
- Optimizing Compilers (scalar & vector C/C++, OpenCL C)
- Advanced Debugger/Profiler
- Functional Simulator
- Cycle-Based Simulator

ISO 26262 & ISO 21434 Compliant

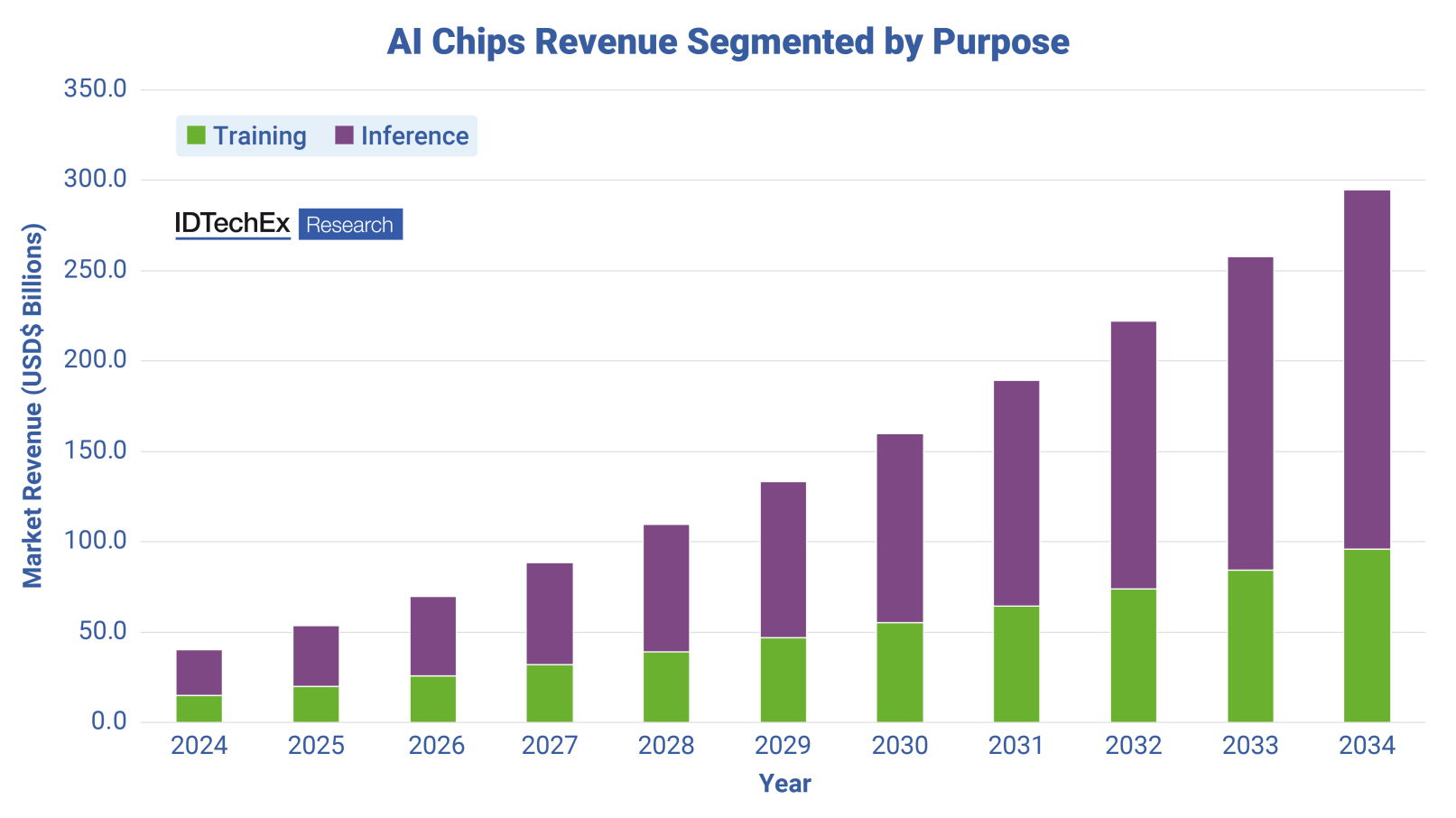
Open-Source Community

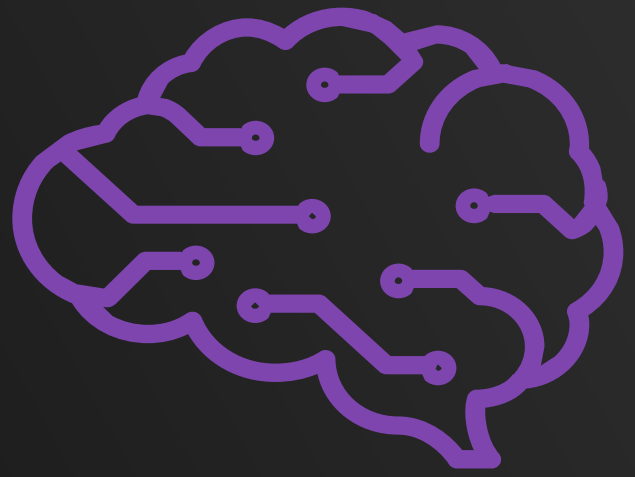
- GCC
- LLVM
- GDB /Binutils
- OpenOCD gdbserver
- QEMU
- RISC-V SBI library
- Linux
- Zephyr
- FreeRTOS
- uBoot
- …

Commercial Partners

AI: The Technology Disrupter

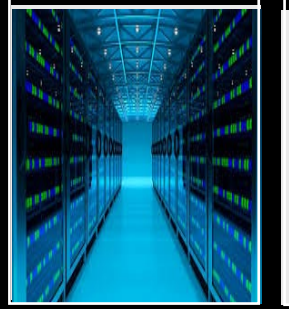
AI Processing Extending from Data Center to the Edge





AI Is Pervasive

Rapid Adoption in Broad Spectrum of Applications Requires PPA-Optimized Solutions

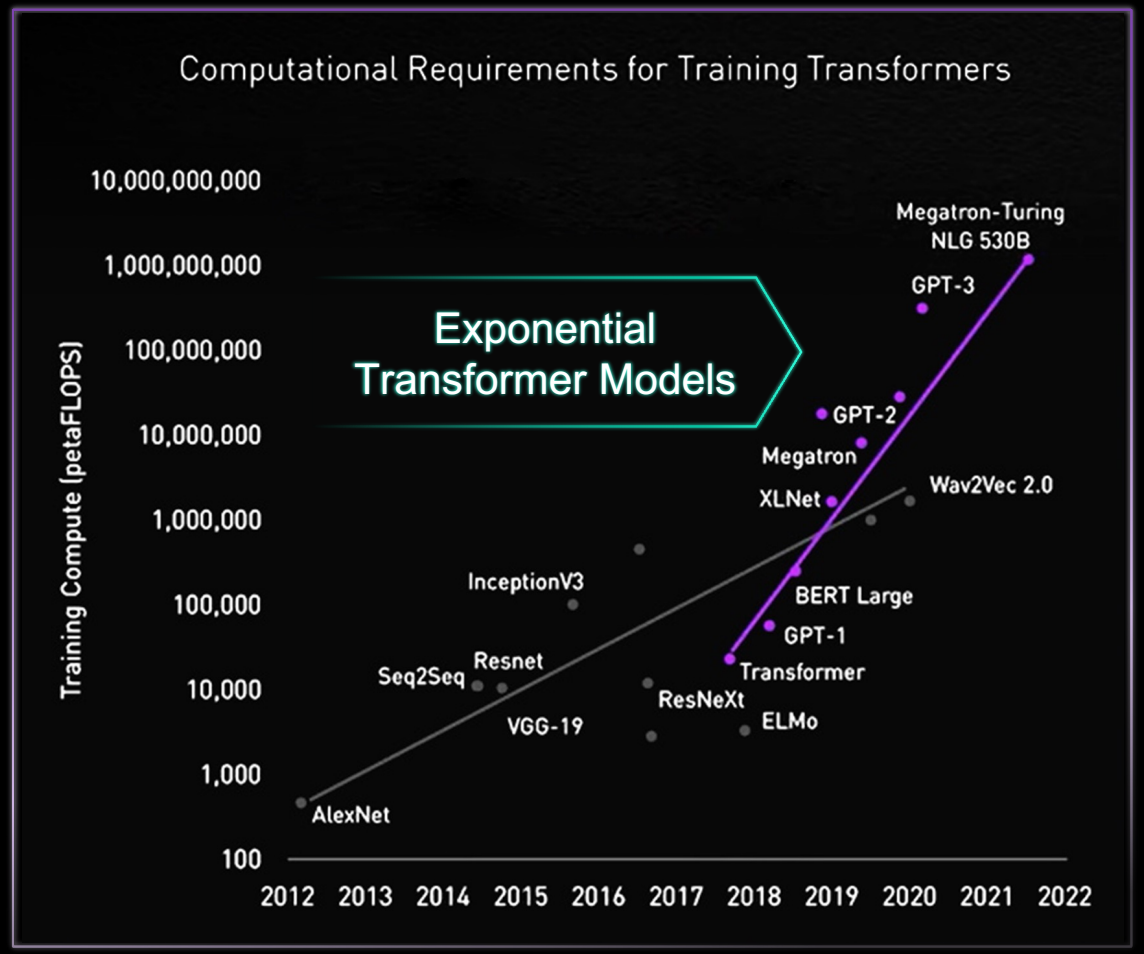
Cloud AI Accelerator	Edge Computing AI Accelerator		On-Device AI					Edge Device Accelerator
Data Centers	Edge Servers	Accelerator Cards	Automotive	Voice Assistants	Mobile & AR/VR	Cameras	Digital TV	IoT /AIoT
								

Performance in TOPS / W →

← Performance in TOPS

Model Compression (pruning, distillation, quantization, etc)

Newer ML Models Further Pushing Limits of Compute



All Models Excluding Transformers:

8X / 2 years

Transformer AI Models:

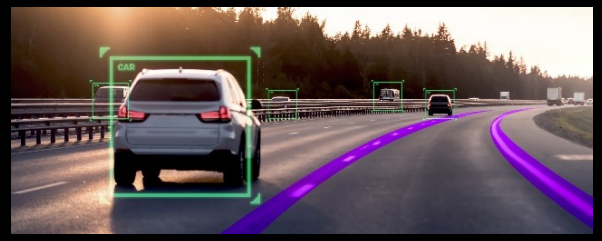
275X / 2 years

Context-Aware Transformer Models Come at a Price

Source: https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/

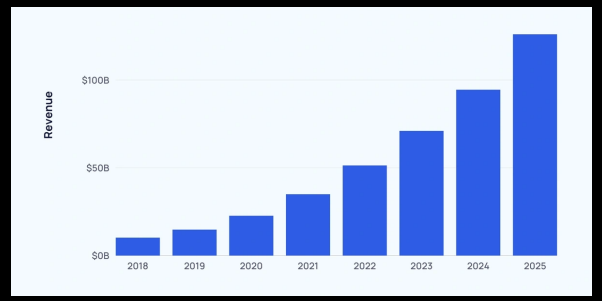
Artificial Intelligence SoC Requirements Evolving

Trends Driving Increased Complexity and Performance Requirements for AI



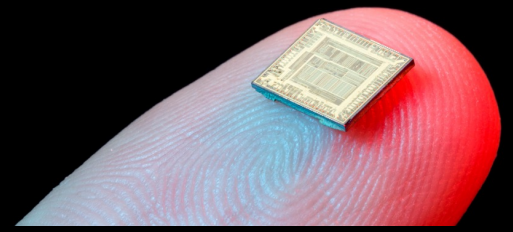
Automotive Safety

There is an arms race toward automotive self-driving autonomy (L2+, L3, L4…) FuSa, Virtualization and Floating-Point Data Types all asked for ADAS.



Proliferation of AI from low- to mid-end use cases to 1000s of TOPS

AI Market adoption and AI performance requirements continue to grow. More entry level AI being added (1-2 TOPS). ADAS L3 and beyond will require more AI computing.



Demand for Power and Area Efficient AI

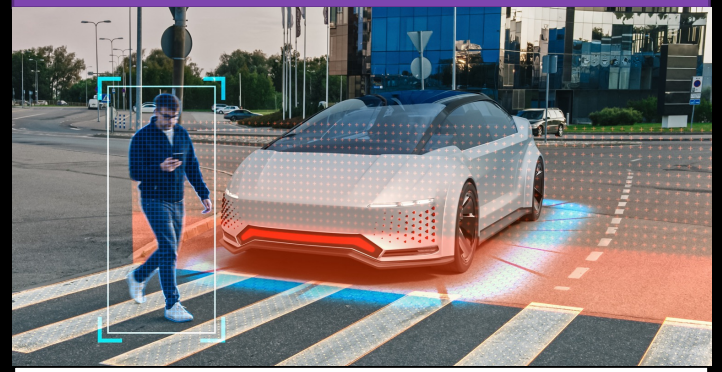
Growing need for strong AI performance in smallest area and power budget – computer, consumer, automotive markets

Evolving AI Research

Emerging Neural Networks (i.e. Transformers, GenAI for NLP, Vision, Speech) require more advanced hardware and software techniques.

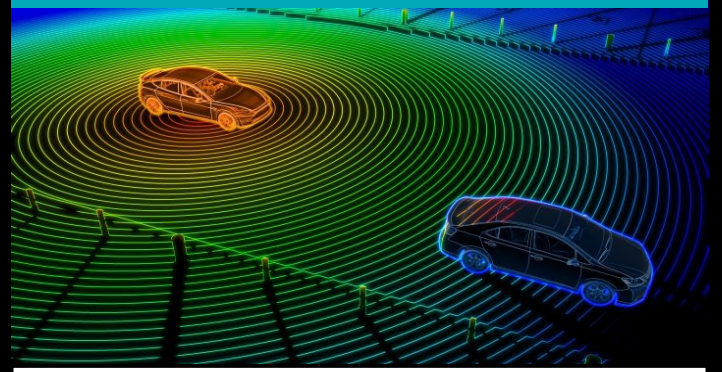
AI in Automotive – FuSa Applications

Embedded Vision



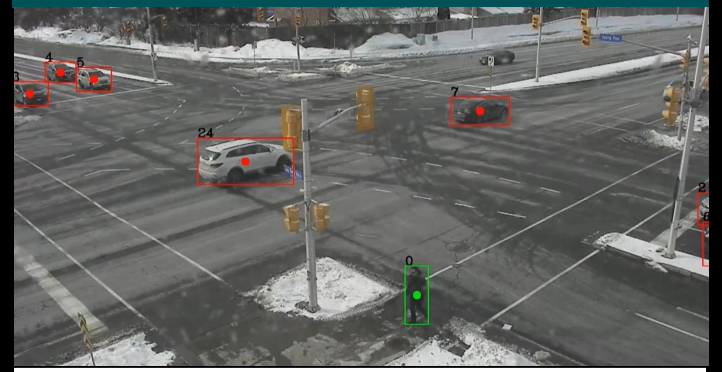
- Analyzes camera data for safety-enhancement and autonomous driving
- Expert-level accuracy in classifying objects in fractions of a second
- Example Use Cases:
  - Lane departure detection
  - Parking assist / self-parking

Radar/LiDAR



- Key components of Level 3+ and autonomous vehicles
- Essential in night driving, conditions of rain and fog
- Example Use Cases:
  - Blind spot detection
  - Collision avoidance systems

Sensor Fusion



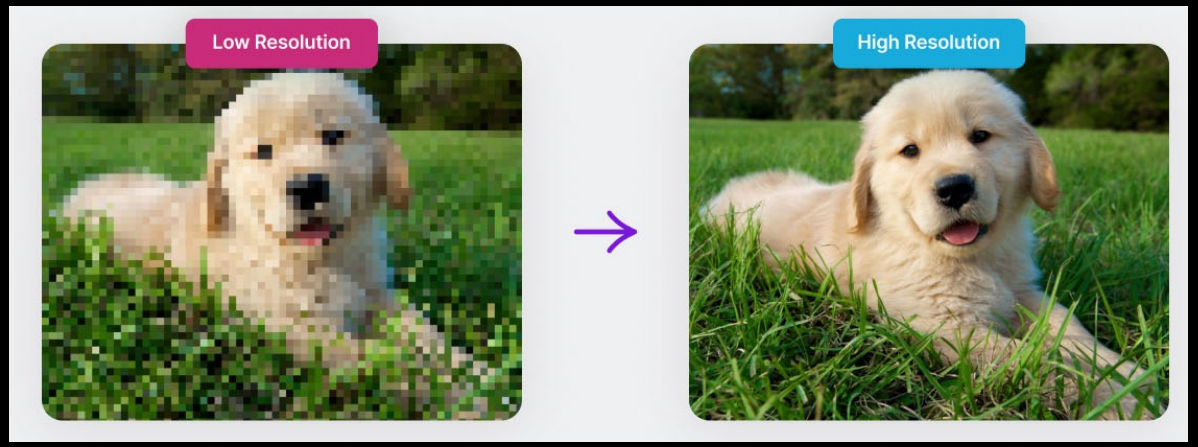
- Combines sensors to make more accurate representations of the environment
- Using multiple sensors together offset weaknesses of individual sensors
- Example Use Cases:
  - Autonomous vehicle nav
  - Smart Intersection

AI for Image Quality Improvement

– NPX6-1K Meets Performance and Power/Area for Image Quality Improvement Needs*

Upscaling Video 4x with NN to save power

- Super resolution reinserts data back into the image to reduce blur or upscale image
- Super resolution requires NN performance
- Upscaling video at destination saves power by reducing data transmitted



Improving ISPs Using NN Denoiser

- Noise is THE limiting factor in capturing low light video
- Applying AI to ISP provides excellent denoising and demosaicing performance

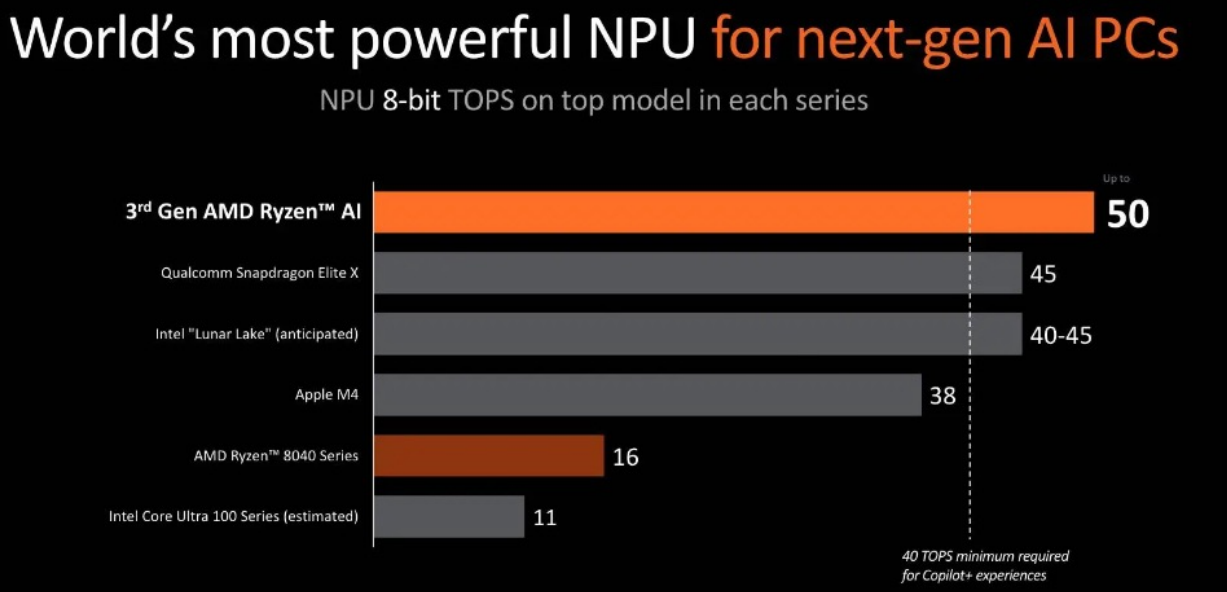
*depending on image frame size and fps requirements

AI Technology Evolution: Edge Devices

	Last 5 years	Ongoing Designs	Next 3 years
High Level M/L Performance	100s of TOPS	Up to 1000 TOPS	2000+ TOPS
Algorithms	CNNs, RNNs	Transformers, GenAI (LLMs/LVMs)	Transformers, GenAI (SLMs, LMMs)
NPU Data Types	INT8	INT8 / INT4 FP16 / BF16	INT4 / INT8 FP4, FP8, OCP MX
Quantization	PTQ (Post Training Quantization)	PTQ w/mixed precision quant.	PTQ, QAT (Quant-Aware Training)
Process Nodes	16 nm / 12 nm	7 nm / 5 nm / 3nm	3nm / 2nm
DRAM Interface	LPDDR5/4/4X	LPDDR5X/5/4X	LPDDR6/5X/5
Multi-Die/Chiplet	N / A	UCIe v1.1	UCIe v1.2
Safety	ISO 26262	ISO 26262 ISO 21434	ISO 26262 ISO 21434

AI LaptopTrends

- Continued increase in AI performance requirements
  - Microsoft Copilot+ requires over 40 TOPs (INT8) *on NPU*
  - NPU performance becoming key PC product differentiator
- Key applications tracking state-of-the-art AI models
  - Large Language Models: GPT-3 to interactive GPT-4o
  - Image to Video Generation: From Stable Diffusion to SORA video generation
  - Transformers remain central construct in GenAI NN models



Source: AMD, May 2024

Generative AI Built With Transformers

- Large-Language Models
  - ChatGPT has exposed a wide audience to AI's capabilities
  - Llama 2 offers scalable solution ( )
- Text-to-Image Models
  - Stable Diffusion, DALL-E, and Midjourney have demonstrated the potential of generative AI to generate images from text descriptions
- Generative AI Inference is moving to the edge
  - 9/11: MLCommons adds GPT-J 6B LLM to MLPerf Edge benchmarks
- Transformers are the building blocks of Generative AI

Stable Diffusion Prompt:
"Cute Dog Typing at a Typewriter German Style"

https://mspoweruser.com/best-stable-diffusion-prompts/

Photography

Stable Diffusion Prompt:
“Macro photography of dewdrops on a spiderweb, with morning sunlight creating rainbows.”

Painting

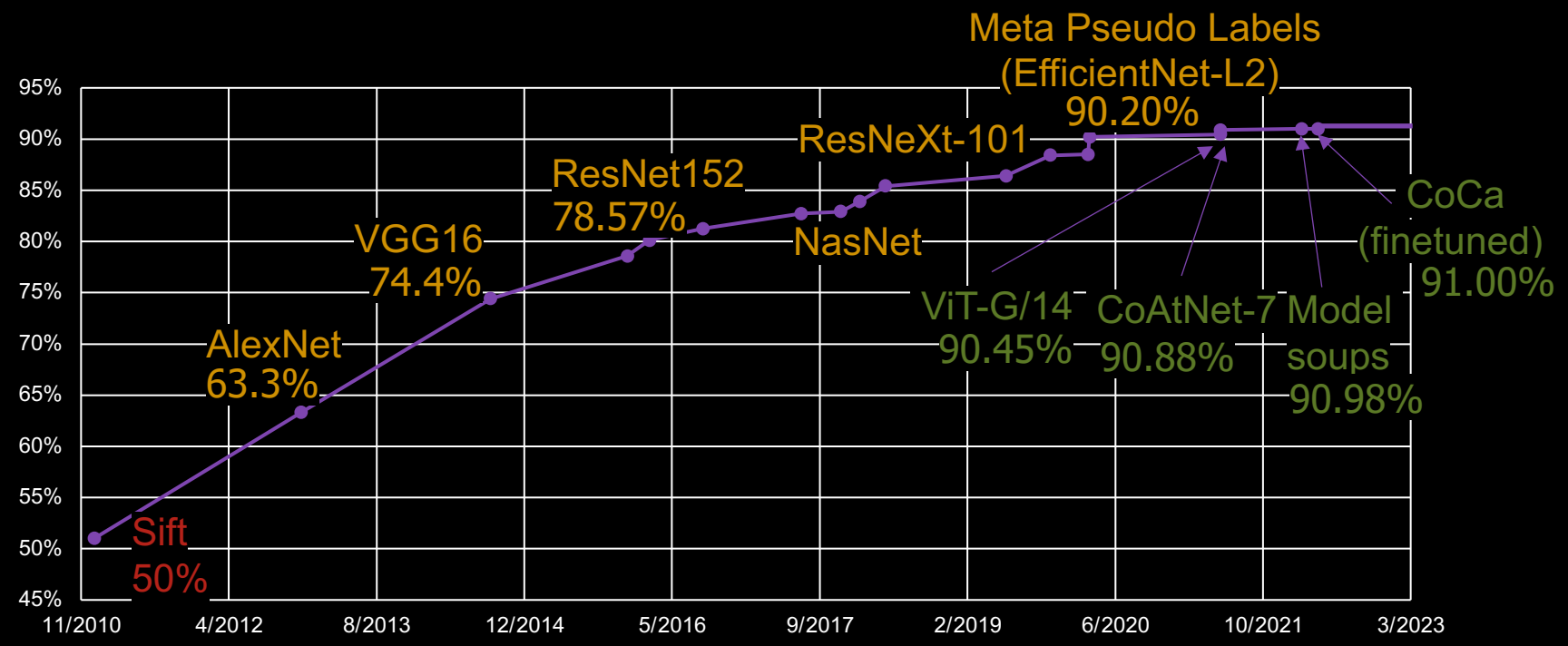
Stable Diffusion Prompt:
“Renaissance-style portrait of an astronaut in space, detailed starry background, reflective helmet. ”

Painting

Stable Diffusion Prompt:
“Pop Art painting of a modern smartphone with classic art pieces appearing on the screen.”

Embedded World Moving to Transformers

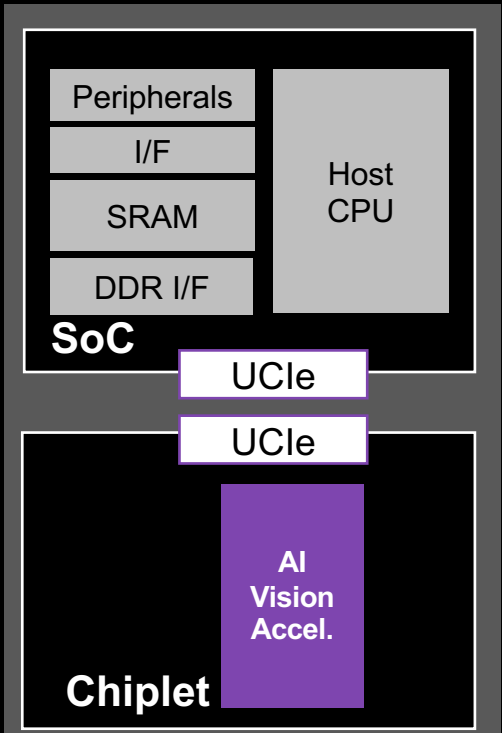
Japanese Automotive OEM: *We will be only running transformers in the future*



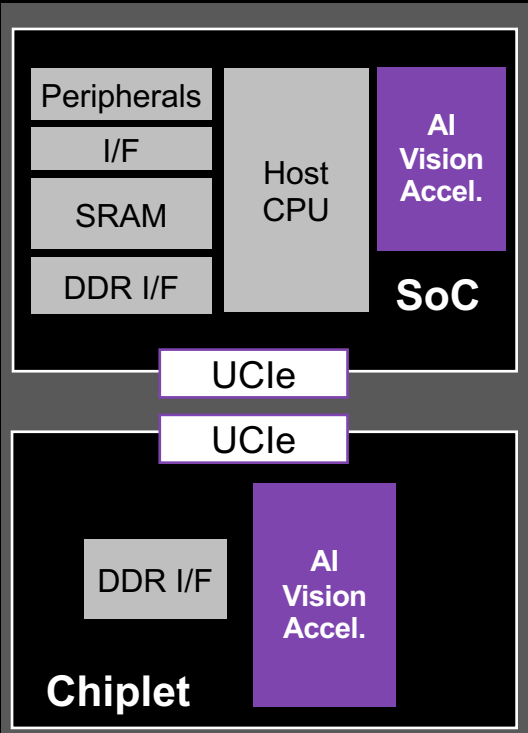
Source: ImageNet Top-1 Accuracy

- Deep learning models originally design for NLP
- State of the art accuracy for vision out of the box
- Attention-based networks better at recognizing more complex objects
- Latest NPUs optimized specifically for high-performance execution of Transformers

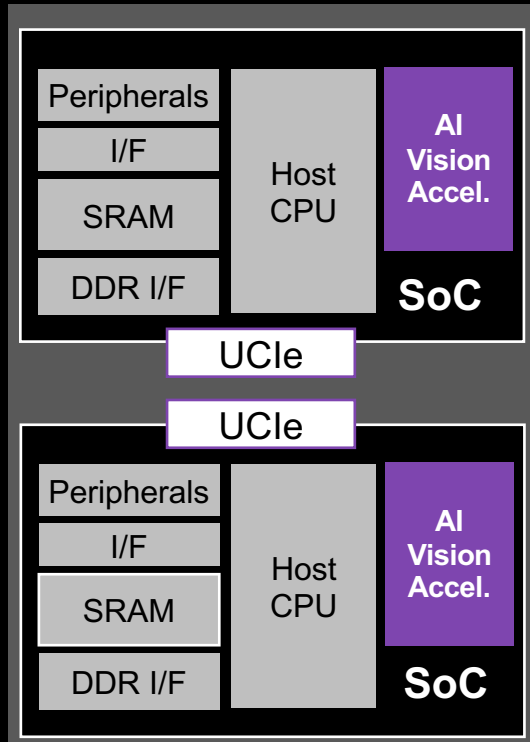
Multi-Die Helps Scale AI Workloads



Scale AI by adding accelerator chiplet to existing SoC (no AI)



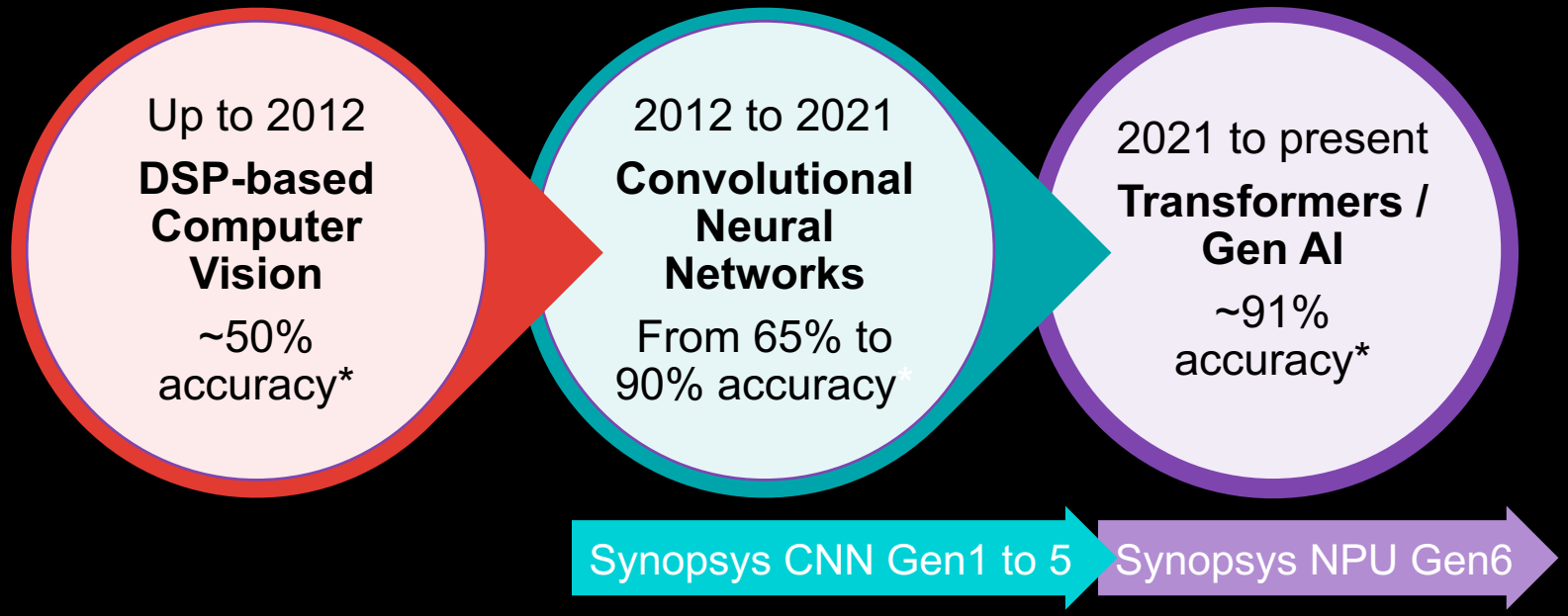
Scale AI by adding accelerator chiplet to existing SoC with AI



Scale AI by duplicating AI-enabled SoC two or more times

- Multi-Die Interfaces
  - UCIe support a strategic focus within Synopsys
- UCIe Die-to-Die Interface
  - Lightweight, low latency & low power “tunnel“ between interconnect fabics in each die
  - Interfaces directly with common fabrics (AXI, CXS, CHI)

ML Moving to Transformers and Gen AI

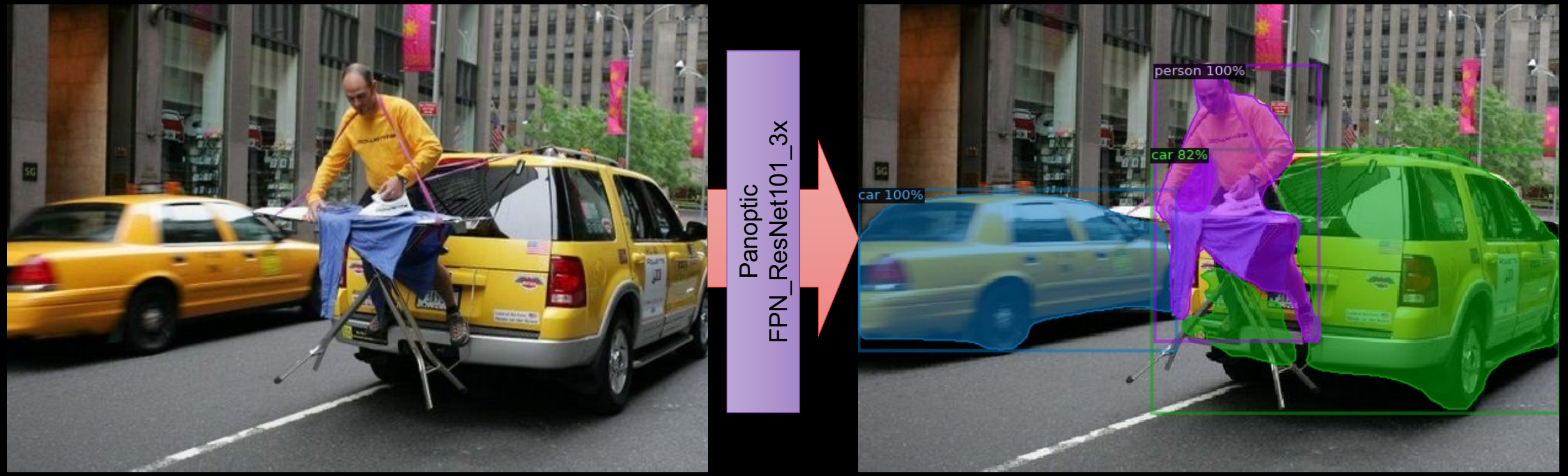


*ImageNet Top-1 Accuracy

Transformers are Replacing CV and RNNs/CNNs and enabling Generative AI

- Synopsys is first company to launch CNN accelerator IP (2014) and first to launch NPU IP with Transformer support (2022)
- Transformers introduce Attention networks which enable emergence of Gen AI
- Generative AI model parameters are orders of magnitude larger than CNNs – requires Transformer support and efficient management of bandwidth

Panoptic Segmentation Using CNNs



Panoptic Segmentation combines predictions from both instance and semantic segmentation in a unified output

Image source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg
Model Used: Detectron2 - COCO-PanopticSegmentation/panoptic_fpn_R_101_3x

Multimodal Transformers

PN_ResNet101_3x

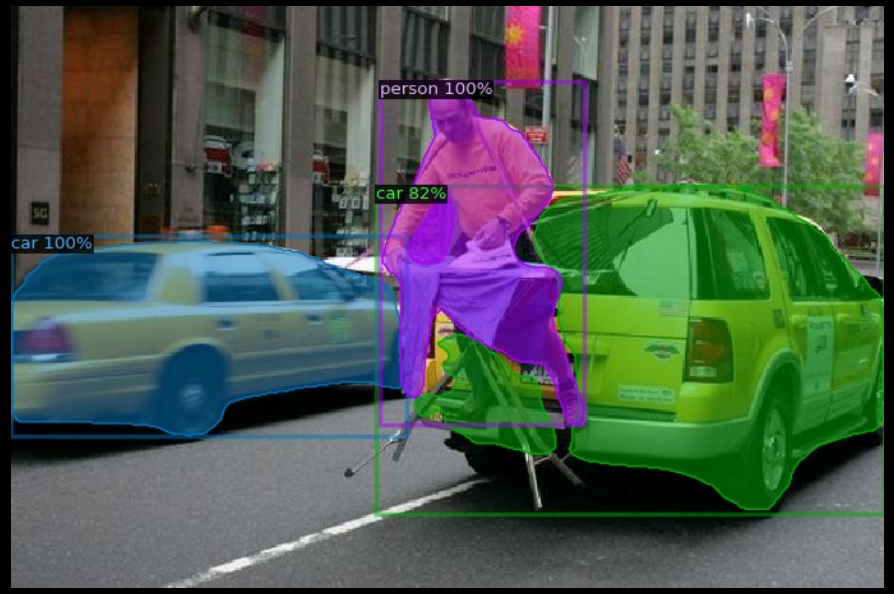
User: What is unusual about this image?

LLaVA: The unusual aspect of the image is that a man is ironing clothes on the back of a yellow minivan while it is on the road. This is an unconventional and unsafe place to perform such an activity, as ironing clothes typically requires a stable surface and appropriate equipment. Ironing clothes in a moving vehicle could lead to potential hazards for both the person doing the ironing and other road users.

LLaVA Language

https://arxiv.org/pdf/2304.08485.pdf

How Much Intelligence Do We Need?



Panoptic Segmentation

User: What is unusual about this image?

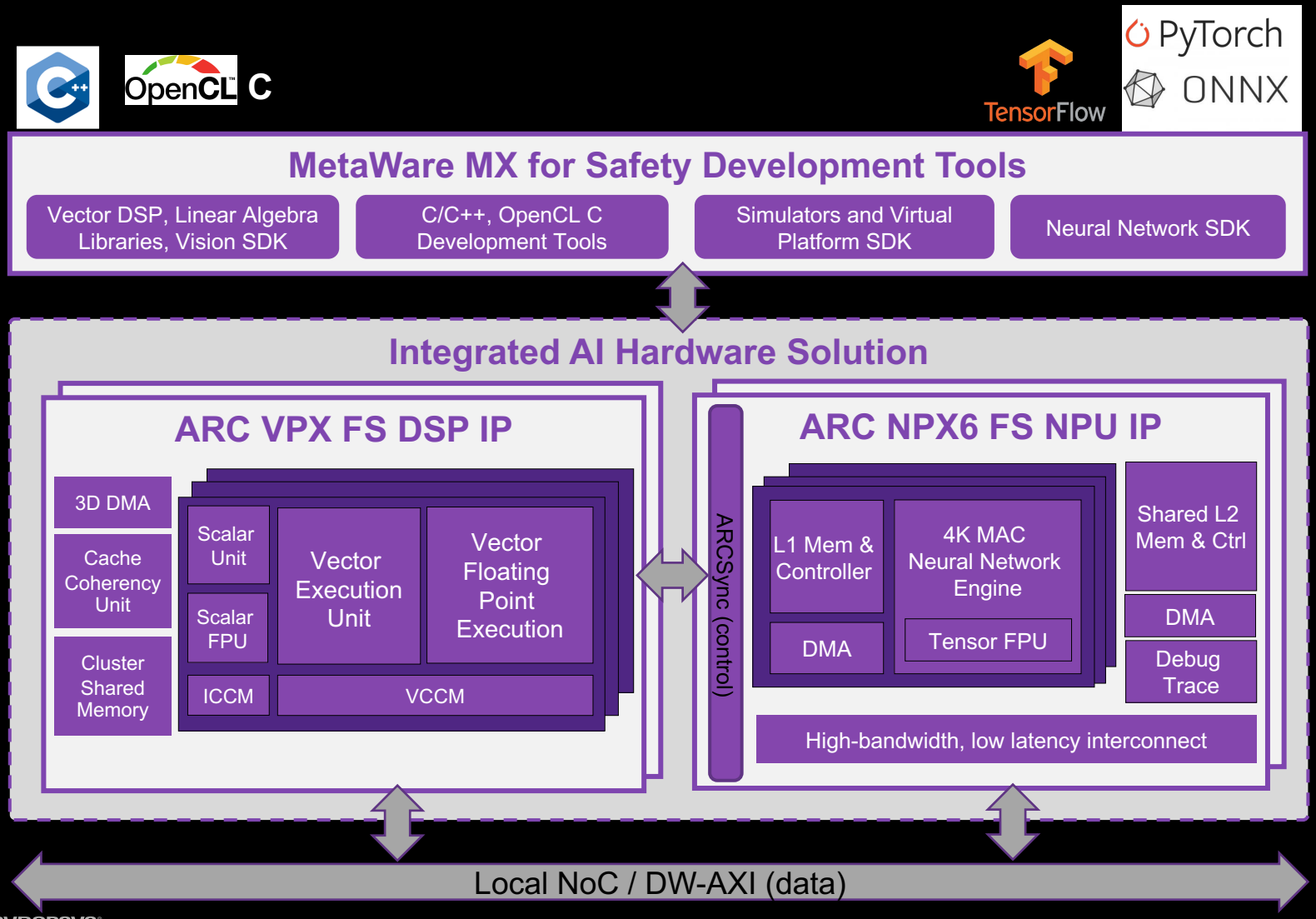
LLaVA: The unusual aspect of the image is that a man is ironing clothes on the back of a yellow minivan while it is on the road. This is an unconventional and unsafe place to perform such an activity, as ironing clothes typically requires a stable surface and appropriate equipment. ….

Multi-modal language/vision

Need to adjust balance
to align with use case requirements

ARC Processor IP & Tools for AI Acceleration

Tightly-Integrated IP with Common SW Development Platform



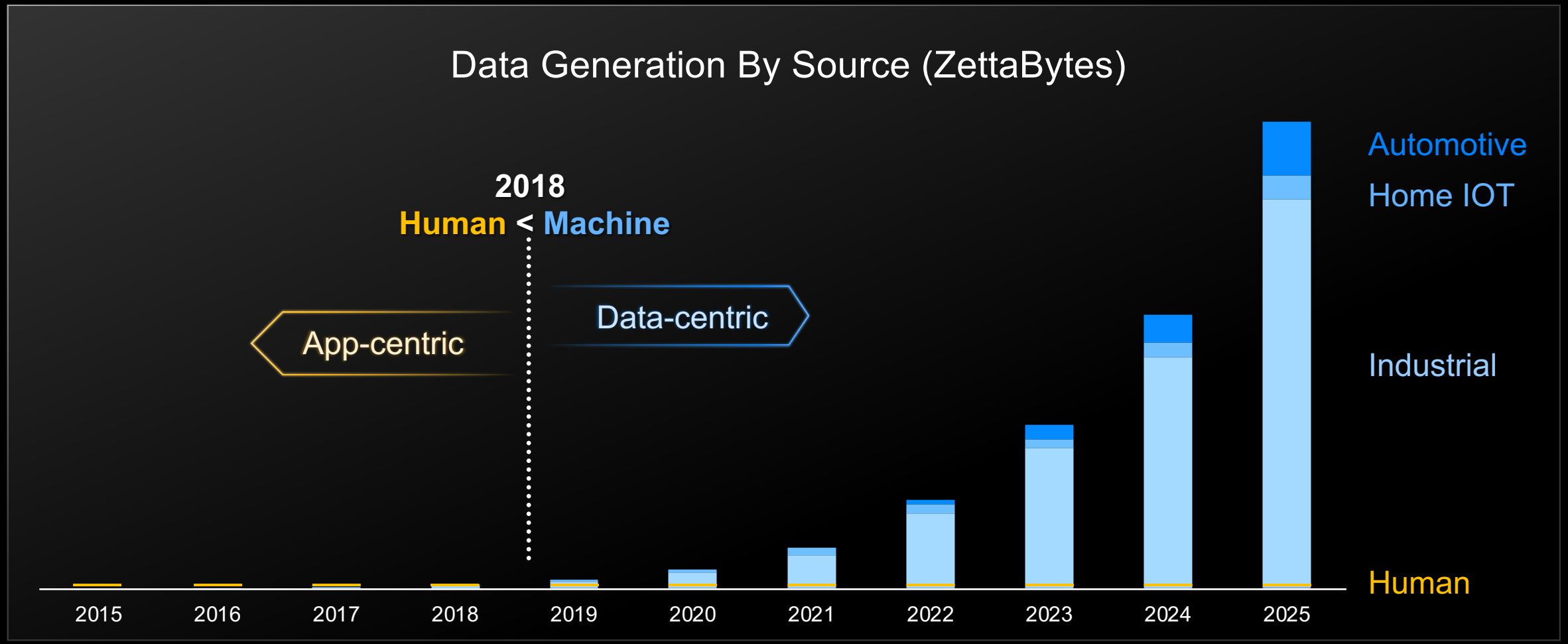
Scalable Architecture Tunable to Specific PPA Requirements

- NPU: 1K → 96K MACs, 1-8 NPUs, 1 to 3000+ TOPS at 1.3GHz 5nm
- DSP: 128/256/512-bit, 1-4 cores
- Floating Point: 0-2 vector FP pipelines, Tensor FPU
- Support for virtual addressing
- DSP & NPU scale independently

Flexible Design Helps Future-Proof SoC for Rapidly Evolving AI

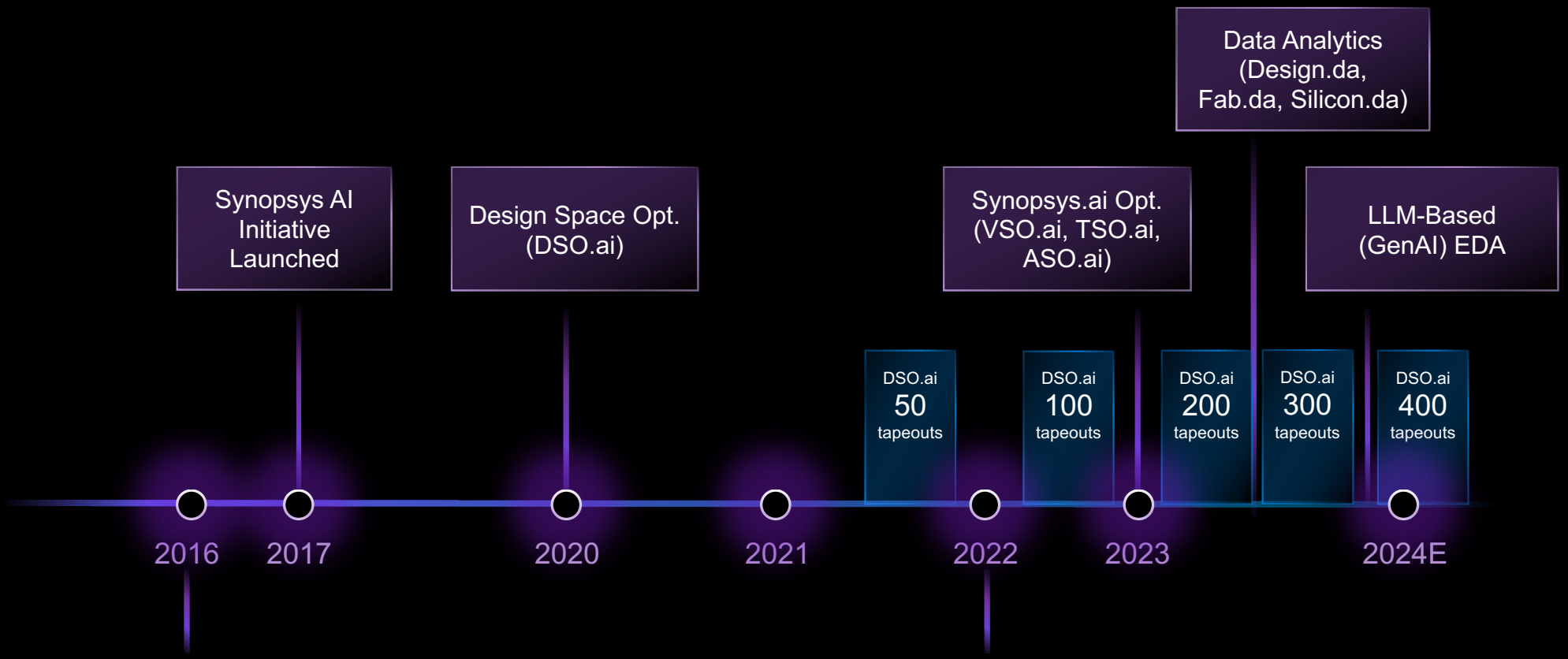
- Data types: INT4, INT8, INT16, INT32, FP32, FP16, Bfloat16
- Multiple languages and machine learning frameworks
- Supports existing and future NNs
- NN SDK optimizes & automates NN mapping, abstracts HW

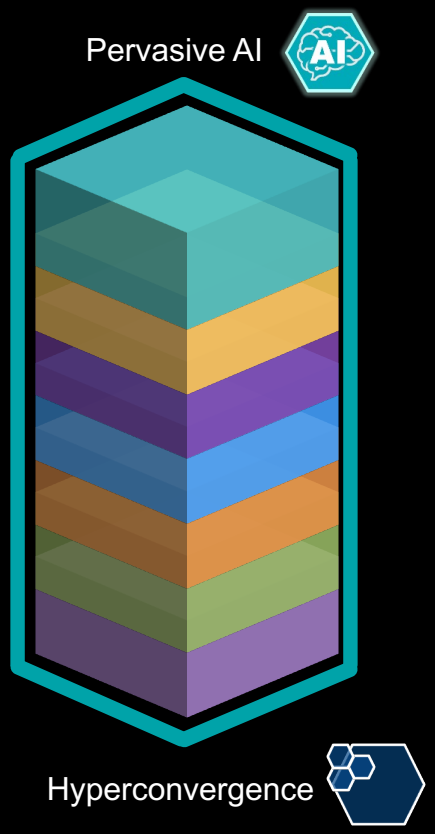
AI Data Explosion



Courtesy: Applied Materials

Blistering Pace of AI innovations





Market Leaders Realizing Significant Gains from Synopsys.ai

DSO.ai*

28nm Image sensor	12% Area Shrink	4nm Mobile SoC	25% Lower Power	FinFET CPU	20% Faster TAT
7nm Automotive CPU	3X Productivity	5nm HPC	4.5% Fmax Boost	6nm Mobile	6.5% Smaller Area

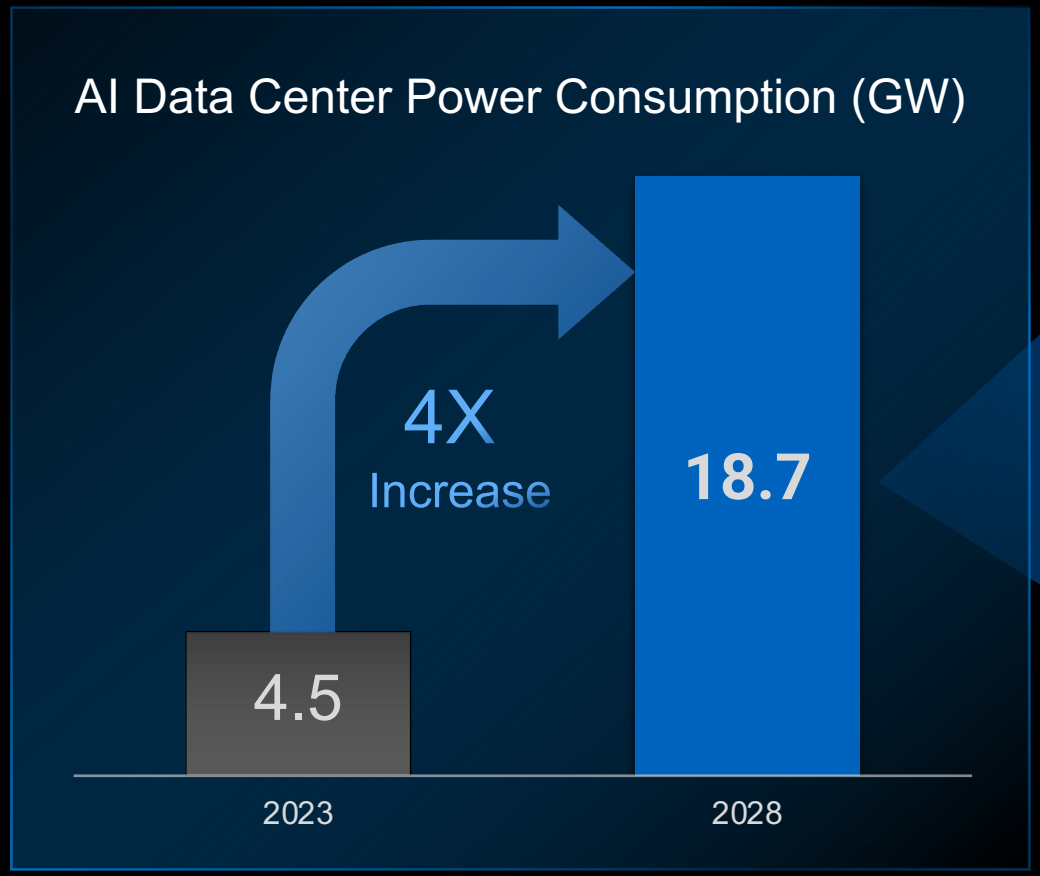
VSO.ai*

HPC	2X Tests Reduction	Automotive SoC	15% Coverage Boost	CPU	2X Faster TAT

Based on results from deployments at:

AMD intel MEDIATEK Microsoft NVIDIA RENESAS SAMSUNG SONY life.augmented

AI Driving Unprecedented Power Consumption



Source: Schneider Electric, December 2023

nVIDIA H100 GPU

700 Watts*

Source: Schneider Electric, December 2023 ** Thermal Design Power*

Energy to train GPT-4

50 GWh

Source: RISE Research Institutes of Sweden, Oct 2023

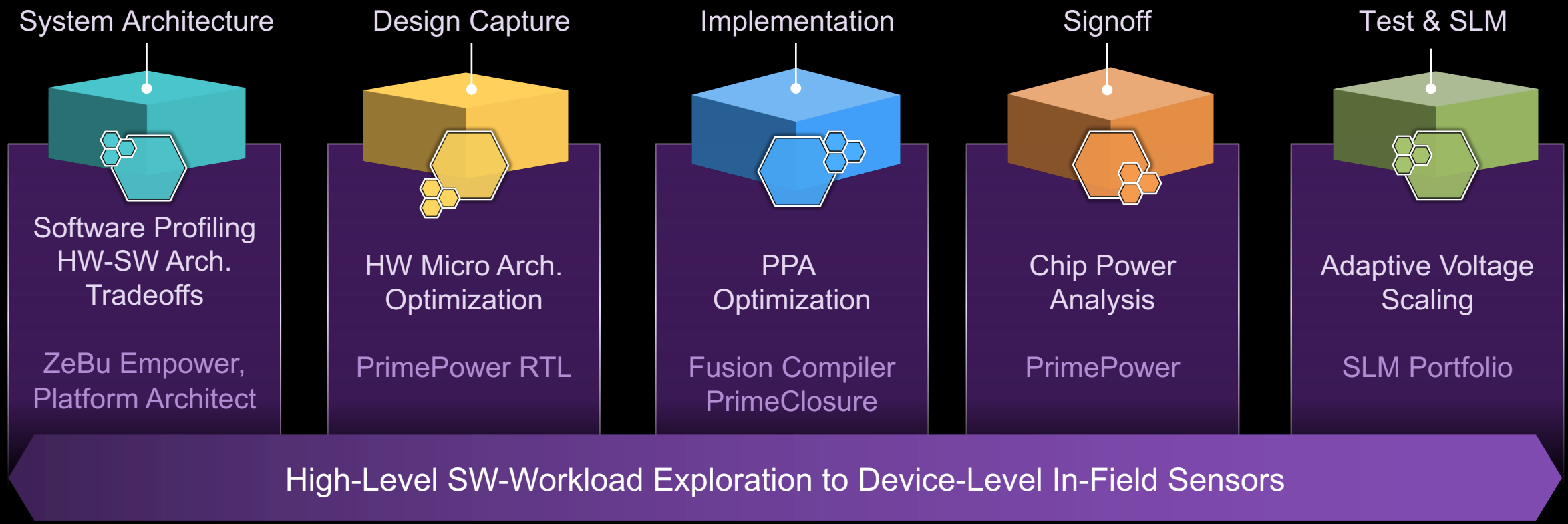
ChatGPT request vs Google search

10X More Power

Source: International Energy Agency, January 2024 report

Software to Device Solution Needed to Address the Magnitude of Power Consumption

Optimizing Power at Every Design Phase



>2X

Savings with Arch. Exploration and SW Profiling of AI Design

20%

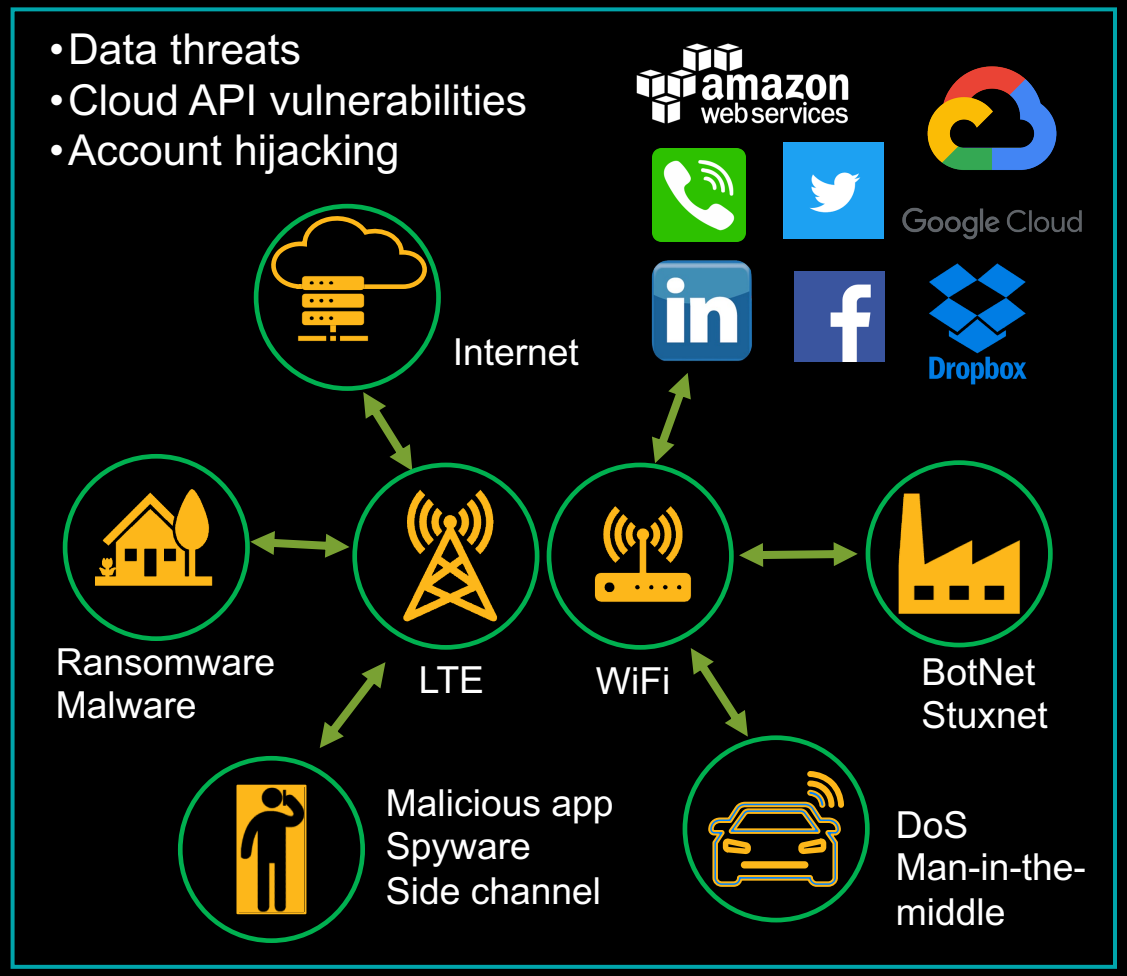
Savings from RTL Changes of NPU Design

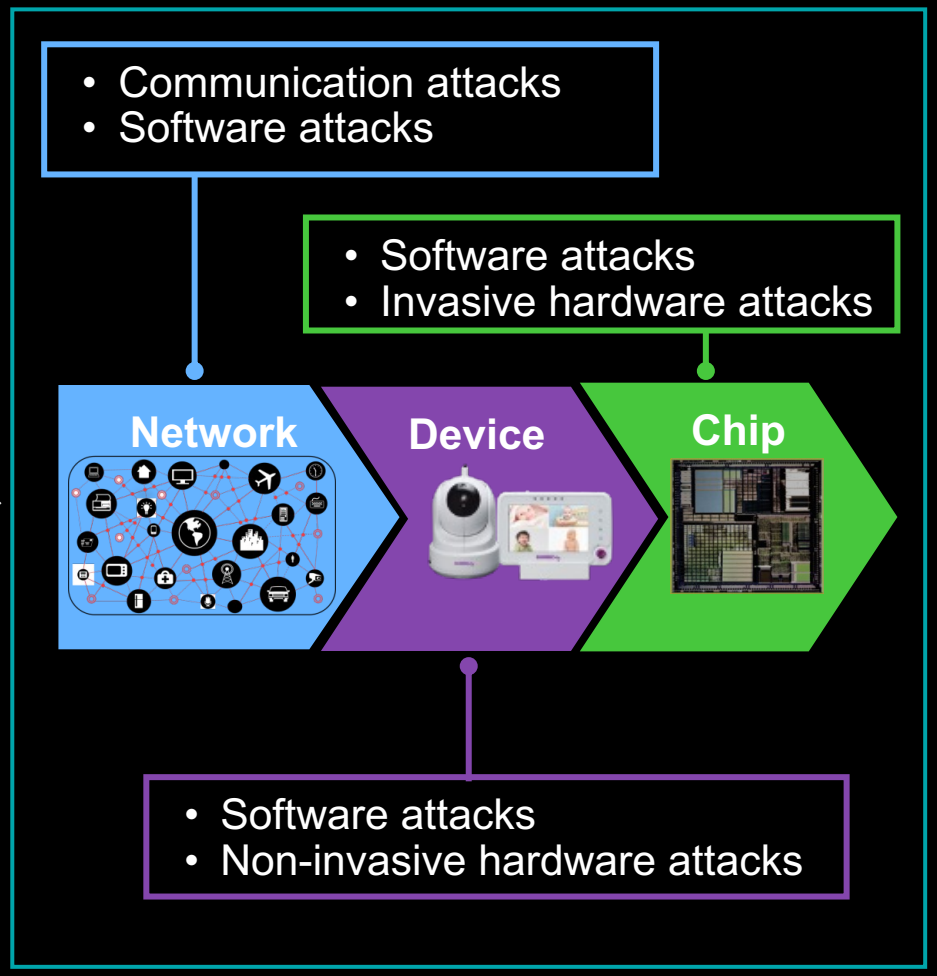
12%

Dynamic Power Savings from Implementation Opt. of HPC Design

All That Data Must Be Protected

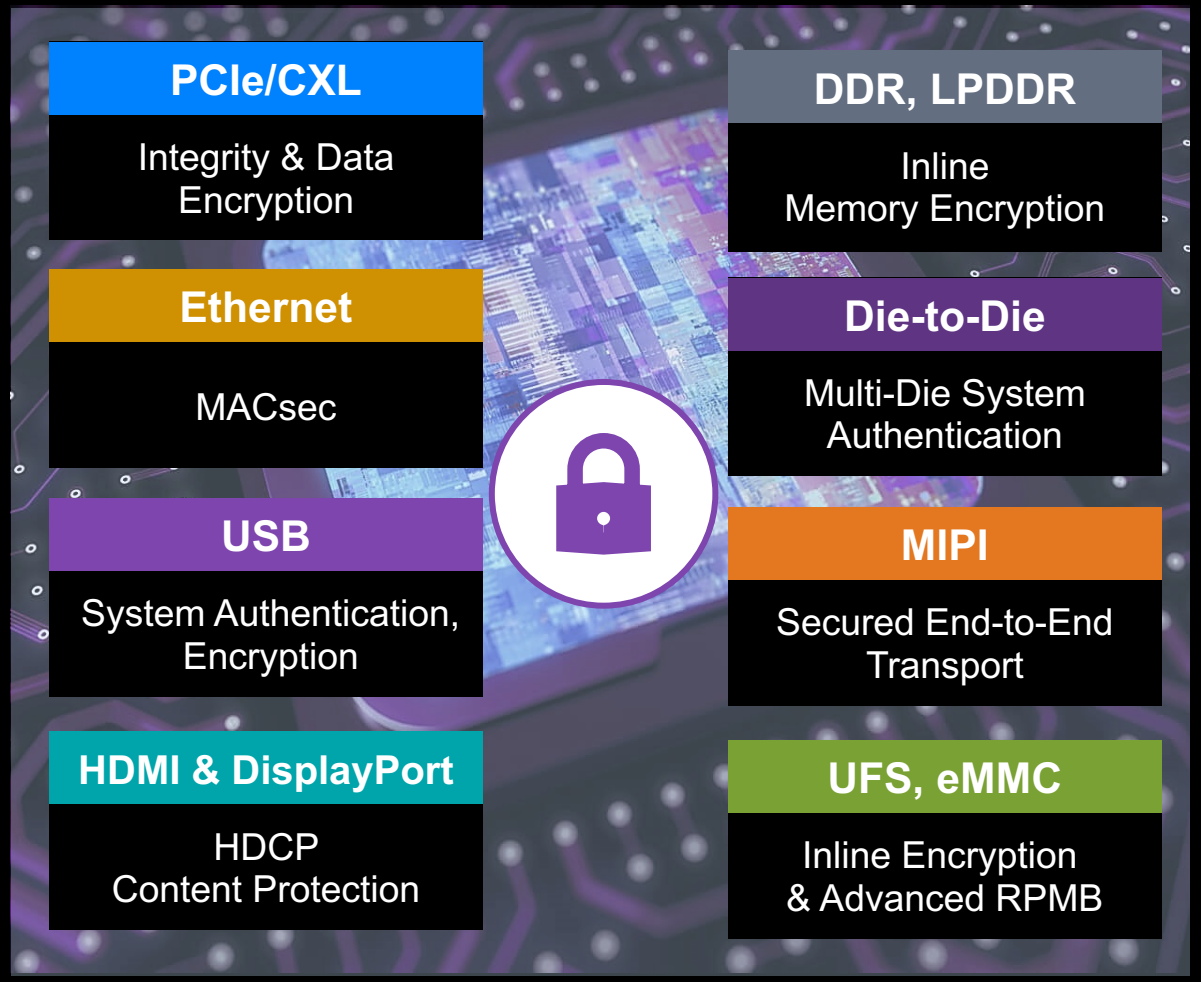
Threats are Coming from Every Direction





Secure SoCs Need Secure Interfaces

Synopsys IP Protects High-Value Data Against Tampering and Physical Attacks



- ✓ **Complete, pre-verified solutions**
  Tightly integrated Controller + Security + PHY enables optimizations for PPA, latency
- ✓ **High grade security**
  Broad, silicon-proven protection for wide range of threats, including side-channel attacks
- ✓ **Standards compliance**
  Tested and certified for compliance to the latest industry interface standards

In Summary

- Significant changes are occurring across many markets, fueling innovation
  - “Smart everything” and ML are creating new challenges and opportunities for semi vendors
  - High quality design IP reduces SoC project risk and time-to-market
- SoC developers are increasingly adopting RISC-V processors
  - Significant ecosystem investments from commercial vendors and the open source community
  - Implementation quality and software tool support are keys to success
- AI is everywhere, rapidly moving from the cloud to the edge
  - NPUs can process ML workloads more efficiently, important for power-constrained devices
  - Vast amounts of data created by AI must be secured, both on-chip and while in transit

Thank You