Key Requirements to Successfully Implement GenAI in Edge Devices — Optimized Mapping to the Enhanced NPX6 Neural Processing Unit IP

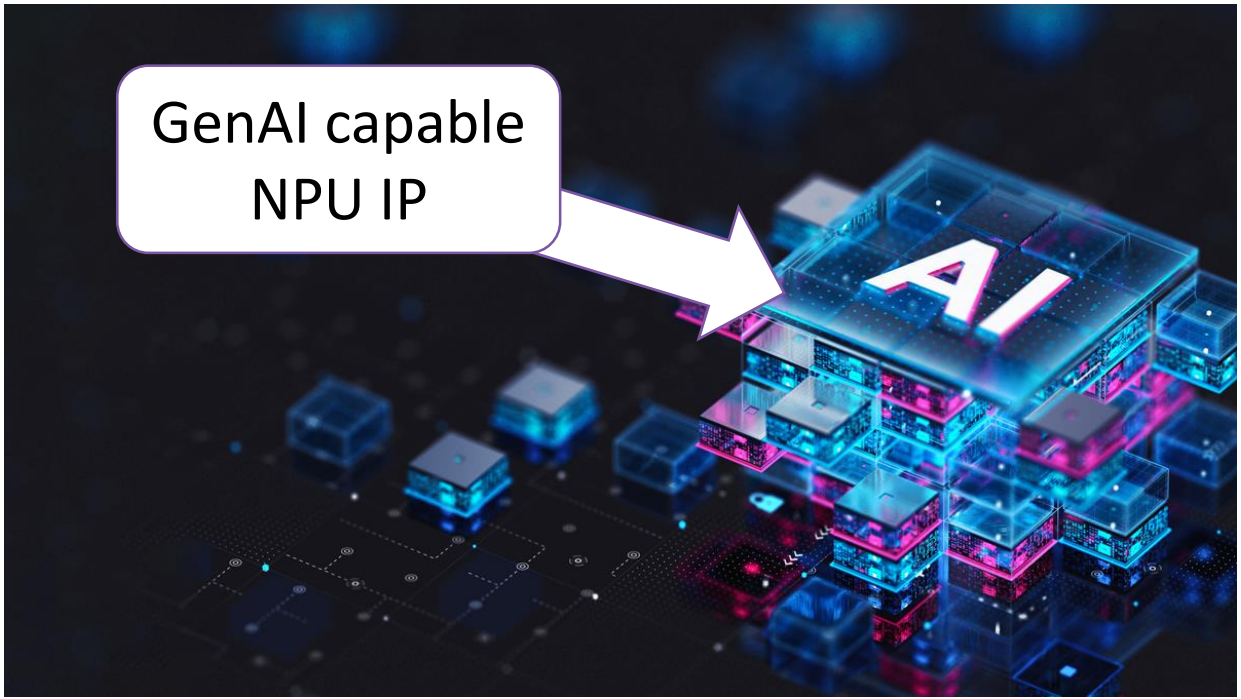
Gordon Cooper

Principal Product Manager

Synopsys

SYNOPSYS®

The Challenge of Fitting GenAI into an Edge Device SoC



Assumptions

- Target solution is an AI-enabled SoC
- GenAI (built on transformer models) capabilities needed
- NPU is needed for transformers / GenAI performance/power efficiency

Extreme Ironing: Panoptic Segmentation Using CNNs

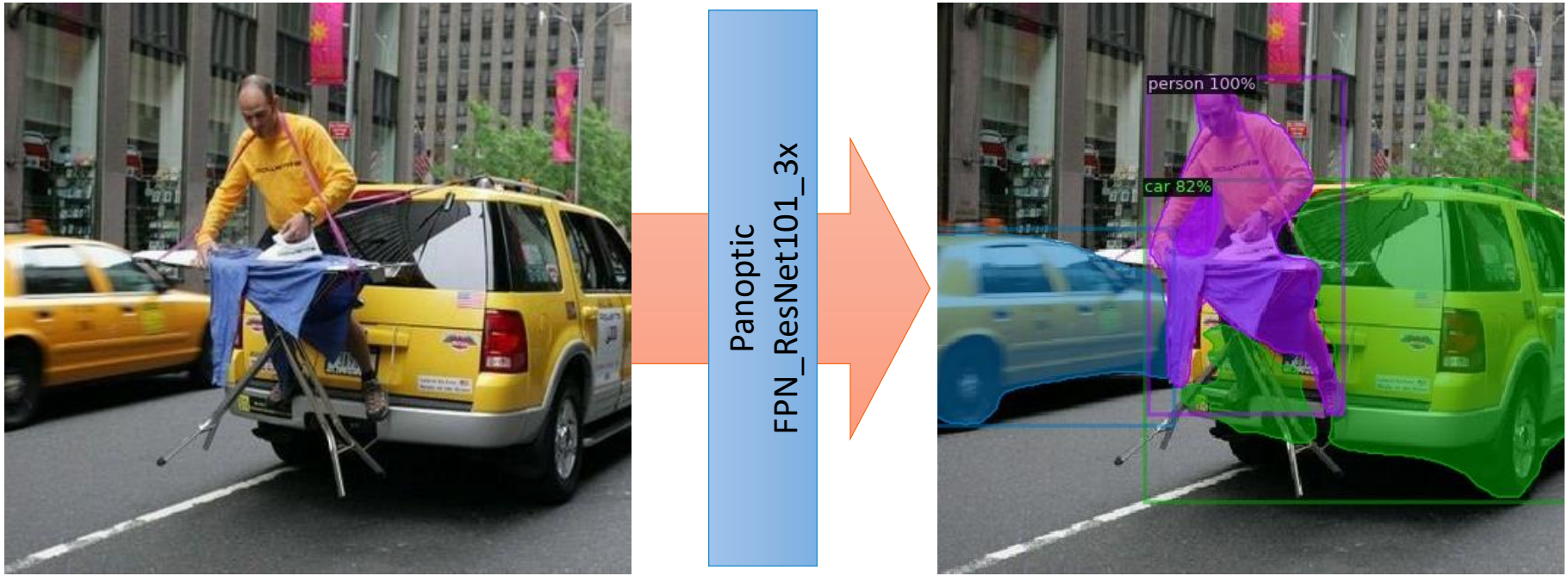


Image source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

Model Used: Detectron2 - COCO-PanopticSegmentation/panoptic_fpn_R_101_3x

Extreme Ironing: Multimodal Transformers Provide Better Contextual Awareness

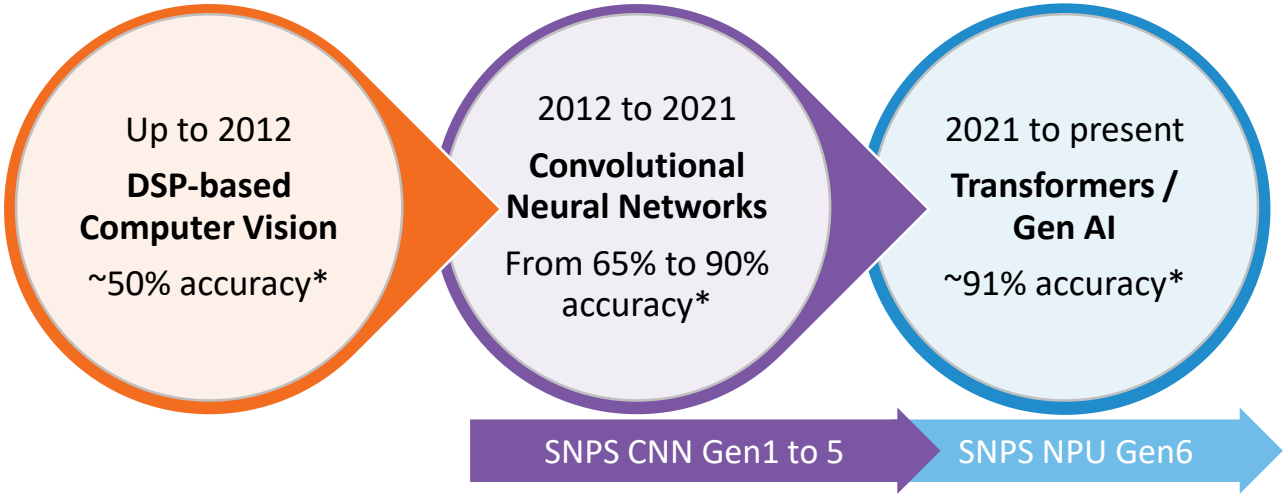
LLaVA
(Large Languag & Vision Assistant)

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.

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

Challenge: AI/ML Technology Evolving (Moving Target)



- Residual connections
- Inception
- Depthwise separable convolutions
- Squeeze and Excitation layers
- New activation functions

MoE
(Mixture-of-Experts)

Uses a collection of smaller expert networks, each specialized in different aspects of the input, to improve performance and efficiency

Concept originated from the 1991 paper *Adaptive Mixture of Local Experts*.

Used in Deepseek, Llama-4, etc.

SYNOPSYS®

**ImageNet Top-1 Accuracy*

Challenge: AI / ML Requirements for AI SoCs Rising

	Last 5 years	Ongoing Designs	Next 3 years
Algorithms	CNNs, RNNs	Transformers, GenAI (Image Gen, LLMs)	Transformers, GenAI (LVMs, LMMs, SLMs)
High End M/L Performance on the edge	100s of TOPS	Up to 1000 TOPS	2000+ TOPS
NPU Data Types	INT8	INT8 / INT4 FP16 / BF16	INT4 / INT8 FP4, FP8, OCP MX
Multi-Die/Chiplet	N / A	UCIe v1.1	UCIe v1.2
Typical Process Nodes*	16 nm / 12 nm	7 nm / 5 nm / 3 nm	3 nm / 2 nm

**ARC Processor IP (NPX6) is process node agnostic*

Challenge: Memory Interface a Chokepoint for GenAI (Especially for edge Devices)

	HBM4	LPDDR5/5x
Common use case	Cloud AI / Training	Edge AI Inference
Max interface bandwidth	1.5+ TB / sec	68 Gbps
Power efficiency (mW/Gbps)	Best	Good
Availability	Poor	Good

- Many customers are avoiding HBM due to cost, limited access to TSMC CoWoS, and DRAM supply issues

Challenge: GenAI Parameters Significantly Larger

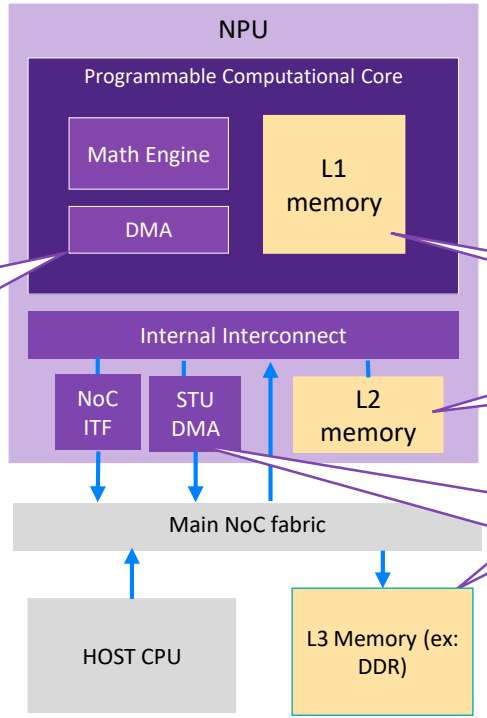
- Generative AI produces compelling results…But parameters required are orders of magnitude larger than CNNs – this makes them bandwidth limited in edge implementations
    - Time to first token
    - Tokens per second

AI Models		Parameters
GPT-4	LLM	1.76 T
LLaVa	LMM	175 B
GPT-3.5	LLM	175 B
Deepseek	LLM	671 B (47B)
Llama 4 Scout	LLM	109 B (17 B)
Llama 2	LLM	7 B / 13 B / 70 B
Llama 3.2	LLM	1 B / 3B / 11 B / 90B
GPT-J	LLM	6 B
GPT 3.5	LLM	1.5 B / 6 B
Deepseek R1 QWEN	LLM	1.5 B / 7B / 14B / 32B
Stable Diffusion	Image Generator	1.5 B
ViT	Vision Transformer	86 M–632 M
BERT-Large	Language Model	340 M
ResNet50	CNN	25 M
Mobile ViT	Vision Transformer	1.7 M

GenAI models <10N parameters

Key Architecture Considerations for NPUs Running GenAI

- ✓ Requires software tools that support HW features and rapid architecture exploration
- ✓ Requires a Programmable Solution Designed for latest Transformers



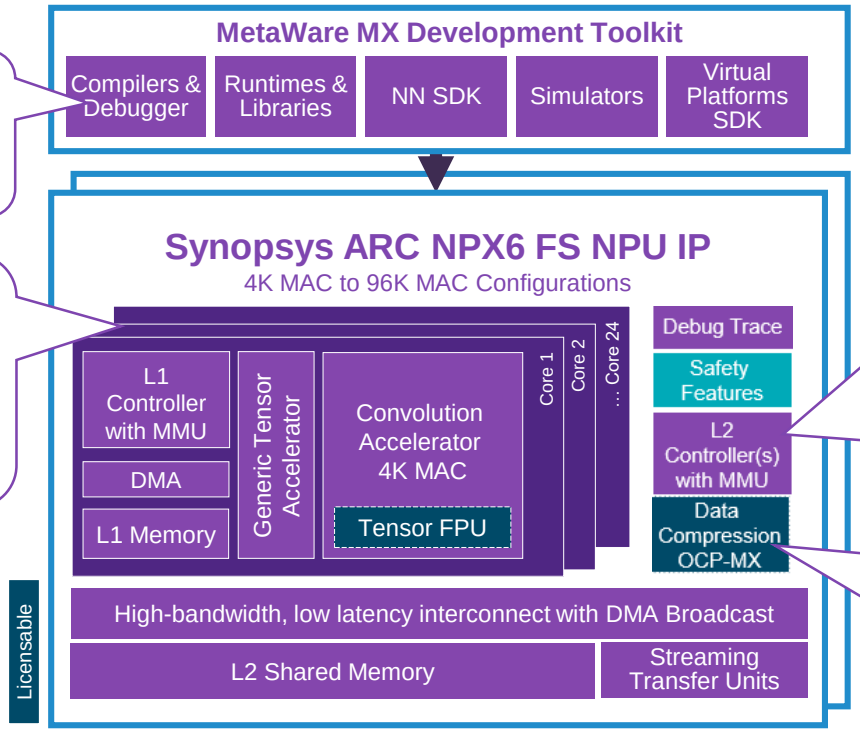
- ✓ Requires Bandwidth reduction in hardware and software
- ✓ Requires multi-level memory management
- ✓ Requires Low-bit resolution support (INT4, FP4, FP6 etc.) for data transfers to minimize BW

ARC NPX6 NPU IP Supports Generative AI for Edge Devices

Trusted software tools scale
- Rapid hardware exploration

Scalable NPX6 processor architecture
- 1 to 24 core NPU w/multi NPU support (3000+ TOPS*)

** 1.3 GHz,5nm FFC worst case conditions using sparse EDSR model*



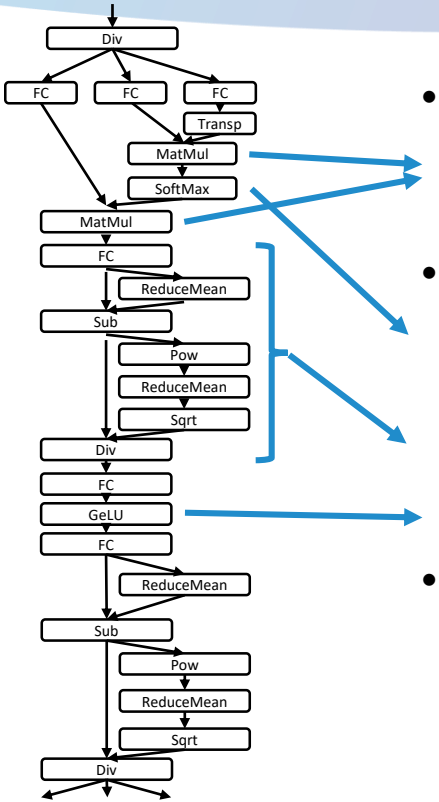
Bandwidth Reduction
- Hardware & SW compression, etc.

Memory Hierarchy –
- high bandwidth L1 and L2 memories
- Powerful data sharing… lowers external memory bandwidth requirements and improves latency

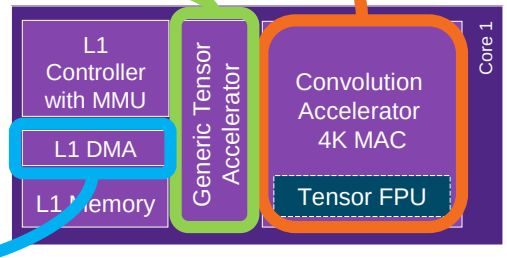
New Data Compression Option
- Supports packing for OCP MX data types, INT

Silicon Proven, Automotive Quality, Synopsys backed

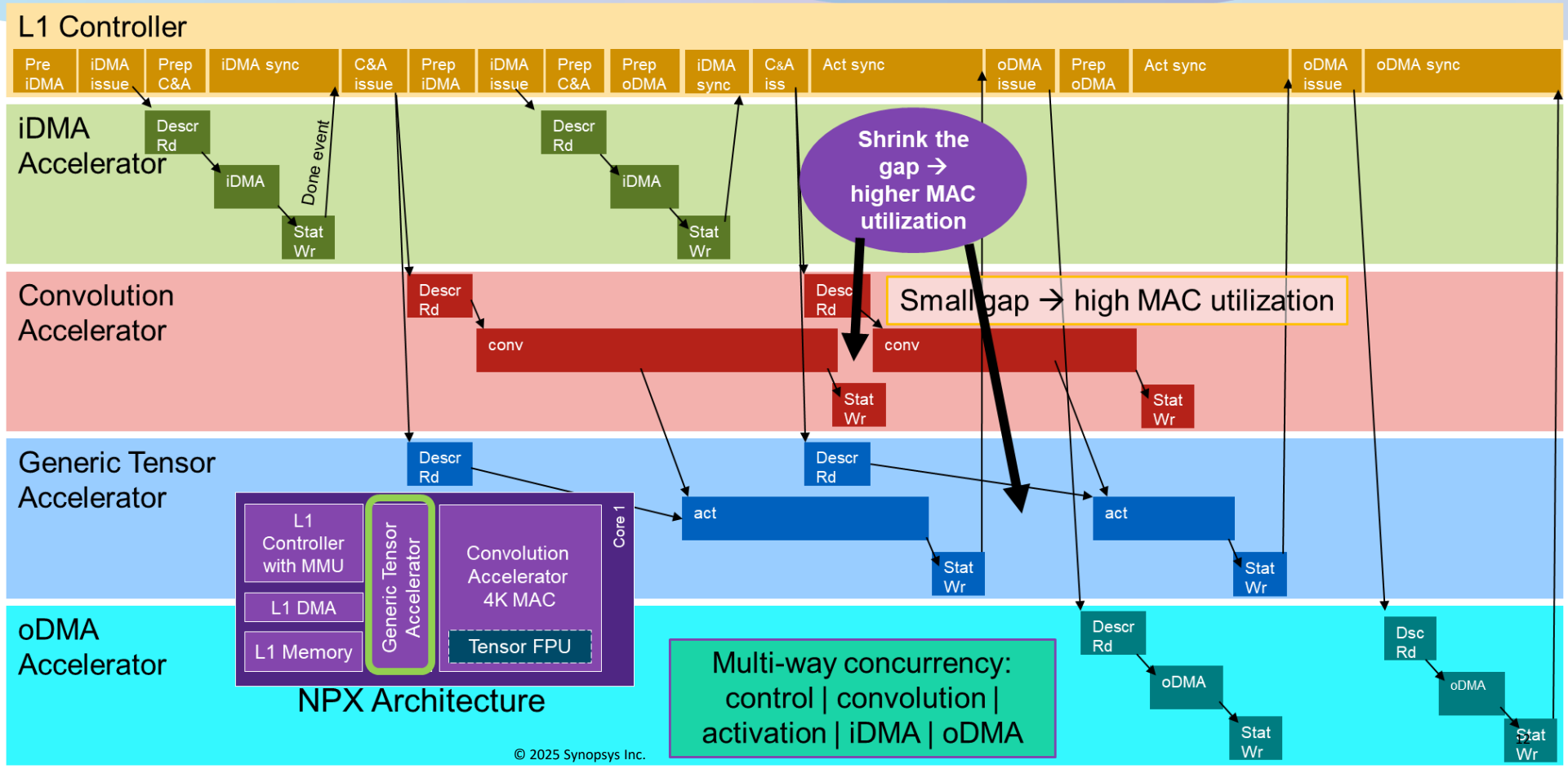
NPX6 Design From Ground up for Transformers Support



- Convolution accelerator feature
  - Support of matrix-matrix multiplications
  - Feature-maps on both operands
- Generic Tensor Accelerator
  - Efficient support for softmax across channels/feature-maps
  - Efficient support for L2 Normalization across feature-maps
  - GeLU support
- L1 DMA – gather support
  - Allows efficient embedding lookups
  - The DMA will read multiple vectors based on a vector of addresses computed by the Generic Tensor Accelerator



Concurrency NPX core (Transformer Optimized)



Enhanced NXP6 NPU IP Supports Many Data Types

Format name	Bits
INT16	16
INT14*	14
INT12*	12
INT10*	10
INT8	8
MXINT8*	8
INT6*	6
INT4*	4

Format Name	Element Type	Bits
FP16	FP16 (E5M10)	16
BF16	BF16 (E8M7)	16
MXFP8*	FP8 (E5M2) FP8 (E4M3)	8
MXFP6*	FP6 (E3M2) FP6 (E2M3)	6
MXFP4*	FP4 (E2M1)	4

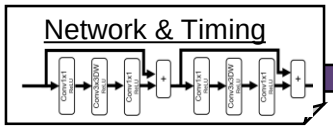
**supported in DMA*

NPX6 Supports Smart Architectural Exploration

IP and SoC-level Architectural Exploration

EMBEDDED VISION SUMMIT

Network & Timing



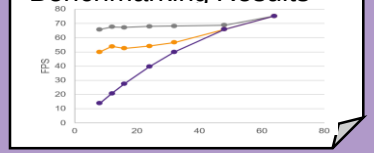
NPX6/model configs:

- # of MACs
- L2 mem size
- Input image size
- DDR Bandwidth
- DDR Latency
- Batch size
- Sparsity
- Quantization

IP Level Performance Analysis

MWMX Analytic Performance Model

Benchmarking Results

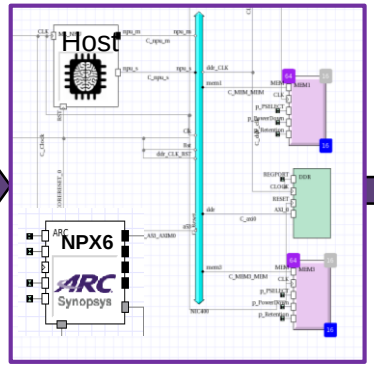


<20% margin of error

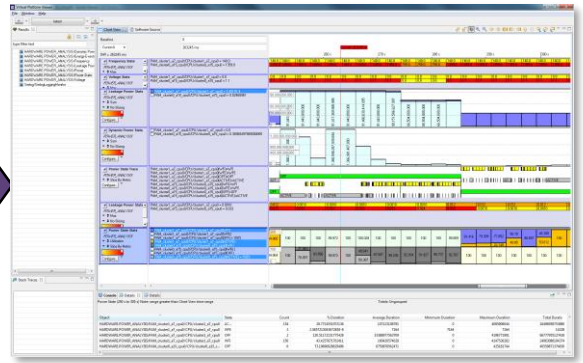
- Throughput
- Latency
- Bandwidths (L2, DDR)
- Energy/Power
- Area
- Stall analysis

Fast iterations (100+)

Integration into Platform Architect



SoC-level Performance Analysis



- Memory architecture analysis
- Interconnect metrics
  - Latency, Throughput
  - Contention, Outstanding transactions
- SoC-level power (*roadmap*)

SYNOPSYS®

NEW

NPX6 Performance, Power and Bandwidth Improvements

EMBEDDED VISION SUMMIT

Enhanced Version of Silicon-proven ARC NPX6 NPU IP family of AI accelerators

- **Transformers Boost** up to 45% better performance on transformer neural network models, accelerating vision and GenAI applications
- **Power Reduction** Up to 10% reduction in power extends battery life and minimizes thermal impact for on-device AI applications
- **AI Data Compression** New option supports input and output of new microscaling (OCP MX) data types, reducing memory footprint and bandwidth pressure for GenAI and other neural networks

SYNOPSYS®

ARC-V Expands on Winning ARC Processor IP Portfolio

Scalable CPU, DSP and AI IP & Tools with Unrivalled PPA Efficiency

EMBEDDED VISION SUMMIT

ARC-V (RISC-V ISA)

RMX Family
Ultra Low Power Embedded

- 32-bit embedded processor, DSP option
- High efficiency 3- and 5-stage pipeline configs

RHX Family
Real-Time Performance

- 32-bit real-time processor, 1-16 cores
- High-speed, dual-issue 10-stage pipeline

RPX Family
Host Processor

- 64-bit host processor, 1-16 cores
- SMP Linux, L2 cache support

Specialty

EV Family
Vision Processor

- Heterogeneous multicore for vision processing
- DNN (Deep Neural Network) Engine

VPX Family
Vector DSP

- SIMD/VLIW design for parallel processing
- Multiple vector FP units for high precision

NPX Family
NPU

- Scalable neural processor units (1K-96K MACs)
- Supports latest AI networks (e.g., transformers)

Classic

EM Family
Embedded MPU

- 3-stage pipeline with high efficiency DSP
- Optimized for low power IoT

SEM Family
Security CPU

- Protection against HW, SW, side channel attacks
- SecureShield to create Trusted Exec Environment

HS Family
High Speed CPU

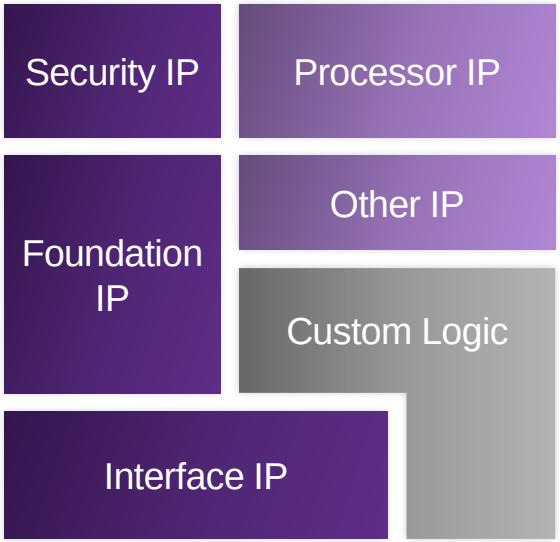
- High performance CPUs, CPU + DSP
- Single- and multi-core configs

Functional Safety (FS) Processors

- Integrated hardware safety features for ASIL compliance across the portfolio (up to ASIL D)
- Accelerates ISO 26262 certification for safety-critical automotive SoCs

SYNOPSYS®

Broadest & Most Advanced IP Portfolio



25 years of investment & commitment

#2 IP provider worldwide

Leader in Foundation IP

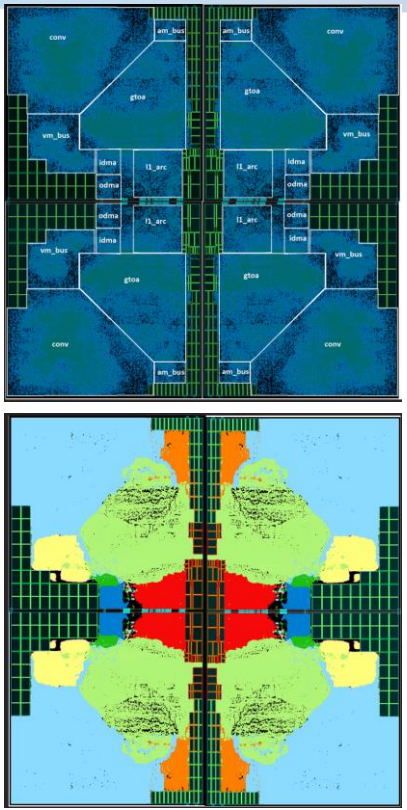
Leader in Interface IP

Growing Processor and Security IP portfolios

Increase productivity and reduce design risk with high-quality Synopsys IP

Summary

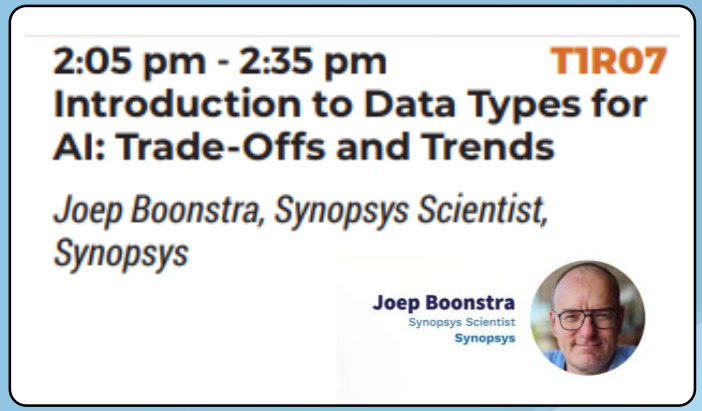
- Transformers lead to state-of-the-art results for vision and speech – and have enabled rise of Generative AI
- Generative AI models can run on NPUs designed for transformers
  - Moving quickly into the embedded space (<10B parameters)
  - Suffers bandwidth bottlenecks due to large parameter size
  - INT4 and MoE-based approaches (like DeepSeek) reduce memory impact
- NPX6 NPU was designed for Transformers and supports Gen AI efficiently
  - Silicon Proven and Scalable solution (includes automotive versions)
  - Enhanced NPX6 NPU IP available now



NPX6-64K layout
(128 dense TOPS at 1 GHz)

Questions?

- **Visit the Synopsys Booth: #717**
- **Check Out the Demos!**
  - Synopsys NN Performance Model Analysis with Platform Architect
  - Visionary.ai Real Time Video Denoiser
  - ADAS NPU Algorithm deployment on working silicon



For more information, please visit:

www.synopsys.com