AI/ML at the Forefront of Semiconductor Evolution: Enhancing Design, Efficiency, and Performance

MPSoC

Yankin Tanurhan

Senior Vice President of Engineering, IP Group

June 18, 2025

The era of pervasive intelligence

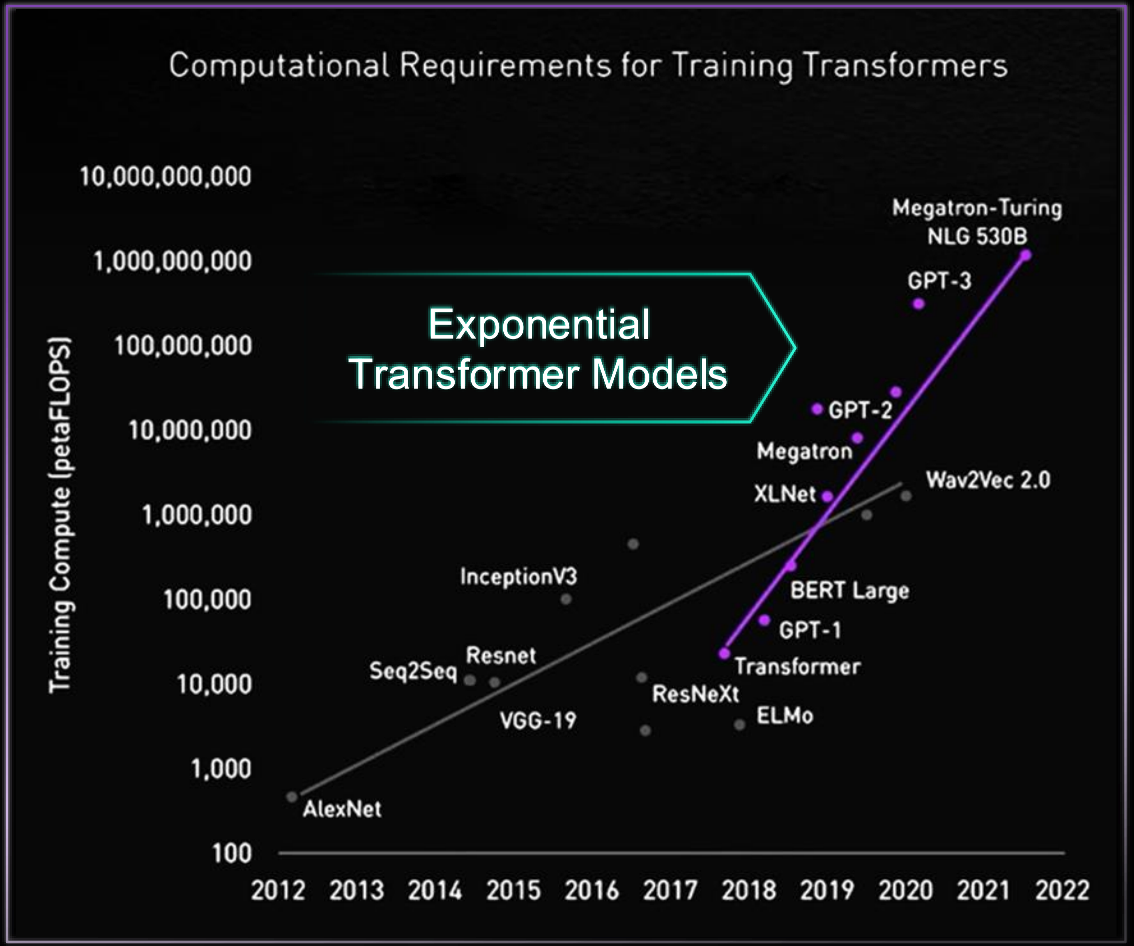
Reinvention of **computing**

Explosion of **intelligent systems**

AI / Datacenter

AI workloads are projected to increase 50x by 2028

AI Transformer Models Further Pushing Limits of Compute



All Models Excluding Transformers:

8X over 2 years

Transformer AI Models:

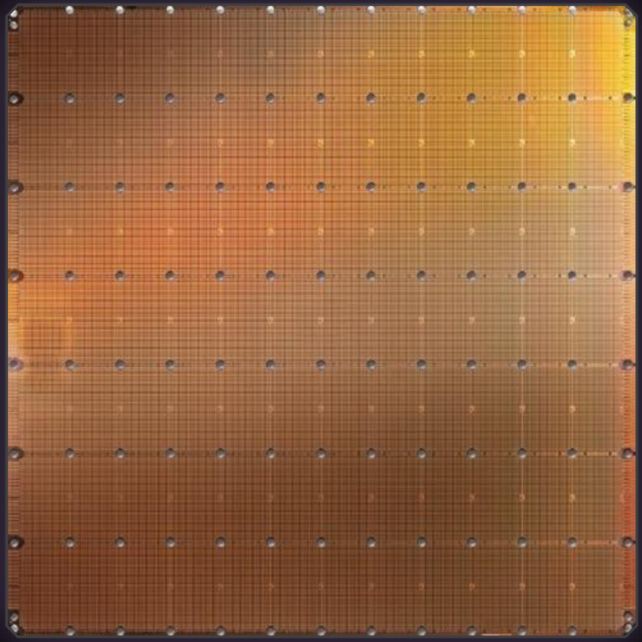
275X over 2 years

Context-Aware Transformer Models Come at a Price

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

Examples of AI “Super-chips”

Cerebras WSE - 2



2.6 trillion transistors

TSMC 7nm

850,000 AI-optimized cores

Graphcore GC200 IPU

59.4Bn transistors

TSMC 7nm @ 823mm^2

1472 independent processor cores

Data center chips for deep learning training and inference

- Trillions of transistors
- Hundreds of thousands of processing elements

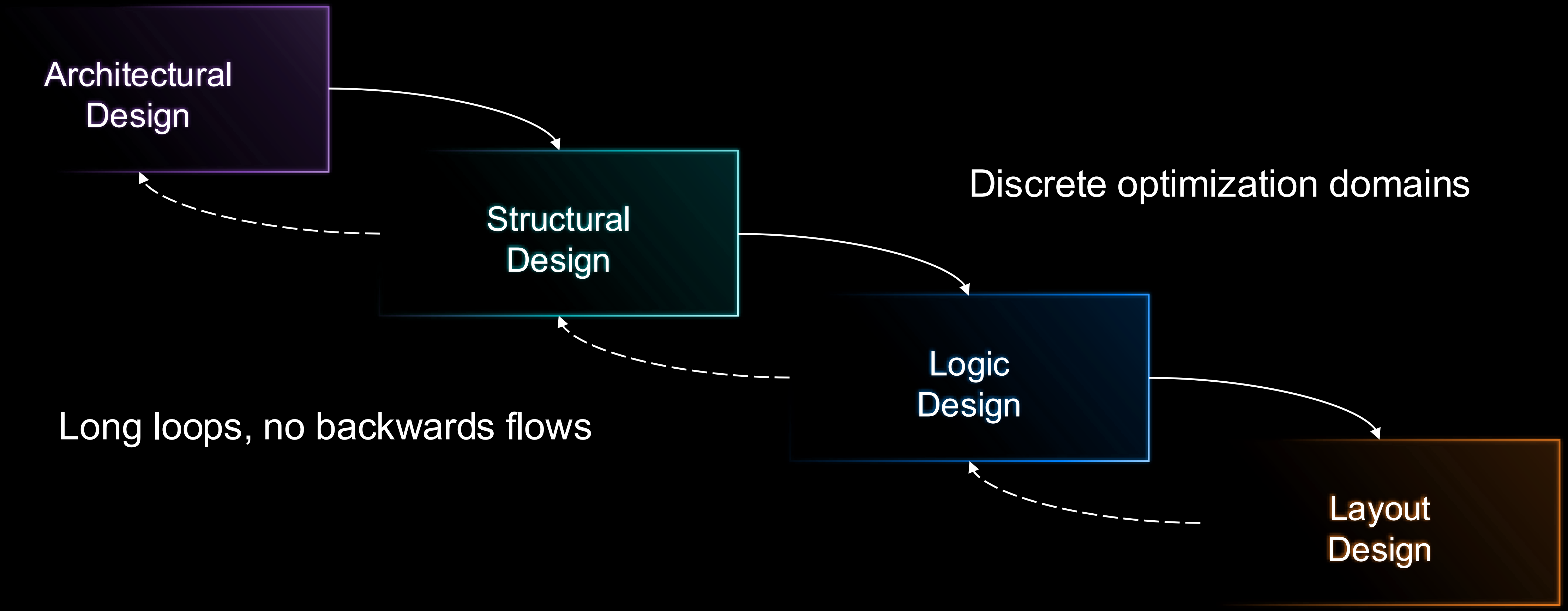
Edge IP (primarily) for deep learning inference

- Mixed scalar/vector/spatial compute
- Ultra energy efficient: Several TOP/s/W

The Power of Generative AI for Chip Design

System Complexity

10^7 X in Design Productivity, Yet Design Takes Time



Discrete optimization domains

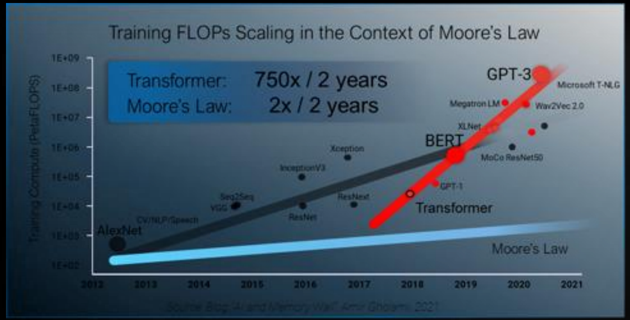
Long loops, no backwards flows

Tens to hundreds of engineers, 24+ months of development

Why AI, Why Now?

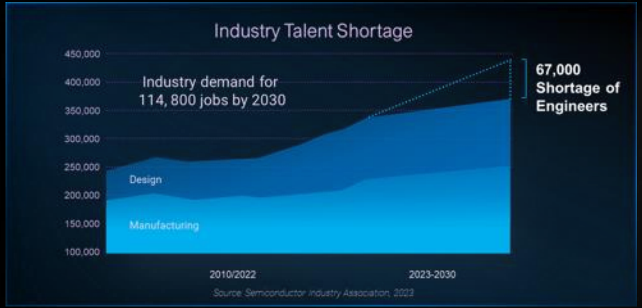
Chip Design Complexity, Cost and Labor Shortage Drive the need for Productivity

Growing Demand & Systemic Complexity



+

30% Talent Shortage By 2030

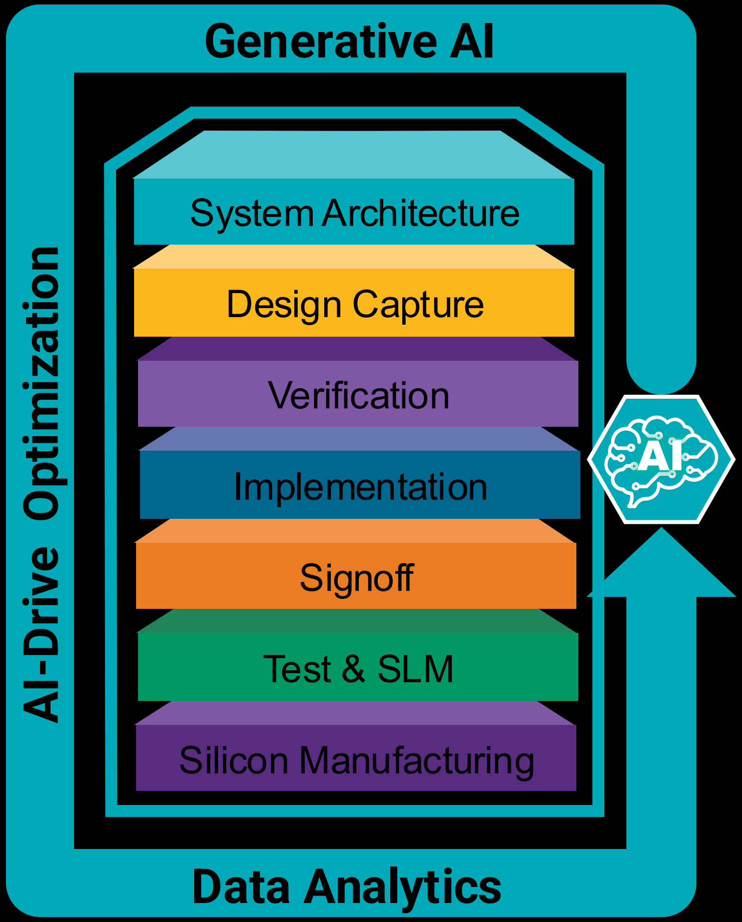


=

Catalyst for AI Design

Chip Design Complexity, Cost and Labor Shortage Drive the Need for Productivity

Synopsys Pioneered EDA AI with Full-Stack Synopsys.ai



AI Optimization

DSO.ai, VSO.ai,

TSO.ai, ASO.ai…

Better, faster results

across all EDA steps

2020

Data Analytics

Design.da, Fab.da

Silicon.da

Rapid root-cause analysis

with expert remedies

2022

Generative AI

Collaborative

Generative

Copilots for EDA engineers

Generation of EDA inputs

2023

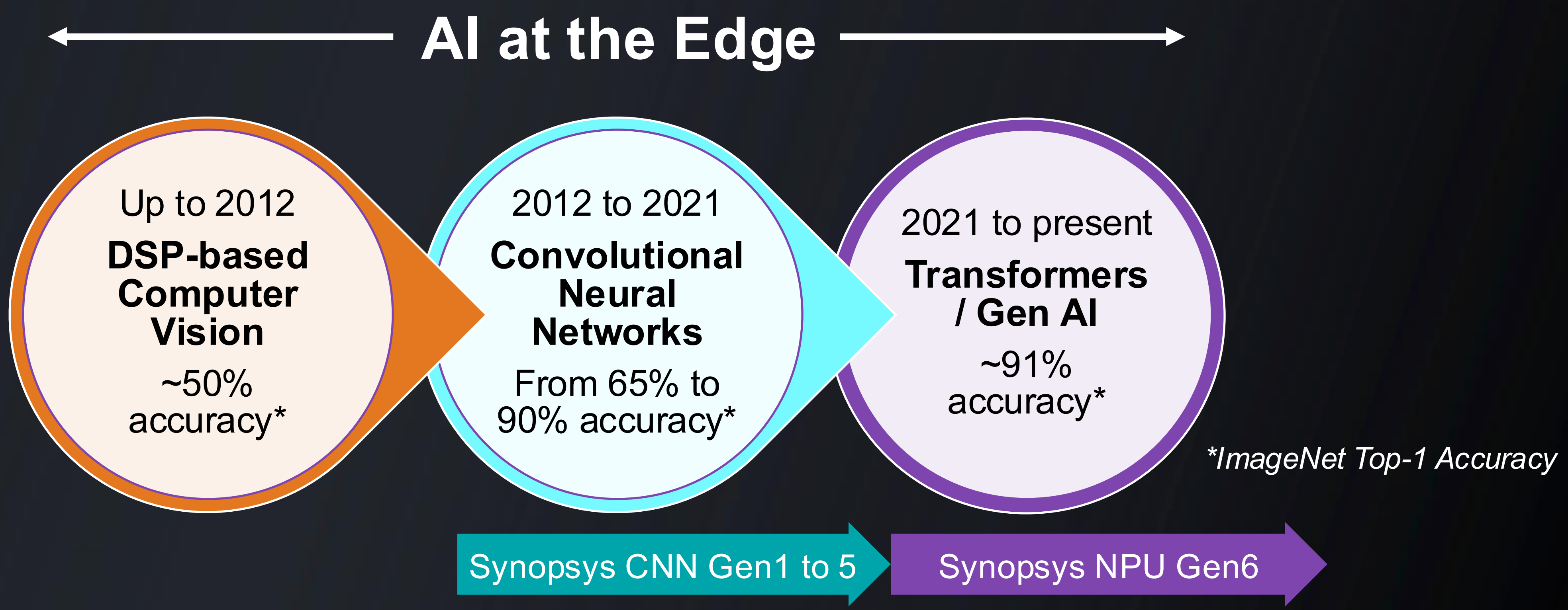
Realizing Significant Gains in SoC Dev’t from Synopsys.ai

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
TSO.ai	Mobile: 45% Pattern Count Reduction	Systems: 18% Pattern Count Reduction	Graphics: 25% Pattern Count Reduction

AMD intel MEDIATEK Microsoft NVIDIA RENESAS SAMSUNG SONY ST

Developing SoCs that Support GenAI on Edge Devices

Transformers and Gen AI Moving from the Cloud to the Edge

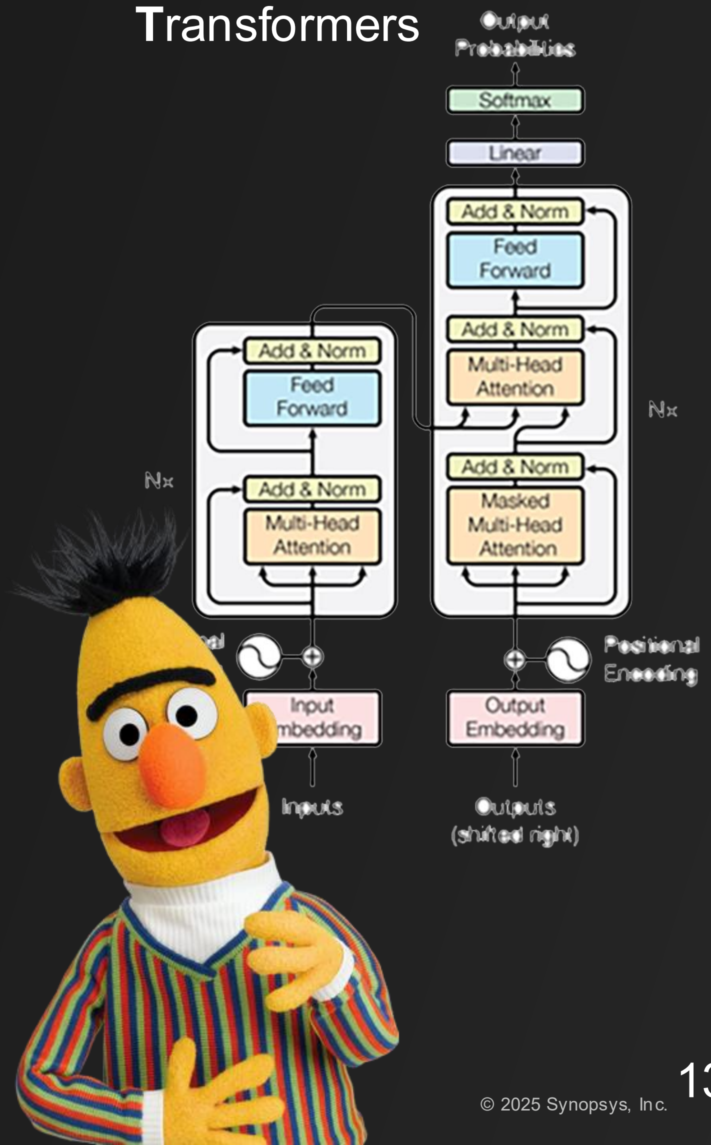


Synopsys was the first company to launch a CNN Accelerator IP (EV6x CNN, 2014) and first to launch NPU IP with Transformer support (NPX6-4K, 2022)

Overview of Transformers

- Transformers are deep learning models primarily used in the field of NLP (and basis for ChatGPT)
- Transformers lead to state-of-the-art results in other application domains of deep learning like vision and speech
  - They can be applied to other domains with surprisingly little modifications
  - Models that combine attention and convolutions outperform convolutional neural networks on vision tasks, even for small models
  - ImageNet accuracy achieved with Transformers surpassed ten years of CNN innovation (90%) in less than two years (>91%)
- Transformers and attention for vision applications are here to stay
  - Real world applications require knowledge that is not easily captured with convolutions

Bidirectional
Encoder
Representations from
Transformers



Transformers Compute Requirements and Model Size

- Compute requirements for early Transformer models are much higher
  - Performance comparison (for same NPU configuration)

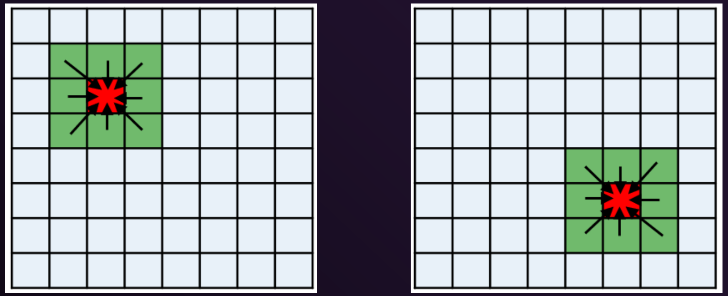
NN Model	Image size	Top 1 Accuracy	Relative GOPS	Relative Frames/sec
CNN-based MobileNetv2	224x224	69.8%	1X	32X
Vision Tranformer ViT_B_16	224x224	84.0%	58X	1X

- All the state-of-the-art models (CNN and Transformers) are huge
  - Approx. 2G parameters
  - Impractical for use in embedded applications

Transformers (Attention) Excels in Capturing More Complex Patterns

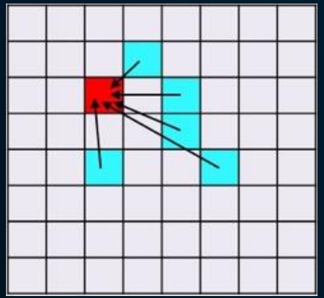
Transformers are Building Blocks of GenAI

Recognizing Cat Fur



Convolution use the same weights for every pixel
This works well for detecting features like cat fur.

Recognizing a Whole Cat



More complex patterns like natural language, or a whole cat, benefit from the attention mechanism in transformers

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

LLaVA Language

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.

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

Challenges for Implementing GenAI on Edge Devices

- Compute requirements for early Transformer models much higher than CNNs
- Generative AI model parameters are orders of magnitude larger than CNNs – this makes them bandwidth limited in edge implementations

Text input

What is the capital of Paris?

Llama2 (LLM)

Text output

Paris is a city, and it does not have a capital as it is itself the capital city of France. Capitals are typically assigned to countries or states, not cities. Paris is known for being a major cultural and economic center in Europe and is one of the most famous cities in the world.

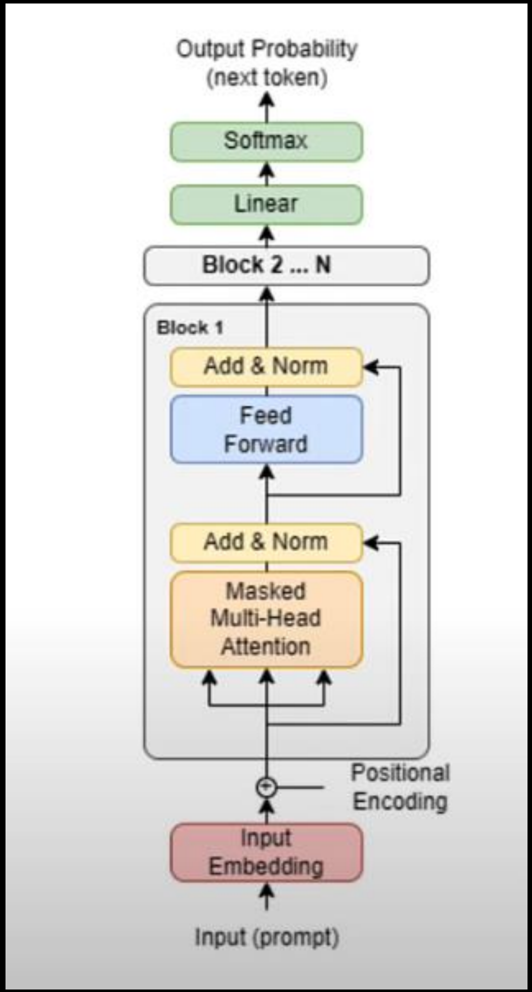
AI Models		Parameters
GPT-4	LLM	1.76T
LLaVa	LMM	175B
GPT-3.5	LLM	175B
Deepseek R1	LLM	37B / 671B
Llama 2	LLM	13B / 70B
Llama 2	LLM	7B
Baichuan	LLM	7B
GPT-J	LLM	6B
GPT 3.5	LLM	1.5B / 6B
Stable Diffusion	Image Generator	1.5B
GigaGAN	Image Generator	1B
ViT	Vision Transformer	86M–632M
BERT-Large	Language Model	340M
ResNet50	CNN	25M
Mobile ViT	Vision Transformer	1.7M

GenAI models <10M parameters

Generative AI - Llama 2 - 7B: Challenges and Solutions

- Large model size →
  compress model with support of INT4 for coefficients and INT16 for feature-maps
- Loading of weights and data to L1/L2 memories from DDR →
  NPX6 uses DDR interface efficiently
- Data locality: keep intermediate data close to NN cores →
  NPX6 exploits high bandwidth L2 memory
- Input prompt processor →
  Efficient mapping using batch processing
- Efficient Attention support →
  NPX6 instruction set is designed for Transformers
- Embedding lookup → DMA Gather support
- Computational complexity of Softmax & Normalization →
  NPX6 leverages Generic Tensor Accelerator,
  designed for flexibility and efficiency of non-MAC-oriented NN operators

➢ **NPX6 performance matches leading public benchmarks for same bandwidth constraint (approx. 30 tokens/sec)**



Memory Interface for AI – Cloud vs Edge

LPDDR features include low power optimizations

	HBM2/2E	LPDDR5
Common use case	Cloud AI	Edge AI
Typical Interface	Octal 128 bit channels (1024 bits total)	Dual/Quad 16 bit channels (32/64 bits total)
Max interface bandwidth	307 → 461 Gbps	51 Gbps
Power efficiency (mW/Gbps)	Highest	High
Interface voltage	1.2V	0.5V / 0.3V
Power Down Mode(s)	No	Yes

- LPDDR uses very little power when not in use
- LPDDR can quickly throttle performance via frequency
- LPDDR5x/5/4x features
  - PLL bypassing
  - Standby state for Receivers
  - Programmable PHY-side ODT strength
  - Support for power down and self refresh per rank
  - DM/DBI tri-stated when not needed
  - LPDDR5 Strobe mode disabled for low data rates
  - Low Latency Standby Modes

AI Technology Evolution: Trends for Edge Devices

	Last 5 years	Ongoing Designs	Next 3 years
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
Typical 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
Algorithms	CNNs, RNNs	Transformers, GenAI (Image Gen, LLMs)	Transformers, GenAI (LVMs, LMMs, SLMs)
Functional Safety	Limited use of AI in automotive	Fast adoption of AI in automotive	Systematic use of AI in automotive

*ARC Processor IP (e.g., ARC NPX6 NPU) is delivered as soft IP so process node agnostic

- Synopsys's broad portfolio of IP includes ARC NPX6, LPDDR and UCIe
- Designing for automotive safety drives a culture of quality across all Synopsys IP products

NPUs Most Efficient Processor for Edge AI/ML

Programmability and Ease of Use Key Considerations Due to Pace of NN Innovation

	CPU	GPU	NPU
Performance	Low	Mid	High
Power	Mid	High	Low
Cost	High	High	Low
Programmability	High	High	Mid

ASIC
Highest
Lowest
Lowest
Low

GenAI Stable Diffusion Images to Minute (512x512, v1-5, 50 steps)

GPU for AI
NVIDIA Titan RTX
>200 W

Similar performance

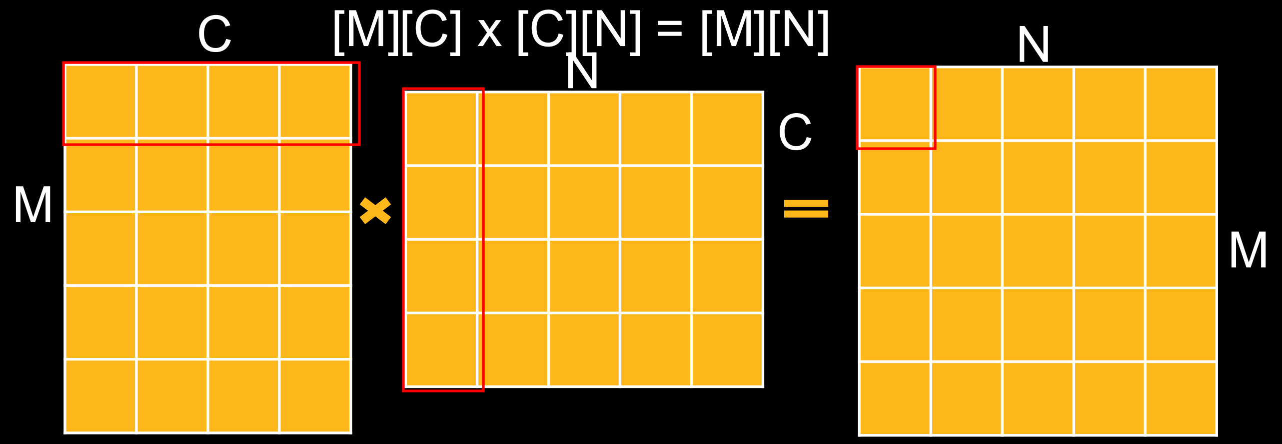
NPU IP
NPX6-32K
<2 W (5 nm)

- China’s DeepSeek generative AI application overtook ChatGPT to become the top app in the Apple App Store
- DeepSeek-V3 model runs on less expensive NVIDIA H800 PCIe graphics cards
- Claims the model was developed for around $6 million
  - The cost per inference is 95-98% lower than that of OpenAI
  - Benchmark tests showed it outperformed Llama 3.1 and matched the performance of GPT-4o
- Key innovation is in the training approach
  - Reinforcement learning
- Use of Mixture of Experts (MoE) for inference model

  Reduces effective model size from 600 B to 32 B weights

AI Memory Wall,
Matrix x Matrix versus Matrix x Vector Computations

Compute Bandwidth: $BW_c = M*N*C$
Memory Bandwidth: $BW_m = (M+N)*C$
Memory/Compute Ratio = $(M+N) / (M*N)$



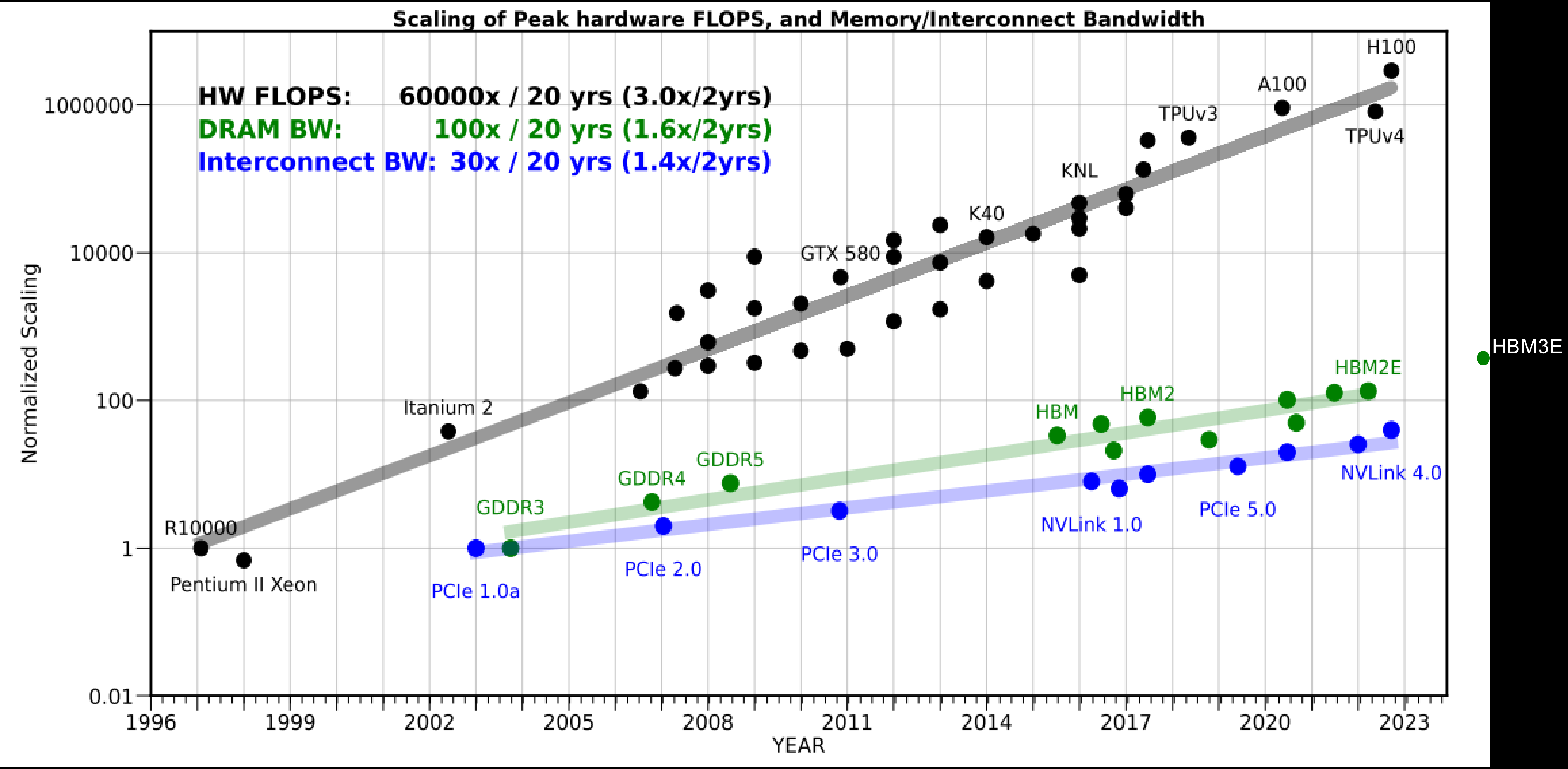
Recent models (LLM,VLM) use more matrix x vector computations

Smaller data-types enable:

- Reduced memory bandwidth, foot-print and reduce latency to load data
- Increased tile sizes in caches → bigger, more efficient matrix multiplication
- More parallelization in hardware

	M=C=	N=	$BW_m : BW_c$	Example applications
matrix*vector	1024	1	1:1	FC layers, MLP
matrix*matrix	1024	32	3 : 100	LLM batched prompt processing
matrix*matrix	1024	64	3 : 200	CNN, limited batch processing
matrix*matrix	1024	1024	1 : 500	CNN, large batch processing

AI Memory Wall, Compute Bandwidth versus Memory Bandwidth



Challenges in Datatype Selections

- Quantization-Aware Training or Post-Training Quantization:
  – Mapping (continuous) values to a limited set of discrete values
  – While minimizing overall accuracy loss
- Inference:
  – In data-centers: option to batch multiple tasks
  – At the edge: limited batching opportunities, limited memory footprint
- Complexity of hardware and software implementation
  – Floating-point more complex than integer & fractional
  – Data compression and decompression

The Quantization Challenge

- Convert continuous infinite input values from a large set to discrete finite output values in a smaller set
- Reduce the precision of calculations to enhance efficiency
- While minimizing impact on training and inference accuracy
- Considerations:
  – What is the quantization range? How to deal with outliers?
  – Uniform vs non-uniform quantization intervals (codebook)
  – Mixed quantization for layers, channels, blocks, weight, activation…
- Perplexity:
  – A measure of how well a language model predicts a sample
  – Higher perplexity → more uncertainty

Micro-Scaling Accuracy: Discriminative Inference

Task	Family	Model	Dataset	Metric	Baseline FP32	MXINT8	MXFP8		MXFP6		MXFP4
							E4M3	E5M2	E2M3	E3M2	
Language Translation	Transformers (Enc-Dec)	Transformer-Base [9]	WMT-17	BLEU Score ↑	26.85	26.64	26.27	25.75	26.38	25.97	22.68
		Transformer-Large [9]			27.63	27.56	27.44	27.02	27.52	27.22	26.33
	LSTM	GNMT [10]	WMT-16		24.44	24.52	24.53	24.45	24.51	24.44	23.75
Language Encoding	Transformers (Enc-Only)	BERT-Base [11]	Wikipedia	F-1 Score ↑	88.63	88.58	88.47	87.04	88.38	88.05	84.94
		BERT-Large [11]			93.47	93.41	93.42	93.32	93.45	93.27	90.97
Image Classification	Vision Transformer	DeiT-Tiny [12]	ImageNet ILSVRC12	Top-1 Acc. ↑	72.16	72.20	71.37	70.11	71.56	70.16	56.72
		DeiT-Small [12]			80.54	80.56	79.83	79.00	80.11	79.04	71.35
	CNN	ResNet-18 [13]			70.79	70.80	69.08	66.16	69.71	66.10	48.77
		ResNet-50 [13]			77.40	77.27	75.94	73.78	76.42	73.75	42.39
		MobileNet v2 [14]			72.14	71.61	65.74	53.50	67.76	53.46	0.25
Speech Recognition	Transformer	Wav2Vec 2.0 [15]	LibriSpeech	WER ↓	18.90	18.83	23.71	21.99	20.63	21.98	42.62
Recommendations	MLPs	DLRM [16]	Criteo Terabyte	AUC ↑	0.803	0.803	0.802	0.801	0.802	0.801	0.7947

Conclusion:

- For direct-cast inference, MXINT8 is effective as replacement for FP32
- Formats with more mantissa bits are more accurate
- Not shown: After fine-tuning, MXFP6_E2M3 accuracy is close to FP32

Wrap Up

- The **Era of Pervasive Intelligence** is upon us
- Synopsys.ai delivers to SoC developers the power of Generative AI across the full EDA stack
- Synopsys NPX6 NPU IP & tools were designed with the future of AI in mind
  - Support for the latest Transformers was key design driver
  - Transformers now being applied to many deep learning domains (e.g., vision, NLP) with surprisingly little modifications
- Generative AI builds on Transformers, moving quickly to the edge Data type selection, compression and quantization is an important aspect of AI
- Smaller data types enable:
  - External memory bandwidth, footprint, power, latency and system cost reduction
  - Cache spilling reduction, bigger tiles for matrix multiplication
  - Hardware parallelization for higher performance
- Future will see more hardware and software optimizations to support

AI,
The Only Way Forward

Thank you