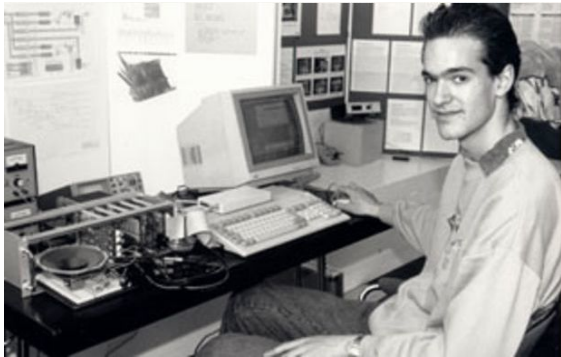


Next 10x in AI Systems, Silicon, Algorithms & Data

October 16, 2024

Erik Norden





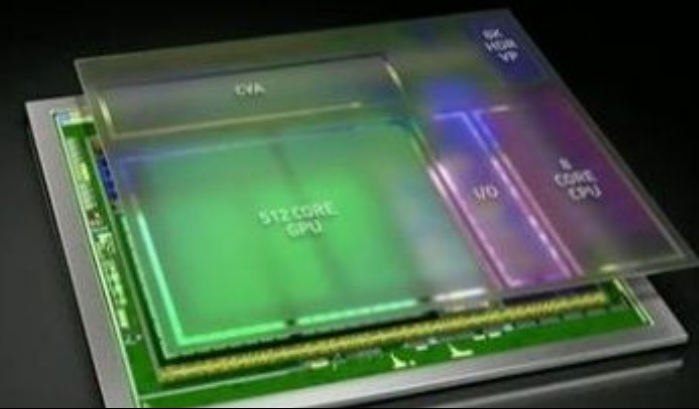
31 patents
2 Hot Chips
presentations

Intel AI: Scale-out,
Retrieval-augmented
generative models,
Reasoning

Bosch
Autonomous
Driving
Research

INTRODUCING XAVIER

AI SUPERCOMPUTER SOC



Kicked off Computer Vision Acceleration at
NVIDIA



Invented
Apple
Neural Engine

64-bit Fusion

architecture
Performance cores process complex tasks faster than ever, while custom efficiency cores handle everyday tasks — helping to deliver a huge leap in battery life.

Neural Engine

for advanced machine learning

The 8-core, Apple-designed Neural Engine is up to 20% faster and uses up to 15% less power. It's a driving force behind the triple-camera system, Face ID, AR apps, and more.

Fastest CPU

in a smartphone

The CPU's two performance cores are up to 20% faster and use up to 30% less power. And its four efficiency cores are up to 20% faster and use up to 40% less power.

Machine Learning

accelerators

Two new machine learning accelerators on the CPU run matrix math computations up to six times faster, allowing the CPU to perform over one trillion operations per second.

Fastest GPU

in a smartphone

The Apple-designed GPU is 20% faster and uses 40% less power. Perfect high-performance gaming and the latest AR experiences.

Core ML 3

for machine learning in

To help developers leverage machine learning power, the A13 Bionic, Core ML 3 works with the Machine Learning Controller to automatically direct tasks to the CPU, Neural Engine,

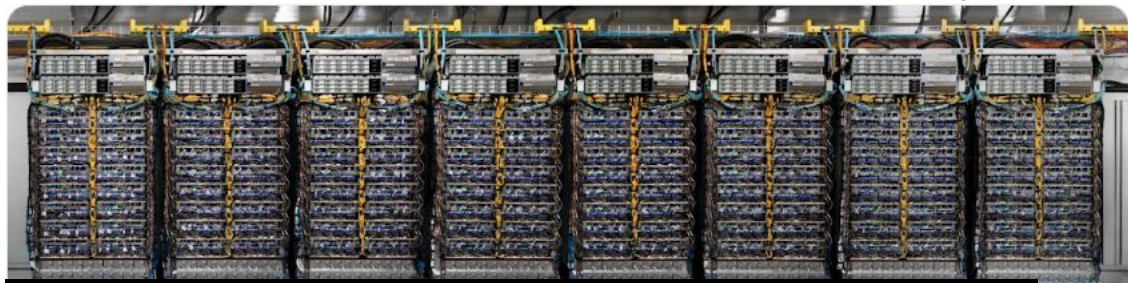
GenAI Startup CTO

Robotics Startup Advisor

Infineon TriCore 2 microarch lead

Two THREADS FOR TriCORE 2

Infineon Reveals TriCore 2 Multithreading Extensions



Google TPU Architecture, Advanced RAG LLM Serving

MICROPROCESSOR
REPORT
www.MPRonline.com

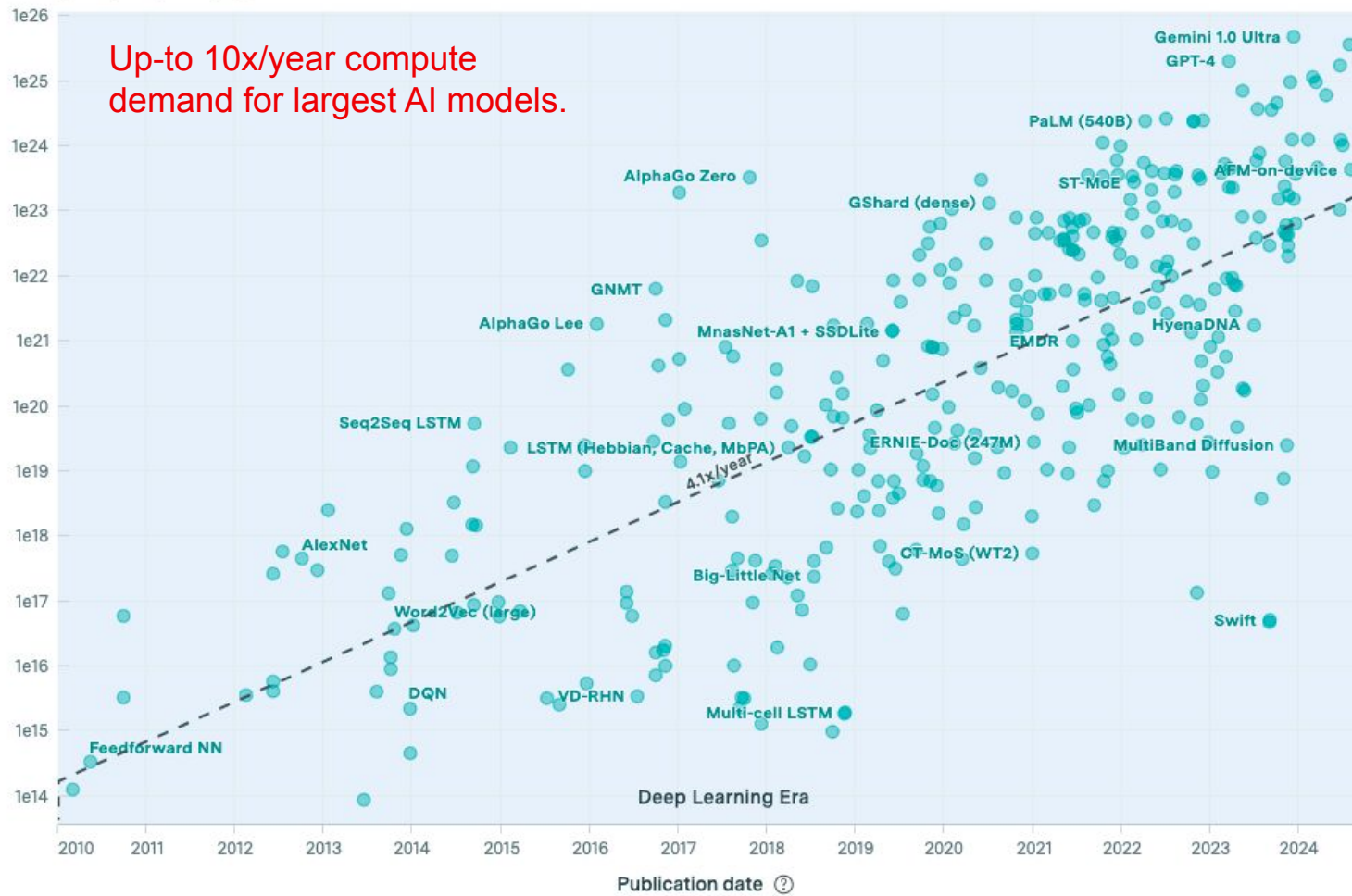
THE INSIDER'S GUIDE TO MICROPROCESSOR HARDWARE

Notable AI Models

Training compute (FLOP) ?

388 estimates out of 827 models

Up-to 10x/year compute demand for largest AI models.



Performance and Efficiency Matters

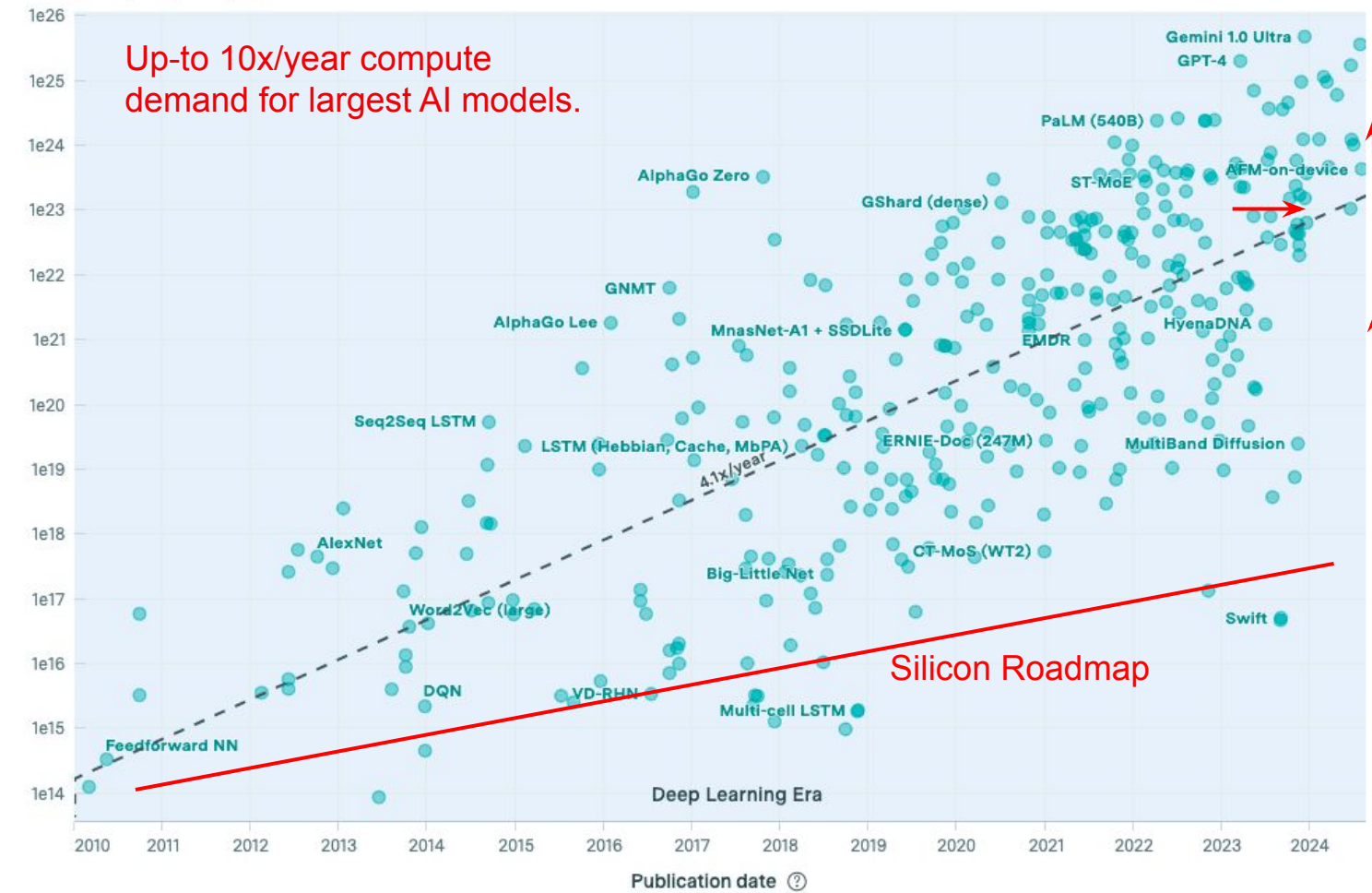
Source: Epoch AI

Z Y P H R A

Notable AI Models

Training compute (FLOP) ?

388 estimates out of 827 models



Performance and Efficiency Matters

Algorithmic Advancements

Scale Out \$\$\$

Packaging

Source: Epoch AI

Z Y P H R A

AI Model Evolution

- Multilayer Perceptrons (pre-2012)
- CNNs (AlexNet 2012) and LSTMs (2014)
- Transformers (2017)
 - Self-Attention Mechanism
 - Parallelizable (faster), Scalable (-> LLMs)

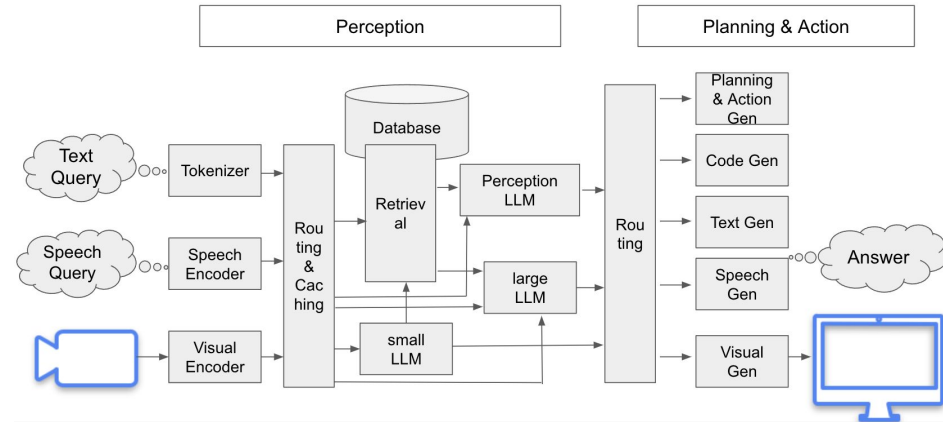
- **Beyond Transformers e.g.**

- Hybrid SSM Attention Architecture

- **Compound AI Systems**

- Combination of AI models with different functionality and sizes.
- Caching, routing.
- Retrieval from vector databases, knowledge graphs. Web search.
- Combined with SW functions. APIs.
- Can be multimodal: text, images, audio, video, other sensor data.
- Can be used for agentic systems.

Compound AI system example:



AI Silicon Evolution

Disaggregated:

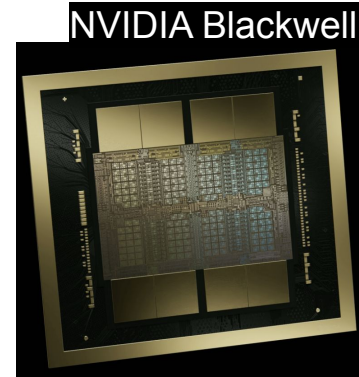
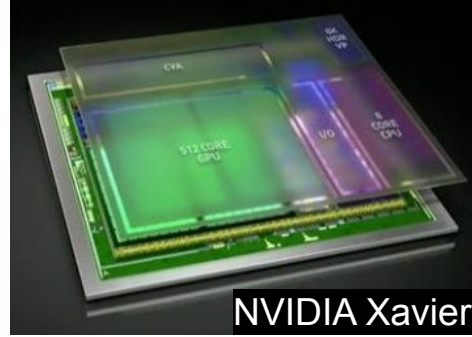
- CPUs (pre-2012)
- GPUs (AlexNet, NVIDIA, 2012), GPUs with AI HW (TensorCore, 2017)
- TPU (Google, 2016), FPGA
- Variants: Cerebras, Tenstorrent, SambaNova, Graphcore, MatX...
Inference only: Groq, d-Matrix, Positron, Etched...

Integrated:

- On-device, AI PC (2017 Apple Neural Engine)
- Edge SOCs (2018 NVIDIA Xavier)

Characteristics e.g.:

- Numerics:
FP32 (AlexNet) -> Float16 or Int8 (Apple Neural Engine, TPUv1, GPU) -> Bfloat16 (Google) -> FP8 -> Custom low precision with block scaling factors.
- Memory technology for efficient inference (SRAM, DDR, HBM, others)
- Chiplets and Packaging



AI Systems - Cloud

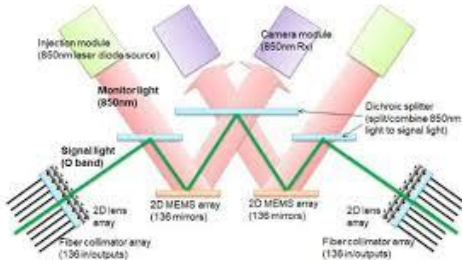
- Originally CPU based (pre-2012).
- GPU cluster, first with central parameter server.
- TPU 3D torus.
- Silicon photonics for cross rack communication.
- OCS - Optical Circuit Switch - for fast reconfiguration.
- Towards liquid cooling.
- Advanced compiler & algorithms for data & tensor parallelism, pipelining, sharding.
- Asynchronous training across clusters.



Google TPU Pod

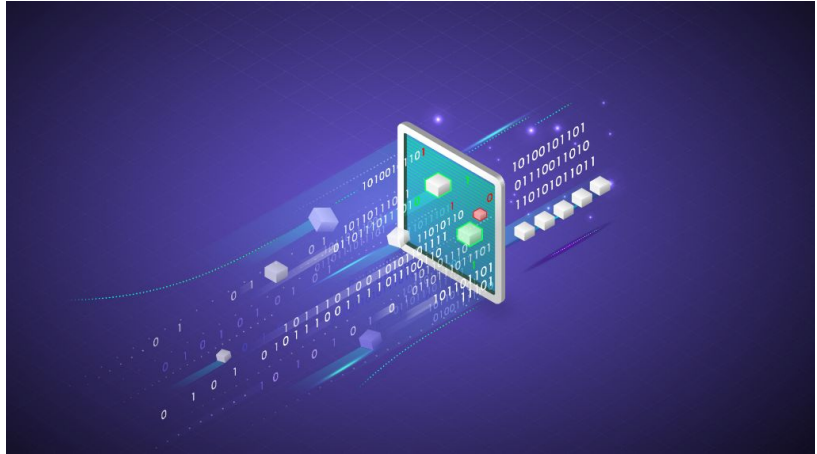


Google OCS



AI Dataset Advancements

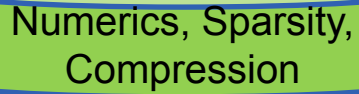
- Small datasets (pre-2012, curated like MNIST, CIFAR)
- ImageNet dataset (2012, large)
- Massive datasets (2018, common crawl, multimodal)
- Current Trend: **Optimized datasets** for pretraining, finetuning, retrieval
 - Data curation tools: deduplication, removal of low quality data, debiasing
 - Synthetic data



AI Key Elements

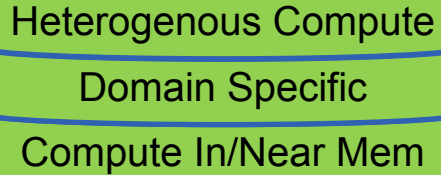


Predict AI Compute Trends, Next Wave



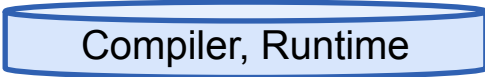
Numerics, Sparsity, Compression

Increase computational efficiency.



Heterogenous Compute
Domain Specific
Compute In/Near Mem

Optimized efficient compute close to the data.



Compiler, Runtime

SW Infrastructure

AI Key Elements

Predict AI Compute Trends, Next Wave

System, Scaleout, Interconnect

Scalable performance.

Numerics, Sparsity, Compression

Increase computational efficiency.

Beyond Dense Matrix Compute

Scalable performance beyond dense matrix WLS.

Heterogenous Compute

Optimized efficient compute close to the data.

Domain Specific

Compute In/Near Mem

Power Mgmt, Cooling

HW Infrastructure

Compiler, Runtime

SW Infrastructure

Security & Privacy

AI Key Elements

Predict AI Compute Trends, Next Wave

Workloads beyond LLMs. Towards higher intelligence and future applications.

System, Scaleout, Interconnect

Scalable performance.

Numerics, Sparsity, Compression

Increase computational efficiency.

Beyond Dense Matrix Compute

Scalable performance beyond dense matrix WLS.

Heterogenous Compute

Optimized efficient compute close to the data.

Domain Specific

Compute In/Near Mem

Silicon Photonics, Memory Technology, Process, Chiplets, Packaging

New technologies and optimizations.

Power Mgmt, Cooling

HW Infrastructure

Dataset Quality

Data

Beyond transformers
Algorithm codesign like
Flash Attention

Algorithms

Compound Systems:
Multimodal, RAG etc.

Compiler, Runtime

SW Infrastructure

Security & Privacy

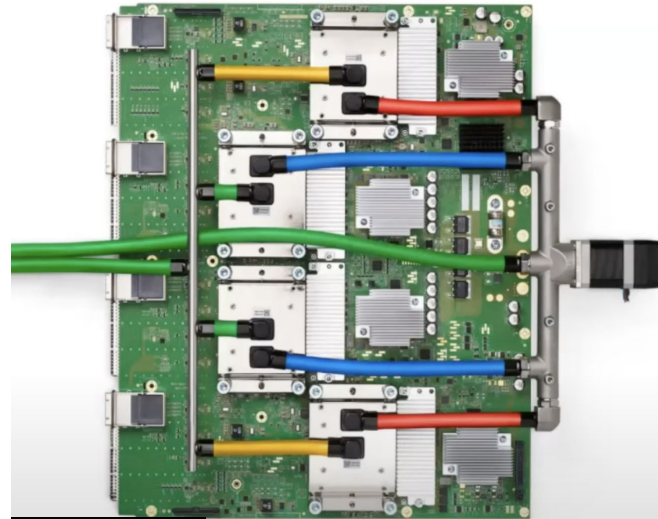
RAS

Google TPUv4

- 4096 TPU chips per Pod
- 64 TPU racks, deployed 8x8
- Total compute >1ExaFLOP BF16



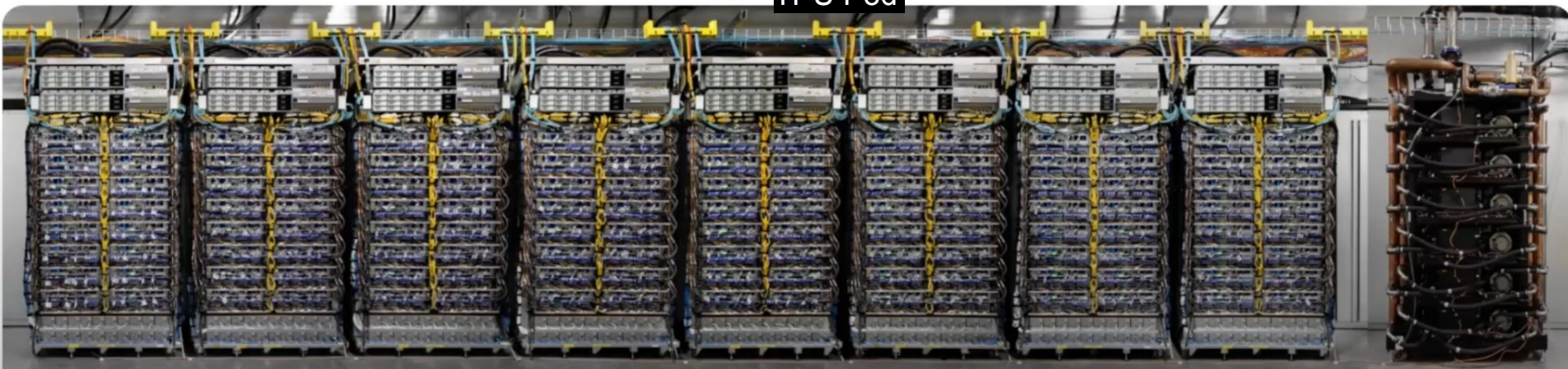
TPU



TPU Board

Source: Google

TPU Pod

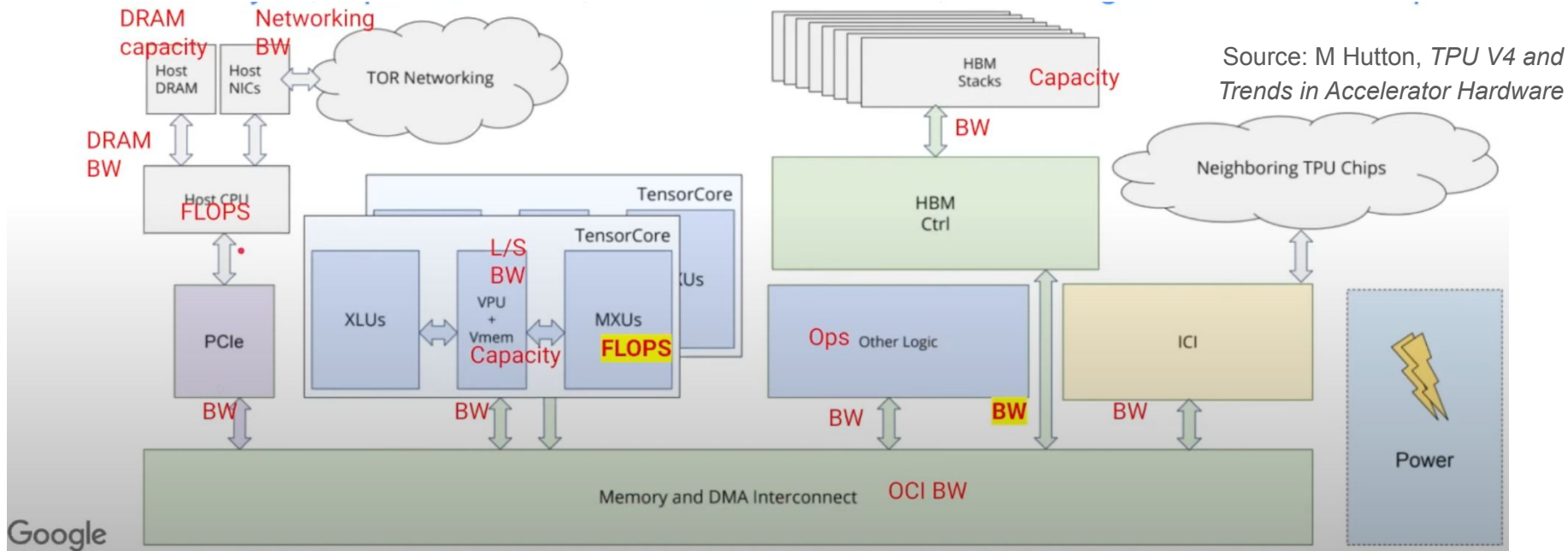


Sizing TPU Chip and System

Predict future workload mix. Use case and workload analysis.

Size BW paths, memory capacity, TPU FLOPS, host compute, various features.

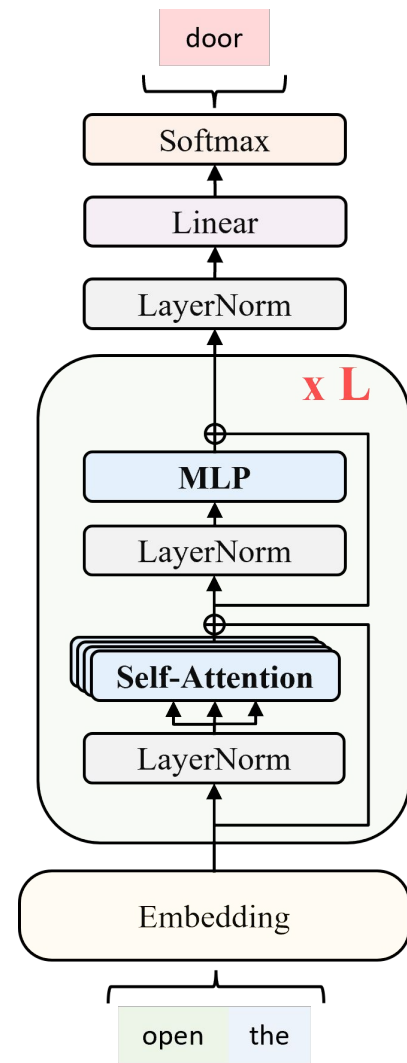
Optimizing Performance/TCO. TCO is Total Cost of Ownership.



Transformer

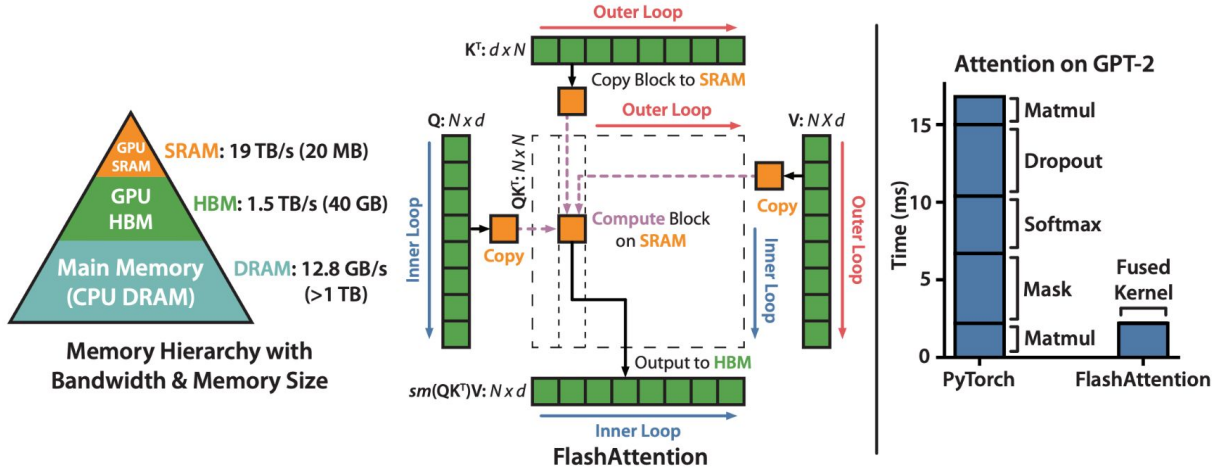
Transformers are currently the main building block in many areas such as text generation, summarization, vision, speech recognition, ...

- Next token prediction (decoder)
- Primarily composed of alternating multi-head self-attention and multilayer perceptron (MLP) blocks.
- Self-attention helps the model understand the relationships between different tokens in a sentence
- Linear and Softmax layers: translate the internal representation into the next token.



Flash Attention

A tiling and fusion optimization of the attention algorithm to minimize HBM bandwidth by avoiding materialization of intermediate results to HBM. There is no loss in quality. Flash attention helps balance HBM bandwidth and compute. This specifically helps with large input sequence lengths.



Left: FlashAttention uses tiling to prevent materialization of the large $N \times N$ attention matrix (dotted box) on (relatively) slow GPU HBM. In the outer loop (red arrows), FlashAttention loops through blocks of the K and V matrices and loads them to fast on-chip SRAM. In each block, FlashAttention loops over blocks of Q matrix (blue arrows), loading them to SRAM, and writing the output of the attention computation back to HBM. Right: Speedup over the PyTorch implementation of attention on GPT-2. FlashAttention does not read and write the large $N \times N$ attention matrix to HBM, resulting in a 7.6 \times speedup on the attention computation. Source: Flash Attention paper

Zyphra Tree Attention

Topology-Aware Decoding for Long-Context Attention on GPU Clusters

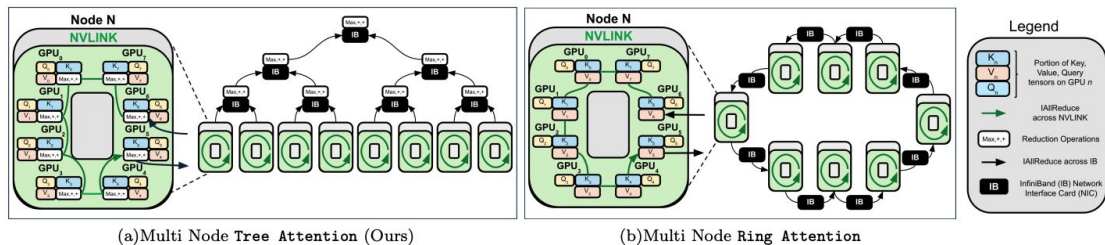
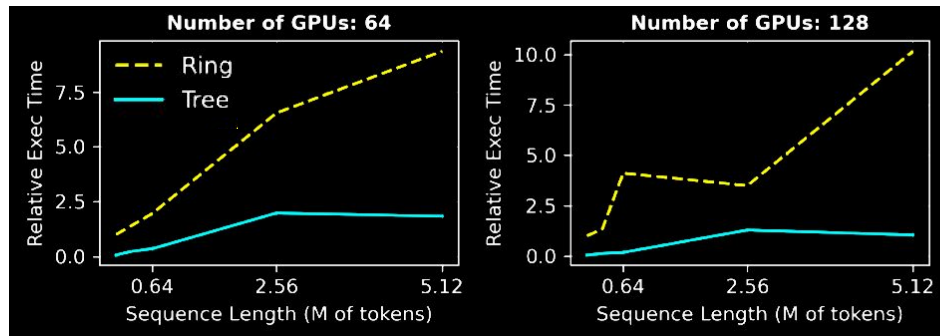
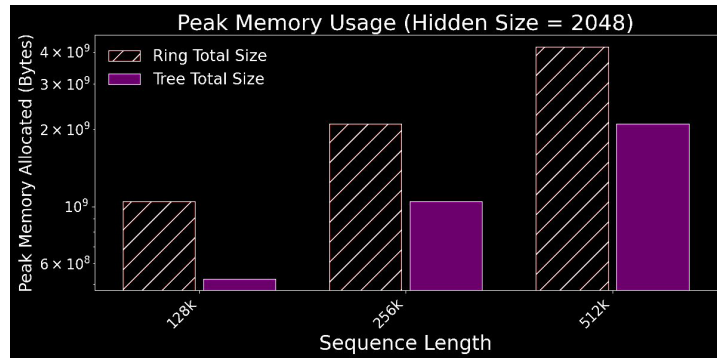


FIG. 1: Ring and Tree Attention Topologies. Due to the associative properties of the logsumexp and max operations of Tree Attention (Fig. 1(a)), it is possible to structure the reduction across the sequence as a tree, requiring asymptotically fewer communication steps than Ring Attention (Fig. 1(b)) as well as less memory and communications volume.

Lower Latency

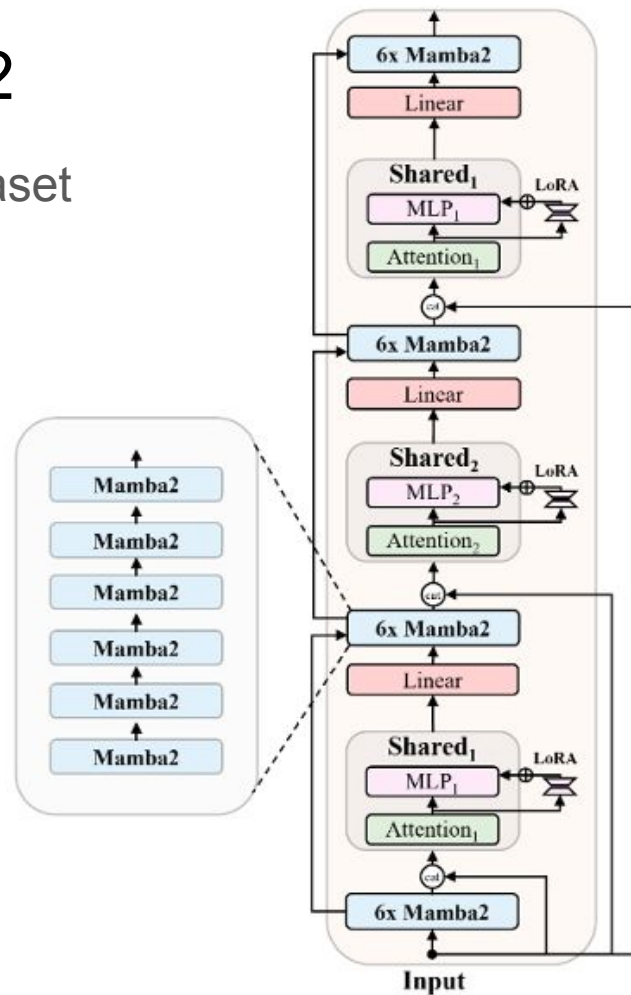
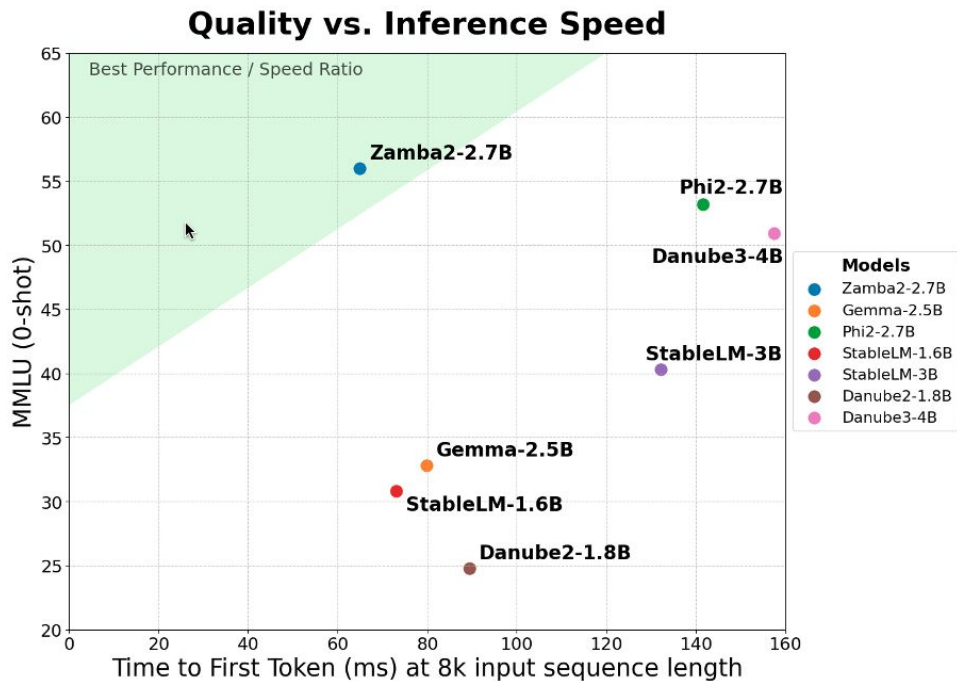


Lower Peak Memory



Beyond Transformers: Zephyr Zamba2

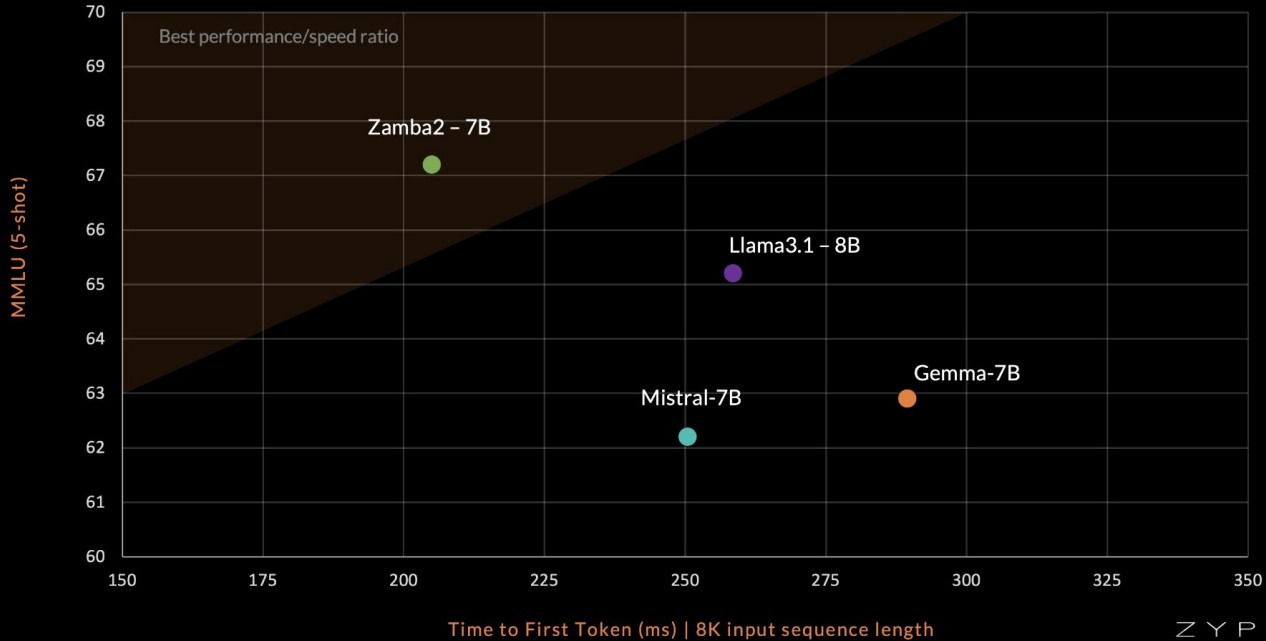
Hybrid SSM Attention Model, trained with custom dataset



Beyond Transformers: Zyphra Zamba2-7B

*just
released!*

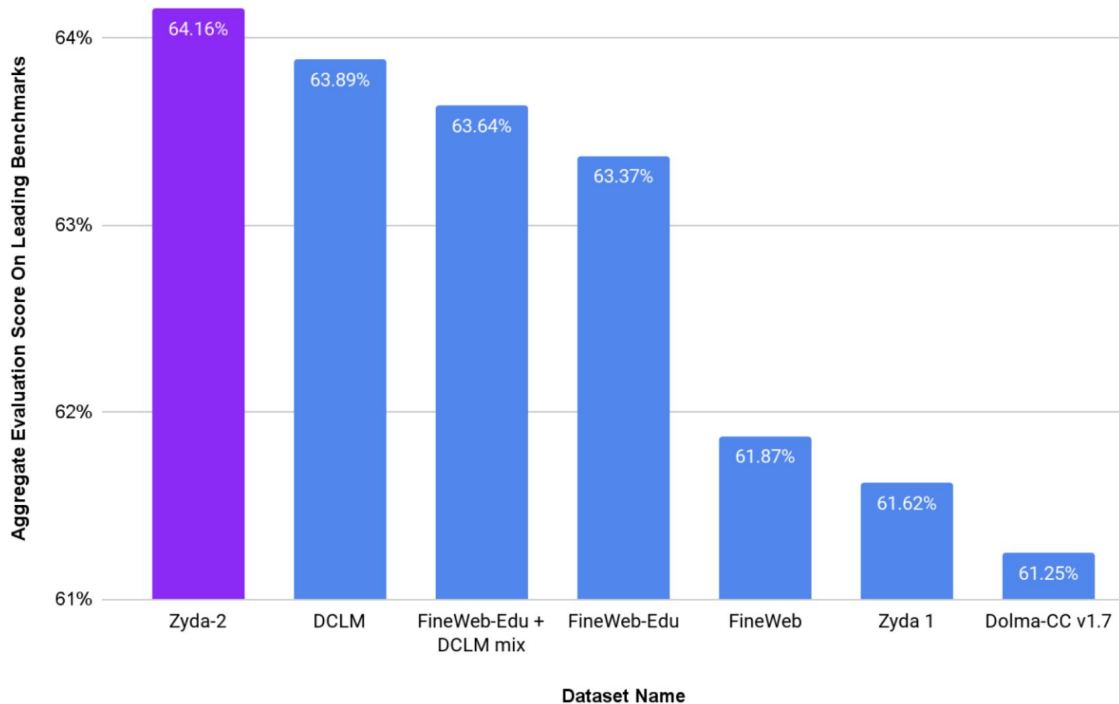
ZAMBA2 7B - QUALITY VS INFERENCE SPEED



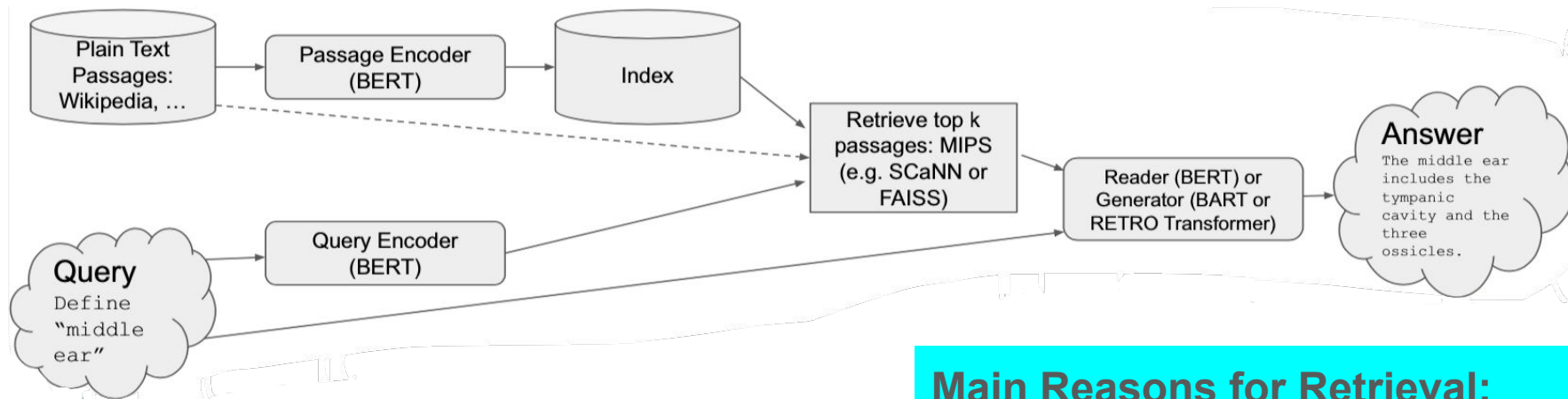
Zyphra Zyda-2 Dataset

*just
released!*

- Open 5T-Token Dataset Processed with NVIDIA NeMo Curator.
- Outperforms existing state-of-the-art open-source language modeling datasets in aggregate evaluation scores.
- Study based on Zamba2-2.7B.
- Aggregate score is a mean of MMLU, Hellaswag, Piqa, Winogrande, Arc-Easy, and Arc-Challenge.



Retrieval-Augmented Transformers: RAG



Augmenting an LLM with a large text database from which additional tokens can be retrieved via nearest neighbor search to enhance the next token.

Information is split into deep learning parameters and the database (e.g. plain text or knowledge graph) to create a more effective and efficient system.

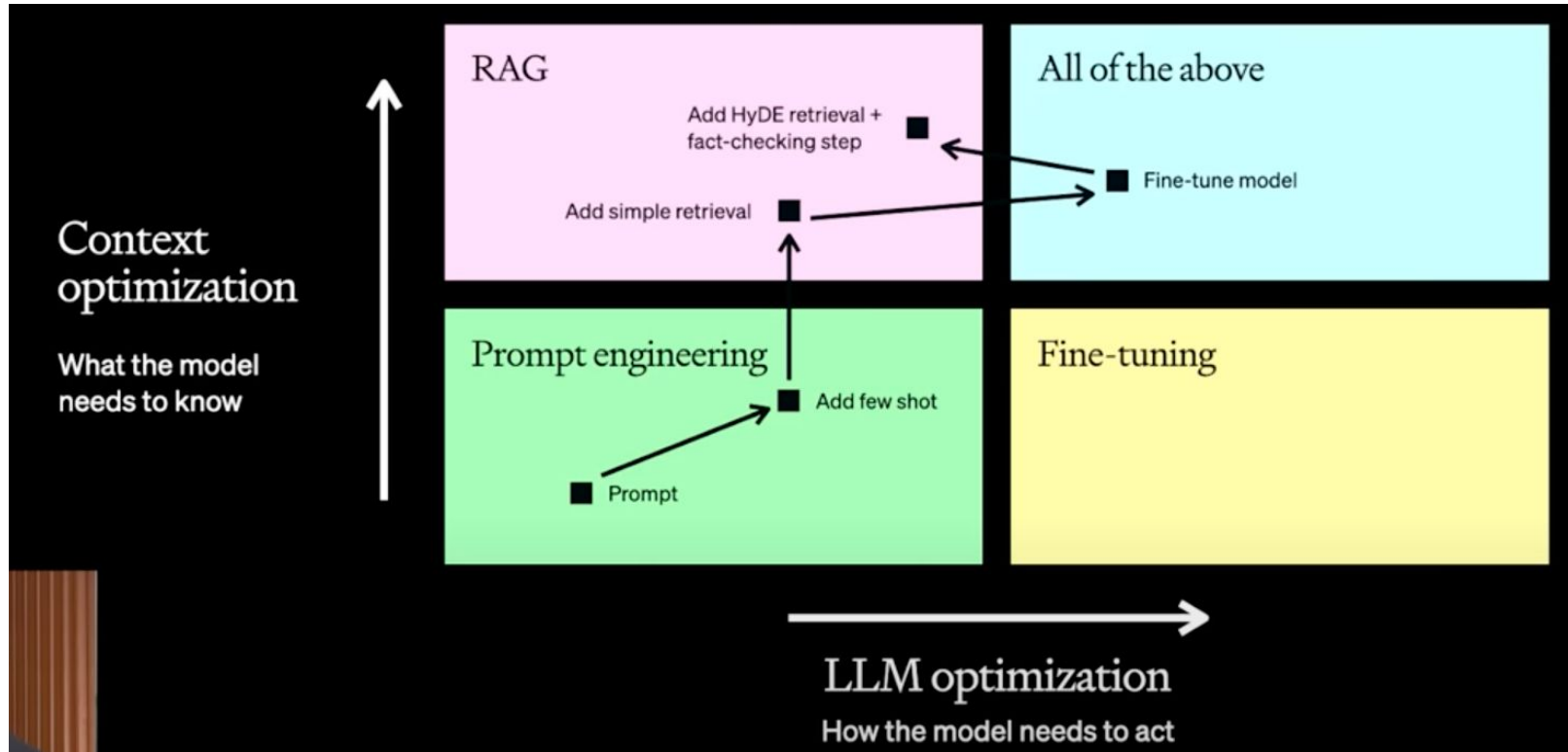
Main Reasons for Retrieval:

Add Custom Data to LLM: extensible without fine-tuning.

Grounding: improves factuality for generative models.

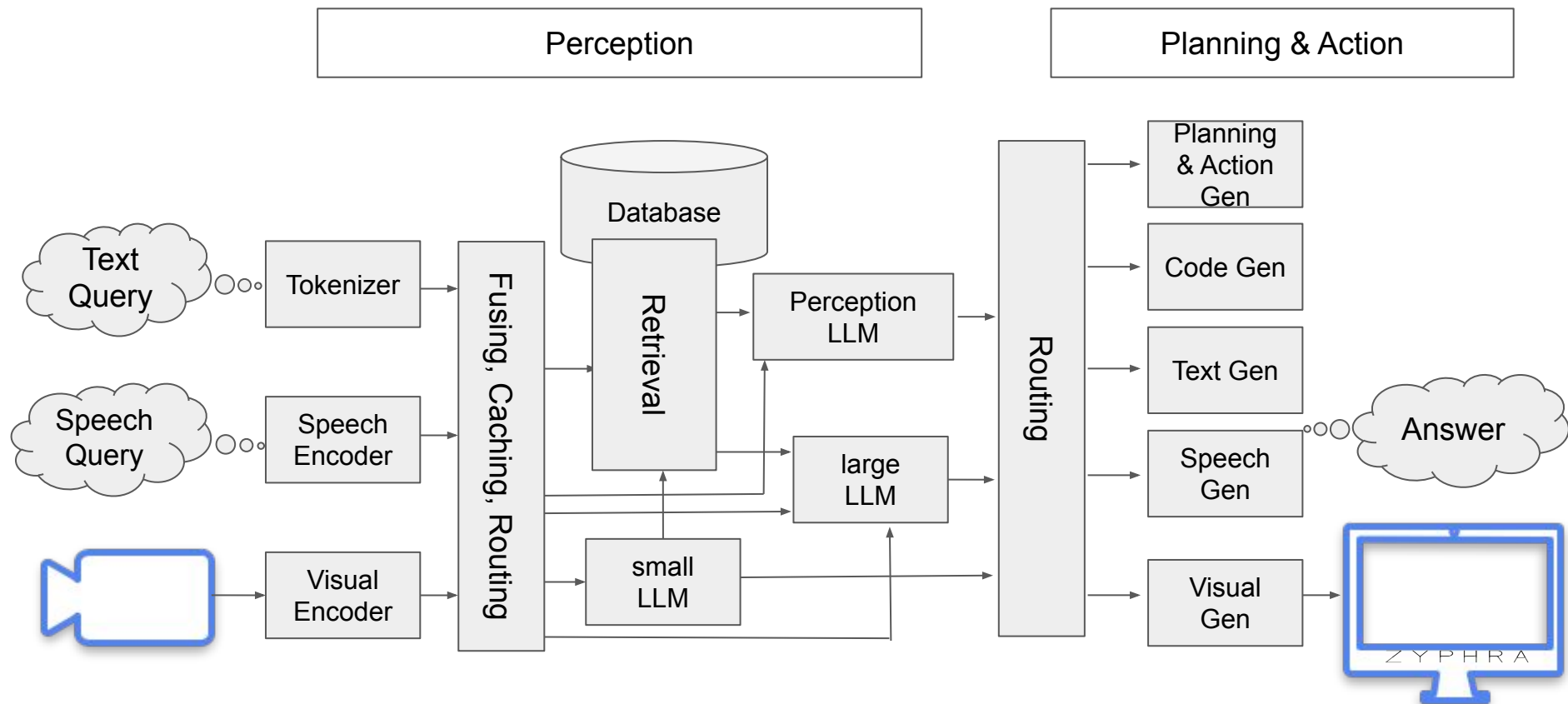
Enhanced Explainability: provides sources for the answer.

LLM Optimization Flow

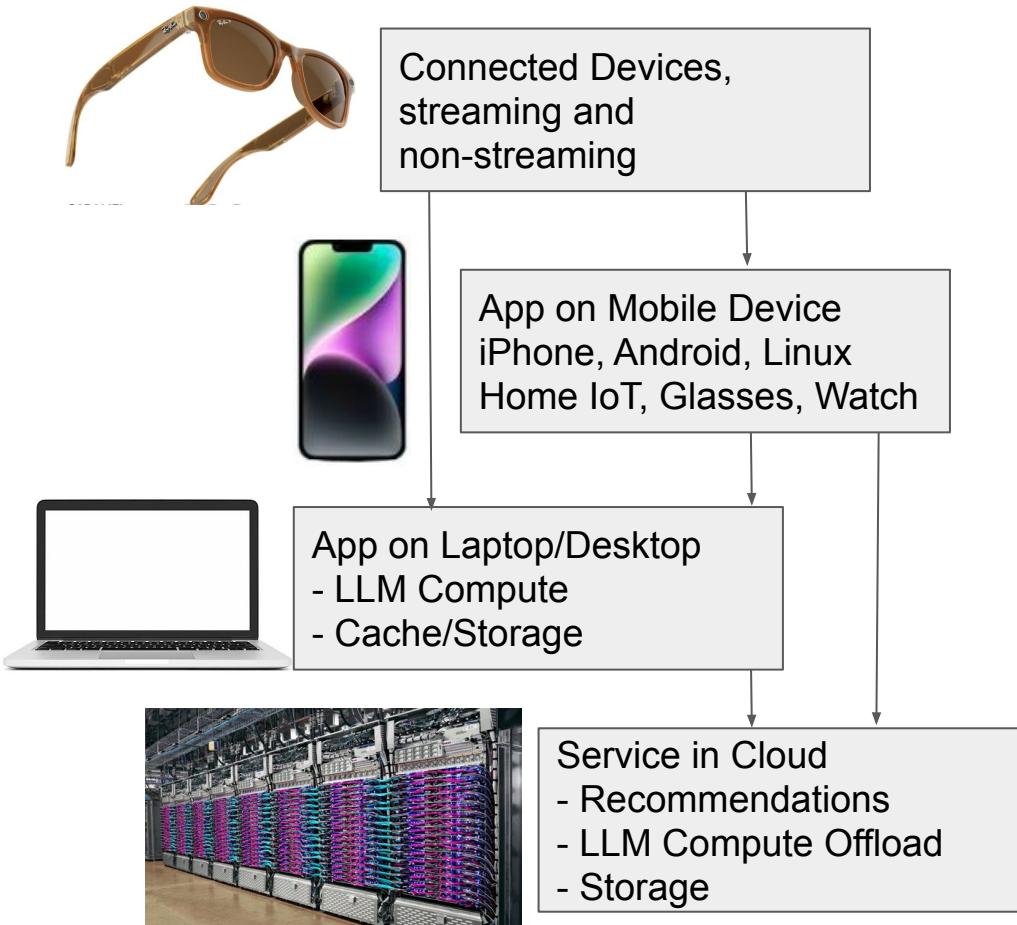


Source: OpenAI, *A Survey of Techniques for Maximizing LLM Performance*

Compound Model



Platform Across Devices



Streaming: e.g. video, audio
Non-streaming: e.g. photos, sensor
data like temperature, position etc.
Compression

Multimodal User Interface
Personalization
Compression
Basic Voice/Audio/small LLMs

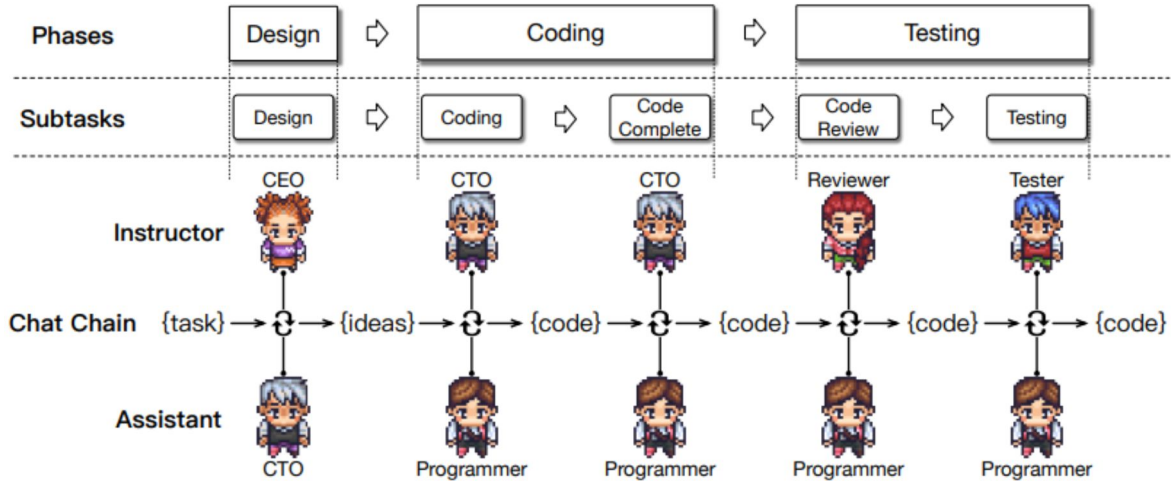
Small and larger GenAI models
Personalization, Agents
Storage
Optional Hub Functionality

Recommendations (Ads etc.)
For subscribers only:

- Small to largest GenAI models
- Extended Storage, Backup

Multimodal Agentic Systems

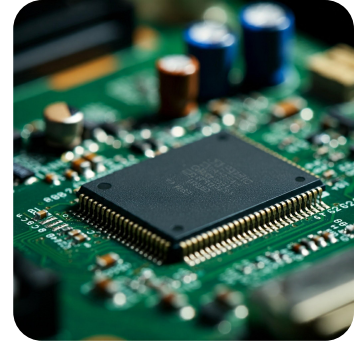
- Planning:
 - Chain-of-thought, tree-of-thought
 - ReACT, self-reflection
- Memory
 - In-context, vector database, knowledge graph, external files
- API access to tools
 - Calculator, task specific tools



Example: chain of agents for SW development

Source: C. Qian et al, *ChatDev: Communicative Agents for Software Development*

Generative AI for Chip Design



- Increasing efforts in academia and industry
- LLM models and agents to significantly improve chip design productivity by providing design assistance as chatbots and copilots and automating more manual design tasks
 - Engineering chat bot to answer questions on how to do certain tasks incl. specific command line generation
 - Assistance on resolving bug reports, assistance on PPA improvements
 - Copilot for design space exploration, hardware code generation, documentation generation
- More domain specific datasets and benchmarks needed
 - Specifications, code, databases, PPA metrics
 - So that LLM models and agents can be trained and optimized.

Conclusions

- Goals: Advance generative AI serving in cloud, edge, on-device. Towards AGI. Advance LLM training in the cloud.
 - Model: Beyond transformers e.g. **Zamba2** hybrid SSM attention model
 - Data: **Advanced datasets** for pretraining, fine-tuning, retrieval
 - Advanced Co-design - e.g. **Tree-attention**
 - **Compound systems** – multi-modal, Graph-RAG, multi-hop reasoning
 - Applications - towards **multimodal agentic systems**
- Innovation across domains is critical to get to the next 10x in generative AI.



ZYPHRA

info@zyphra.com