MLSys

The Seventh Annual Conference
on Machine Learning and Systems

AI and ML at Capital One

Leveraging standardized cloud platforms for data management, model development, and operationalization, we use AI and ML to look out for our customers' financial well-being, help them become more financially empowered, and better manage their spending.

LEARN MORE

AI RESEARCH PRIORITIES

Anomaly Detection	Natural Language Processing	Behavior Models	Deep Learning for Event Prediction
Foundation Models	Privacy & Accessibility	Graph Networks	Large Language Models

MLSys

The Seventh Annual Conference
on Machine Learning and Systems

MatX: high throughput chips for LLMs

Reiner Pope, cofounder and CEO

MatX focuses on:

maximizing **performance/$**

on **large models**

MatX *does not* focus on:

small models

small deployments

This saves a lot of silicon!

ML/HW codesign

ML research is critical to the company:

- numerics
- memory bandwidth
- …

Great research needs a great codebase:

- matx.com/seqax

matx.com/jobs

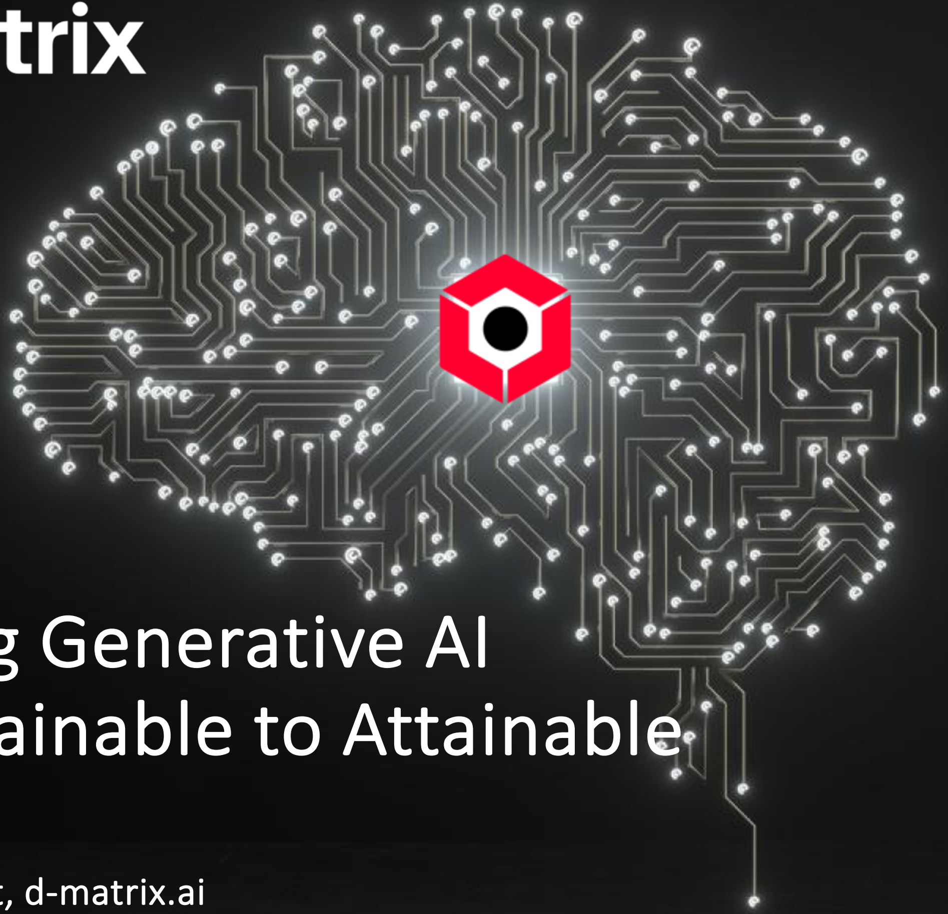
Meet the team **Tuesday, May 14th from 5-8pm** at the Hyatt for drinks and food!

Visit matx.com/meetmatx to register and get more information

MLSys

The Seventh Annual Conference
on Machine Learning and Systems

d-Matrix



Transforming Generative AI from Unsustainable to Attainable

Sree Ganesan
Vice President of Product, d-matrix.ai

Unique Challenges of Generative Inference

Models are large (billions of parameters) and context lengths are growing

→ Requires *more* memory capacity and *more* compute capacity

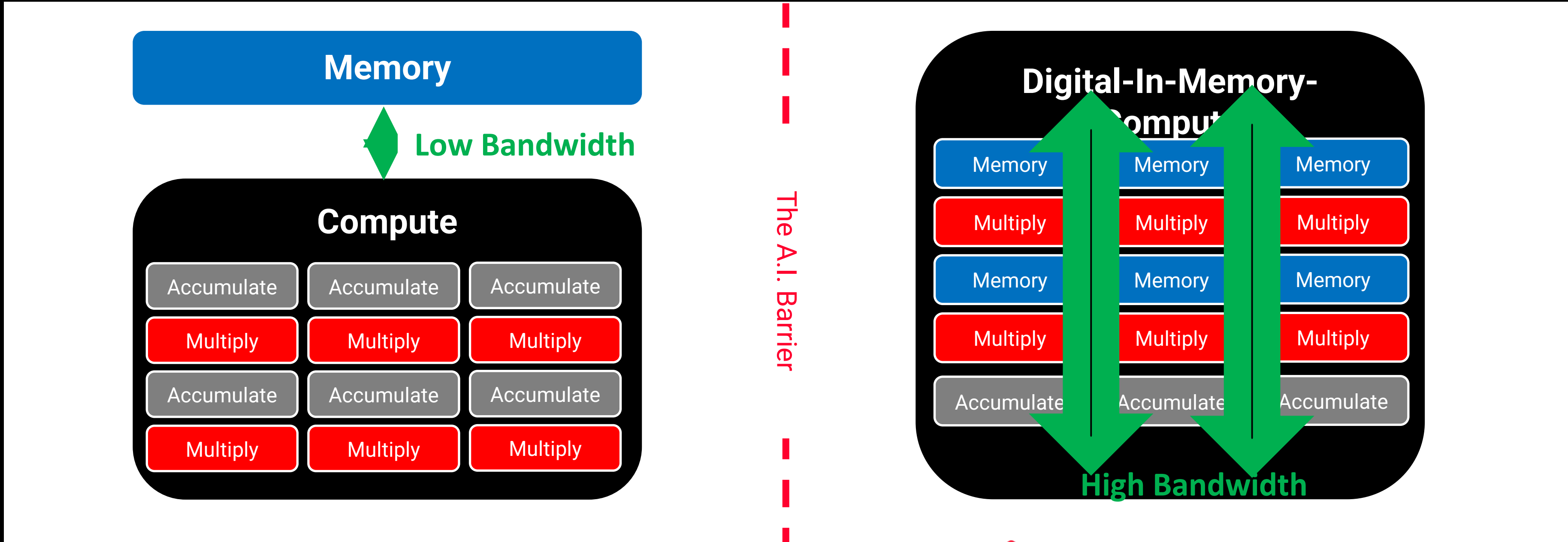
Prompt processing is compute bound & token generation is memory bound

→ Requires *high* memory bandwidth and *high* peak compute capability

These exacerbate the pain points of cost, power and performance

The d-matrix inference solution is built from the ground-up to accelerate generative inference

A New Computing Paradigm is Needed



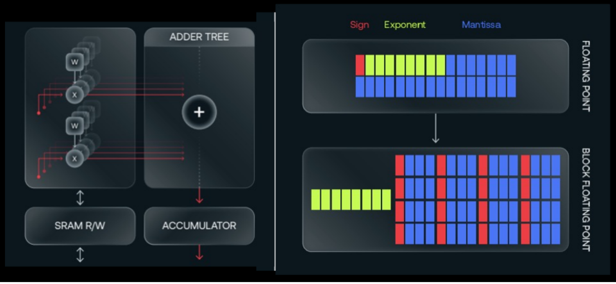
Traditional architecture

d-Matrix architecture

The d-Matrix Advantage

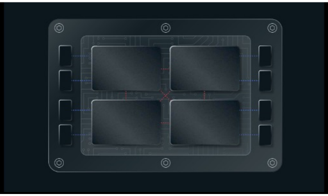
Circuits & Numerics

- Digital In-Memory Compute (DIMC)
- Block Float Sparsity
- Compression



Chiplets & Advanced Packaging

- 2D, 3D Stacking
- Logic, Memory Co-package



Software

- Easy to Use
- Performant, Scalable

Transforming Generative AI from Unsustainable to Attainable

Performant

Cost efficient

Power efficient

Interactivity (tokens/s/user)

Perf-TCO ($ per token)

MLSys

The Seventh Annual Conference
on Machine Learning and Systems

SambaNova Systems Overview

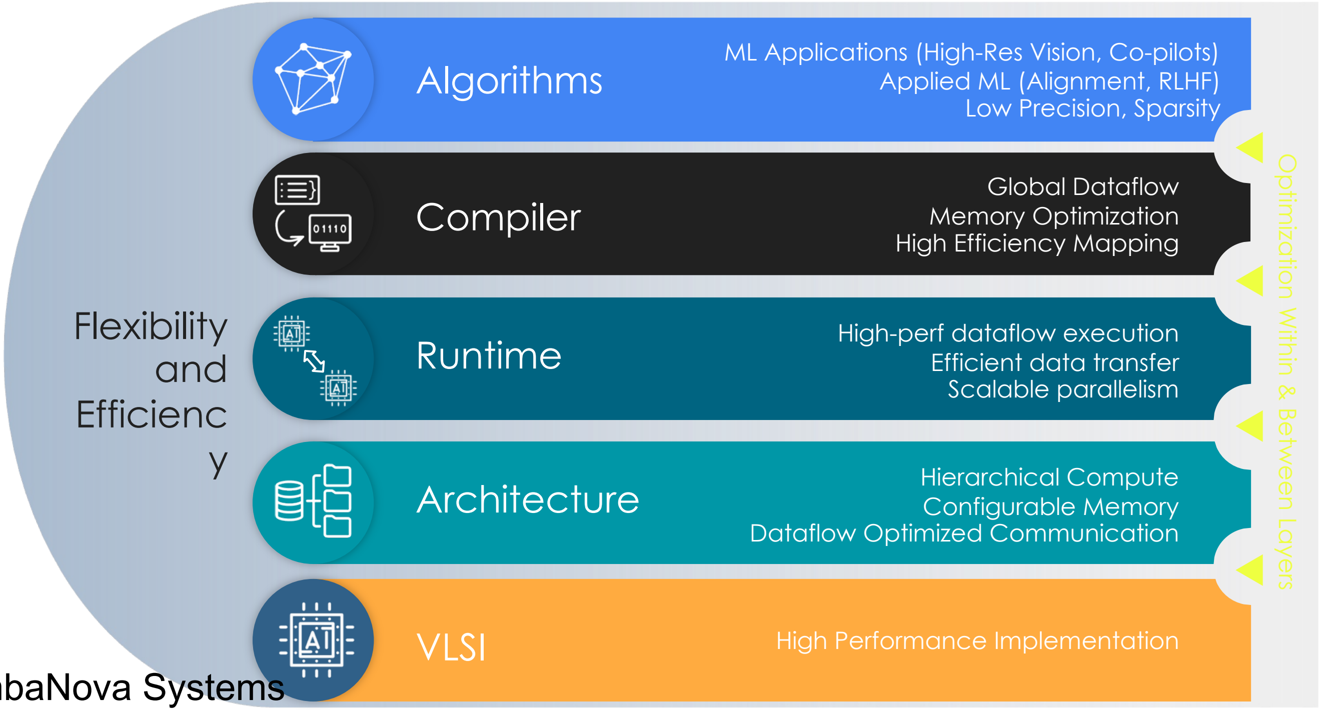
Full stack co-engineering yields optimizations where best delivered with the highest impact

Kunle Olukotun
Professor EE/CS
Stanford
University

Chris Ré
Professor CS
Stanford University

Rodrigo Liang
CEO
ex-SVP Oracle SPARC CPU

Reconfigurable Dataflow Architecture (RDA)



Resources

Developer Website and Past Publications

VentureBeat

SambaNova announces new AI Samba-CoE v0.2 that already beats Databricks DBRX

SiliconANGLE

SambaNova debuts 'composition of experts' AI model with 1T+ parameters

MarkTechPost

This AI Paper from SambaNova Presents a Machine Learning Method to Adapt Pretrained LLMs to New Languages

The rapid advancement of large language models has ushered in a new era of natural language processing capabilities.

SambaNova

https://sambanova.ai › blog › using-mixed-precision-on-...

Using Mixed Precision on RDUs

Job Opportunities

- Computer Vision
- Large Language Models (LLMs)
- Multimodality
- Compiler
- System Software
- Computer Architecture
- Physical Design / VLSI

MLSys

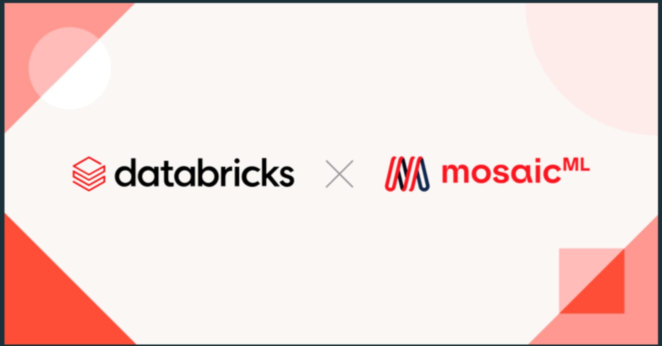
The Seventh Annual Conference
on Machine Learning and Systems

databricks

Databricks Mosaic Research: MLSys 2024

Abhi Venigalla, NLP Architect

Databricks Mosaic Research



- Help orgs build+serve custom AI models …
- … Using their own unique data …
- … As efficiently as possible.

- What do we need? Reusable tools, infrastructure, recipes.
  - DBRX: https://huggingface.co/databricks/dbrx-instruct
  - Composer: https://github.com/mosaicml/composer
  - Megablocks: https://github.com/databricks/megablocks
  - StreamingDataset: https://github.com/mosaicml/streaming
  - Lilac: https://www.lilacml.com/

- Why open-source software/models? Feedback, testing, trust, ownership.

Research -> Production

Idea #1: Hardware is changing rapidly, check assumptions!

- What networking bandwidths / block sizes are available in the cloud **today**?
  - H100: 3200 Gbps/Node, in blocks of 2k+ GPUs

- How much memory, memBW is there?
  - 8xH100: 640 GB HBM
  - 8xB100, 8xMI300X: 1.5 TB HBM
  - 72xB200 (NVL72): 13.5 TB HBM, + another 17 TB LPDDR5X

- What data types are being hardware accelerated?
  - Today: BF16, FP8
  - Soon: BF16, FP8, **MXFP4**

- If you see an excess of a quantity (FLOPs/BW/compute/memory), can you modify the workload to take advantage of that excess to deliver better quality/latency/something else?

Research –> Production

Idea #2: Work together with scaling laws

- Show that an idea **scales well at small budgets** (0.01 –> 0.1 –> 1)
- To convince folks that it will work at larger budgets (1 –> 10 –> 100)
- Might require new metrics – continuous rather than discrete scores

- Also: good tools/models/simulators are useful at every scale
  - Personal request: Please build an **LLM online inference simulator!!**

- E.g. many-shot ICL, why didn't we catch this sooner? 1 –> 10 –> 100 –> (today) 1000 shot
  - In-Context Learning with Long-Context Models: An In-Depth Exploration: https://arxiv.org/pdf/2405.00200

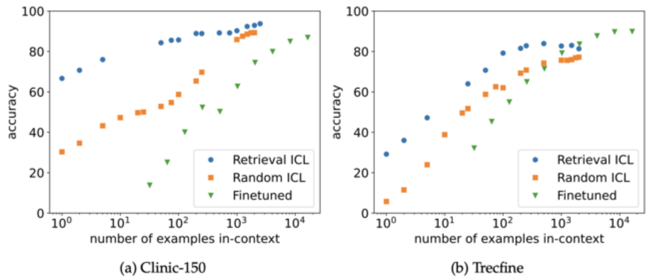


Figure 2: Comparing retrieval ICL, random selection ICL, and finetuning on two representative datasets. Finetuning sometimes, but not always, exceeds ICL at high numbers of demonstrations. Note that, while retrieval ICL uses the listed number of examples in context, it assumes access to the larger test set to draw examples from (Perez et al., 2021). Results on other datasets are in Appendix C.

MLSys

The Seventh Annual Conference
on Machine Learning and Systems

THE
FASTEST
CLOUD FOR
GEN AI

BUILT ON
LEADING AI
RESEARCH

together.ai

WE BELIEVE THE FUTURE OF AI IS OPEN SOURCE

01 TRANSPARENCY

Inspect how models are trained and what data is used to increase accuracy and minimize risk

02 CONTROL

Customize models with proprietary data to build applications that serve your needs

03 PRIVACY

Maintain complete data privacy by storing data locally or in our secure cloud

Innovations

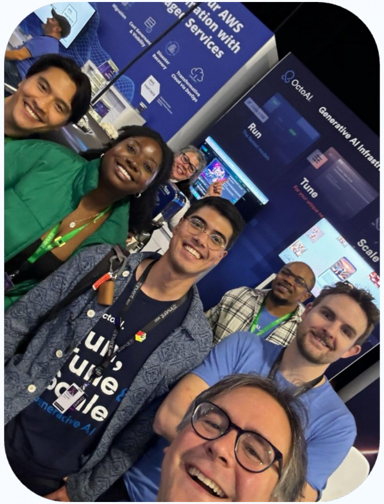
- FLASH ATTENTION 2
- COCKTAIL SGD
- SUB-QUADRATIC ARCHITECTURES
- REDPAJAMA OPEN DATA & MODELS

Products

- TOGETHER INFERENCE
- TOGETHER FINE-TUNING
- TOGETHER GPU CLUSTERS
- TOGETHER CUSTOM MODELS

MLSys

The Seventh Annual Conference on Machine Learning and Systems



OctoAI's mission is to enable customers to benefit from the latest AI innovations by offering efficient, customizable, and reliable AI systems.

Founded 2019 in Seattle, WA
UW-CSE Spin-off

100 employees, 50% in Seattle, rest across the globe

Built on deep expertise in AI systems, with foundational open source traction (Apache TVM, MLC-LLM, etc)

$132M seed/A/B/C from Madrona, Amplify, Addition, Qualcomm and Tiger Global

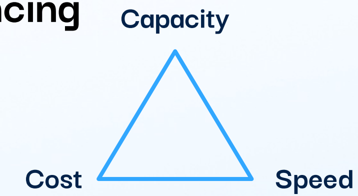
OctoAI Stack: Integrated and Composable

OctoAI.

turnkey genAI inference

model ensembles

balancing



multi-tenant and private environments

APIs usable by general developer

Solutions: Media Gen | Text Gen | Compute Service | Trust and Safety

Model Portfolio: Text | Image | Video | Embeddings | Custom Models

AI Systems Optimized Cross-Stack for Inference

- Model optimization & compilation
- Custom serving layer (batching, multiplexing,..)
- Fine tuning
- Hardware selection
- Config parameter search
- Model asset management (LoRAs, checkpoints, etc)

Serverless experience: Req load balancing & routing | Cross-cloud support

SaaS: Google Cloud | AWS | Serveless

Self-hosted (OctoStack): snowflake | VPC | GPU Cloud | On-Prem | Edge

AI Compute GPUs (NVIDIA & AMD) **Accelerators** (TPUs, Inferentia, Trainium) **Edge** (WebGPU, iOS, Android)

AI Systems Stack for User Outcomes

Our goal: **flexible**, **fast** inference stack with **rapid time-to-market**

Achieved by:

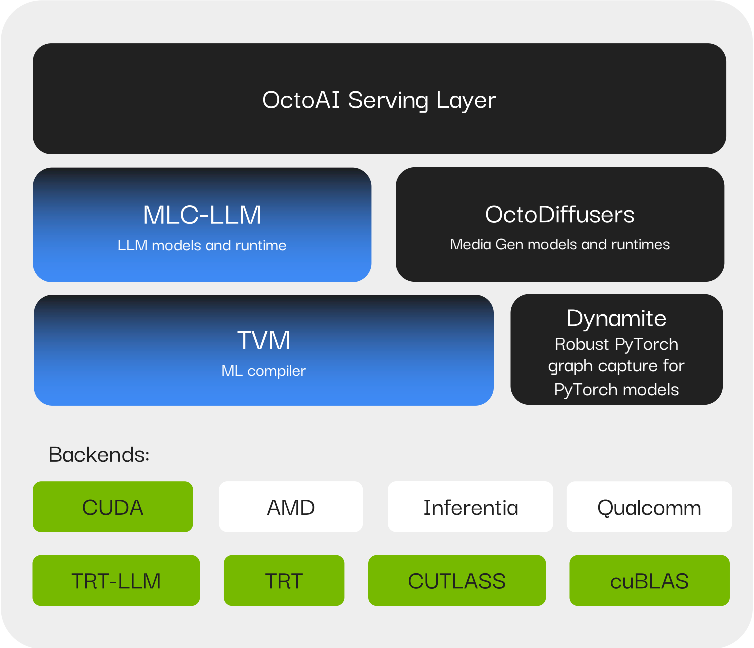
- Leveraging and extending native SW software stack
- Building graph capture, compiler, and runtime systems
- Open source TVM and MLC-LLM w/ community

Example: Enabled high-value use cases on NVIDIA ahead of native support:

- Large-scale LoRA for image gen
- High-performance structured JSON output
- Mixtral and Llama3 on OctoAI

100+ customers in production, 10s of B of tokens/day, millions of images/wk, exciting customer use cases.

OctoAI Optimized Inference Stack



Our stack enables broad hardware target coverage and

Fast time-to-market through:

- Combining existing kernels, libraries, and compilers
- More robust graph capture of PyTorch Ops

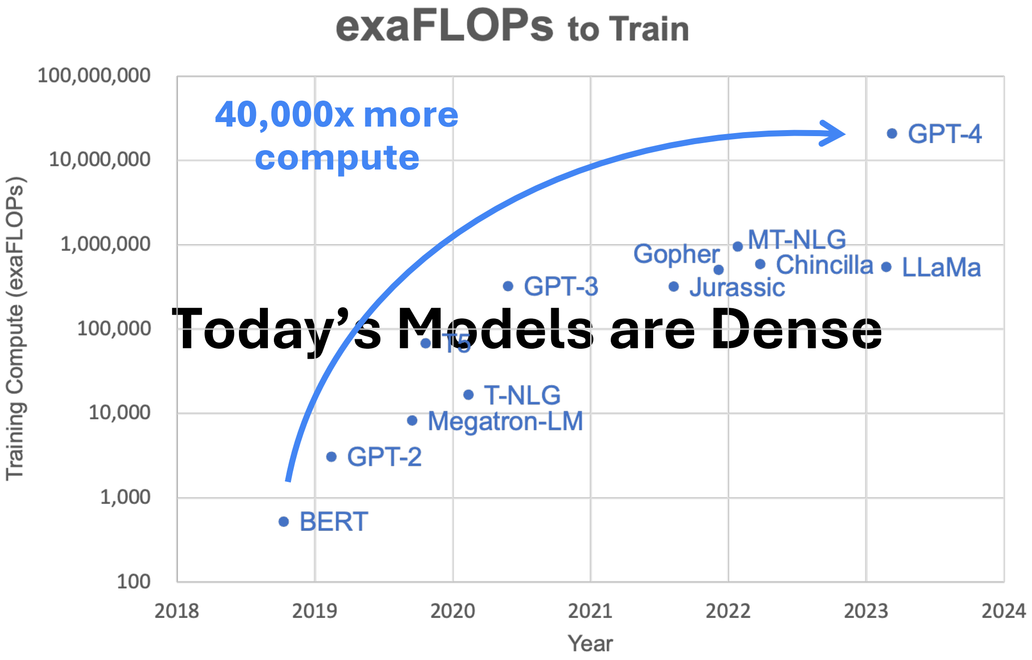
MLSys

The Seventh Annual Conference
on Machine Learning and Systems

Accelerating AI with Wafer-Scale Computing

ML on Cerebras

From multi-lingual LLMs to healthcare chatbots to code models. State-of-the-Art quality

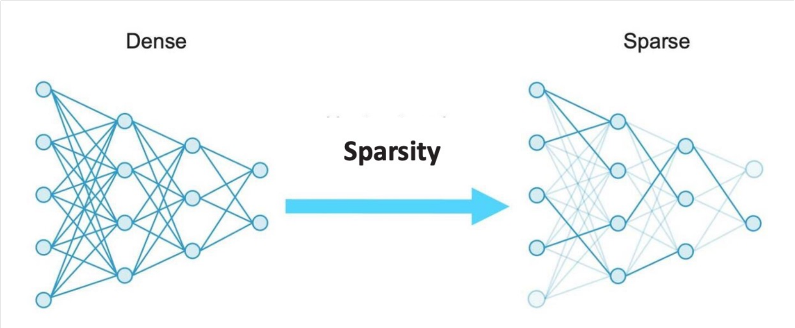


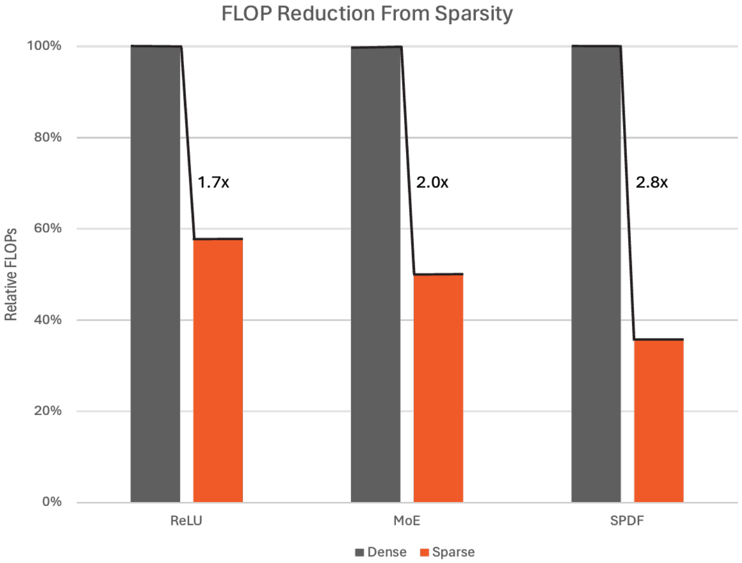
Today's Models are Dense

Scaling today's frontier models is unsustainable !!

WSE: Co-designed with Sparsity for Scale

- **Sparsity opportunities are everywhere**
  - e.g. ReLU, Mixture of Experts, Weight Sparsity
  - Not all HW can take advantage of all forms of sparsity





- **Wafer-Scale Memory Bandwidth built for Sparsity**
  - Low data reuse ⇒ high mem bw
- **WSE accelerates all forms of sparsity**
  - Static and dynamic sparsity
  - Structured and unstructured sparsity
  - Weight and activation sparsity

- **Recent Sparsity Publications**
  - **Sparse-IFT**: Sparse Iso-FLOP Transformations for Maximizing Training Efficiency *(to appear at ICML, 2024)*
  - Enabling **High-Sparsity Foundational Llama** Models with Efficient Pretraining and Deployment *(arXiv, 2024)*
  - **SPDF**: Sparse Pre-training and Dense Fine-tuning for Large Language Models *(UAI, 2023)*

Find out more at our Booth

[1] Li et al., The Lazy Neuron Phenomenon: On Emergence of Activation Sparsity in Transformers, 2023
[2] Jiang et al., Mixtral of Experts, 2024
[3] Thangarasa et al., SPDF: Sparse Pre-training and Dense Fine-tuning for Large Language Models, 2023

MLSys

The Seventh Annual Conference
on Machine Learning and Systems

MLSys

DIAMOND SPONSOR

Meta

PLATINUM SPONSORS

together.ai

Lambda

GOLD SPONSORS

aws

Capital One

d-Matrix

cerebras

snowflake

AMD

SambaNova SYSTEMS

Determined AI

SILVER SPONSORS

MATX

Bai du 百度

OctoAI.

DeepMind

BOOK PUBLISHER

COCALC

now the essence of knowledge