Supercharged AI Inference on Modern CPUs

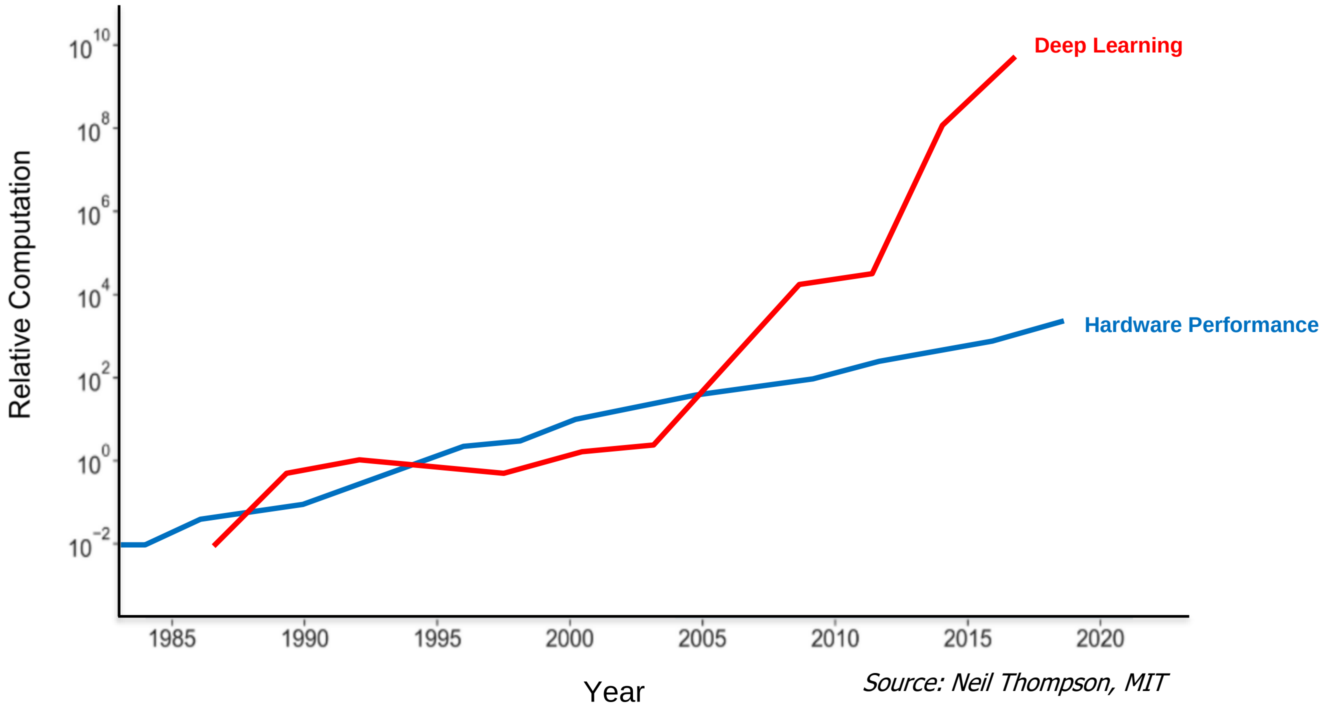
Lawrence Spracklen
Subutai Ahmad

Numenta

HotChips 2023

For Over 30 Years, AI Driven By Brute Force Compute

Computing power demanded by Deep Learning



Source: Neil Thompson, MIT

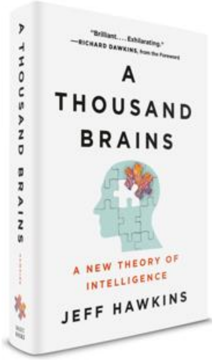
Numenta

Dramatically improve AI using discoveries from neuroscience

- Founded in 2005 by Jeff Hawkins and Donna Dubinsky

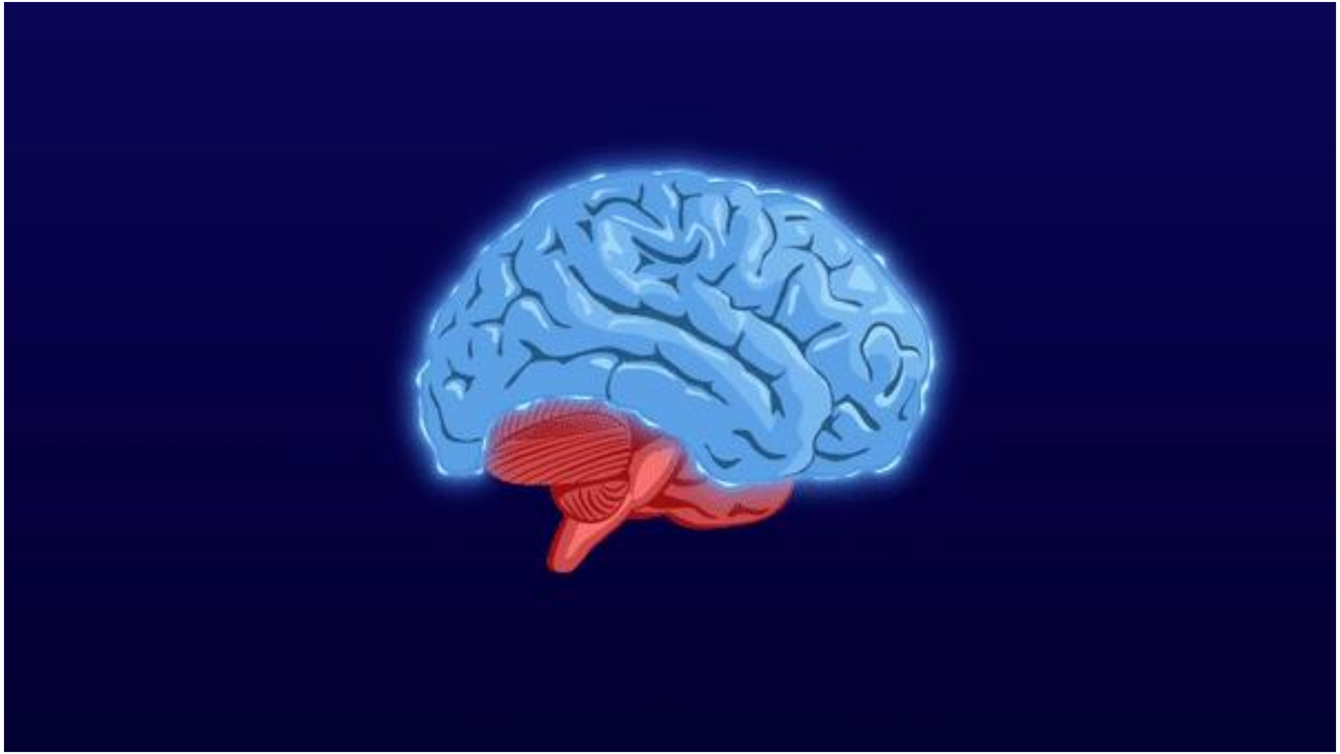
- Mission: reverse engineer the neocortex and apply neocortical principles to AI
  - Two decades of neuroscience research yielded breakthrough AI technology

- Generative AI Platform launch in September
  - 10x - 150x cost/speed improvements across all LLM models
  - Highly scalable deployment of LLMs on CPUs with >10X price/performance
  - Key partnerships with Intel, Oracle, Weights and Biases, and others

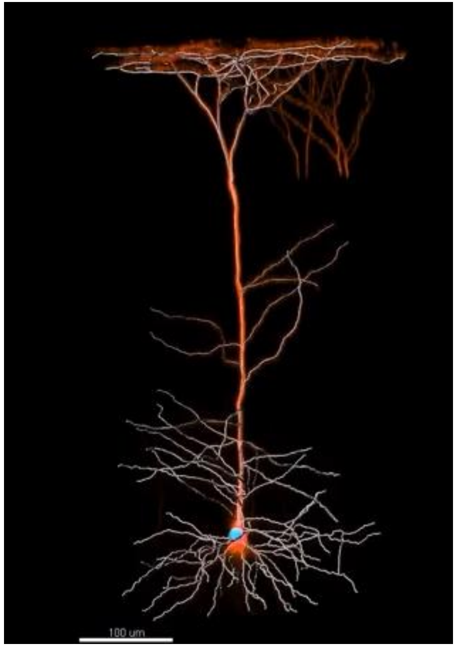

"Brilliant. . . . Exhilarating."
—RICHARD DAWKINS, from the Foreword
A THOUSAND BRAINS
A NEW THEORY OF INTELLIGENCE
JEFF HAWKINS
A THOUSAND BRAINS

*One of Bill Gates'
Top 5 Books of 2021*

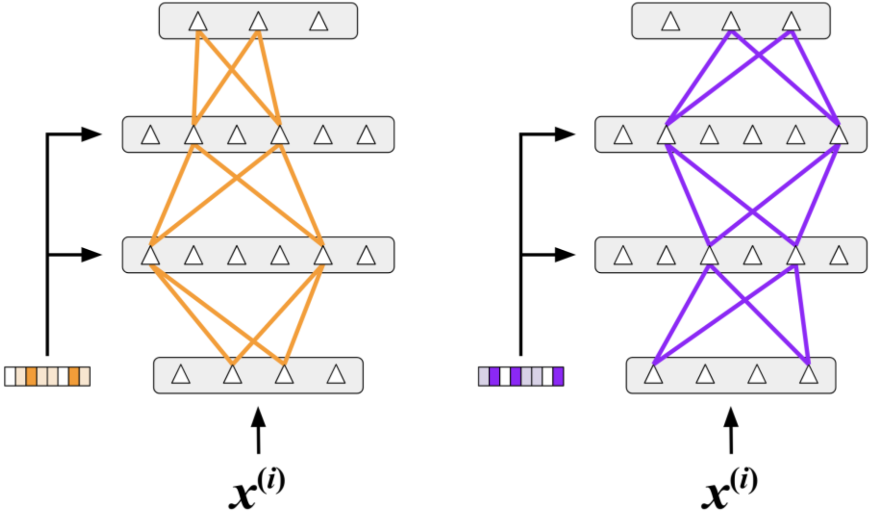
Can Neuroscience Improve AI?



Biological Neurons Are Complex



Pyramidal neuron

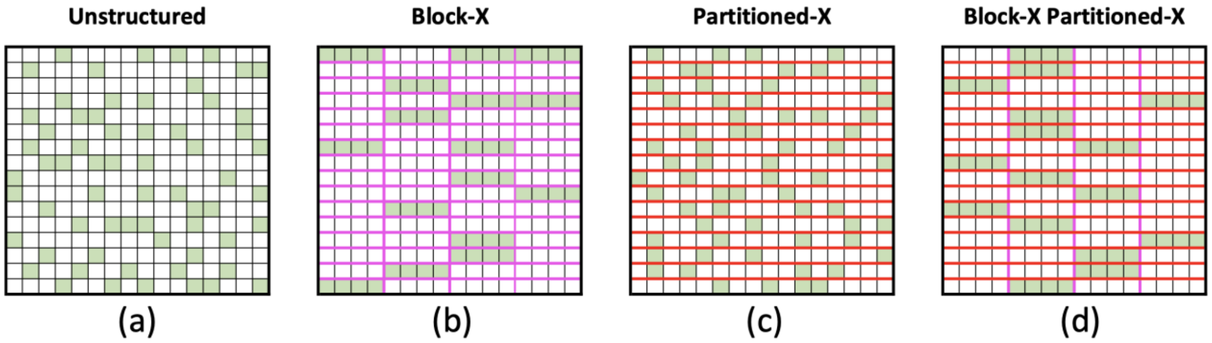


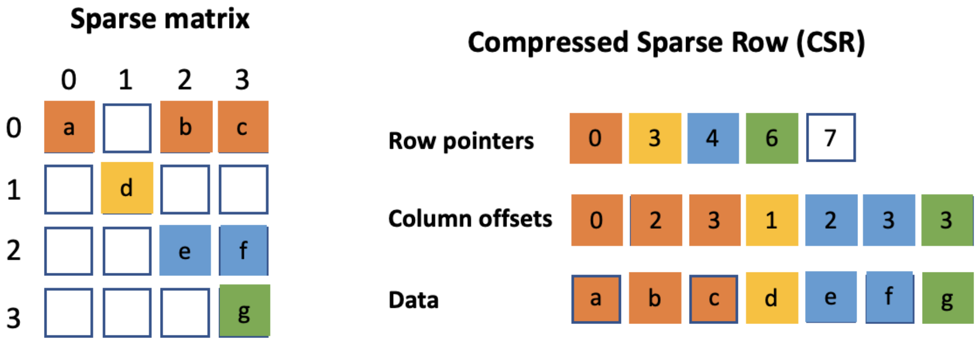
Biological networks are highly sparse and context sensitive

Video: Smirnakis Lab, Baylor College of Medicine

Numenta

Sparsity: Opportunities and Challenges





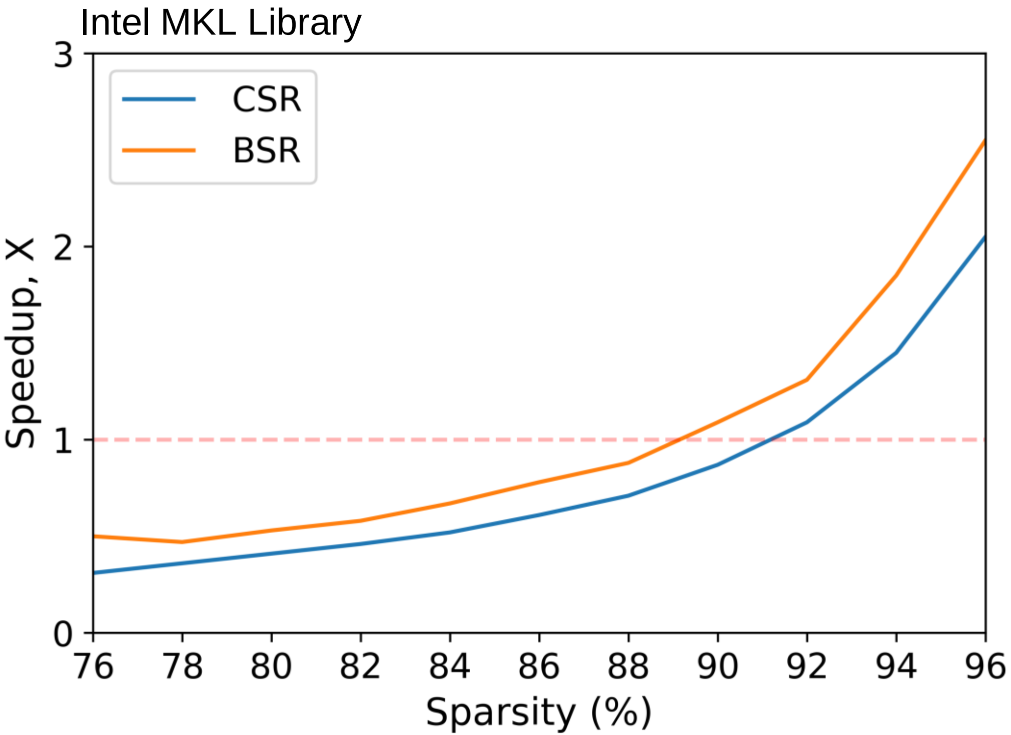
Sparse matrix

	0	1	2	3
0	a		b	c
1		d		
2			e	f
3				g

Compressed Sparse Row (CSR)

Row pointers	0	3	4	6	7		
Column offsets	0	2	3	1	2	3	3
Data	a	b	c	d	e	f	g

Sparsity Today



Problems With GPUs For Inference

- Inflexible programming model
    - Difficult and time consuming to program

- Implementing multi-tenant solutions presents challenges
    - Resource allocation, performance, and scalability concerns

- Co-processor architecture introduces challenges
    - Dual memory architecture leads to slow startup for large models / datasets

- Handling asynchronous requests with low-latency is challenging

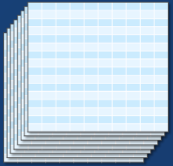
- Mixed CPU+GPU infrastructure challenging for many IT departments

Intel® Advanced Matrix Extensions (AMX) built-in for AI

“Tiles”

2D Register Files

8 tiles
16-rows * 64B

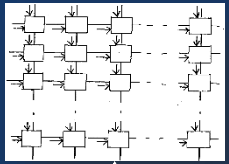


Store bigger chunks of data in each core.

+

“TMUL”

Tile Matrix Multiply



1024 BF16 ops/cycle

Instructions that compute larger matrices in a single operation.

AMX Opportunities

- Significant computational improvements over AVX512
- Significant potential **once tiles have been loaded**
  - 16x32x32 BF16 matrix multiplication in 16-clks
  - 1x32x32 BF16 matrix multiplication in 9-clks
- Critical to hide tile loads to maximize compute

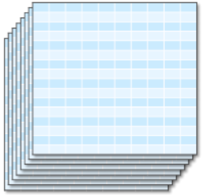
- Possible to use AVX512 in parallel with AMX
  - Conversion of FP32 results back to BF16 for subsequent processing
  - Any necessary data swizzling
  - Other algorithmic requirements (e.g., SoftMax etc.)

- Assumes user wants to perform dense matrix multiplications…..

Per Core
AMX Support

“Tiles”

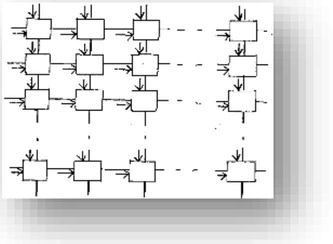
2D Register Files



+

“TMUL”

Tile Matrix Multiply

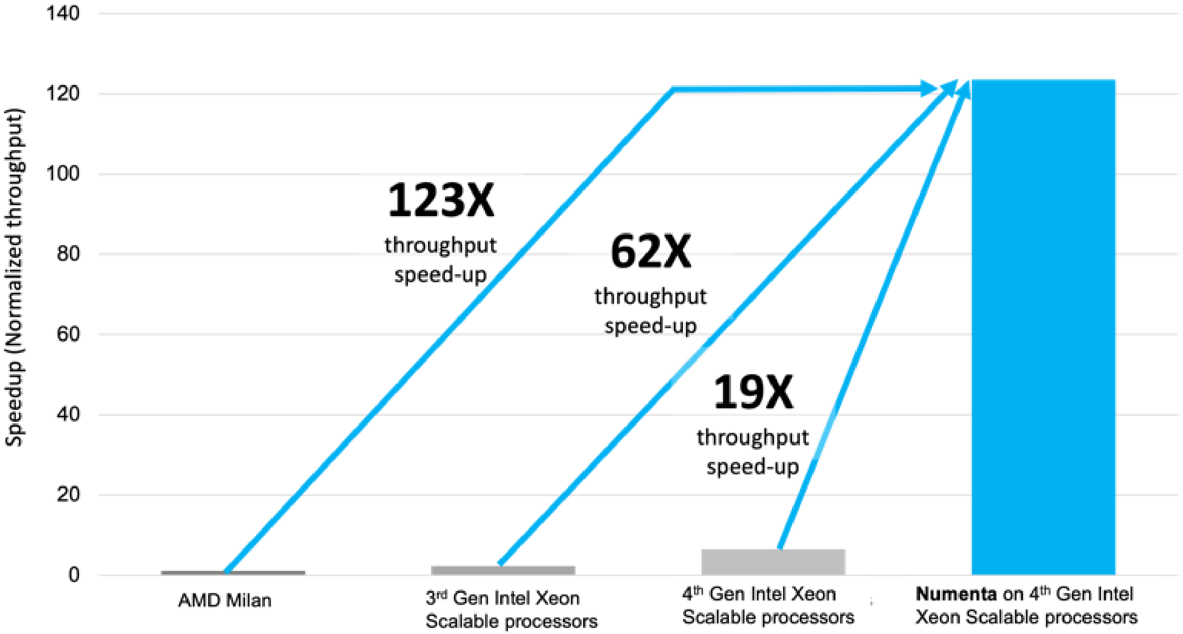


Generative vs. Non-Generative AI: Both Required

	Generative AI (GPT-like)	Non-Generative AI (BERT-like)
What	**Creates new text**	**Understands existing text**
How	Models create original, human-like responses	Models analyze, interpret, and find answers within text
Pros	• Creativity • Flexibility	• Accuracy • Price / Performance • Safety & Control
Cons	• Unreliable • Slow and expensive	• Can’t do long contexts
Examples	• Create chatbot responses • Translations • Summarization	• Compare and classify text • Identify sentiment of a document • Find answers to questions in document collections • Extract entities

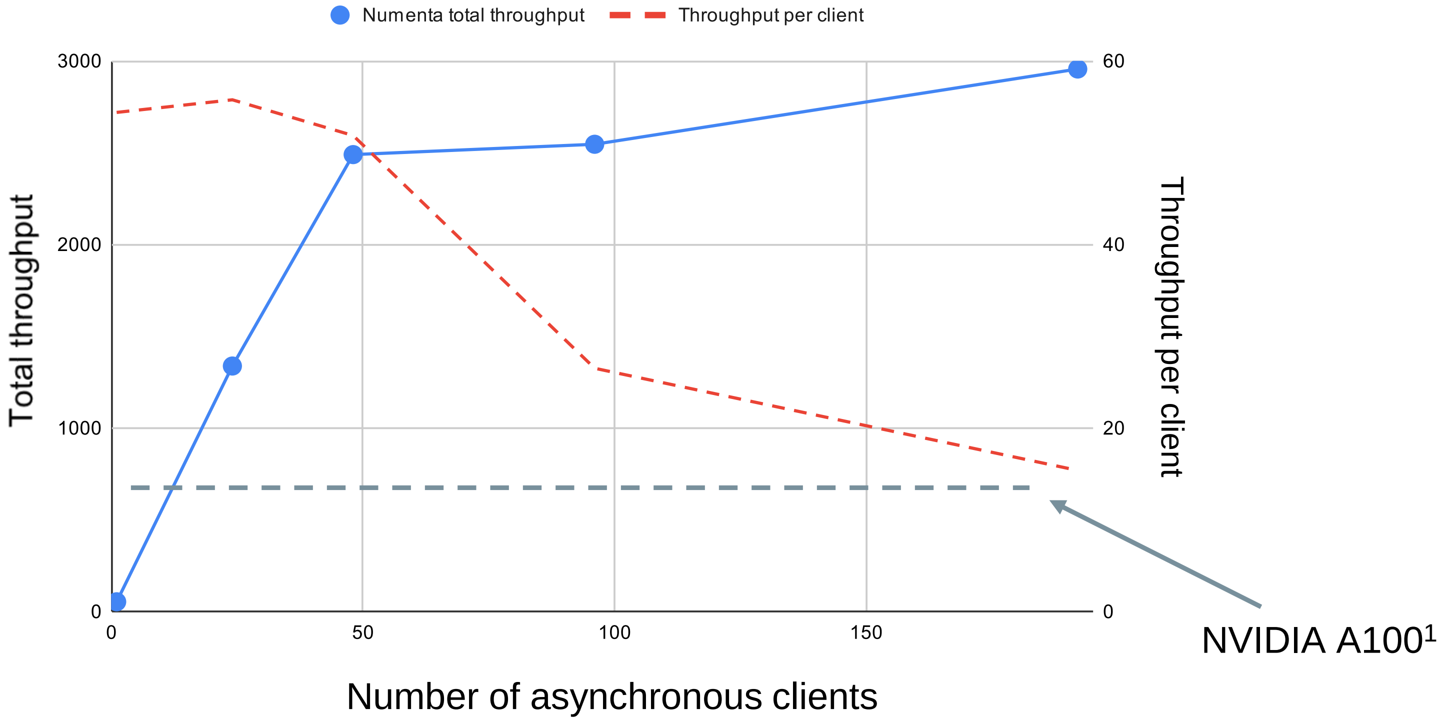
Numenta

Large Throughput Increases With AMX + Numenta



BERT-large, seq_len =64; 56-core SPR; AWS M6i.32xlarge [32 core Ice lake]; AWS M6a.48xlarge [48 core AMD Milan]

Throughput With Asynchronous Clients



BERT-Large, Seq len 64, BS=1
System: AWS m7i.48xlarge
96-core, 4th Gen Xeon

[1] https://github.com/NVIDIA/DeepLearningExamples/tree/master/TensorFlow/LanguageModeling/BERT#inference-performance-nvidia-dgx-a100-1x-a100-40gb

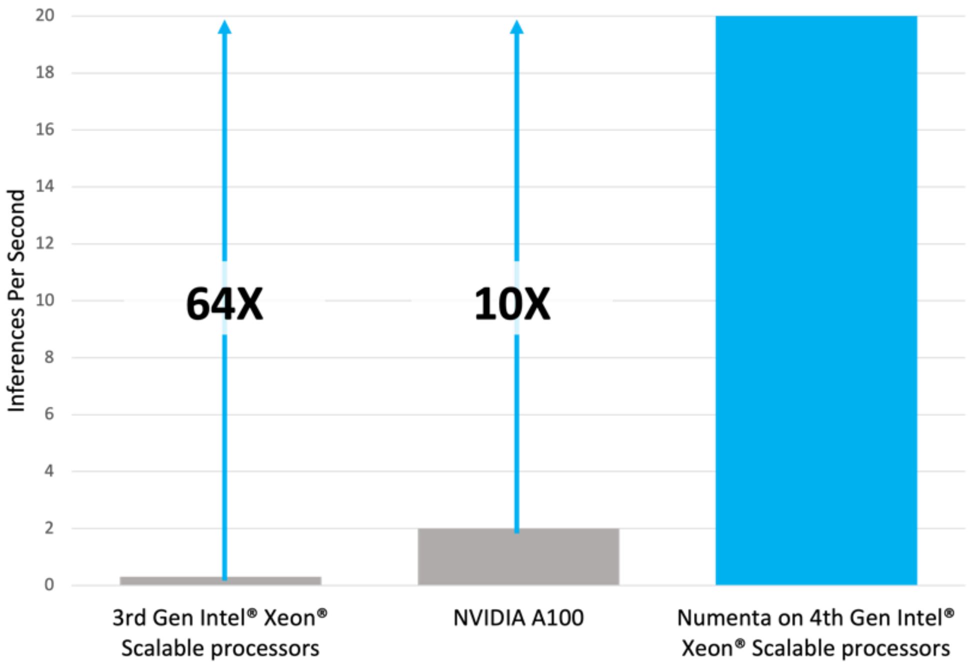
Generative AI Increases Compute Even More

From millions to billions parameters	Iterative build out	Increased cost with increased context
20X-1000X larger than other LLMs such as BERT	Entire model must be run many times to generate each result	As the amount of context increases, so does the complexity of the task

200 - 1000 X # tokens X context length

= 10,000 – 100,000 times more compute

Scaling GPT Models



Numenta + AMX delivers

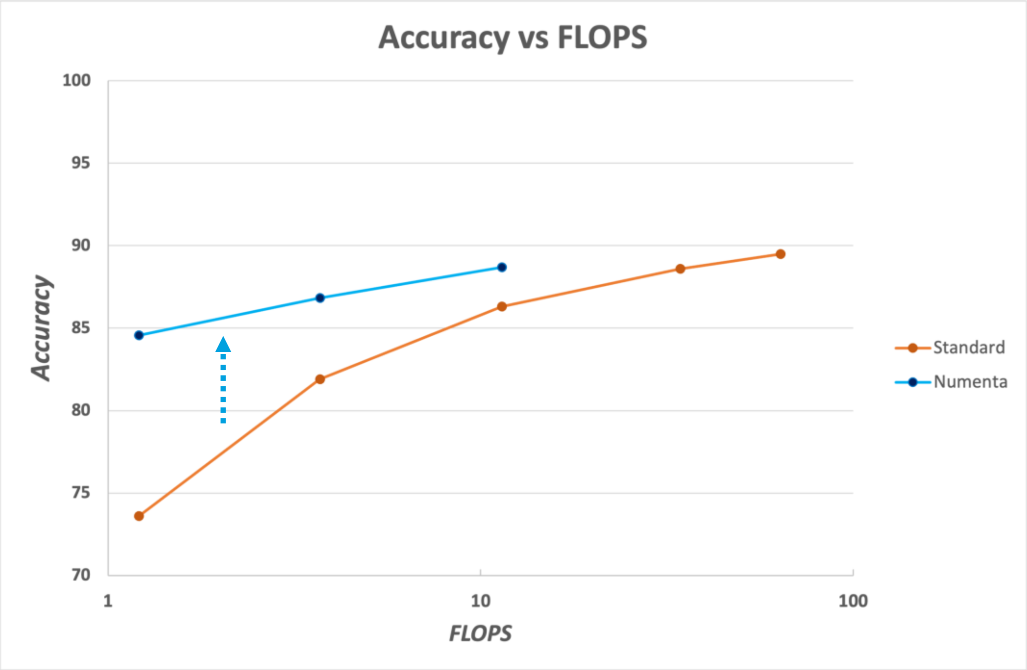
- **10X throughput** of NVIDIA A100
- **Latencies <.5 second**

Results shown for 32 input tokens, 32 output tokens, GPT-J-6B

Numenta Shifts AI Accuracy Scaling Laws

- In AI accuracy increases with network size

- At a fixed compute cost, we achieve significantly higher accuracies



Data from (Tay et al., 2022)

Evolution of AI and Hardware Architectures

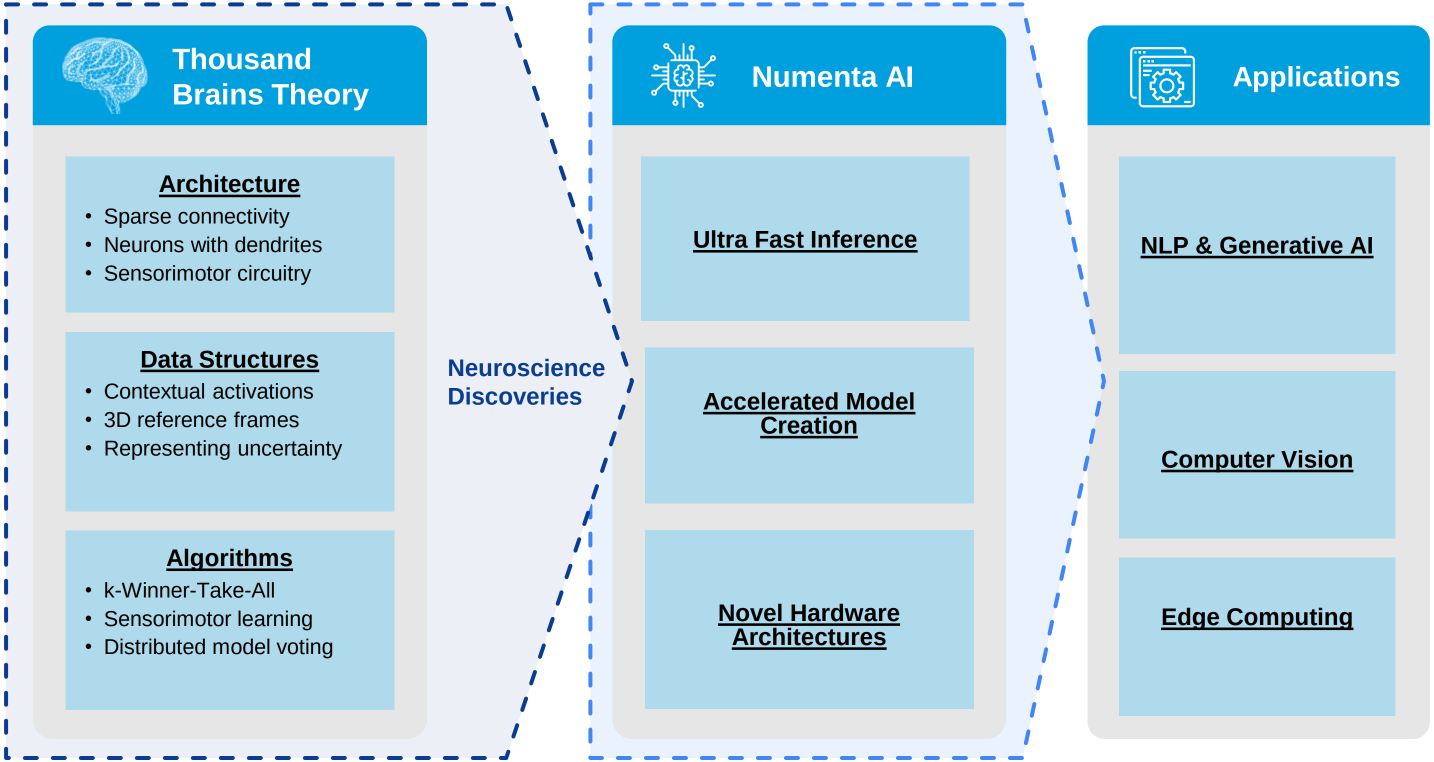
- AMX can provide significant performance gains for LLMs
  - Simple programming model accelerates development

- Matmul primitives are powerful, but complicate novel architectures
  - Many common sparsity techniques are incomputable

- For large models & sequence lengths, memory bandwidth is performance limiter
  - Use of HBM helps -- 3X throughput improvements

- Evolution of AI
  - Sparsity introduces irregularity – rigid instructions as in tensor cores introduce problems
  - Will require completely new architectural components

Neuroscience as a Technology



Thousand Brains Theory

Architecture

- Sparse connectivity
- Neurons with dendrites
- Sensorimotor circuitry

Data Structures

- Contextual activations
- 3D reference frames
- Representing uncertainty

Algorithms

- k-Winner-Take-All
- Sensorimotor learning
- Distributed model voting

Neuroscience Discoveries

Numenta AI

Ultra Fast Inference

Accelerated Model Creation

Novel Hardware Architectures

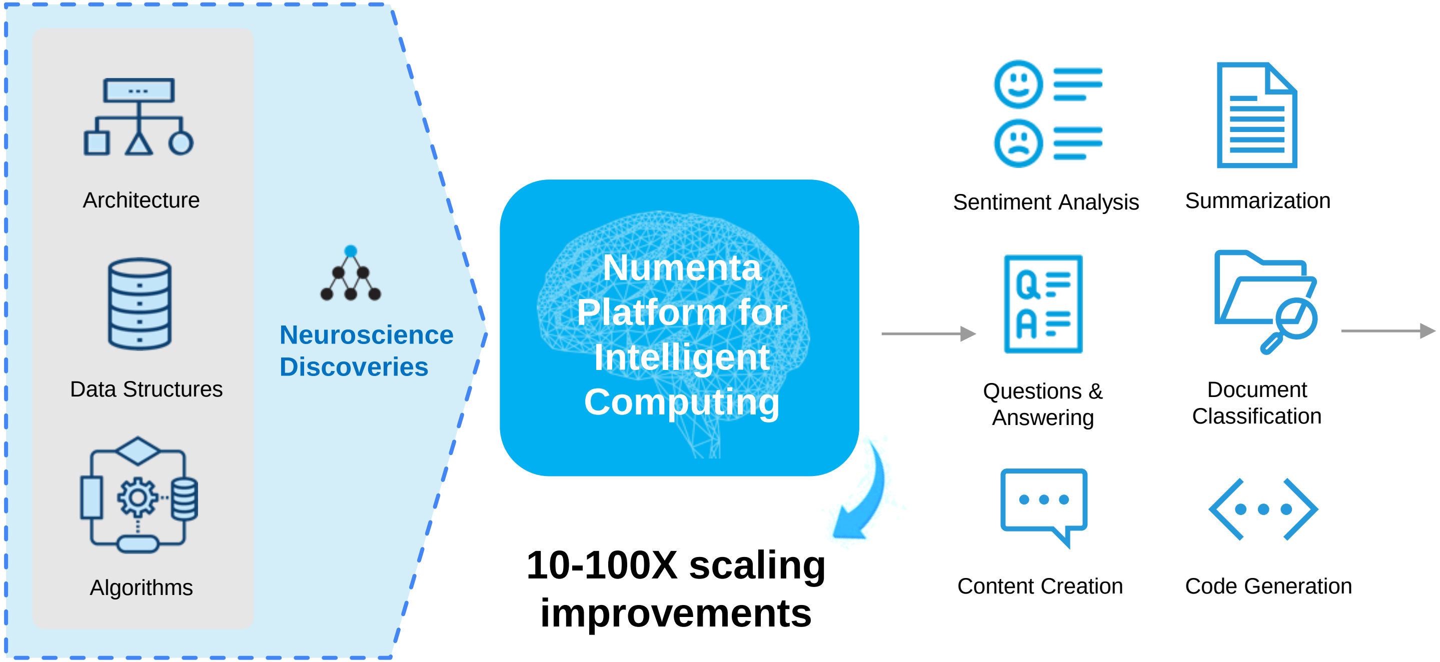
Applications

NLP & Generative AI

Computer Vision

Edge Computing

Numenta: Scalable and Secure Deployment of LLMs



Architecture

Data Structures

Algorithms

Neuroscience Discoveries

Numenta Platform for Intelligent Computing

10-100X scaling improvements

Sentiment Analysis

Summarization

Questions & Answering

Document Classification

Content Creation

Code Generation

Conversational Chatbots

Customer Service

⋮

Contract Analysis

Numenta

Summary

- State of AI today
  - Inference and training have very different requirements
  - With smart algorithms, CPUs are ideal for AI inference workloads. Lack of GPUs not a problem.

- Neuroscience shows us the future of AI
  - Extremely low power, highly sparse, dynamic routing of information
  - Training and inference will merge with continual learning

- The future of AI is not just faster and faster matmuls
  - Critical to have a flexible programming model
  - Modern CPUs illustrate the directions we need to go

Questions? Contact us: sahmad@numenta.com

Thank You!