DL Efficiency Through Technology: Hardware-Software Interactions

NASEM AI Workshop

Implications of Artificial Intelligence-Related Data Center Electricity Use and Emissions

November 13, 2024

Bill Dally

Chief Scientist and SVP of Research, NVIDIA Corporation

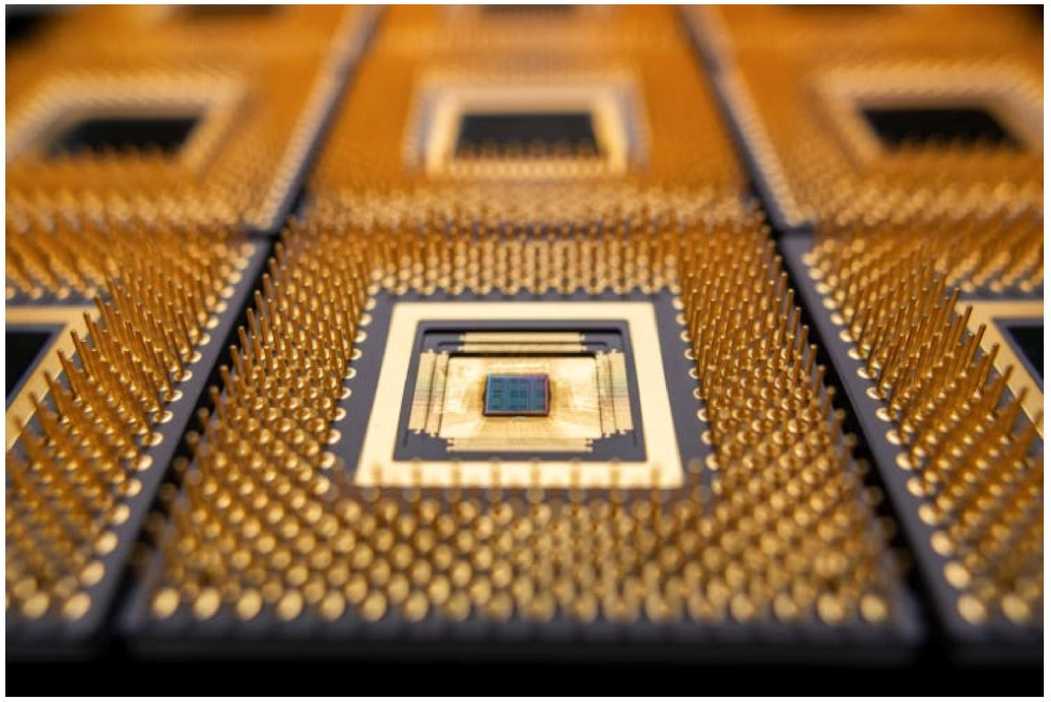
Adjunct Professor of CS and EE, Stanford University

OUTLOOK | 17 October 2024

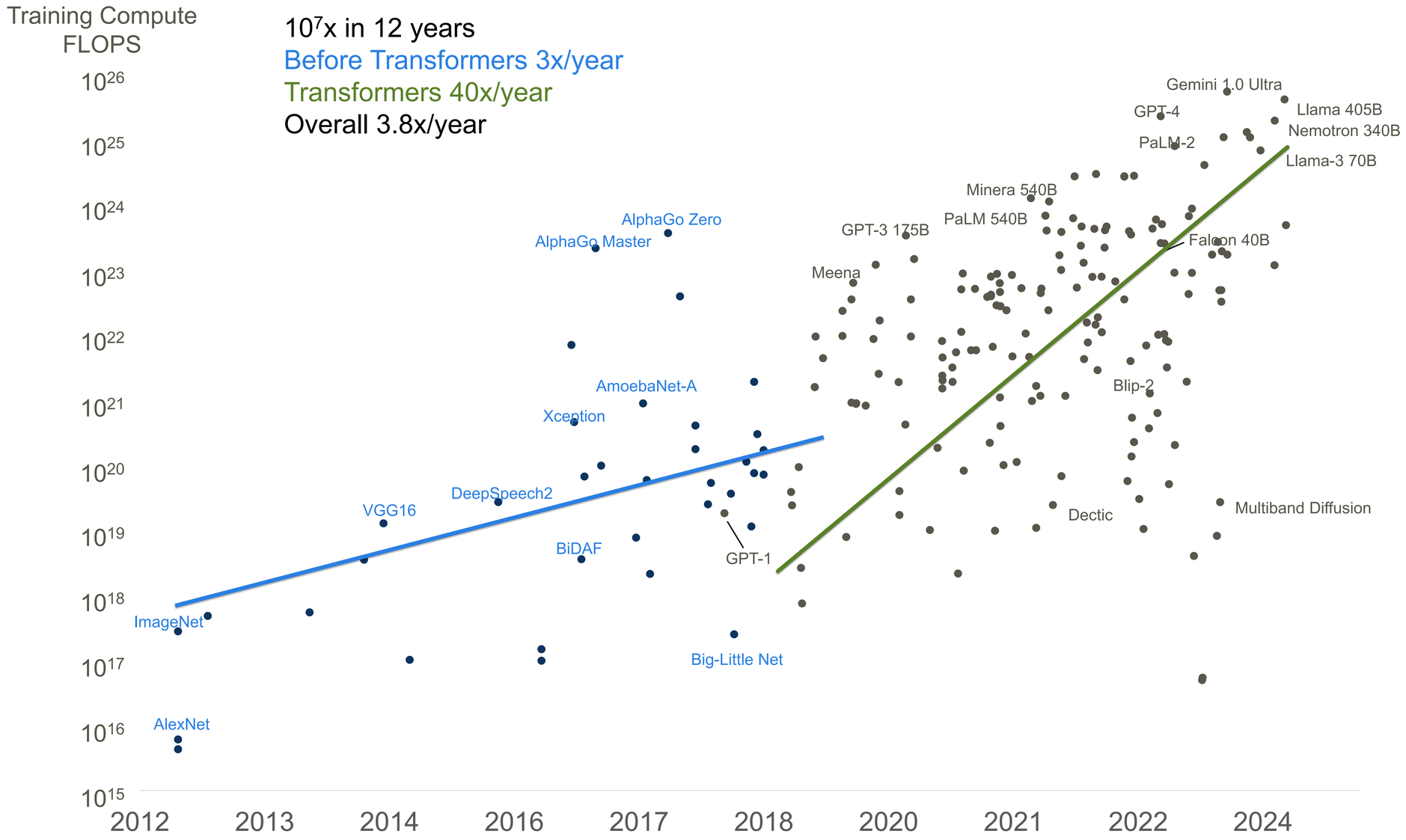
Fixing AI's energy crisis

Hardware that consumes less power will reduce artificial intelligence's appetite for energy. But transparency about its carbon footprint is still needed.

By Katherine Bourzac



FLOPS to Train a Model vs Calendar Year

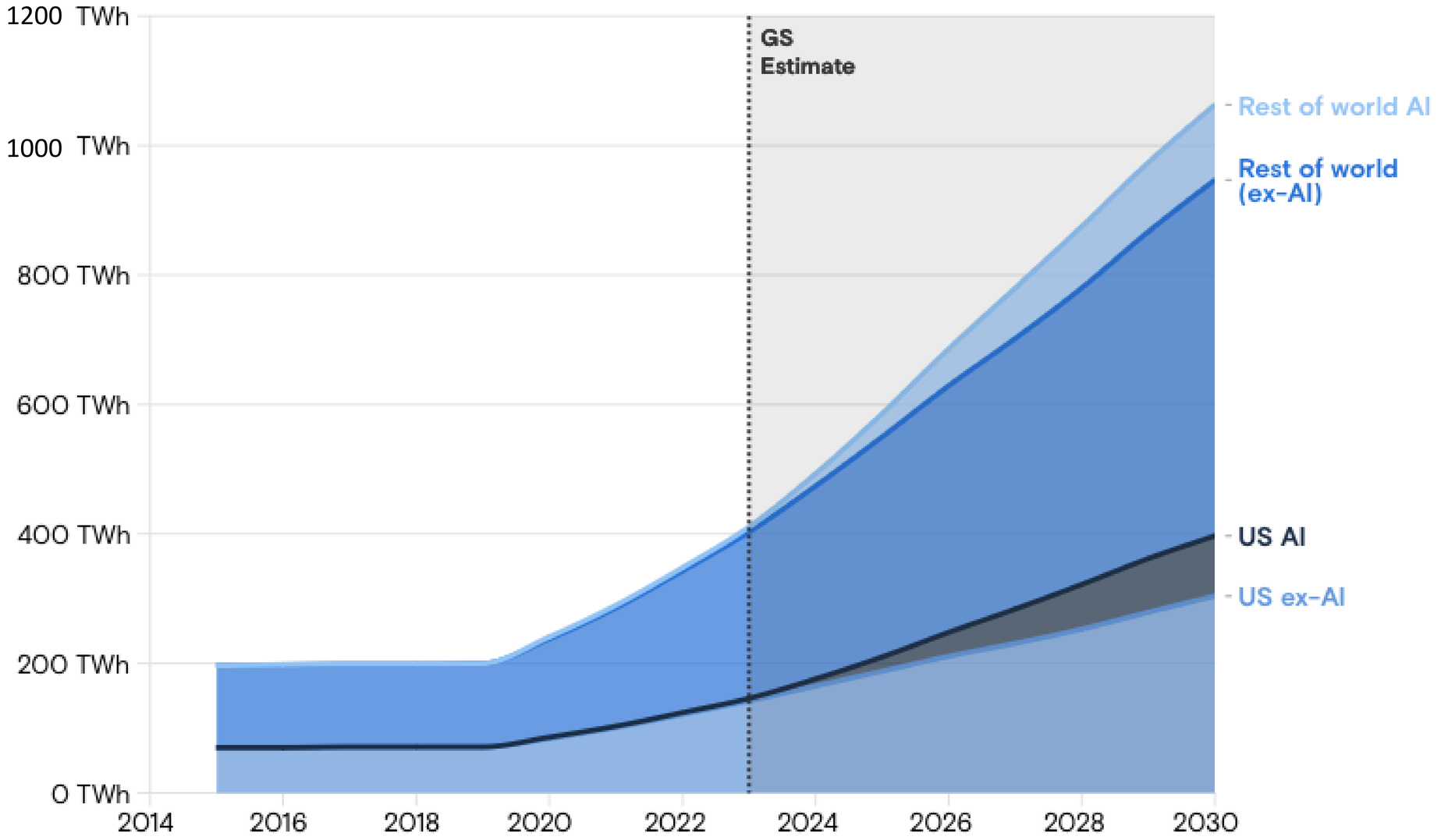


Each model is taking more Ops (training and inference)

We are also deploying more models

for more applications

Data center power demand



Source: Masanet et al. (2020), Cisco, IEA, Goldman Sachs Research

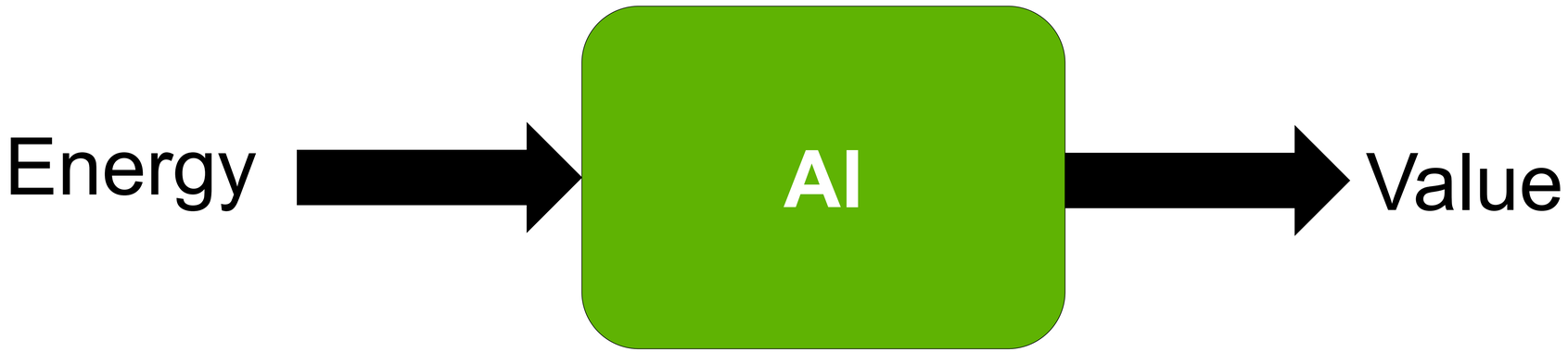
Goldman Sachs

Trends driving increased deployment of AI

Larger LLMs

More applications: medicine, education, science, entertainment, engineering, software corporate productivity

Multiple-pass inference (chain-of-thought, eg., GPT4-o1)


Energy
AI
Value

Huge increase in demand has been offset by large increase in efficiency

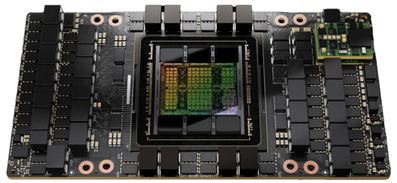
Single-Chip Inference Performance – 5000x in 12 years
Efficiency 1250x in 12 years

B200
20,000

Int 8 TOPS

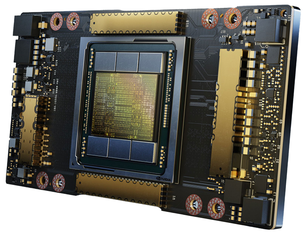
4500.00
4000.00
3500.00
3000.00
2500.00
2000.00
1500.00
1000.00
500.00
0.00

4/1/12
8/14/13
12/27/14
5/10/16
9/22/17
2/4/19
6/18/20
10/31/21
3/15/23



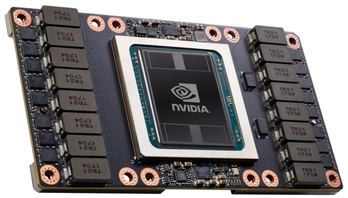
H100
4000.00

Sparsity

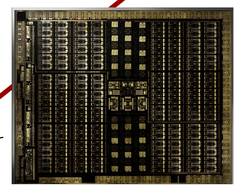


A100
1248.00

FP16
HMMA



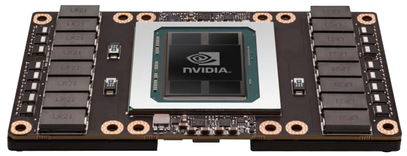
INT8
IMMA



Q8000
261.00

V100
125.00

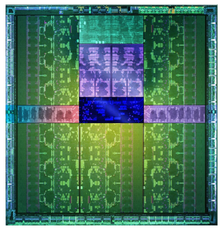
FP16
HDP4



M40
6.84

P100
21.20

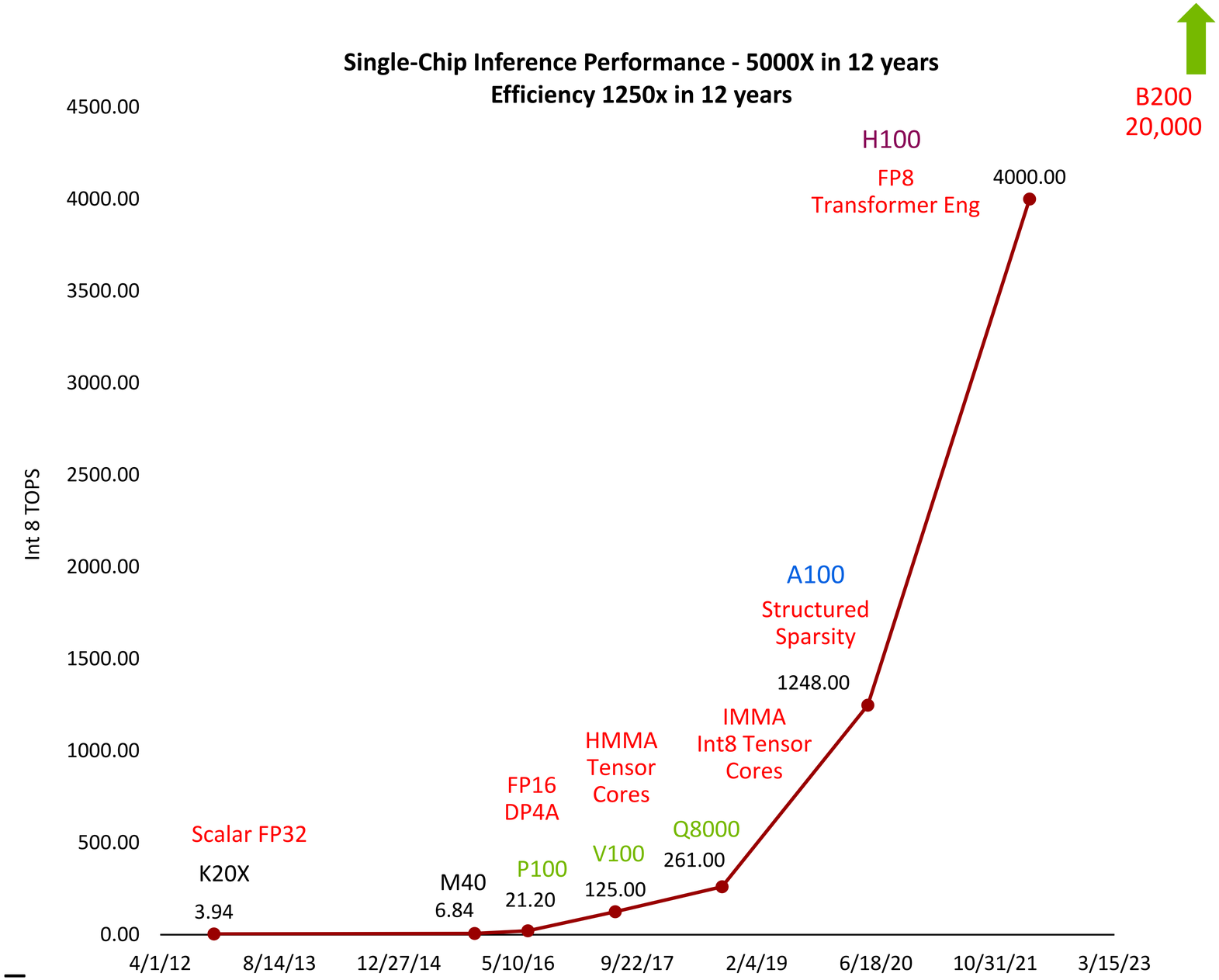
FP32
FMA



K20X
3.94

Gains from

- Number Representation
  - FP32, FP16, Int8, FP4
  - (TF32, BF16)
  - ~16x, 32x
- Complex Instructions
  - DP4, HMMA, IMMA
  - ~12.5x
- Process
  - 28nm, 16nm, 7nm, 5nm, 4nm
  - ~2.5x, 3x
- Sparsity ~2x
- Die Size 2x
- Model efficiency has also improved – overall gain > 1000x

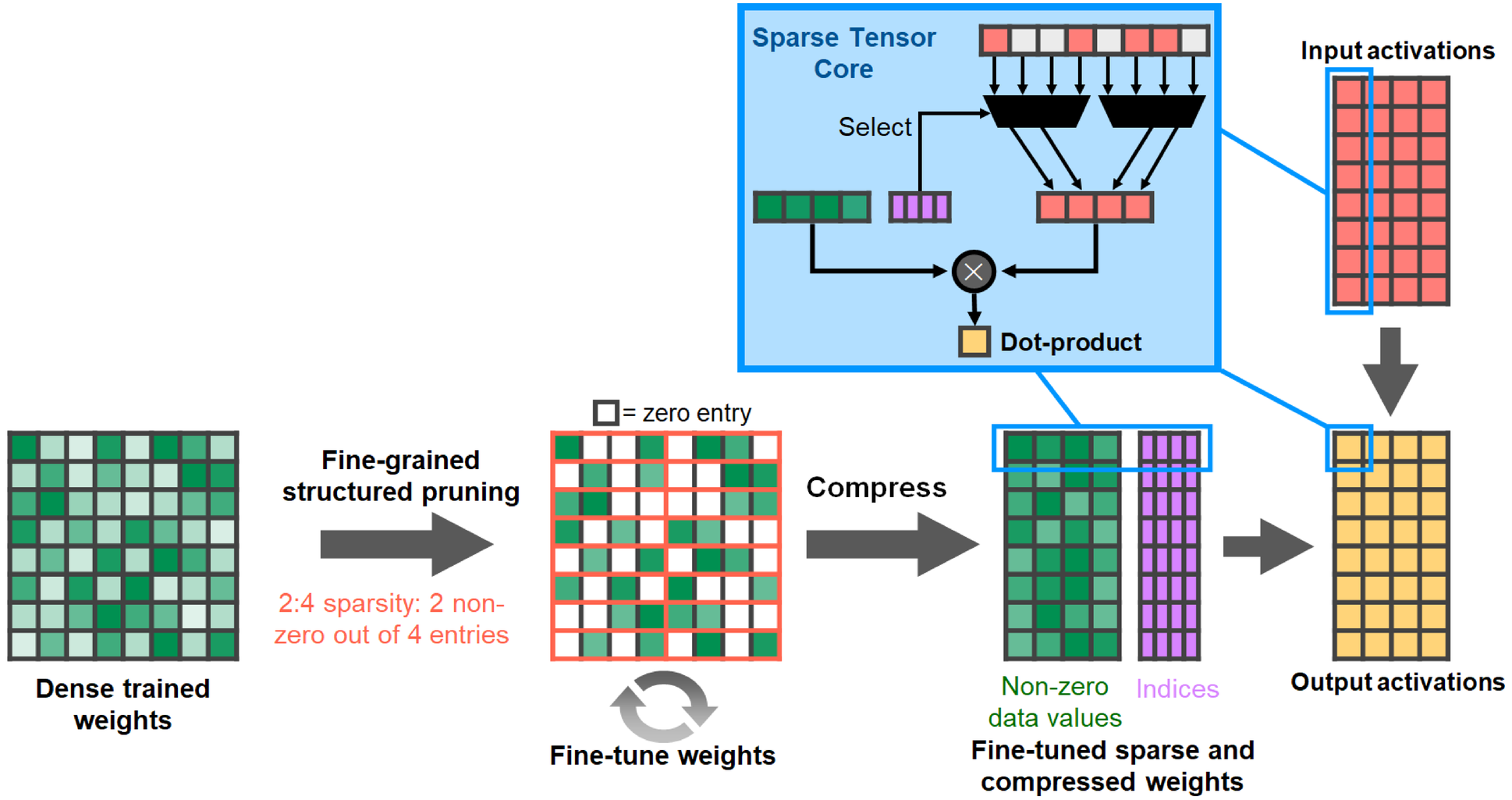


Specialized Instructions Amortize Overhead

Operation	Energy**	Overhead*
HFMA	1.5pJ	2000%
HDP4A	6.0pJ	500%
HMMA	110pJ	22%
IMMA	160pJ	16%

*Overhead is instruction fetch, decode, and operand fetch – 30pJ
**Energy numbers from 45nm process

Structured Sparsity

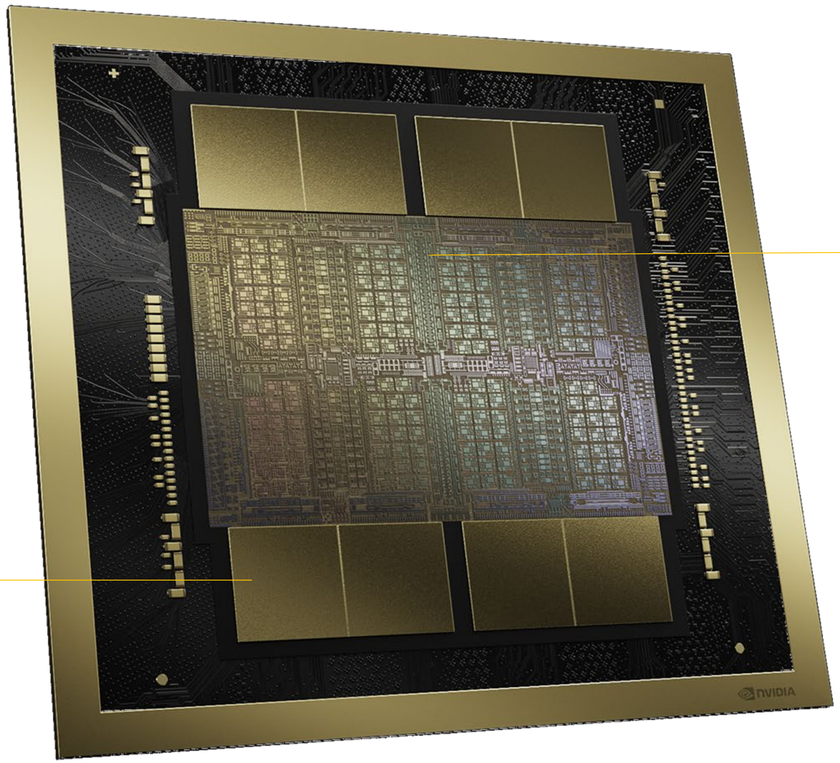


Mishra, Asit, et al. "Accelerating sparse deep neural networks." *arXiv preprint arXiv:2104.08378* (2021)

NVIDIA A100 Tensor Core GPU Architecture whitepaper

Blackwell B200

The Two Largest Dies Possible—Unified as One GPU



2 reticle-limited dies operate as One Unified CUDA GPU

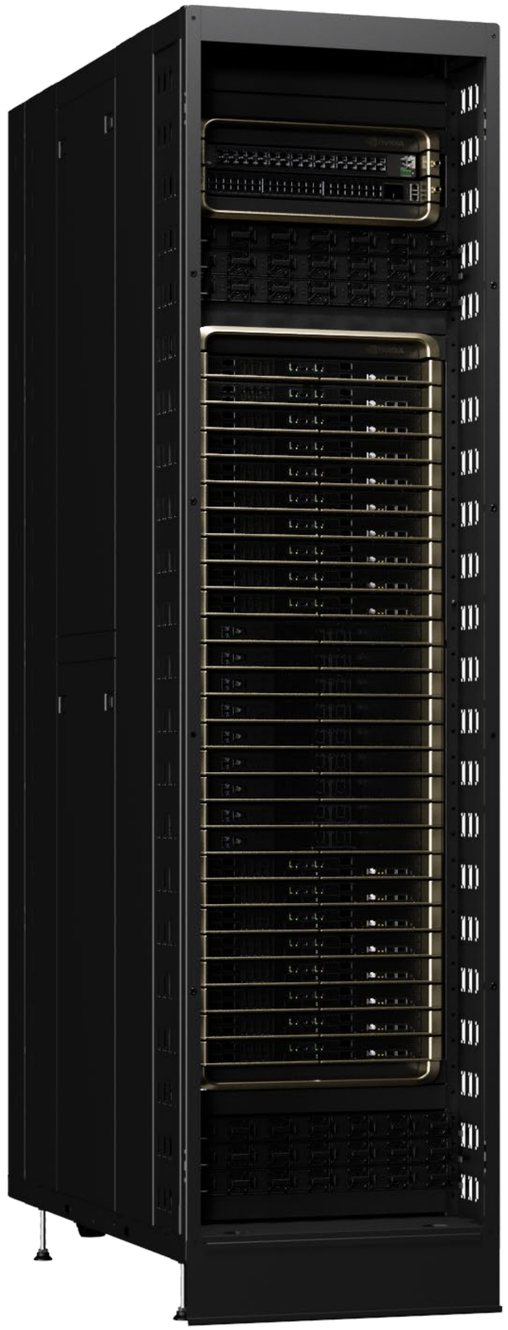
NV-HBI 10TB/s High Bandwidth Interface

Full performance. No compromises

4X Training | 30X Inference | 25X Energy Efficiency & TCO

Fast Memory
192GB HBM3e

10 PetaFLOPS FP8 | 20 PetaFLOPS FP4
192GB HBM3e | 8 TB/sec HBM Bandwidth | 1.8TB/s NVLink



GB200 NVL72

Delivers New Unit of Compute

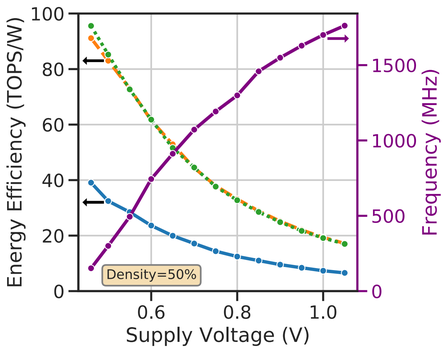
GB200 NVL72	36 GRACE CPUs 72 BLACKWELL GPUs Fully Connected NVLink Switch Rack

Training	720 PFLOPs
Inference	1.4 EFLOPs
NVL Model Size	27T params
Multi-Node All-to-All	130 TB/s
Multi-Node All-Reduce	260 TB/s

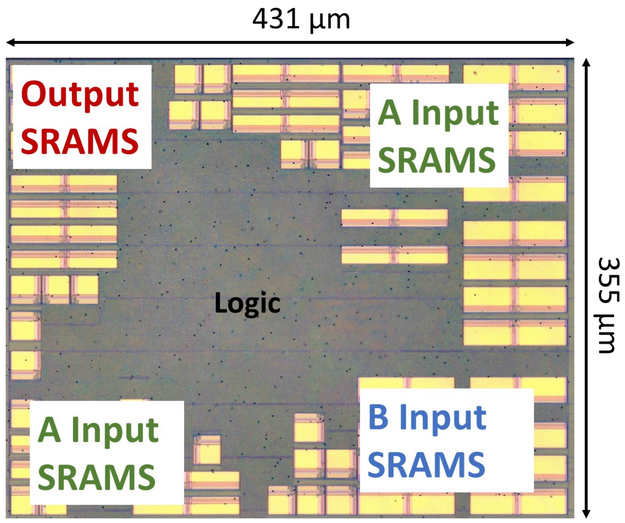
Energy-efficient DL Inference accelerator

Transformers, VS-Quant INT4, TSMC 5nm

- Efficient architecture
  - Used MAGNet *[Venkatesan et al., ICCAD 2019]* to design a low-precision DL inference accelerator for Transformers
  - Multi-level dataflow to improve data reuse and energy efficiency
- Low-precision data format: VS-Quant INT4
  - Hardware-software techniques to tolerate quantization error
  - Enable low cost multiply-accumulate (MAC) operations
  - Reduce storage and data movement
- Special function units



- 95.6 TOPS/W with 50%-dense 4-bit input matrices with VSQ enabled at 0.46V
- 0.8% energy overhead from VSQ support with 50%-dense inputs at 0.67V

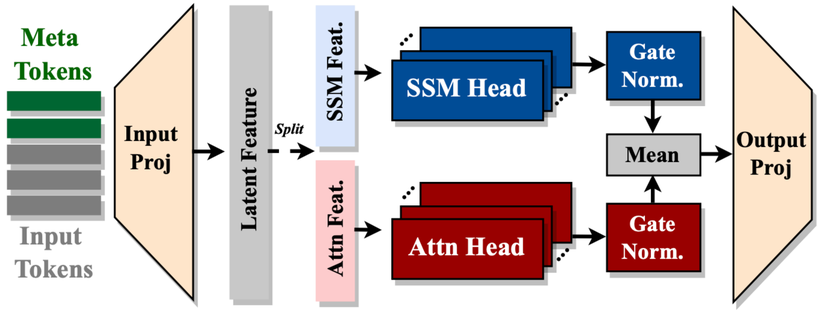


- TSMC 5nm
- 1024 4-bit MACs/cycle (512 8-bit)
- 0.153 mm^2 chip
- Voltage range: 0.46V – 1.05V
- Frequency range: 152 MHz – 1760 MHz

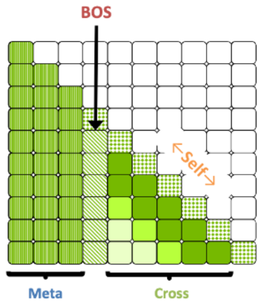
[Keller, Venkatesan, et al., “A 95.6-TOPS/W Deep Learning Inference Accelerator with Per-Vector Scaled 4-bit Quantization in 5nm”, JSSC 2023]

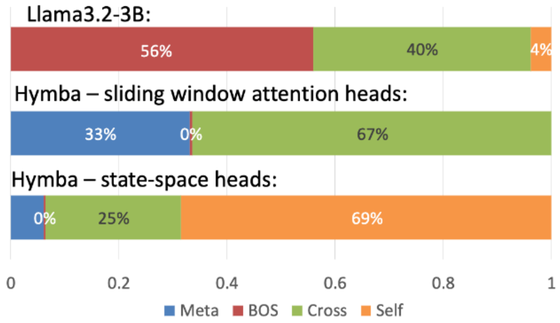
More efficient models

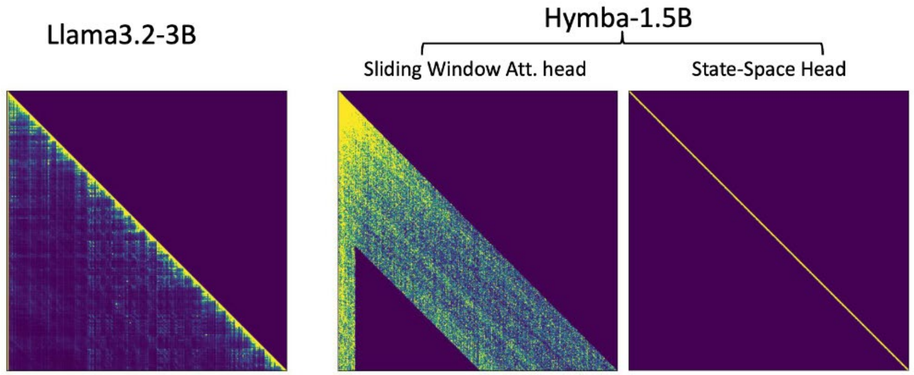
Hymba: Hybrid-Head Architecture for Small Language Models

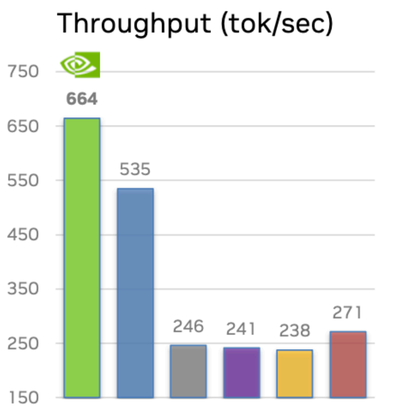


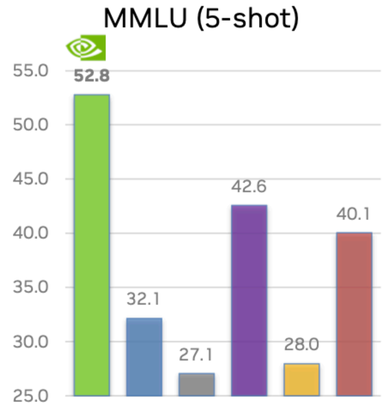
	Inference	Memory cost	Long context
Transformer	Slow (scales quadratically with seq length)	Large (scales linearly with seq length)	Great (memory storage)
State-Space Models (Mamba)	Fast (scales linearly with seq length)	Small (Grows linearly)	Bad (everything in fixed memory)
Hymba (Our)	Fast (approximately linear)	Small (fixed memory + tiny linear memory)	Great (because of hybrid memory)

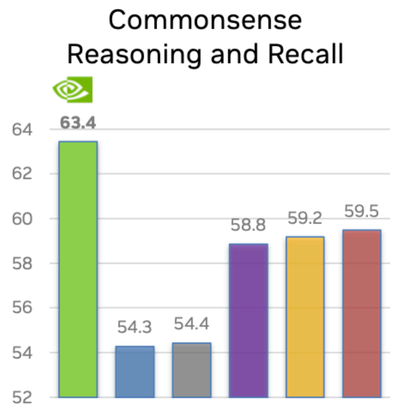


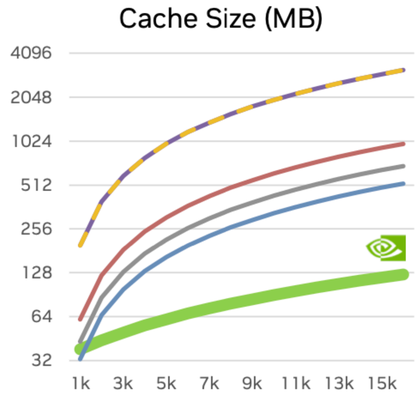






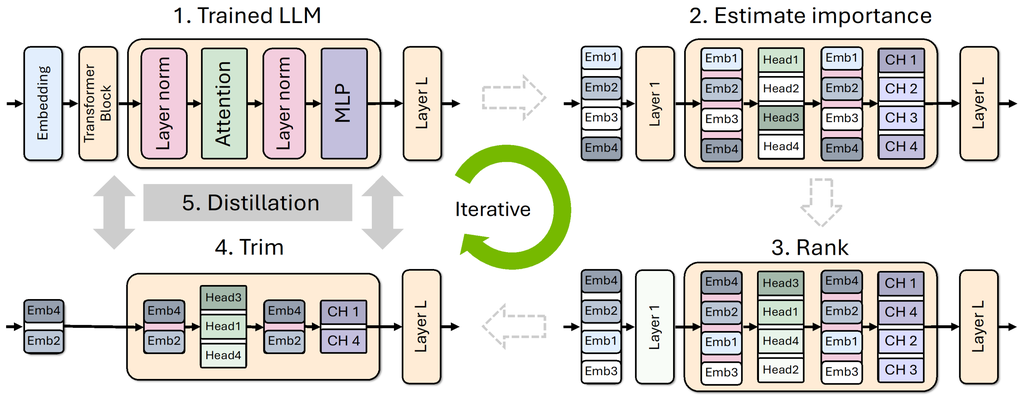


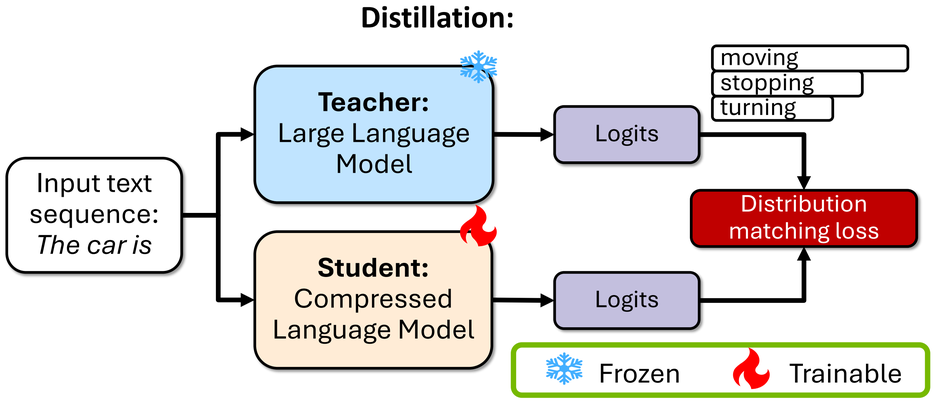


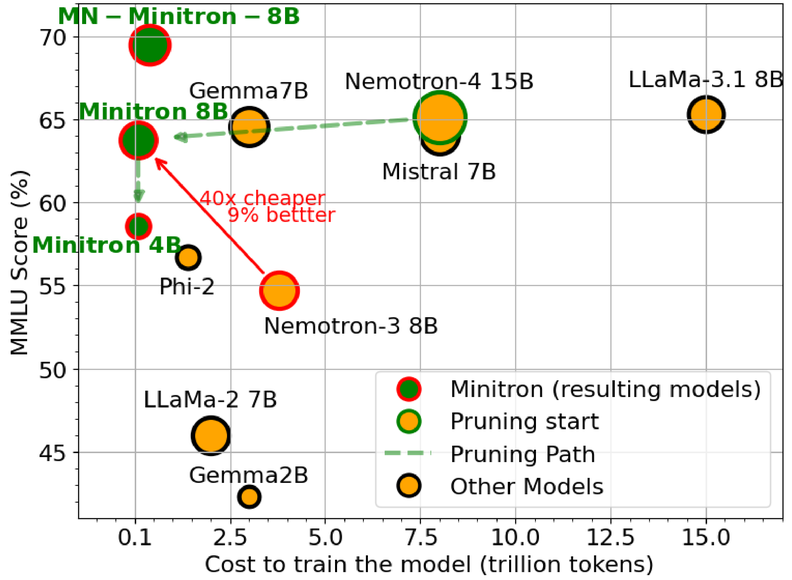


NVIDIA/Hymba-1.5B | META/LLaMA 3.2-1B | Apple/OpenELM-1B | Microsoft/Phi-1.5 | HuggingFace/SmolLM-1.7B | H2O/danube2-1.8B

Minitron: Compact Models via Pruning and Distillation







Benchmark	LLaMa-3	Mistral 7B	Gemma 7B	**MN-Minitron 8B**
Training tokens	15T	8T	6T	**0.4T**
General knowledge	65	64.1	64	**69.5**
Summarization	31	4.8	17	**32**
Reasoning	78	78.5	78	**80.4**
Coding	28	28.7	32	**36.2**

Details: *"LLM Pruning and Distillation in Practice: The Minitron Approach",* NVIDIA, arxiv: https://arxiv.org/abs/2408.11796, 2024

Conclusion

Conclusion

- Energy demands of deep learning are growing rapidly
  - Larger models, more data – training cost growing 40x/year
  - More applications, wider deployment
  - Deeper inference – chain of thought
- Demand partially offset by improved efficiency
- 1250x Improvement in hardware efficiency
  - Number representation
  - Sparsity
  - Complex instructions (MMA)
- Large gains in software efficiency
  - Hybrid state-space/transformer models
  - Distillation to smaller models, specialized models
- Efficiency will continue to improve
  - Better sparsity, number representation, data movement (10x)
  - More efficient, specialized models (10-100x)