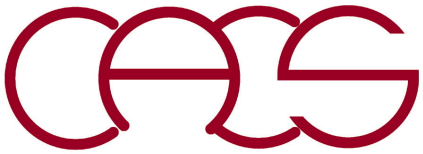
Parallel Programming: Now What?

Aiichiro Nakano

Collaboratory for Advanced Computing & Simulations
Department of Computer Science
Department of Physics & Astronomy
Department of Quantitative & Computational Biology
University of Southern California

Email: anakano@usc.edu

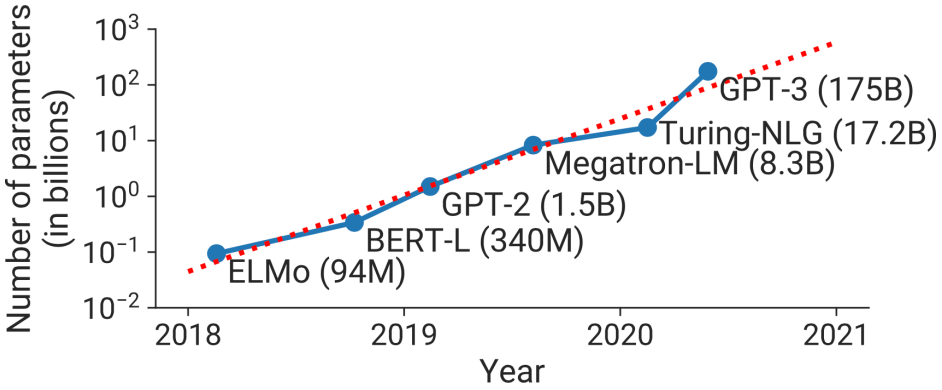
So what? Learned the current (MPI+OpenMP+CUDA) & emerging (MPI+OpenMP target) parallel programming languages

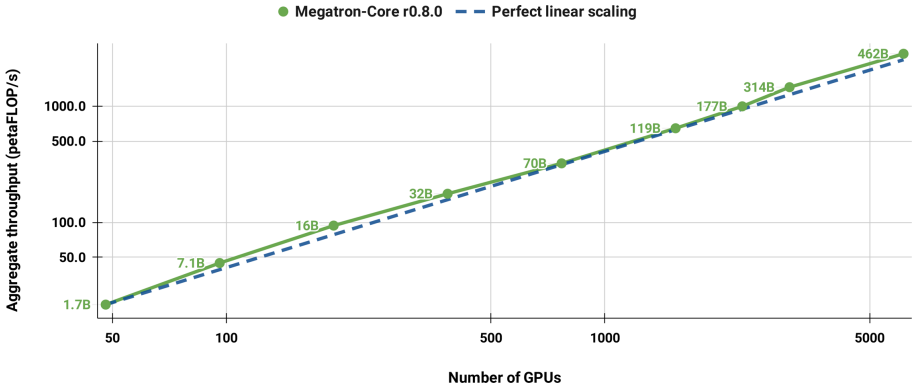


Now what?

Extreme-Scale Deep Learning

- **Trillion-parameter deep-learning (DL) model has been trained on 3000+ GPUs by Microsoft-NVIDIA team**





Narayanan *et al.*, "Megatron-LM," *SC21*
https://aiichironakano.github.io/cs596/Narayanan-MegatronLM-SC21.pdf
https://github.com/NVIDIA/Megatron-LM (distributed training, flash attention, *etc.*)

- **MegatronLM uses ZeRO (zero redundancy optimizer) system to eliminate memory redundancy & improve training speed**

Rajbhandari *et al.*, "ZeRO," *SC20*
https://aiichironakano.github.io/cs596/Rajbhandari-ZeRO-SC20.pdf

Test-drive and profile for final project?

Google Tensor Processing Unit

- **Google's tensor processing unit (TPU) accelerators are available on cloud**
- **XLA (accelerated linear algebra) is a compiler for TensorFlow applications on TPU**

https://cloud.google.com/tpu

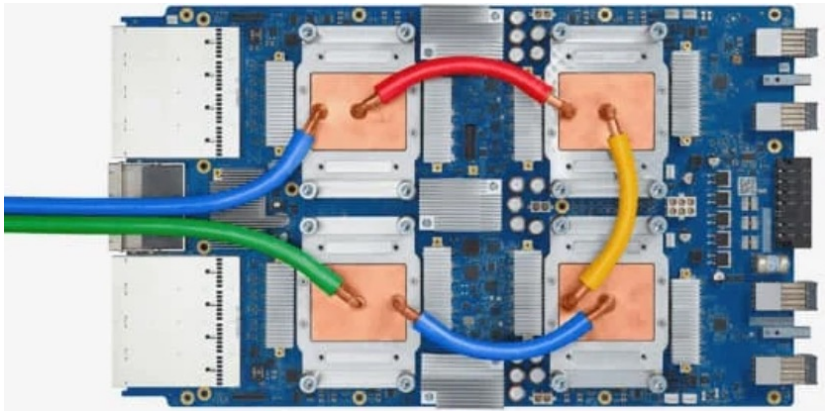
- **For physics-informed machine learning (ML), use JAX software built on Autograd (automatic differentiation)—both on GPU & TPU**

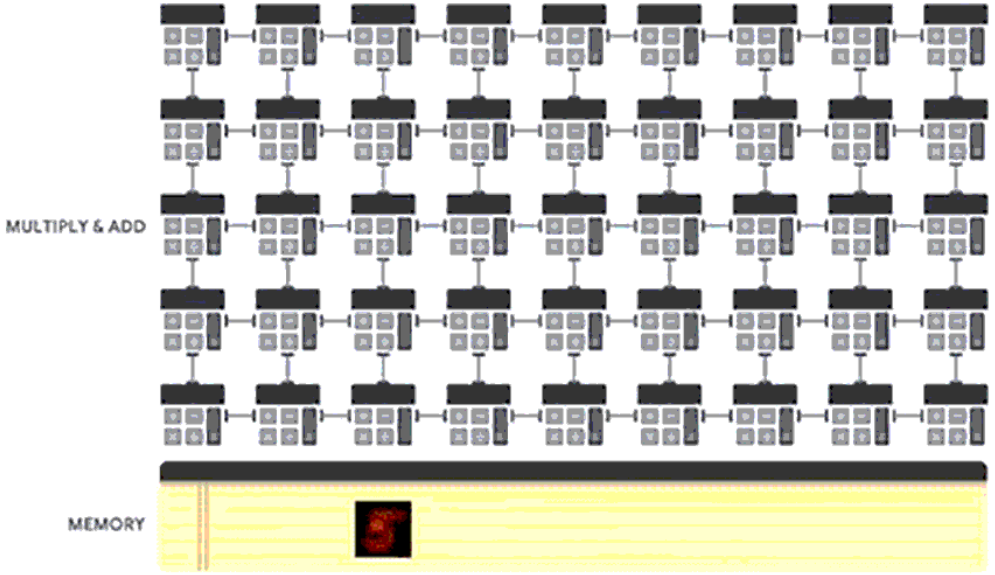
https://github.com/google/jax

w.r.t. model parameters

- **JAX-MD is an accelerated, differentiable molecular dynamics engine**

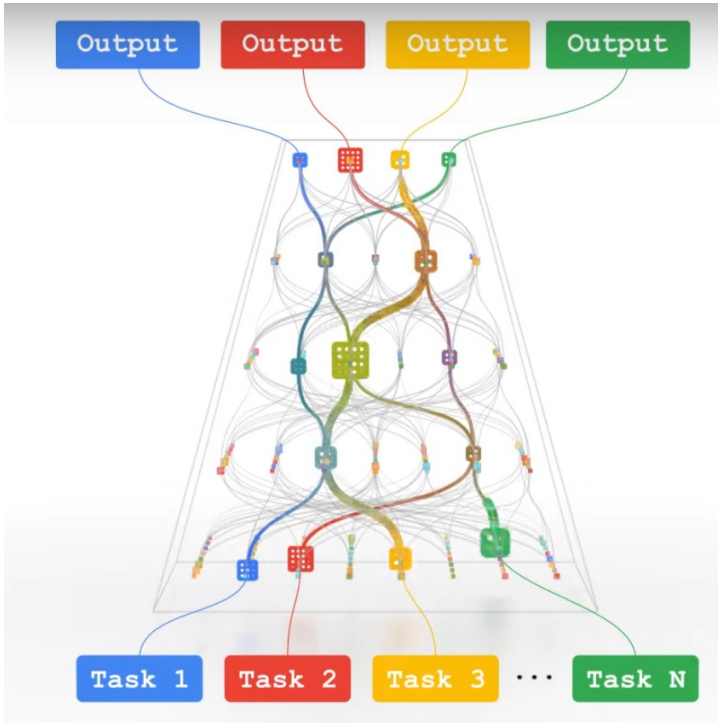
https://github.com/google/jax-md





Google's Pathways to AI Future

- **Pathways—a new AI architecture—will handle many tasks at once, learn new tasks quickly and reflect a better understanding of the world for human-like general AI**



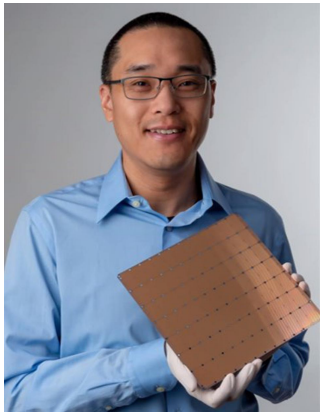
Jeff Dean, "AI isn't as smart as you think — but it could be," *TED Talk*
https://www.ted.com/talks/jeff_dean_ai_isn_t_as_smart_as_you_think_but_it_could_be

"Introducing Pathways: a next-generation AI architecture"
https://blog.google/technology/ai/introducing-pathways-next-generation-ai-architecture/

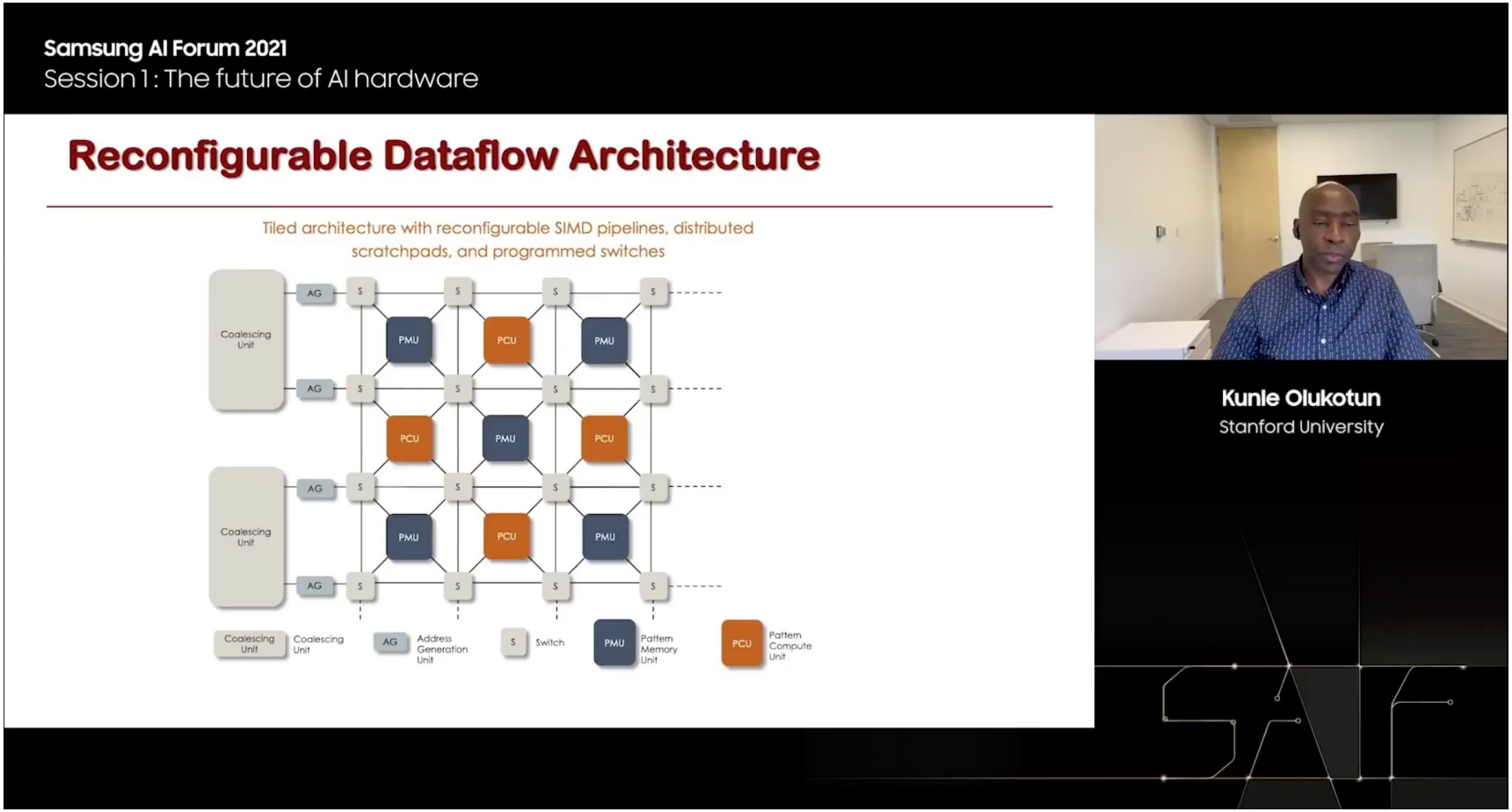
GPU & TPU Are No Good

- **It's sparsity: A lot of "multiply by zero" operations degrade speed & power efficiency**

  cf. "Selectable sparsity" on Cerebras AI chip
  https://cerebras.net/



- **Need new architectures & programming models**



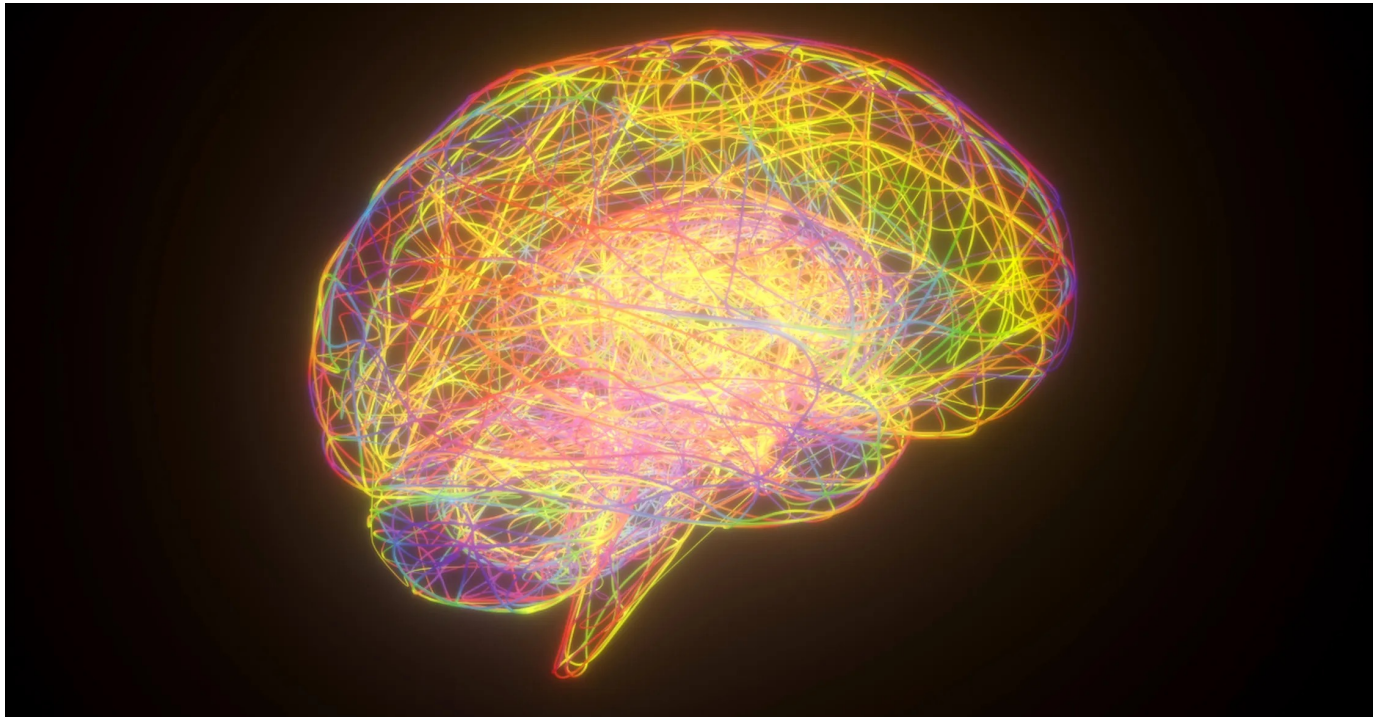
Will AI Destroy the Earth?

WIRED

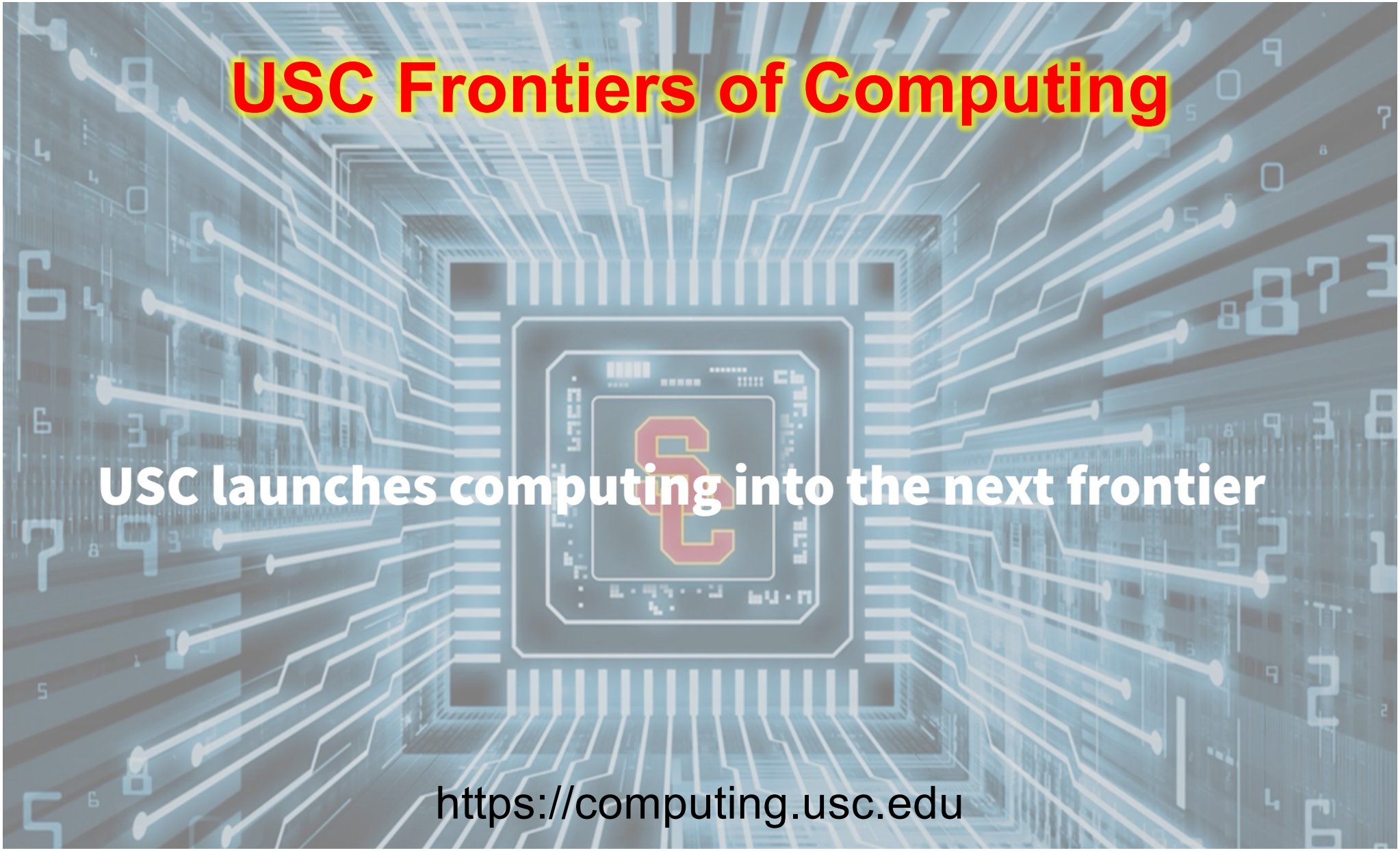
REECE ROGERS GEAR JUL 11, 2024 6:30 AM

AI's Energy Demands Are Out of Control. Welcome to the Internet's Hyper-Consumption Era

Generative artificial intelligence tools, now part of the everyday user experience online, are causing stress on local power grids and mass water evaporation.



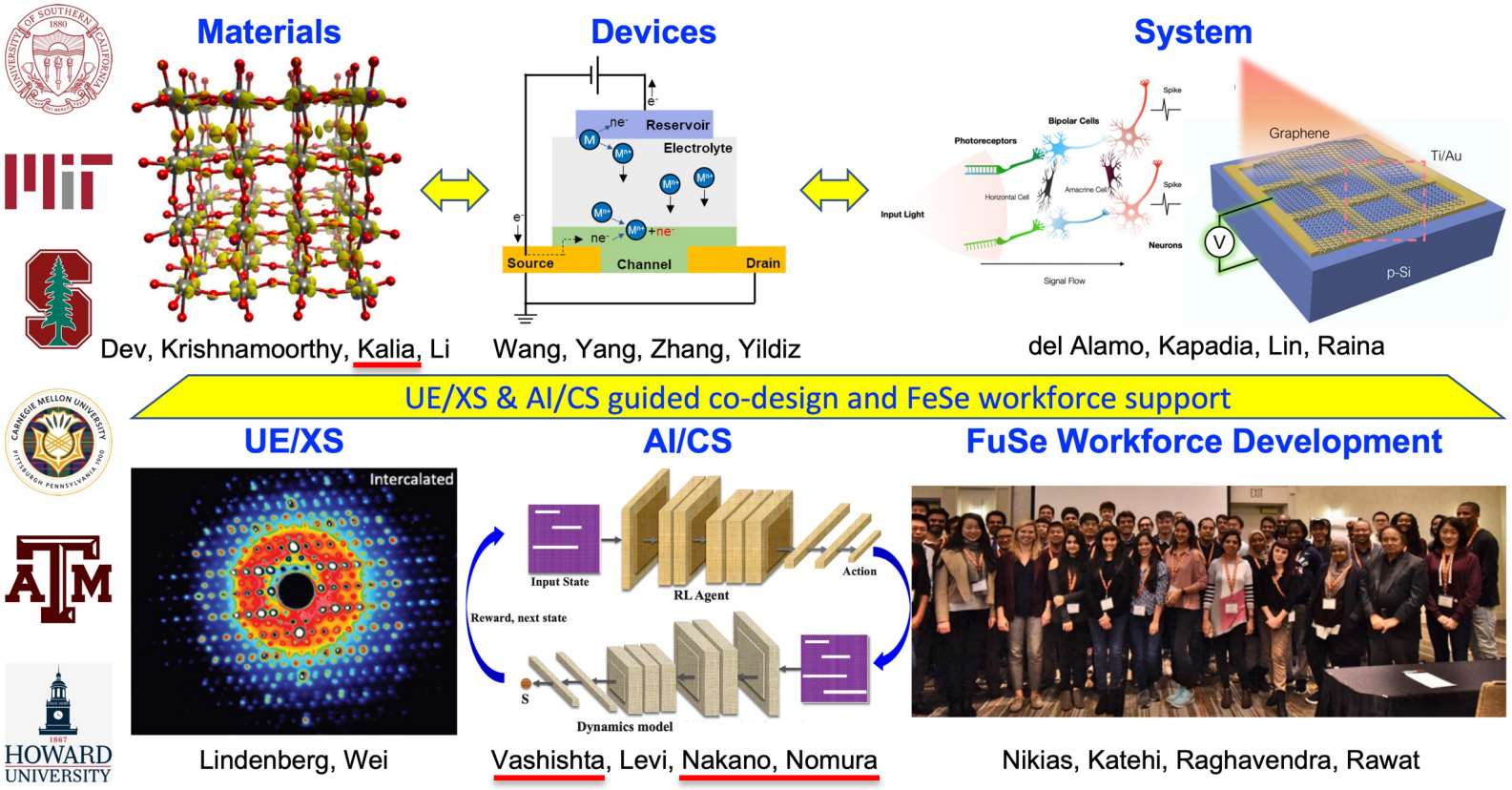
https://www.wired.com/story/ai-energy-demands-water-impact-internet-hyper-consumption-era/



- **$1 billion+, 10 years initiative**
- **New School of Advanced Computing**
- **30 senior & 60 junior & mid-level hires**

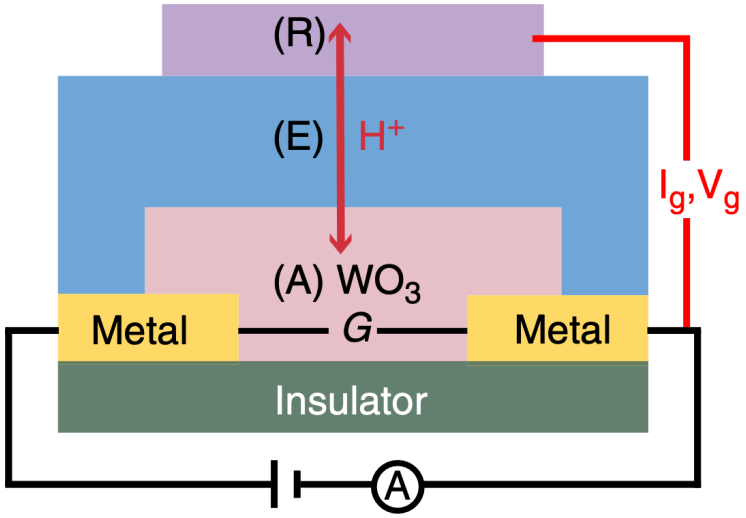
Future Semiconductors

- **USC-MIT-Stanford-CMU-TAMU-Howard team received a U.S. National Science Foundation (NSF) Future Semiconductors (FuSe) teaming grant (TG) for aJ in-sensor computing for sustainable AI future**
  Award #2235462, PI–Priya Vashishta (Mar. 15, 2023 - Mar. 14, 2025)



AttoJoule Neurons

- **Identify atomistic & electronic mechanisms of emerging attoJoule (aJ) in-sensor neuromorphic computing without external power:**
  1. **Protonic synapse that is deterministic & high speed with aJ energy consumption**
  2. **Retinal neurons for in-sensor image computing without external power**



Conductivity switch in H-doped WO_3

Protonic synapse (MIT)

Onen *et al.*, *Science* **377**, 539 ('22)

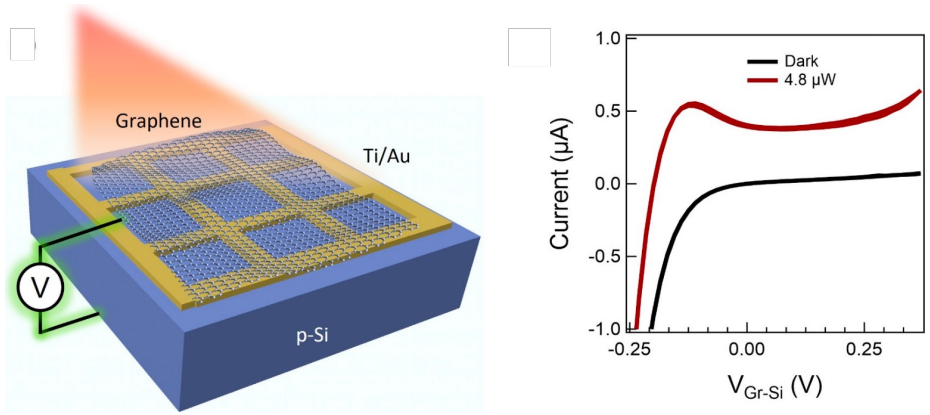


Photo-induced negative differential resistance (NDR) in graphene/Si

Oscillatory retinal neural network (USC)

Ahsan *et al.*, *ACS Nano* **18**, 23785 ('24)

Hokyo *et al.*, *J. Phys. Chem. Lett.* **15**, 9226 ('24)

- **Using our breakthrough AI-driven simulation technologies:**
  1. **Nonadiabatic quantum molecular dynamics (NAQMD) to study photoexcitation dynamics involving electrons & nuclei**
  2. **Machine learning (ML)-based neural-network quantum molecular dynamics (NNQMD) with state-of-the-art accuracy, speed & robustness: Allegro-Legato**

Ibayashi *et al.*, *ISC 2023*, DOI: 10.1007/978-3-031-32041-5_12

Computer Science Perspective

"Intelligent Heuristics Are the Future of Computing"

SHANG-HUA TENG, University of Southern California (USC), USA

Back in 1988, the partial game trees explored by computer chess programs were among the largest search structures in real-world computing. Because the game tree is too large to be fully evaluated, chess programs must make heuristic strategic decisions based on partial information, making it an illustrative subject for teaching AI search. In one of his lectures that year on AI search for games and puzzles, Professor Hans Berliner — a pioneer of computer chess programs[1] — stated:

"*Intelligent heuristics are the future of computing.*"

As a student in the field of the theory of computation, I was naturally perplexed but fascinated by this perspective. I had been trained to believe that "Algorithms and computational complexity theory are the foundation of computer science." However, as it happens, my attempts to understand heuristics in computing have subsequently played a significant role in my career as a theoretical computer scientist. I have come to realize that Berliner's postulation is a far-reaching worldview, particularly in the age of big, rich, complex, and multifaceted data and models, when computing has ubiquitous interactions with science, engineering, humanity, and society. In this article,[2] I will share some of my experiences on the subject of heuristics in computing, presenting examples of theoretical attempts to understand the behavior of heuristics on real data, as well as efforts to design practical heuristics with desirable theoretical characterizations. My hope is that these theoretical insights from past heuristics — such as spectral partitioning, multilevel methods, evolutionary algorithms, and simplex methods — can shed light on and further inspire a deeper understanding of the current and future techniques in AI and data mining.

ACM Trans. Intell. Syst. Technol. ('23)
https://doi.org/10.1145/3627708