Parallel Programming: Now What?

Aiichiro Nakano

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

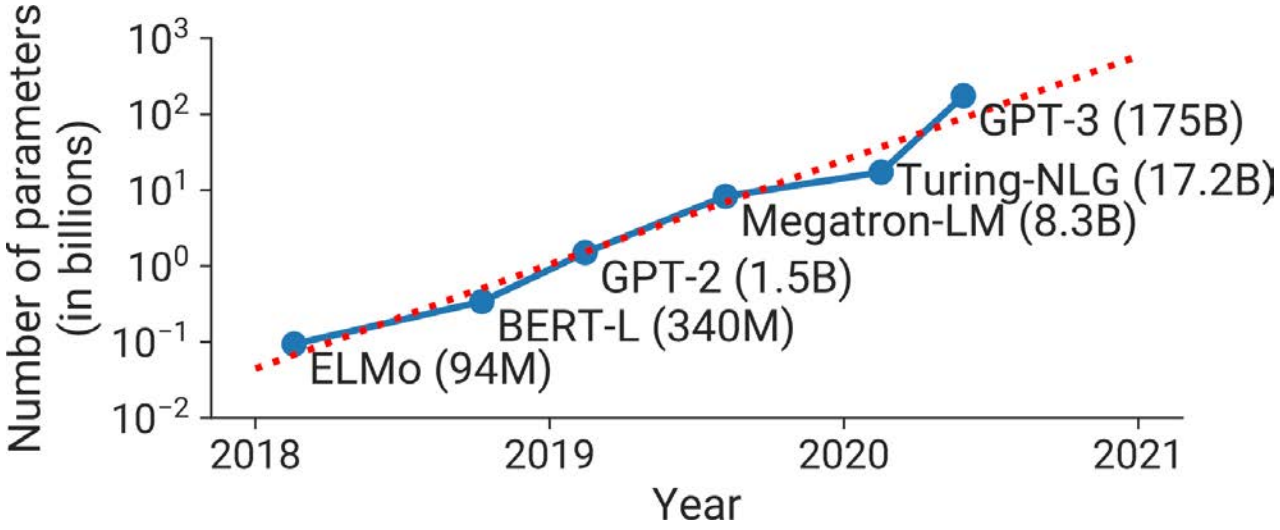
Email: anakano@usc.edu

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

- **MegatronLM used 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

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

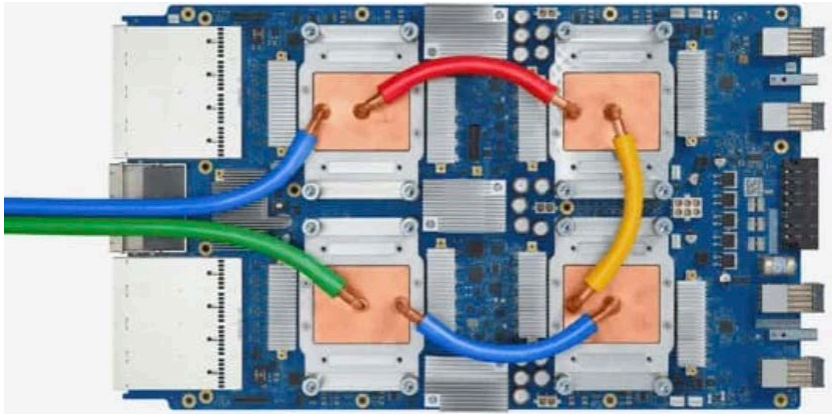
- **For physics-informed machine learning (ML), use JAX software built on Autograd (automatic differentiation)—both on GPU & TPU**

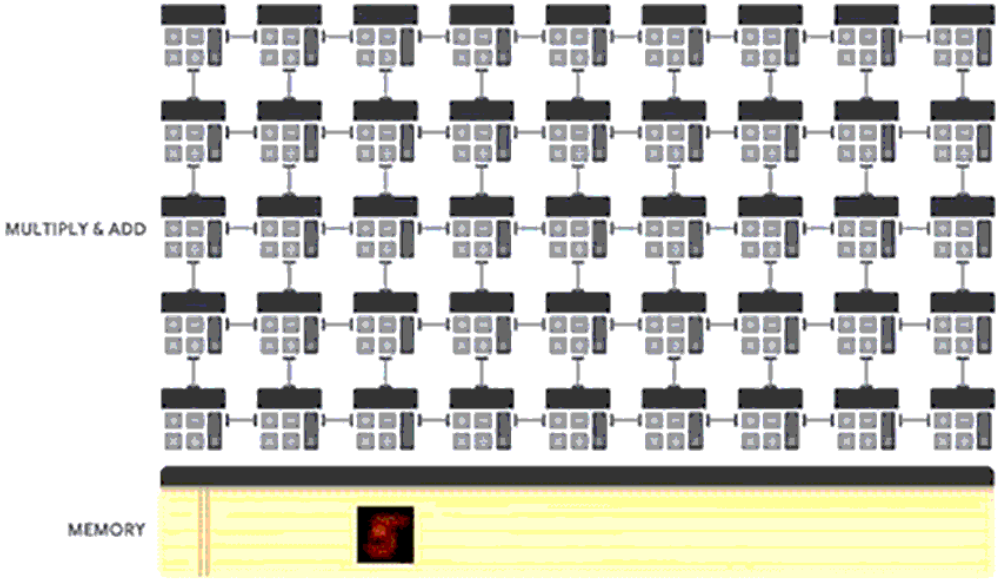
https://github.com/google/jax

w.r.t. model parameters

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

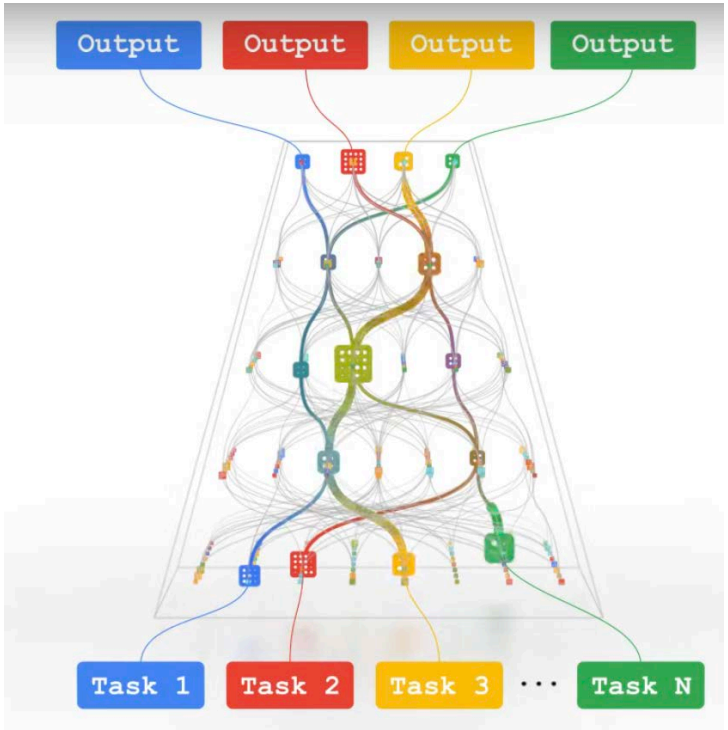
https://github.com/google/jax-md





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**



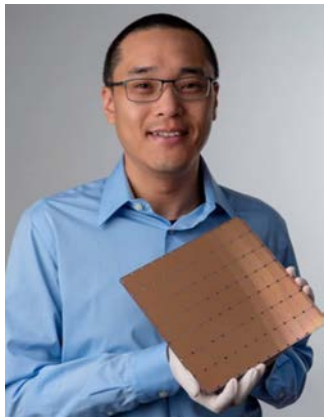
Jeff Dean, "AI isn't as smart as you think — but it could be," *TED Talk*
https://www.ted.com/talks/jeff_dean_ai_isn_t_as_smart_as_you_think_but_it_could_be

"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**