Accelerating Discovery

12/15/20 5:31 PM

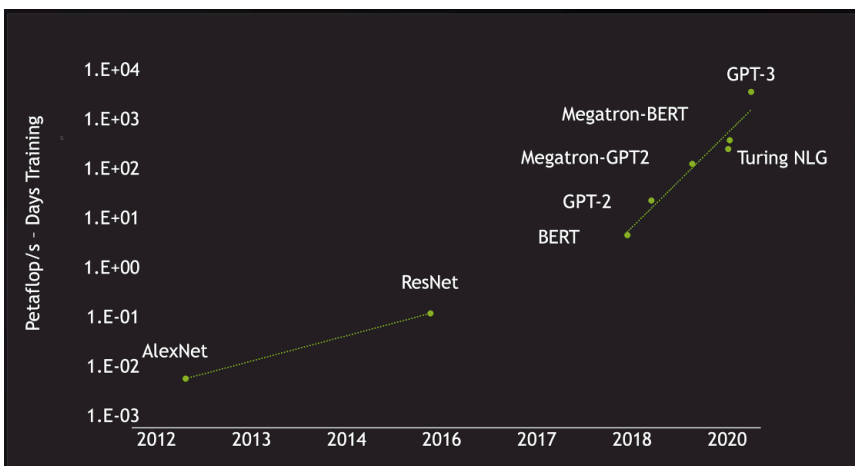
Speaker: William Dally (Chief Scientist NVIDIA, Stanford)

DL Application Breakthroughs

- DeepMind: Alpha-fold
- NVIDIA: Fluid simulation

DL Ingredients

- Hardware
- Large amount of data
- Neural Network models
-



GPUs (under NVIDIA A100)

New Sparsity Acceleration
New Multi-Instance GPU
3rd Gen NVLINK and NVSWITCH

- Structural sparsity brings additional speedups
  - Dense Matrix - Sparse Matrix - A100 Tensor Core (half units are zeroes, skip half of the compute core)
- The Selene Supercomputer
  - NVIDIA's DGX SuperPOD Deployment
  - One of the fastest and most efficient supercomputers on the planet, built in under one month

Key here: trade-off between dynamic complexity and accuracy

The Road Ahead

Number system - logarithmic numbers

- Multiplication are "free" for logarithmic systems

Cost of operations

- One of the biggest challenges is access to memory is very expensive

Importance of staying local

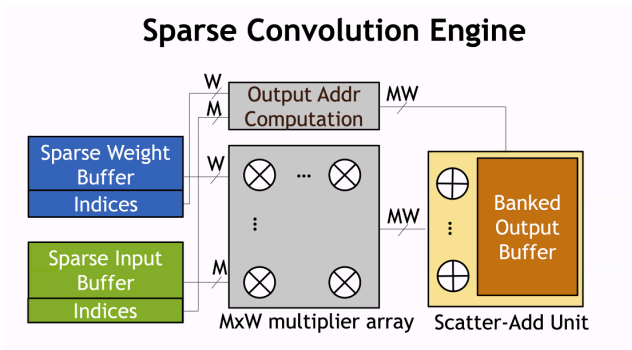
- Communication energy (moving data to core system)
- One way to improve localization is pruning DNN

Accelerators

EIE system - pruning network

SCNN: Sparse Convolution

- Only compute where
-



MAGNET

- Modular Accelerator Generator

DataFlow Options

- Weight stationary v.s. Output Stationary
- Add a weight stationary (WS) -> Multi-Level DataFlows

Conclusions