Accelerators for Machine Learning of Big Data

Ran Ginosar

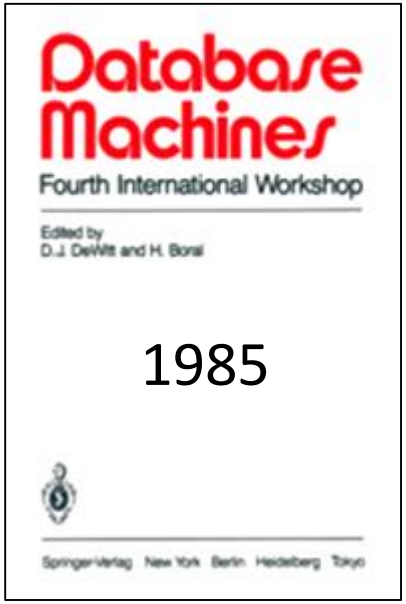
Technion, Israel

The “new” slogans

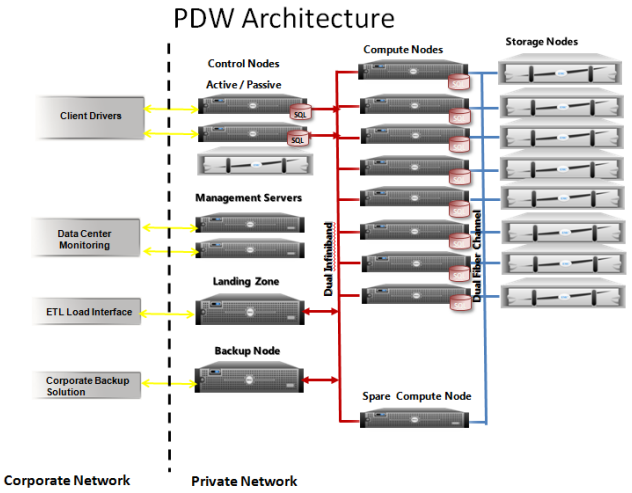
- Send the query to the data
- Process near memory
- Process in memory
- Processor in the disk

- So far, they mostly apply to small data

Processing near disks: Database Machines

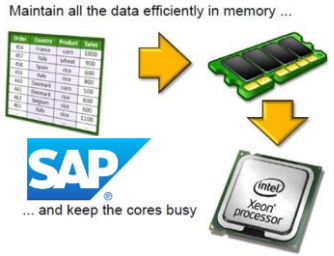


Microsoft



Processing near memory: In-memory database → analytics

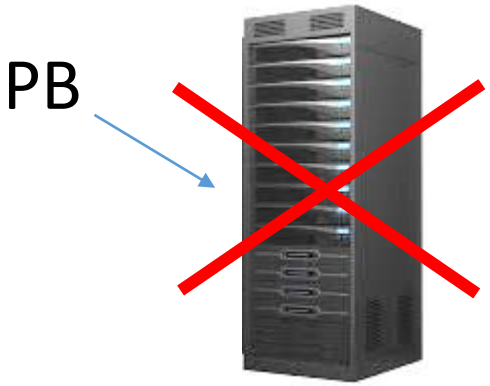
Goodyear MPP
1983



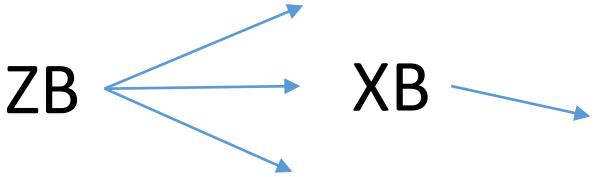
BIG DATA is not small data

- Accelerators for big data
  are not accelerators for small data
- The challenges are different
- Small data: performance, power
- Big data: energy

What is big data?



- Doesn't fit on a server
- At least an entire data center
  - Or many centers



- Example: global med/healthcare data
  - 10B persons, 10TB/person (omics, records, streams) = 100ZB

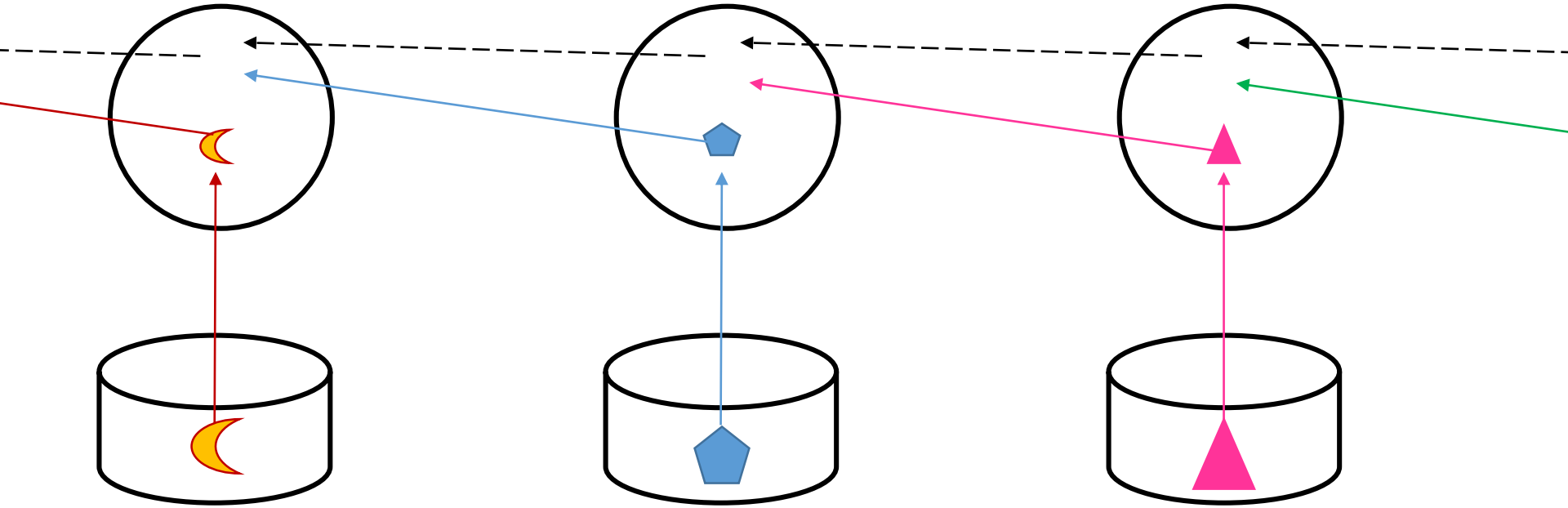
What is machine learning for big data?

- Entire data is needed
- Entire data is cross-associated
  - Multi-variance questions:
  - What features predict a specific disease at > 99%?
- N records, K features: $O(N^K)$
  - Imagine N=100 million persons, K=10 million features…
  - Interim algorithms hope for $O(N^m)$, m small integer
- Compute energy is easy
- Moving BIG DATA to BIG DATA is BIG ENERGY

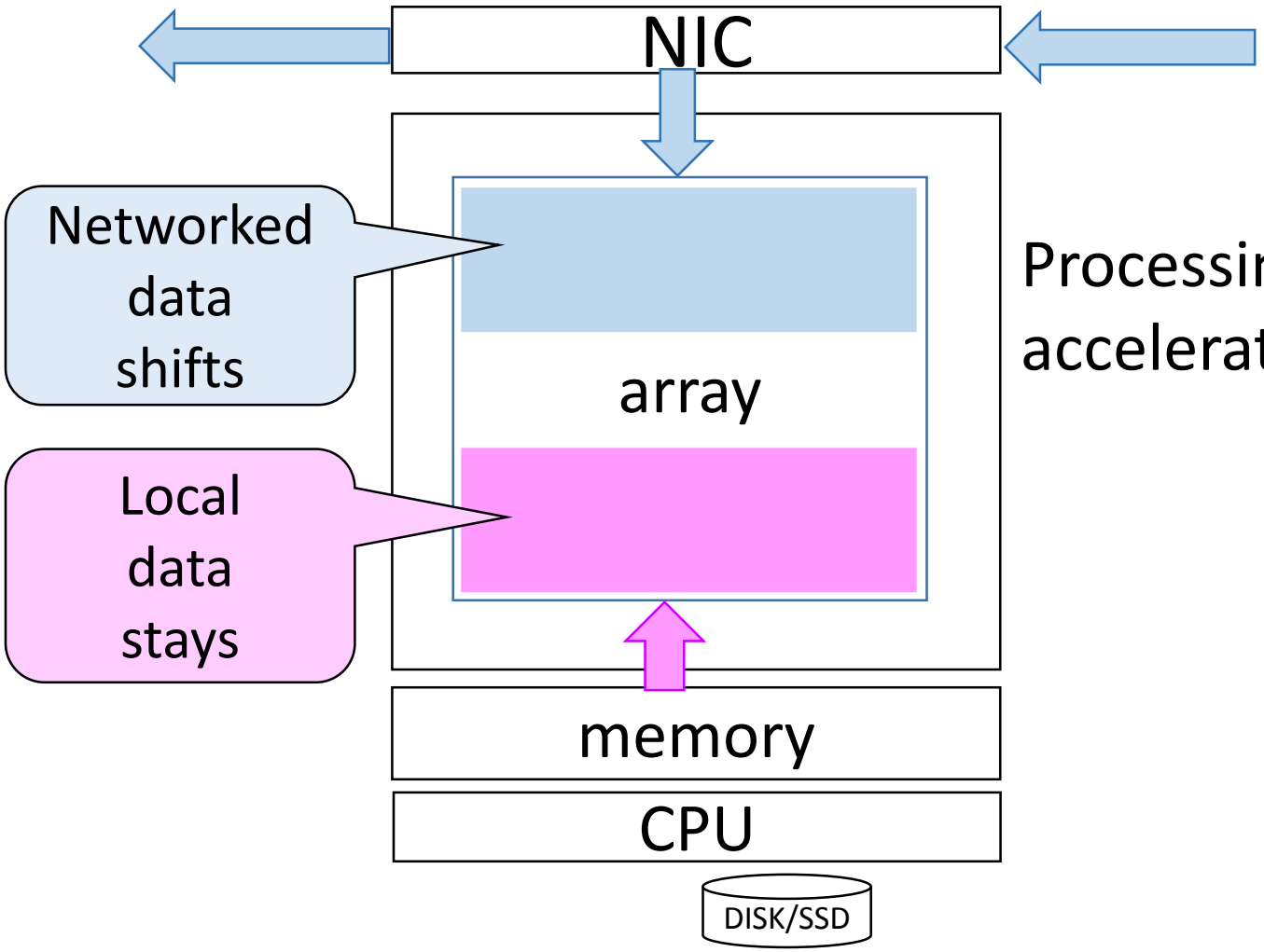
Energy

- Compute:
  - One SPFP: 20 pJoule
  - ExaFLOP: 20 M-Joule
  - Take 0.2 sec on 100 MWatt data center
- Move the data only once within a data center:
  - One bit, Chip-to-chip: 50 pJoule
  - One ExaByte, within one data center: 50 Tera-Joule
  - Take **6 days** on 100 MWatt data center

The BDML data center



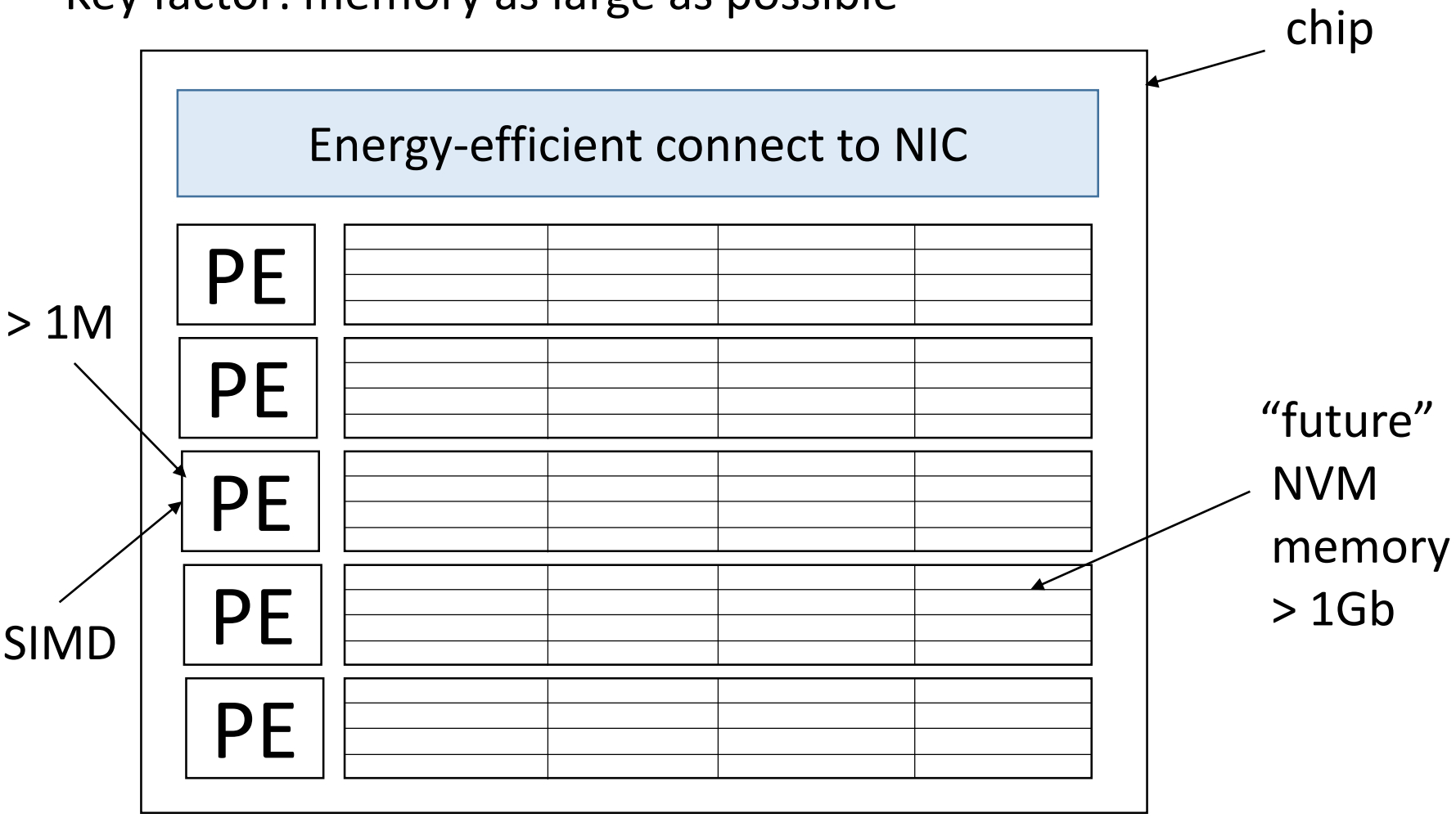
The BDML server node



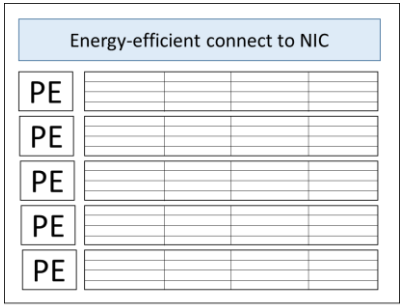
Processing-in-memory accelerator

The BDML accelerator

Key factor: memory as large as possible



The BDML accelerator



- Studied memristor-based architecture with 100Mb
  - Goal is 10 Gb. Need 3D integration
- Simulated sparse matrix-matrix multiplication
- Showed 10 GFLOPS/watt
  - CPU, GPU are less than 1 GFLOPS/watt
  - Goal is 1 TFLOPS/watt

Summary

- Rather than bring the accelerator close to memory, accelerator IS the memory
- Lots of distant data must be brought to it, to be compared with local data. Should minimize energy for data moves
- Accelerators with larger memory help minimize overall energy of BDML