Cheaper, Faster Computing

with

hardware accelerators and NVM storage

Sang-Woo Jun

Assistant Professor

Department of Computer Science

University of California, Irvine

UCI

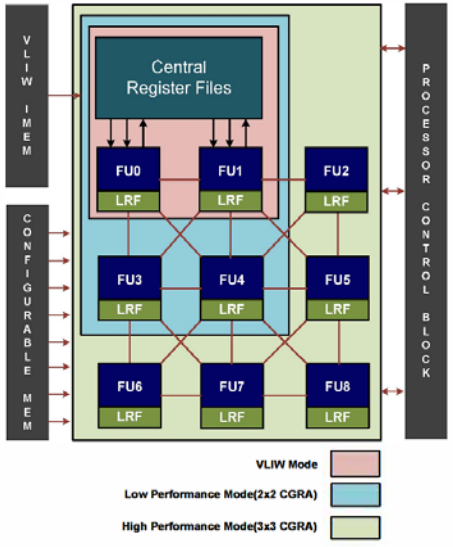
2018-10-05

About Me

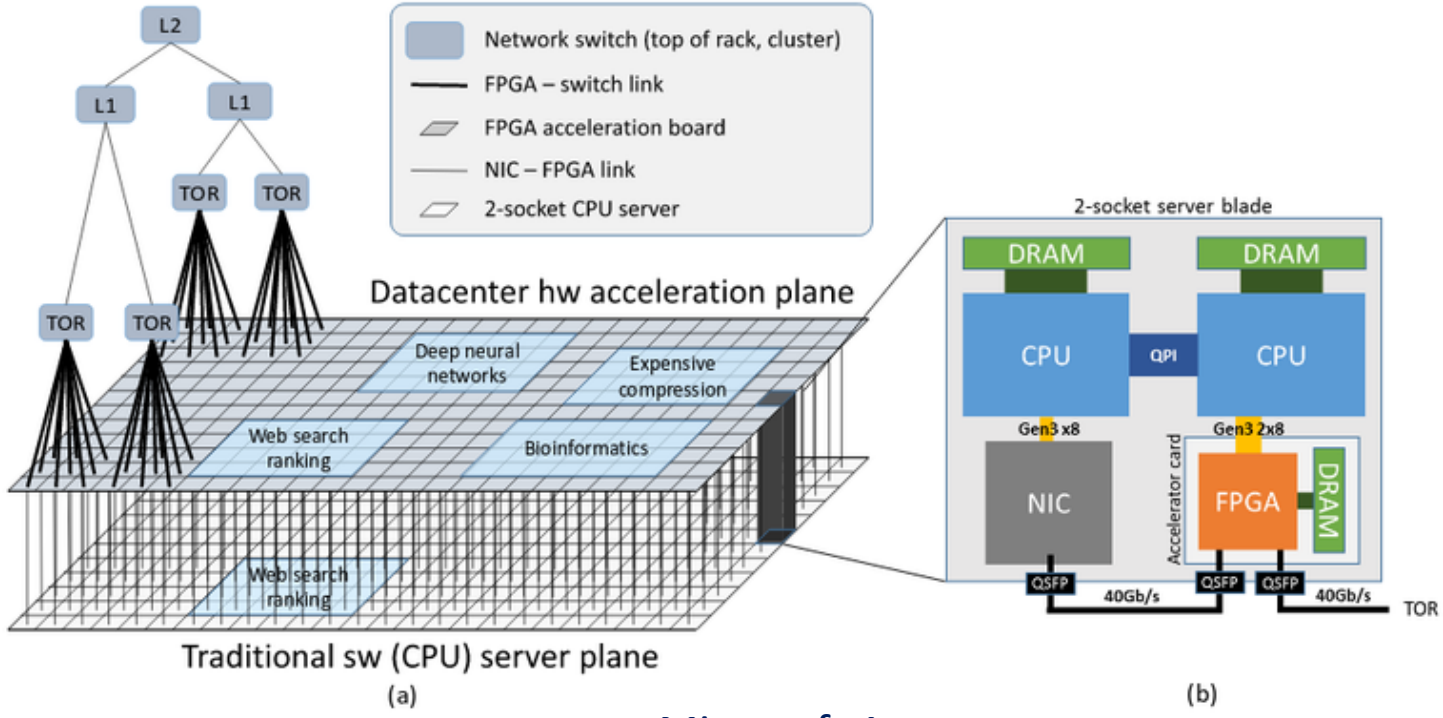
- Sang-Woo Jun
- Ph.D. (2018) @ MIT
- Research Interests
  - Systems architecture
  - Accelerators
  - NVM storage
  - Applications!
    - Graphs, Bioinformatics, Machine learning…
- Some Nice Papers
  - (ISCA, VLDB, FAST, FPGA, …)
- Some Nice Media Coverage
  - Engadget, The Next Platform, …

Exciting Time to Be a Compute Architect

Google TPU

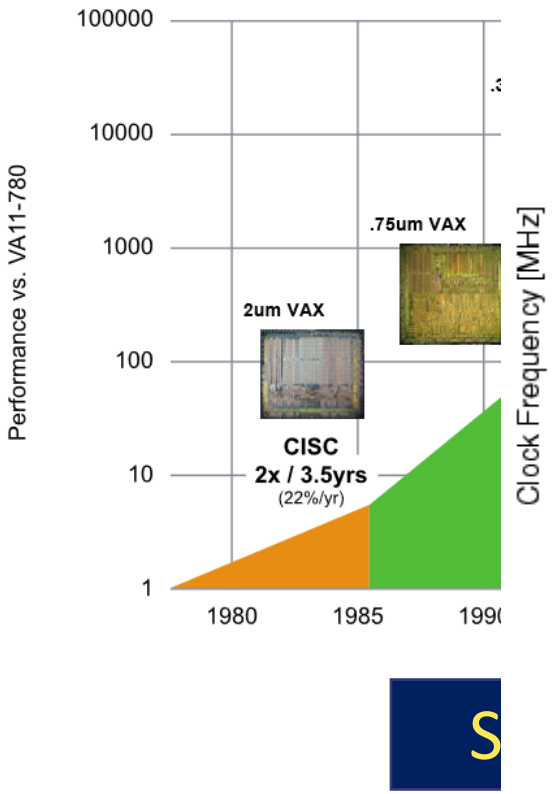


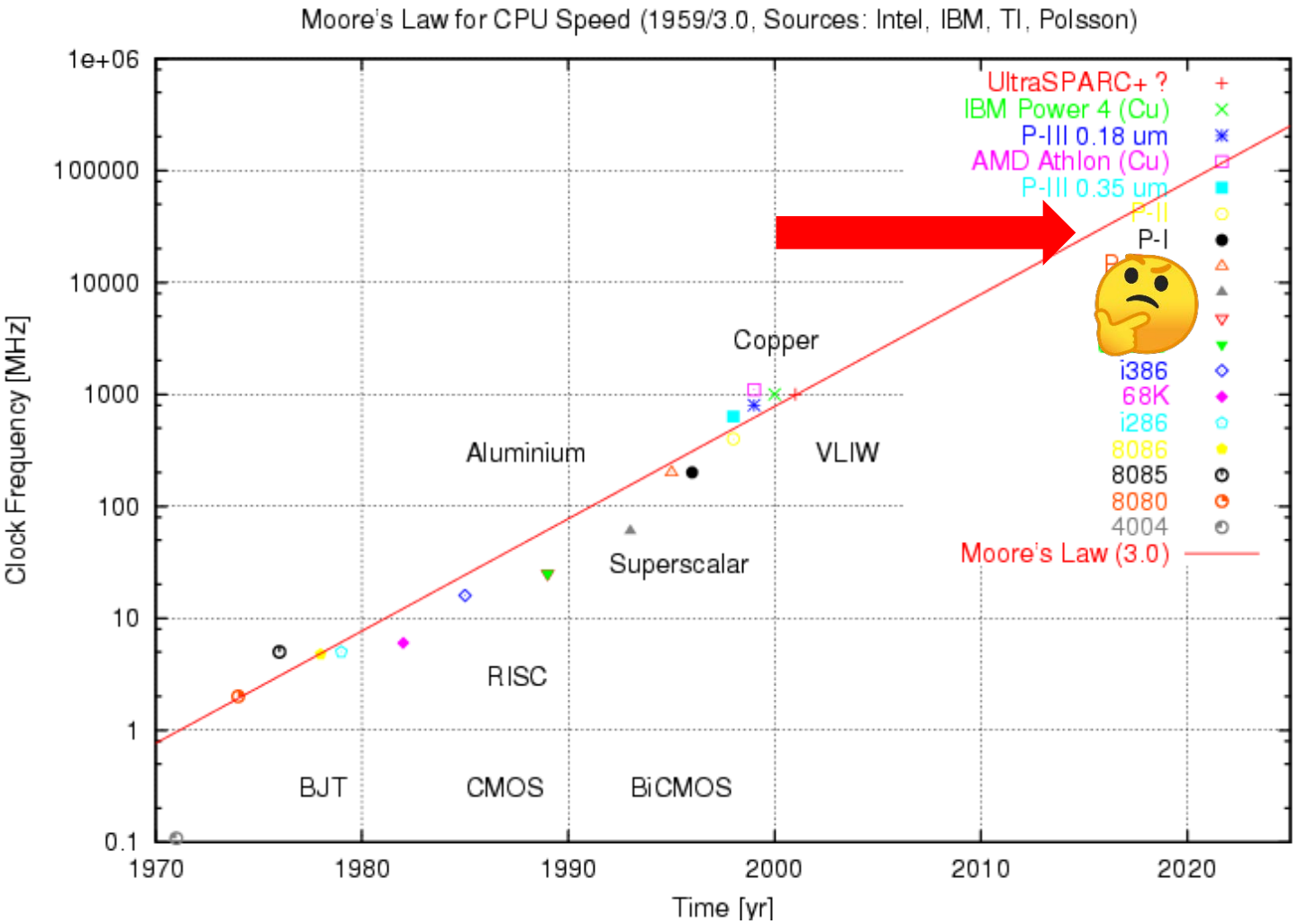
Samsung Reconfigurable Processor

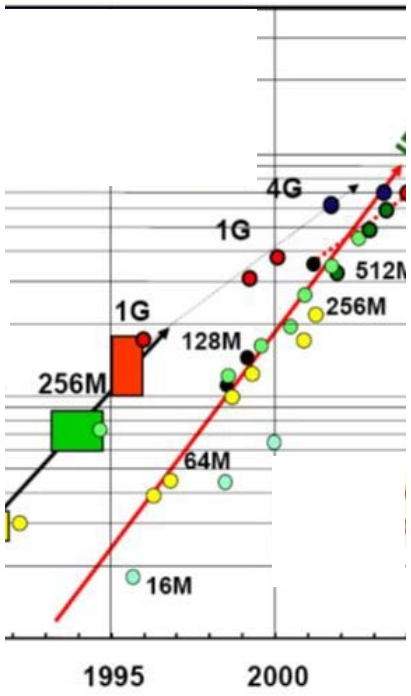


Microsoft Azure

A Computer – Some History



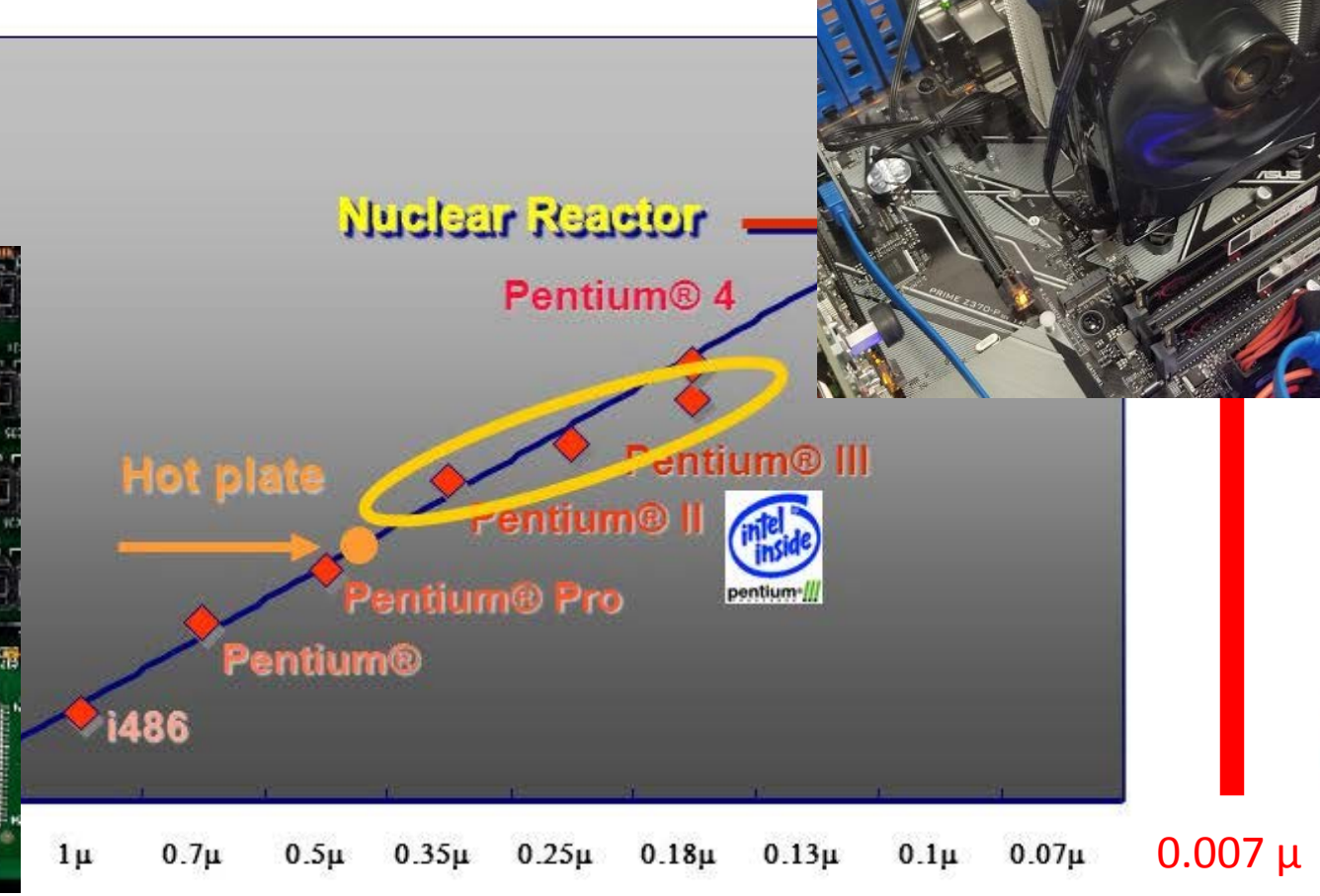




John Hennessy and David Patterson, "Computer Architecture: A Quantitative Approach", 2018 (Cropped)

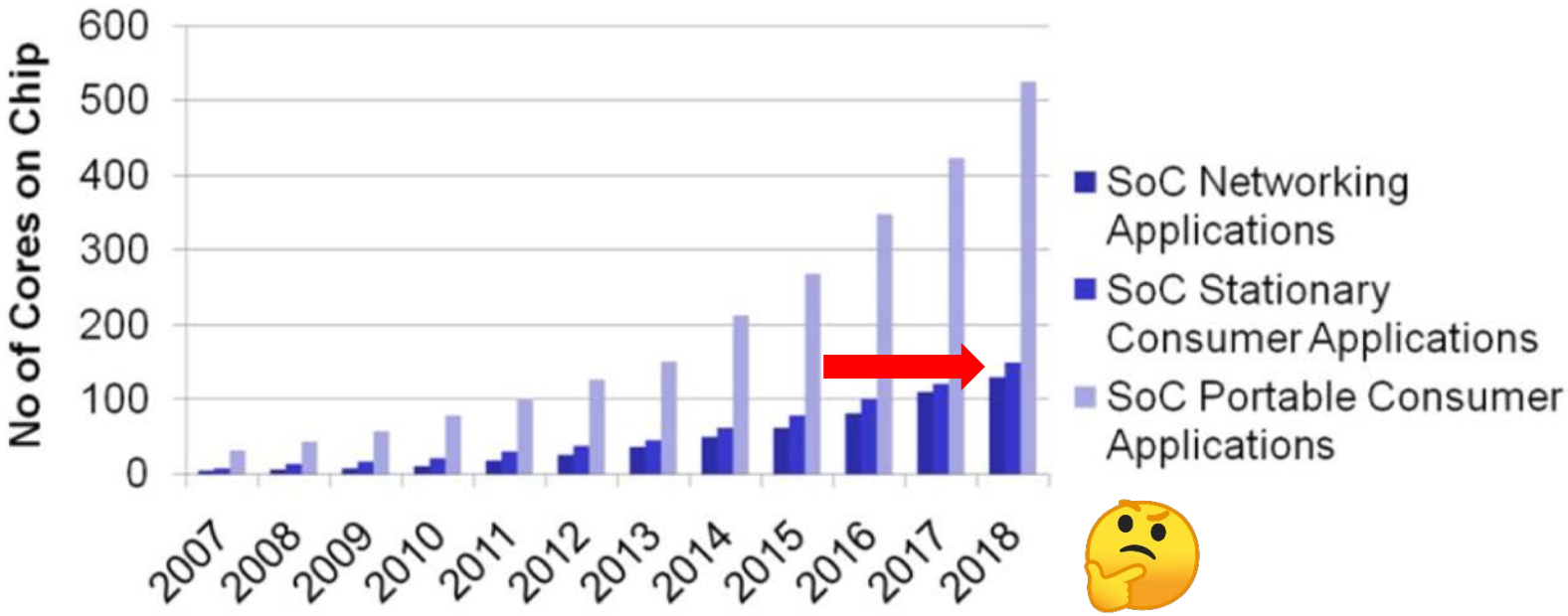
Bon-jae Koo, "Understanding of semiconductor memory architecture", 2007 (Cropped)

Running Into the Power Wall

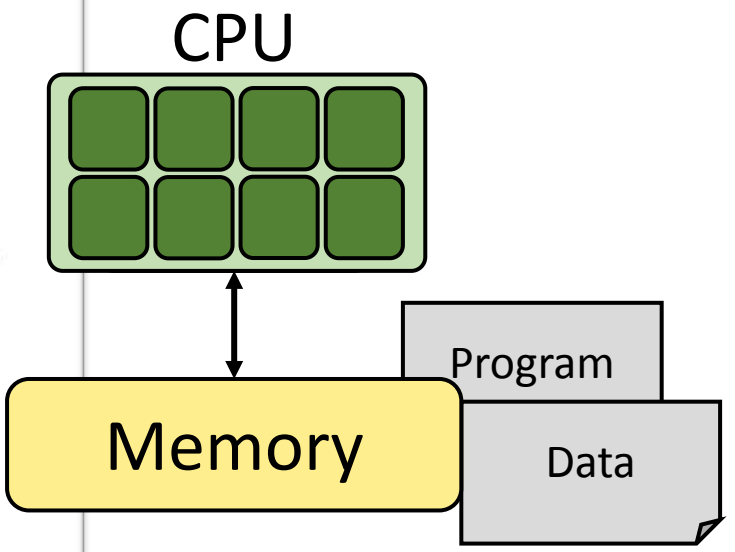


* "New Microarchitecture Challenges in the Coming Generations of CMOS Process Technologies" – Fred Pollack, Intel Corp. Micro32 conference key note - 1999.

Crisis Averted With Manycores?

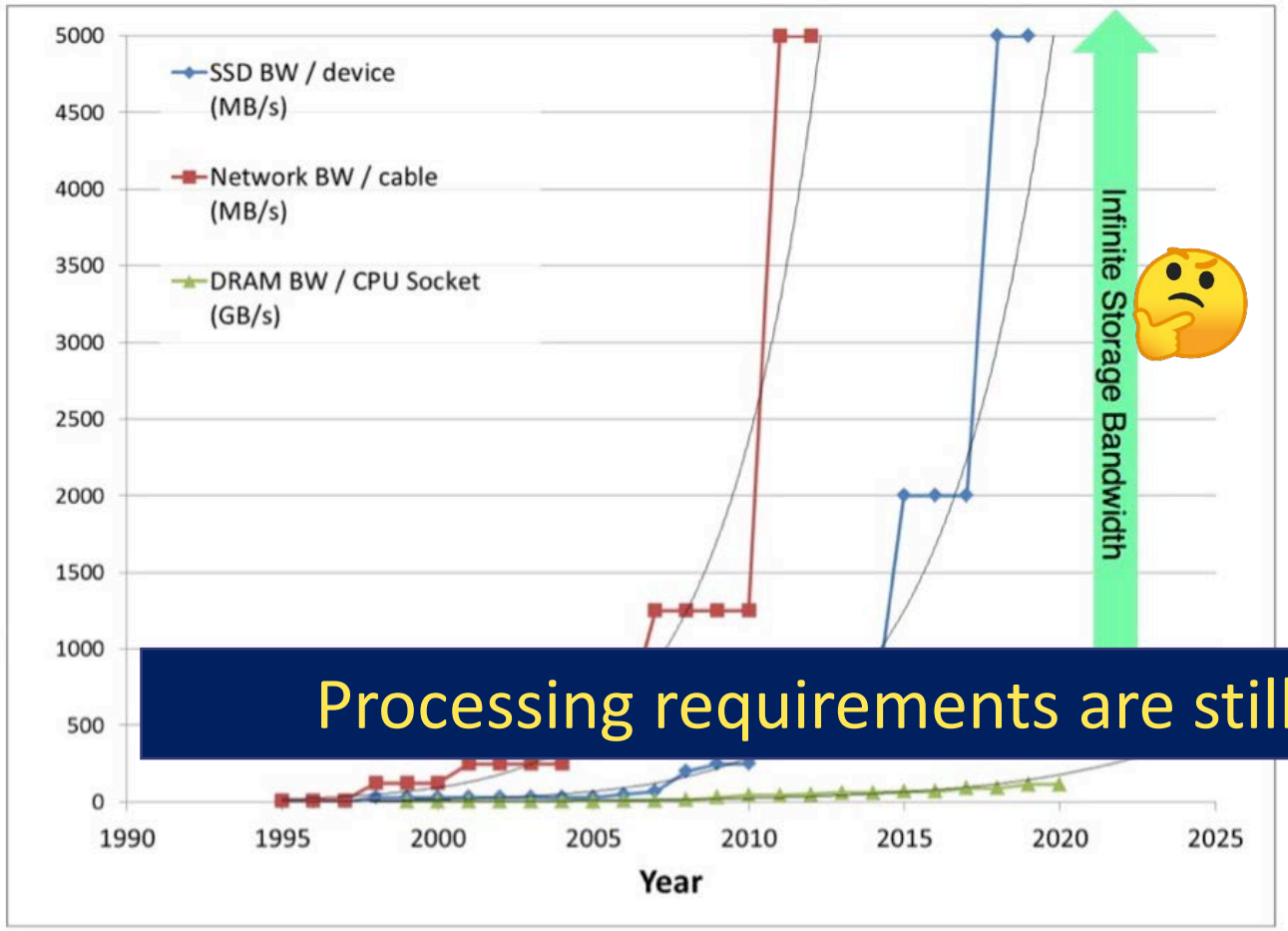


Source:
International Roadmap for Semiconductors 2007 edition (http://www.itrs.net/)



Bernd Hoefflinger, “ITRS 2028—International Roadmap of Semiconductors”, 2015

Memory/Storage Worries Too!



"[…] per gigabit (Gb) has declined from **$11** in 2006 to less than **$1** [in 2013]"

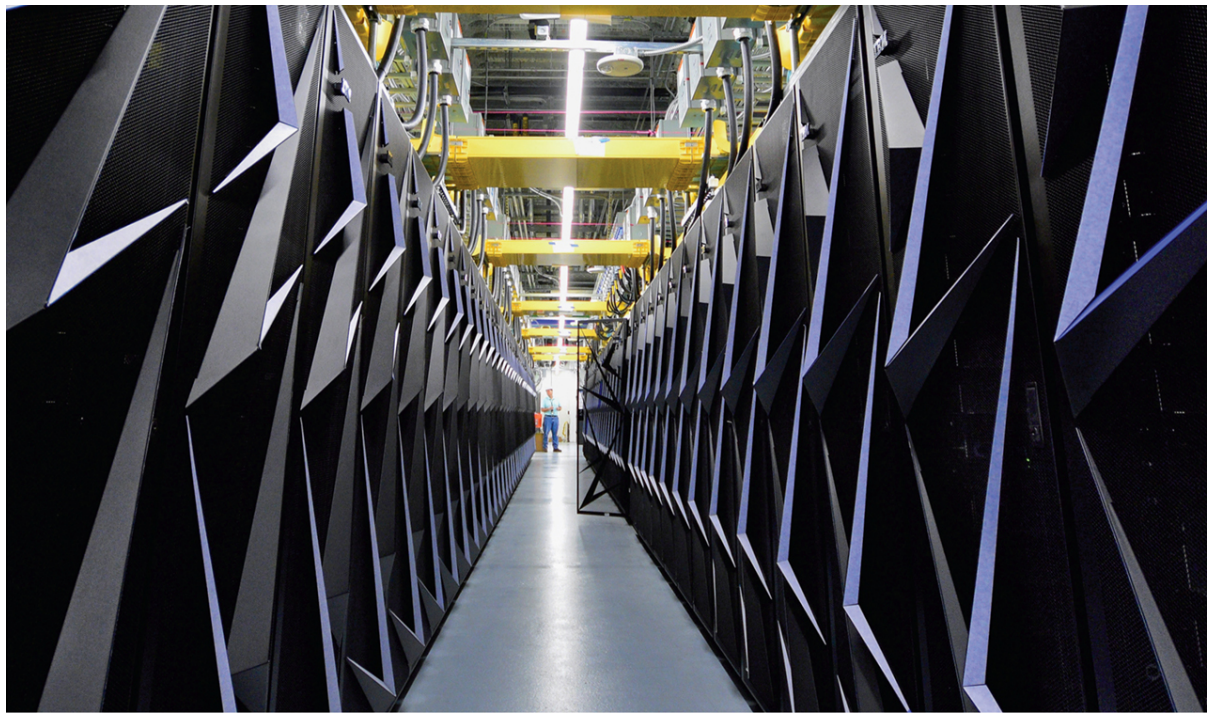
We are still around **$0.5 - $1/Gb** as of 2018

Processing requirements are still increasing exponentially!

Western Digital, "CPU Bandwidth – The Worrisome 2020 Trend", 2016

The Exascale Challenge

Department of Energy requests an exaflop machine by 2020



1,000,000,000,000,000,000 floating point operations per second

Using 2016 technology, 200 MW

MIT Research nuclear reactor

6 MW

Lynn Freeny, Department of Energy

Smaller Challenges Near Us

Smartphones

IoT Devices

AI Assistants

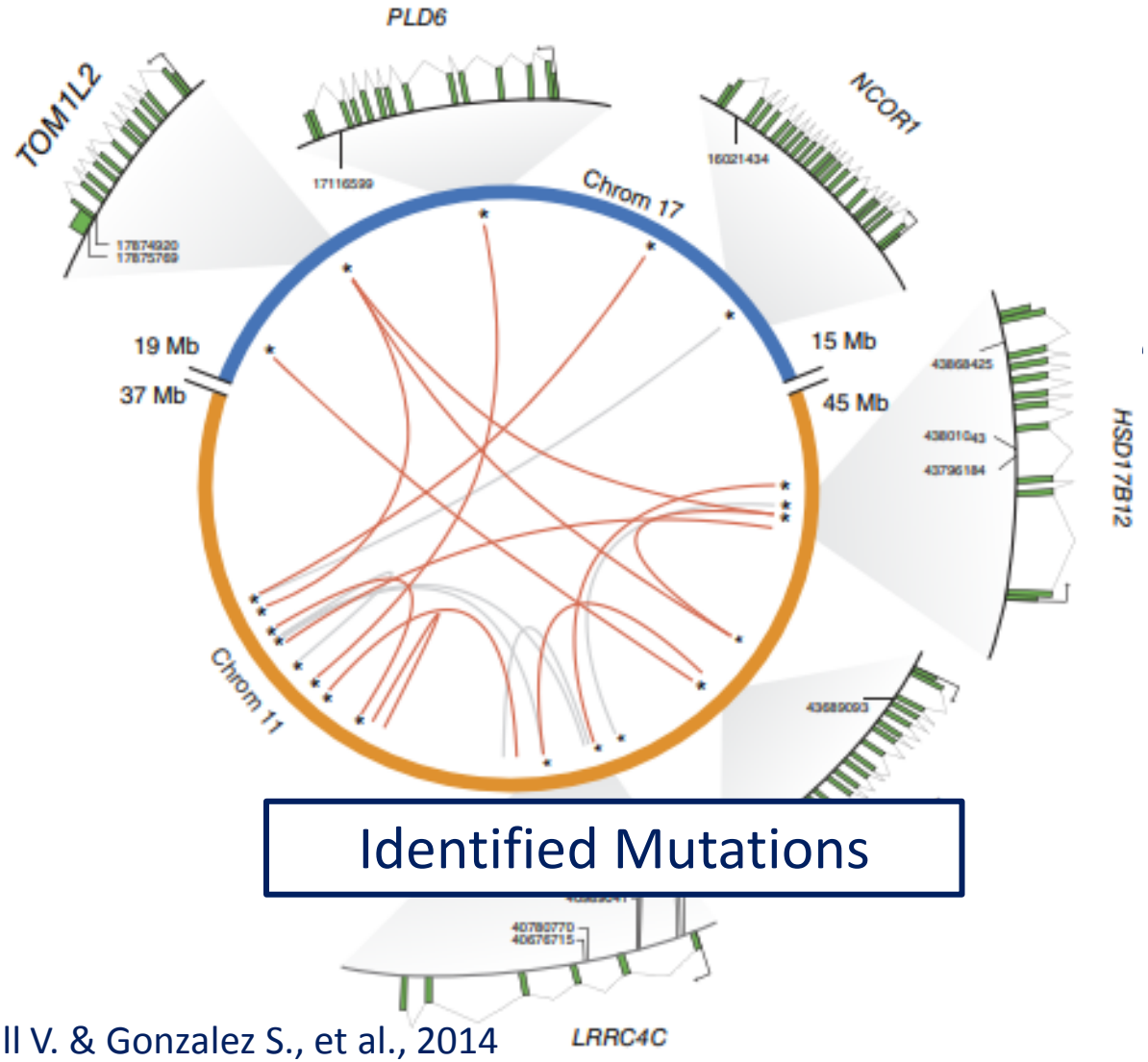
No Better Time to Be an Architect!

“There are Turing Awards waiting to be picked up
if people would just work on these things.”
—David Patterson, 2018

Photo: Peg Skorpinski,UC Berkeley

A Big Data Application: Personalized Genome

Cancer Patient



Identified Mutations

"Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads," Moncunill V. & Gonzalez S., et al., 2014

Cluster System for Personalized Genome

$100,000

7,000 Watts

A Cheaper Alternative Using Hardware-Accelerated SSD

+

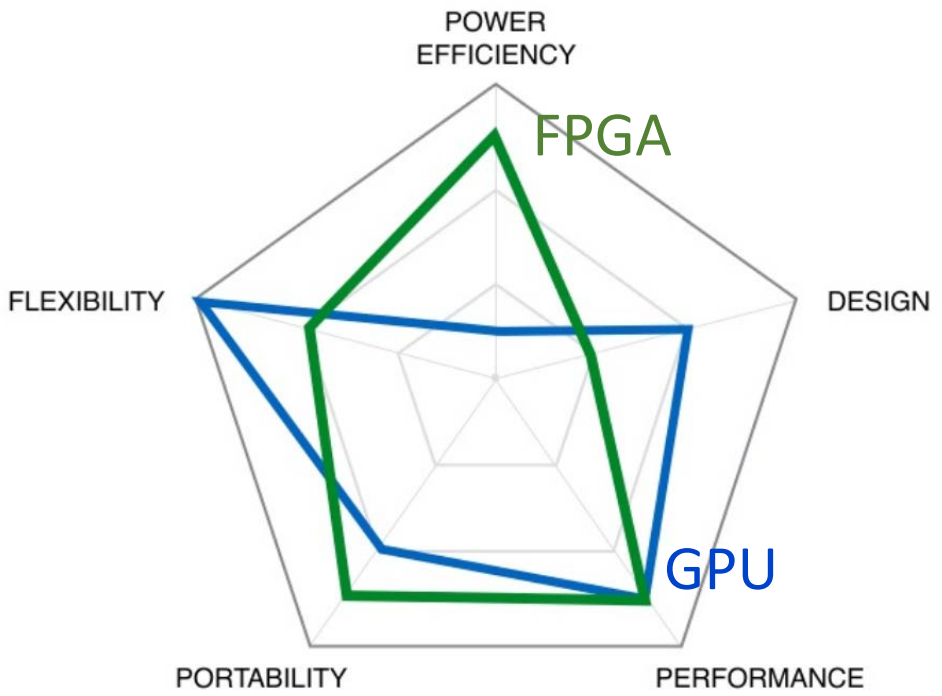
+

$2,000

80 Watts

Reconfigurable Hardware Acceleration

Field Programmable Gate Array (FPGA)



Program application-specific hardware

High performance, Low power

Reconfigurable to fit the application

amazon web services EC2

Azure

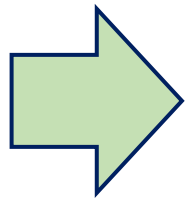
intel CYCLONE 10 inside

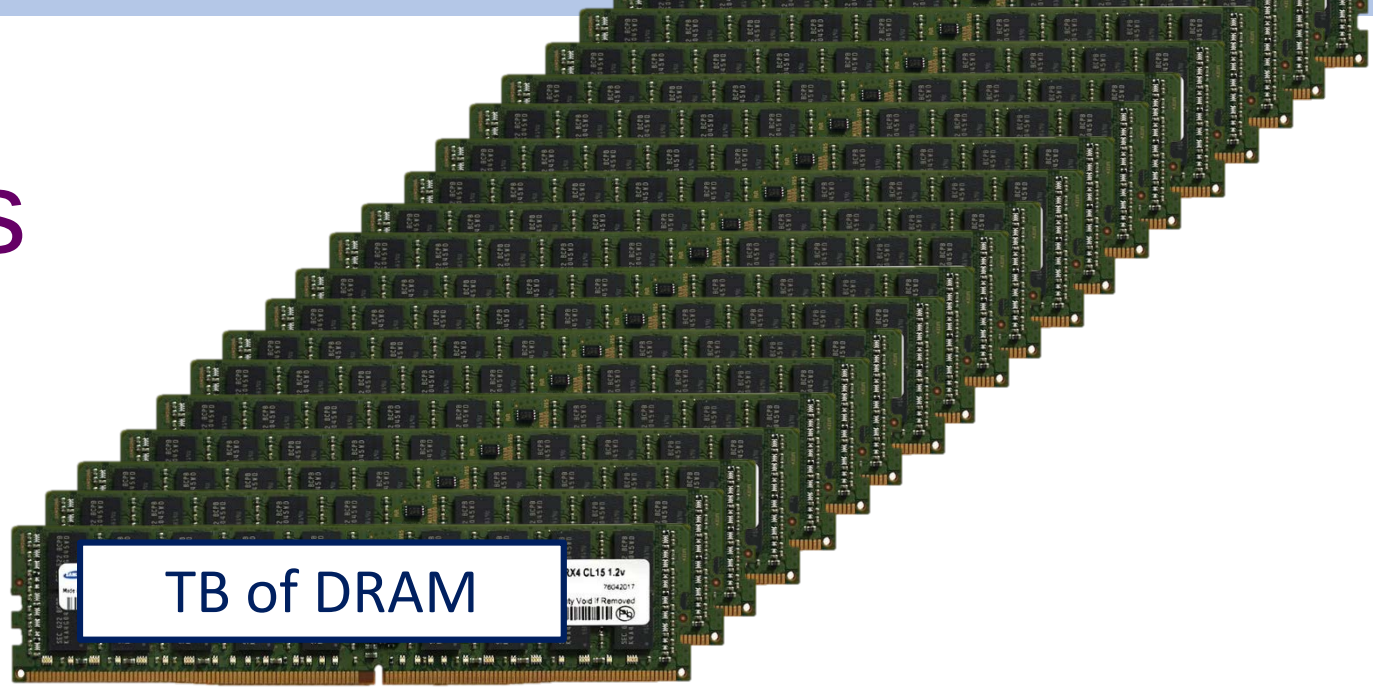
Bracco Filippo, “Rationale behind FPGA”, 2017

Storage for Analytics

Fine-grained,
Irregular access

Terabytes in size



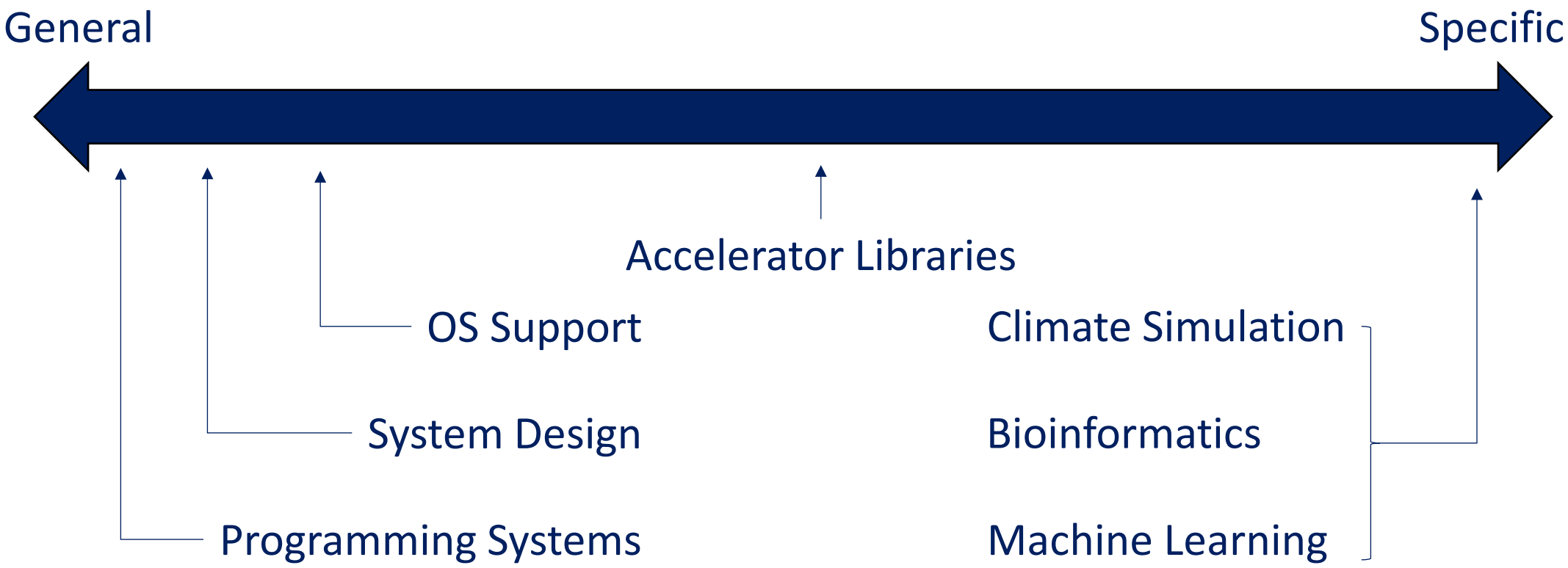


$$$ $8000/TB, 200W

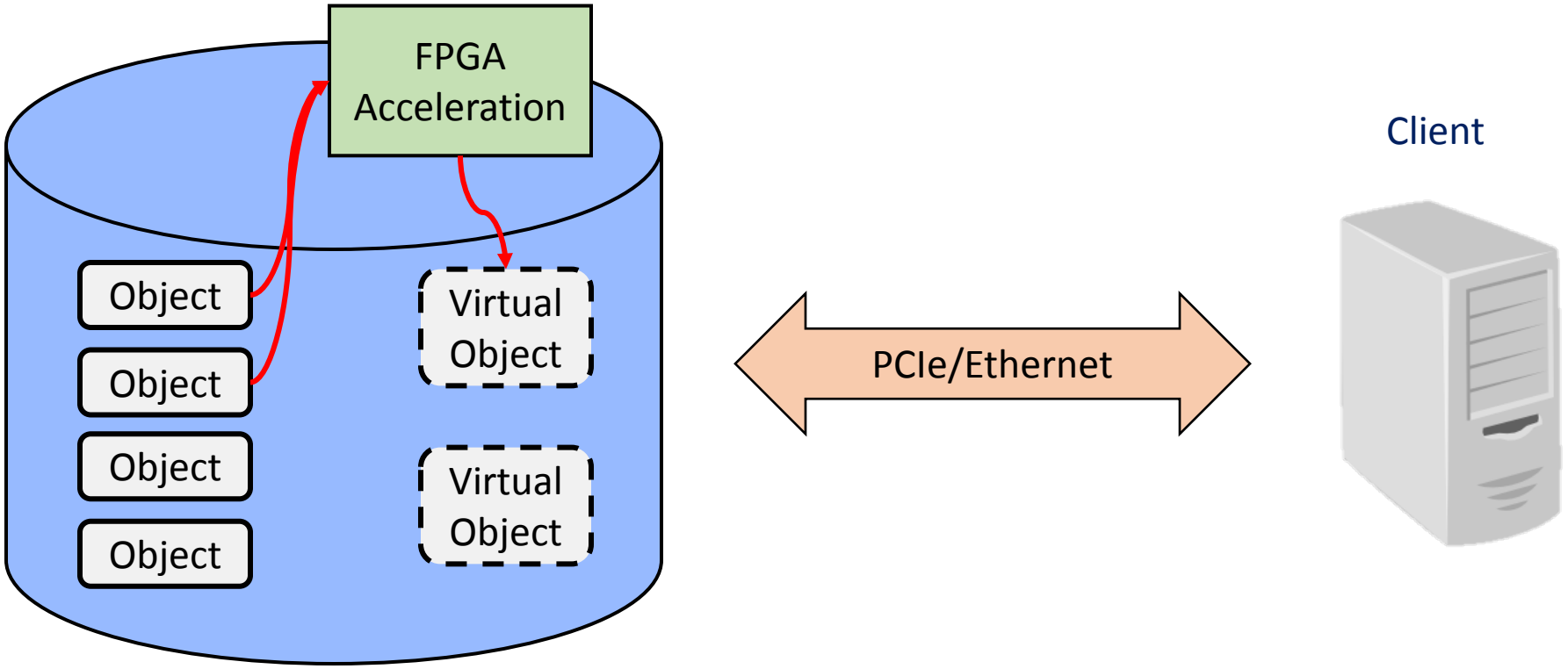
Our goal:

$ $500/TB, 10W

Research Topics Galore

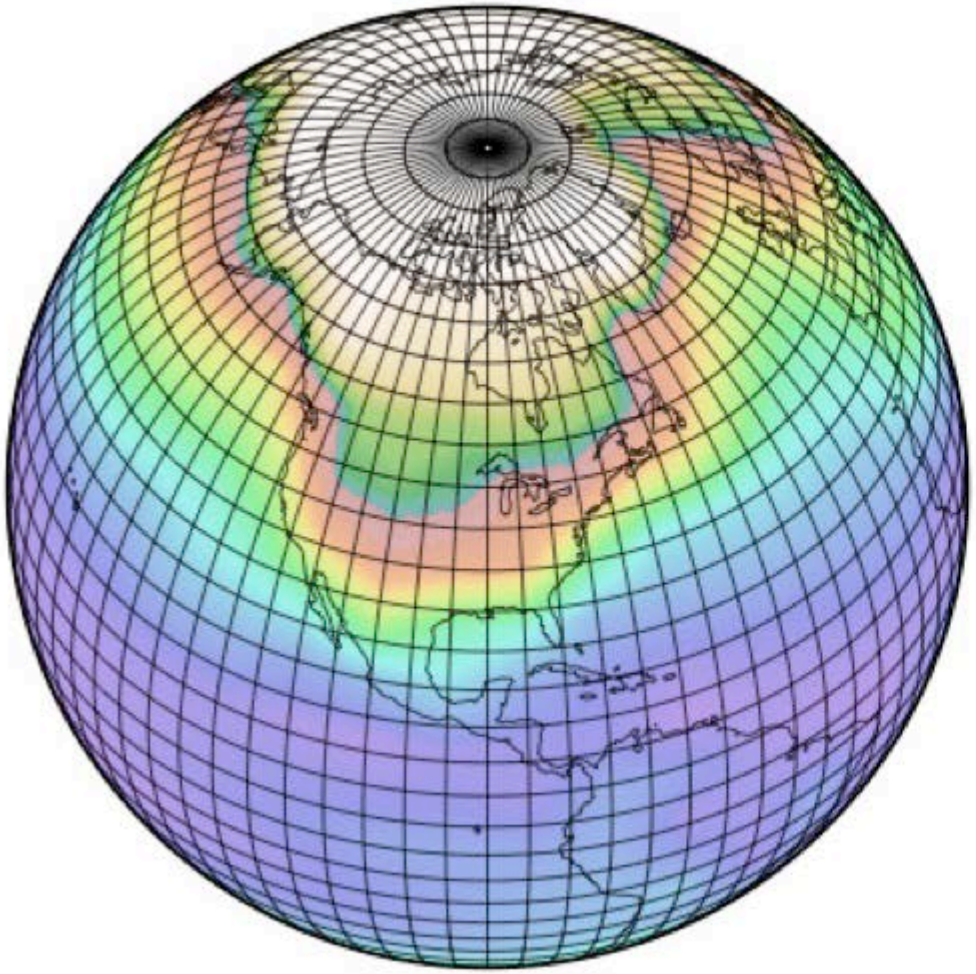


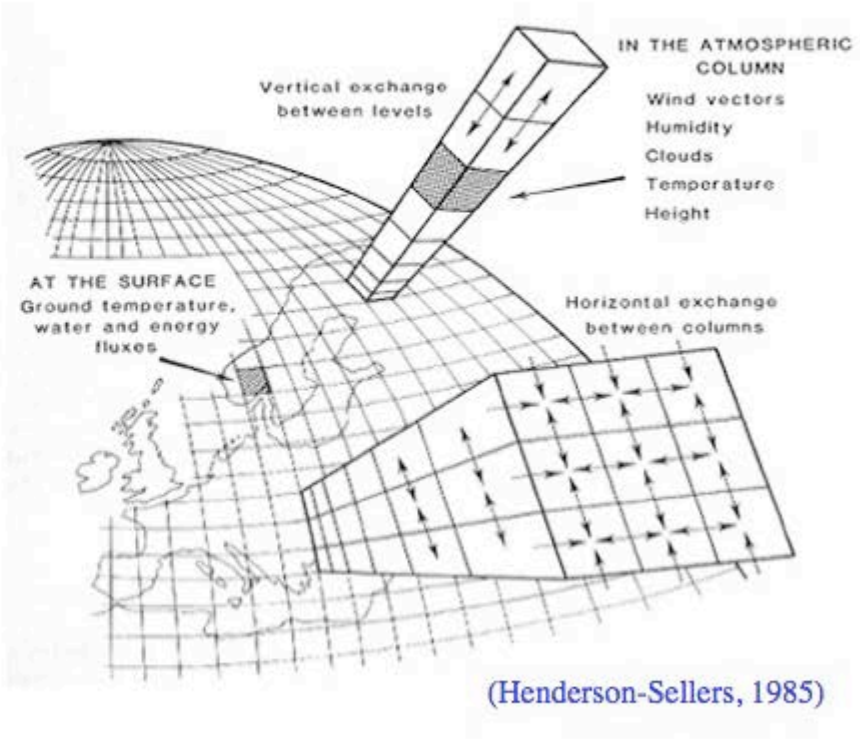
Project: Accelerated Object Storage



- Storage exposes high-level object store abstraction to software
- Computation offloaded to accelerator using “virtual objects”, not breaking object store abstraction

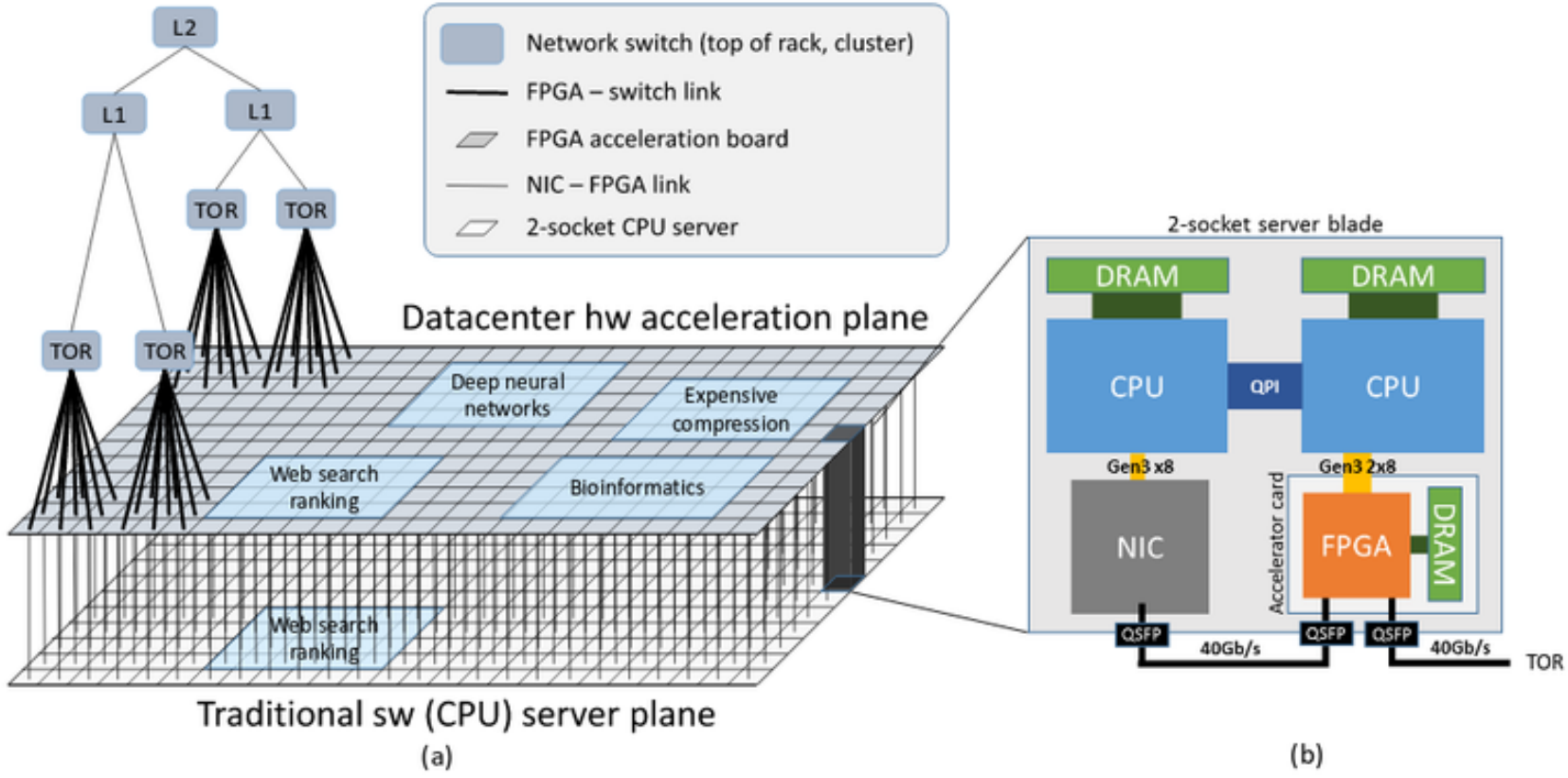
Project: Accelerating Stencil Computation for Climate Simulation





(Henderson-Sellers, 1985)

Project: Distributed FPGA Cluster



Project:
Applications For Accelerator Platform

- Platform for efficient fine-grained acceleration
- Goal: 10x performance against baseline
- Claim: Easy to develop!

- Candidate applications: Dynamic Time Warping, Smith-Waterman, Cosine Similarity, N-body simulation, …

Ideas?

Things To Come!

Thank you!