CS294: RISE
Logistics, Overview, Trends

Joey Gonzalez, Joe Hellerstein, Raluca Popa, Ion Stoica

August 29, 2016

Instructors

Joseph E. Gonzalez
jegonzal@cs.berkeley.edu

Joseph Hellerstein
hellerstein@cs.berkeley.edu

Raluca Ada Popa
raluca.popa@cs.berkeley.edu

Ion Stoica
istoica@cs.berkeley.edu

Goal of this Class

Bootstrap RISE research agenda

- Start new projects or work on existing ones

Read related work in the areas relevant to RISE Lab

- ML, Security, Systems/Databases, Architecture

Allow people from one area learn about state-of-the-art research in other areas → key to success in an interdisciplinary effort

Course Information

Course website is:

- https://ucbrise.github.io/cs294-rise-fa16/
  - It is on Github so you can contribute content!
- We will be adding a few more updates today and tomorrow

We will be using Piazza for discussion about the class

- https://piazza.com/berkeley/fall2016/cs29420/home

Tentative Lecture Format (not today!)

First 1/3 of each lecture presented by faculty
- Second 2/3 covers papers presented by students

Reading assignments should be up several weeks in advance
- All students are required to read all papers

All students must answer short questions on google form
- Student will prepare 15 minute presentations on selected paper
- We will post on Piazza about how to signup later this week
- Address the questions in the form
- Identify key insights, strengths and weaknesses, and implications on RISE research agenda

Grading Policy

50% Class Participation

- Answer questions, join discussion, and present papers

10% Initial Project Proposal Presentation

- Presented in class on 10/17

20% Final Project Presentation

- During class final exam 12/12

20% Final Project Report

- Emailed to instructors 12/16 by 11:59 PM

Rest of This Talk

Reflect on how

- Application trends (i.e., user needs & requirements)
- Hardware trends

have impacted the design of our solution

How we can use these lessons to design new systems in the context of RISE Lab

The Past and The Lessons

2009: State-of-the-art in Big Data

Hadoop

- Large scale, flexible data processing engine
- Fault tolerant
- Batch computation (e.g., 10s minutes to hours)

Getting rapid industry traction:

- High profile users: Facebook, Twitter, Yahoo!, …
- Distributions: Cloudera, Hortonworks
- Many companies still in austerity mode

2009: Application Trends

Interactive computations, e.g., ad-hoc analytics

- SQL engines like Hive and Pig drove this trend

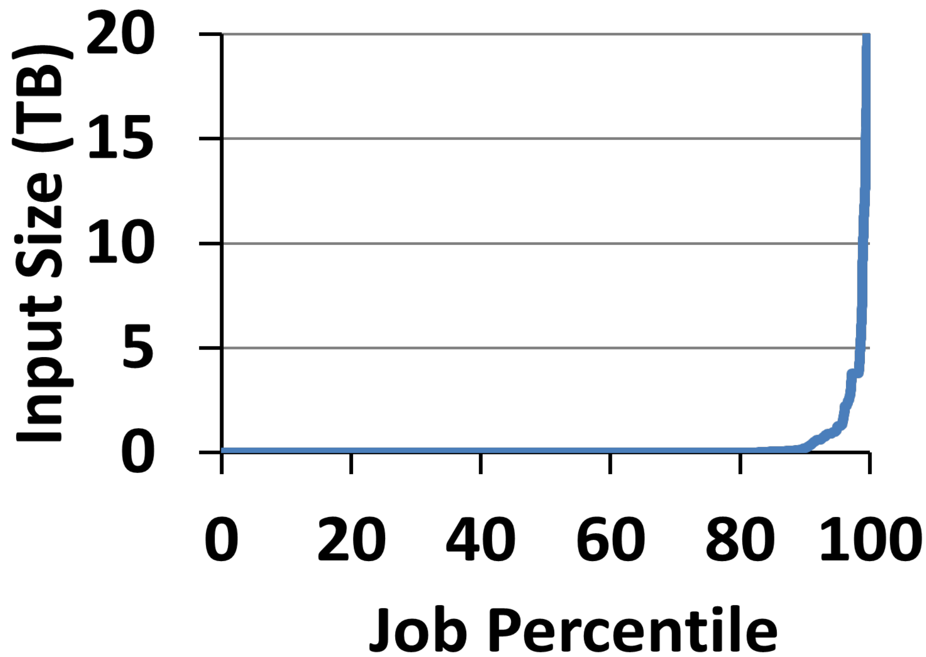
Iterative computations, e.g., Machine Learning

- More and more people aiming to get insights from data

2009: Application Trends

Despite huge amounts of data, many working sets in big data clusters fit in memory

Inputs of 96% of Facebook jobs fit in memory*



*G Ananthanarayanan, A. Ghodsi, S. Shenker, I. Stoica, "Disk-Locality in Datacenter Computing Considered Irrelevant", HotOS 2011

2009: Application Trends

Memory (GB)	Facebook (% jobs)	Microsoft (% jobs)	Yahoo! (% jobs)
8	69	38	66
16	74	51	81
32	96	82	97.5
64	97	98	99.5
128	98.8	99.4	99.8
192	99.5	100	100
256	99.6	100	100

*G Ananthanarayanan, A. Ghodsi, S. Shenker, I. Stoica, ”Disk-Locality in Datacenter Computing Considered Irrelevant”, HotOS 2011

2009: Application Trends

Memory (GB)	Facebook (% jobs)	Microsoft (% jobs)	Yahoo! (% jobs)
8	69	38	66
16	74	51	81
32	96	82	97.5
64	97	98	99.5
128	98.8	99.4	99.8
192	99.5	100	100
256	99.6	100	100

*G Ananthanarayanan, A. Ghodsi, S. Shenker, I. Stoica, "Disk-Locality in Datacenter Computing Considered Irrelevant", HotOS 2011

2009: Hardware Trends

Memory still growing with Moore's law

I/O throughput and latency stagnant

- HDD dominating data clusters as storage of choice

2009: Trends Summary

Users require interactivity and support for iterative apps

Majority of working sets of many workloads fit in memory

Memory capacity still growing fast, while I/O stagnant

2009: Our Solution: Apache Spark

In-memory processing

Generalizes MapReduce to multi-stage computations

- Fully implements BSP model

2009: Challenges & Solutions

Low-overhead resilience mechanisms →

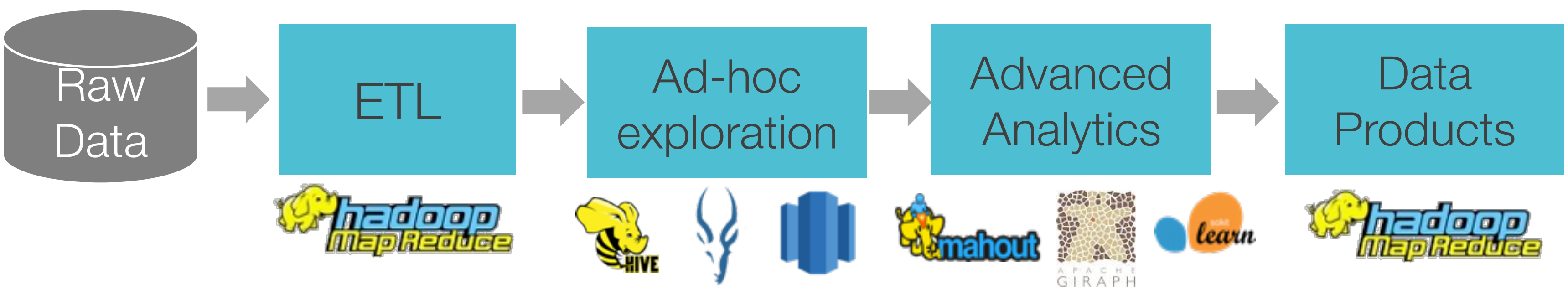
- Resilient Distributed Datasets (RDDs)

Efficiently support for ML algos →

- Share data between stages via memory
- Powerful and flexible APIs: map/reduce just two of over 80+ APIs

2012: Application Trends

People started to assemble e2e data analytics pipelines



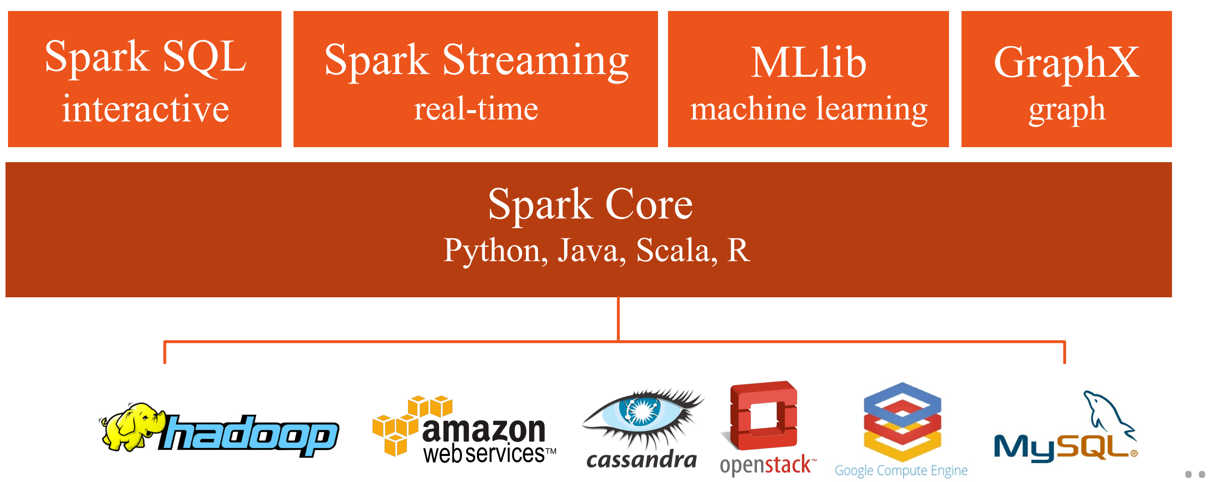
Need to stitch together a hodgepodge of systems

2012: Our Solution: Unified Platform

Support a variety of workloads

Support a variety of input sources

Provide a variety of language bindings



2015: Application Trends

New users, new requirements

Spark early adopters

hadoop
Users

Understands
MapReduce
& functional APIs

Data Engineers
Data Scientists
Statisticians
R users
PyData …

2015: Hardware Trends

Memory capacity continue to grow with Moore's law

Many clusters and datacenters transitioning to SSDs

- DigitalOcean: SSD only instances since 2013

CPU growth slowing down → becoming the bottleneck

2015: Our Solution

Move to schema-based data abstractions, e.g., DataFrames

- Familiar to data scientists, e.g., R and Python/pandas
- Allows us to in-memory store data in binary format
  - Much lower overhead
  - Alleviates/Avoids JVM's garbage collection overhead

Project Tungsten

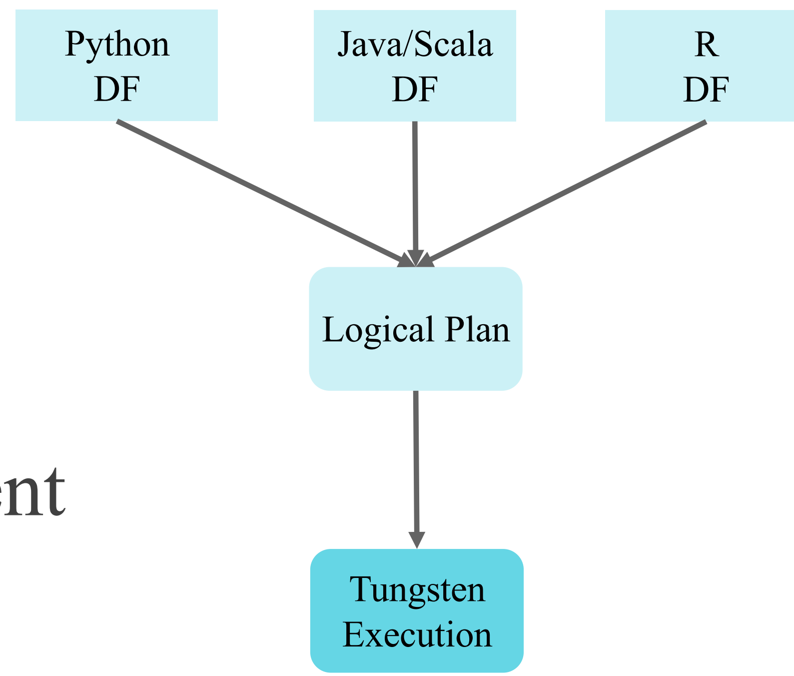
2015: Project Tungsten

Substantially speed up execution by optimizing CPU efficiency, via:

(1) Runtime code generation

(2) Exploiting cache locality

(3) Off-heap memory management



What’s Next for RISE Lab?

Overview

Application trends

Hardware trends

Challenges and techniques

Application Trends

Data only as valuable as the decisions and actions it enables

What does it mean?

- Faster decisions better than slower decisions
- Decisions on fresh data better than on stale data
- Decisions on personal data better than on aggregate data

Application Trends

Real-time decisions

decide in ms

on live data

with strong security

Application Trends

Real-time decisions

decide in ms

on live data

the current state of the environment

with strong security

Application Trends

Real-time decisions

decide in ms

on live data

the current state of the environment

with strong security

privacy, confidentiality, integrity

Applications	Quality	Latency		Security
		Decision	Update	
Zero-time defense	sophisticated, accurate, robust	sec	sec	privacy, integrity
Parking assistant	sophisticated, robust	sec	sec	privacy
Disease discovery	sophisticated, accurate	sec/min	hours	privacy, integrity
IoT (smart buildings)	sophisticated, robust	sec	min/hour	privacy, integrity
Earthquake warning	sophisticated, accurate, robust	ms	min	integrity
Chip manufacturing	sophisticated, accurate, robust	sec/min	min	confidentiality, integrity
Fraud detection	sophisticated, accurate	ms	min	privacy, integrity
"Fleet" driving	sophisticated, accurate, robust	sec	sec	privacy, integrity

Addressing these challenges, the goal of next Berkeley lab:
RISE (Real-time Secure Execution) Lab

Research areas

Systems: parallel computation engines providing msec latency and 10k-100K job throughput

Goal: develop Secure Real-time Decision Stack, an open source platform, tools and algorithms for real-time decisions on live data with strong security

Security: achieve privacy, confidentiality, and integrity without impacting performance

Overview

Application trends

Hardware trends

Challenges and techniques

Moore’s law is slowing down

MIT Technology Review

Topics+ Top Stories

Computing

Intel Puts the Brakes on Moore’s Law

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video

Archive > Volume 530 > Issue 7589 > News Feature > Article

NATURE | NEWS FEATURE

عربي

The chips are down for Moore’s law

The semiconductor industry will soon abandon its pursuit of Moore's law. Now things could get a lot more interesting.

E

TECHNOLOGY QUARTERLY
AFTER MOORE'S LAW

Double, double, toil and trouble

What does it mean?

CPUs affected most: only 15-20%/year perf. improvements

- More complex layouts, harder to scale
- Exploring these improvements hard → parallel programs

Memory: still grows at 30-40%/year

- Regular layouts, stacked technologies

Network: grows at 30-50%/year

- 100/200/400GBpE NICs at horizon
- Full-bisection bandwidth network topologies

CPUs is the bottleneck and it's getting worse!

What does it mean?

CPUs affected most: only 15-20%/year perf. improvements

- More complex layouts, harder to scale
- Exploring these improvements hard → parallel programs

Memory: still grows at 30-40%/year

- Regular layouts, stacked technologies

Network: grows at 30-50%/year

- 100/200/400GBpE NICs at horizon

Memory-to-core ratio increasing

e.g., AWS: 7-8GB/vcore → 17GB/vcore (X1)

Unprecedented hardware innovation

From CPU to specialized chips:

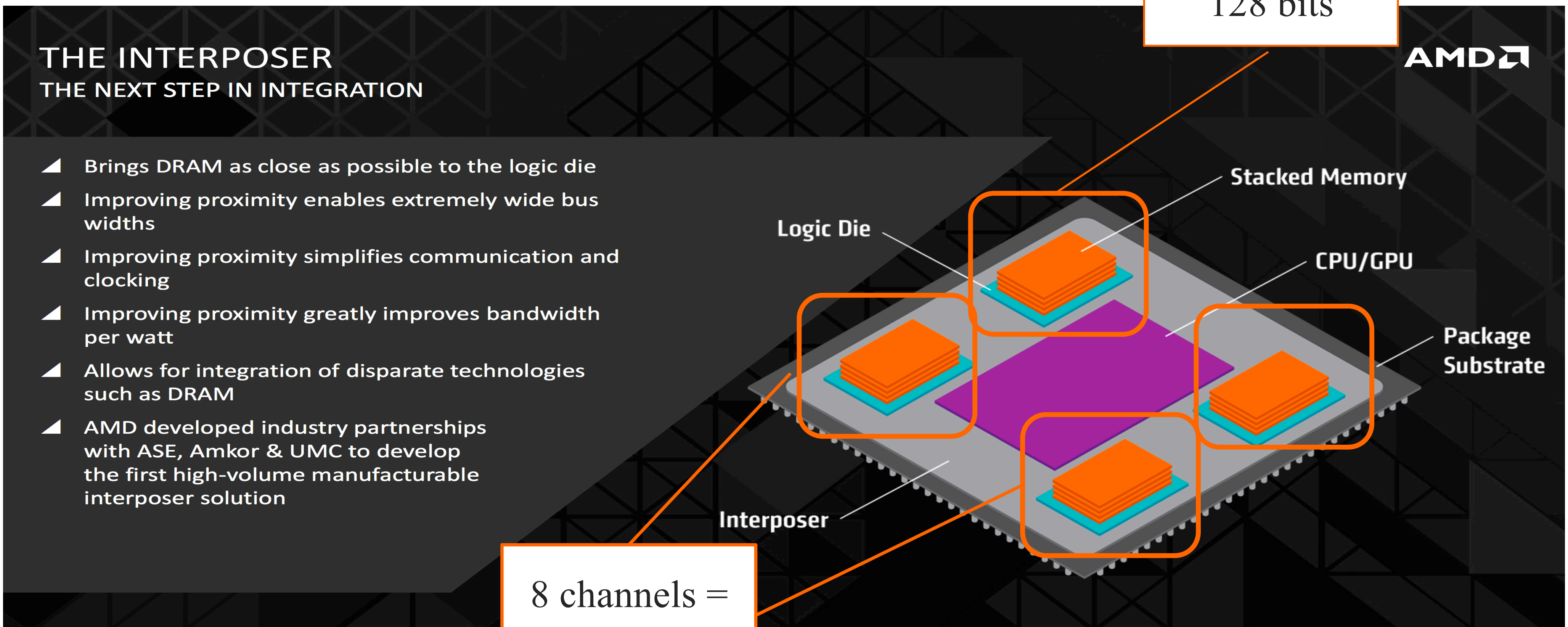
- GPUs, FPGAs, ASICs/co-processors (e.g., TPU)
- Tightly integrated (e.g., Intel's latest Xeon integrates CPU & FPGA)

New memory technologies

- HBM (High Bandwidth Memory)

High Bandwidth Memory (HBM)

2 channels @ 128 bits



8 channels = 1024 bits

High Bandwidth Memory (HBM)

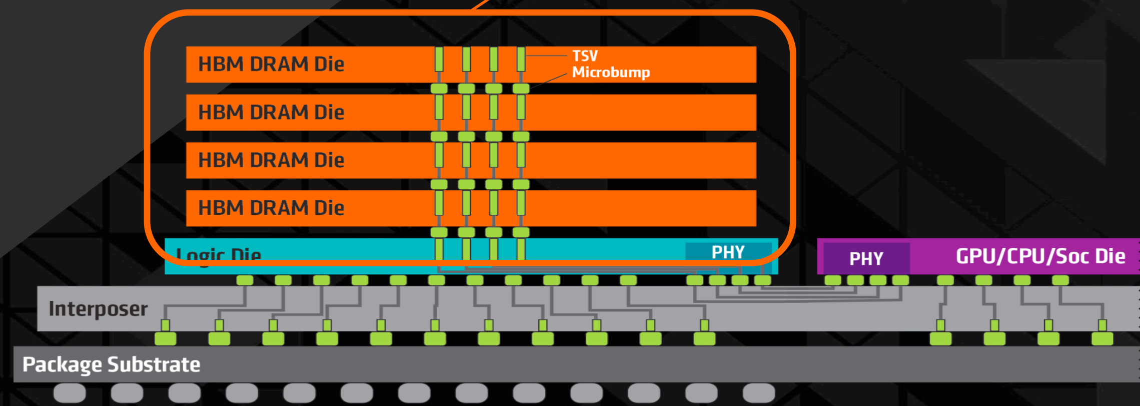
8 stacks =
4096 bits →
500 GB/sec

HIGH-BANDWIDTH MEMORY

DRAM BUILT FOR AN INTERPOSER

AMD

- A new type of memory chip with low power consumption and an ultra-wide bus width
- Many of those chips stacked vertically like floors in a skyscraper
- New interconnects, called "through-silicon vias" (TSVs) and "μbumps", connect one DRAM chip to the next
- TSVs and μbumps also used to connect the SoC/GPU to the interposer
- AMD and SK Hynix partnered to define and develop the first complete specification and prototype for HBM



Unprecedented hardware innovation

From CPU to specialized chips:

- GPUs, FPGAs, ASICs/co-processors (e.g., TPU)
- Tightly integrated (e.g., Intel's latest Xeon integrates CPU & FPGA)

New memory technologies

- HBM2: 8 DRAM chips/package → 1TB/sec

Unprecedented hardware innovation

From CPU to specialized chips:

- GPUs, FPGAs, ASICs/co-processors (e.g., TPU)
- Tightly integrated (e.g., Intel's latest Xeon integrates CPU & FPGA)

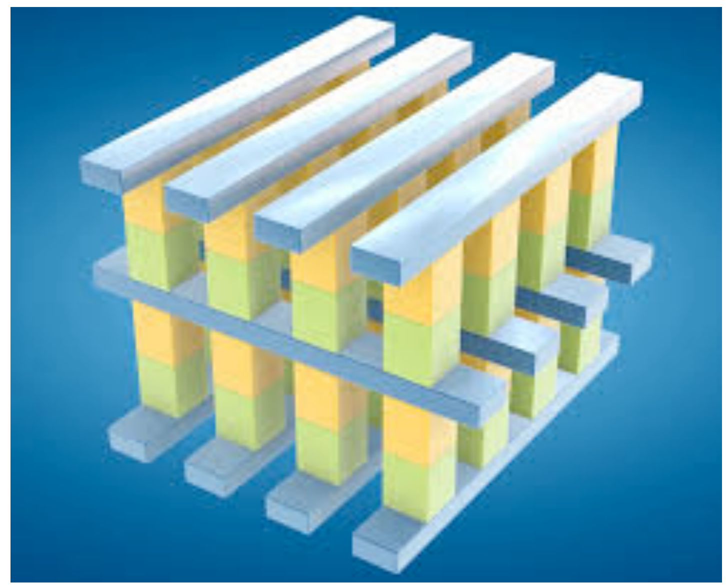
New memory technologies

- HBM2: 8 DRAM chips/package → 1TB/sec
- 3D XPoint

3D XPoint Technology

Developed by Intel and Micron

- Announced last year; products released this year



Characteristics:

- Non-volatile memory
- 2-5x DRAM latency!
- 8-10x density of DRAM
- 1000x more resilient than SSDs

Unprecedented hardware innovation

From CPU to specialized chips:

- GPUs, FPGAs, ASICs/co-processors (e.g., TPU)
- Tightly integrated (e.g., Intel's latest Xeon integrates CPU & FPGA)

New memory technologies

- HBM2: 8 DRAM chips/package → 1TB/sec

"Renaissance of hardware design" – David Patterson

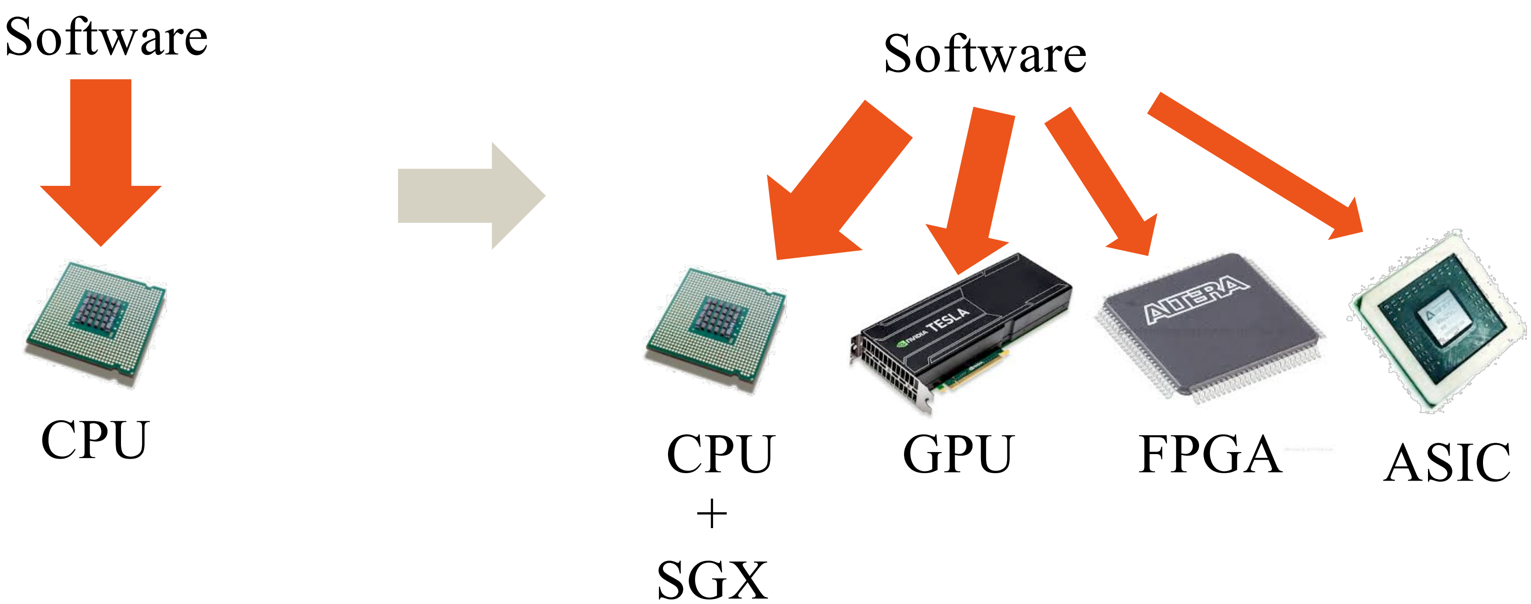
Overview

Application trends

Hardware trends

Challenges and techniques

Complexity – Computation



Complexity – Memory

2015

L1/L2 cache	~1 ns
L3 cache	~10 ns
Main memory	~100 ns / ~80 GB/s / ~100GB
NAND SSD	~100 usec / ~10 GB/s / ~1 TB
Fast HHD	~10 msec / ~100 MB/s / ~10 TB

2020

L1/L2 cache	~1 ns
L3 cache	~10 ns
HBM	~10 ns / ~1TB/s / ~10GB
Main memory	~100 ns / ~80 GB/s / ~100GB
NVM (3D Xpoint)	~1 usec / ~10GB/s / ~1TB
NAND SSD	~100 usec / ~10 GB/s / ~10 TB
Fast HHD	~10 msec / ~100 MB/s / ~100 TB

Complexity – more and more choices

Basic tier: A0, A1, A2, A3, A4
Optimized Compute : D1, D2, D3, D4, D11, D12, D13
D1v2, D2v2, D3v2, D11v2,…
Latest CPUs: G1, G2, G3, …
Network Optimized: A8, A9
Compute Intensive: A10, A11,…

Microsoft AZURE

t2.nano, t2.micro, t2.small
m4.large, m4.xlarge, m4.2xlarge, m4.4xlarge, m3.medium, c4.large, c4.xlarge, c4.2xlarge, c3.large, c3.xlarge, c3.4xlarge, r3.large, r3.xlarge, r3.4xlarge, i2.2xlarge, i2.4xlarge, d2.xlarge d2.2xlarge, d2.4xlarge,…

Amazon EC2

n1-standard-1, ns1-standard-2, ns1-standard-4, ns1-standard-8, ns1-standard-16, ns1highmem-2, ns1-highmem-4, ns1-highmem-8, n1-highcpu-2, n1-highcpu-4, n1-highcpu-8, n1-highcpu-16, n1-highcpu-32, f1-micro, g1-small…

Google Cloud Engine

Complexity – more and more constraints

Latency

Accuracy

Cost

Security

Techniques of conquering complexity

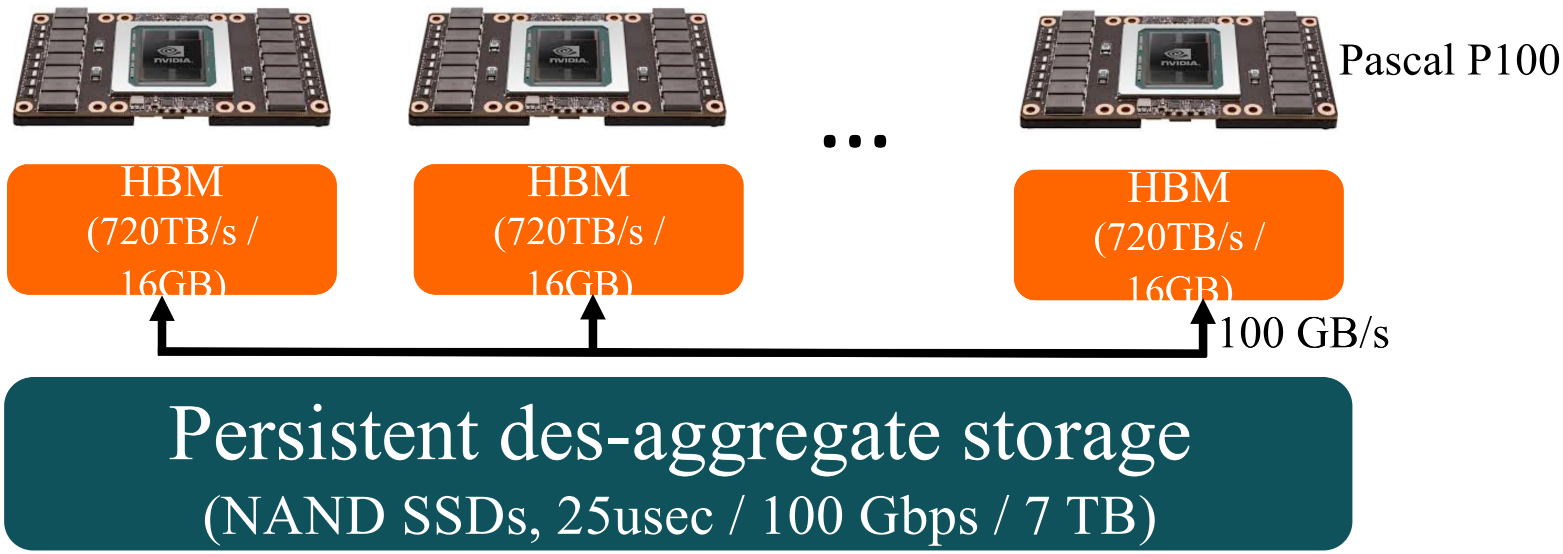
Use additional choices to simplify!

Expose and control tradeoffs

Don’t forget “tried & true” techniques

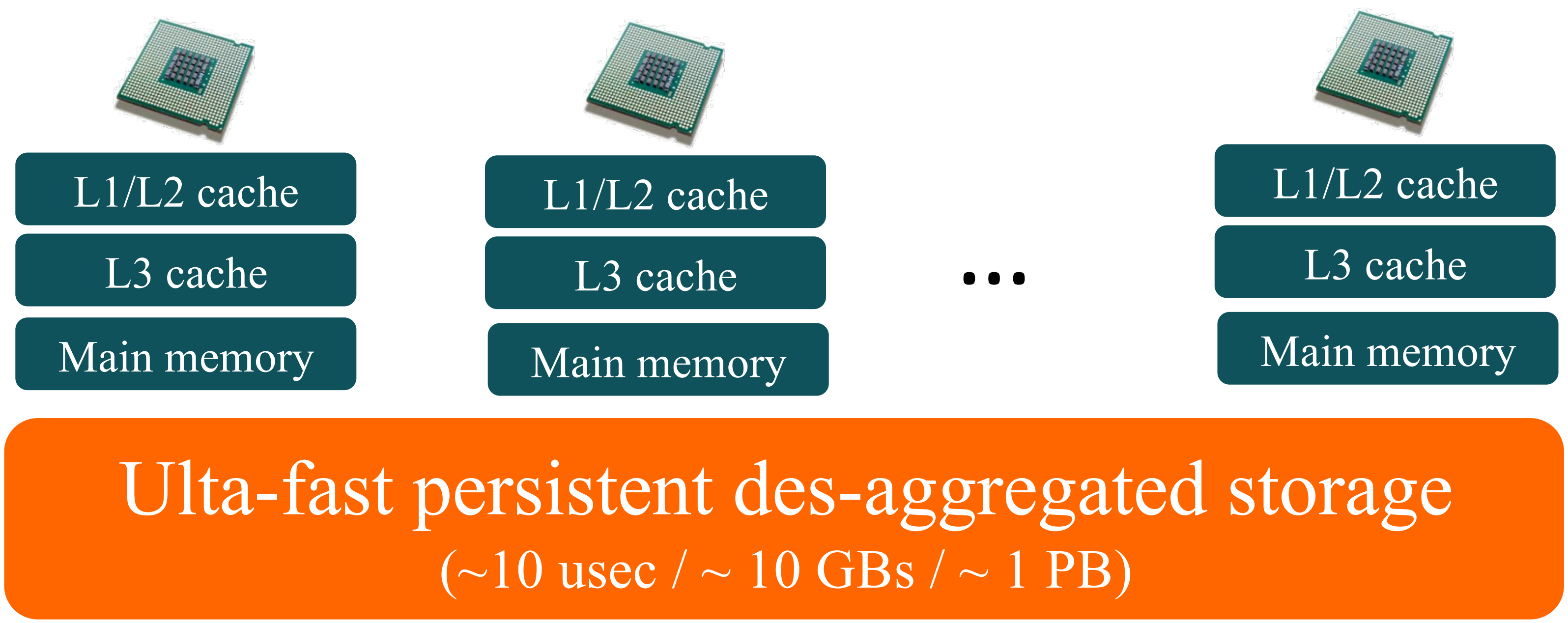
Use choices to simplify!

Example: NVIDIA DGX-1 supercomputer for Deep Learning



Use choices to simplify!

Possible datacenter architecture (e.g., FireBox, UC Berkeley)



Expose and control tradeoffs

Latency vs. accuracy

- Approximate query processing (e.g., BlinkDB)
- Decompose ML algos: light weight, ensemble and correction model (e.g., Clipper)

Latency (response time) vs. cost

- Predict response times given configuration (e.g., Earnest)

Security vs. latency vs. accuracy

- E.g., CryptDB, Opaque

"Tried & true" techniques

Sampling

- Scheduling (e.g., Sparrow), computation (e.g., BlinkDB), storage (e.g., KMN)

Batching

- Scheduling (e.g., Drizzle)

Speculation:

- Replicate time-sensitive requests/jobs (e.g., Dolly)

Incremental algorithms

- Updates (e.g., IndexedRDDs), and Machine Learning (e.g., Clipper)

Summary

We are at an inflection point both in terms of apps and hardware trends, and RISE lab is at the intersection of it

Many opportunities

Be aware of “complexity”: use myriad choices available to simplify!