Looking Ahead

(or we are mostly done)

Announcements

• May 4: Poster/Presentation

• May 11: Project Write-Up
   L75 pages, single column
   → similar information to posters
   (but more details)
   → please use pictures/visuals

• May 13: 4PM - 5:50PM Exam
   → same room
   → covers entire semester

Today: Trends & Summary

- A little bit of reading the tea leaves to see what
Aims:
1. Understand how things we looked at change.
2. Understand what remains the same.

Slowdown in scaling.

Dennard scaling ended in ~2006.

Parallellism core to scaling → multi-core, distributed computation.
Communication starts becoming a bottleneck.

Multi-core → coherence traffic.

Distributed → compression, serialization, net. protocol, deser.

22-27% of server cycles.
- Transistor Scaling/Moore's Law Slowing Down Since 2010, Ending Soon (?)

- Slowdown In Core Count Growth
  - Slowdown In New Processor Features
    - SIMD
    - VI-X, Virtualization, ...
    - AES-NI
    - ...

Some Consequences

Goals
- Improve Sequential Compute Time
- Reduce Communication
- Better Parallelization

- Specialized Hardware Accelerators
  - TPU
  - Programmable Asics
  - GPUs
  - ...

- Reconfigurable Hardware
  - FPGAs

Growing Data
- **More Sensors**
  - **Smart Phone** (ca. 2006) := Ubiquitous

- **More Connectivity**
  - A larger # of devices are now connected
  - Quality & Availability is still uneven
    - Varies economically & geographically

- **Storage Costs have shrunk**
  - Technology shifts, etc.

**Consequences**

- **More applications rely on statistical methods**
  - Machine Learning
  - Data Mining
    - Crime Prevention
    - Video Analytics
    - Medicine
    - ...

- Cost of centralizing data seems to be increasing
Almost all major providers (Google, AWS...) worried about global WAN capacity during pandemic.

Most (FB, NSFT, Google, AWS) now own their own global WANS.

Building & upgrading is slow, need fancy ships.

[Will return to proposals to address]

**Legality/other changes**

- Privacy laws
  
  - GDPR
  
  - German
  
  - CCPA
  
  - ...

- Data sovereignty
  
  - EU
  
  - China
  
  - Russia?
  
  - ...

- Tied to above?

- Impact on application semantics?
Accountability

Cost/Effect of Bugs/
Hacks/Outages

Growing concern about impact on critical infrastructure

Work on

- Verifying correctness
- Auditing
- Securing

But can we enforce?

Trends to address (more data + legal concerns)

Edge computing

Place computation in more geographic regions

\[ \text{Computation Resources} \]

\[ \text{Trade-off} \]

\[ \# \text{of Locations} \div \text{Proximity to Population} \]

Many questions

- Who owns resources?
- Who manages them?
- Who has access?
- How shared?
→ How stuff works?

→ What hardware?

→ What consistency, durability, isolation, ...
  guarantees?

→ Is this sufficient/necessary for reducing demand on the WAN?

→ Is this sufficient/necessary for meeting privacy/legal challenges?
Trends: New Algorithms & Data Structures

- Previously: Differential Privacy & Federated Learning
  (NYT article today complaining about use in census)

- Approximation ← Delay Communication To When Capacity Is Available + Compute On Partial Results

- CRDT ← A Framework To Construct
  ▸ Data types that converge despite communication delays

(Handwritten by Google Docs)
Challenges

+ Interpretability of Results?

+ Boons on Intermediate Answers?

Scheduling, Where Art Thou

+ Scalability --- Deal with more resources
  
  More spread out resources

...
Core Challenges

\[ \Rightarrow \]

Combination of locally optimal decisions are not globally optimal

When, where, how to give up global optima
ADDITIONAL CONCERN

Economic & Environmental Cost of Computation