Scalable Systems and Algorithms

Prof. Anthony D. Joseph
adj@berkeley.edu
April 25, 2022

About Me

- MIT EECS Bachelors and Masters, and CS PhD
  - 5th year Masters program with co-op internship
  - PhD focus on mobile computing and disconnected operation

- Chancellor's Professor in Electrical Engineering and Computer Science
  - Joined Berkeley in 1998
  - Core Faculty in Center for Computational Biology
  - Faculty Director of Fung Institute for Engineering Leadership, https://funginstitute.berkeley.edu
  - Campus Cyber-Risk Responsible Executive

About Me

- Current research areas:
  - Fog Robotics (edge computing)
  - Secure Machine Learning (SecML)
  - DETER security testbed
- Previous research areas:
  - Modin (drop-in Pandas replacement)
  - Cancer Genomics/Precision Medicine (ADAM)
  - Cloud computing (Apache Mesos /Apache Spark)
  - Peer-to-Peer networking (Tapestry)
  - Mobile computing and Wireless/Cellular networking (Iceberg)
  - Scalable Internet Services (Ninja)

About Me

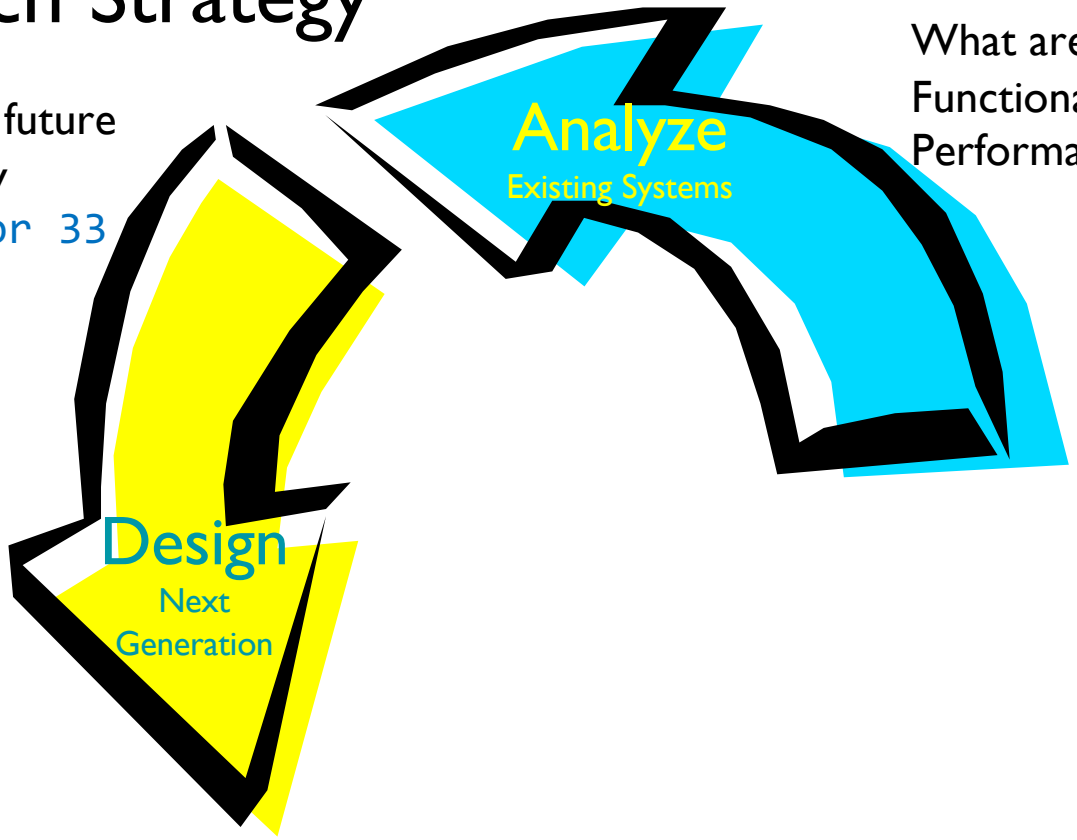
- Some Outside Activities
  - Big Data and Apache Spark BerkeleyX MOOCs: '15/'16 >240k students with >11% finishing
  - Unite Genomics co-founder (focused on rare diseases treatments)

My Research Strategy

What are the issues?
Functional limitations?
Performance bottlenecks?

My Research Strategy

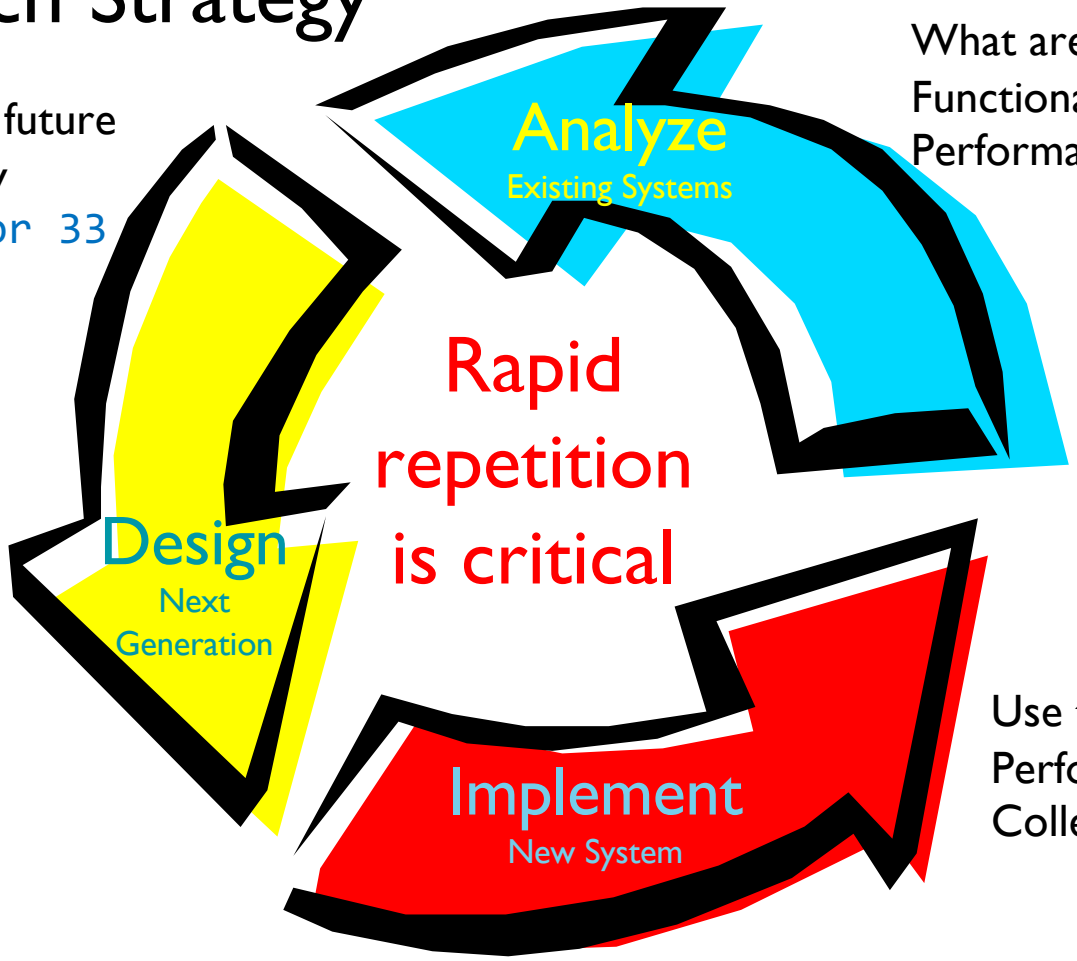
Think 10 years into the future
Be willing to go blue sky
Risk Xerox PARC `error 33`



What are the issues?
Functional limitations?
Performance bottlenecks?

My Research Strategy

Think 10 years into the future
Be willing to go blue sky
Risk Xerox PARC `error 33`



What are the issues?
Functional limitations?
Performance bottlenecks?

Use time travel!
Perform rapid prototyping
Collect user feedback

Two Projects

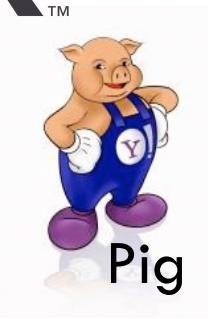
- Mesos
- Big Data Genomics

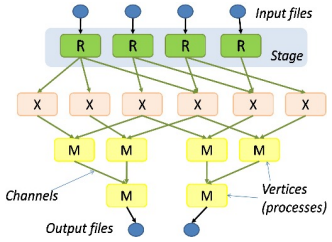
Apache MESOS™

Benjamin Hindman, Andy Konwinski, Matei Zaharia, Ali Ghodsi,
Anthony D. Joseph, Randy Katz, Scott Shenker, Ion Stoica

Motivation

Google
Pregel





Dryad

CIEL

S4 distributed stream computing platform

Google
Percolator

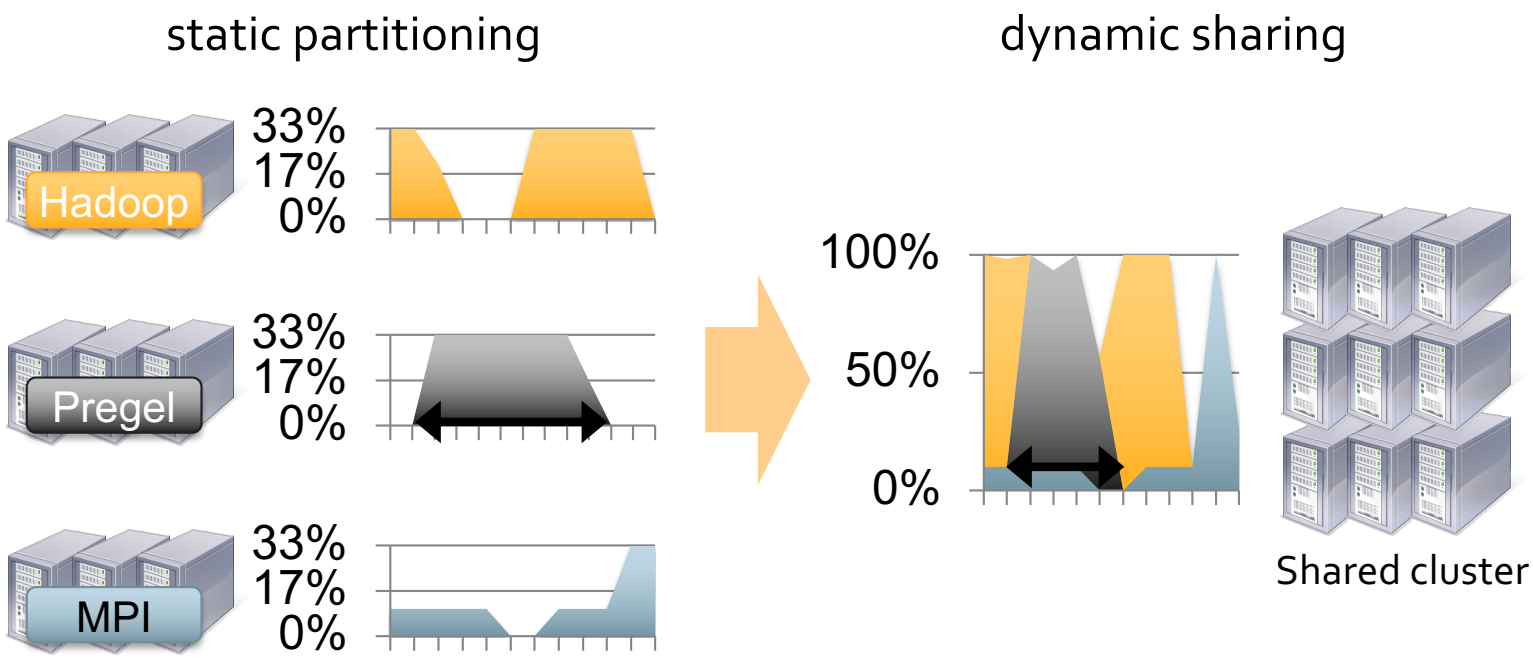
observation. Many distributed apps; no single one optimal for all use cases.

Want to run multiple applications in a single cluster

…to *maximize utilization*

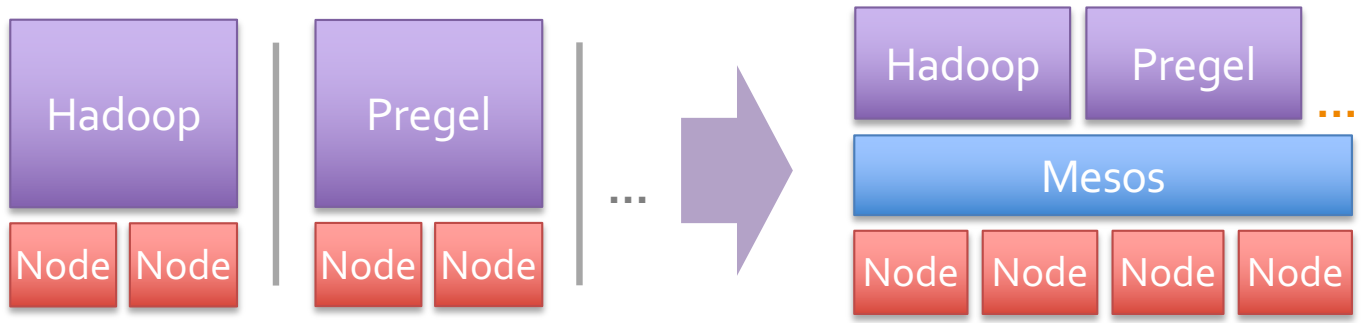
…to *share data*

Mesos aims to make it easier to build distributed applications/frameworks and share cluster resources


static partitioning
dynamic sharing
Hadoop
Pregel
MPI
33%
17%
0%
33%
17%
0%
33%
17%
0%
100%
50%
0%
Shared cluster

Solution

Mesos, a common layer over which diverse applications can run



Run multiple instances and/or versions of the *same* application

Build *specialized applications* targeting particular problem domains

Mesos Goals

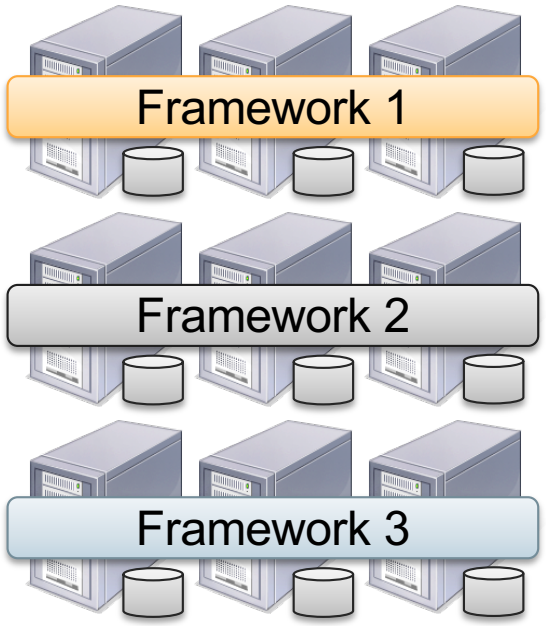
- High utilization of resources
- Diverse applications
- Scalability to 50-100K nodes
- Fault tolerance

Two-Level Model

- Mesos: controls resource allocations to applications/frameworks
- Applications/frameworks: make decisions about what to run
- Mesos allocator influenced by resource requests from applications/frameworks

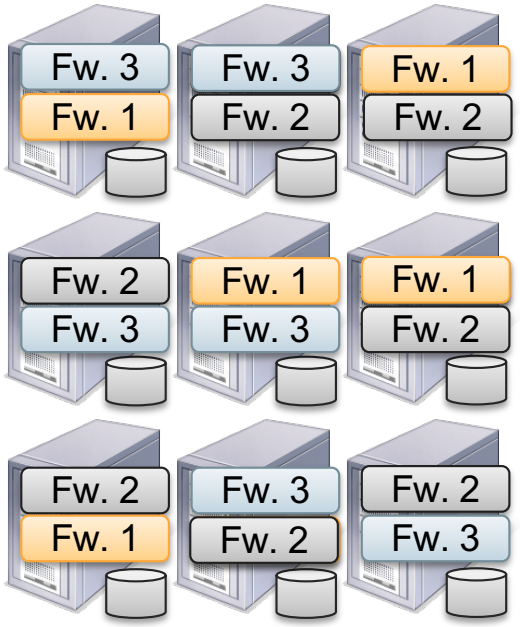
design element 1. Fine-grained sharing

Coarse-Grained Sharing (HPC):



Storage System (e.g. HDFS)

Fine-Grained Sharing (Mesos):



Storage System (e.g. HDFS)

+ Improved utilization, responsiveness, data locality

design element 2. Resource offers (vs. global scheduler)

Global Scheduler

Frameworks express needs in a specification language, global scheduler matches them to resources

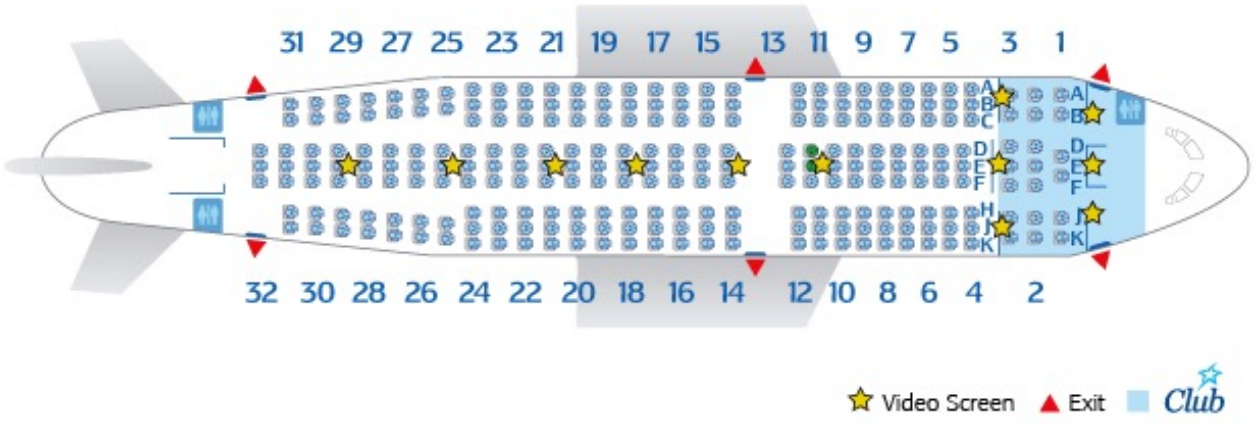
+ Can make optimal decisions
– Complex: language must support all framework needs
– Difficult to scale and to make robust
– Future frameworks may have unanticipated needs

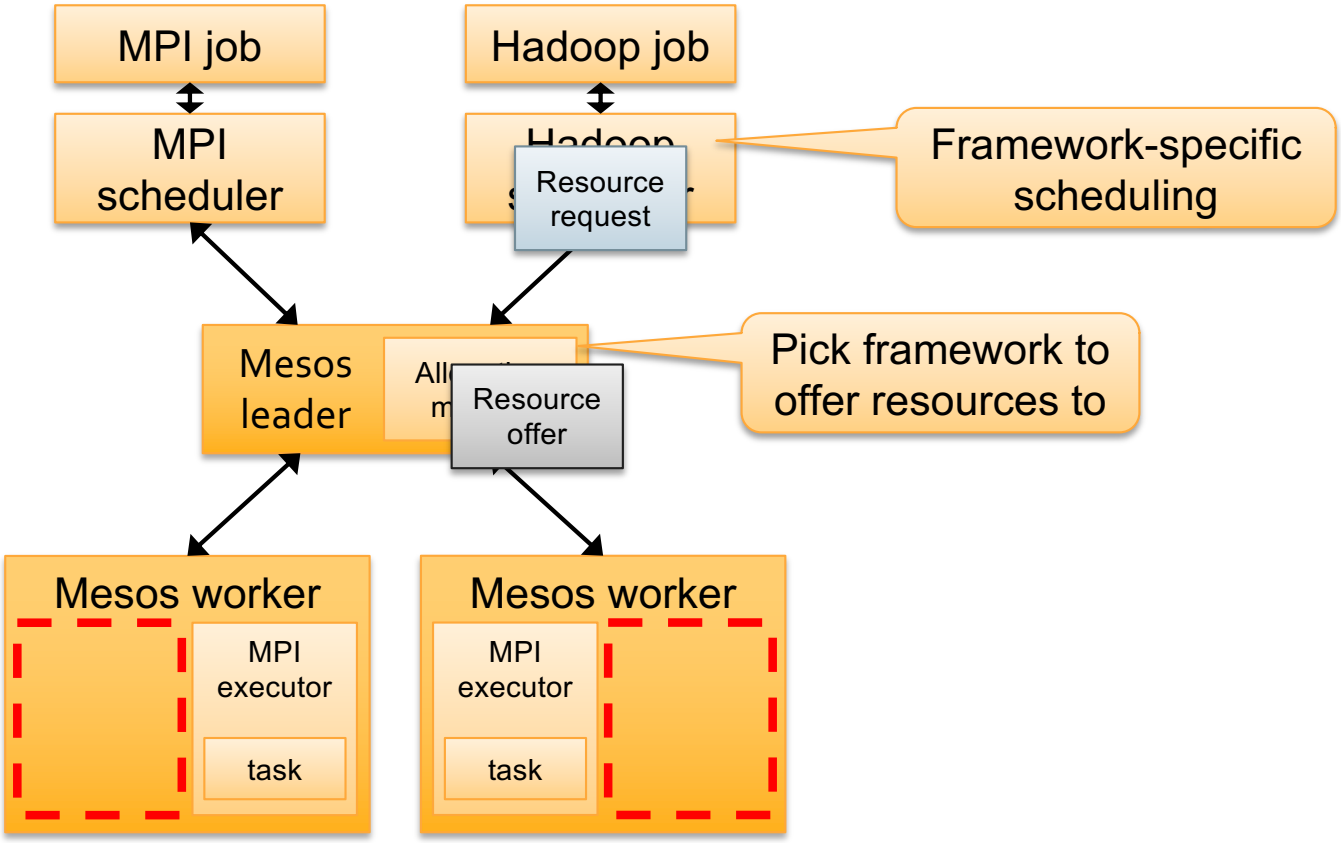
Mesos: Resource Offers

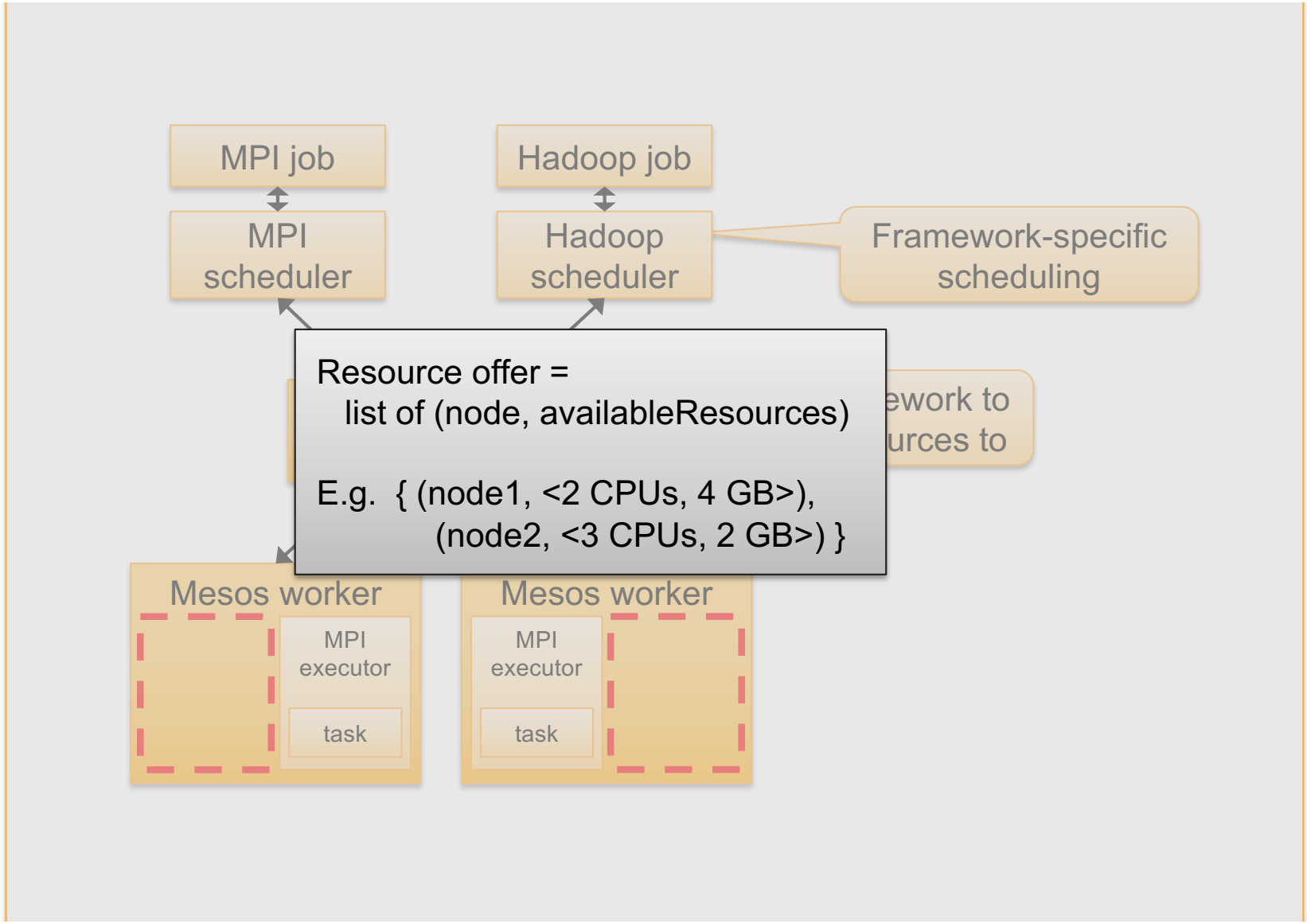
- Offer available resources to frameworks, let them pick which resources to use and which tasks to launch

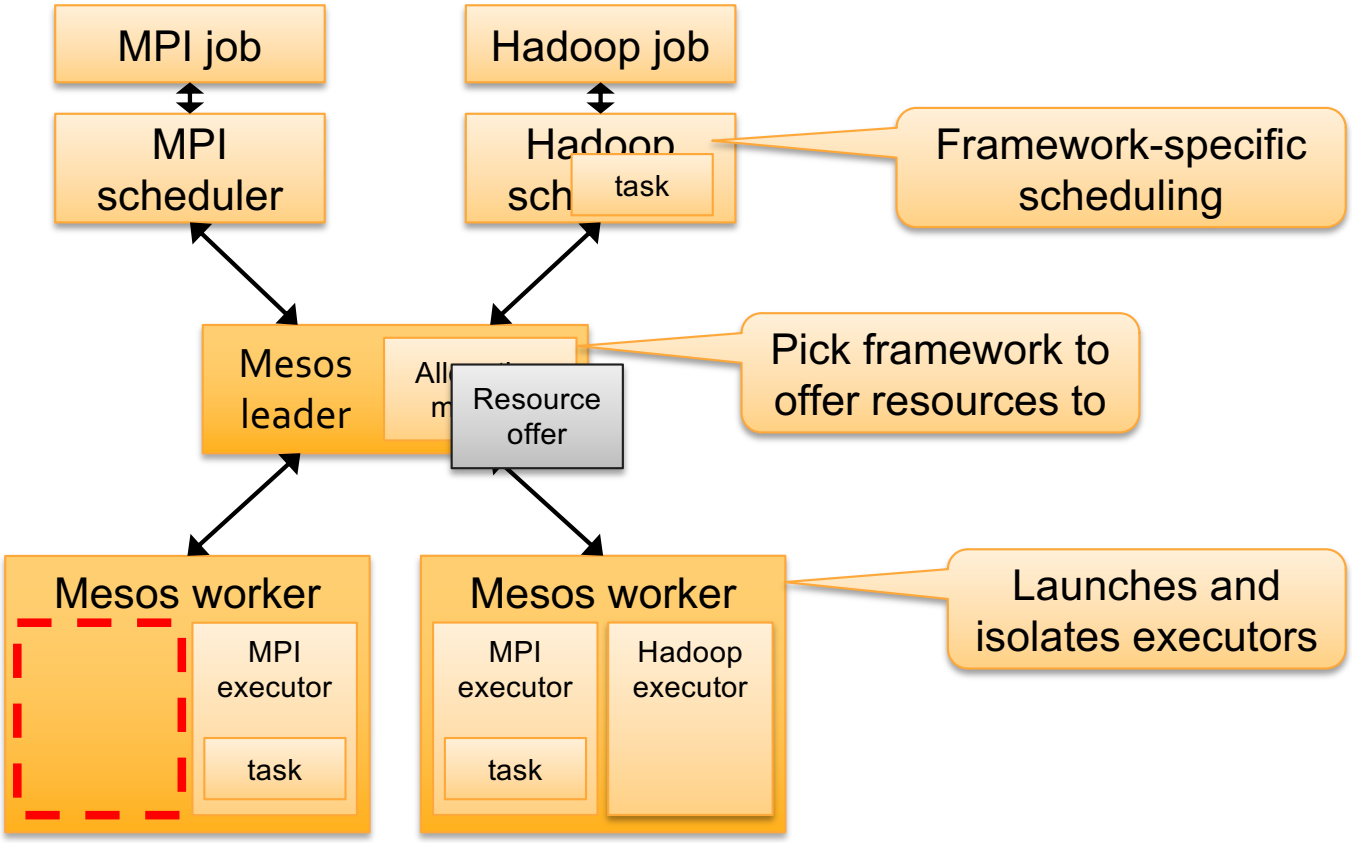
  + Keeps Mesos simple, lets it support future frameworks

  – Decentralized decisions might not be optimal




MPI job
Hadoop job
MPI scheduler
Resource request
Framework-specific scheduling
Mesos leader
Resource offer
Pick framework to offer resources to
Mesos worker
MPI executor
task
Mesos worker
MPI executor
task


MPI job
Hadoop job
MPI scheduler
Hadoop scheduler
Framework-specific scheduling
Resource offer =
list of (node, availableResources)
E.g. { (node1, <2 CPUs, 4 GB>),
(node2, <3 CPUs, 2 GB>) }
Mesos worker
MPI executor
task
Mesos worker
MPI executor
task


MPI job
Hadoop job
MPI scheduler
task
Framework-specific scheduling
Mesos leader
Resource offer
Pick framework to offer resources to
Mesos worker
MPI executor
task
Mesos worker
MPI executor
task
Hadoop executor
Launches and isolates executors

Resource Requests

A framework can construct a collection of requests that specify either a particular host, some particular amount of resources, or both

Resource Request Examples

Desired Resource(s)	Created Resource Request
Specific machine with any number of resources	Only specifies the host (and not any resources)
Any machine with a particular number of resources	Only specifies the resources (and not any host)
A specific machine with a specific number of resources	Specifies both
A rack with any number of resources	Collection of requests, one per host in rack (resources unset)
A rack with a specific number of resources	Collection of requests, one per host in rack, specifying desired per host resources

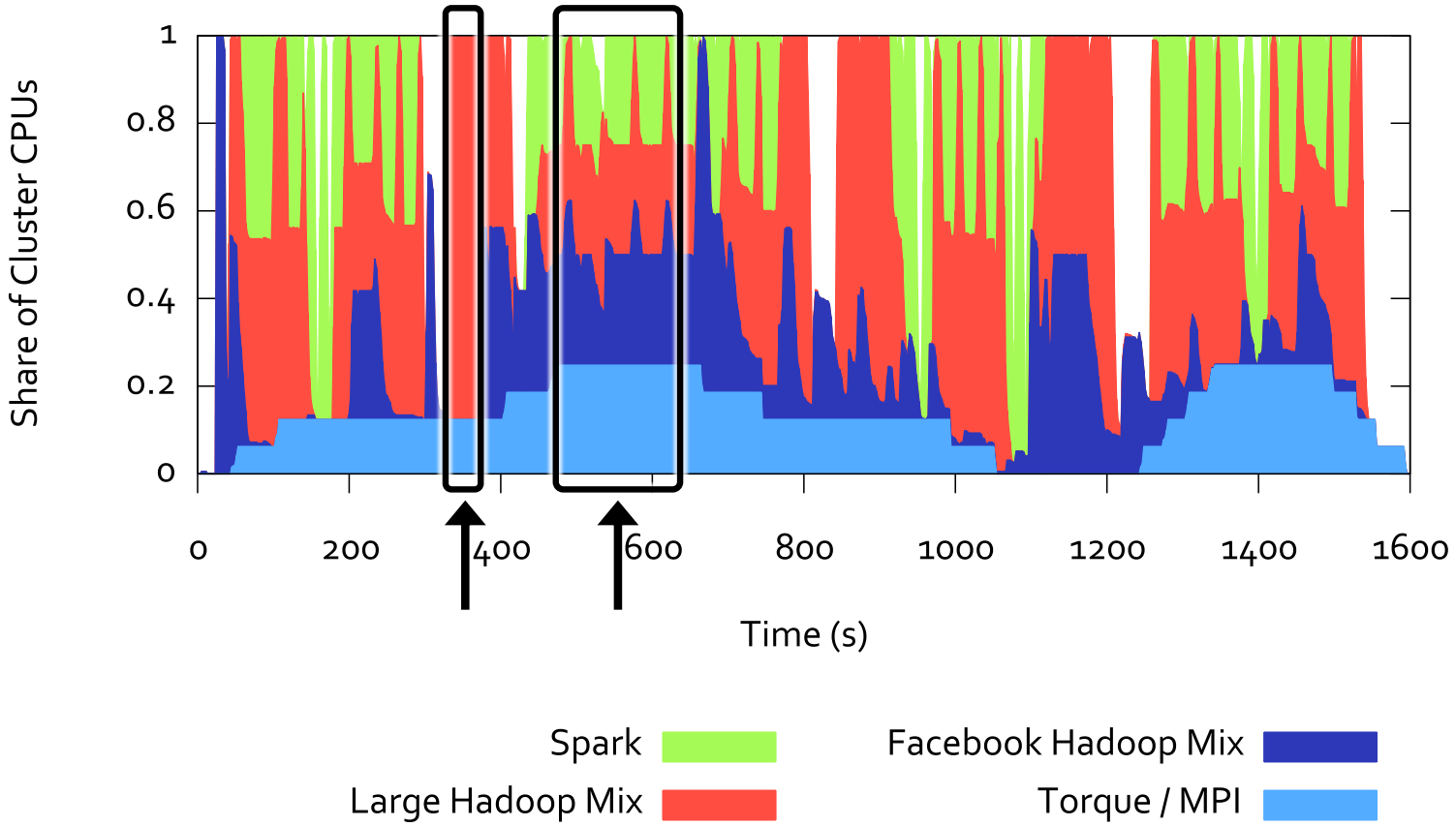
Fault Tolerance

- Mesos leader has only *soft state*: list of currently running frameworks and tasks
- Rebuild when frameworks and workers re-register with new leader after a failure
- **Result:** fault detection and recovery in ~10 sec

Framework Isolation

- Mesos uses OS isolation mechanisms, such as Linux containers and Solaris projects
- Containers currently support CPU, memory, IO and network bandwidth isolation
- Future work: other isolation mechanisms (e.g., lightweight virtualization)

evidence that sharing helps


Share of Cluster CPUs
1
0.8
0.6
0.4
0.2
0
0
200
400
600
800
1000
1200
1400
1600
Time (s)
Spark
Large Hadoop Mix
Facebook Hadoop Mix
Torque / MPI

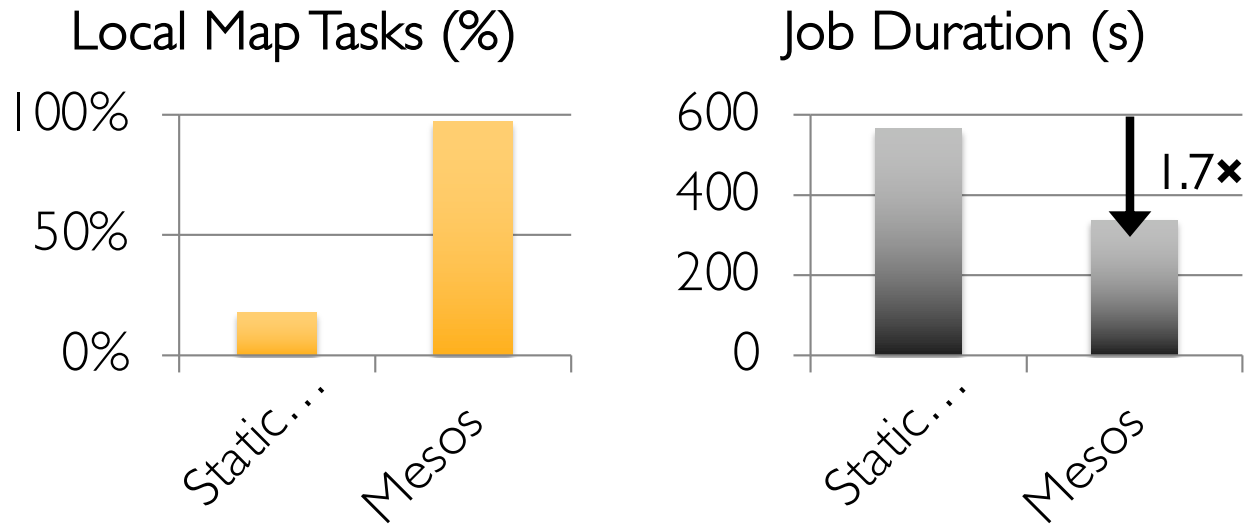
vs. static allocation (each app gets 25% of nodes)

Framework	Speedup on Mesos
Facebook Hadoop Mix	1.14×
Large Hadoop Mix	2.10×
Spark	1.26×
Torque / MPI	0.96×

Data Locality with Resource Offers

Ran 16 instances of Hadoop on a shared HDFS cluster

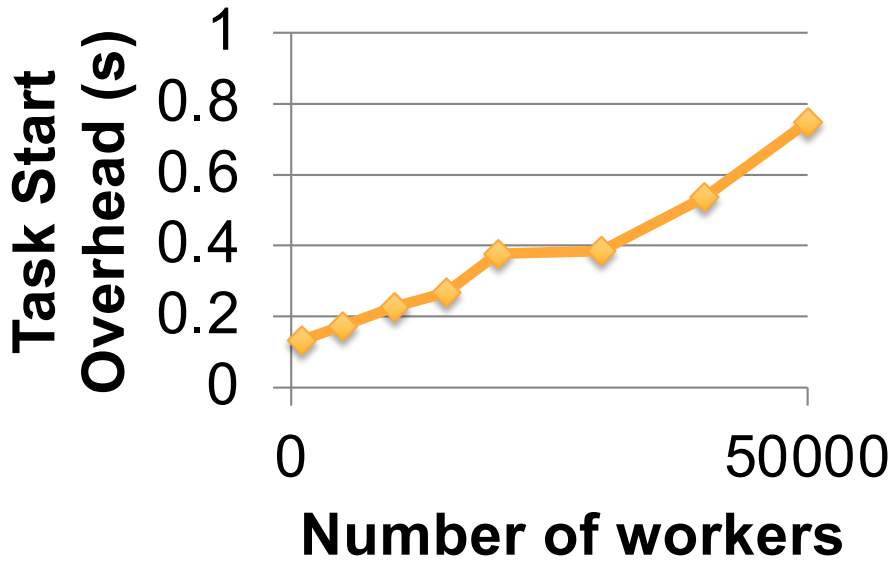
Used delay scheduling [EuroSys '10] in Hadoop to get locality (wait a short time to acquire data-local nodes)



Scalability

Mesos only performs *inter-framework* scheduling (e.g. fair sharing), which is easier than intra-framework scheduling

Result:
Scaled to 50,000 emulated workers, 200 frameworks, 100K tasks (30s len)



Mesos Summary

- Mesos shares clusters efficiently among diverse frameworks thanks to two design elements:
    - **Fine-grained sharing** at the level of tasks
    - **Resource offers**, a scalable mechanism for application-controlled scheduling
- Enables co-existence of current frameworks and development of new specialized ones
- In use at Twitter, Netflix, UC Berkeley, Conviva and UCSF

Big Data in 2020

View from 2011

Almost Certainly:

- Create a new generation of big data scientist
- A real datacenter OS
- ML becoming an engineering discipline
- People deeply integrated in big data analysis pipeline

If We're Lucky:

- System will know what to throw away
- Generate new knowledge that an individual person cannot

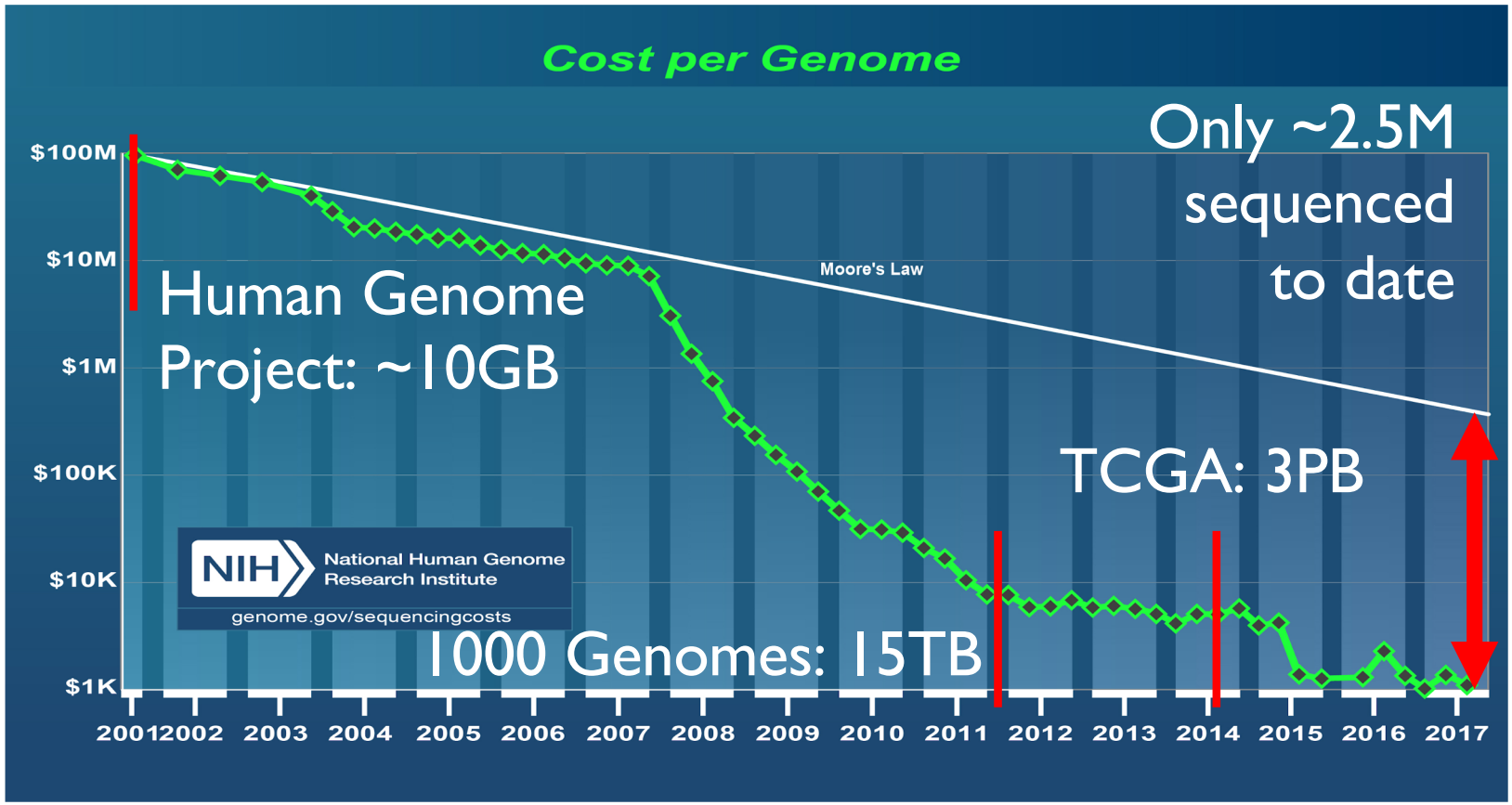
https://github.com/bigdatagenomics/adam

Democratizing scalable bioinformatics tools

Acknowledgements

- **UC Berkeley:** Matt Massie, Timothy Danford, André Schumacher, Jey Kottalam, Karen Feng, Eric Tu, Alyssa Morrow, Niranjan Kumar, Frank Nothaft, Ananth Pallaseni, Michael Heuer, Justin Paschall, Taner Dagdelen, Devin Petersohn, Anthony D. Joseph, Dave Patterson
- **Mt. Sinai:** Arun Ahuja, Neal Sidhwaney, Ryan Williams, Michael Linderman, Jeff Hammerbacher
- **GenomeBridge:** Carl Yeksigian
- **Cloudera:** Uri Laserson, Tom White
- **Microsoft Research:** Ravi Pandya, Bill Bolosky
- **UC Santa Cruz:** Benedict Paten, David Haussler, Hannes Schmidt, Beau Norgeot, Audrey Musselman-Brown, John Vivian
- And many other open source contributors, especially Neil Ferguson, Andy Petrella, Xavier Tordior, Deborah Siegel, Denny Lee
- Total of >70 contributors to ADAM/BDG from >12 institutions

Motivation: Moore's Law



Single Whole Human Genome Data Sizes

	Input	Pipeline Stage	Output
SNAP	3GB Fasta 200GB Fastq	Alignment	100GB BAM
ADAM	250GB BAM	Pre-processing	200GB ADAM
Avocado	200GB ADAM	Variant Calling	10MB ADAM

Whole Genome Sequence Analysis Challenges

- Data exploding and genomic analyses are slow and expensive
  - 2025: Genomes 40 EB/year versus YouTube videos 2EB/year
- Can accelerate with lots of computers, but ad hoc process
  - Most existing tools are designed to run on a single computer
- Clinical (timeliness), Research (large data scalability) challenges
  - Many open questions: genotype vs phenotype, gene vs environment, ...
- Only a few institutions can afford to work with national-scale datasets
  - Million Veterans Program, NIH All of Us, UK 100K, Singapore 5M

A Precision Medicine First

- Joshua Osborn (14 years old)
  - Visited ER 3 times in 4 months with encephalitis (acute inflammation of the brain)
  - Hospitalized after 3rd visit and placed into a medically induced coma due to uncontrollable seizures

A Precision Medicine First

- Joshua Osborn (14 years old)
  - Visited ER 3 times in 4 months with encephalitis (acute inflammation of the brain)
  - Hospitalized after 3rd visit and placed into a medically induced coma due to uncontrollable seizures

- Doctors tried traditional medicine approaches
  - Over 100 viral/fungal/bacterial pathogens cause encephalitis - *treatment depends on pathogen*
  - Given many different *pathogen-specific* tests
  - 1 cm^3 of brain tissue biopsied

A Precision Medicine First

- Last resort – sent spinal fluid to UCSF for DNA sequencing
  - Needle in a haystack! Most of the genetic data (hay) is Joshua's
- Dr Charles Chiu's analysis pipeline (includes BDG software)
  - Discarded human data (99.984% of data), and harmless viruses/bacteria
- 80% of remaining non-human data from rare *Leptospira santarosai* bacteria
  - Joshua had been given pathogen-specific CDC test, but it was negative
  - Test later determined to be broken!
- Joshua given correct antibiotics, out of coma in week, discharged 4 weeks

The NEW ENGLAND JOURNAL of MEDICINE

HOME | ARTICLES & MULTIMEDIA | ISSUES | SPECIALTIES & TOPICS | FOR AUTHORS | CME

ORIGINAL ARTICLE
BRIEF REPORT

June 4, 2014

Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing

Michael R. Wilson, M.D., Samia N. Naccache, Ph.D., Erik Samayoa, B.S., C.L.S., Mark Biagtan, M.D., Hiba Bashir, M.D., Guixia Yu, B.S., Shahria… Ph.D., Robert Sokolic… Kurt D. Reed, M.D., … Henderson, M.D., P…
June 4, 2014 | DOI: …

Abstract | Ar…

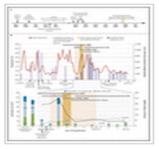
More than half … laboratory test… establishing a … can be difficult… discover emer… generation seq… largely unexpl… bacterial cause… contributed dir… outcome.

CASE REPOR…

A 14-year-old boy with severe combined immunodeficiency (SCID) caused by adenosine deaminase deficiency and partial immune reconstitution after he had undergone two haploidentical bone marrow transplantations initially presented to the emergency department in early April 2013 after having had headache and fevers, with temperatures up to 39.4°C, for 6 days (Figure 1A). He was admitted to the hospital and discharged 1 day later after resolution of his fever and headache.

The patient's outpatient medications included monthly infusions of intravenous immune globulin for hypogammaglobulinemia and trimethoprim–sulfamethoxazole or atovaquone for prophylaxis against *Pneumocystis jirovecii* pneumonia. He had no known sick contacts but did have three pet cats. He had gone on a missionary trip to Puerto Rico during the first 2 weeks of August 2012 (Figure 1A), where he swam in a river and the ocean. Notably …

FIGURE 1



Clinical Course of the …

The New York Times

In a First, Test of DNA Finds Root of Illness

By CARL ZIMMER JUNE 4, 2014

Joshua Osborn, 14, lay in a coma at American Family Children's Hospital in Madison, Wis. For weeks his brain had been swelling with fluid, and a battery of tests had failed to reveal the cause.

The doctors told his parents, Clark and Julie, that they wanted to run one more test with …inal fluid for …halitis.

…here so …University …n —

…gnals an …quencing …e to yield

… Slezak, …l

…efore …useful: Not …uld lead to … Lyme disease.

Diagnosis is a crucial step in medicine, but it can also be the most difficult. Doctors usually must guess the most likely causes of a medical problem and then order individual tests to see which is the right diagnosis.
The guessing game can waste precious time. The causes of some conditions, like encephalitis, can be so hard to diagnose that doctors often end up with no answer at all.

"About 60 percent of the time, we never make a diagnosis" in encephalitis, said Dr. Michael R. Wilson, a neurologist at the University of California, San Francisco, and an author of the new paper. "It's frustrating whenever someone is doing poorly, but it's especially frustrating when we can't even tell the parents what the hell is going on."

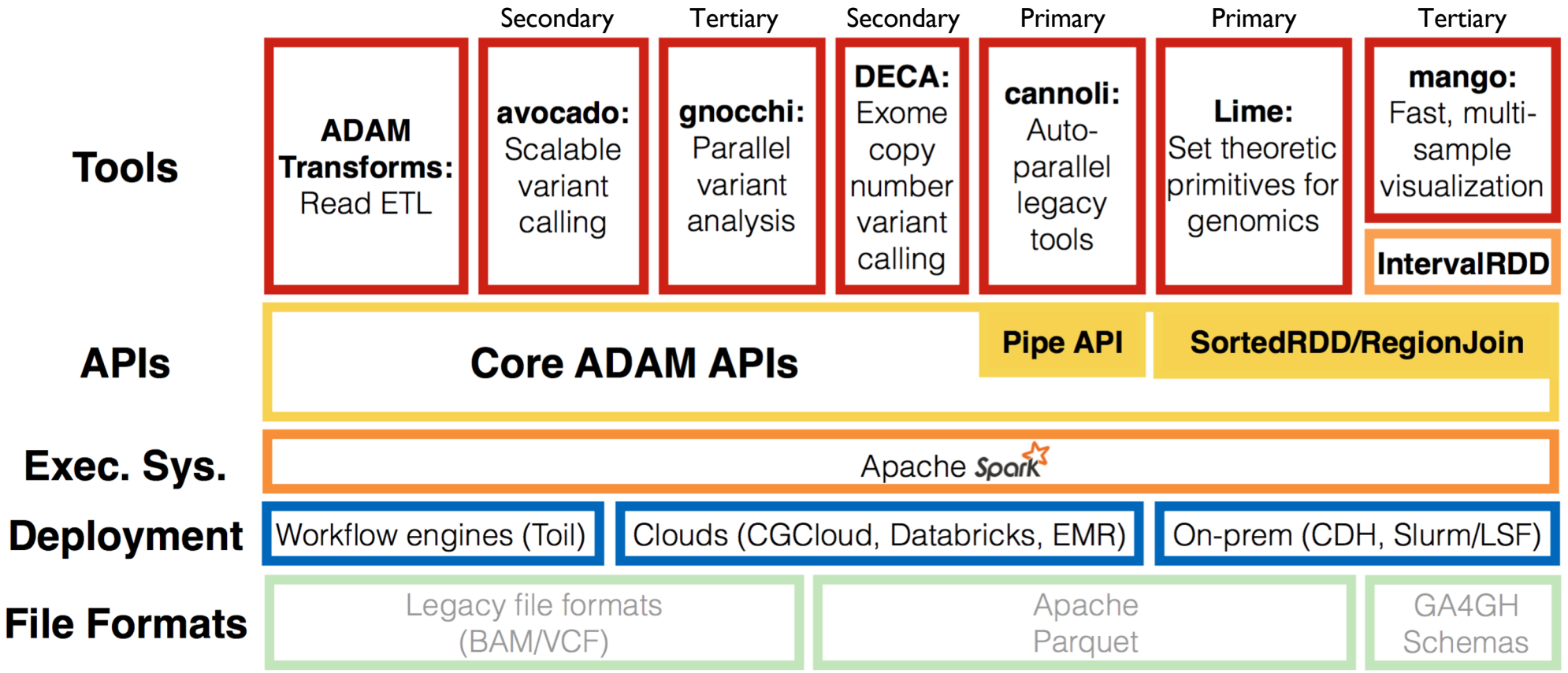
- US: 15,000 cases a year of encephalitis with 2,000 deaths (>70% of deaths underdiagnosed)
- Worldwide: 70,000 diagnosed cases a year with 25,000 deaths

https://amplab.cs.berkeley.edu/2014/06/04/snap-helps-save-a-life/

Parallel Genomic Analysis with ADAM

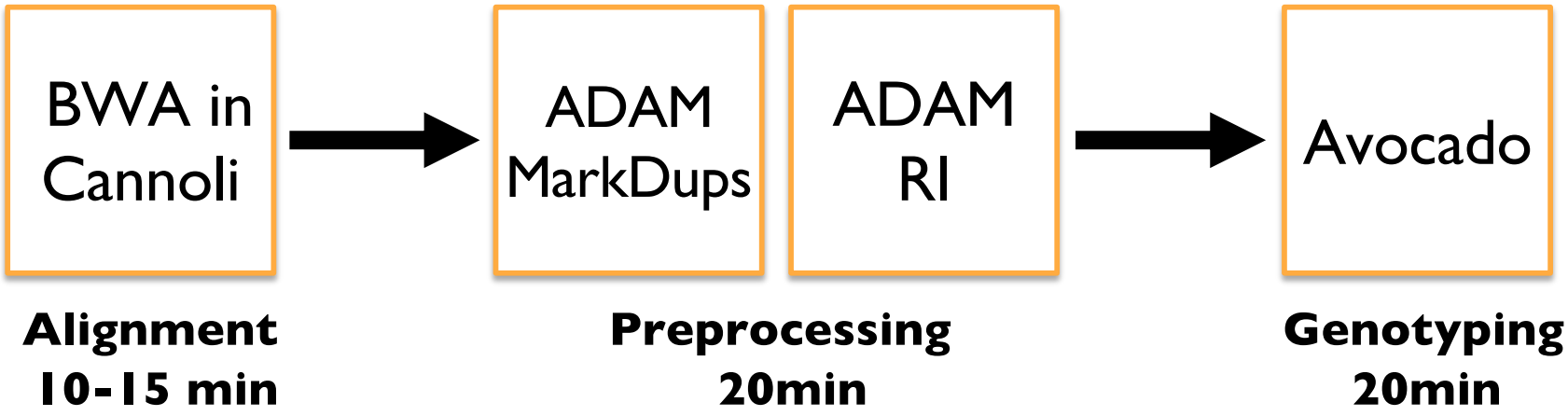
- Open source, high perf, distributed genomic analysis library
  - ADAM handles data management and partitioning
- Bio users write queries in familiar languages: Python and R
  - Batch and exploratory analysis of all types of genomic data
- Reuses existing tools by automatically parallelizing them
  - Many bioinformatics tools follow a streaming pattern...

Big Data Genomics Open Source Software Stack



End-to-end variant calling in ADAM

50-55 minutes to align and variant call WGS, $10-20

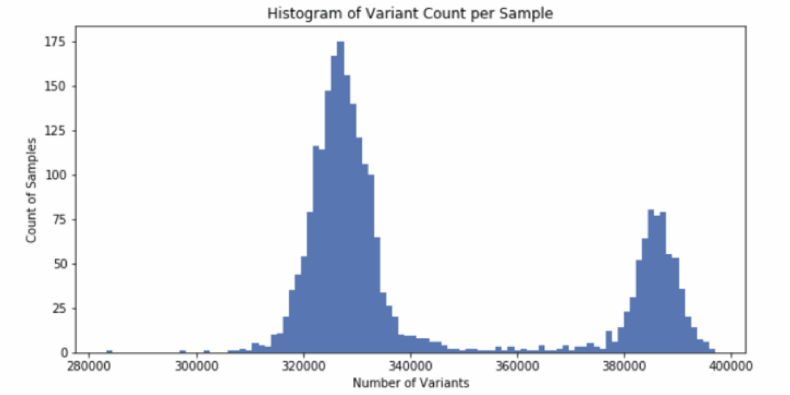


- Highly accurate primary and secondary analysis
- Over 17x faster than GATK3 and 2-3x faster than GATK4

Genomic Exploratory Data Analysis using Mango

```
ax, results = VariantsPerSampleDistribution(spark, genotypes).plot()

ax.set_title("Histogram of Variant Count per Sample")
ax.set_xlabel("Number of Variants")
ax.set_ylabel("Count of Samples")
plt.show()
```



- Interactively build ML models and apply them to genomic datasets
  - Example: Transcription Factor Binding Site prediction
- Visualize results on 10 TB dataset in < ½ second

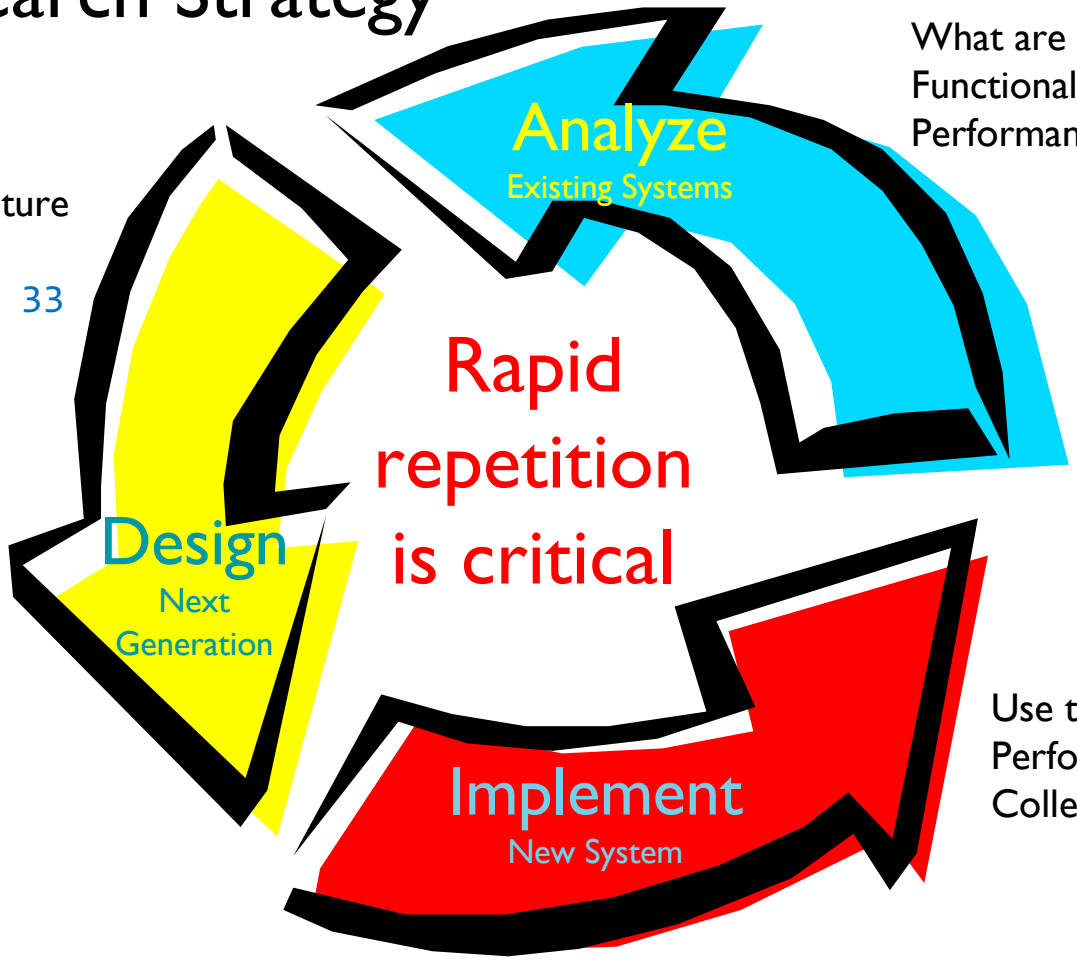
BDG Summary

- Research genomics use is scale-sensitive
  - Many countries have population-scale sequencing efforts
- Clinical genomics use case is high-accuracy, time-sensitive
  - Accurately diagnosing infections in minutes...
- Open source tools like ADAM help democratize genomics use
  - Rapid WGS processing and analyses
  - Interactive, population-scale data exploration

Undergrads: Remember Dan Garcia's LAUGH metaphor!

- Lean In and take the initiative to connect with faculty and engage in research
- Academic soulmate: find your best project buddies and try to take your project courses together
- Underground project: work on personal projects you're really interested in during free time
- Give back as a TA or reader or lab assistant or volunteer
- Have fun!

Grads: Research Strategy



What are the issues?
Functional limitations?
Performance bottlenecks?

Think 10 years into the future
Be willing to go blue sky
Risk Xerox PARC `error 33`

Use time travel!
Perform rapid prototyping
Collect user feedback

Thanks!