Cook and Spark

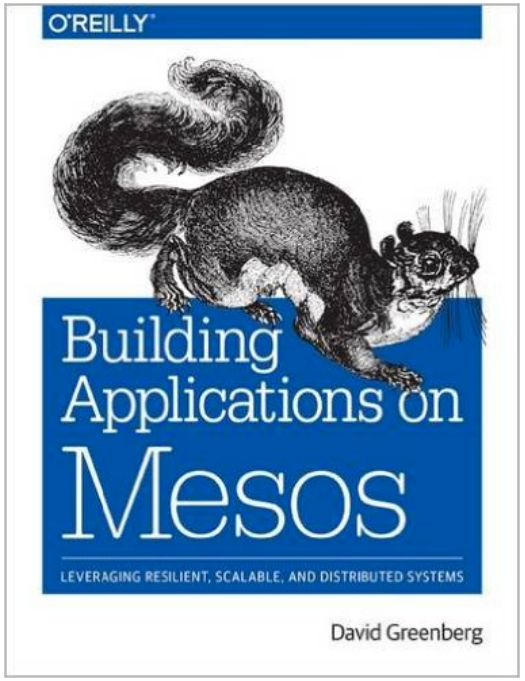
David Greenberg

Two Sigma

Who am I?

- Architected project to build a massive multi-datacenter Mesos cluster
- Built custom framework, leveraging open source software
- Author of upcoming book

Building Applications on Mesos



Plan for Today

- What is Spark?
- What is Cook?
- How can we use them together?
- Cook on its own?

Plan for Today

- What is Spark?
- What is Cook?
- How can we use them together?
- Cook on its own?

What is Spark?

- Map Reduce
- Medium Data
- Better Hadoop
  - In Memory*

Spark Workflow

- Write interactively at REPL
- Submit to cluster in production

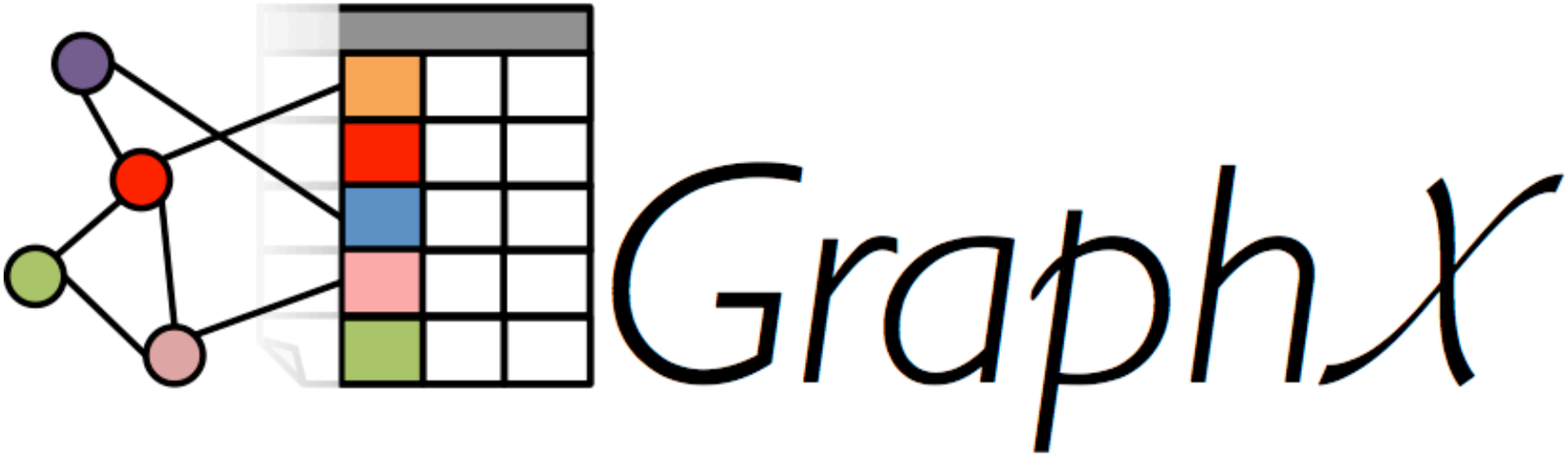
```
mbo@mbo-ubuntu-vbox:~/mbo/spark$ MASTER=spark://localhost:7077 ./spark-shell
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 0.9.0-SNAPSHOT
      /_/

Using Scala version 2.9.3 (Java HotSpot(TM) Client VM, Java 1.6.0_45)
Initializing interpreter...
Creating SparkContext...
Spark context available as sc.
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

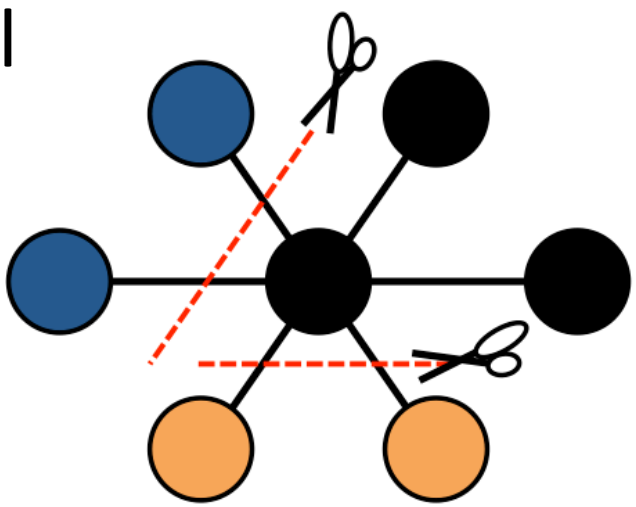
Tools Built on Spark

- GraphX

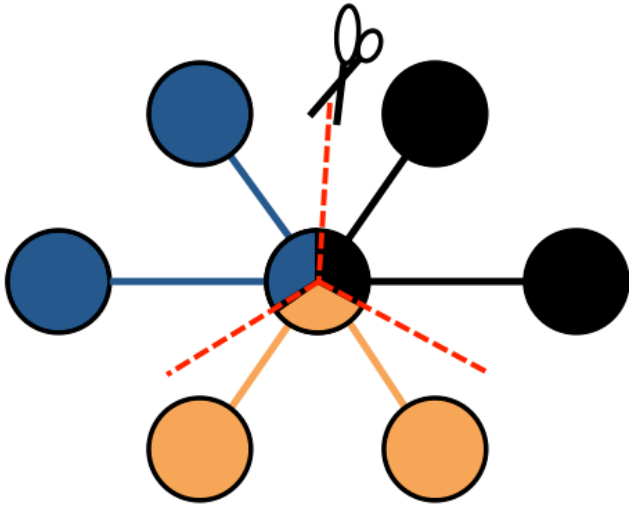


GraphX

- All kinds of graph algorithms
- PageRank
- Pregel
- Joins



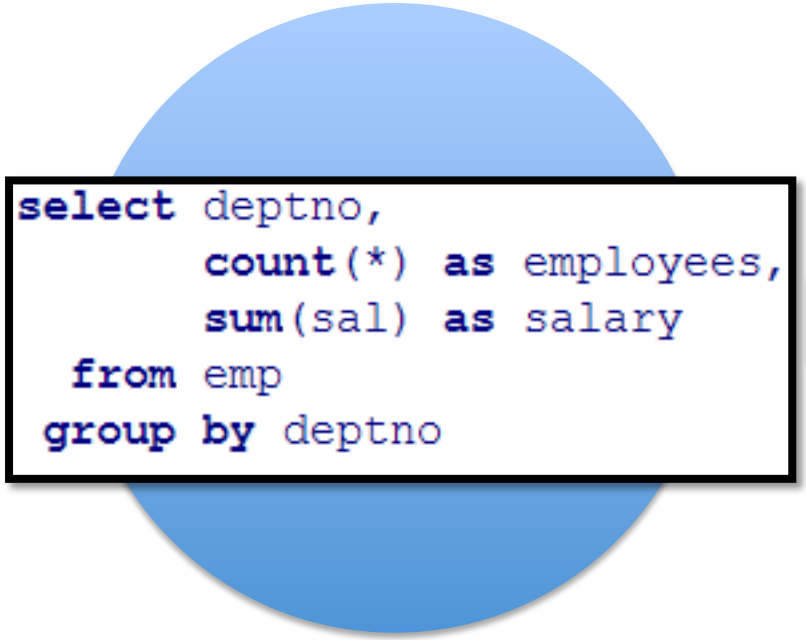
Edge Cut



Vertex Cut

Tools Built on Spark

- GraphX
- SparkSQL



```
select deptno,
       count(*) as employees,
       sum(sal) as salary
  from emp
 group by deptno
```

Tools Built on Spark

- GraphX
- SparkSQL
- MLlib

```
points = spark.textFile("hdfs://...")
              .map(parsePoint)

model = KMeans.train(points, k=10)
```

Also, it's about 100x faster

Tools Built on Spark

- GraphX – Fast, parallel graph analysis
- SparkSQL – Run SQL on datasets without indices or an RDBMS
- MLlib – Modern machine learning library

Spark on Mesos

Works great for one user!

No story for multiple users on one Mesos cluster…

Plan for Today

- What is Spark?
- What is Cook?
- How can we use them together?
- Cook on its own?

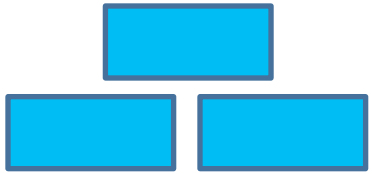
Cook

- Preemptive Job Scheduler
- What should we optimize for?
  - Latency vs Throughput
  - User happiness
  - Cumulative Resource Shares (see Li Jin's Mesoscon 2015 talk)

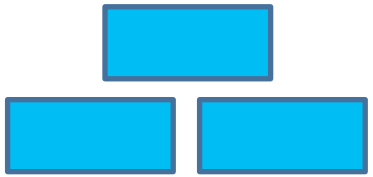
Preemption: Intuition

Waiting

Running



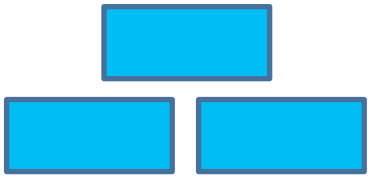
Ava



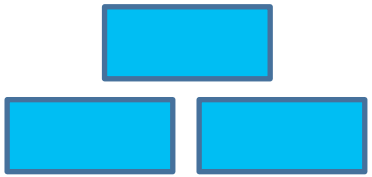
Bartley

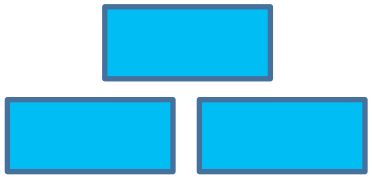
Preemption: Intuition

Waiting



Running





Ava

Bartley

Conor

Preemption: Intuition

Waiting



Running



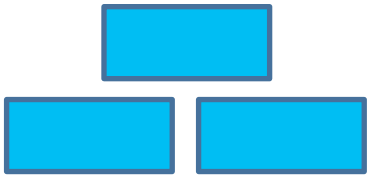
Ava

Bartley

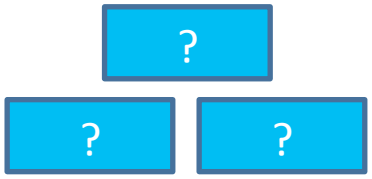
Conor

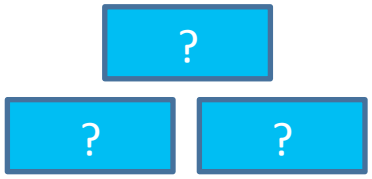
Preemption: Intuition

Waiting



Running





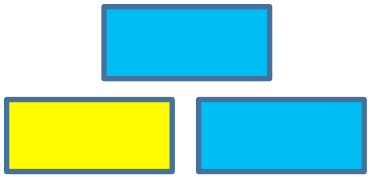
Ava

Bartley

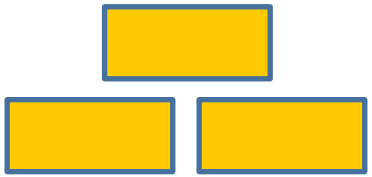
Conor

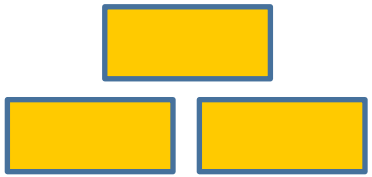
Preemption: Intuition

Waiting



Running





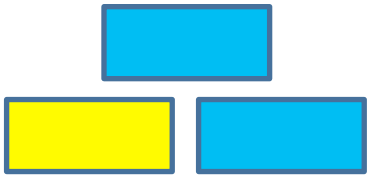
Ava

Bartley

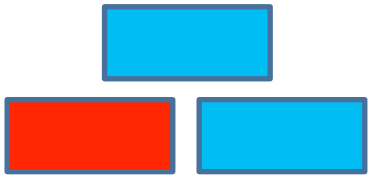
Conor

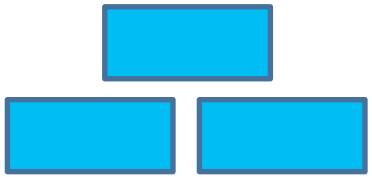
Preemption: Intuition

Waiting



Running





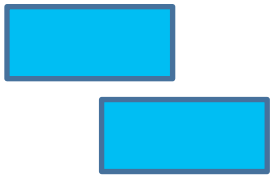
Ava

Bartley

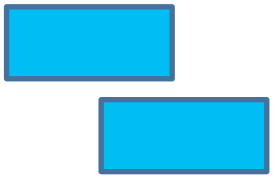
Conor

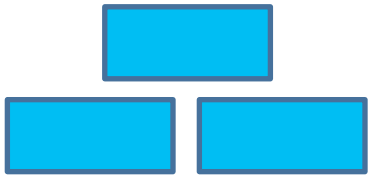
Preemption: Intuition

Waiting



Running





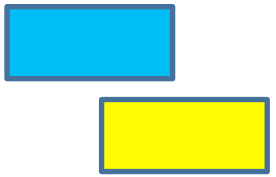
Ava

Bartley

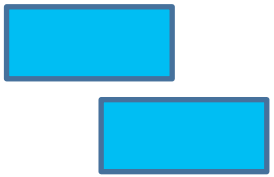
Conor

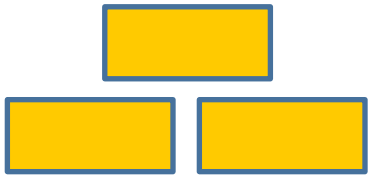
Preemption: Intuition

Waiting



Running





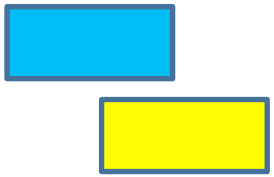
Ava

Bartley

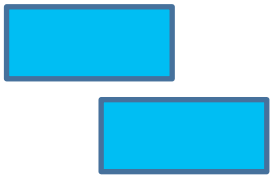
Conor

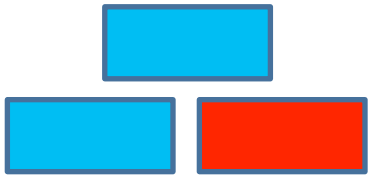
Preemption: Intuition

Waiting



Running





Ava

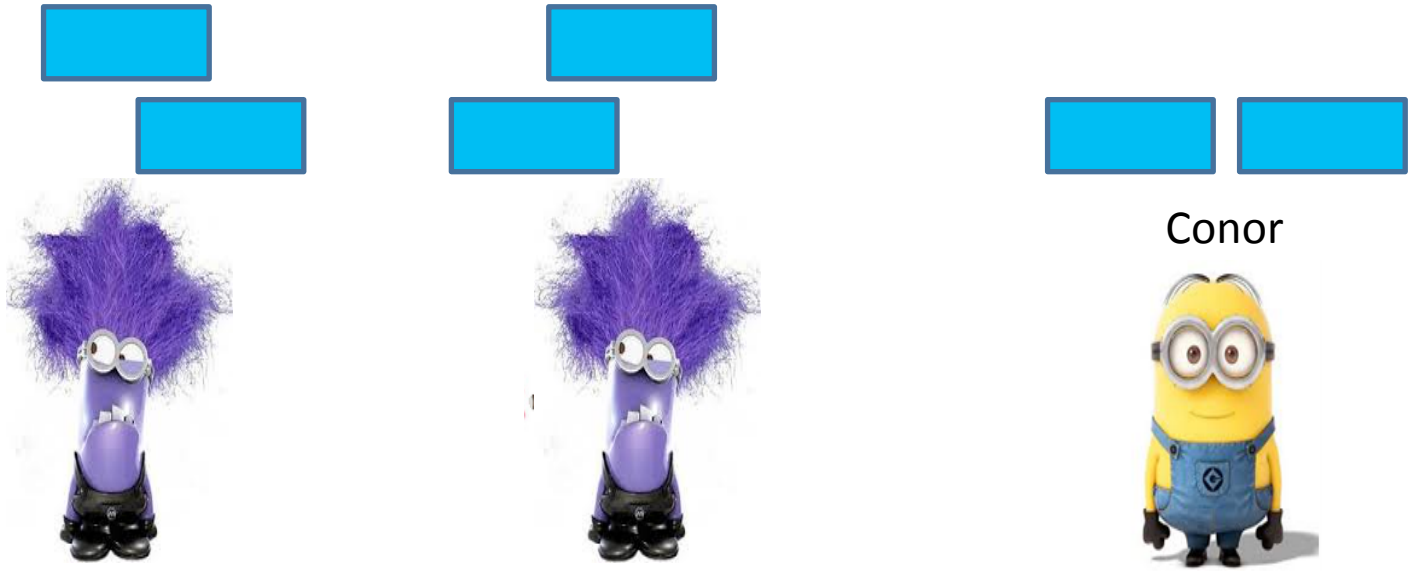
Bartley

Conor

Preemption: Intuition

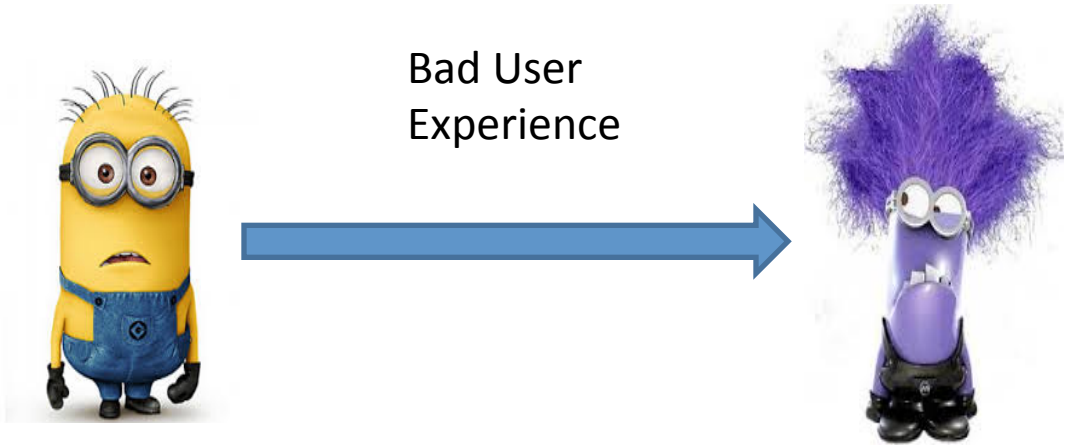
Waiting

Running



Problem

- Not all tasks are equal
  - We just preempted some important tasks!



Preemption: Intuition

Waiting

¥¥
¥¥¥
¥¥¥¥

Running

€
€€
€€€

Ava

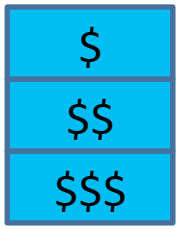
£££
££££
£££££

Bartley

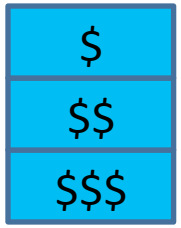
Conor

Preemption: Intuition

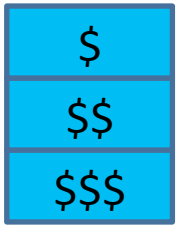
Waiting



Running



Ava

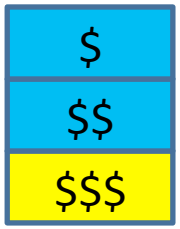


Bartley

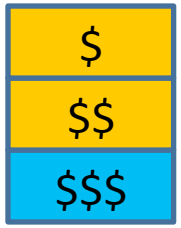
Conor

Preemption: Intuition

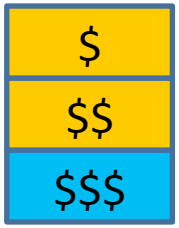
Waiting



Running



Ava

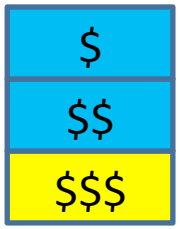


Bartley

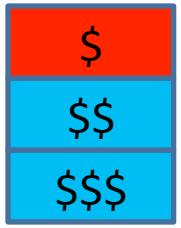
Conor

Preemption: Intuition

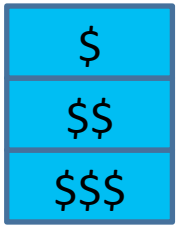
Waiting



Running



Ava



Bartley

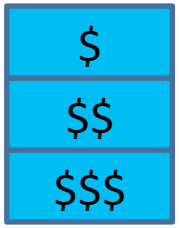
Conor

Preemption: Intuition

Waiting

Running

Ava



Bartley

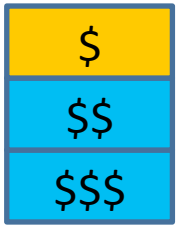
Conor

Preemption: Intuition

Waiting

Running

Ava



Bartley

Conor

Preemption: Intuition

Waiting

Running

Ava

Bartley

Conor

Preemption: Intuition

Waiting

€

£££

¥¥

Running

€€

€€€

Ava

£££

£££££

Bartley

¥¥¥

¥¥¥¥

Conor

Optimizing for User Happiness Works!

- Do our best to satisfy everyone interactively first
- We reduced complaints about capacity from 10/week to ~0/week

Let's Talk Specifics

- 2 level queue—fair then priority
  - Ex each user, or each team
- Cook runs "jobs"—idempotent, retriable commands
- All scheduling is automatic

Plan for Today

- What is Spark?
- What is Cook?
- How can we use them together?
- Cook on its own?

Spark + Cook

1. Apply the patch
2. Connect to Cook:

   ```
   cook://dgrnbrg:passwd@cook.example.com
   ```

3. No changes necessary, just awesomeness

No Changes Necessary

- We've been using this at Two Sigma for several months
- Multitenant Spark on Mesos has been solved

Plan for Today

- What is Spark?
- What is Cook?
- How can we use them together?
- Cook on its own?

Cooking Independently

- 100% Open Source, Apache Licensed
- Written in Clojure with Datomic
- Bundles all dependencies
- Running in production >1 year

Batch Jobs

What if I don't use Spark?

- Java API
- Rest API
- Submit, query, monitor, and kill jobs

Built for Production

- High Availability
- Built-in safety valves and checks
- Detailed internal performance and cluster metrics
  - JMX, Riemann, Graphite
- HTTP Basic and Kerberos Auth

Upcoming Features

- Fenzo for Task Placement
  - Constraints
  - Bin-packing
  - Launching Speed
- More Client Libraries
  - Go, C++, Python

Plan for Today

- What is Spark?
- What is Cook?
- How can we use them together?
- Cook on its own?

Spark

Like Hadoop but faster and cooler

Cook

Dynamically shares your cluster with advanced technology

Try it!

- Download Cook: http://github.com/twosigma/cook
- Run Spark on it
- Report bugs
- Give feedback