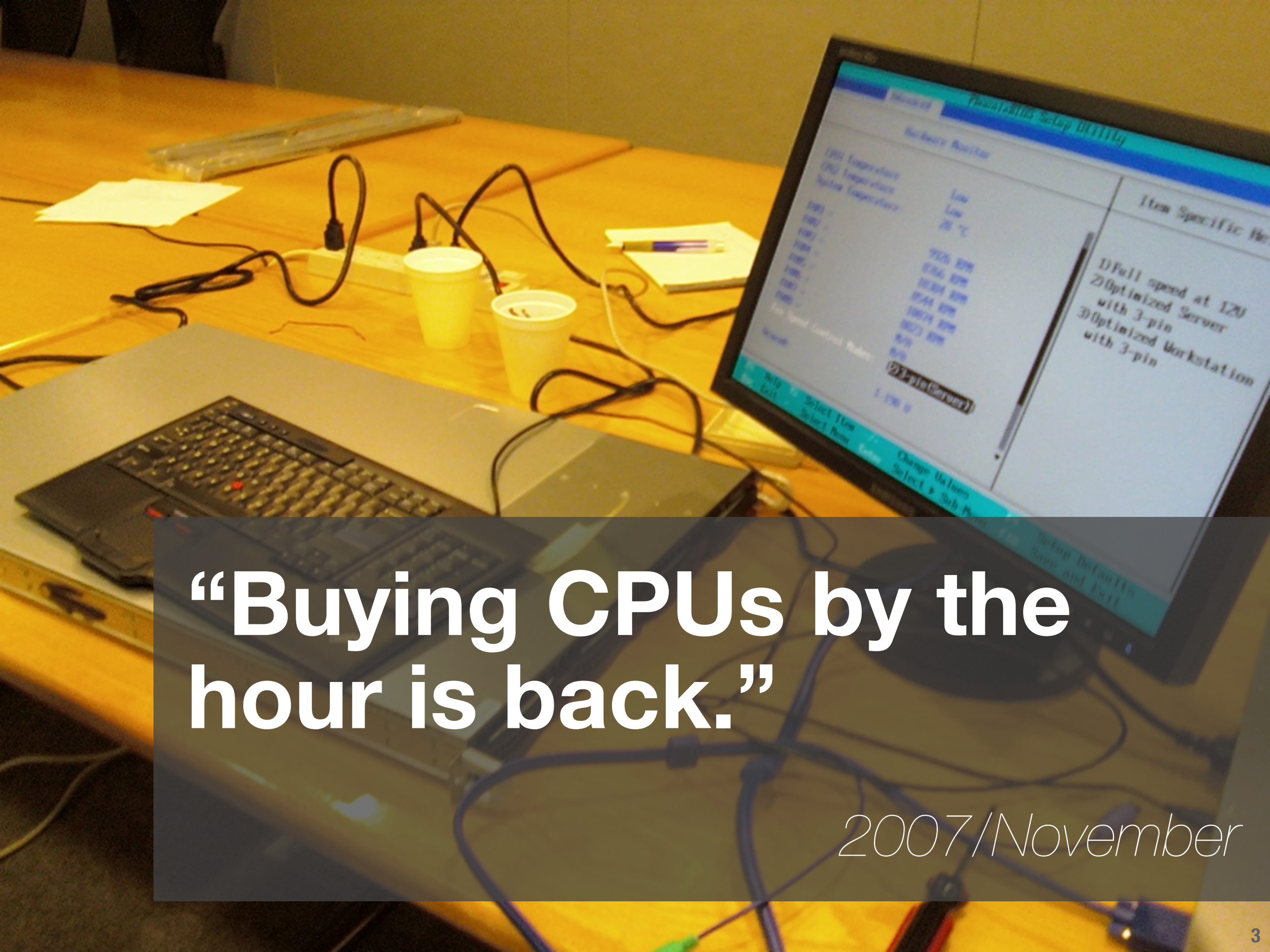
Converged IT for Life Sciences

Scientific Computing in the Cloud

Bioinformatics and Big Iron



“Buying CPUs by the hour is back.”

2007/November

Cloud Computing

Hyper-converged Infrastructure

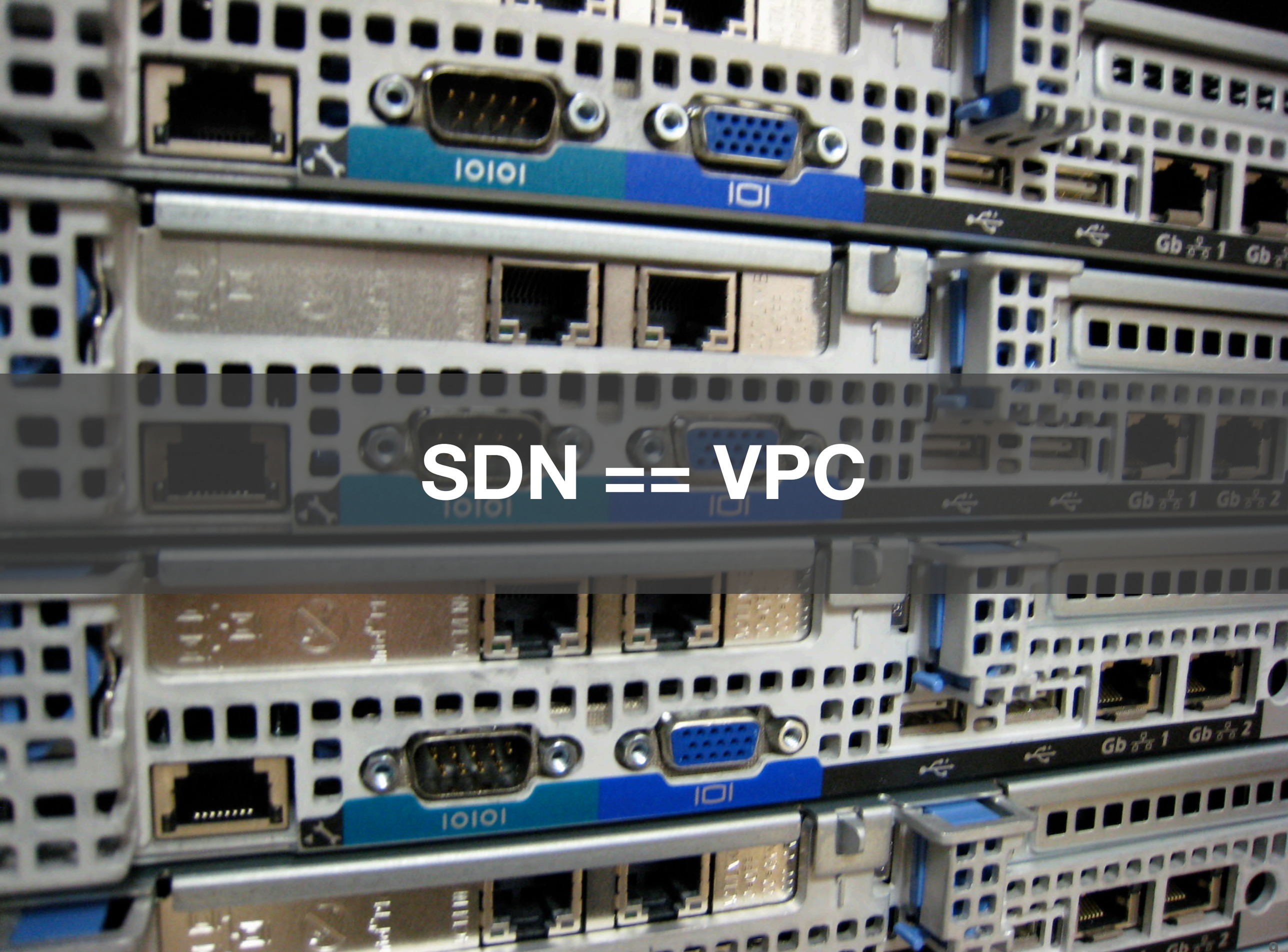
What it is

- **Converged Infrastructure is…**
  - **Computing on Storage?**
  - **Storage on Compute?**
- **Infrastructure as Code**
- **Scriptable Infrastructure**
- **Software Defined Infrastructure**

Building Blocks

Everything has an API

Software Defined Networking

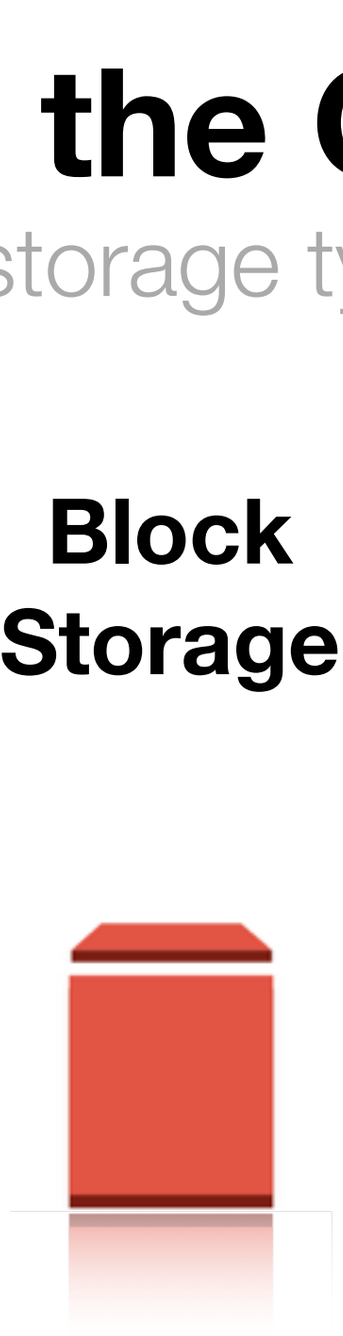


SDN == VPC

Software Defined Storage

Storage in the Cloud

Choose the right storage type for your application

Object Storage	Block Storage	Data Stores	Filesystems
			
Amazon S3	Amazon EBS	DynamoDB	Amazon EFS

Infrastructure as Code

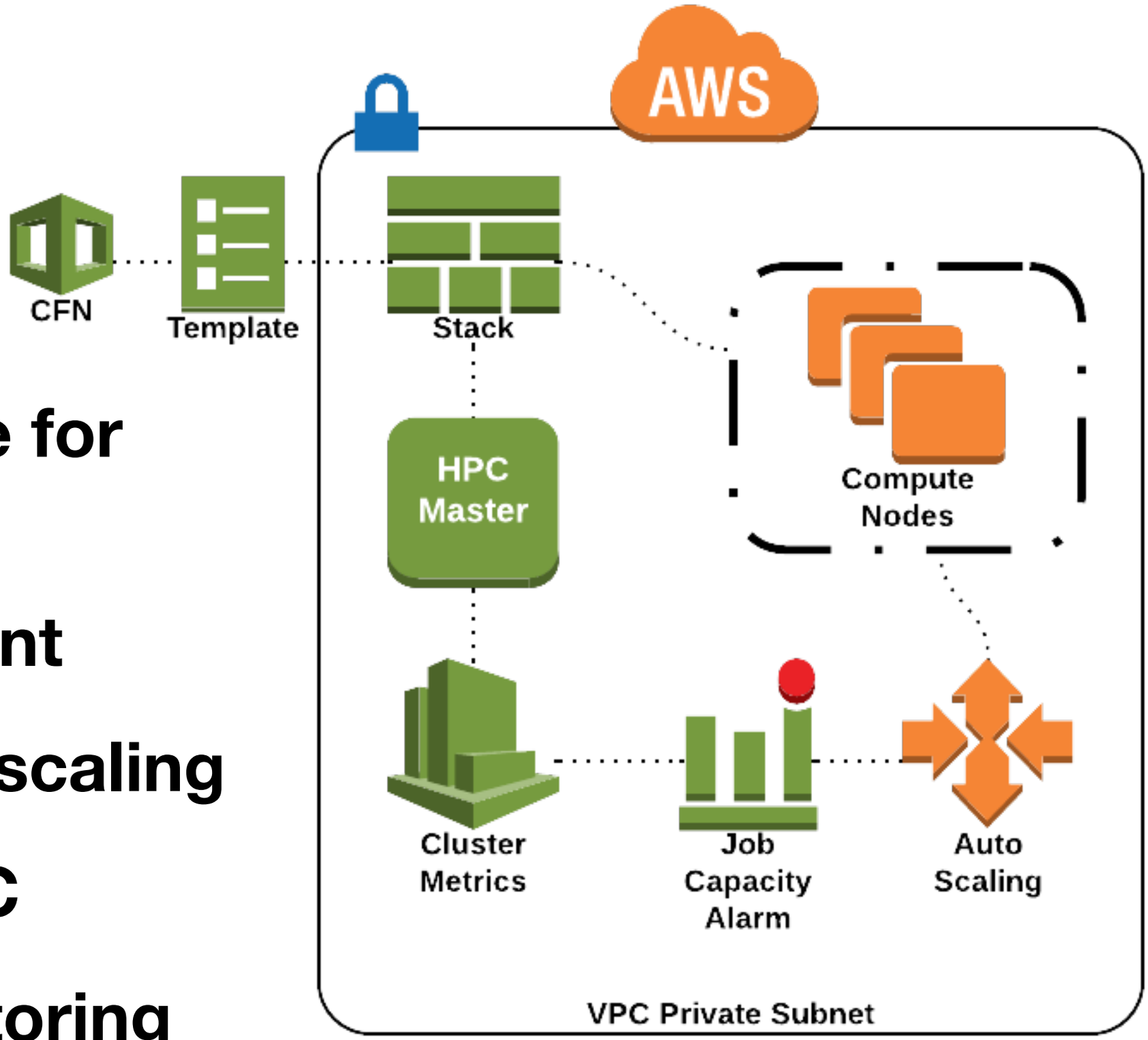
AWS CloudFormation

- **Provisioning and Orchestration service**
- **Declarative JSON syntax with dependency injection**
- **Create, Update, and Delete Stacks**
- **Stacks are parameterized and idempotent**
- **Stack resources are namespace and tagged**

Infrastructure as Code

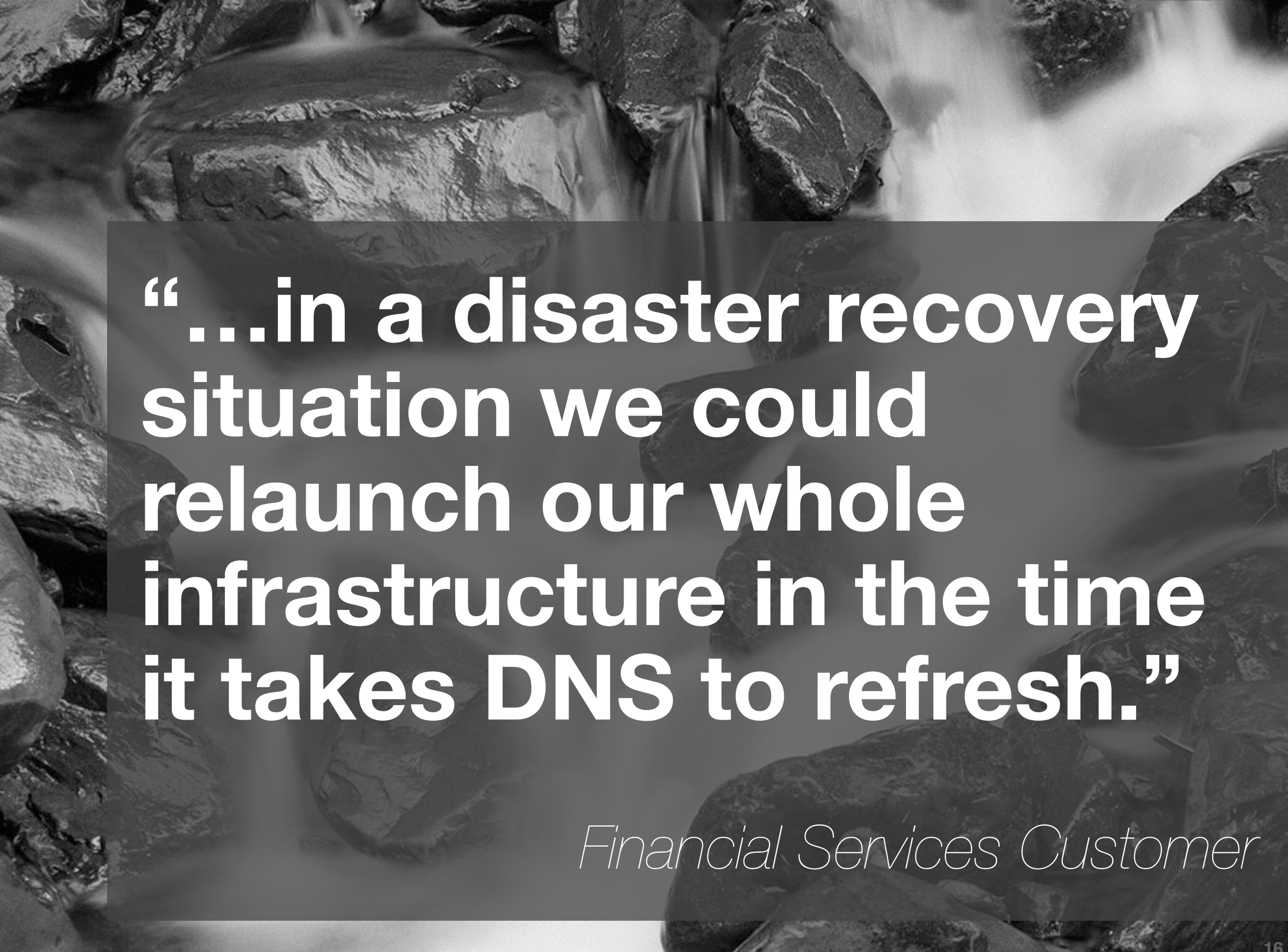
CFNCluster

- **CFN template for HPC clusters**
- **Stateless client**
- **Custom Autoscaling**
- **Supports VPC**
- **Built-in monitoring**



Tools and Techniques

Configuration Management



“…in a disaster recovery situation we could relaunch our whole infrastructure in the time it takes DNS to refresh.”

Financial Services Customer

Tool Building vs Automation

Tool Building vs Automation

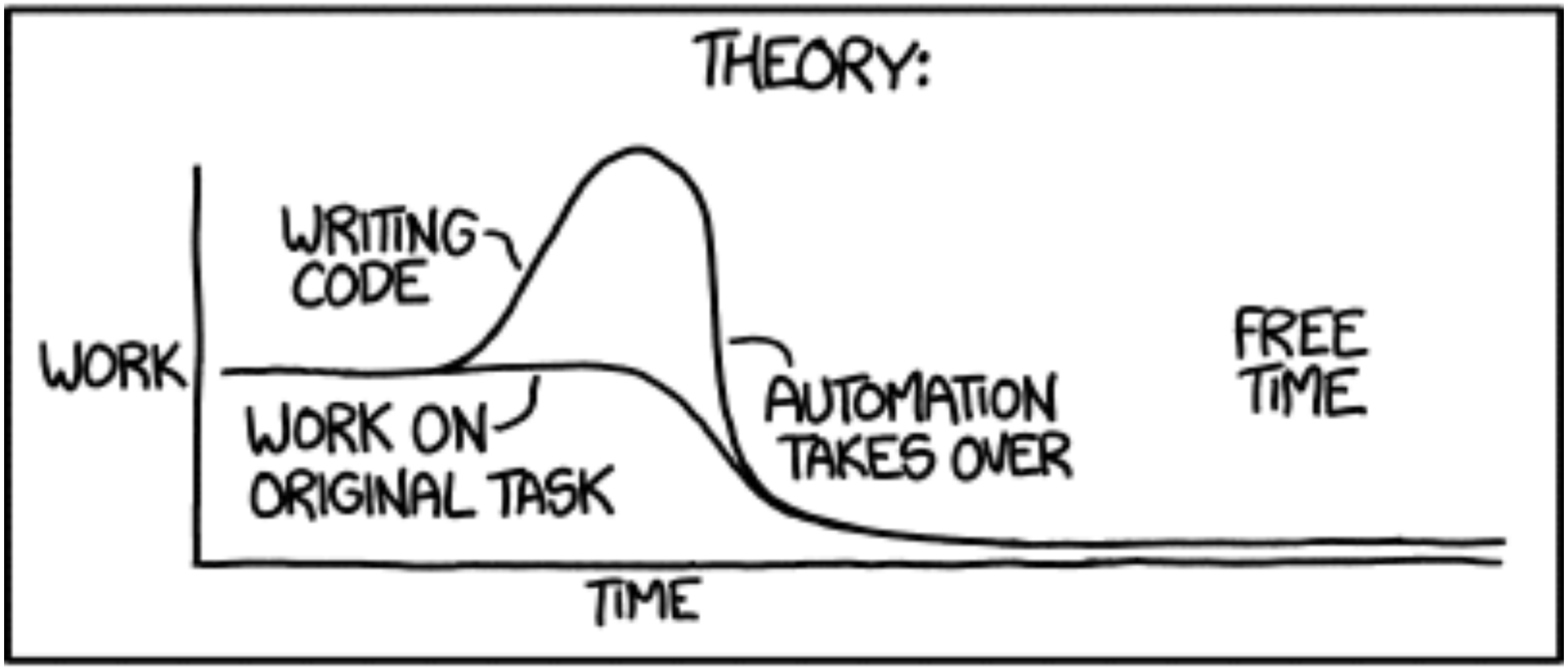
Tool Building

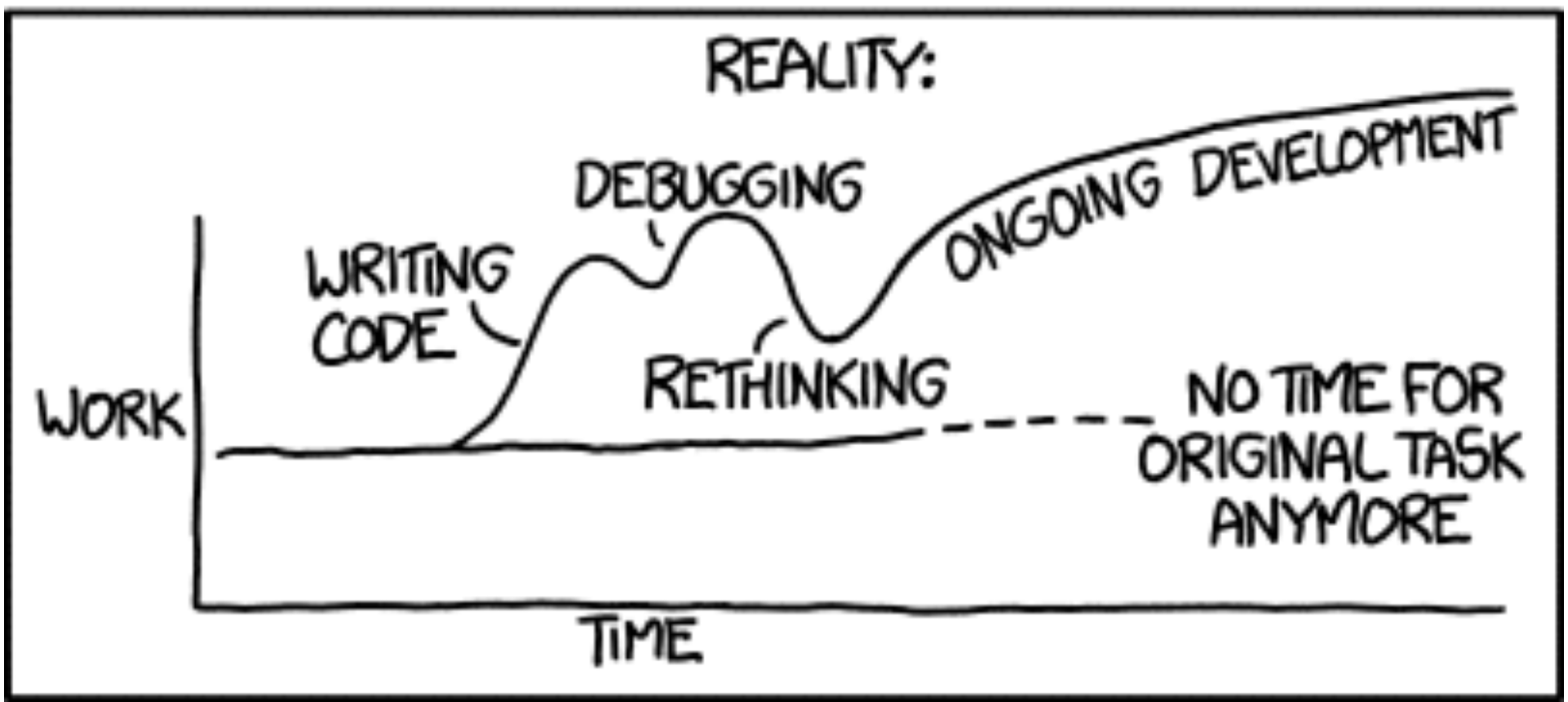
- Makes hard or impossible tasks easier
- Tools can be semi-automatic
- Can be used in unexpected ways
- Tools can be combined and composed with other tools

Automation

- Removes manual intervention completely
- Reduces human error
- Requires time and effort to maintain

"I SPEND A LOT OF TIME ON THIS TASK.
I SHOULD WRITE A PROGRAM AUTOMATING IT!"

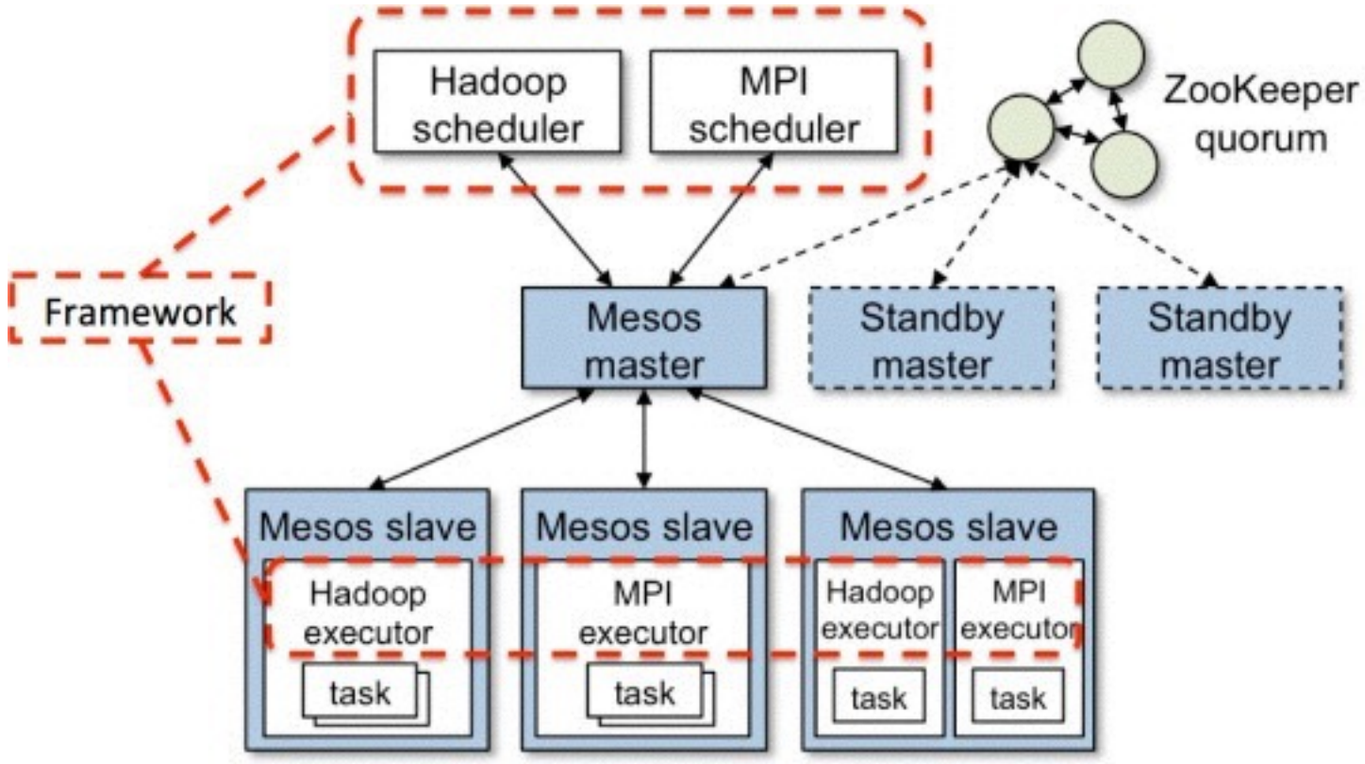




Apache Mesos

Software Defined Data Centers

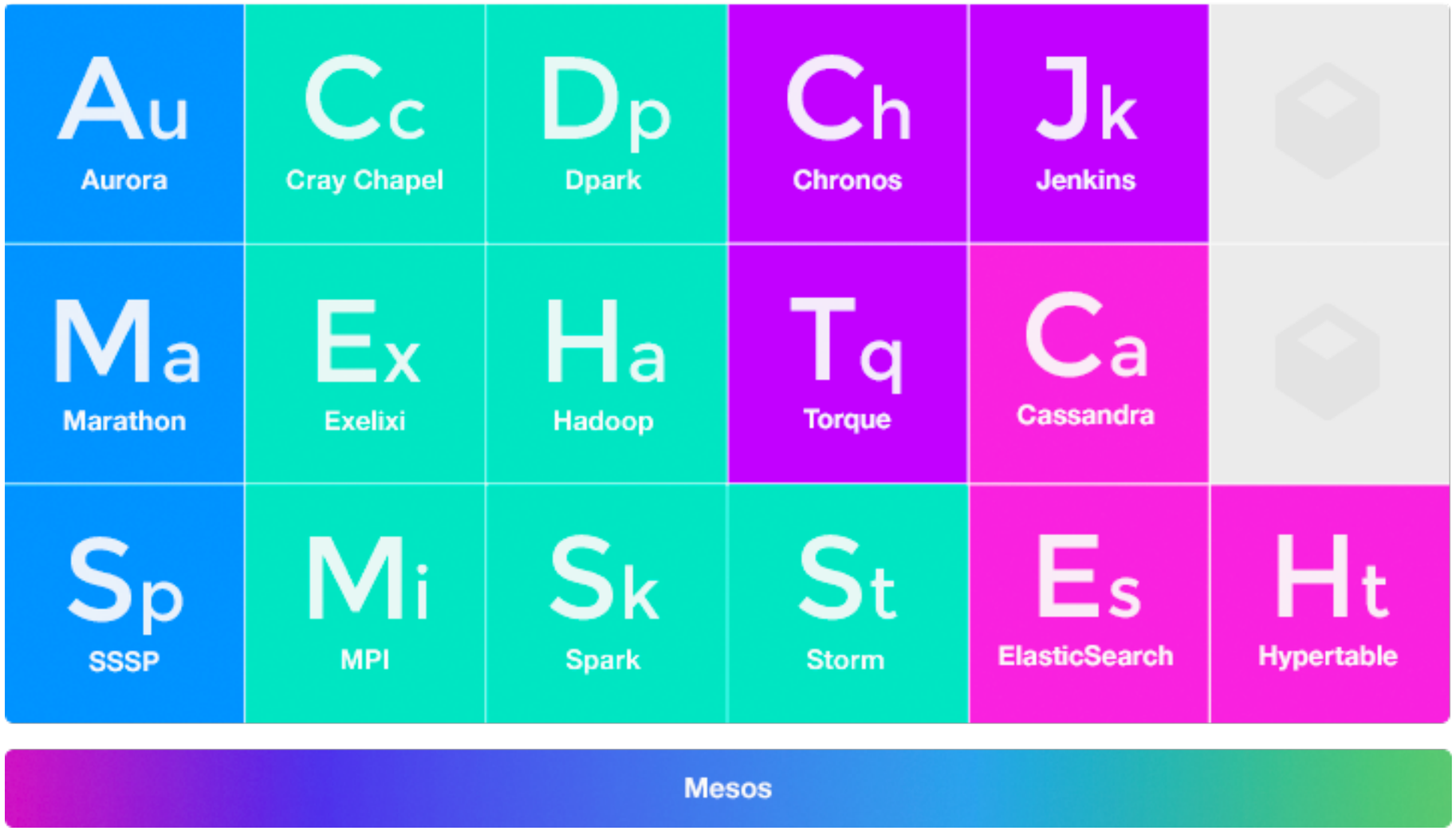
- **Two-level scheduler architecture**
- **Solves the multi-framework problem**



http://people.csail.mit.edu/matei/papers/2011/nsdi_mesos.pdf

Apache Mesos

Software Defined Data Centers



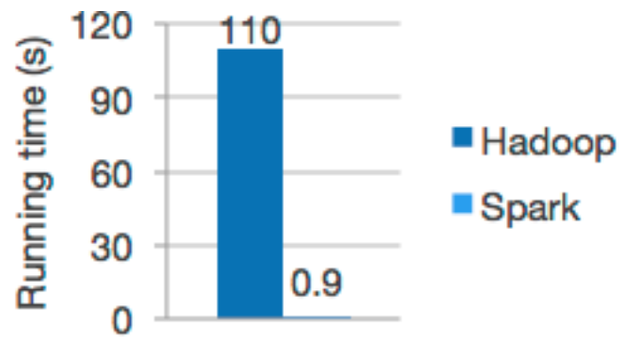
Scalable Software

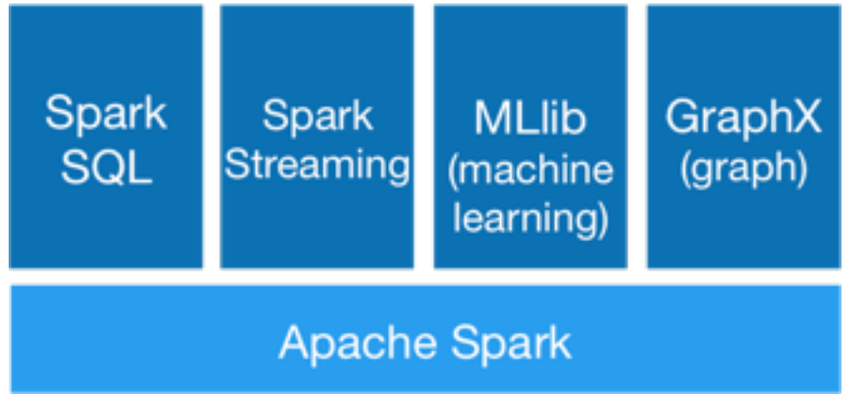
Hadoop and friends

Apache Spark

Lightning-fast cluster computing

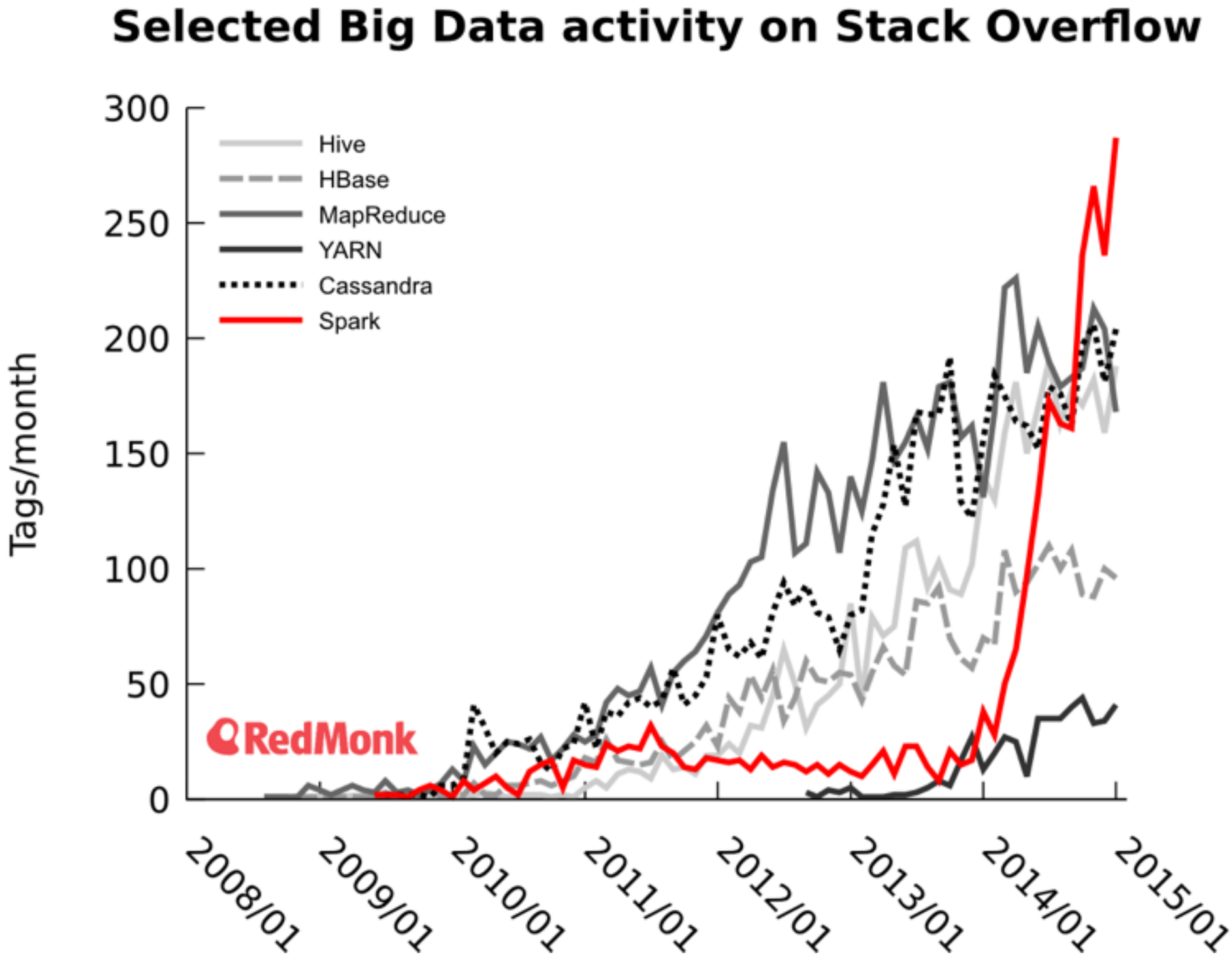
- **Provides high-level API's for Java, Scala, Python**
- **Rich set of tools for SQL, machine learning, graph processing, and streaming**
- **https://github.com/googlegenomics/spark-examples**





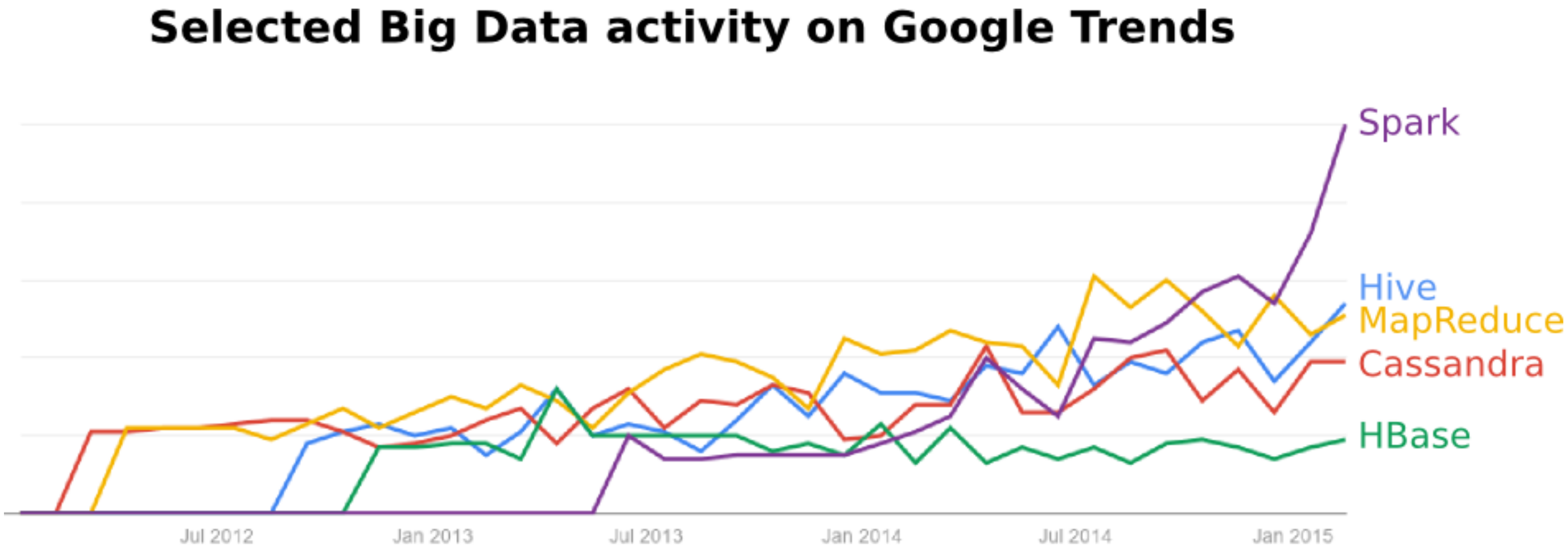
Apache Spark

Lightning-fast adoption



Apache Spark

Lightning-fast adoption

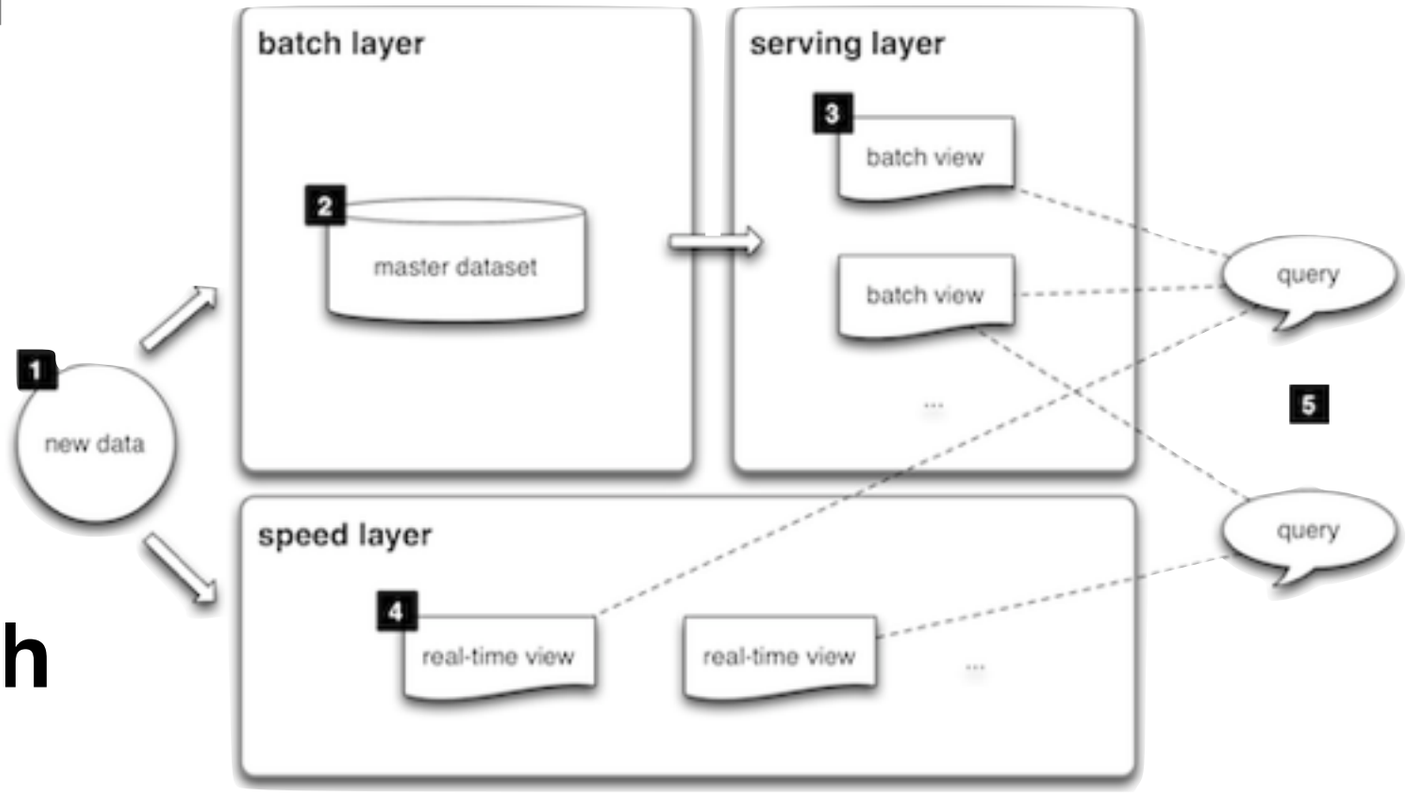


λ = f(all data)

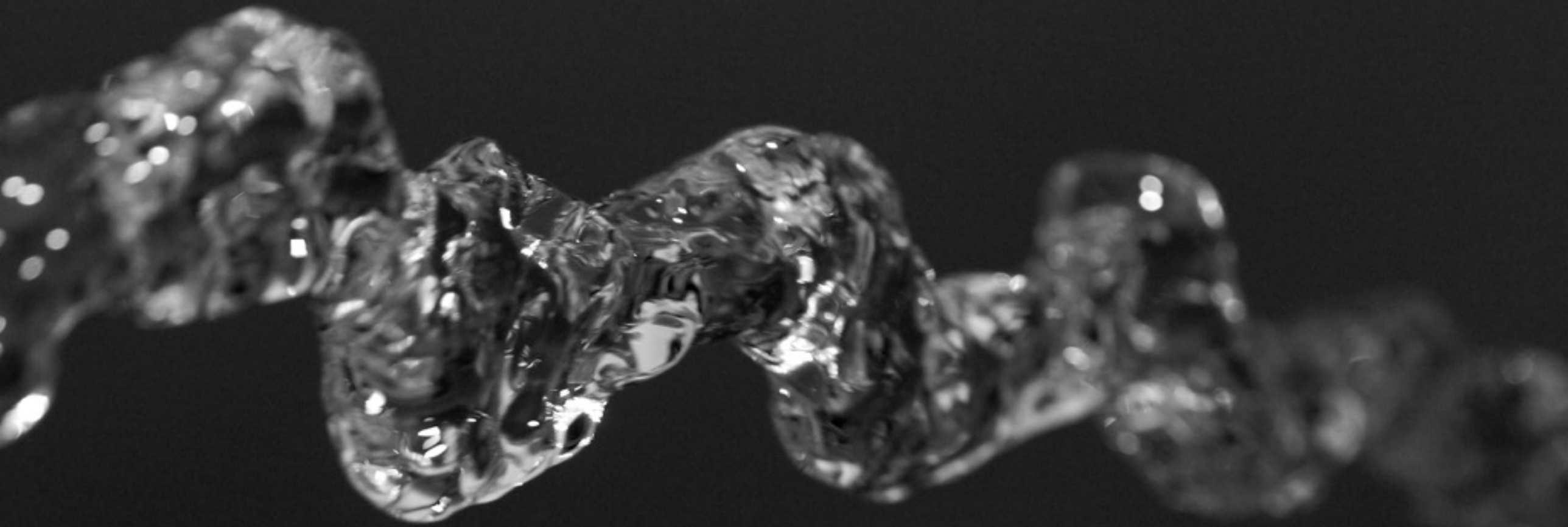
Lamba Architecture

generic, scalable and fault-tolerant data processing architecture

- **All data is dispatched to batch and speed layers**
- **Serving layer indexes the batch views**
- **Speed layer deals with recent data only**
- **Any incoming query can be answered by merging results from batch and speed layers**



lamba-architecture.net



Go with the Flow

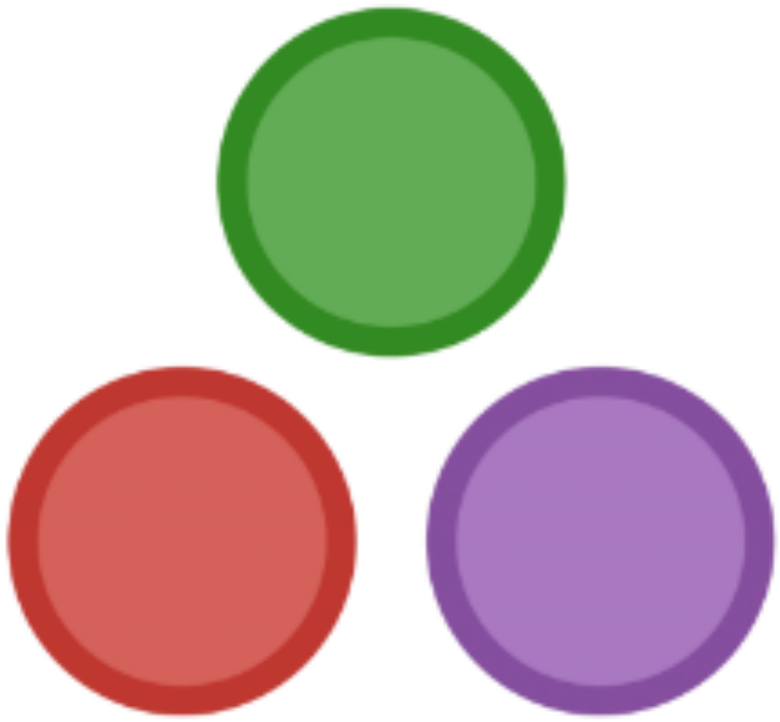
Data flow frameworks

- **AWS Kinesis**
- **AWS Lambda**
- **Google Cloud Data Flow**
- **DIY**

I <3 Julia

A fresh approach to technical computing

- **Homoiconic; Dynamic type system**
- **Designed for parallelism and distributed computation**
- **MATLAB-like syntax and extensive math library**
- **Call C functions directly**
- **Call Python functions**
- **IJulia Notebook**
- **Open Source**



Converged IT

Platform

- **Resource management**
- **Automatic storage allocation**
- **Service discovery**
- **Coordination**

Converged IT

Principals

- **Everything is distributed**
- **Everything is monitored**
- **Everything has an API**
- **Everything runs in a container**
- **Everything has a lifecycle policy**

Monitoring and Logging

- ***“What gets measured, gets managed.”***
- **Metrics, metrics, metrics**
- **Won’t know you needed it until it’s too late**
- **CloudWatch, CloudTrail**
- **ELK Stack**
  - Elasticsearch, Logstash, Kibana

Smooth is Fast

Think Parallel

--maxcpus=1000000

1,000,000 CPU's

Warehouse Scale Computing

google.com/datacenters

The Datacenter as a Computer

Concepts

Replication	Improve performance and reliability. Updates are more complicated.
Partitioning	Splitting data into smaller fragments across a large number of machines.
Load Balancing	Dynamically adjust load by selecting which servers dispatch to a request new to.
Health checking	Check connection level responsiveness and abort long-running requests.
Compression	Application specific. Fit as much of the working set as possible in DRAM.
Eventual Consistency	Relax consistency requirements for limited periods provided the system eventually returns to a stable state.

Pattern Recognition

Pattern Recognition

Design for humans

- **Humans are experts at pattern recognition**
  - Chunking
- **Languages and Frameworks do matter**
  - Building the right abstractions
  - Human Factors
  - Human Computer Interaction

Learn By Doing

The End