Cloud-scale Sequence Analysis

Michael Schatz

March 18, 2013

NY Genome Center / AWS

Outline

1. The need for cloud computing

2. Cloud-scale applications

3. Challenges and opportunities

Big Data in Bioinformatics

<Insert Moore's Law Graph Here>

Huge need for:

trimming/qc,

aligning,

variant detection,

de novo assembly,

expression quantification,

peak finding

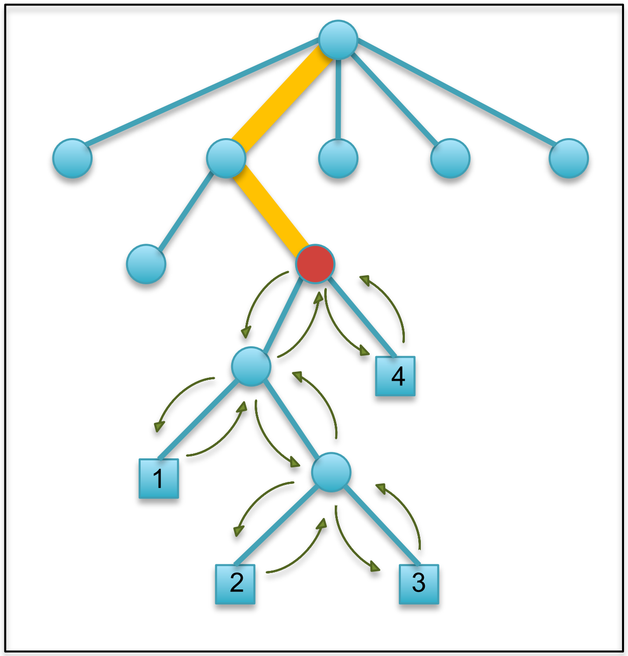
clustering

…

MUMmerGPU

http://mummergpu.sourceforge.net

- Index reference using a suffix tree
  - Each suffix represented by path from root
  - Reorder tree along space filling curve
- Map many reads simultaneously on GPU
  - Find matches by walking the tree
  - Find coordinates with depth first search
- Performance on nVidia GTX 8800
  - Match kernel was ~10x faster than CPU
  - Search kernel was ~4x faster than CPU
  - End-to-end runtime ~4x faster than CPU



- Cores are only part of the solution.
- Need storage, fast IO
- Locality is king

High-throughput sequence alignment using Graphics Processing Units.
Schatz, MC, Trapnell, C, Delcher, AL, Varshney, A. (2007) BMC Bioinformatics 8:474.

Web-Scale Information Processing

Jimmy Lin
The iSchool
University of Maryland

Monday, January 28, 2008

Material adapted from slides by Christophe Bisciglia, Aaron Kimball, & Sierra Michels-Slettvet, Google Distributed Computing Seminar, 2007 (licensed under Creation Commons Attribution 3.0 License)

Hadoop MapReduce

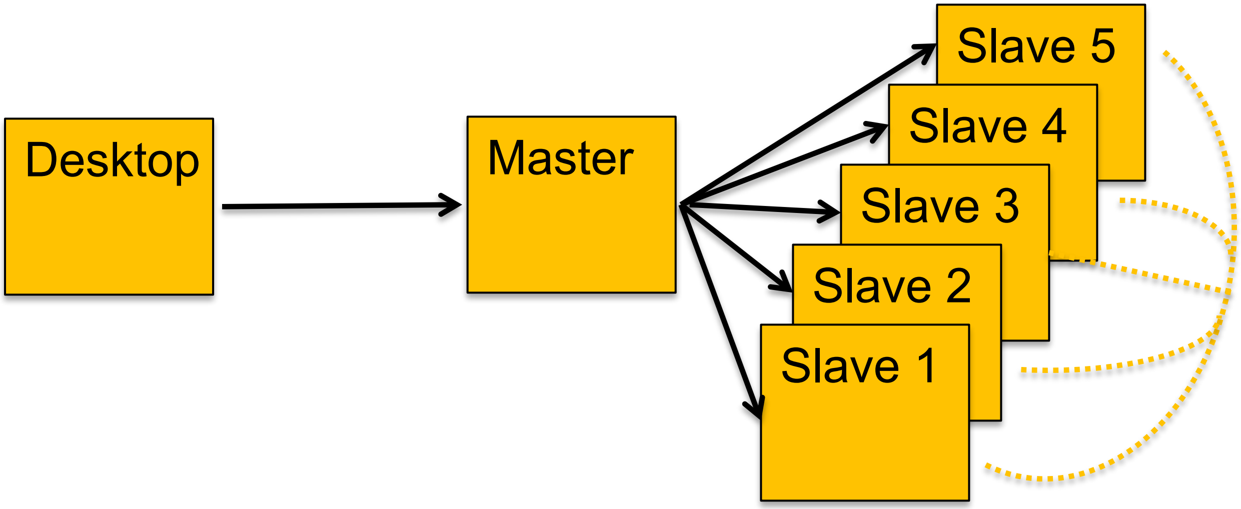
http://hadoop.apache.org

- MapReduce is Google's framework for large data computations
  - Data and computations are spread over thousands of computers
    - Indexing the Internet, PageRank, Machine Learning, etc… (Dean and Ghemawat, 2004)
    - 946PB processed in May 2010 (Jeff Dean at Stanford, 11.10.2010)
  - Hadoop is the leading open source implementation
    - Developed and used by Yahoo, Facebook, Twitter, Amazon, etc
    - GATK is an alternative implementation specifically for NGS

- Benefits
  - Scalable, Efficient, Reliable
  - Easy to Program
  - Runs on commodity computers

- Challenges
  - Redesigning / Retooling applications
    - Not Condor, Not MPI
    - Everything in MapReduce

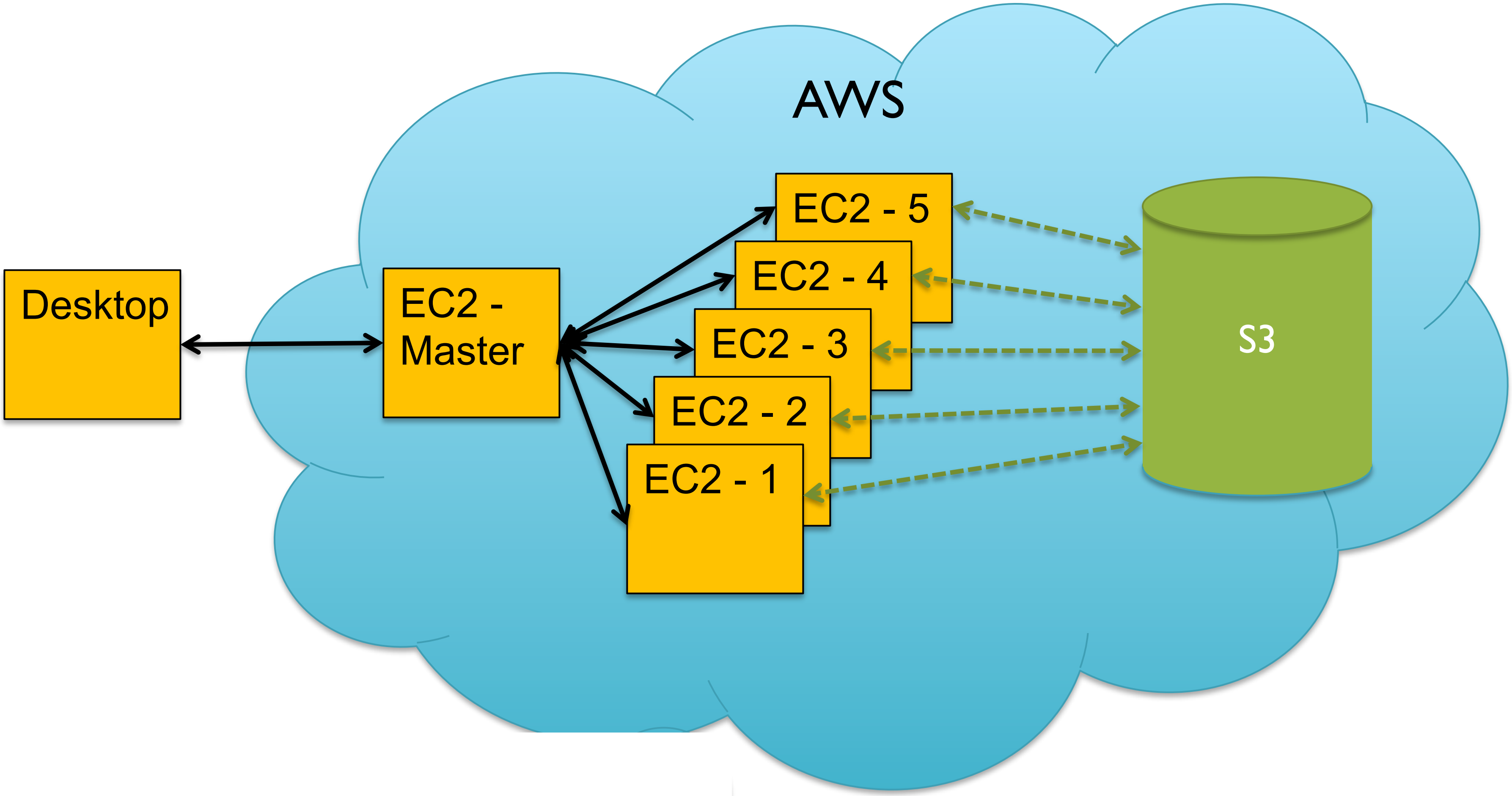
System Architecture



- Hadoop Distributed File System (HDFS)
  - Data files partitioned into large chunks (64MB), replicated on multiple nodes
  - Computation moves to the data, rack-aware scheduling

- Hadoop MapReduce system won the 2009 GreySort Challenge
  - Sorted 100 TB in 173 min (578 GB/min) using 3452 nodes and 4x3452 disks

Hadoop on AWS



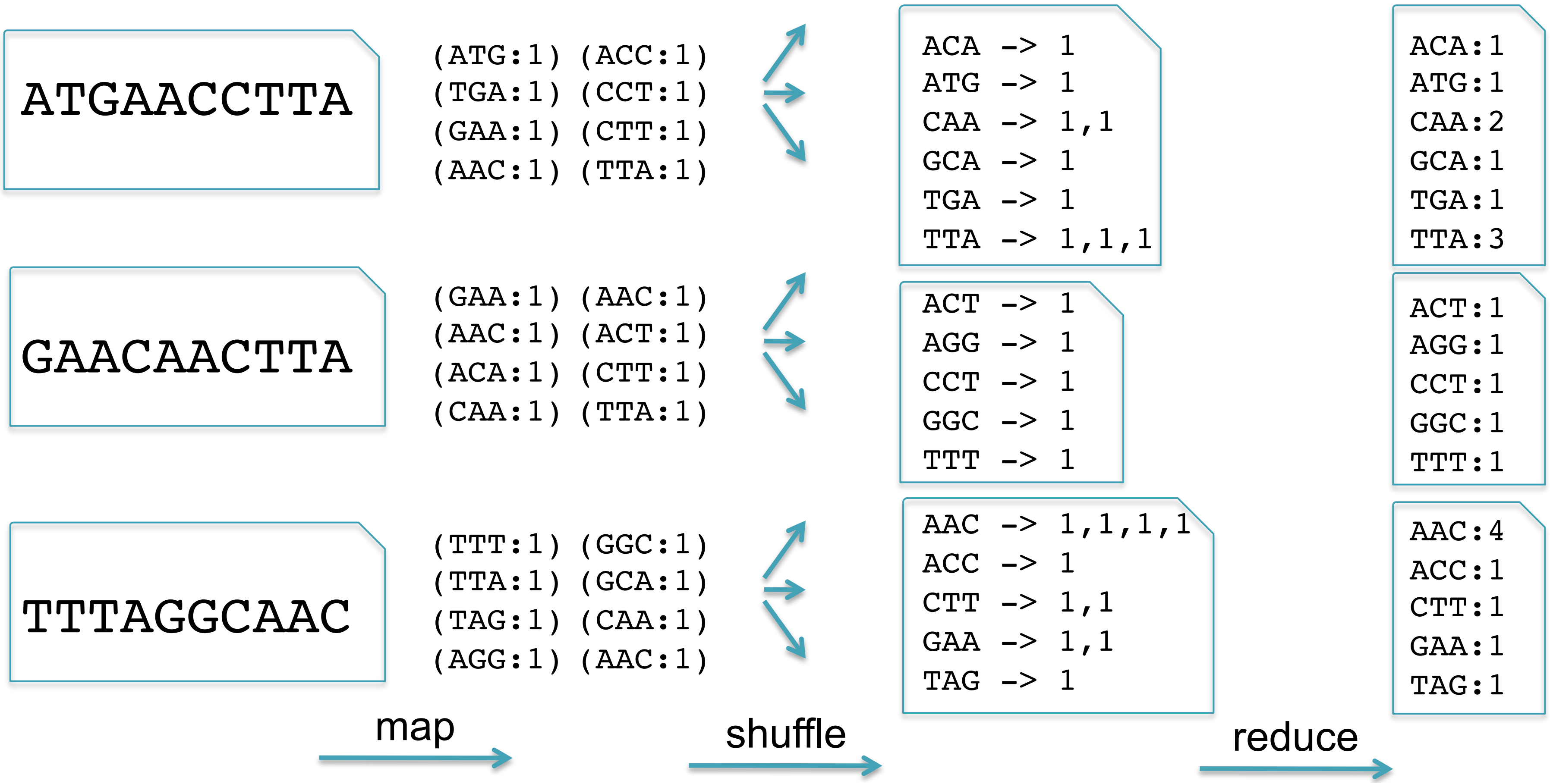
If you don't have 1000s of machines, rent them from Amazon

- After machines spool up, ssh to master as if it was a local machine.
- Use S3 for persistent data storage, with very fast interconnect to EC2.

K-mer Counting

- Application developers focus on 2 (+1 internal) functions
  - Map: input ➔ key:value pairs
  - Shuffle: Group together pairs with same key
  - Reduce: key, value-lists ➔ output

Map, Shuffle & Reduce
All Run in Parallel



```
ATGAACCTTA

(ATG:1) (ACC:1)
(TGA:1) (CCT:1)
(GAA:1) (CTT:1)
(AAC:1) (TTA:1)
```

```
GAACAACTTA

(GAA:1) (AAC:1)
(AAC:1) (ACT:1)
(ACA:1) (CTT:1)
(CAA:1) (TTA:1)
```

```
TTTAGGCAAC

(TTT:1) (GGC:1)
(TTA:1) (GCA:1)
(TAG:1) (CAA:1)
(AGG:1) (AAC:1)
```

```
ACA -> 1
ATG -> 1
CAA -> 1,1
GCA -> 1
TGA -> 1
TTA -> 1,1,1
```

```
ACT -> 1
AGG -> 1
CCT -> 1
GGC -> 1
TTT -> 1
```

```
AAC -> 1,1,1,1
ACC -> 1
CTT -> 1,1
GAA -> 1,1
TAG -> 1
```

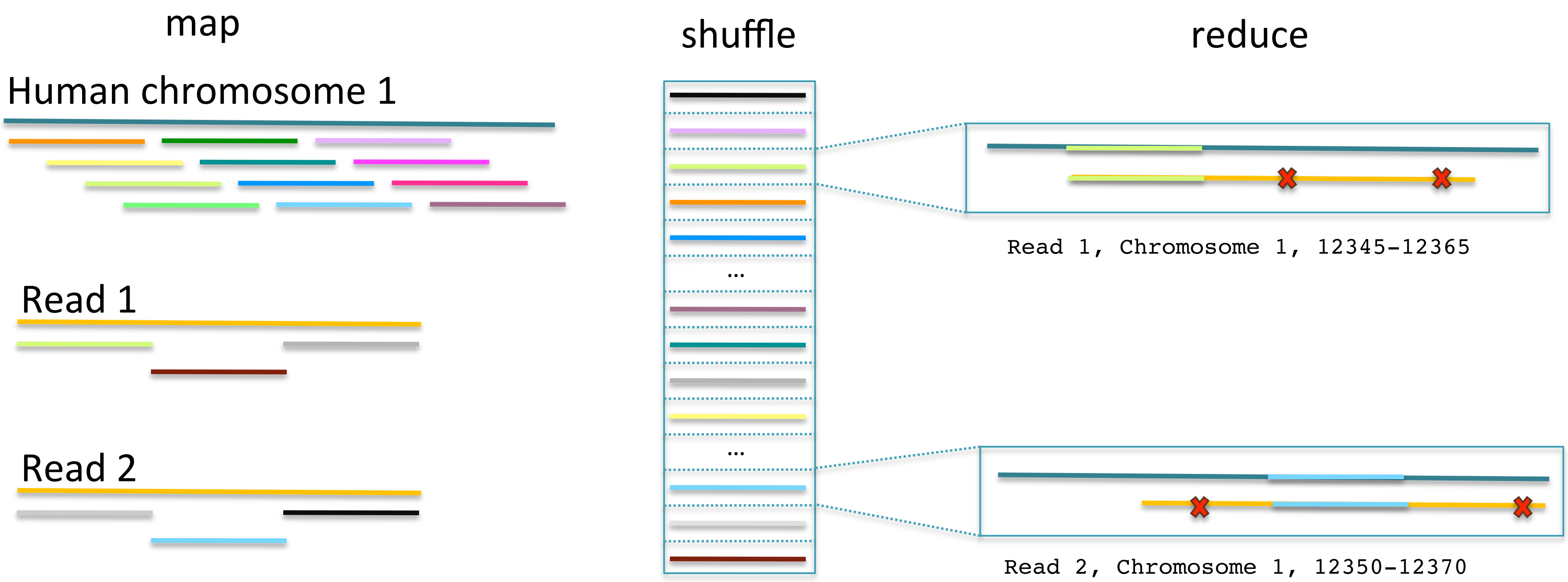
```
ACA:1
ATG:1
CAA:2
GCA:1
TGA:1
TTA:3
```

```
ACT:1
AGG:1
CCT:1
GGC:1
TTT:1
```

```
AAC:4
ACC:1
CTT:1
GAA:1
TAG:1
```

CloudBurst

1. Map: Catalog K-mers
   - Emit k-mers in the genome and reads
2. Shuffle: Collect Seeds
   - Conceptually build a hash table of k-mers and their occurrences
3. Reduce: End-to-end alignment
   - If read aligns end-to-end with ≤ k errors, record the alignment



map

shuffle

reduce

Human chromosome 1

Read 1

Read 2

…

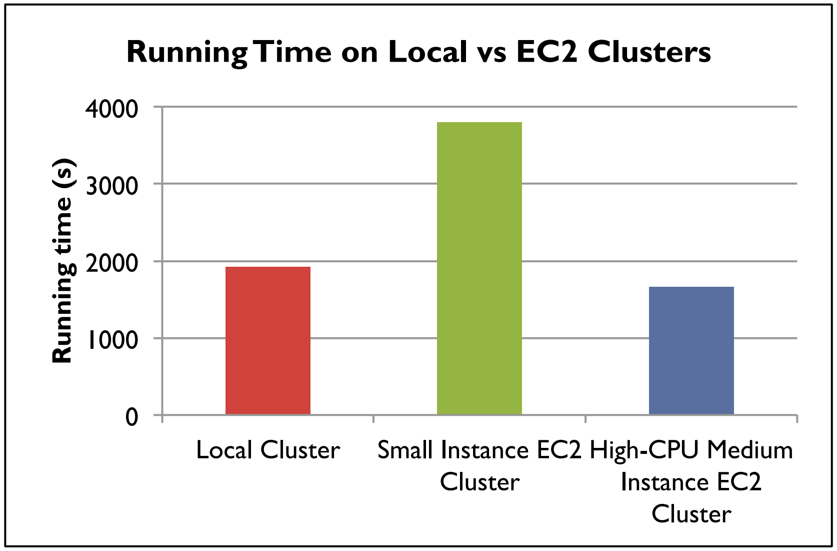
…

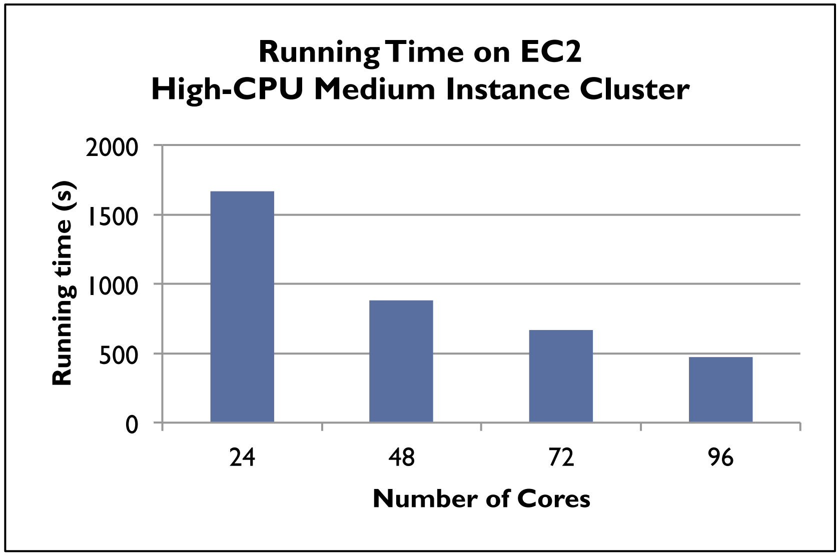
```
Read 1, Chromosome 1, 12345-12365
```

```
Read 2, Chromosome 1, 12350-12370
```

CloudBurst: Highly Sensitive Read Mapping with MapReduce.
Schatz, MC (2009) *Bioinformatics*. 25:1363-1369

AWS EC2 Performance



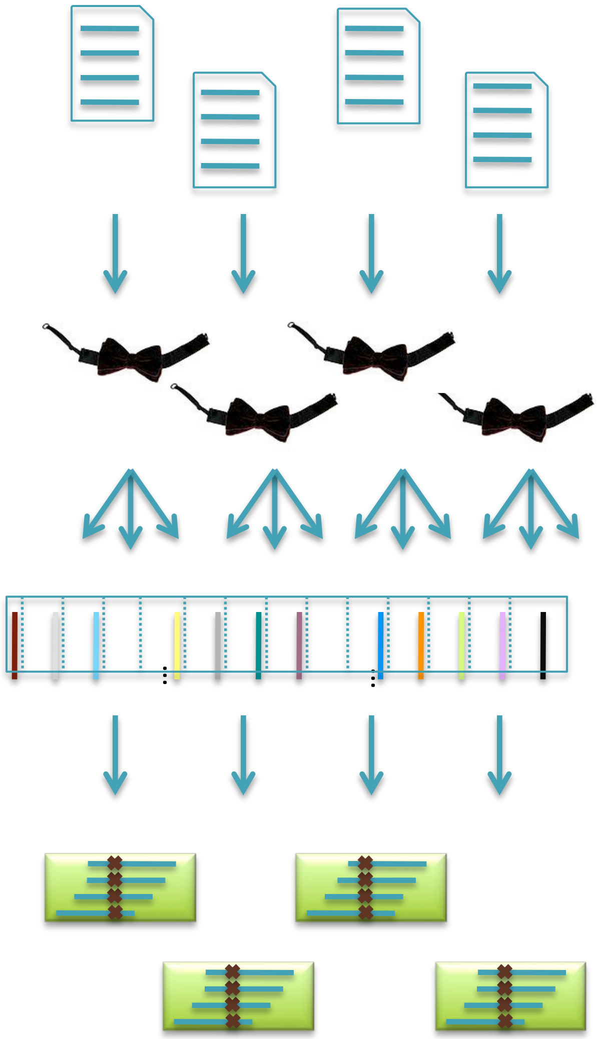


- CloudBurst running times for mapping 7M reads to human chromosome 22 with at most 4 mismatches on the local and EC 2 clusters.
  - The 24-core Amazon High-CPU Medium Instance EC2 cluster is faster than the 24-core Small Instance EC2 cluster, and the 24-core local dedicated cluster.
  - The 96-core cluster on AWS was 100x faster than serial RMAP.

- Cloud can be very effective for genomics
- When computing at scale, space is time
- Implementing from scratch is expensive

Crossbow

http://bowtie-bio.sourceforge.net/crossbow



- Align billions of reads and find SNPs
  - Reuse software components: Hadoop Streaming
- Map: Bowtie (Langmead *et al.*, 2009)
  - Find best alignment for each read
  - Emit (chromosome region, alignment)
- Shuffle: Hadoop
  - Group and sort alignments by region
- Reduce: SOAPsnp (Li *et al.*, 2009)
  - Scan alignments for divergent columns
  - Accounts for sequencing error, known SNPs

Searching for SNPs with Cloud Computing.
Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL (2009) *Genome Biology.* **10**:R134

Performance in Amazon EC2

	Asian Individual Genome		
Data Loading	3.3 B reads	106.5 GB	$10.65
Data Transfer	1h :15m	40 cores	$3.40
Setup	0h : 15m	320 cores	$13.94
Alignment	1h : 30m	320 cores	$41.82
Variant Calling	1h : 00m	320 cores	$27.88
End-to-end	4h : 00m		$97.69

Discovered 3.7M SNPs in one human genome for ~$100 in an afternoon.
Accuracy validated at >99%

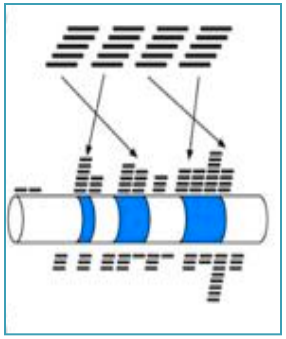
- Very compelling example of cloud computing in genomics
- Transfer takes time, but totally depends on institution
- Need more applications!

Hadoop for NGS Analysis

Myrna

Cloud-scale differential gene expression for RNA-seq

Expression of 1.1 billion RNA-Seq reads in ~2 hours for ~$66



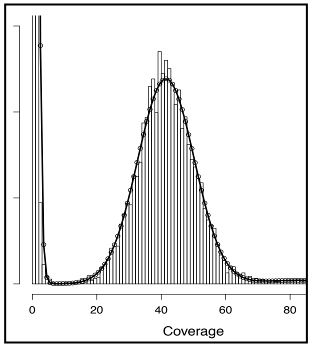
(Langmead, Hansen, Leek, 2010)

http://bowtie-bio.sf.net/myrna/

Quake

Quality-aware error correction of short reads

Correct 97.9% of errors with 99.9% accuracy



(Kelley, Schatz, Salzberg, 2010)

http://www.cbcb.umd.edu/software/quake/

Contrail

Assembly of Large Genomes Using Cloud Computing

Quickly assemble the human genome with hundreds of commodity cores

(Schatz, 2010)

http://contrail-bio.sf.net/

Genome Indexing

Rapid Parallel Construction of Genome Index

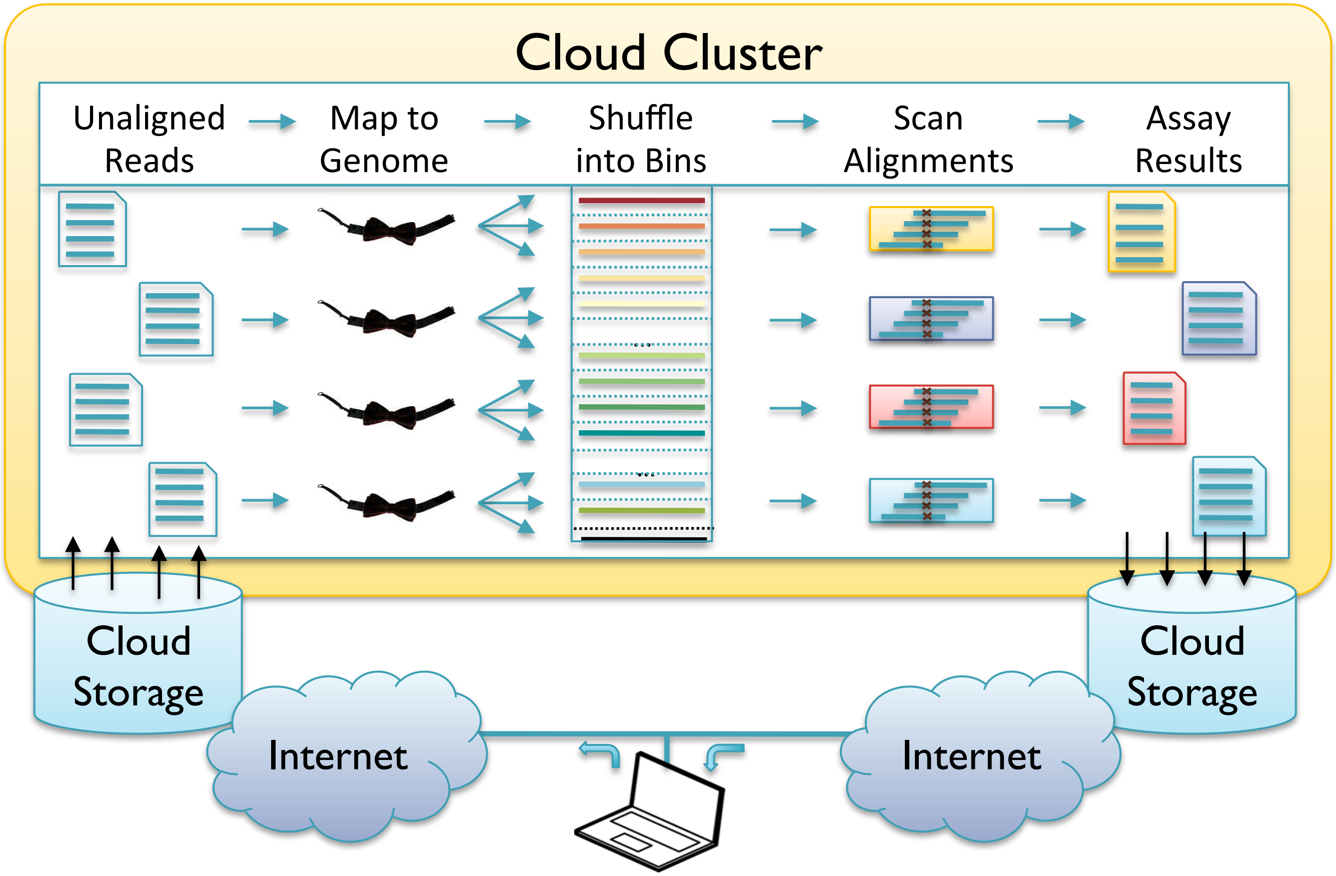
Construct the BWT of the human genome in 9 minutes

$GATTACA
A$GATTAC
ACA$GATT
ATTACA$G
CA$GATTA
GATTACA$
TACA$GAT
TTACA$GA

(Menon, Bhat, Schatz, 2011)

http://code.google.com/p/genome-indexing/

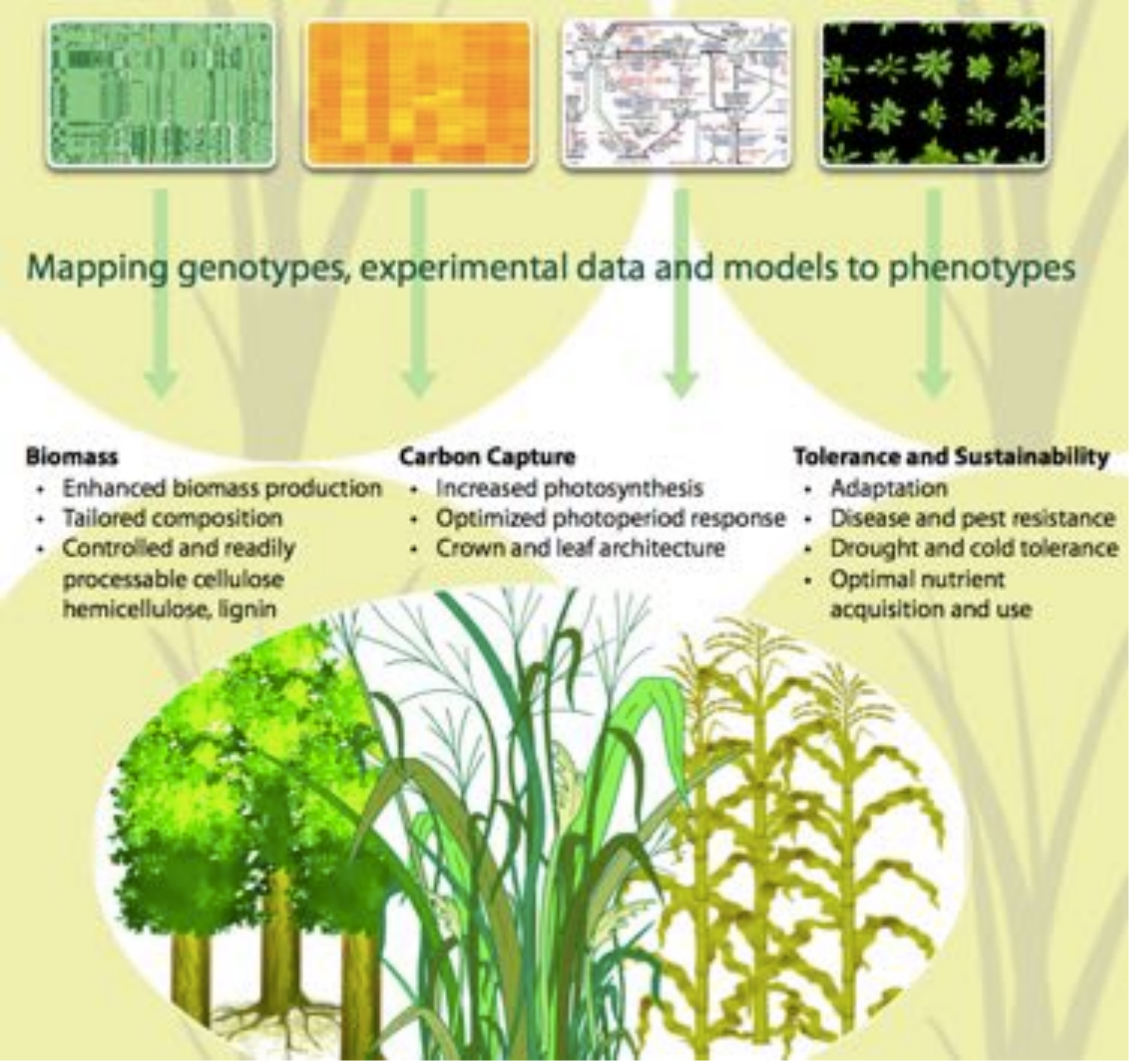
Map-Shuffle-Scan for Genomics



- Genomics+Cloud is very effective
- Need more applications, users, and a scientific goal

Cloud Computing and the DNA Data Race.
Schatz, MC, Langmead B, Salzberg SL (2010) *Nature Biotechnology.* **28**:691-693

Systems Biology Knowledgebase



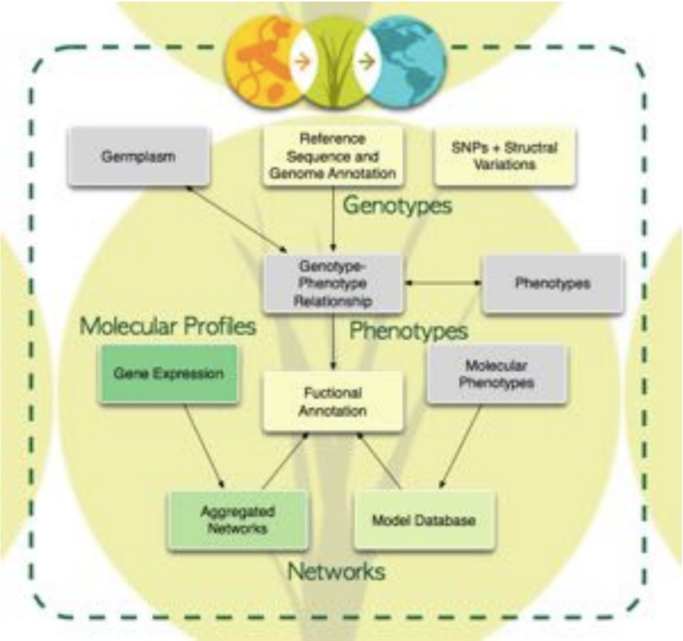
Mapping genotypes, experimental data and models to phenotypes

Biomass
- Enhanced biomass production
- Tailored composition
- Controlled and readily processable cellulose hemicellulose, lignin

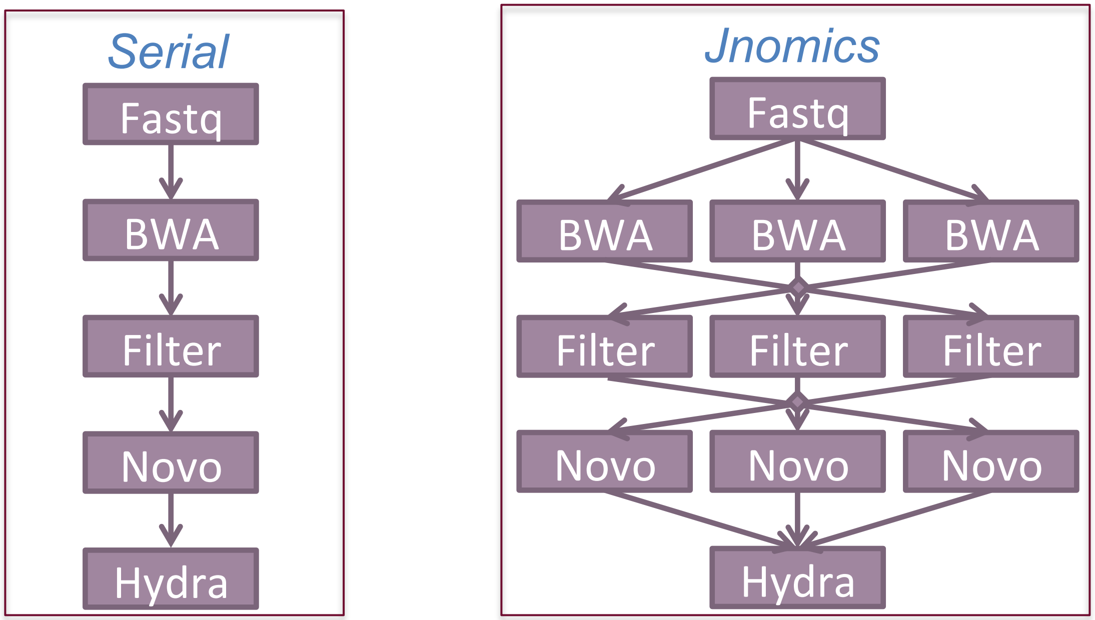
Carbon Capture
- Increased photosynthesis
- Optimized photoperiod response
- Crown and leaf architecture

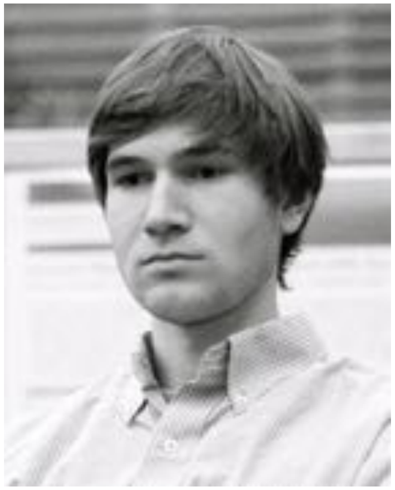
Tolerance and Sustainability
- Adaptation
- Disease and pest resistance
- Drought and cold tolerance
- Optimal nutrient acquisition and use

Model development
Hypothesis testing
Knowledge Synthesis



Jnomics: Cloud-scale genomics





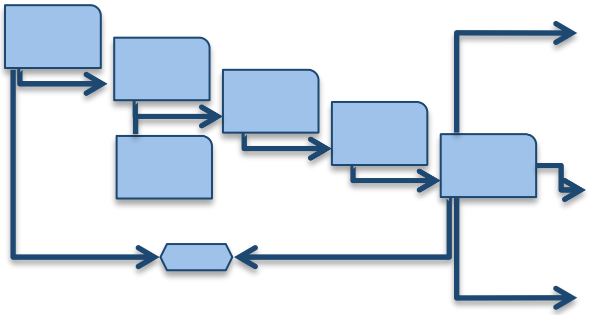
- Rapid parallel execution of data-intensive analysis
  - FASTX, BWA, Bowtie2, Novoalign, SAMTools, Hydra
  - Sorting, merging, filtering, selection, clustering, correlating
  - Supports BAM, SAM, BED, fastq

Answering the demands of digital genomics
Titmus, MA, Gurtowski, J, Schatz, MC (2012) *Concurrency & Computation*

Jnomics API 1.0

Genotyping API

- **Bowtie**: Launch alignment task with Bowtie
- **BWA**: Launch alignment task with BWA
- **SNPCalling**: Launch SNPcalling task with SAMTools
- **SortAlignments**: Launch task to sort by chromosome



Job API

- **ClusterStatus**: return basic status of cluster (jobs running, nodes available, etc)
- **JobStatus**: Given a JobID, returns current status
- **ListJobs**: List JobID running with a given username
- **KillJob**: Kills a given JobID

Data API

- **List**: List files in a directory
- **Fetch**: Fetch files from HDFS
- **Put**: Put files into HDFS
- **RM**: Delete files on HDFS
- **FetchBAM**: On-the-fly conversion to BAM
- **PutFastq**: Put reads into HDFS with conversion

Notes:

- All calls are authenticated with KBase username/password

Maize Population Analysis

Align & call SNPs from 131 maize samples
1TB fastq / 408Gbp input data

	Serial	KBase cloud (small)	KBase Cloud (large)
Config	1 core (1 node)	210 cores (15 nodes)	854 cores (61 nodes)
Bowtie2	1311 hr*	19.5 hr	5 hr
Sort	58 hr*	N/A	N/A
Samtools	58 hr*	3.5 hr	1.5 hr
End-to-End Speedup	1427 hr* 1x	23 hr 62x	6.5 hr 219x

*estimated time

Summary

Staying afloat in the data deluge means computing in parallel

- Hadoop + Cloud computing is an attractive platform for large scale sequence analysis, computation, and collaboration

Diversity is the biggest barrier to adoption

1. Diversity of applications
   - Long tail distribution of critical to experimental
2. Diversity of requirements
   - Storage, Network, IO, cache, RAM, cores
3. Diversity of data
   - Datatypes, scale, formats, available bandwidth
4. Diversity of users
   - Super-scripters to point-and-click users

Acknowledgements

<u>Schatz Lab</u>

Giuseppe Narzisi
Shoshana Marcus
Jeremy Lewi
Hayan Lee
Rob Aboukhalil
Mitch Bekritsky
Charles Underwood
James Gurtowski
Rushil Gupta
Avijit Gupta
Shishir Horane
Deepak Nettem
Varrun Ramani
Piyush Kansal
Alejandro Wences
Eric Biggers
Aspyn Palatnick

<u>CSHL</u>

Hannon Lab
Iossifov Lab
Levy Lab
Lippman Lab
Lyon Lab
Martienssen Lab
McCombie Lab
Ware Lab
Wigler Lab

IT Department

<u>UMD/JHU</u>

Steven Salzberg
Mihai Pop
Ben Langmead
Cole Trapnell

amazon webservices™

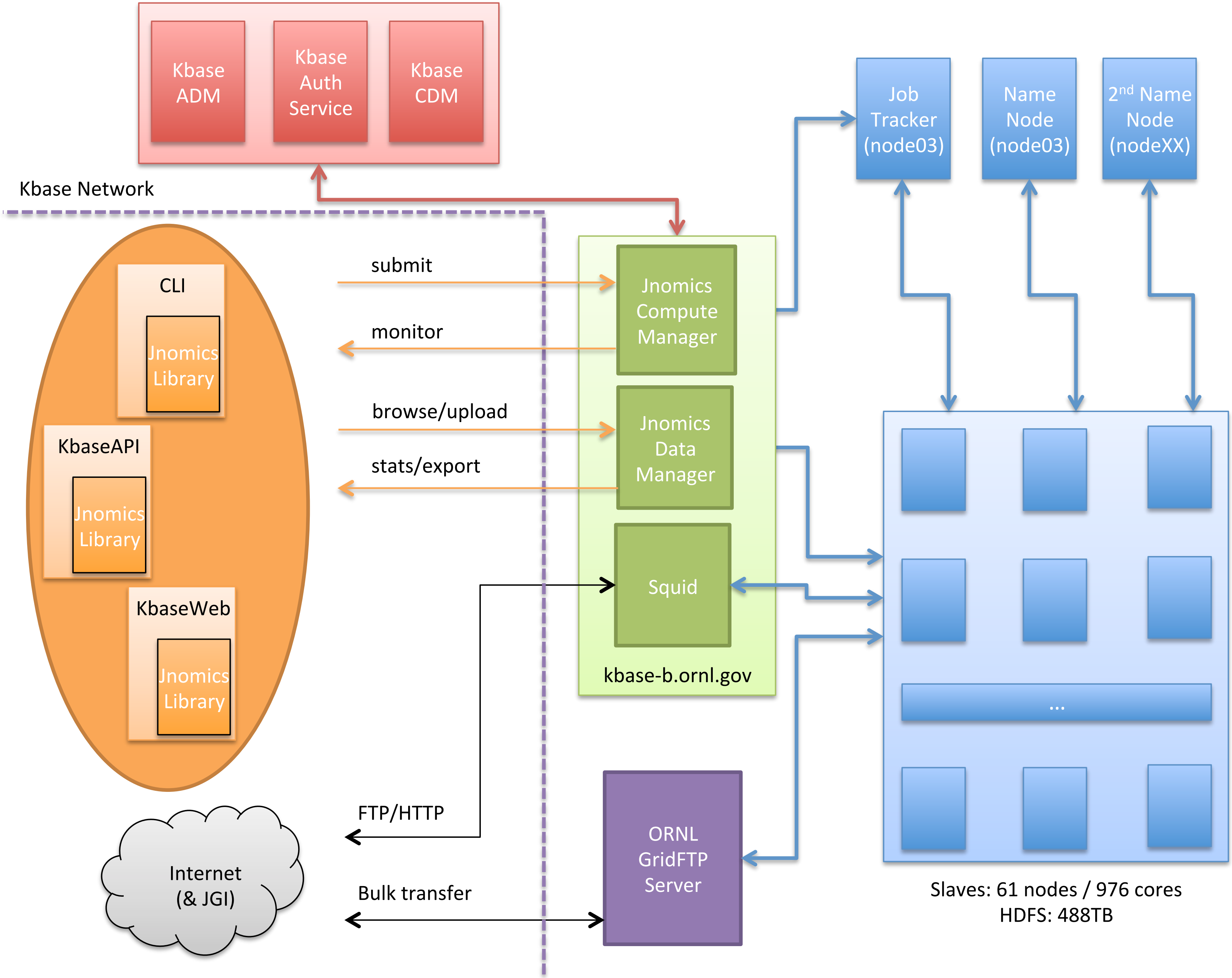
National Human Genome Research Institute

U.S. DEPARTMENT OF ENERGY

NSF

Thank You!

http://schatzlab.cshl.edu
@mike_schatz


Kbase ADM
Kbase Auth Service
Kbase CDM
Kbase Network
CLI
Jnomics Library
KbaseAPI
Jnomics Library
KbaseWeb
Jnomics Library
submit
monitor
browse/upload
stats/export
Jnomics Compute Manager
Jnomics Data Manager
Squid
kbase-b.ornl.gov
Job Tracker (node03)
Name Node (node03)
2nd Name Node (nodeXX)
...
Slaves: 61 nodes / 976 cores
HDFS: 488TB
FTP/HTTP
Bulk transfer
Internet (& JGI)
ORNL GridFTP Server