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.
Web-Scale Information Processing

Jimmy Lin
The iSchool
University of Maryland

Monday, January 28, 2008

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

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 United States See http://creativecommons.org/licenses/by-nc-sa/3.0/us/ for details
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
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
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 \( \leq k \) errors, record the alignment

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

AWS EC2 Performance

- 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.
# Performance in Amazon EC2

<table>
<thead>
<tr>
<th>Step</th>
<th>Time</th>
<th>Cores</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Loading</td>
<td>3.3 B reads</td>
<td>106.5 GB</td>
<td>$10.65</td>
</tr>
<tr>
<td>Data Transfer</td>
<td>1h :15m</td>
<td>40 cores</td>
<td>$3.40</td>
</tr>
<tr>
<td>Setup</td>
<td>0h : 15m</td>
<td>320 cores</td>
<td>$13.94</td>
</tr>
<tr>
<td>Alignment</td>
<td>1h : 30m</td>
<td>320 cores</td>
<td>$41.82</td>
</tr>
<tr>
<td>Variant Calling</td>
<td>1h : 00m</td>
<td>320 cores</td>
<td>$27.88</td>
</tr>
<tr>
<td>End-to-end</td>
<td>4h : 00m</td>
<td></td>
<td>$97.69</td>
</tr>
</tbody>
</table>

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*

(Menon, Bhat, Schatz, 2011)  
http://code.google.com/p/genome-indexing/
Map-Shuffle-Scan for Genomics

Cloud Cluster

- Unaligned Reads → Map to Genome → Shuffle into Bins → Scan Alignments → Assay Results

Cloud Computing and the DNA Data Race.

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

Model development
Hypothesis testing
Knowledge Synthesis

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

Germline
- Reference Sequence and Genome Annotation
- SNPs + Structural Variations

Genotypes
- Genotype-Phenotype Relationship
- Functional Annotation

Molecular Profiles
- Gene Expression
- Functional Annotation

Phenotypes
- Aggregated Networks
- Model Database

Networks

DOE Systems Biology Knowledgebase
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*
**Jnomic 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

<table>
<thead>
<tr>
<th></th>
<th>Serial</th>
<th>KBase cloud (small)</th>
<th>KBase Cloud (large)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Config</strong></td>
<td>1 core (1 node)</td>
<td>210 cores (15 nodes)</td>
<td>854 cores (61 nodes)</td>
</tr>
<tr>
<td><strong>Bowtie2</strong></td>
<td>1311 hr*</td>
<td>19.5 hr</td>
<td>5 hr</td>
</tr>
<tr>
<td><strong>Sort</strong></td>
<td>58 hr*</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Samtools</strong></td>
<td>58 hr*</td>
<td>3.5 hr</td>
<td>1.5 hr</td>
</tr>
<tr>
<td><strong>End-to-End Speedup</strong></td>
<td>1427 hr*</td>
<td>23 hr</td>
<td>6.5 hr</td>
</tr>
</tbody>
</table>

*estimated time
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

Summary

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

Schatz Lab
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

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

IT Department

UMD/JHU
Steven Salzberg
Mihai Pop
Ben Langmead
Cole Trapnell

amazon web services™
National Human Genome Research Institute
U.S. DEPARTMENT OF ENERGY
NSF
Thank You!

http://schatzlab.cshl.edu
@mike_schatz