

Technical Challenges of Cloud Computing

Narayan Desai & Michael Schatz

Beyond the Genome
Oct 13, 2010

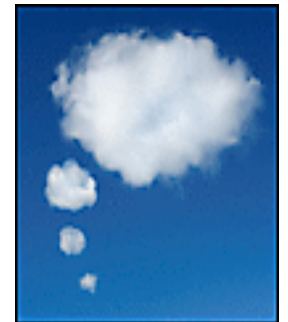


Data Challenges

- HiSeq 2000: ~ 200 Gbp / 8 day run
 - ~200GB compressed (gzip) : ~303KB/s
 - Total Volume: Sequences, QV, Read Ids
- 200 GB Transfer Time [75% efficiency]:
 - Cable/Fios: 10 Mbit/s [0.94 MB/s] = 61 hrs
 - CSHL: 100 Mbit/s [9.4 MB/s] = 6 hrs
 - USB 2.0: 480 Mbit/s [45 MB/s] = 76 min
 - LAN: 1 Gbit/s [96 MB/s] = 36 min
- Crossbow:
 - 106 GB, 1007 fq.gz files
 - Many small files, easier to parallelize/recover/verify
 - EBI – Amazon (US East) in 75min [~24MB/s]
 - Parallel transfer:
 - 40 simultaneous FTPs
 - Very late at night
 - Saturated- no benefit using more

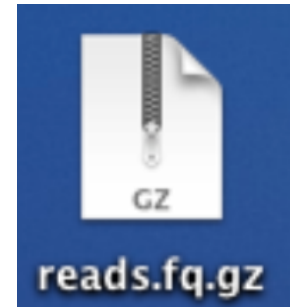


~42 DVDs / run



Data Solutions

- Reduce volume to transfer
 - Compressed fastq files or binary encoding
 - Short read ids
 - Semantic reductions:
 - Don't send/store uninformative sequences?
 - Store compressed de Bruijn Graph?
- Make best use of existing connections
 - Parallel FTP, TCP Window Size, Aspera
 - Wanted: S3/Hadoop transfer with recovery & MD5
- Establish new connections
 - Upgrade to Gigabit+ Internet
 - > \$100k / year
 - Sneakernet solutions
 - Reduced fee for AWS Import/Export
 - Ship DNA to sequencing centers colocated with cloud
 - Complete Genomics, BGI



5:1
compression

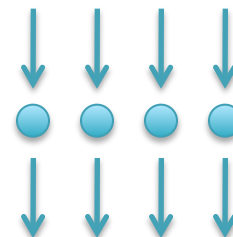


NYC – LA
@ 37 Tbit/s

Cloud Programming Models

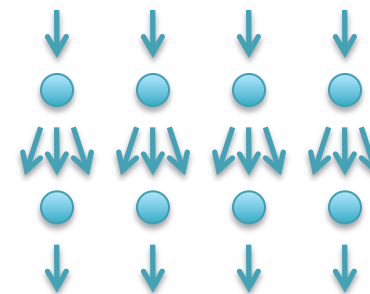
Batch computing: Condor, Grid Engine, Amazon SQS

- Programming Model: Relatively easy, but restricted
- Challenges: Scheduling, Load Balancing, Fault Tolerance
- Resources: Sufficient local memory & cores, fast file system



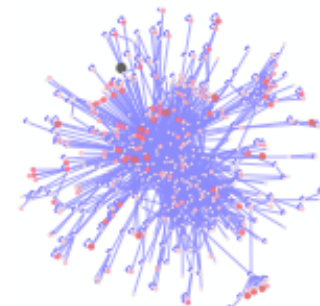
Loosely coupled: Hadoop, Dryad, Amazon EMR

- Programming Model: More complicated, more expressive
- Challenges: Parallel Communication
- Resources: 4+ cores, 1 TB / core disk, 4 GB / core RAM
 - Cloudera Recommendations: <http://bit.ly/bj2lec>



Tightly coupled: MPI, Pregel, Hadoop

- Programming Model: Most complicated, most expressive
- Challenges: Parallel Algorithms
- Resources: High Bandwidth, low latency interconnects
 - Amazon Cloud Compute Instance Type



Cloud Programming Challenges



- Distributed computing challenges
 - Foreign/restricted programming model
 - Distributed debugging
 - Input/output formats, compressed formats
 - Distributing files to hundreds of nodes
- Parallel Algorithms Challenges
 - Amdahl's Law & Load Balance
 - Speedup is never as good as you want
 - Graph searching is notoriously difficult & slow
 - Emerging models (Pregel) look promising
- Shared environment dumping the queue
 - Schedule job on "broken" machine
 - (<1GB disk free, no free RAM, etc)
 - Job crashes right away -> No jobs running -> repeat
 - Virtualization protects users from each other

Institutionalization

- Usability & Support
 - Documentation and Predictability
 - User Handholding
- Policy
 - Privacy, Security
 - IRB/HIPPA for patient/sensitive data
- Accountability
 - Who is on the line for making it all work?
 - Maintaining continuity is hard