Cloud Computing Introduction

David Levine

July 20, 2018

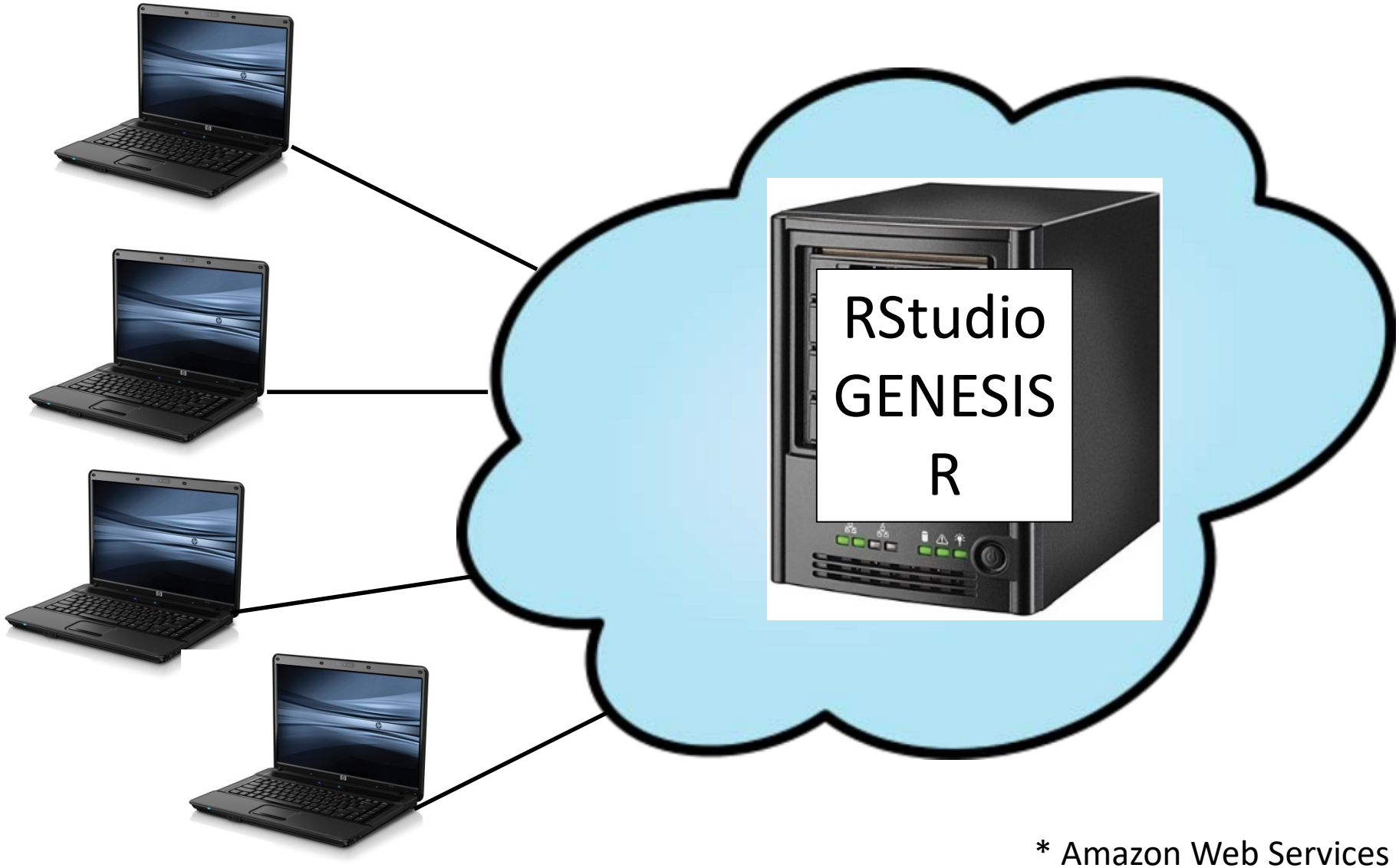
Not powerful enough for WGS data

Data Set	Samples	Variants
workshop	1,126	25,760
freeze.1c	2,643	112,275,224
freeze.2a	9,109	140,980,783
freeze.3a	16,558	185,970,832
freeze.4	18,526	219,154,455
freeze.5	64,960	581,967,553
freeze.6		817,626,115

- Memory
- CPU
- Disk space

Single server OK for smaller data sets

Workshop: Server access via AWS*



* Amazon Web Services

Large WGS data sets belong on a cluster

What is a Cluster?

- Hardware
  - Many computers (instances) each with
    - Multiple processors (cores)
    - Own shared memory
  - Shared file system
  - Network connectivity
- Software
  - Linux OS
  - Queuing system (SGE)
  - Jobs execute independently
- Pros: Many cores and lots of memory
- Cons: Responsible for managing parallelism

* Standard distributed-memory

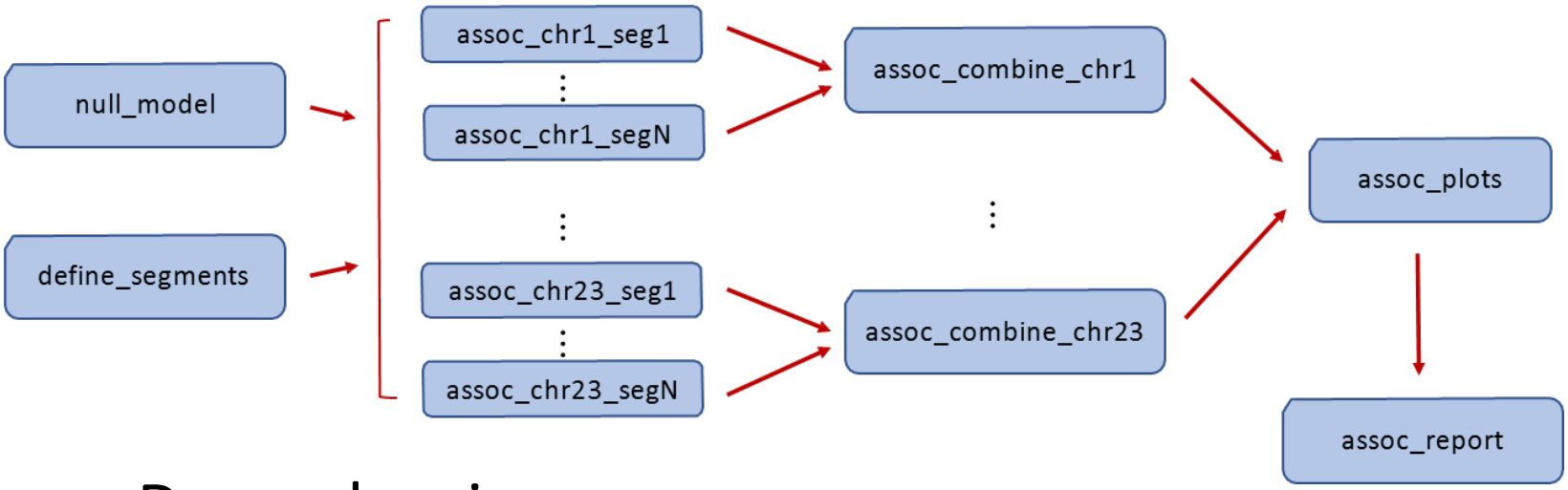
Where to get a cluster?

- Owning is expensive, so rent (Cloud)
- Pros
  - No/low infrastructure costs
  - Pay per use model
  - Scalable with increasing data set sizes
  - Variety of computers (RAM, CPU, disk, GPU)
  - Minimal management
  - Automatic software updates
  - Reliability and disaster recovery

Where to get a cluster?

- Owning is expensive, so rent (Cloud)
- Cons
  - Ongoing monthly costs
  - Pay for debug runs, failed runs, instances left running
  - You are your own IT person (or still need one)
  - Manage much of your own security
  - Extra effort to minimize costs
  - Cloud vendor lock-in

Managing Pipeline Parallelism



- Dependencies
- Synchronization
- Heterogeneity
- Autoscaling
- Retry

Cloud environments add cost complexity

Managing Pipeline Parallelism

- Explicit management (command line tools)
  - Python, JSON
  - AWS Batch
- Embedded in a genomics application (GUI)
  - Seven Bridges, DNAnexus
  - Mitigate complexity
  - Centralize data access

WGS major computational need

- Run one time
  - VCF to GDS file conversion
- Run a few times
  - Relatedness analysis
- **Run many times**
  - **Association testing**

What influences cloud costs

- No. samples
- No. variants & filtering
- No. variants per aggregation unit
- Algorithm: Single variant, Aggregate
- Implementation: sparse matrices, fastSKAT
- Cloud hardware used (cores, RAM, disk)

AWS cloud benchmarks

Analysis	Number Samples	Number Variants	Time (hh.mm)	Parallel Segments	Standard Cost	Spot Cost	Note
GRM	25,077	415,235,243	40:12	22		$58.12	SNV
Single Variant	25,077	46,534,015	0:44	27	$34.60	$10.80	MAC > 10
SKAT	25,077	268,368,508	6:51	27	$311.50	$97.20	MAF <= 1%

- N subjects, M variants
- Single variant tests
  - RAM $O(N^2)$
  - CPU $O(MN)$
- SKAT tests
  - RAM $O(N^2)$
  - CPU $O(M^2N)$