Clouds in Bioinformatics

Rob Knight

HHMI and University of Colorado at Boulder

How we use infrastructure clouds

- New instruments collect vast datasets: investigators care about time-to-result (especially for diagnostics/pathogen discovery), demand is highly elastic
- Workflow services (Galaxy, CIPRES, QIIME) deployed to cloud allow rapid processing of sequence data for tasks including genome assembly, microbial community analysis
- Of increasing interest: cloud-enabled GUIs that allow biologists to run analyses themselves using vast compute resources

Appeal of infrastructure clouds?

- Allows large resources (e.g. EC2) to be used on-demand at trivial cost compared to buying resource that handles peak load (e.g. 1000 cores to process 1.6 TB from a HiSeq run overnight; more for interactive use)
- Greatly reduced hardware support and systems administration burden compared to purchasing own machine
- 3rd parties using our software can easily pay for their own compute cycles, rather than drawing from our resources/allocations
- Machine image easily used by end-users with no prior cloud experience (installation on arbitrary clusters a nightmare)

Cloud Challenges

- Difficult to get large datasets into the cloud: must currently mail drives, need larger pipes!
- Unreliability. Have had to rewrite code to deploy on EC2, mostly to deal with random node and network failures, checkpointing apparently more critical than on traditional cluster.
- Existing clouds suited to only a subset of bio tasks: higher-memory configurations with local storage especially needed in order to minimize problems with I/O-bound tasks

Compare to other outsourcing

- Academic collaborations: clouds e.g. EC2 and Magellan vastly easier to use than native install on arbitrary and surprisingly configured cluster. Resource availability much better especially near grant deadlines. Disadvantage is cost (commercial) or reliability (current academic systems)
- Advantages relative to TG: ease of use, actually works for our applications (e.g. taskfarming and other loosely coupled tasks, especially long-running tasks, not priority for TG), administrators responsive, homogeneous across resources in a given cloud. No disadvantage from my perspective.
- Advantages relative to commercial: transparency of how result was obtained, ability to reproduce computation. No disadvantage from my perspective.