# Cloud Sobriety

*Technical challenges in mapping*

*Informatics to the cloud*

Chris Dagdigian

2010 NHGRI Cloud Workshop

BIOTEAM
Enabling Science

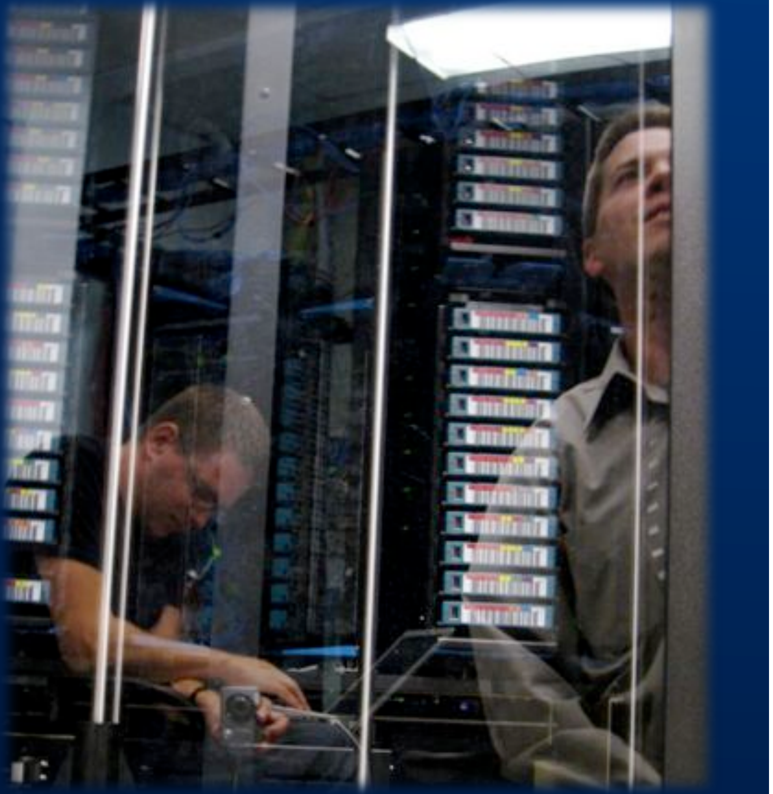
# Welcome To Day 2

- Excellent talk lineup for today
- Focus on implementation, architecture & deployment challenges
- I'll be giving a brief overview before turning the floor over to the real experts

BIOTEAM
Enabling Science

# Who I am

- Part of the BioTeam
  - Bioinformatics → HPC & Research IT nerd
- Our business:
  - Bridging the gap between science & high performance IT



BIOTEAM
Enabling Science

# Why I'm here

- Doing "science" on the cloud since 2007
- Heavy IaaS user
  - Amazon AWS
- Can speak from multiple viewpoints:
  - Cloud User/ consumer
  - Vendor/integrator

BIOTEAM
Enabling Science

# Understand My Bias

- I'm an infrastructure geek
  - My building blocks are compute, storage & network services, not software or service platforms
    - I care about "Utility Computing" or "IaaS"
- I don't particularly care about
  - Platform-as-a-Service ("PaaS")
  - Software-as-a-Service ("SaaS")
  - IaaS providers with < 200,000 cores under active management
- Amazon Web Services is the only provider who can meet all of my needs today
  - I am quite mercenary in technology choices though …
  - If a better solution comes along I'll switch in an instant

BIOTEAM
Enabling Science

# Cloud Informatics Challenges

**Architectural**

*Science != facebook*

**Technical**

*Adventures in data movement & virtualization*

**Political**

*Of kingdoms & sysadmins*

BIOTEAM
Enabling Science

# Architectural Challenges

*Infrastructure clouds were not built for people like us*

# Architectural Challenges

- Cloud designed for large internet-scale services
- Delivered via:
  - Loosely coupled, asynchronous services
  - Significant replication & load balancing tricks
  - Eventual consistency model

- Not ideal for our needs:
  - We are used to tightly coupled & fast systems
  - We happily trade reliability & availability for additional performance & throughput
  - Scientists see eventual consistency as evil

BIOTEAM
Enabling Science

# Architectural Challenges

- Virtual everything is slow
  - Performance is sacrificed to provide the foundational services required by the extreme internet-scale Web 2.0 crowd
  - Particularly problematic in life science informatics where we are often performance bound by the speed of our storage systems

BIOTEAM
Enabling Science

# Architectural Challenges

- Radical effect on HPC & Grid Computing:
  - Many of us use large HPC clusters & compute grids within our organization
    - Large systems *shared* by multiple users, groups, workflows & projects; Platform LSF or Sun Grid Engine software to enable the shared infrastructure resource
  - Clouds allow *dedicated resources for every user, problem, workflow & project*
    - Turns traditional methods & practices upside down

# Technical Challenges

*Data movement & HPC hassles in the cloud …*

BIOTEAM
Enabling Science

# Technical Challenges

- Mentioned in talks both today & yesterday
- No time to get really deep & technical
- Brief comments on
  - Data Movement
  - Networks
  - Storage
  - Documentation & How-To pitfalls

BIOTEAM
Enabling Science

# Technical Challenges

- Data Movement
  - #1 issue/concern
  - Internet vs. FedEx?
  - One-way or bidirectional?
  - Not just the size of your pipe …
    - Physical location matters as well

- Networks
  - No control over topology
  - Some nasty surprises for HPC people & software
  - Software VPNs for unifying network space work
    - … but it's an insane hassle to set up and manage
    - Amazon VPC not quite there yet

# Technical Challenges

- Storage
  - It's slow. Absolute fact.
  - Various methods to mitigate or work around
  - Among top 3 implementation challenge in most workflows we've seen

- "Bad" Documentation
  - Just like the "beowulf cluster" days
  - Most material written for an entirely different audience
    - Following some 'best practice' advice can actually hinder scientific workflows

BIOTEAM
Enabling Science

# Political Challenges

*Now it gets really complicated …*

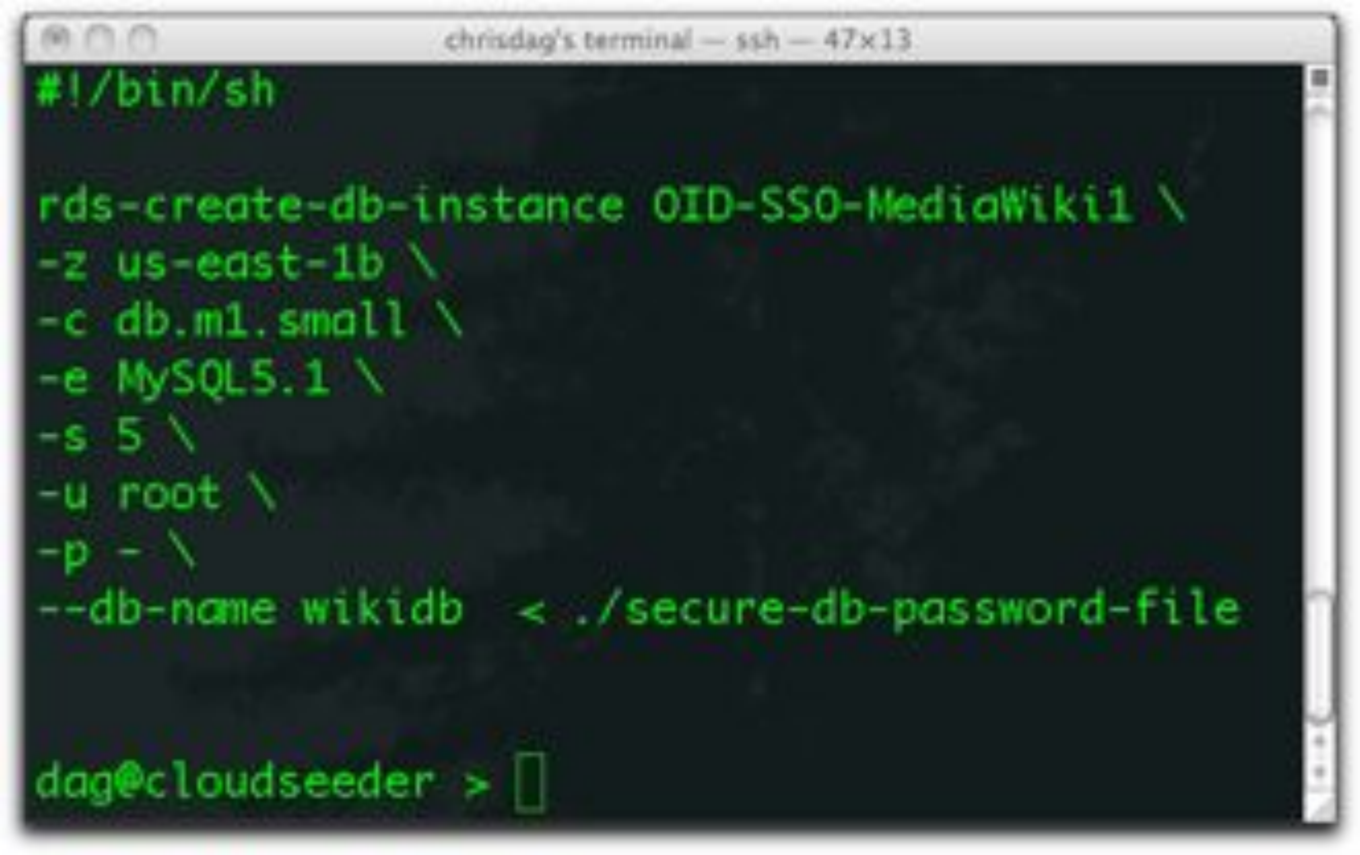
# Political Issues

- Clouds raise significant internal issues
  - CapEx vs. OpEx issues
  - Who pays? How do we pay? Who monitors?
  - When do you port legacy apps to “the new cloud way” ?
  - What does the support model look like?
  - What does the development model look like?
- Often encounter these issues:
  - IT staff protecting internal empires
  - Incredibly difficult to accurately track *true fully loaded internal costs* of local infrastructure
    - And if you can’t do this, how can you claim the cloud will save money?

BIOTEAM
Enabling Science

# The elephant in the room …

# “Scriptable Infrastructure”



```
#!/bin/sh

rds-create-db-instance OID-SSO-MediaWiki1 \
-z us-east-1b \
-c db.m1.small \
-e MySQL5.1 \
-s 5 \
-u root \
-p - \
--db-name wikidb  < ./secure-db-password-file


dag@cloudseeder > 
```

This single command will start a 5GB managed MySQL database in the Amazon cloud for $0.11/hour. The database is ***automatically*** patched, managed and backed up. Planned enhancements include auto-scaling & snapshots.  THIS IS A BIG DEAL.

BIOTEAM
Enabling Science

# Scriptable Infrastructure

- *What happens to IT roles when anyone with a web browser can instantly launch (and manage) a complex cluster, software pipeline or massive database?*

- Radical restructuring of the lines between
  - Research staff & Investigators
  - IT Operations Staff
  - IT Support Staff

# Scriptable Infrastructure

- For the first time some of our IT infrastructure might be 100% virtual and entirely controllable via scripts and APIs
- Anyone can drive this stuff, especially motivated researchers
- My prediction:
  - The role of “Systems Administrator” is going to change
  - More focus on toolsmithing, scripting, troubleshooting
  - Significant focus on enabling end users to be effective and self-supporting (as much as possible)
  - Interesting times ahead …

# Quick Security Thoughts …

# Quick Security Thoughts …

1. Microsoft, Google & Amazon have better operating, audit and network security controls than you do.
2. I am suspicious of people demanding cloud security practices that they themselves have failed to deploy on their own infrastructure
3. Cloud providers will happily answer your deepest technical security questions

# End;

- Thanks!
- Time for the more detailed talks

- Presentation slides will appear here:
  - http://blog.bioteam.net

- Comments/feedback:
  - chris@bioteam.net

BIOTEAM
Enabling Science