Utility Computing For Cynics

2009 Amazon NYC Start-Up Tour

Chris Dagdigian

BioTeam Inc.

BIOTEAM
Enabling Science

Topics for Today

Who we are:
Why we agreed to speak

Why we use AWS:
How we came to drink the Kool-aid

Pharma Example:
Protein Docking on AWS

BIOTEAM
Enabling Science

BioTeam Inc.

- **Independent Consulting Shop:**
  Vendor/technology agnostic

  Distributed entity - no physical office
- **Staffed by:**
  - Scientists forced to learn
    High Performance IT to conduct research
  - Many years of industry & academic experience
- **Our specialty:**
  Bridging the gap between Science & IT

BIOTEAM
Enabling Science

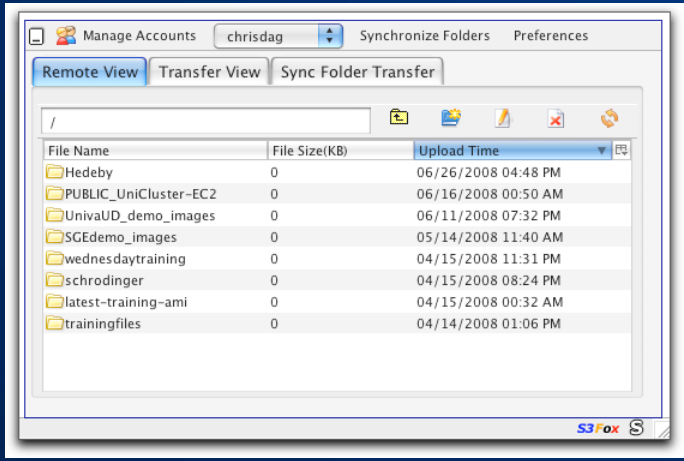
Setting The Stage

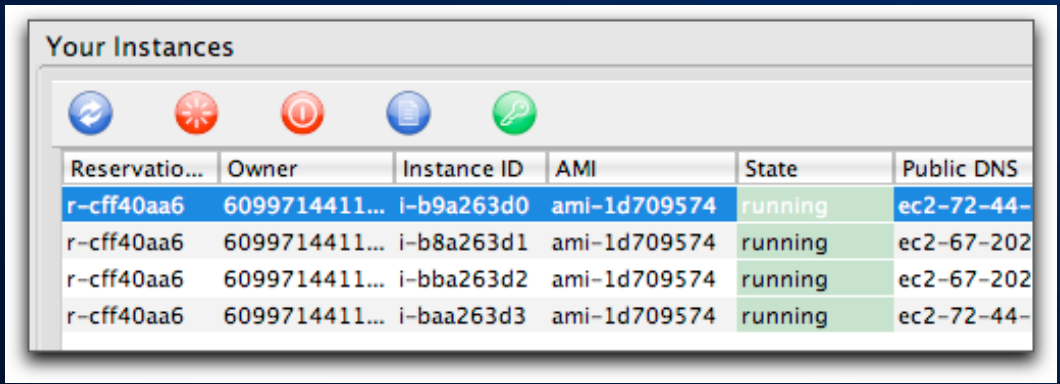
- **Burned by "OMG!! GRID Computing" Hype**
  In 2009 will try hard never to use the word "*cloud*"
  in any serious technical conversation. Vocabulary matters.
- **Understand My Bias:**
  - Speaking of "utility computing" as it resonates with *infrastructure* people
  - My building blocks are servers or groups of systems, not software stacks, developer APIs or commercial products
  - Goal: Replicate, duplicate, improve or relocate complex systems

BIOTEAM
Enabling Science

Lets Be Honest

- **Not rocket science**
- **Fast becoming accepted and mainstream**
- **Easy to understand the pros & cons**





BIOTEAM
Enabling Science

Tipping Point: Hype to Reality

- 2007: Individual staff experimentation all year
  - Including MPI applications (mpiblast)
- Q1 2008:
  - Realized that every single BioTeam consultant had *independently* used AWS to solve a customer facing problem
  - No mandate or central planning, it just happened organically

BIOTEAM
Enabling Science

BioTeam AWS Use Today

- **Running Our Business**
  Development, Prototyping & CDN
  - Effective resource for tech-centric firms
- **Grid Training Practice**
  - Self-organizing Grid Engine clusters in EC2
  - Students get root on their own cluster
- **Proof Of Concept Projects**
  - UnivaUD - UniCluster on EC2
  - Sun - SDM ‘spare pool’ servers from EC2
- **Directed Efforts on AWS**
  ISV software ports, Pharma clients, etc.

BIOTEAM
Enabling Science

HPC & AWS: Whole new world

- **For cluster people some radical changes**
  Years spent tuning systems for shared access
  - Utility model offers *dedicated* resources
  - EC2 not architected for *our* needs
  - Best practices & reference architectures will change
- **Current State: Transition Period**
  - Still hard to achieve seamless integration with local clusters & remote utility clouds
  - Most people are moving entire workflows into the cloud rather than linking grids
  - Some work being done on ‘transfer queues’

BIOTEAM
Enabling Science

Real World Example

Protein Engineering on Amazon AWS

BIOTEAM
Enabling Science

Protein Engineering with AWS

- **Pfizer Biotherapeutics & Bioinnovation Center**
  - **Giles Day, Pfizer**
  - **Adam Kraut, BioTeam**

- **Problem:**
  - Antibody models can be created in a few hours on a standard workstation
  - Full-atom refinement of each model using Rosetta++ requires 1000 CPU hours
  - 2-3 months required *per-model* on existing Pfizer research cluster
  - Cluster subject to unpredictable loads

BIOTEAM
Enabling Science

Protein Engineering with AWS

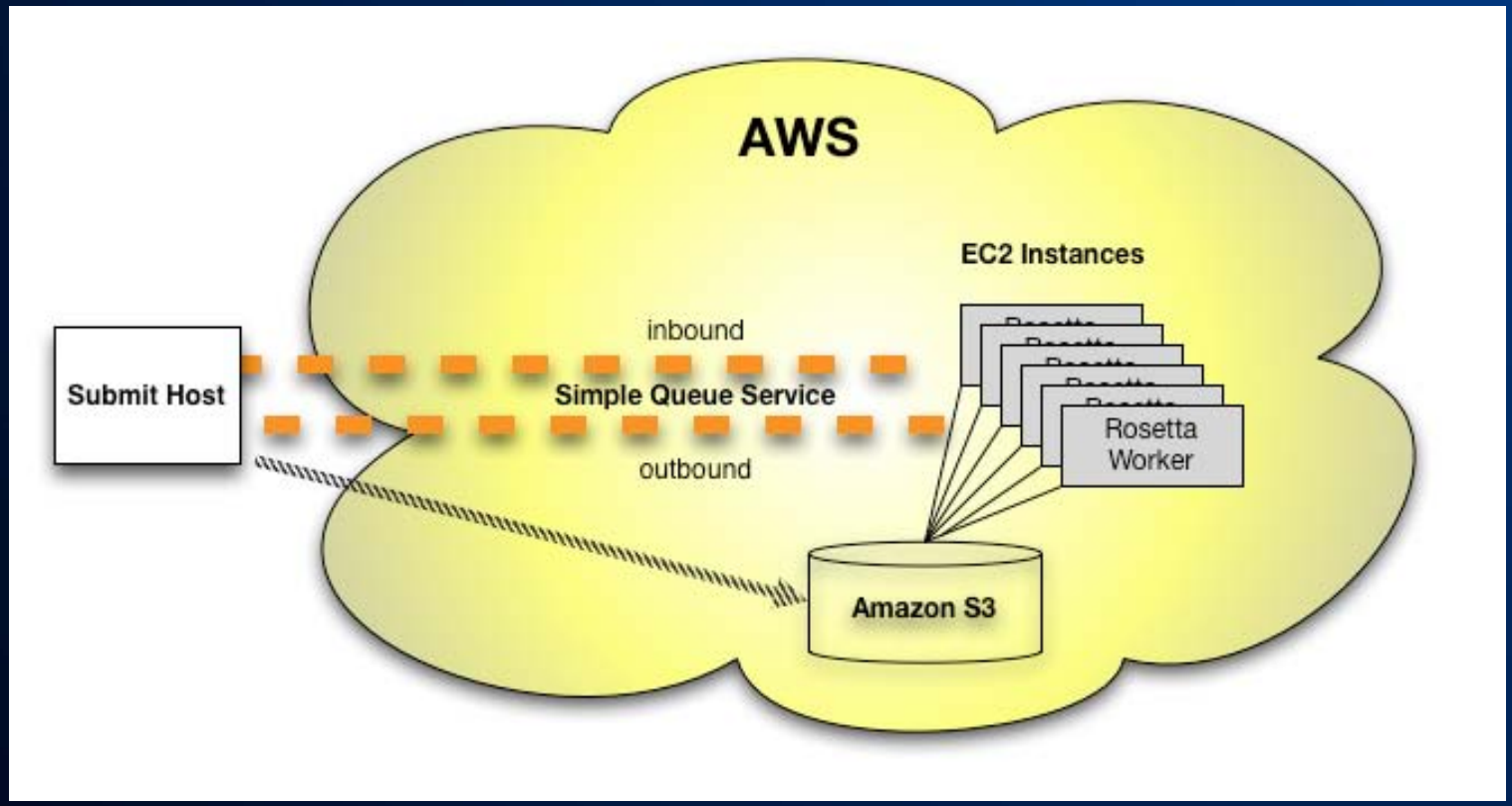
- **1000 CPU Hour Antibody Refinement Problem**
  - **Using Rosetta++ (Davd Baker, Uwash)**
- **Huge Opportunity for Pfizer:**
  - Deliver antibody model refinement results in *one day* rather than 2-3 months
- **Ideal AWS Candidate:**
  - CPU bound
  - Low data I/O requirements
  - Free up cluster for I/O bound workloads

BIOTEAM
Enabling Science

Protein Engineering with AWS

- Borrows heavily from RightScale & AWS published best practices
- Inbound/Outbound SQS queues
- Job specification in JSON format
- Data for each work unit in S3 buckets
- Custom EC2 AMI
- Workers pull from S3, push back when finished
- Job provenance/metadata stored in SimpleDB
- Independent work units allow dynamic allocation of Worker instances

BIOTEAM
Enabling Science



BIOTEAM
Enabling Science

Looking Ahead

My $.02

BIOTEAM
Enabling Science

Utility Computing Pet Peeve

Security

Don’t want to belittle security concerns:

- But whiff of hypocrisy is in the air …
  - Is your staff *really* concerned or just protecting turf?
  - It is funny to see people demanding security measures that they don’t practice internally across their own infrastructure

BIOTEAM
Enabling Science

Utility Computing Pet Peeve

Security

My take:

- Amazon, Google & Microsoft probably have better internal operating controls than you do
- All of them are happy to talk as deeply as you like about all issues relating to security
- Do your own due diligence & don’t let politics or IT empire issues cloud decision making

BIOTEAM
Enabling Science

State of Amazon AWS

New features are being rolled out fast and furious

But …

- EC2 nodes still poor on disk IO operations
- EBS service can use some enhancements
  - Many readers, one-writer on EBS volumes would be fantastic
- Poor support for latency-sensitive things and workflows that prefer tight network topologies

This matters because:

- Compute power is easy to acquire
- Life science tends to be IO bound
- Life science is currently being buried in data

BIOTEAM
Enabling Science

Bulk Data Ingest/Export

- How do I load 1TB/day into AWS?
  - AWS ImportExport Service still “beta”,
  - No Export yet
  - Very excited about this though
- My field is looking for answers
  - Need “cheap and deep” store(s)
  - Currently buried by lab instruments that produce TB/day volumes
    - Next-Gen DNA Sequencing
    - 3D Ultrasound & other imaging
    - Confocal microscopy
    - Etc.

BIOTEAM
Enabling Science

Bulk Data Ingest/Export

Big Scary Concern:

- BioTeam asked yesterday to help move 30TB from Amazon S3 onto physical media for client delivery
- We put a server 1 hop away from a major Boston internet backbone, started transfers …
- Best S3 download speed so far:
  30 MB/sec sustained, bursting to 50 MB/sec

  Possibly not good enough for terabyte scale data…

- My Fear:
  - Amazon may be the only practical supplier of bulk import/export services
  - Unless you are big enough to peer directly
  - *Ever priced out GbE internet connectivity? Whoa …*

BIOTEAM
Enabling Science

If ingest problem can be solved …

- I think there may be petabytes of life science data that would flock to utility storage services
  - Public and private data stores
  - Mass amount of grant funded study data
  - Archive store, HSM target and DR store
  - “Downloader Pays” model is compelling for people required to share large data sets

BIOTEAM
Enabling Science

Terabyte Wet Lab Instrument

BIOTEAM
Enabling Science

Cautionary Tale: 180TB kept on desk

The life science “data tsunami” is no joke

BIOTEAM
Enabling Science

Next-Gen & Potential AWS use

What this would mean:

- Primary analysis onsite; data moved into remote utility storage service after passing QC tests
- Data would rarely (if ever) move back
- Need to reprocess or rerun?
  - Spin up “cloud” servers and re-analyze in situ
  - Terabyte data transit not required

Summary:

- Lifesci data; 1-way transit into the cloud
- Archive store or public/private repository
- Any re-study or reanalysis primarily done in situ
- Downside: replicating pipelines & workflows remotely
- Careful attention must be paid to costs

BIOTEAM
Enabling Science

Cloud Sobriety

McKinsey presentation "Clearing the Air on Cloud Computing" is a must-read

- Tries to deflate the hype a bit
- James Hamilton has a nice reaction:
  - http://perspectives.mvdirona.com/

Both conclude:

- IT staff needs to understand "the cloud"
- Critical to quantify your own internal costs
- Do your own due diligence

BIOTEAM
Enabling Science

End;

- Thanks!
- Comments/feedback:
  - chris@bioteam.net

BIOTEAM
Enabling Science