

# Commercial Clouds for Bioinformatics

Beyond The Genome

Oct 11-13 2010, Harvard Medical School



[chris@bioteam.net](mailto:chris@bioteam.net) - <http://www.bioteam.net>

# Hi Kaitlyn!

- Fulfilling my “Uncle’s Duty” to embarrass a niece who is home from school today
- World Arthritis Day – Oct 12
- <http://www.worldarthritisday.org/>



# Who am I?

- I'm from the BioTeam
  - Independent consulting shop
  - Staffed by scientists forced to learn IT to get our own research done
- Found a fun business niche
  - Bridging the “gap” between science, IT & high performance computing
- This matters today because ...
  - We've been doing production informatics work on Amazon AWS since 2007
  - Can speak from multiple AWS perspectives (Customer, Developer, Integrator)



*Scene from ancient history in a cloud-enabled world...*

# Why I drank the Kool-Aide

- Laziness
- Beauty
- Agility
- Money



*AWS is not the real reason I often visit Seattle, shhhhhh!*

# Laziness

- Larry Wall's 1<sup>st</sup> Great Virtue:
  - *"... the quality that makes you go to great effort **to reduce overall energy expenditure. It makes you write labor-saving programs that other people will find useful...**"*
- Scriptable IT Infrastructures are the latest boon for the perennially lazy (like myself)



*Note subtle Amazon product plug above ...*



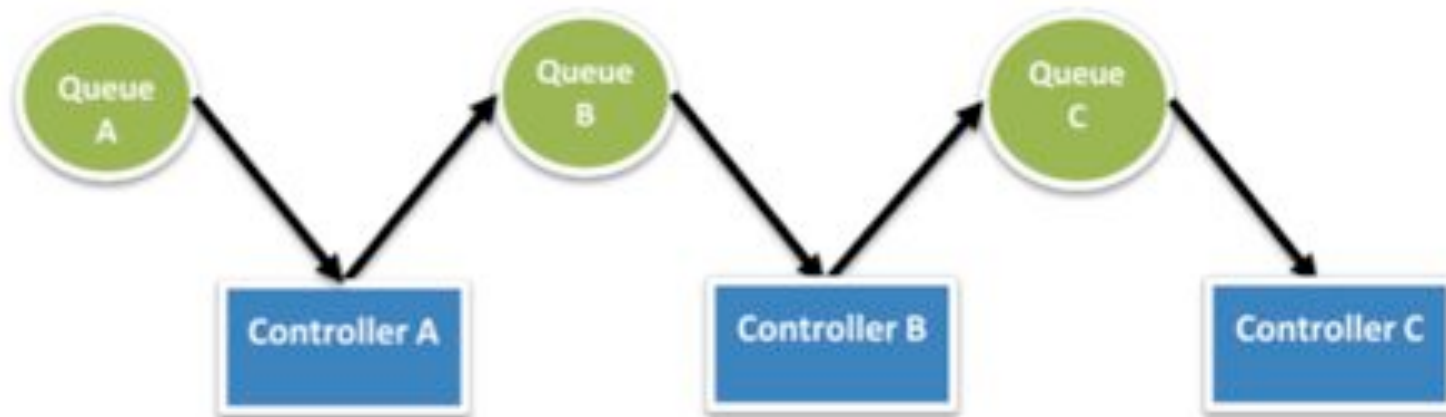
# Beauty

- Call me a nerd but this stuff is amazing:
  1. “Scriptable Datacenter(s)”
  2. Orchestrating complex systems & workflows with a few lines of code
  3. Infrastructure managed like source code



*Can you believe that the Broad Institute @ MIT let me into their telco closets & machine rooms?*

# Agility



# Money

- If a cloud talk occurs without mention of \$
  - ... did it actually happen?

Greetings from Amazon Web Services,

This e-mail confirms that your latest billing statement is available on the AWS web site. Your account will be charged the following:

Total: \$101.68

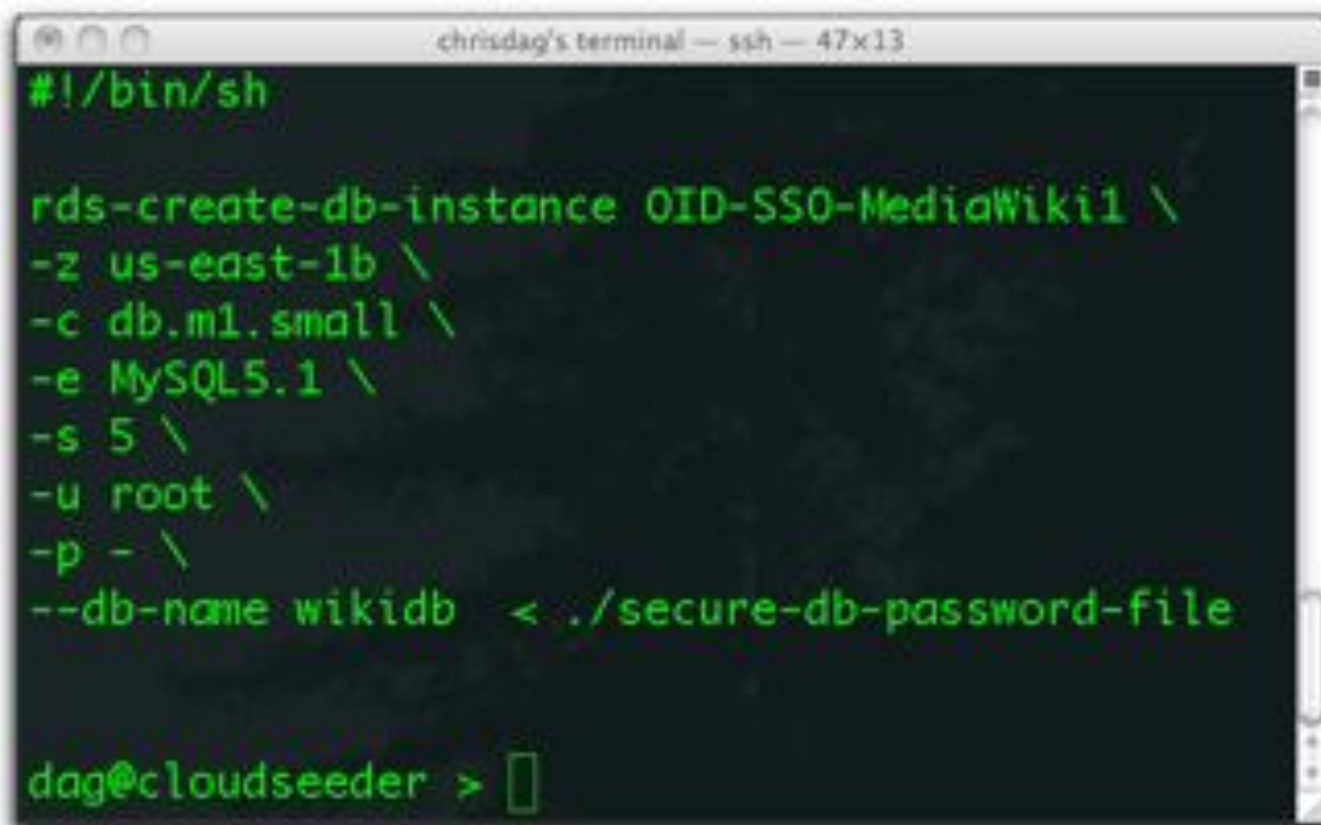
Please see the Account Activity area of the AWS web site for detailed account information:

<http://aws-portal.amazon.com/gp/aws/developer/account/index.html?action=activity-summary>



# Laziness

## “Scriptable Infrastructure” is a BIG DEAL



```

chrisdag's terminal — ssh — 47x13
#!/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.

# It's ALL scriptable ...

- Servers
- Storage
- Operating System(s)
- Network
- Provisioning
- Management
- Monitoring & Scaling
- Accounting

THIS is why we are using commercial clouds for bioinformatics! The pervasive automation available on the cloud has yet to spread far into our own datacenters.

# Not hype. Real.

- Every facet of our IT infrastructure can now be automated and remotely controlled via simple scripts and API calls
- Benefits go way above and beyond simple IT Operations work, server “lights out management” features and what local VM systems provide
- *Nirvana for lazy nerds like myself*
  - *Concentrate on getting my work done, not babysitting my datacenter racks*

```

top - 10:10:40 up 22:21, 1 user, load average: 0.10, 0.16, 0.18
Tasks: 102 total, 1 running, 101 sleeping, 0 stopped, 0 zombie
Cpu(s): 0.0% us, 4.0% sy, 0.0% ni, 96.0% id, 0.0% wa, 0.0% hi, 0.0% si
Mem: 2523776k total, 2507368k used, 16408k free, 41660k buffers
Swap: 589816k total, 208k used, 589608k free, 1984196k cached

  PID USER      PR  NI  VIRT  RES  SHR  S  %CPU  %MEM    TIME+  COMMAND
 20252 dag        16   0 6172 1140 836  R   4.0   0.0   0:00.08 top
    1 root        16   0 4772  640 536  S   0.0   0.0   0:02.75 init
    2 root        34  19   0     0   0  S   0.0   0.0   0:00.04 ksoftirqd/0
    3 root         5 -10   0     0   0  S   0.0   0.0   0:01.11 events/0
    4 root         5 -10   0     0   0  S   0.0   0.0   0:00.05 khelper
    5 root         7 -10   0     0   0  S   0.0   0.0   0:00.00 kthread
    6 root        13 -10   0     0   0  S   0.0   0.0   0:00.00 kacpid
   86 root         5 -10   0     0   0  S   0.0   0.0   0:32.12 kblockd/0
   87 root        15   0   0     0   0  S   0.0   0.0   0:00.00 khubd
  110 root        20   0   0     0   0  S   0.0   0.0   0:00.00 pdflush
  111 root        15   0   0     0   0  S   0.0   0.0   0:13.78 pdflush
  112 root        16   0   0     0   0  S   0.0   0.0   0:28.31 kswapd0
  113 root         7 -10   0     0   0  S   0.0   0.0   0:00.00 aic/0
  259 root        24   0   0     0   0  S   0.0   0.0   0:00.00 kseriod
  492 root        19   0   0     0   0  S   0.0   0.0   0:00.00 scsi_eh_0
  514 root        15   0   0     0   0  S   0.0   0.0   1:20.84 kjournald
 1228 root         6 -10   0     0   0  S   0.0   0.0   0:00.00 kauditd
  
```

So easy an Ipad can control it.

# Scriptable Infrastructure

- For the first time some of our IT infrastructure might be 100% virtual and entirely controllable via scripts and APIs
- It's not rocket science
- Anyone can drive this stuff
  - Especially motivated researchers
  - **... and this is what is driving informatics onto cloud platforms**



# Beauty

## Scripted Infrastructure – Creating AMI Bundles

```
ec2-bundle-vol -d /mnt \  
-k /root/.ec2/pk-6LGHW [REDACTED] .WHU7.pem \  
-c /root/.ec2/cert-6LGHW [REDACTED] .HSHWHU7.pem \  
-u 6099-7144-1117 \  
-r i386 -p 32bit_ManualChef
```

# BO

## Scripted Infrastructure – Upload bundle to S3 bucket

```
ec2-upload-bundle -b cloudtraining/32bmanchef \  
-m /mnt/32bit_ManualChef.manifest.xml \  
-a AKIAJFXS50PTB52QFJFA \  
-s BLr2w[REDACTED]0xp+GaA5/5dv
```

# BORI

## Scripted Infrastructure – Register & receive new AMI ID

```
ec2-register -d "32bit CentOS, BioTeam Chef Managed Server (Manual Registration)" \  
-n "32b_chef_manreg" \  
-K /root/.ec2/pk-6L[REDACTED]HU7.pem \  
-C /root/.ec2/cert-6L[REDACTED]SWHU7.pem \  
-a "i386" \  
cloudtraining/32bmanchef/32bit_ManualChef.manifest.xml
```

# BORING.

# Beauty

- “Scriptable Infrastructure” is just the baseline
  - The cool stuff happens when we build on top of these capabilities
- AWS enables us to **orchestrate** vast arrays of complex systems, pipelines, workflows & applications
  - *Without leaving the hammock*
- *Orchestrated systems working in concert are a beautiful thing.*





# Money

# Money

For anyone seriously looking at IaaS Cloud Platforms

- Can't escape it
- Critical to have a solid understanding of the financial issues
- Viewed from many different angles:
  - **Save** money & increase efficiency
  - **Convert** CapEx to OpEx
  - **Enhance** existing capability (elasticity, agility)
  - **Enable** new capabilities

# Why Commercial Clouds

# Why Commercial?

- I don't care where you work
- Nobody in this room can match the internet-scale operators with respect to:
  - Scale
  - Engineering Resources
  - Rate of change

# Scale

- Can you build datacenters all over the world with PUE values getting closer and closer to 1.0?
- Are your datacenters chiller-less because you can shed and route load all over the globe?
- Can you have one employee per {XX,XXX} servers?
- Do you have exabytes of spinning disk? Operating 500K servers? 1M cores?
- How much leverage do you have with your server vendor over hardware component efficiency?



Yahoo's new "Chicken Coop" datacenter design



# Scale

- Primary benefit of IaaS platforms is economic
  - At massive scale, commercial providers can sell robustly engineered services to us cheaper than we can afford to do it ourselves\*\*
    - While still earning healthy profit margins...
- This is why trends favor the commercial providers

*\*\* If we are honest about the true cost of delivering IT services to our customers*

# Engineering Resources

- Google, MS, Amazon, etc. all have long experience running robust & resilient services in an incredibly hostile networking environment
- All of them can also afford to hire dedicated teams of smart people that do nothing but chase down 1% efficiency gains wherever they might be found
- Great example & resource, James Hamilton's blog:
  - <http://perspectives.mvdirona.com/>

# Rate of Change

- Using Amazon Web Services as example ...
- The rate at which new services are rolled out and existing services are improved is insane
- Extremely difficult to match
- Basically means everyone else plays catch-up
  - *Can your “private cloud” match this?*

# 1: AWS Rate of Change Example

- Dec 2009
  - **Amazon VPC launch**
  - **AWS Spot Instance launch**
  - Windows Server 2008, SQL Server 2008 support
  - **AWS Import/Export launch**
  - US-West AWS region launch
- Feb 2010
  - SimpleDB consistency enhancements
  - Reserved Instances (Windows)
  - **m2.xlarge EC2 instance type**
  - **AWS Consolidated Billing**
  - S3 Object Versioning

*The AWS Blog is a great resource: <http://aws.typepad.com/aws/>*

## 2: AWS Rate of Change Examples

- March 2010
  - **S3 Import/Export**
    - **Raw drive support**
  - **S3 Versioning**
  - Combined bandwidth pricing
  - Reverse DNS for elastic IPs
- April 2010
  - SNS Service beta
  - RDS Europe launch
  - Singapore AWS Region w/ 2 availability zones launched

*The AWS Blog is a great resource: <http://aws.typepad.com/aws/>*



## 3: AWS Rate of Change Examples

- May 2010
  - **RDS Multi-AZ Deployment**
  - **S3 Reduced Redundancy Storage (RRS) launch**
  - **RDS support in AWS Console**
- June 2010
  - Elastic Map Reduce Updates
  - **S3 Import/Export API**
  - CloudFront HTTPS support
  - **S3 support in AWS Console**
  - CloudWatch metrics for EBS volumes
  - SSL support for RDS

*The AWS Blog is a great resource: <http://aws.typepad.com/aws/>*

## 4: AWS Rate of Change Examples

- July 2010
  - **SQS Enhancements**
    - 100K req/month for free; Configurable message size & retention period
  - More RDS integration into AWS Console
  - **S3 per-bucket access policies!**
  - **cc1.4xlarge instance types!**
  - VPC access control & config generators
  - S3 RRS support in AWS Console
  - More S3 – SNS Integration
    - S3 Buckets can now send messages to SNS topics
  - Enhanced CloudFront log data
  - Support for custom Linux kernels on EC2
  - Penetration Testing Policy & Resource
- August 2010
  - RDS moves to Mysql 5.1.49 w/ InnoDB plugin
  - RDS Reserved Instance Launch

*The AWS Blog is a great resource: <http://aws.typepad.com/aws/>*

## 5: AWS Rate of Change Examples

- September 2010
  - **EC2 Price Reduction**
  - VPC support in AWS Console
  - EC2 Micro-instance Launch
  - **S3 Import/Export support for 8TB storage devices**
  - **Amazon Linux AMI Launch**
  - **EC2 “bring your own keypair” support**
  - EC2 idempotent instance creation
  - EC2 Resource Tags
  - EC2 describe-instances filters
- October 2010
  - RDS price reduction & read replicas
  - SNS integrated with AWS management console

*The AWS Blog is a great resource: <http://aws.typepad.com/aws/>*

# Bioinformatics on the cloud

# Heard this before?

*Problems in mapping informatics to the cloud:*

- **Architecture**
  - IaaS platforms not built for our HPC use cases
- **Performance**
  - “Virtual everything” comes at a price
- **Data**
  - Data movement still a fantastic pain
  - Still difficult to make storage “go fast” on cloud
- **Networks, Networking & Message Passing**
  - Still awkward

# Architecture Issues

- Complaint
  - IaaS platforms built for massive internet scale operation. Heavily biased towards resilient, highly distributed & loosely connected services
  - We need fast, tightly coupled systems and will happily trade some reliability to get this ...
- Status in Late 2010
  - Getting better
  - AWS “compute cluster” instances are a huge step forward
  - The market is clearly listening to HPC audience



# Performance Issues

- Complaint
  - Multi-tenant virtual platform causes performance hit; cloud performance is variable & hard to instrument
- Status Late 2010:
  - IaaS platforms slowly getting better
  - WE are getting better much faster!
  - Emerging body of life-science cloud best practices is growing rapidly
  - Still finding the benefits outweigh the negatives

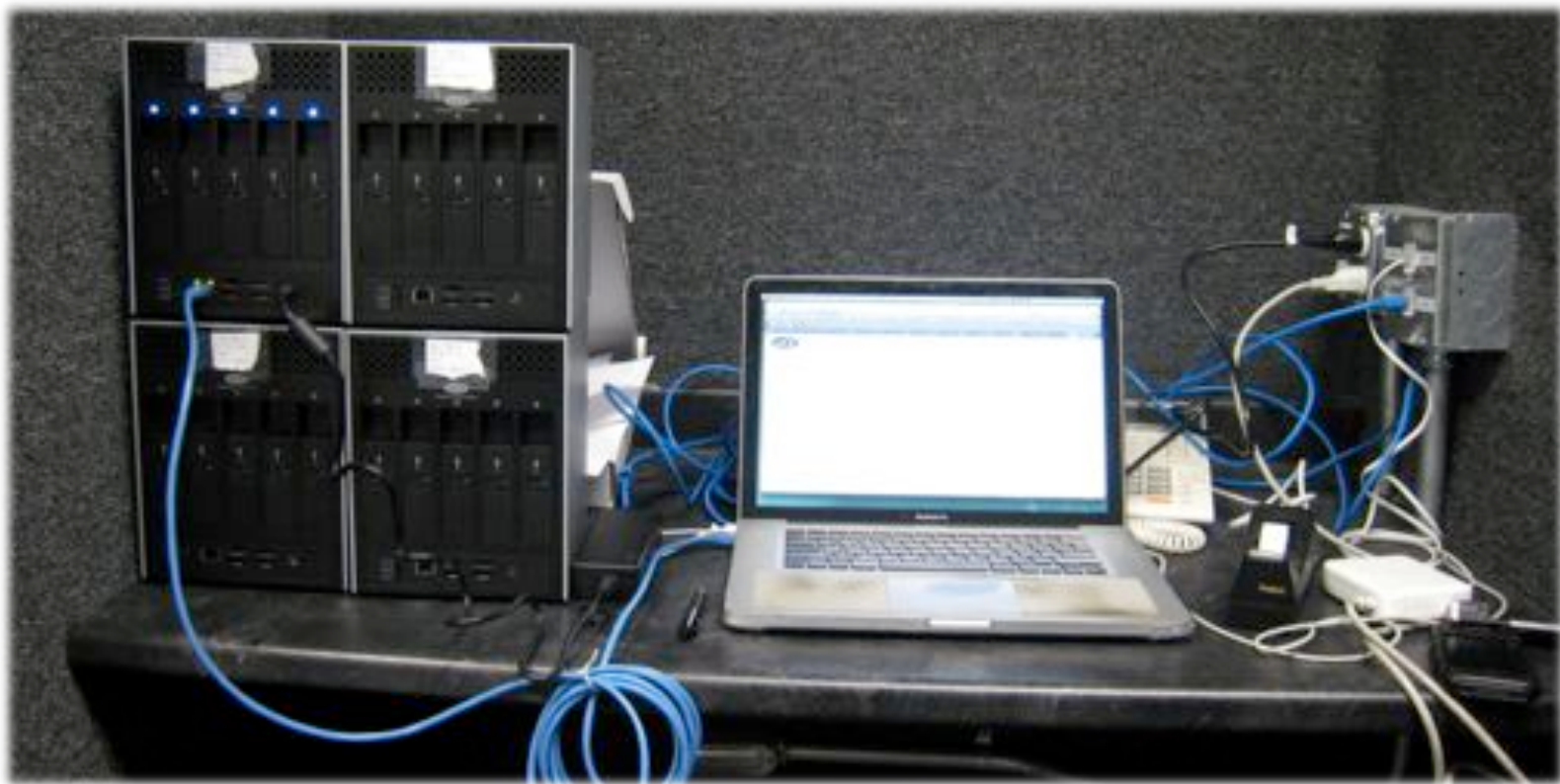
# Data Issues

- Complaint
  - Data movement still #1 technical challenge
  - Harder problem than it initially appears
- Status Late 2010:
  - Incremental improvements seen w/ best practices for physical & network-based movement
  - Hopeful for 2011:
    - Expansion of peering options with cloud providers
    - Cloud providers joining high speed research networks

# Picture Tour!

Physical data movement in the real world ...

## Physical Data Transfer Station -1



## Physical Data Transfer Station - 2





## Physical Data Transfer Station - 3





## SATA Toasters!



## Physical storage of @ scale data movement materials



## Physical storage example - 2





# Network Issues

- Complaint
  - Still nasty to run MPI or latency sensitive codes
  - Subnet issues & fan-out of EC2 servers
  - Everybody still rolling their own software VPN overlays
- Status Late 2010:
  - VPC and Elastic IP far more usable now
  - Hadoop-friendly Bioinformatics apps gaining steam
  - cc1.4xlarge instance types can help
    - Grouped for close network topology
    - Full bisectional (and not oversubscribed) 10GigE

# Wrapping Up

Bioinformatics on IaaS as of late 2010 ...

# Late 2010: Status Update

- Fast becoming mainstream & accepted
  - Just about everyone is experimenting/trialing
  - Identifying the obstacles is not hard
- We are well past the initial learning curves
  - ... *wow did I do some dumb stuff back in '07* ...
- Handling “legacy” workflows is easier than ever
  - MIT Starcluster, AWS compute cluster instances



## Late 2010: Status Update, cont.

- Easier and easier to run apps in a 'cloudy' way
  - Reference architectures & best practices emerging
  - Elastic Map Reduce, CycleComputing, RightScale
- I personally believe that commercial cloud platforms currently hold more promise than internal/private clouds
  - Cost, complexity, feature & economic benefits less clear except for edge/niche cases
  - Still hard to uncover the nuggets of truth from the immense piles of marketing BS being shoveled our way

end;

- Questions?
- Comments/feedback welcome;
- Watch <http://blog.bioteam.net> to see our cloud efforts unfold

<chris@Bioteam.net>

Shameless Plug:

BioTeam AWS Cloud Training

<http://healthtech.com/cloud>

Boston: Nov 1-2, 2010

San Fran: Feb 21-22, 2011

