

HPC From The Trenches

2010 BioITWorld Conference & Expo

Track 1 – IT Infrastructure & Hardware



chris@bioteam.net - <http://www.bioteam.net>

Welcome!

Logistics & News

- Volcano Victim
 - Phil Butcher's talk

Why I'm here

- 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 because ...
 - We see many organizations of varying type, size & structure
 - Gives us good perspective on current state of the industry



Disclaimer

- Spent 10+ years solving IT and informatics challenges in demanding production environments
 - This does not mean that I'm an expert in *anything*
- I am not / trying hard not to be:
 - "a visionary"
 - "a pundit"
 - "an industry expert"
- I speak personally about my views and experiences out in the field
 - Filter my words accordingly!



Today's Remarks

Observations

- What we've seen

Trends

- Current Trends
- Emerging Trends

Issues

- Headaches ahead

Observations

What we've seen over the last year ...

Observations

Hardware

- 10 Gigabit Ethernet
 - Not spreading as fast as I thought
 - Still mainly used for:
 - Linking storage to network
 - Linking edge switches to network core
 - Linking network core with Blade enclosure(s)

Observations

Hardware

- Blades have not taken over
 - Blades vs. rackmount form factor still ongoing
 - In new datacenter projects we've seen both done very well
- Choice driven by:
 - Preferred methods for reducing operational burden
 - Preferred vendor agreements
 - Facility issues
 - Mainly power density per floor tile

Observations

CPU Architectures

- x86_64 still the overwhelming platform choice
- Dual-socket, quad-core Intel Nehalem used to be the server “sweet spot” in almost all cases
- This seems to be changing ...
 - AMD is back in the game
 - New trends in base hardware config are emerging

Observations

CPU Architectures

- CPU selection methodologies seem to be changing
- Previous
 1. Benchmark with science apps & buy what works best
 2. Sub-select on performance/watt or performance/\$
- Now:
 - Non-science drivers heavily influencing CPU selection
- Why?
 - Enterprise needs take precedence:
 - Virtualization platform standard
 - Socket/core preferences

Observations

Distributed Resource Management

- Sun Grid Engine & Platform LSF still going strong
 - Somewhat surprising to me
 - ... expected steady state & eventual decline
 - Oracle/Sun merger also complicating things
- Based on 2010 consulting ...
 - Grid Engine project count will be higher than 2009
 - New deployments & revamps of existing systems

Observations

Private Internal Compute Clouds

- Still stupid in 2010
 - 5% useful, 90% empty hype & cynical marketing
- Two types of “private clouds” observed:
 1. Marketers excreting the “c-word” onto the same VMWare/Xen virtualization methods many of us have been using for ages
 2. Thinly veiled sales pitch from people aiming to gut and replace everything in your datacenter

Observations

Private Internal Compute Clouds

- Hype aside, obstacle is mainly legacy/technical
 - “cloud benefits” come from serious automation and live migration of running servers & services
 - Currently both Xen & VMWare can only do live migration within a subnet:
 - How many of you run a single flat subnet that spans your entire datacenter?
 - ... your entire campus?
 - ... span multiple datacenters?

Observations

Cloud Computing

- No longer considered edgy, cool or radical
- Used by few, investigated by most
- Almost ready to be called mainstream part of the IT resource toolkit within life science informatics
- Fills a need/niche, not a total solution

Observations

Cloud Computing

- No longer cool? How did we get here?
 - Enterprise long trending away from monolithic tightly coupled applications
 - Web services, SOA's and loosely-coupled systems becoming common
 - Hypervisor-based virtualization proven, not-scary and absolutely mainstream
 - Many public SaaS & PaaS cloud success stories
 - **... feels like just another step in evolution of enterprise & research IT**

Observations

Cloud Computing

- Even in the most conservative IT shops
 - Virtualization no longer seen as risky or scary
 - Hypervisor methods delivering clear benefit
 - Google Apps, etc. showing *SaaS* is more than hype
- Thus
 - Significant curiosity in efforts that virtualize, automate and commoditize ***infrastructure***
 - More than just server virtualization
 - More than just a development platform
 - “Infrastructure-as-a-Service” (*IaaS*) taking off

Observations

Tier 1 Storage

- Chance that some Tier 1 storage vendors may slip
- Starting to see significant differentiation in features, performance and density
- Some vendors falling behind on supporting density that customers are asking for
 - If a storage vendor can't fit a petabyte (or more) of disk in a single rack they better have a good reason
 - ... because competitors are shipping such systems today

Flops Failures & Freakouts

Storage war stories from 2009-2010

#1 - Unchecked Enterprise Architects

- Scientist: *"My work is priceless, I must be able to access it at all times"*
- Storage Guru: *"Hmmm...you want H/A, huh?"*
- System delivered:
 - Small (< 50TB) Enterprise FC SAN
 - Asynchronous replication to remote DR site
 - Can't scale, can't do NFS easily
 - ~\$500K/year in support & operational costs

#1 - Unchecked Enterprise Architects

- Lessons learned
- Corporate storage architects may not fully understand the needs of HPC and research informatics users
- End-users may not be precise with terms:
 - “Extremely reliable” means “no data loss”, not 99.999% uptime at a cost of millions
- When true costs are explained:
 - Many research users will trade a small amount of uptime or availability for more capacity or capabilities

#2 - Unchecked User Requirements

- Scientist: *"I do bioinformatics, I am rate limited by the speed of file IO operations. Faster disk means faster science. "*
- Storage Guru: *"Hmm. You want speed, huh?."*
- System delivered:
 - Budget blown on top tier 'Cadillac' system
 - Fast *everything*
- Outcome:
 - System fills to capacity in 9 months, zero budget left

#2 - Unchecked User Requirements

- Lessons learned
 - End-users demand the world
 - Necessary to really talk to them and understand their work, needs and priorities
- You will often find
 - The people demanding the “fastest” storage don’t have actual metrics to present
 - Many groups will happily trade some level of performance in exchange for a huge win in capacity or capability

#3 - D.I.Y Cluster/Parallel File systems

- Common source of storage unhappiness
- Root cause:
 - Not enough pre-sales time spent on design and engineering
- System as built:
 - Not enough metadata controllers
 - Poor configuration of key components
- End result:
 - Poor performance or availability

#3 - D.I.Y Cluster/Parallel File systems

- Lessons learned:
 - Software-based parallel or clustered file systems are non-trivial to *correctly* implement
 - Essential to involve experts in the initial design phase
 - *Even if using 'open source' version ...*
 - Commercial support is essential
 - *And I say this as an open source zealot ...*

Current Trends

Current Trends

“Fat” Hardware

- Significant increase in purchases of “fat” systems
 - 32+ CPU cores, 128GB RAM (or much more)
 - Available from many vendors these days
- Previously we saw these systems mainly with people doing annotation & assembly
- ... now we see many groups using them
 - Lots of applications for high-mem or big-SMP

Current Trends

“Fat” Clusters

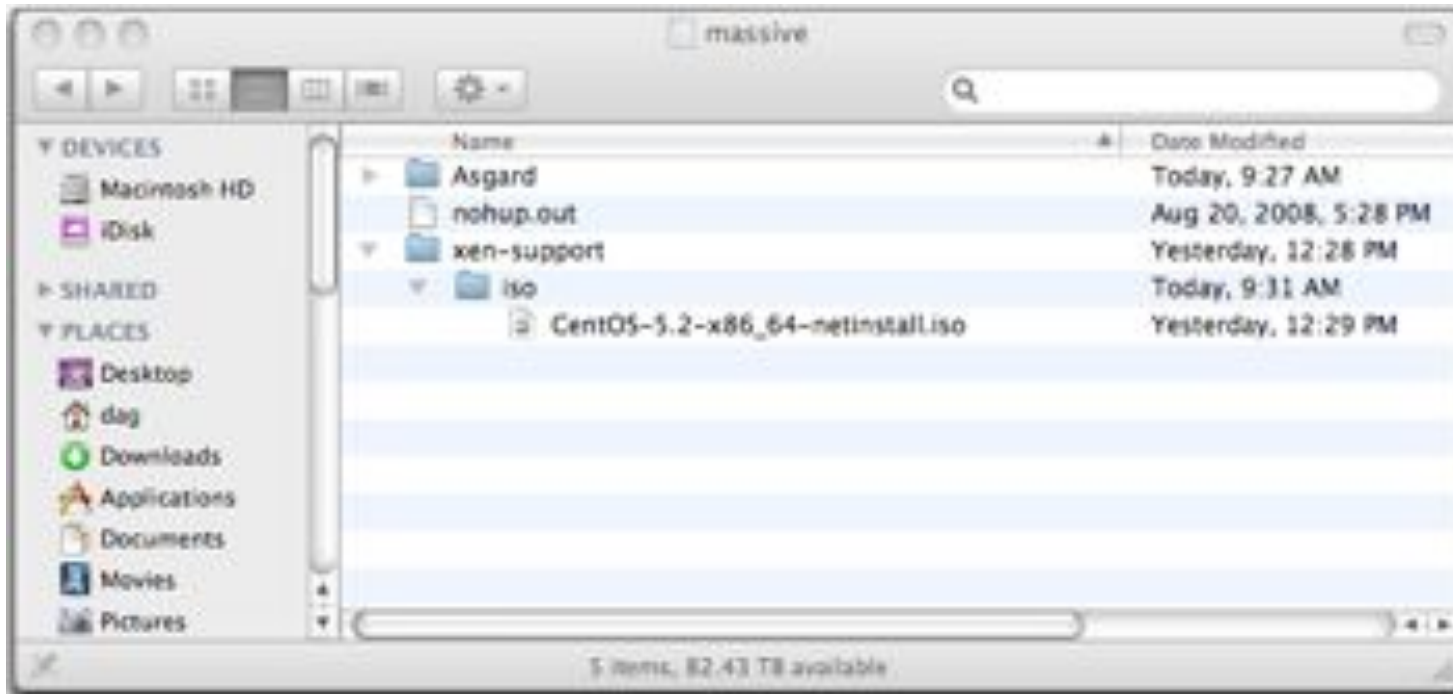
- Cluster node sweet spot generally:
 - Dual-socket, quad-core Intel Nehalem
 - Bought by the dozens or hundreds
- Some people are
 - Building small clusters (2-6 nodes)
 - ... built from “fat” systems
- And in some cases ...
 - A single “fat” system can replace a small legacy cluster

Current Trends

Storage

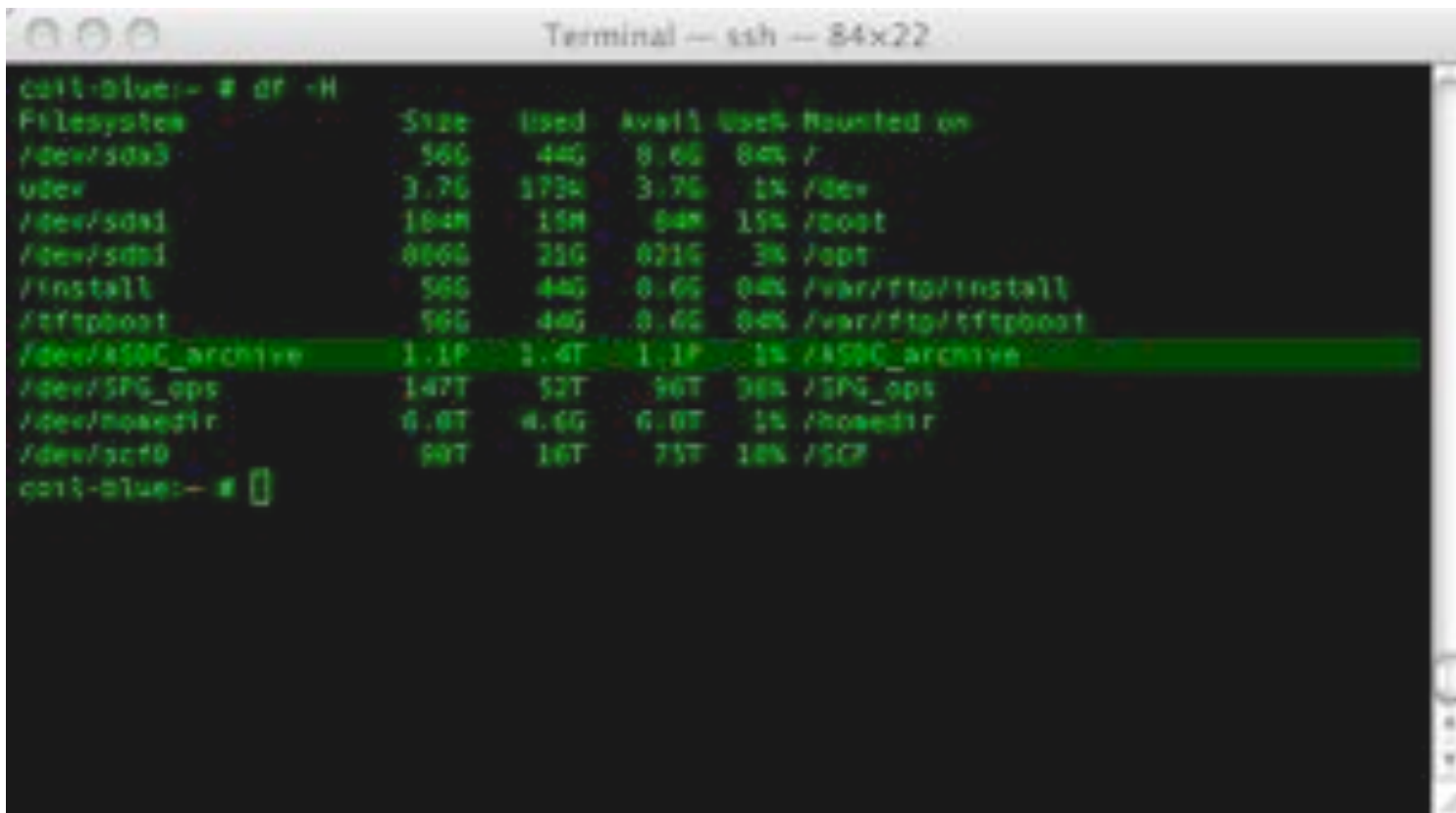
Single Namespace

- 82.4 free terabyte of space in this folder
- Very satisfying



Single Namespace

- 1.1 Petabytes free space
- Even more satisfying

A terminal window titled "Terminal — ssh — 84x22" showing the output of the command "cat@blue:~# df -H". The output is a table with columns: Filesystem, Size, Used, Avail, Use%, and Mounted on. The row for "/dev/xsdc_archive" is highlighted in green, showing 1.1P available space. Other filesystems include /dev/sda3, udev, /dev/sda1, /dev/sda1, /install, /tftpboot, /dev/SPG_ops, /dev/hwmedir, and /dev/scf0.

```
cat@blue:~# df -H
Filesystem      Size  Used Avail Use% Mounted on
/dev/sda3        56G   44G   8.6G   84% /
udev            3.7G  173k  3.7G    1% /dev
/dev/sda1       184M    15M    84M   15% /boot
/dev/sda1       886G   25G  821G    3% /opt
/install        56G   44G   8.6G   84% /var/ftp/install
/tftpboot       56G   44G   8.6G   84% /var/ftp/tftpboot
/dev/xsdc_archive 1.1P  1.4T  1.1P    1% /xsdc_archive
/dev/SPG_ops    147T   52T   96T   36% /SPG_ops
/dev/hwmedir    6.0T   4.6G   6.0T    1% /hwmedir
/dev/scf0       90T   16T   75T   18% /SCF
cat@blue:~#
```

User Expectation Management

- End users still have no clue about the true costs of keeping data accessible & available
- “I can get a terabyte from Costco for \$220!” (Aug 08)
- “I can get a terabyte from Costco for \$160!” (Oct 08)
- “I can get a terabyte from Costco for \$124!” (April 09)
- “I can get a terabyte from NewEgg for \$84!” (Feb 10)
- IT needs to be involved in setting expectations and educating on true cost of keeping data online & accessible



Storage Trends

In 2008

- First 100TB single-namespace project
- First Petabyte+ storage project
- 4x increase in “technical storage audit” work
- First time witnessing 10+TB catastrophic data loss
- First time witnessing job dismissals due to data loss
- Data Triage discussions are spreading well beyond cost-sensitive industry organizations



Storage Trends

In 2009

- More of the same
 - 100TB not a big deal any more
 - Even smaller organizations are talking (or deploying) petascale storage



Storage Trends

Now in 2010 ...

- Peta-scale is no longer scary
- A few years ago 1PB+ was somewhat risky and involved significant engineering, experimentation and crossed fingers
 - Especially single-namespace
- Today 1PB is not a big deal
 - Many vendors, proven architectures
 - Now it's a capital expenditure, not a risky technology leap



Storage Trends

Now in 2010 ...

- Worrisome Trend
 - Significant rise in storage requirements for post-instrument downstream experiments and mashups
 - The decrease in instrument generated data flows may be entirely offset by increased consumption from users working downstream on many different efforts & workflows
 - ... *this type of usage is harder to model & predict*



Emerging Trends

Don't hold me to these ...

Potential Trends

Cloud PaaS in the lab

- Cloud platform services well established
- Potentially a very good fit for:
 - Inventory tracking & sample management
 - LIMS systems
 - Protocol or experiment management
- It would be nice to replace those dusty vendor-locked and hard to support PC systems scattered under various lab benches ...

Potential Trends

Cloud Storage

- Google, Amazon, Microsoft, etc. all operate at efficiency scales that few can match
 - Cutting-edge containerized data-centers with incredible PUE values
 - Fast private national and trans-national optical networks
 - Rumors of “1 human per XX,000 servers” automation efficiency, etc.
 - Dozens or hundreds of datacenters and exabytes of spinning platters

My Argument

- Not a single person in this room can come anywhere close to the IT operating efficiencies that these internet-scale companies operate at every day
- Someone is going to eventually make a compelling service/product offering that leverages this ...

Potential Trends

Cloud Storage

- Cheap storage is easy, we all can do this
- Geographically replicated, efficiently managed cheap storage is not easy
 - ... or not cheap
- When the price is right ...
 - I see cloud storage as being a useful archive or deep storage tier
 - Probably a 1-way transit
 - Data only comes “back” if a disaster occurs
 - Data mining & re-analysis done in-situ with local ‘cloud’ server resources if needed
- Not ready for prime time yet – ask me again in 2011
 - Many questions, concerns & issues – all valid
 - May never be ready for production/enterprise use
 - My gut feeling is that majority of obstacles are surmountable

Issues

Major & minor things I expect to bother me in 2010

Issues

Virtualization & “Fat” nodes

- People are well underway with full-on virtualization strategies
- However, some of the “fat” nodes have CPU and memory resources that exceed what a hypervisor can easily provision

Issues

Internet speeds & cloud performance

- Not all of us can get on heavily subsidized high speed research networks
- Our connection to the internet & external collaborator sites is becoming more and more important
- Even worse, it's not just the size of the pipe that matters. Your location & peering arrangements matter.
 - BioTeam/Boston -> Amazon S3 Storage Cloud
 - Full utilization of available pipe (we were the bottleneck)
 - Sanger/UK -> Amazon S3 Storage Cloud
 - 10% utilization of available circuit speed

Issues

Exploding size of downstream data

- The next-gen DNA sequencing data deluge will eventually go away
- However
 - Storage consumption by researchers working on variations & mashups of this data is rising fast
 - This use case is much harder to model & predict than output from a lab instrument
 - Will cause headaches in 2010 and beyond
- Expect some words from Matthew Trunnell on this

Issues

Cloud Best Practices, HowTo's & online documentation

- Feels just like the Beowulf-cluster days
 - Available documentation simply **wrong** for the needs of the life science community
- If you followed the internet advice:
 - You'd have ended up building a cluster architected primarily for running latency-sensitive MPI applications
 - Useful but not the best design most for life science informatics requirements & use cases

Issues

Cloud Best Practices, HowTo's & online documentation

- Same thing is happening with cloud practices
- Example Google Search:
 - Improving performance of Amazon EBS storage
 - What you'll find:
 - Docs mainly written by hardcore database people
 - ... who mainly care about random IO performance
 - This is not generally our primary concern
 - Following the online recommendations might cost extra money while yielding little in the way of actual performance gain

Issues

Accurate accounting for the true cost of IT operations & services

- Becoming essential to know in detail what it costs to maintain, run, staff & operate your IT infrastructure
- This is quite hard and subject to political manipulation in some cases
- If you can't do this accurately ...
 - Impossible to know if alternatives or cloud approaches are worth pursuing

Issues

Accurate accounting for the true cost of IT operations & services

- Example from yesterday's cloud workshop
- Amylin Pharmaceuticals talk:
 - Huge effort to construct a spreadsheet that tracked the real cost of delivering each IT service
 - When compared against actual budget, the spreadsheet was accurate to within \$2K out of a 20M budget!
 - Incredible benefits from this data
 - Realized it cost \$2M/year to run HR internally
 - Many IT staff simply "keeping lights running" and not driving business or scientific success
 - Information shared widely, senior managers really did not like being associated with the most expensive services
 - ... lead to organizational changes & operational methods that deliver huge recurring savings

Issues

Cloud Politics & the changing role of research IT

Issues

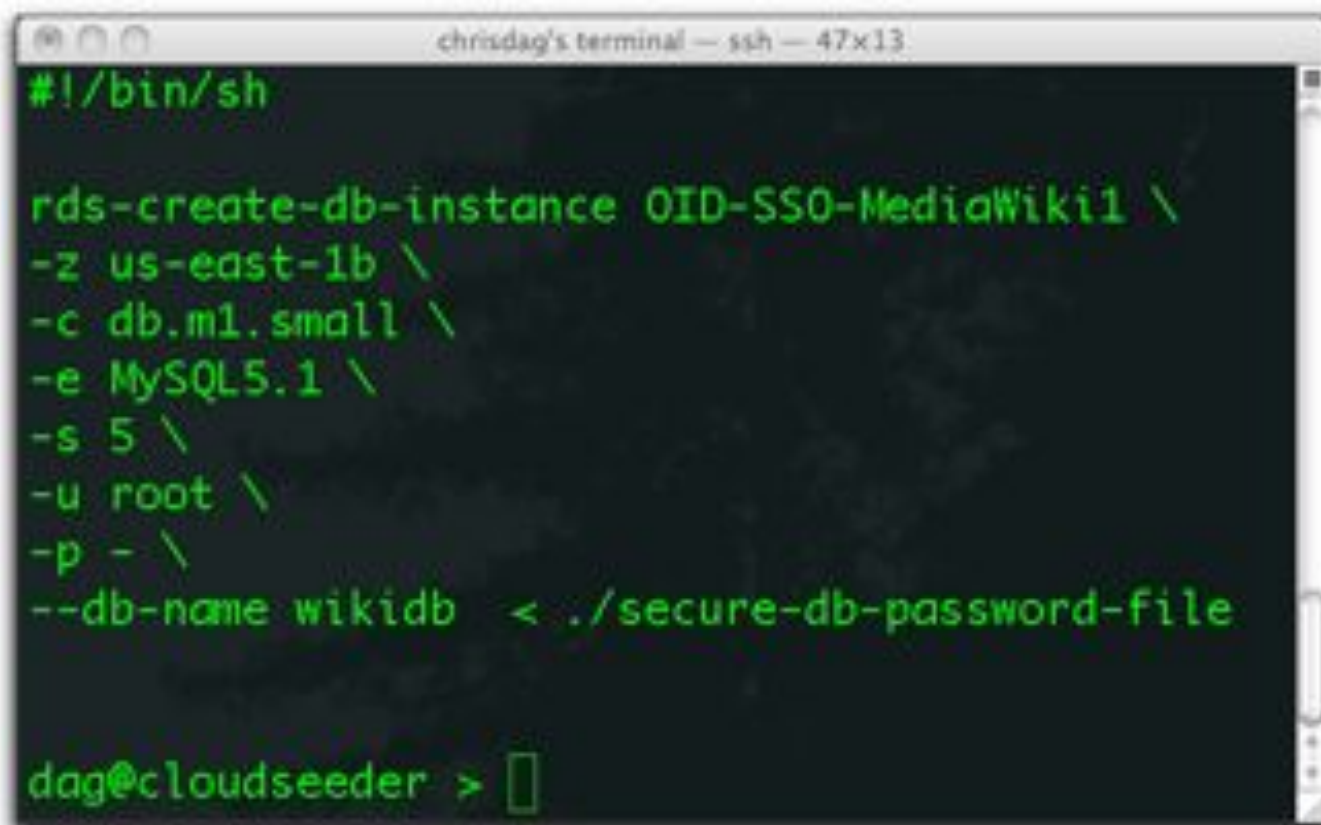
Clouds raise internal issues ...

- 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 see ...

- 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?

“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.

Politics & Scriptable IT

- *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 ...

And with that ...

end;

- Thanks!
- Talk slides will be up on <http://blog.bioteam.net> shortly
- Comments/feedback - <chris@bioteam.net>