HPC From The Trenches

2011 BioITWorld Conference & Expo

Track 1 – IT Infrastructure & Hardware

BIOTEAM
Enabling Science

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

Who is this guy?

- 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 Topics

- Business/Market Observations
- Server, Facility & Datacenter
- Storage (current state)
- Storage (what comes next)
- Cloud

Biz Observations

Gartner we ain't

Business / Meta Observations

- BioTeam neatly straddles “science”, “software” and “high performance IT” as a bespoke consulting shop
- Our consultants get pulled in the direction of any large scale shifts in industry trends or practices
- Looking at what our staff has been asked to do over the past 12 months can sometimes give good info on what pain/problems our market is experiencing

Business / Meta Observations

- In 2010
  - ~4 BioTeam staff engaged almost full time on issues regarding data handling, data management and multi-instrument Next-Gen sequencing handling
  - Ranging from **BioTeam MiniLIMS** deployments to general consulting/support/assistance with science, software & process issues related to sequencing & resequencing efforts

Business / Meta Observations

- In 2010
  - Two staff engaged almost full time on infrastructure, storage and facility related projects
    - Dwan: Big infrastructure & facility projects for Fortune 20 companies and .GOV customers
    - Dag: 40% infrastructure, 20% storage, 20% cloud
  - One consultant full time on Amazon Cloud projects
    - Adam K

What that tells us

Data & Data Management

- Next-gen sequencing still causing a lot of pain when it comes to data handling, storage, organization & integration
- As sequencing continues to be commoditized, this will likely only get worse

What that tells us …

Storage

- Storage is still a problem in 2011
- But not as bad as it was in the past…
    - We are not “putting out fires” as much
    - Far more storage refresh/redesign or long term storage planning efforts
    - More on this later …

What that tells us …

Datacenter & Facility Projects

- Significant action in datacenter/facility areas
  - Large and small organizations are making major changes to facilities and their plans for future facilities
  - More and more orgs are securing colocation space, often for power density reasons
  - **A few projects are being driven by urgent scientific requirements but many appear to be the natural result of cyclic IT refresh/re-examine cycles**

Server & Facility Observations

Server Density & CPU Core Proximity

Danger Ahead

- Starting to see more facility issues causing rifts between IT/facility managers and scientific researchers
- Problems are being caused by issues and arguments over CPU density, core counts and preferred server specs
- Causing real stress & real problems in organizations today

Datacenter Rack Example

IT Manager View, example 1

- Using standard (conservative) dual-socket 1U servers we can get 6-12 CPU cores per rack unit
- At ~230watts per server we can easily stick 40 servers in a standard rack
- **This gets us to 480 CPUs per cabinet and 3,840 CPU cores for an 8-cabinet row**
- All within a reasonable 13kW power envelope per rack

Datacenter Rack Example

IT Manager View, example 2

- Using moderate-density 1U “Twin-Servers” we can get 12-24 CPU cores per 1U rack space
- At ~440 watts per server we can put 28 servers in, leaving 14U unused space in each cabinet
  - 14U unused space is not too much ‘waste’
- **This gets us to 672 CPUs per cabinet and 5,376 CPU cores for an 8-cabinet row**
  - *Without using exotic packaging or blades*
- All within a reasonable 13kW power envelope per rack

Meet the enemy …

Aaaargh! screams the IT manager

- Quad-socket, 12-core AMD CPUs (48 CPU cores total)
- 32 DIMM slots for up to 512GB of physical memory
- Fits neatly in 2U of rack space and costs less than $16,000
- Oh yeah, this box has …
- **1400 watt power supplies**
- **< Gulp . >**

48 Cores & 256GB RAM in 2U

Meet the enemy …

This is a problem

- 2U boxes with 1400 watt power supply demands are problematic for some datacenters
- Unless you have been planning for blades or other high density environments …
- You might be lucky to fill 1/3 of your cabinet before exceeding the available power or cooling envelope

Your scientists want these.

Even worse …

- **Researcher demands for these 48-core / large memory nodes are entirely justified**
  - Still hard to get access to large memory systems from cloud or IaaS providers
  - **Hard to satisfy these needs with VMs and Blades**
  - The Mem:Core ratio still works out for general unprofiled informatics apps & use cases

Even worse (continued) …

- **Researcher demands for these 48-core / large memory nodes are entirely justified …**
  - Mammalian genome assembly requires large memory systems (256GB and higher usually)
  - Our crap code is still not MPI aware or re-architected to run under MapReduce frameworks
    - So having lots of cores and memory available **inside a single chassis** is often the **only solution** for our inefficient, non-distributed legacy codes & algorithms

In a nutshell

- Datacenter operators must carefully balance efficient use of (limited) physical space with available power & cooling constraints
- Easy to do when our "sweet spot" was dual-socket quad-core servers with 24-48GB RAM
- Much harder to do when scientists have legit needs for large memory & large CPU in a single server chassis
- Hard to deploy these types of systems on Virtualized or Blade infrastructures as well

Datacenters

- We see lots of datacenters in any given year
- My favorite facility in the last year belongs to:
  - biogen idec
- Bad news:
  - Confidentiality agreements prohibit me from telling you what I saw there (I don't know what they don't want you to know!)
- Good news:
  - Mike Russo from Biogen is speaking today at 3:45pm re: "**Biogen Idec's Data Center Redesign**"

Storage Landscape

Storage Sub-Topics

1. Meh, the sky is not falling
2. OMG! The sky IS falling!
3. What comes next?

The sky is not falling

No need to panic.

Life Science Data Deluge

- Scare stories and shocking graphs are getting pretty tiresome
- We've had "terabyte instruments" in our wet labs since 2004
  - … and somehow we've managed to survive
- Next few slides
  - Try to explain why storage does not scare me all that much in 2011

Sky is not falling

1. You are not the Broad Institute or Sanger Center

- Overwhelming majority of us do not operate at Broad/Sanger levels
  - *Those folks are adding 200TB/week in Tier1 storage*
- We still face challenges but the scale/scope is well within the bounds of what traditional IT technologies can handle
- Large cohort w/ [1-4] or [4-12] NGS instruments
- We've been doing this for years
  - Many vendors, best practices, "war stories", proven methods and just plain "people to talk to…"

Sky is not falling

2. Instrument Sanity Beckons

- Yesterday: .TIFF overload
- Today: RTA, in-instrument data reduction
- Tomorrow: Basecalls, BAMs & Outsourcing
- Day-after-tomorrow: Write directly to the cloud

Sky is not falling

3. Peta-Scale Storage is not really exotic or unusual anymore

- Peta-scale storage has not been a risky exotic technology gamble for years now
- Today it's just an engineering & budget exercise
  - ▫ Multiple vendors don't find petascale requirements particularly troublesome and can deliver proven systems within weeks
  - ▫ < $1M will get you 1PB from several top vendors
- However, still HARD to do BIG, FAST & SAFE
  - ▫ Hard but solvable, many resources & solutions out there

On the other hand …

OMG!!!! The sky IS falling!

Ok, we might have to panic a little bit.

The Sky IS Falling!

1. @!*#&^@ Scientists

- As instrument output declines …
- Downstream storage consumption by end-user researchers is increasing rapidly
- Each new genome generates new data mashups, experiments, data interchange conversions, etc.
- WAY, way harder to do capacity planning against human beings vs. instruments

The Sky IS Falling!

2. @!*#&^@ Scientific Leadership

- Sequencing is being commoditized
  - Same path for other high-data instruments
- NOBODY simply banks the savings
- EVERYBODY buys more instruments
  - 50% price reduction = 50% more platforms
- BioTeam late-2010 anecdotes from small/mid labs:
  - 10 HiSeq's not too extraordinary
  - 20 genomes to Complete Genomics no big deal

The Sky IS Falling!

Gigabases vs. Moore's Law

BIG SCARY GRAPH

The Sky IS Falling!

3. Something is going to break

- This is not sustainable
  - Downstream consumption exceeding instrument data reduction
  - Commoditization yielding more platforms
  - Chemistry moving faster than IT infrastructure
  - What the heck are we doing with all this sequence?

Storage Landscape

Some interesting vendor movement in the last year or so

Positive Storage Trends

- In my storage-focused talks I usually show slides covering "**flops, failures & freakouts**"
  - Compiled list of storage disasters we've witnessed
- Majority of these disasters have two root causes:
  - Storage vendors slow to offer solutions more tailored to our industry need …
  - Forcing storage customers to make difficult "budget vs. burden" decisions
    - What can I afford to ***buy***?
    - What can I afford to ***manage***?

Positive Storage Trends

What has changed?

- My #1 storage problem is largely resolved
- In the past, if I needed “petabyte capable” storage I was forced into the most extreme end of the highest-performing storage systems on the market
- Designed for supercomputing sites where raw performance is ***everything***
- ***Man that stuff is expensive!***

Positive Storage Trends

What has changed?

- Life science people need peta-scale ***capable*** systems because we can't predict what our researchers are going to need 2-3 years down the road
  - *Remember that lab protocols are changing way faster than our IT refresh cycles!*
- **Difficult choices to make:**
  - **Spend lots of money on an exotic system?**
  - **Buy something smaller but risk having to throw it away in 18 months?**

Positive Storage Trends

What has changed?

- Peta-capable storage much easier now …
- I'll use three vendors as an example
    - Isilon
    - BlueArc
    - Panasas

Isilon

Abusing our NL nodes …

- In the last two consulting projects where I thought Isilon would be a good fit …
- I ***started*** my baseline configuration from the Isilon NL series storage nodes
- "NL" stands for Near Line and is intended to be used in archival or secondary tiers

Isilon, continued

NL Series To the rescue

- We start with the NL series density-optimized storage nodes as our **PRIMARY TIER**
- We swap out or add additional faster or more specialized Isilon kit as business or scientific need dictates
- This works great even if Isilon sales reps are not thrilled

BlueArc

Mercury NAS heads FTW

- In the past, if I wanted Peta-scale storage from a NAS appliance vendor …
- Forced to purchase the top-tier filer heads in order to get the capacity headroom I needed
- … if we went cheaper we ran the risk of obsoleting our expensive filer head before its time
- With BlueArc this meant using their Titan series filer line …

BlueArc, continued

Mercury NAS heads FTW

- All of this changed with the introduction of BlueArc Mercury line
- Cheaper than the Titan series hardware, aimed at the midrange of the market
- Yet still capable of scaling into 1PB, 2PB, 4PB range
- … means I no longer fear having to retire/replace expensive filer head if capacity demands explode suddenly

Panasas

Hybrid PAS systems …

- Panasas launched PAS-12
- *"Fastest HPC Storage System In The World!"*
- You can see one on the show floor today
- However, the PAS-12 is likely overkill for many of my storage requirements …

Panasas, continued

Hybrid PAS systems …

- When I mentioned this concern to Panasas people …
- They were already way ahead of me
- Their approach:
  - Use PAS-12 director blade
  - Mix in PAS-8 and PAS-9 storage blades as needed
- This hybrid PAS-12/PAS-N approach looks very interesting

Storage: What comes next …

Next 12 months are going to be fun.

What comes next

Same rules still apply in 2011 and beyond …

- Accept that science changes faster than the underlying IT infrastructure
- Be glad you are not Broad/Sanger
- Flexibility, scalability and agility become the key requirements of research informatics platforms
- Shared/concurrent access is still the overwhelming storage use case
    - NAS, NAS, NAS

What comes next

In 2011 …

- Many peta-scale *capable* systems will be deployed
  - **Majority** will operate in *the **hundreds-of-TBs** or smaller* range
- Far more aggressive "data triage" will continue to be seen in higher-scale organizations
  - ".BAM only!"
- Even more data will sit untouched & unloved
  - Opportunities for tiers, HSM & even tape

What comes next

In 2011 …

- Broad, Sanger and others will pave the way with respect to metadata-aware & policy driven storage frameworks
  - And we'll shamelessly copy a year or two later
- Distributed Bio is going down this road via iRODS
  - Well worth keeping an eye on them

What comes next?

NFS v4.1 & pNFS to the rescue?

- I'm sort of an NFS bigot
  - ▫ It's simple
  - ▫ It works
  - ▫ Leverages commodity ***everything***
  - ▫ Supports way more use cases than traditional SAN
  - ▫ Far less complex than parallel & distributed filesystems
  - ▫ Easy to see the limitations
    - Easy to work around them
    - … and if not, easy to see when "something else" needed

NFS (v4.1) to the rescue?

- NFS has long been an issue for us
  - Most use it for it's simplicity & engineer around the bottlenecks and performance limits
  - When that is not possible:
    - Some people go GPFS/LUSTRE
    - Some people glom NAS gateway(s) onto SANs
    - Some people go scale-out NAS via specialized solutions
  - Rule of thumb:
    - Start with NFS, engineer something else if NFS can't handle the technical requirements

pNFS: What's the big deal?

- 2011 will see the widespread release of parallel NFS efforts that have been in development since 2005
- An actual industry standard
  - RFC 5661
  - http://tools.ietf.org/html/rfc5661
- It's out there!
  - BlueArc demo pNFS @ SC 2010
  - Fedora Core 13 has NFS v4.1 client/server RPMs

pNFS: What does it mean for me?

- For management types:
  - Storage investment becomes less of a scary risk
  - Industry standard pNFS offers the excellent ***scale-out*** and ***scale-up*** previously seen only if you were willing to go exclusive with a storage provider running a proprietary solution
- For end users:
  - **Performance**. Parallel FS users might not see much change but people switching from NFS to pNFS should see huge gains

So what?

- Not very different from how proprietary or experimental solutions have been used for ***years***
- *What makes it different:*
  - □ *An actual industry standard*
  - □ ***Native pNFS clients in your OS stack*** *(!!!!)*
  - □ *A common client for multiple storage backends*
  - □ *Fewer support issues (one hopes)*
- **“Parallel file system scaling & performance with the simplicity of standard NFS”**

pNFS: My take

- This should become a big deal in our space
- It's the right time for all of this stuff to converge
  - … straight onto 10 Gigabit Ethernet
- I love the fact that I should be able to get levels of performance via commodity solutions that previously were only available as highest-end solutions from proprietary storage platforms

pNFS: My take, continued

- Not sure what the rate of adoption will be
  - … FC13 is OK but when will it show up in RHEL?
  - Kick the tires in 2011, deploy in 2012-13?
- Feels right to me
  - Best features of parallel & distributed file systems baked into a converged standard with OS-native access clients

The C-word

Can we really talk@ BioIT without mentioning cloud computing?

Nope.

Private Cloud Thoughts

Private Internal Compute Clouds

- Just as stupid in 2011
  - 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

Private Clouds Are Dumb

My $.02 of course!

- Multi-tenant storage + VMWare
  - Wow. Big deal. Got anything interesting?
- Most effective use requires an almost total datacenter refit (good for vendors, bad for you)
- Always going to play feature catch-up with the public IaaS providers

Private Clouds Are Dumb

My $.02 of course!

- Ignores the primary benefits of public IaaS
- **Engineering**:
  - MS, Amazon, Google have years of experience running massive-scale systems in incredibly hostile environments.
  - It may look “easy” to us but under the hood is some seriously complex engineering
  - Does your private cloud vendor have the same level of engineering & deployment experience?
  - Can you afford to hire people who do nothing but squeeze out additional operational efficiencies?

Private Clouds Are Dumb

My $.02 of course!

- Ignores the primary benefits of public IaaS
- **Resiliency & Redundancy**:
  - ▫ How many datacenters are you deploying your private cloud across? How many continents? How many flood plains? How many earthquake zones?
  - ▫ What are you paying for bandwidth?
  - ▫ What PUE are your datacenters operating at?
  - ▫ Can you repurpose idle capacity and keep your total infrastructure utilization above 90%?

Private Clouds Are Dumb

My $.02 of course!

- Ignores the primary benefits of public IaaS
- **Financial savings:**
  - Core benefit of public IaaS comes from the fact that a small number of companies are doing this stuff on an obscene globe-spanning scale
  - These companies can sell us well-engineered IaaS primitives (storage, compute, etc.) cheaper than we can do it (correctly/safely/reliably) ourselves
  - Amazon, MS & Google operate at a scale and level of operational efficiency that **you can not match**

Private Clouds Are Dumb

Wrapping up this topic ..

- I’m being nasty/cynical for a reason. The entire situation feels eerily similar to the multi-site GRID computing BS from the ’90s
  - ▫ Way too much hype, way too little usefulness for industry users
- Private clouds will be great for some people:
  - ▫ Supercomputing sites & people with sovereign nation funding
  - ▫ Academic sites with lots of inexpensive labor
  - ▫ Fortune 100 / large companies that actually have many datacenters and tens of thousands of internal “customers” to service

Public Clouds (IaaS)

Sticking my neck out again …

- For the past few years I've used weasel words to suggest that Amazon AWS may end up ruling the pure IaaS cloud world
  - *"the window for competitors to catch up is shrinking …"*
- I'm ready to stop being so wishy-washy
- **Amazon IS the infrastructure cloud**
- **Period.**

Public Clouds (IaaS)

Sticking my neck out again …

I'm ready to stop being so wishy-washy

- **Amazon IS the infrastructure cloud**
- **Period.**

Amazon Web Services

Rate of Change Example

- I am going to flash through the next 6 presentation slides
- The main point is to show you the rate at which Amazon Web Services:
  1. Rolls out entirely new products and services
  2. Adds significant enhancements to existing services
- My take:
  - ▫ At this point I can't see anyone matching or even catching up …

AWS Rate of Change Examples

- 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/

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/

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/

AWS Rate of Change Examples

- July2010
  - **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/

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
    - MapReduce live resizing
    - Load Balancing w/ SSL
- November 2010
    - **GPU instance types on EC2!**
    - AWS ISO 27001 Certification

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

AWS Rate of Change Examples

- December 2010
  - Route 53 DNS Service
  - **VMWare .vmdk Import launch**
  - RDS reserved instance support
  - AWS Import/Export in Singapore
- January 2011
  - Elastic Beanstalk in AWS Console
  - Elastic Beanstalk Eclipse integration
  - Simple Email Service (SES) launched
  - CloudWatch in AWS Console
- Feb 2011
  - **IAM users can now login to AWS web console**
  - Website hosting on AWS S3
  - EBS 'force detach'
  - EC2 termination protection
  - **CloudFormation in AWS Console**

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

AWS Rate of Change Examples

- March 2011
  - **EC2 console enhancements**
    - Change instance type, shutdown behavior & user-data of stopped EBS-backed servers
  - **Major VPC Enhancements***
    - **Internet gateways**
    - **No hardware required**
  - **EC2 VPC Dedicated Instance Launch**

- April 2011
  - **??**

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

Cloud Wrap-Up

- IaaS cloud computing in the life sciences is no longer exotic or novel.
  - It's here. It's useful. Real people are doing real work on it with real benefit.
  - There is an entire Track devoted to Cloud at this very meeting
- Most people in this room are already using it or have done the due-diligence and "tire kicking" to see if it is something worth pursuing
- The real fun stuff is happening at a different level

Cloud Wrap-Up, continued

- The real action lies in what people are doing w/ “**scriptable infrastructure**” and “**cloud orchestration**” on top of IaaS providers
- Often with Opscode Chef as the orchestration master
  - Rackspace, Penguin or Amazon .. Chef handles it all!
- The other real action is what this is going to do to the **careers** and **job descriptions** of IT, engineering and application development staff
- These topics are worth a talk (or even a workshop) of their own.

And with that …

end;

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