



Data Management & Storage for NGS

XGen Congress Short Course
2010

Speakers for Today

Chris Dagdigian

Storage for NGS

Giles Day

Pfizer Informatics for NGS

Adam Kraut

Structured Data Management

Word of warning

- Known for speaking very quickly and having a large slide deck
- Also an unrepentant PowerPoint fiddler
 - Slides I deliver today are different from your packets
 - Final version will be posted on <http://blog.bioteam.net/>



BioTeam Inc.

- **Independent Consulting Shop:**

Vendor/technology agnostic

- **Staffed by:**

- Scientists forced to learn High Performance IT to conduct research

- **Our specialty:**

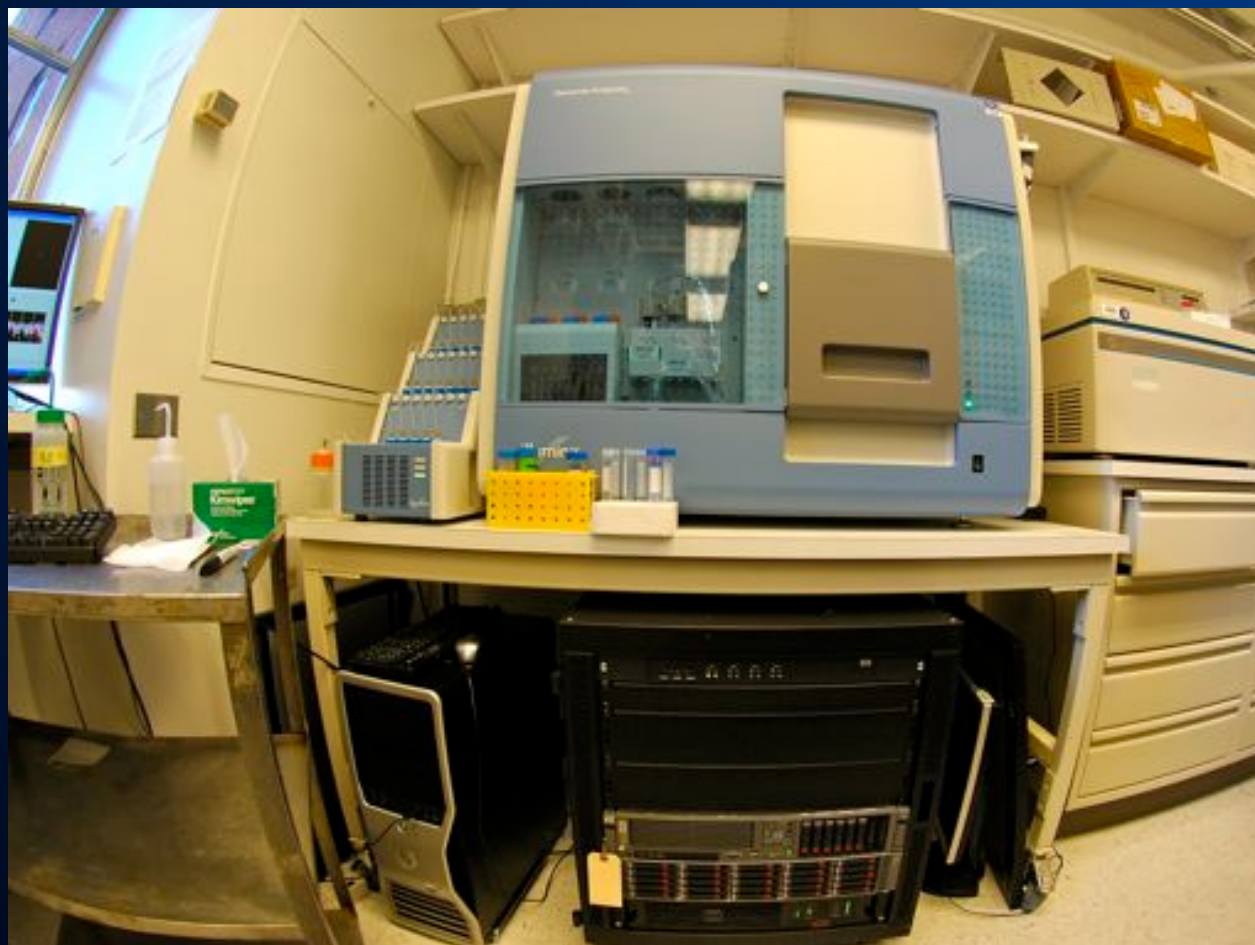
Bridging the gap between Science & IT



Science Driven Storage

Infrastructure Photo Tour

Example: Point solution for NGS



Self-contained lab-local cluster & storage for Illumina

chris@bioteam.net

Example: Small core shared IT



100 Terabyte storage
system and 10 node / 40
CPU core Linux Cluster
supporting multiple NGS
instruments

Example: Large Genome Center



Setting the stage

- Data Awareness
- Data Movement
- Data Management
- Storage & Storage Planning
- Storage Requirements for NGS
- Putting it all together ...

The Stakes ...



180+ TB stored on lab bench

The life science “data tsunami” is no joke.

Flops, Failures & Freakouts

When research storage goes bad ...

#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:**
 - 40TB Enterprise FC SAN
 - Asynchronous replication to remote DR site
 - Can't scale, can't do NFS easily
 - \$500K/year in maintenance 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."*
- **System delivered:**
 - Budget blown on top tier 'Cadillac' system
 - Fast *everything*
- **Outcome:**
 - System fills to capacity in 9 months

#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 ...*

Science Driven Storage

Back on track ...

Data Awareness

- First principals:

- Understand chemistry changes faster than IT
- Understand the data you will *produce*
- Understand the data you will *keep*
- Understand how the data will *move*

- Second principals:

- One instrument or many?
- One vendor or many?
- One lab/core or many?

Data You Produce

- Important to understand data sizes and types on an instrument-by-instrument basis
 - How many instrument runs per day/week?
 - What IT resources required for each basecall made?
- Will have a significant effect on storage performance, efficiency & utilization
- Where it matters:
 - Big files or small files?
 - Hundreds, thousands or millions of files?
 - Does it compress well?
 - Does it deduplicate well?

Data You Produce

- Cliché NGS example
 - *Raw* instrument data
 - ◆ Massive image file(s)
 - *Intermediate* pipeline data
 - ◆ Raw data processed into more usable form(s)
 - *Derived* data
 - ◆ Results (basecalls & alignments)
 - ◆ Wiki's, LIMS & other downstream tools

Data You Will Keep

- Instruments producing terabytes/run are the norm, not the exception
- Data triage is real and here to stay
 - Triage is the norm, not the exception these days
 - I think the days of “unlimited storage” are likely over
- *What bizarre things are downstream researchers doing with the data ?*
- Must decide what data types are kept
 - And for how long ...

Data You Will Keep

- Raw data \Rightarrow Result data
 - Can involve 100x reduction in some cases
- Result data \Rightarrow Downstream derived data
 - Often overlooked and trend-wise the fastest growing area
 - Researchers have individual preferences for files, formats and meta-data
 - Collaborators have their own differences & requirements
 - The same data can be sliced and diced in many ways when used by different groups

General Example - Data Triage

- Raw Instrument Data

- Keep only long enough to verify that the experiment worked (7-10 days for QC)

- Intermediate Data

- Medium to long term storage (1year to forever)
- Tracked via Wiki or simple LIMS
- Can be used for re-analysis
 - ◆ Especially if vendor updates algorithms

- Result Data

- Keep forever

Applying the example ...

■ Raw Instrument Data

- Instrument-attached local RAID
- Cheap NAS device
- Probably not backed up or replicated

■ Intermediate Data

- Almost certainly network attached
- Big, fast & safe storage
 - ◆ **Big** for flexibility & multiple instruments
 - ◆ **Fast** for data analysis & re-analysis
 - ◆ **Safe** because it is important data & expensive to recreate

■ Result Data

- Very safe & secure
- Often enterprise SAN or RDBMS
- Enterprise backup methods

NGS Vendors don't give great advice

- Skepticism is appropriate when dealing with NGS sales organizations
 - Essential to perform your own diligence
- Common issues:
 - Vendors often assume that you will use only their products; interoperability & shared IT solutions are not their concern
 - May lowball the true cost of IT and storage required if it will help make a sale

Data Movement

■ Facts

- Data captured does not stay with the instrument
- Often moving to multiple locations
- Terabyte volumes of data could be involved
- Multi-terabyte data transit across networks is rarely trivial no matter how advanced the IT organization
- Campus network upgrade efforts may or may not extend all the way to the benchtop ...

Data Movement - Personal Story

- One of my favorite '09 consulting projects ...
 - Move 20TB scientific data out of Amazon S3 storage cloud
- What we experienced:
 - Significant human effort to swap/transport disks
 - Wrote custom DB and scripts to verify all files each time they moved
 - ♦ Avg. 22MB/sec download from internet
 - ♦ Avg. 60MB/sec server to portable SATA array
 - ♦ Avg. 11MB/sec portable SATA to portable NAS array
 - At 11MB/sec, moving 20TB is a matter of *weeks*
 - *Forgot to account for MD5 checksum calculation times*
- Result:
 - Lesson Learned: data movement & handling took 5x longer than data acquisition

Data Movement: Recommendations

- Network & network design matters
- Gigabit Ethernet has been a commodity for years
 - Don't settle for anything less
- 10 Gigabit Ethernet is reasonably priced
 - We still mostly use this for connecting storage devices to network switches
 - Also for datacenter to lab or remote building links
 - 10GbE to desktop or bench top not necessary
 - ◆ *10GbE to nearby network closet may be*
- Portable disk enclosures might be appropriate
 - Remember to account for time needed for copying and checksum activities
 - Safe & secure storage is important

Data Movement: Recommendations

- Don't bet your experiment on a 100% perfect network
 - Instruments writing to remote storage can be risky
 - Some may crash if access is interrupted for any reason
 - Stage to local disk, then copy across the network
- Network focus areas:
 1. Instrument to local capture storage
 2. Capture device to shared storage
 3. Shared storage to HPC resource(s)
 4. Shared storage to desktop
 5. Shared storage to backup/replication

Data Management

- Very important
 - Especially if multiple IT & storage systems involved
- There is no universal solution
 - Too many variables in protocol, data & research flows
- We have seen many different methods adopted
 - LIMS, Wiki, Spreadsheets, etc.
 - All have pros and cons
- Choosing
 - Flexibility is key
 - ◆ Chemistry or SOP might change faster than a typical LIMS design lifecycle can handle
 - ◆ **All solutions are useless if unused**

Storage Requirements for NGS

What features do we actually need?

“Must Have”

- High capacity & scaling headroom
- Variable file types & access patterns
- Multi-protocol access options
- Concurrent read/write access

“Nice to have”

- Single-namespace scaling
 - No more “/data1”, “/data2” buckets ...
 - Horrible cross mounts, bad efficiency
- Low Operational Burden
- Appropriate Pricing*
- “A la cart” feature and upgrade options

Capacity

- Chemistry/instruments improving faster than our IT infrastructure
 - Flexibility is essential to deal with this
- If we don't address capacity needs:
 - Expect to see commodity NAS boxes crammed into lab benches and telco closets
 - Expect hassles induced by island of data
 - Backup issues (if they get backed up at all)
 - ... and lots of USB drives on office shelves ...

Remember The Stakes ...



180+ TB stored on lab bench

The life science “data tsunami” is no joke.

File Types & Access Patterns

- Many storage products are optimized for particular use cases and file types
- Problem
 - Life Science & NGS can require them all:
 - ◆ Many small files vs. fewer large files
 - ◆ Text vs. Binary data
 - ◆ Sequential access vs. random access
 - ◆ Concurrent reads against large files

Multi-Protocol Is Essential

- The *overwhelming* researcher requirement is for *shared* access to *common* filesystems
 - Especially true for next-gen sequencing
 - Lab instrument, cluster nodes & desktop workstations all need access the same data
 - This enables automation and frees up human time
- Shared storage in a SAN world is non-trivial
- Storage Area Networks (SANs) are not the best storage platform for discovery research environments

Storage Protocol Requirements

- NFS

- Standard method for file sharing between Unix hosts

- CIFS/SMB

- Desktop access
- Ideally with authentication and ACLs coming from Active Directory or LDAP

- FTP/HTTP

- Sharing data among collaborators

Concurrent Storage Access

- Ideally we want read/write access to files from
 - Lab instruments
 - HPC / Cluster systems
 - Researcher desktops
- If we don't have this
 - Lots of time & core network bandwidth consumed by data movement
 - Large & possibly redundant data across multiple islands
 - Duplicated data over islands of storage
 - Harder to secure, harder to back up (if at all ...)
 - Large NAS arrays start showing up under desks and in nearby telco closets

Data Drift: Real Example

- Non-scalable storage islands add complexity

- Example:

- Volume “Caspian” hosted on server “Odin”
- “Odin” replaced by “Thor”
- “Caspian” migrated to “Asgard”
- Relocated to “/massive/”

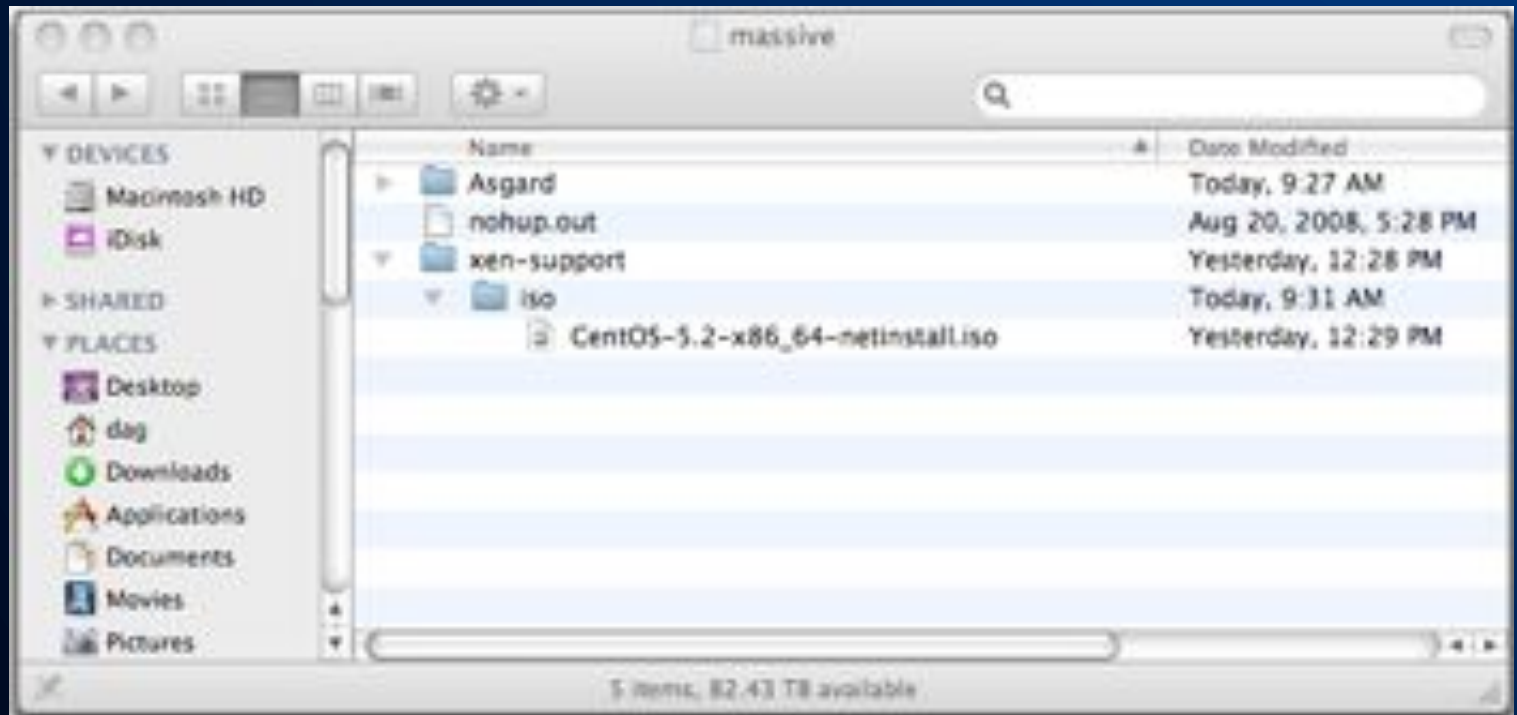
- Resulted in file paths that look like this:

`/massive/Asgard/Caspian/blastdb`

`/massive/Asgard/old_stuff/Caspian/blastdb`

`/massive/Asgard/can-be-deleted/do-not-delete...`

Single-namespace is valuable



Things To Think About

An attempt at some practical advice ...

Storage Landscape

- Storage is a commodity in 2010
- Cheap storage is easy
- Big storage getting easier every day
- Big, cheap & *SAFE* is much harder ...
- Traditional backup methods may no longer apply
 - Or even be possible ...

Storage Landscape

- Still see extreme price ranges
 - Raw cost of 1,000 Terabytes (1PB):
 - ◆ \$125,000 to \$4,000,000 USD
- Poor product choices exist in all price ranges

Poor Choice Examples

- On the low end:
 - Use of RAID5 (unacceptable in 2009)
 - Too many hardware shortcuts result in unacceptable reliability trade-offs

Poor Choice Examples

- And with high end products:
 - Feature bias towards corporate computing, not research computing - pay for many things you won't be using
 - Unacceptable hidden limitations (size or speed)
 - Personal example:
 - ◆ \$800,000 70TB (raw) Enterprise NAS Product
 - ◆ ... *can't create a NFS volume larger than 10TB*
 - ◆ ... *can't dedupe volumes larger than 3-4 TB*

One slide on RAID 5

- I was a RAID 5 bigot for many years
 - Perfect for life science due to our heavy read bias
 - Small write penalty for parity operation no big deal
- RAID 5 is no longer acceptable
 - Mostly due to drive sizes (1TB+), array sizes and rebuild time
 - *In the time it takes to rebuild an array after a disk failure there is a non-trivial chance that a 2nd failure will occur, resulting in total data loss*
- In 2009
 - Only consider products that offer RAID 6 or other “double parity” protection methods
 - Even RAID 6 is a stopgap measure ...

Research vs. Enterprise Storage

- Many organizations have invested heavily in centralized enterprise storage platforms
 - Natural question: *Why don't we just add disk to our existing enterprise solution?*
 - This may or may not be a good idea
 - ♦ NGS capacity needs can easily exceed existing scaling limits on installed systems
 - ♦ Expensive to grow/expand these systems
 - ♦ Potential to overwhelm existing backup solution
 - ♦ NGS pipelines hammering storage can affect other production users and applications

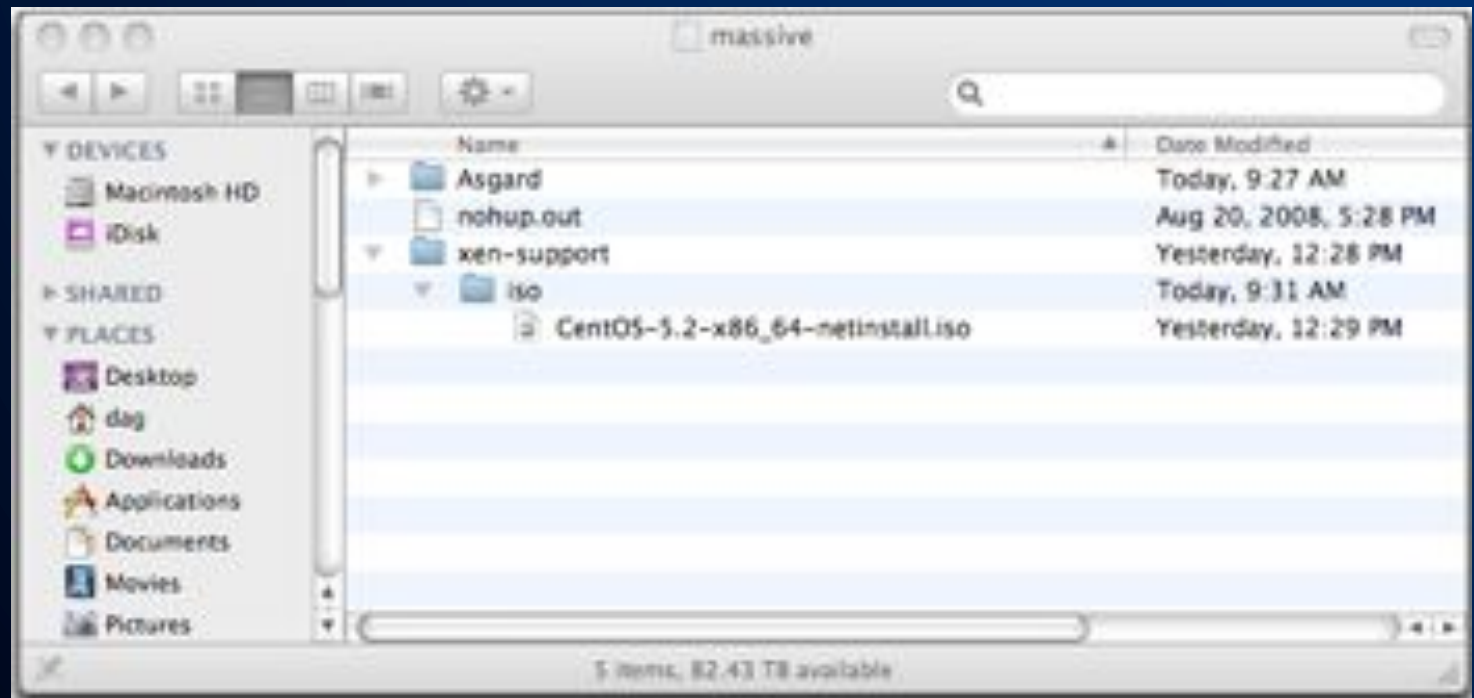
Research vs. Enterprise Storage

- Monolithic central storage is not the answer
- There are valid reasons for distinguishing between enterprise storage and research storage
- Most organizations we see do not attempt to integrate NGS process data into the core enterprise storage platform
 - Separate out by required features and scaling needs

Observations & Trends

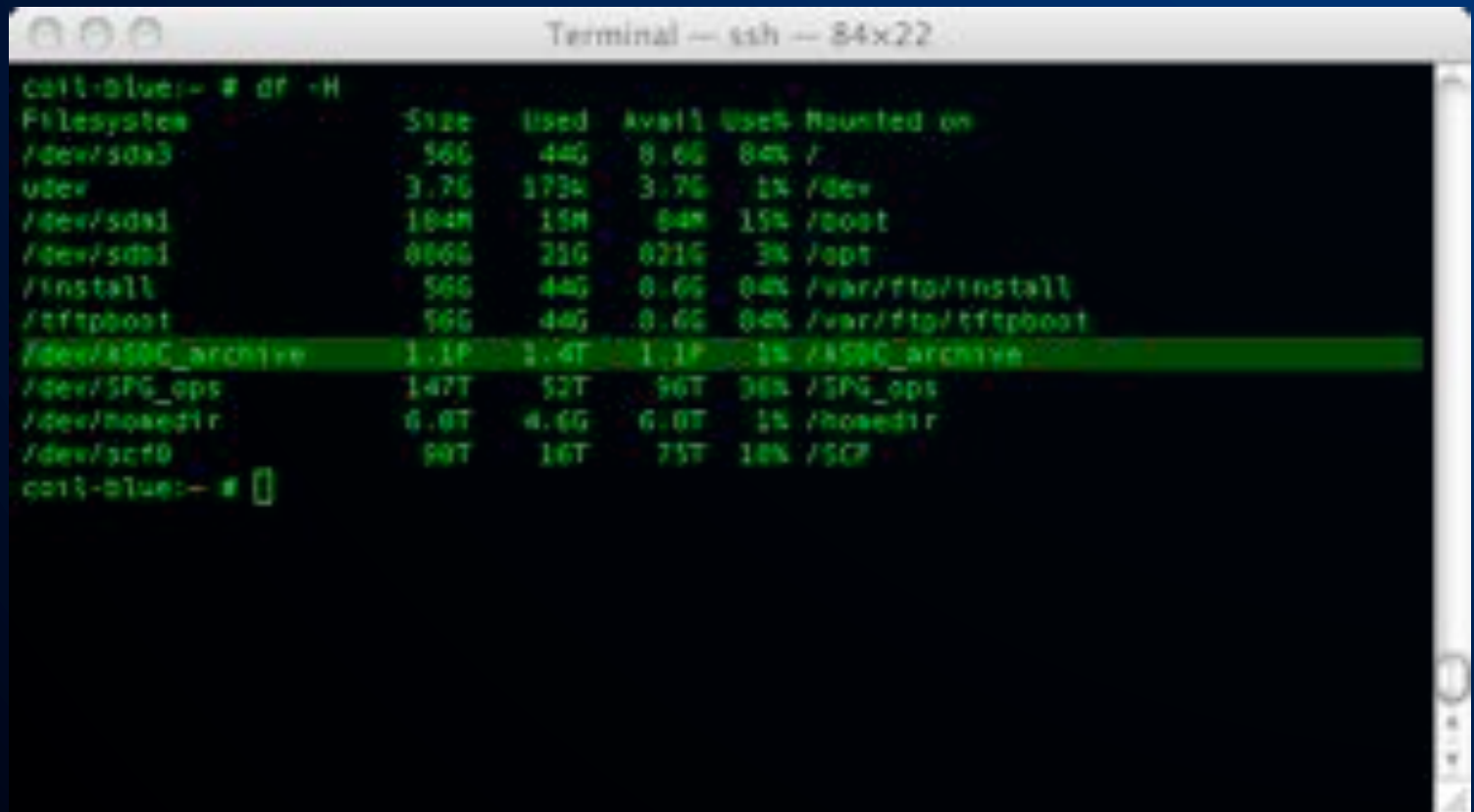
Trends: Single Namespace

- 82TB - Very Satisfying



Trends: Single Namespace

- 1PB - *More Satisfying*



A terminal window titled "Terminal — ssh — 84x22" showing the output of the command `cat@blue:~# df -H`. The output is a table of disk usage for various filesystems. The row for `/dev/ASDC_archive` is highlighted in green, showing 1.1P used and 1.1P available. The other rows show usage in GB, MB, or TB.

Filesystem	Size	Used	Avail	Use%	Mounted on
/dev/sda3	56G	44G	8.6G	84%	/
udev	3.7G	173K	3.7G	1%	/dev
/dev/sda1	164M	15M	84M	15%	/boot
/dev/sdb1	886G	21G	821G	3%	/opt
/install	56G	44G	8.6G	84%	/var/ftp/install
/tftpboot	56G	44G	8.6G	84%	/var/ftp/tftpboot
/dev/ASDC_archive	1.1P	1.1T	1.1P	1%	/ASDC_archive
/dev/SPG_ops	147T	52T	96T	36%	/SPG_ops
/dev/boxedir	6.0T	4.6G	6.0T	1%	/boxedir
/dev/scf0	90T	16T	75T	18%	/SCF

Single Namespace Matters

- Non-scalable storage islands add complexity
- Also add “data drift”

- Example:

- Volume “Caspian” hosted on server “Odin”
- “Odin” replaced by “Thor”
- “Caspian” migrated to “Asgard”
- Relocated to “/massive/”

- Resulted in file paths that look like this:

/massive/Asgard/Caspian/blastdb

/massive/Asgard/old_stuff/Caspian/blastdb

/massive/Asgard/~~can-be-deleted~~/do-not-delete...

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



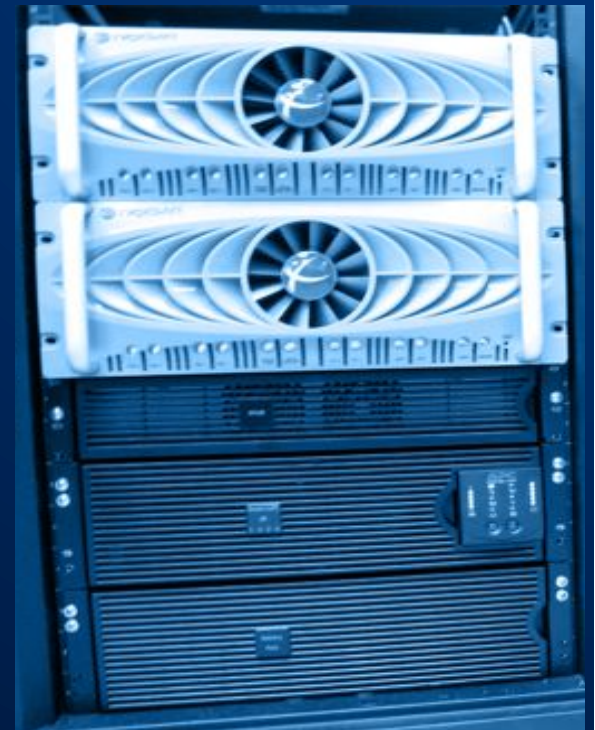
Storage Trends

- In 2009 ...
 - More of the same
 - 100TB not a big deal any more
 - Even smaller organizations are talking (or deploying) petascale storage
- Witnessed spectacular failures of Tier 1 storage vendors:
 - \$6M 1.1PB system currently imploding under a faulty design.
 - \$800K NAS product that can't supply a volume larger than 10TB
 - Even less with dedupe enabled



Going into 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



Going into 2010 ...

- Biggest 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*



Cloud Storage

I'm a believer (maybe)

Why I drank the kool-aid

- I am known to be rude and cynical when talking about over hyped “trends” and lame cooption attempts by marketing folk
 - Wide-area Grid computing is an example from dot com days
 - “Private Clouds” - another example of marketing fluff masking nothing of actual useful value in 2010
- I am also a vocal cheerleader for things that help me solve real customer-facing problems
 - *Cloud storage might actually do this ...*

Cloud Storage: Use Case 1

- Amazon AWS “downloader pays” model is extremely compelling
- Potentially a solution for organizations required to make large datasets available to collaborators or the public
 - Costs of local hosting, management & public bandwidth can be significant resource drain
 - Cloud-resident data sets where the downloader offsets or shares in the distribution cost feels like a good match

Cloud Storage: Use Case 2

- Archive, deep or cold storage pool
- Imagine this scenario:
 - Your 1PB storage resource can't be backed up via traditional methods
 - Replication is the answer
 - However just to be safe you decide you need:
 - ◆ Production system local to campus
 - ◆ Backup copy at Metro-distance colo
 - ◆ Last resort copy at WAN-distance colo
 - Now you have 3PB to manage across three different facilities
 - ◆ *Non trivial human, facility, financial and operational burden costs ...*

Cloud Storage: Use Case 2

- James Hamilton has blogged some interesting figures
 - Site: <http://perspectives.mvdirona.com>
 - Cold storage geographically replicated 4x can be achieved *at scale* for \$.80 GB/year (and falling quickly)
 - With an honest accounting of all your facility, operational and human costs can you *really* approach this figure?

Cloud Storage: Use Case 2

- 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 hypothesis:
 - ♦ 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 ...

Cloud Storage: Use Case 2

- Cheap storage is easy, we all can do this
- Geographically replicated, efficiently managed cheap storage is not very 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

Final Thoughts

- Yes the “data deluge” problem is real
- Many of us have large-scale storage issues today
- “Data Deluge” & “Tsunami” are apt terms
- But ...

Final Thoughts

- But:
 - The problem does not feel as scary as it once did
 - Many groups have successfully deployed diverse types of petascale storage systems - Best practice info is becoming available
 - Chemistry, reagent cost, data movement & human factors are natural bottlenecks
 - Data Triage is an accepted practice, no longer heresy
 - Data-reduction starting to happen within instruments
 - Customers starting to trust instrument vendor software more
 - We see large & small labs dealing successfully with these issues
 - There are now many ways to tackle IT requirements from established vendors in 2010

End;

- Thanks!
- Lots more detail coming in next presentations
- Comments/feedback:
 - chris@bioteam.net