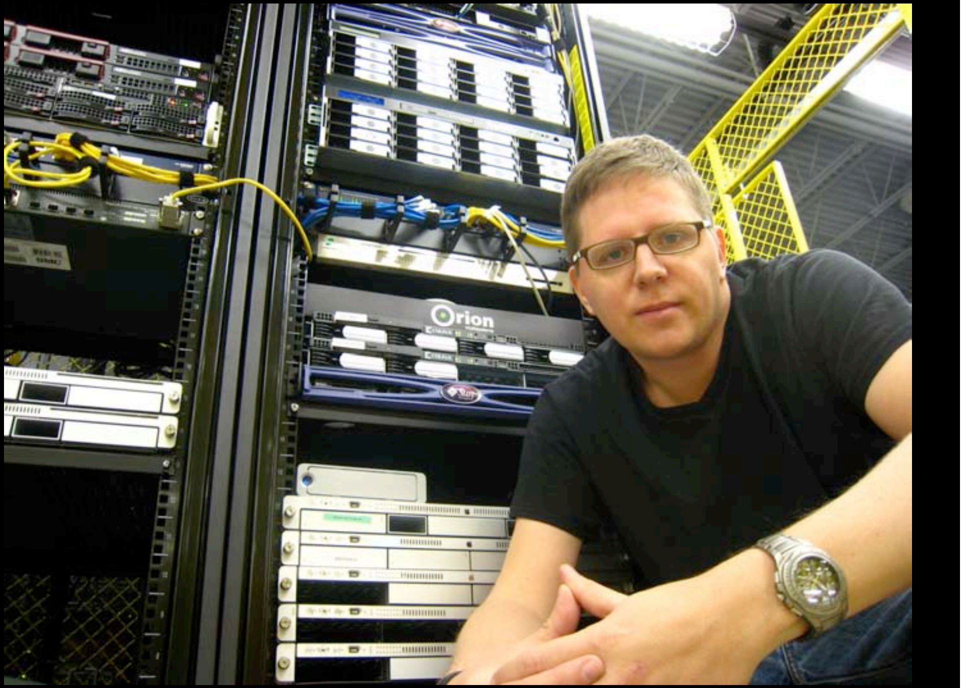
“Trends from the trenches”

Subtitle:
‘40 slides in 15 minutes`

Chris Dagdigian
Biomedical HPC Leadership Summit 2008

Who is this guy?

- I'm Chris
  - 'dag@sonsorol.org' (public)
  - 'chris@bioteam.net' (corporate)
- I work for the BioTeam
- I am:
  - A total infrastructure geek
- Grid Engine Zealot
  - http://gridengine.info
- Long OSS involvement
  - http://bioperl.org
  - http://xml-qstat.org



BioTeam Inc.

- Independent Consulting Shop
  - http://www.bioteam.net
  - Vendor/technology agnostic
- Staffed by
  - Scientists forced to learn HPC IT
  - Many years of industry & academic experience
- Our specialty
  - Bridging the gap between Science & IT

Why this talk?

- BioTeam
  - Often a resource for labs and workgroups that don't have their own supercomputing centers and IT empires
  - All types of clients (Gov, EDU, Biotech, Pharma, Fortune-20, Vendors, etc.)
- Thus:
  - We are in a good position to see where IT meets science in "real world" production settings

Beware.

- I’m known for talking fast and carrying a large slide deck
- Goal for this talk:
  - 15 minutes!
- Unrepentant powerpoint fiddler
- Updated presentations & data:
  - http://blog.bioteam.net

Additional disclaimer

- Content of this talk may be inappropriate for this particular audience
- Most BioTeam clients *don’t* have 7 figure IT budgets, petabyte SANs and dedicated datacenters
- Will discuss problems that simply don’t exist for the largest Bio-HPC centers

Hardware & Networking

Observed Trends: Hardware

- CPU wars
  - 2008 - No change since ‘07
    - We still benchmark with real apps & data when possible
- Small Cluster Market
  - 2007: “… this market is going away”
  - 2008: Pretty much dead
    - Multi-core chassis rule the land
    - We’ll see how MS HPC Server does …

Observed Trends: Hardware

- Cool in 2007: Cooling/HVAC
- Cool in 2008: Green power features
  - Favorite ‘08 example
    - AutoMAID on NexSan SATABeast

Observed Trends: Networking

- 10 Gigabit is now mainstream
- In 2007
  - Connect storage to networks
  - Connect switches together
- In 2008 …
  - Not a lot of change; still mostly:
    - Storage to network & switch to switch
- 2009 Prediction
  - Server-to-server 10 Gigabit will mainstream

Observed Trends: Networking

- Infiniband is now mainstream in 2008
  - Fun to watch the price curve trends
  - ‘08 Uptick: Customers purchasing it inappropriately
    - Blame: Unethical or clueless resellers/vendors
  - Still low adoption rate in Bio HPC

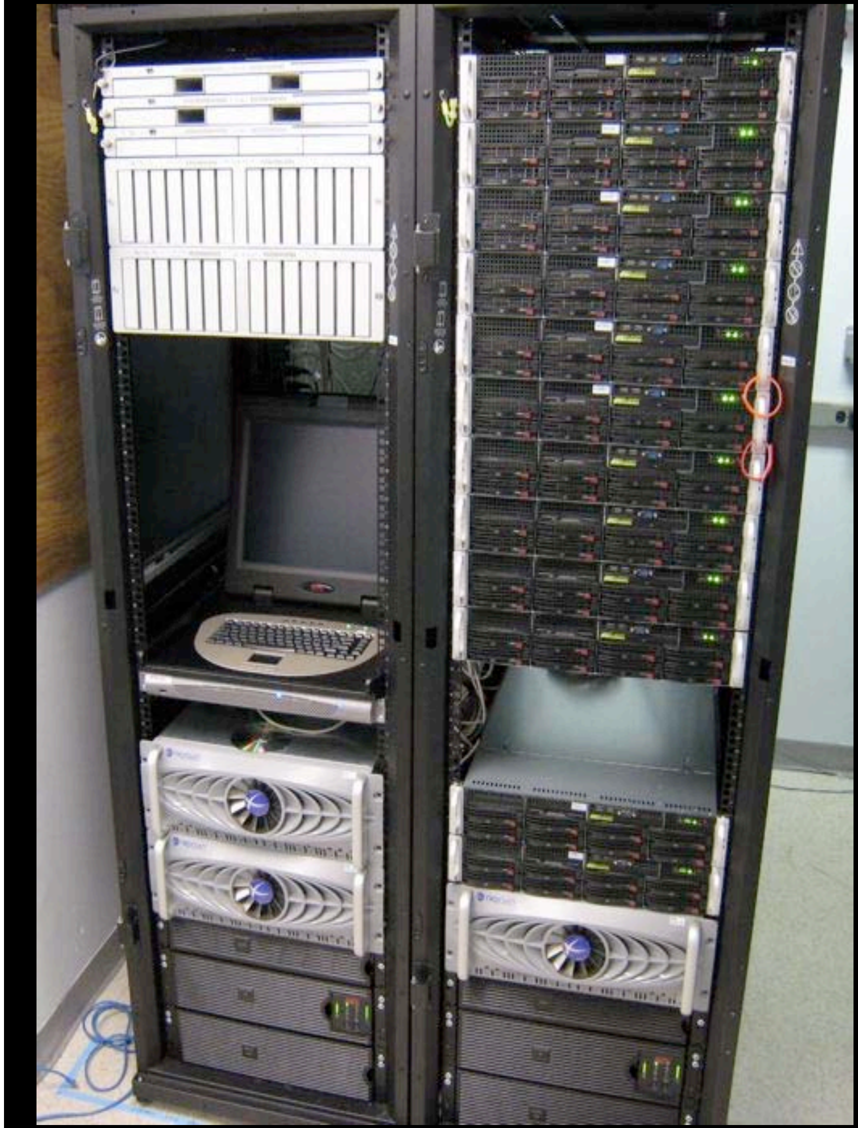
- Still used more for parallel/cluster storage than for application (MPI) traffic …

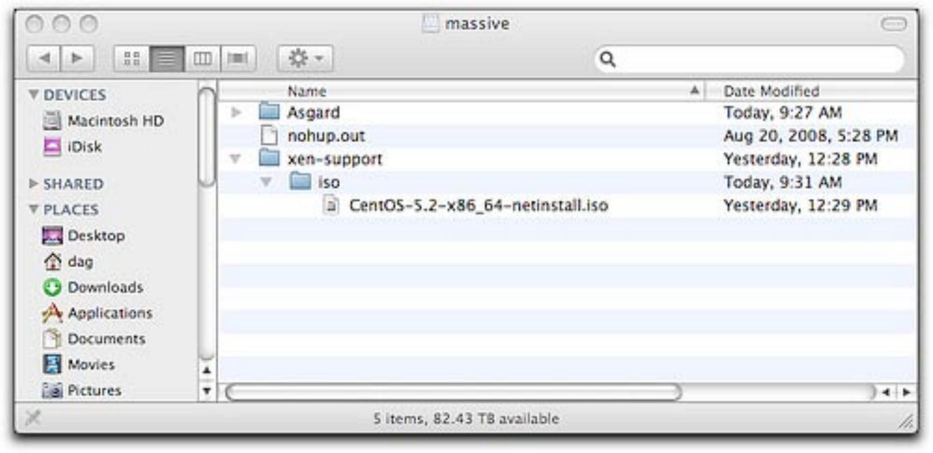
Storage

Storage Trends 2008

- What BioTeam saw in 2008
  - First 100TB single-namespace project
  - First Petabyte-scale 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

100TB for BioDefense





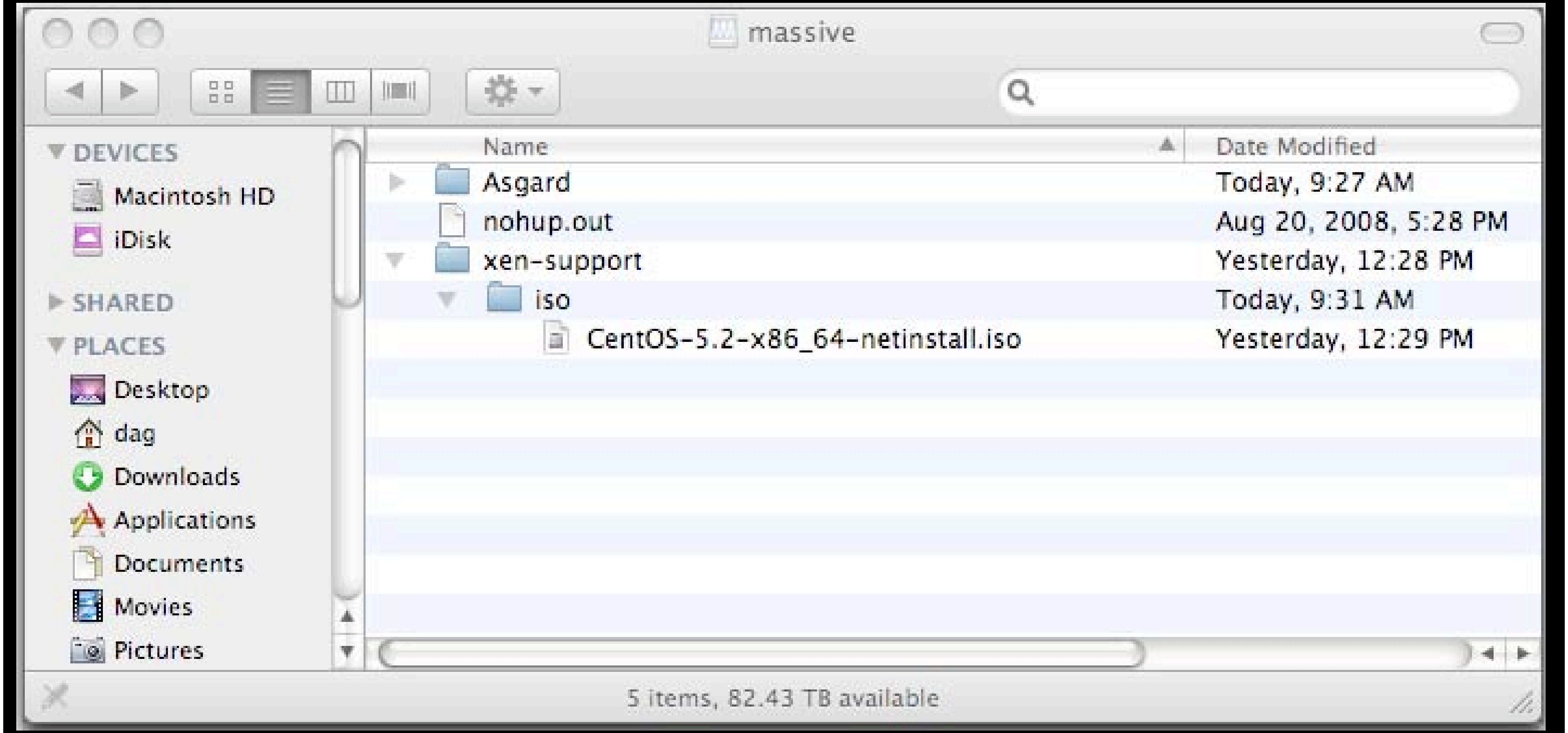
massive

Name	Date Modified
Asgard	Today, 9:27 AM
nohup.out	Aug 20, 2008, 5:28 PM
xen-support	Yesterday, 12:28 PM
iso	Today, 9:31 AM
CentOS-5.2-x86_64-netinstall.iso	Yesterday, 12:29 PM

5 items, 82.43 TB available

- Picture taken 9/19/08
- 100 TB Raw / 87 TB Usable
- Single namespace ("/massive")
  - Commodity SATA disk
  - Fiber attached
  - RHEL + Cluster GFS

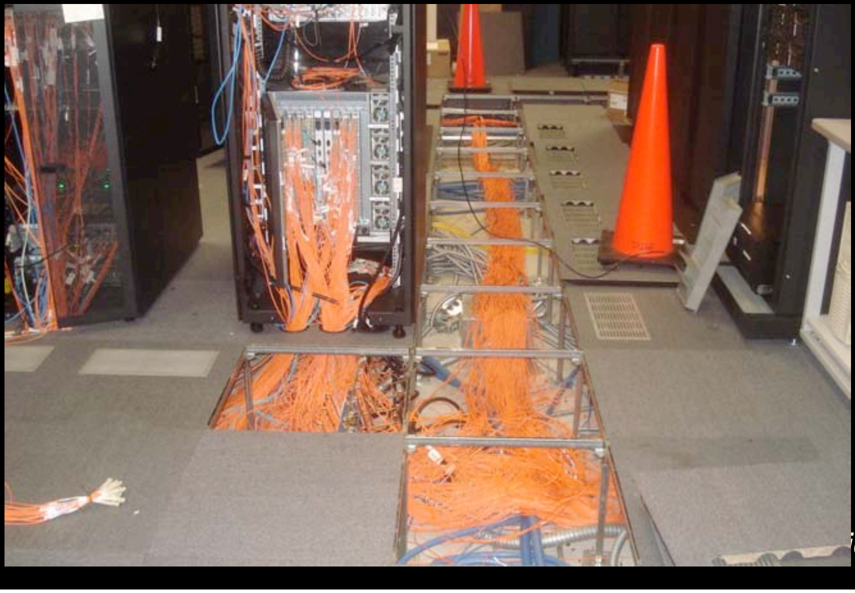
82TB Folder. Very satisfying.



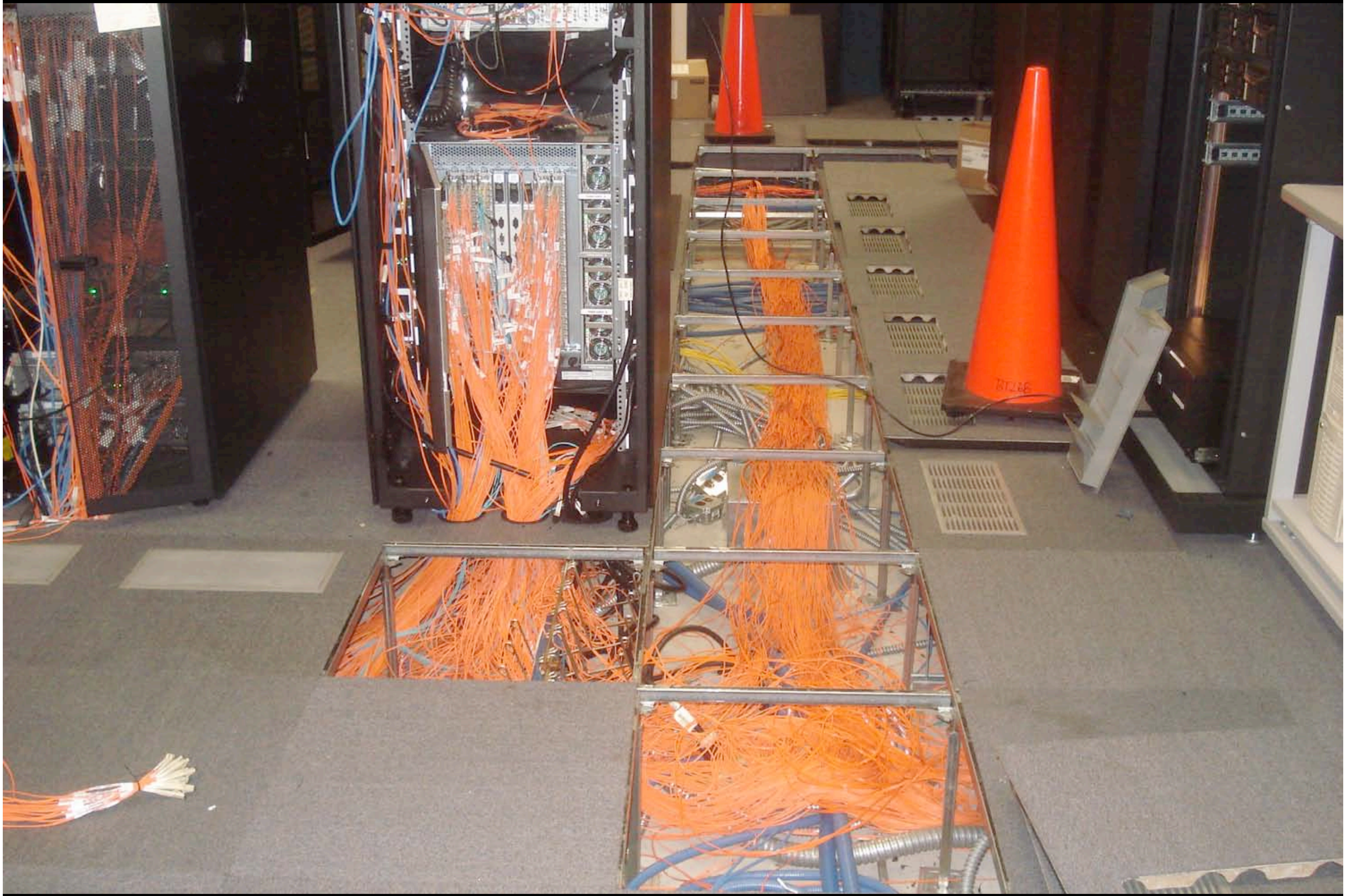
Name	Date Modified
Asgard	Today, 9:27 AM
nohup.out	Aug 20, 2008, 5:28 PM
xen-support	Yesterday, 12:28 PM
iso	Today, 9:31 AM
CentOS-5.2-x86_64-netinstall.iso	Yesterday, 12:29 PM

5 items, 82.43 TB available

Petabyte$^+$ for Science



- Picture taken 9/2/08
- 1.2PB usable / 1.8PB raw
- Fibre connected
  - 384+ fibre ports
- 2,560 individual disk drives
  - 16 disks per chassis
  - 10 chassis per rack
  - 16 racks of disks
- IBM Linux servers, mixed P6 and x86 CPUs to support legacy codes
- Filesystem: IBM GPFS



General Observations

- Storage is “cheap” & getting cheaper
- Operational costs seem to be remaining the same
- Backup & data continuity costs are exploding
    - I am personally in awe of the backup experts who are still staying afloat in the age of 1TB SATA disks …

Observations cont.

- End users 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)*
- IT needs to be involved in setting expectations and educating on true cost of keeping data online & accessible
- Organizations need forward looking research storage roadmaps

Observations cont.

- The rise of “terabyte instruments” is already having a major disruptive influence on existing environments
  - We see individual labs deploying 100TB+ systems
  - If a lab needs 100TB, what does your organization need?
- I was wrong when I said
  - *“petabyte scale storage needs will appear within the decade …”*
  - That time is now for some large organizations

Capacity Dilemma: Data Triage

- Data Triage
  - The days of unlimited storage for research are likely over
  - Rate of consumption increasing unsustainably
  - First saw triage acts in 2007 (industry client)
  - Becoming acceptable practice in 2008
- Why delete data?
  - Given full lifecycle cost of data, sometimes repeating the experiment is cheaper than storing the results forever
- Triage Example (Solexa data)
  - Image data kept for ~7 days as QC/QA measure; then deleted
  - Derived data (bead intensity reads) kept forever

Capacity Dilemma: Data Triage

- This is going to be a huge problem
  - Storing the data is not hard; protecting it and keeping it ‘available’ is the seemingly unsustainable part
  - Researchers generally come on board once they understand the true cost of keeping data available
- Data Triage seems unavoidable
  - … Unless disruptive technology for nearline, archive or HSM storage appears
- IT Organizations are unqualified to make final triage calls
  - Active participation by research staff is essential

Data Loss: Lessons learned

- The event
  - Metadata for clustered SAN file system irrevocably corrupted
  - 10+ TB of scientific data lost forever
  - 3 people fired and counting …
- Simplified root cause:
  - RAID controller hits bad block; pulls data from parity, DOES NOT write back down to a new block
  - Operators may have ignored disk/chassis warnings
  - Eventually: double disk failure on metadata LUN
  - Rebuild onto spare disk fails due to missing inode data that had been stored on bad blocks but not rewritten elsewhere. Remaining parity data not sufficient for 100% rebuild

Data Loss: Lessons learned

- Observing this event taught us some lessons:
  - We no longer use RAID5 on large filesystems
  - Everything on RAID6 or other double-parity system
  - Mandatory use of SNMP & email reporting
  - Disk handling: Replace drive on warning
  - We reject storage/controller products that are not proactive about disk scrubbing and consistency checking

Observed Trends: Backup

- My IT nightmare every year for the last decade
- 2007
  - Backup products not keeping up with daily advances in storage capacity promoted by vendors
- 2008
  - Became something of a sick joke
  - Storage products leave backup products in the dust
    - Almost too far ahead to even attempt to keep up
- 2009 Conversations / Potential trend
  - Complete re-think of backup paradigm
  - New expectations, new procedures in an age where "nightly full" will never happen again

The most terrifying trend …

What should be keeping you up at night

{ Unchanged since 2007! }

Terrifying trend: Terabyte Instruments

- 2007 was the tipping point
- We now have individual researchers with individual instruments that can:
  - *… generate terabyte scale data streams in a single experiment*
- Previously:
  - Terabyte data problems were at the workgroup, lab or organizational level

Terrifying: Terabyte Instruments

- The problem in a nutshell:
  - Individual researchers and/or single instruments are now capable of generating terabyte scale data *in a single experiment*.
    - Examples:
      - Confocal microscopy & Next generation DNA sequencers
  - These instruments are “cheap”
    - Easily affordable by grant-funded individuals and small labs
  - And …
    - Researchers don’t buy “just one” of these machines
    - Researchers may want to run them 24/7

Terrifying: Terabyte Instruments

- Why this is such a big deal
  - This is a nightmare even for the “big” centers with dedicated datacenters, large SANs and very competent IT staff
  - Imagine the effect on small organizations
    - The infrastructure and staff to support terabyte scale experimentation simply does not exist
  - Also
    - Researchers may be budgeting for the instrument and reagents but not the IT/operation requirements
    - Instrument vendors may be (intentionally or otherwise) downplaying the true infrastructure and operational costs of these instruments

Terrifying: Terabyte Instruments

- Is this your future?
  - Multi-terabyte storage resources in every wet lab?
  - *Sun Thumpers for all!*
- Tough decisions ahead
  - Centralized vs. decentralized data capture & movement
- This will effect *everyone* doing HPC "Bio IT"

Amazon EC2

In 2009 I vow never to use the word “cloud” in any serious technical conversation …

Cl**d Computing

- Amazon EC2 ***is*** the cl**d
  - Everyone else is playing catch-up
  - … or fooling themselves
- Remember:
  - I am known somewhat as an “anti-Grid” crank
  - Because:
    - The few successful multi-site “GRIDs” are operated by Fortune-10 firms or labs backed by sovereign funding
      - … and others with 7 or 8 figure IT budgets
      - Everything else is just empty hype and unmet expectations

Cl**d Computing

- Why I drank the EC2 Cool-Aid
  - Saw it, used it, solved actual customer problems with it
- BioTeam & Amazon EC2
  - Late 2007
    - Initial experimentation & test cases
  - Early 2008
    - By March, every single BioTeam consultant had independently used EC2 to solve a customer facing problem
  - Late 2008
    - Commercial and OSS application EC2 integration requests are coming in almost weekly
  - 2009 Prediction
    - Industry will pressure more and more ISVs to support EC2 model
    - Industry may form consortium to centralize development, porting and "best practices"

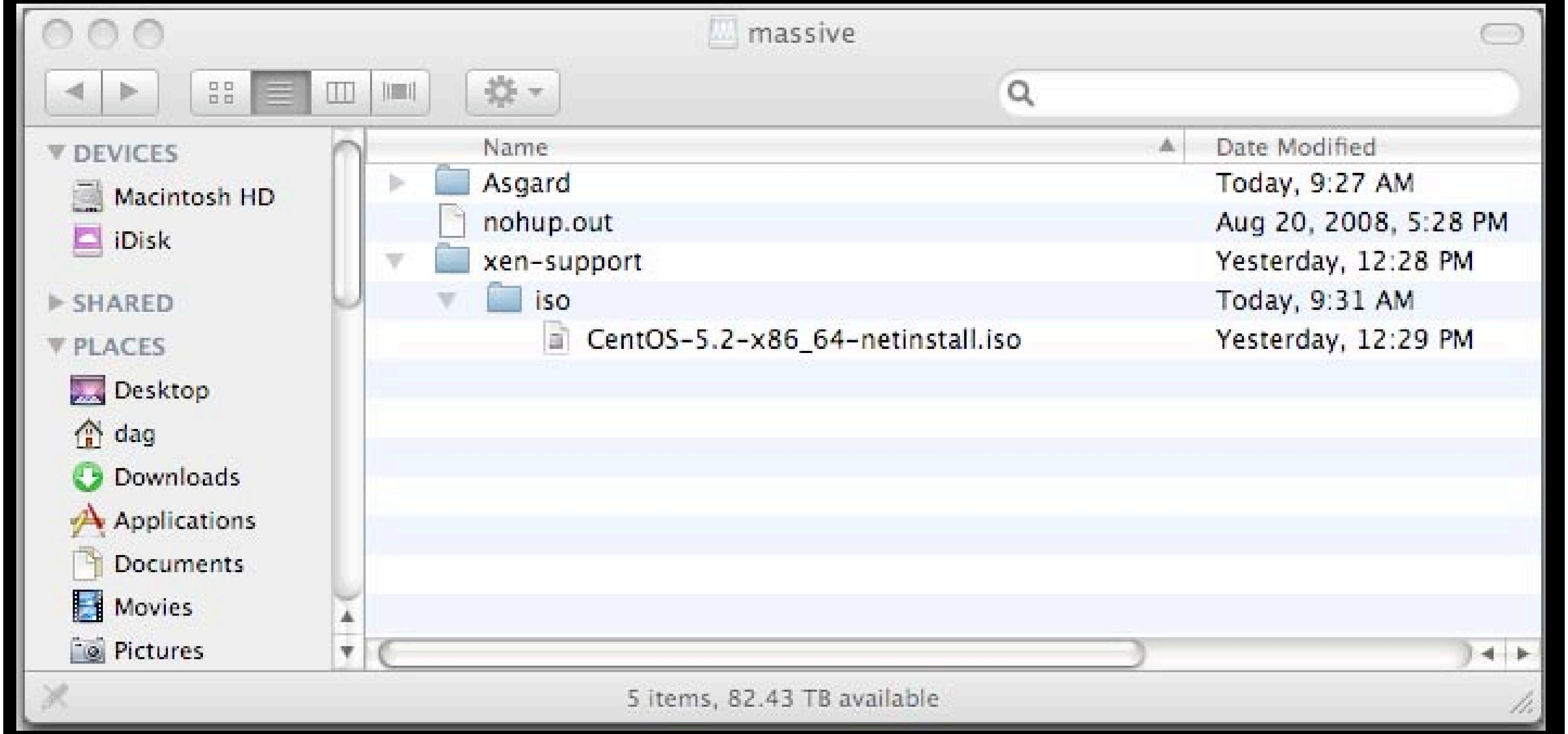
Cl**d Computing

- Some Additional Amazon EC2 Resources
  - http://blog.bioteam.net
  - We have written up two recent projects:
    - "Deploying Univa UniCluster into EC2"
    - "Grid Engine + Service Domain Management in EC2"
  - … more on the way

Coolest things I’ve seen…

2008 edition

Cant get enough of this …



Name	Date Modified
Asgard	Today, 9:27 AM
nohup.out	Aug 20, 2008, 5:28 PM
xen-support	Yesterday, 12:28 PM
iso	Today, 9:31 AM
CentOS-5.2-x86_64-netinstall.iso	Yesterday, 12:29 PM

5 items, 82.43 TB available

Coolest in 2008 ….

- Canadian bioscience firm
    - Email me if you want their name
        - (I need to ask permission first)
    - Core workflow
        - Large scale genome-wide association studies
- What I witnessed:
    - The first real, usable and practical example of multi-site "Grid" computing done by mere mortals that I have ever seen

Coolest in 2008 ….

- Workflow
  - Data & CPU intensive pipelines
    - 80+ step workflows with complex interdependencies are not unusual
- Cluster(s):
  - Small Linux cluster onsite
  - Larger Linux cluster @ metro colo facility
  - As-needed contract with an IBM Deep Computing facility thousands of miles away

Coolest in 2008 ….

- Three clusters, three DRM software layers:
  - Local Linux cluster: Platform LSF
  - Metro-scale colo cluster: Sun Grid Engine
  - IBM facility: SLURMM or LoadLeveler **

Coolest in 2008 ….

- All three systems seamlessly integrated via Platform Process Manager*
  - Registered workflows are packaged together with required data
    - Commercial compression applied
  - Platform PM handles distribution, execution and guaranteed task completion

** A human must decide to activate the IBM facility*

Coolest in 2008 ….

- Why this matters
  - Company run by mere mortals
  - Same resources, staff, budget as you and I
  - And yet …
    - *Three clusters spanning LAN, Metro and WAN scale distances with diverse DRM layers all being productively used to Do Science*
    - *It. Just. Works.*

End;

- Thanks!

- Presentation slides will appear here:
  - http://blog.bioteam.net

- Comments/feedback:
  - “chris@bioteam.net”