Bioinformatics: Adventures and Observations

Chris Dwan cdwan@bioteam.net
Bioteam (http://bioteam.net)

St. Thomas University
September 15, 2010

Agenda

- About your speaker
- DNA sequencing
- High performance computing for science
- Technologies to watch
- Career thoughts for the CS crowd

In the mind of the beginner, there are many possibilities.
In the mind of the expert, there are few.

Your Speaker

University of Michigan, Ann Arbor

- 1996 BS, Computer Science
- 2000 MS, Computer Science / Artificial Intelligence

1996 – 2000: ERIM International / Veridian / General Dynamics

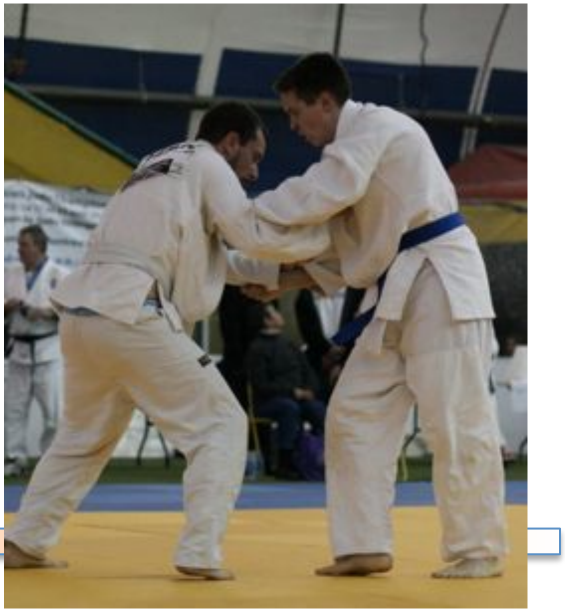
- Machine learning: Teaching missiles to attack tanks, rather than trucks. Also, locating unexploded ordinance.

2000 – 2004: Center for Computational Genomics and Bioinformatics

- University of Minnesota service / research group
- Built my first cluster, took my first international consulting engagement
- Free tuition = graduate coursework in Biology.

2004 - present: Bioteam

- Employee #1 (of four). Company is now up to nine people.
- Many hats. Currently direct all consulting and professional services



Bioteam Inc.

Independent Consulting Shop

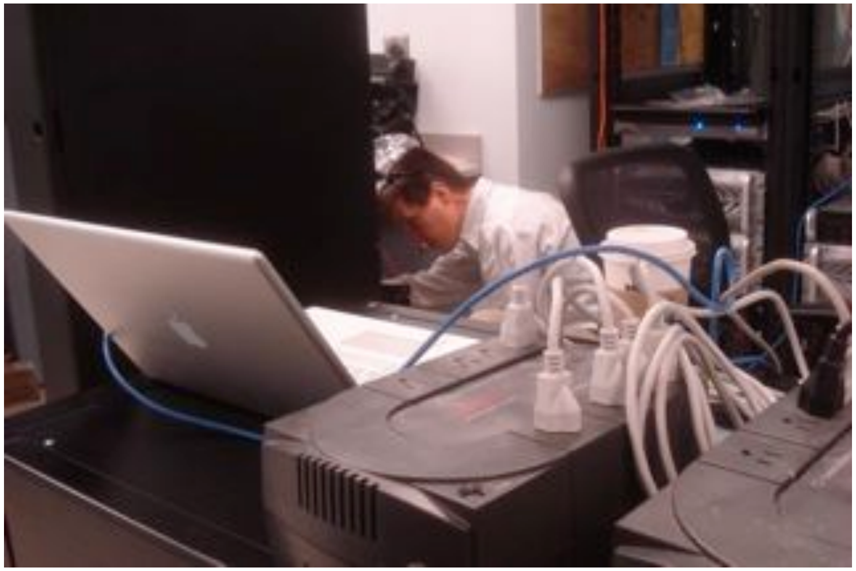
- Started in 2002
- Vendor/technology agnostic
- No financial incentives for any technology we recommend.

Staffed by:

- Scientists forced to learn High Performance IT
- Many years of industry & academic experience

Our specialty:

- Bridging the gap between science & IT



Bioteam Offerings

Consulting

- Technical assessment
- Custom software development / parallelization / tuning
- Computing architecture / design / purchase / build-out / support
- Contract bioinformatics analysis

Software

- Inquiry
- WikiLIMS
- MiniLIMS

High points from the last decade

Many high points …

- Round the world in 11 days
- Visits to Plum Island and other scary research facilities
- Name brand customer list: MIT, NASA, Harvard, Yale, Center for Disease Control and Prevention, …
- Acting as a peer and advisor to PhDs, MDs, National Academy of Science members …
- Chance to work with emerging technology

There are down sides …

- Emerging technology doesn't always work
- 700+ emails in my INBOX on day one.
- 90+ days on the road, year to date 2010
- Re-inventing major pieces of what one might learn in business school

Why Put The Biology First?

"Bioinformatics is full of pitfalls for those who look for patterns or make predictions without a thorough understanding of where biological data comes from and what it means"

Nevin Young PhD

Professor, UMN

What is bioinformatics?

My professor of Genetics at UMN: "We already have a term for the application of mathematical and computational models to the mechanisms of genetic inheritance: That term is 'Genetics.'"

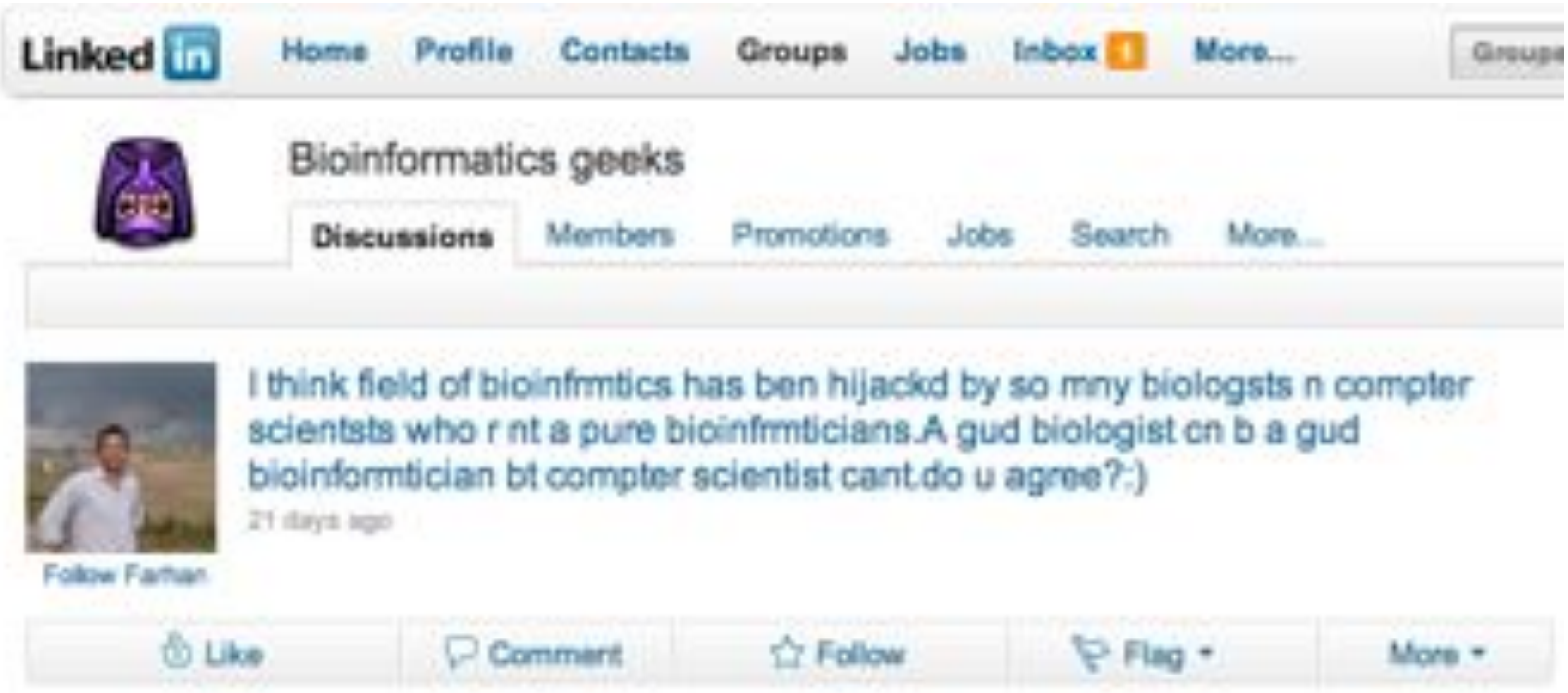
Unknown: "Bioinformatics is a truly wondrous field in which highly trained and skilled individuals from an incredible variety of backgrounds meet to treat each other with mutual scorn and disdain."

My opinion:

- In the 1950's we saw "computational physics." Now we see "computational biology."
- In 2000, it was rare for a biology undergrad to take even one programming course. In 2010, it is commonplace.
- "Bioinformatics" is a midpoint on the path between "biology" and "biology."

More Succinctly: Bioinformatics is biology

Just don’t be this guy



I think field of bioinfrmtics has ben hijackd by so mny biologsts n compter scientsts who r nt a pure bioinfrmticians.A gud biologist cn b a gud bioinformtician bt compter scientist cant.do u agree?:)

21 days ago

More seriously, communication is hard

DNA Sequencing

DNA Sequencing Technology

- 1869: DNA discovered
- 1937: Regular crystal structure ("A stupid tetranucleotide")
- 1952: Identified as genetic material
- 1953: Double helix structure discovered.

  "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material"
- 1957: Central dogma (DNA – RNA – Protein)
- 1975: "Sanger" sequencing (chain termination)
- 2003: 2nd generation: High throughput technologies
- 2010: 3rd Generation: Single molecule technologies

Sanger method DNA sequencing

- Purify DNA, Prime, Trim, and amplify region to be sequenced
- Shear / Generate strings of all possible lengths [1 .. 1000] anchored to a known starting point
- Stick identifying labels on the terminal residue
- Sort by weight in a gel or capillary
- Fundamentally limited to ~1,000 base pairs at a read
- Requires substantial work to prepare the correct subsequence

The ABI 3730 is a workhorse of DNA sequencing.

2nd Generation DNA Sequencing

Two concurrent changes:

- 10^3 fold increase in sequencing speed
- 10^3 fold decrease in per base pair cost to sequence.
- $5k - $7k for an instrument run that yields billions of base pairs
- Substantially less pre-instrument prep work per base pair.

Three major vendors, each with a slightly different technology

- 454 / Roche (GS-20, Titanium):
  - ~500bp reads, ~8 hour run time, ~40GB primary data per run
- Illumina (GA, GA2, HiSeq), ABI SOLiD (v1 – v4)
  - “short” reads (~20bp), 3 – 7 day run times, 1 – 2 TB primary data per run

The Data Deluge (2009 – already dated)

- http://www.politigenomics.com/next-generation-sequencing-informatics

Instrument	Raw Images	Sequence	Run Time	Data Rate
454 FLX	0.01 Tb	1 Gb	8 hours	10 Tb / year
454 Ti	0.04 Tb	4 Gb	6 hours	36 Tb / year
Solexa GA	0.5 Tb	1 Gb	3 days	58 Tb / year
Solexa GA 2	1.1 Tb	7.5 Gb	3 days	135 Tb / year
SOLiD 1	1.8 Tb	1 Gb	6 days	109 Tb / year
SOLiD 2	2.5 Tb	4 Gb	5 days	182 Tb / year

4/27/09:

- 80% of data generated at the Broad Institute is from next generation sequencing machines.
- 100 – 200TB per month growth.

cdwan@bioteam.net

454 Instrument

Illumina and SOLiD Instruments

Human Genome Project

- 1990: Human Genome Project kickoff
  - ~$3x10^9$ unique base pair locations (two instances per individual)
  - BAC libraries ($1.5x10^5$ bp at a time) distributed worldwide
  - ~$3x10^9$ project cost (initial estimates $3/bp, adjusted to $1/bp)
- 1998: Private effort kickoff (Celera Genomics):
  - Shotgun Sequencing
  - $3x10^8$ ($0.1/bp): New technology and data from the public effort.
- 2000: “rough draft” completed
- 2001: Draft Human Genome Published by public and private efforts
- 2003: “complete sequence” published
- 2006: “last chromosome” “complete sequence” published
  - Francis Collins is now director of the NIH
  - Craig Venter is now constructing self replicating organisms de-novo
- 2010: Clinical / research sequencing at $10k - $40k per individual.

Workflow in 2nd generation sequencing

Preliminary data acquisition (chemistry and image capture)

"Base calling": Generally done **once per sample**. Reduces data by ~10x

- Open question: Must we retain primary image data? Consensus is "no," except for regulatory concerns.

"Assembly": Combine "reads" (potentially from many instrument runs), into "contigs"

Downstream analysis

- Homology search
- Annotation
- … course reading …

Genome Sequencing, September 2010

Demonstration exercise for a pathogen research group

- "Unknown" pathogen sample delivered to lab
- Sample preparation: ~2 hours
- 454 Instrument run: ~8 hours
- Preliminary data analysis (base calling): ~5 hours
- Genome assembly: ~4 hours
- Downstream analysis, characterization, report writing, etc: ~12 hours

~31 hours from sample to near complete genome sequence plus detailed analysis report, in a small lab.

3^{rd} generation DNA sequencing: Q4 2010

Single molecule technologies

- Oxford Nanopore
- Pacific Biosystems (Pac-Bio)
- Ion Torrent
- Helicos

Another several orders of magnitude in both speed and price per base pair. Current estimates are “a whole human genome in 3 hours.”

In production “real soon now.” High error rates seem to be a problem.

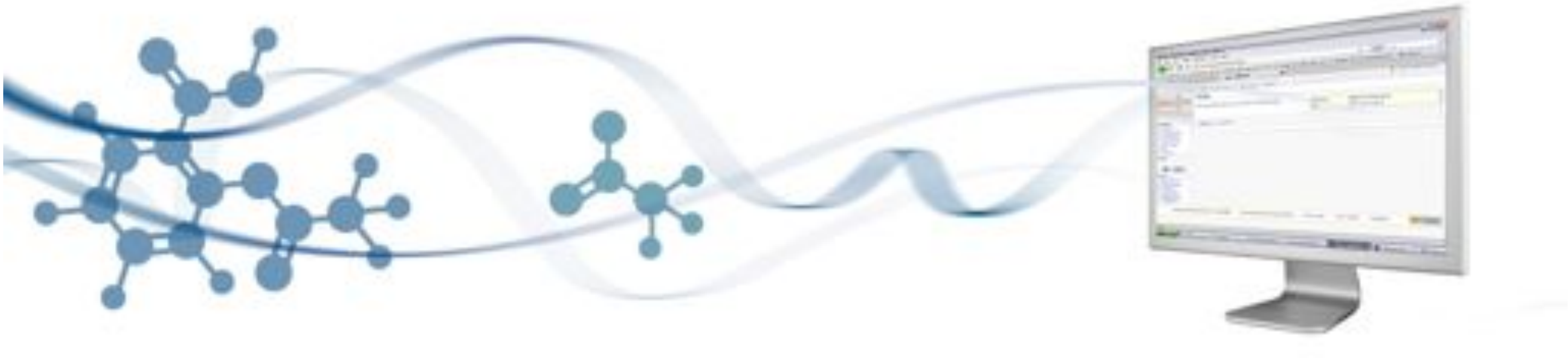
Please help me kill the term “next-next generation.” Use “single molecule” or “3rd generation” instead.

Beyond DNA

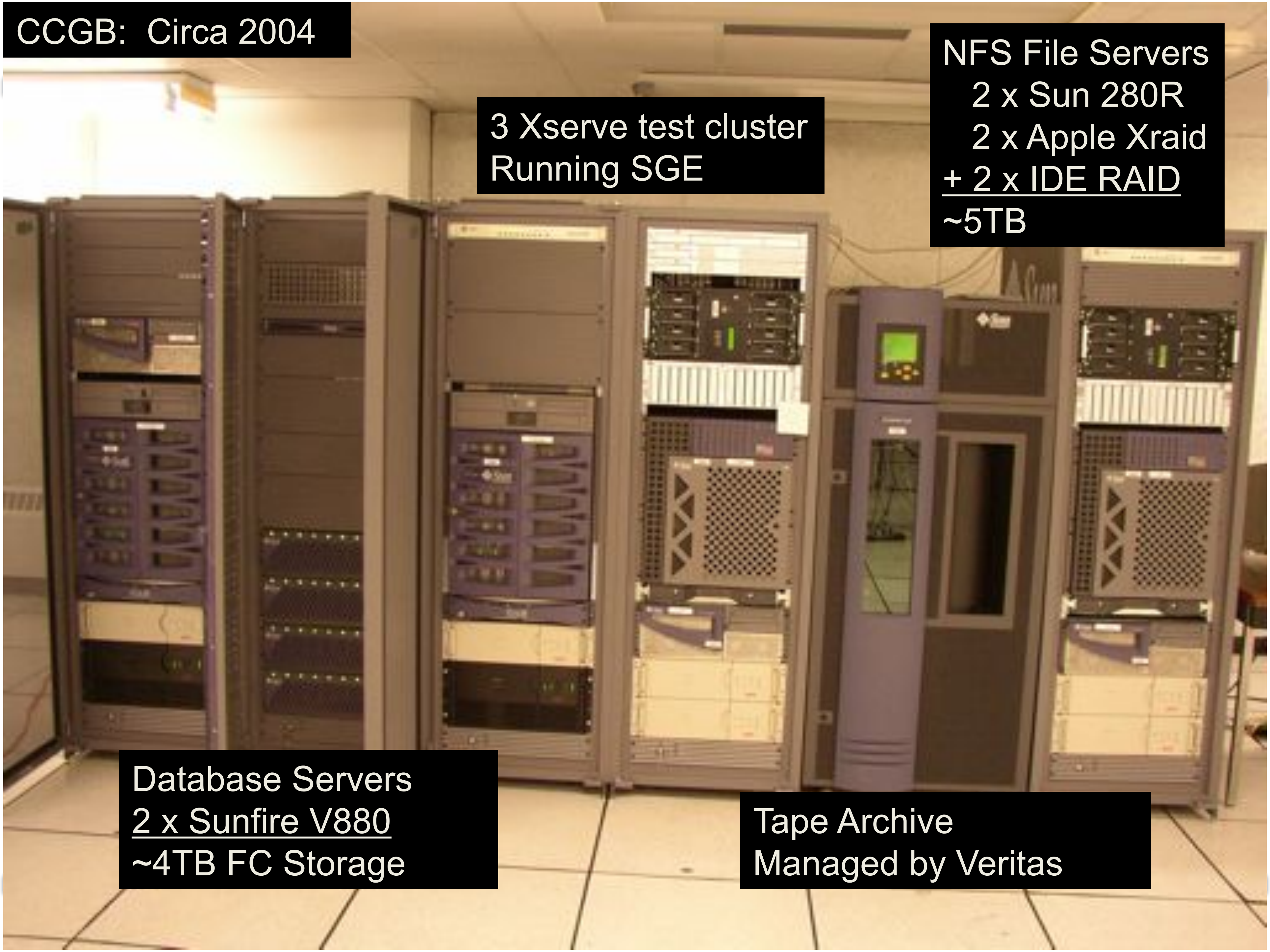
- …

Mysterious little instrument …

High Performance Computing



CCGB: Circa 2004

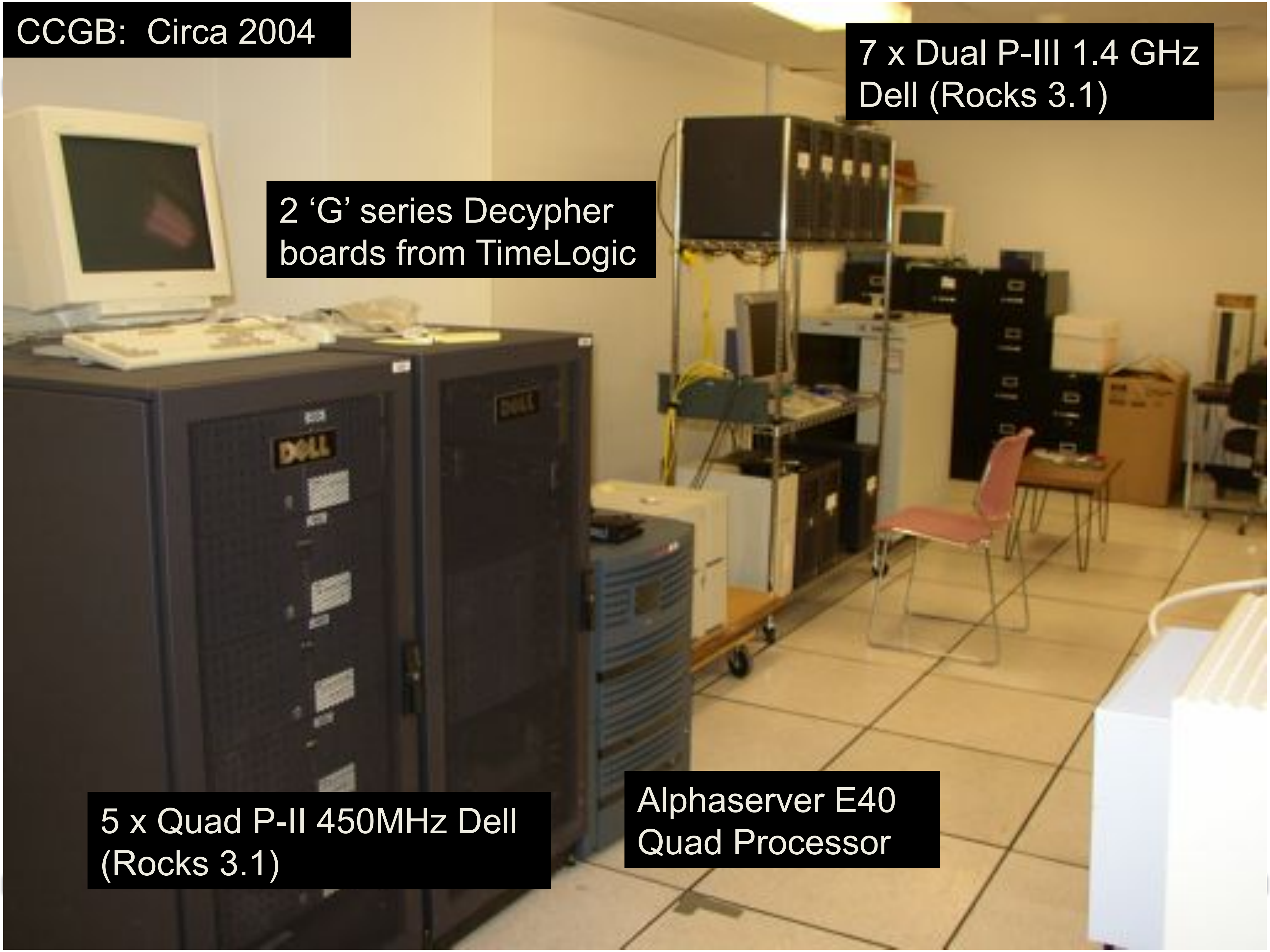


3 Xserve test cluster
Running SGE

NFS File Servers
2 x Sun 280R
2 x Apple Xraid
+ 2 x IDE RAID
~5TB

Database Servers
2 x Sunfire V880
~4TB FC Storage

Tape Archive
Managed by Veritas



CCGB: Circa 2004

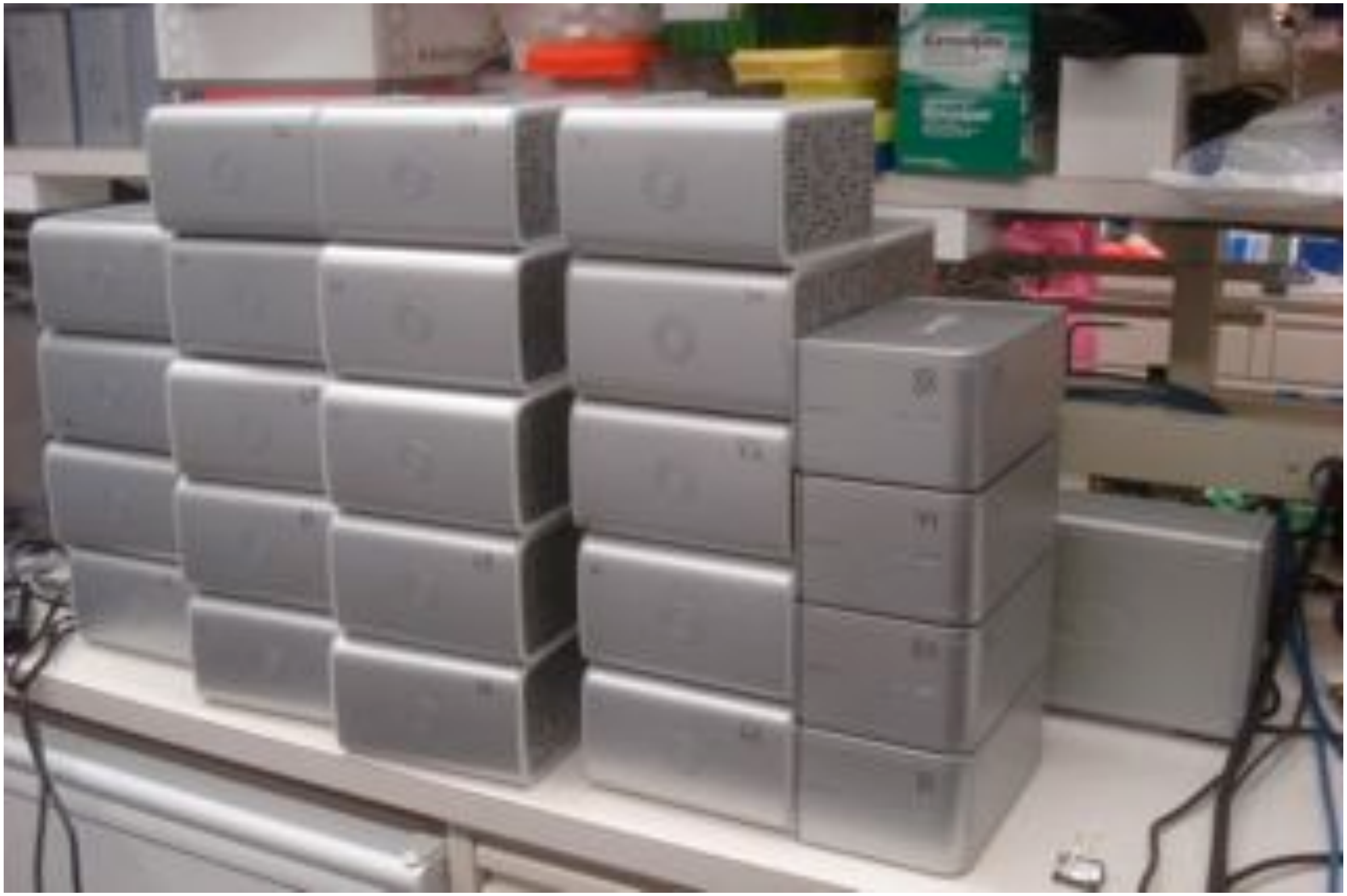
2 ‘G’ series Decypher boards from TimeLogic

7 x Dual P-III 1.4 GHz Dell (Rocks 3.1)

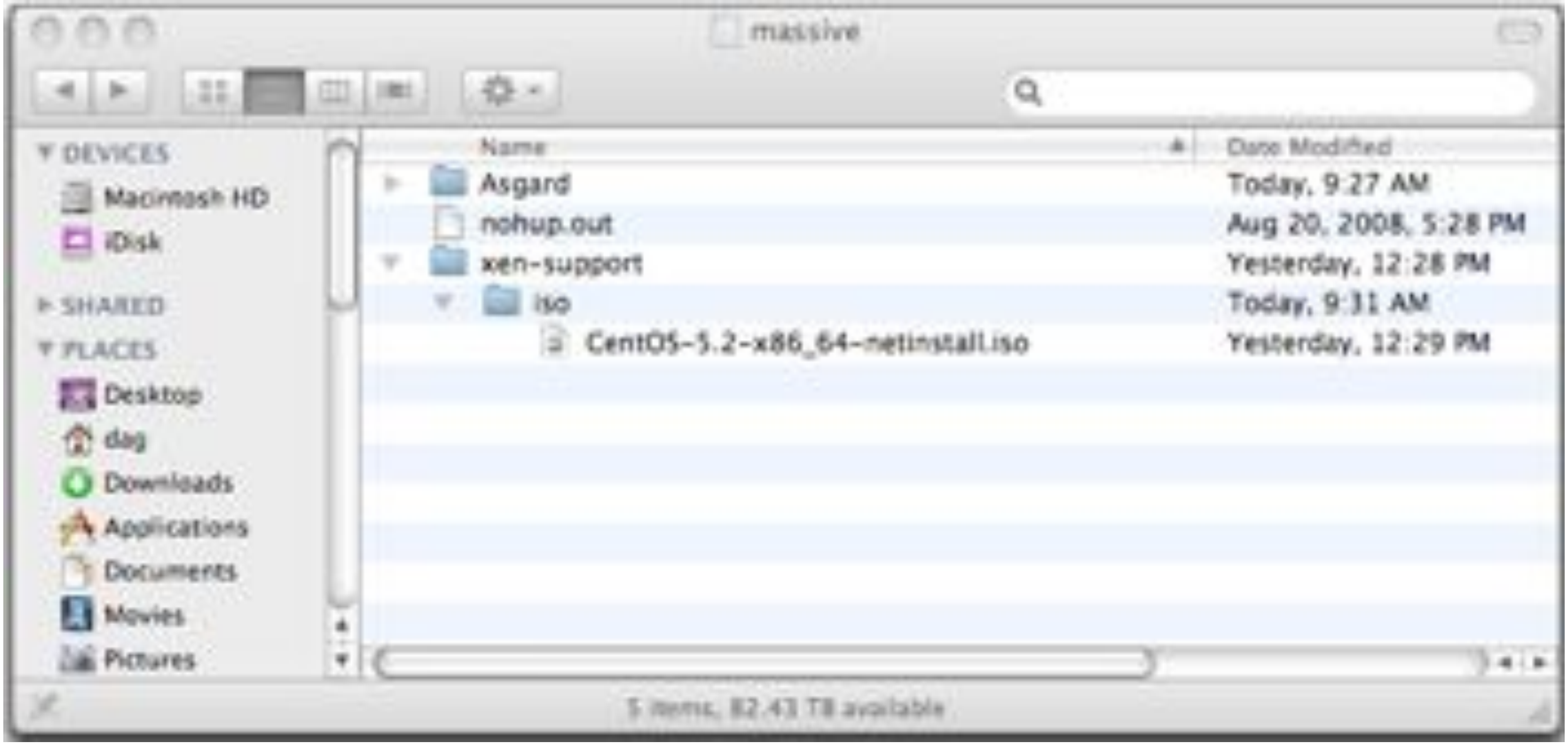
5 x Quad P-II 450MHz Dell (Rocks 3.1)

Alphaserver E40 Quad Processor

This is not data storage



82 TB Folder. Very satisfying

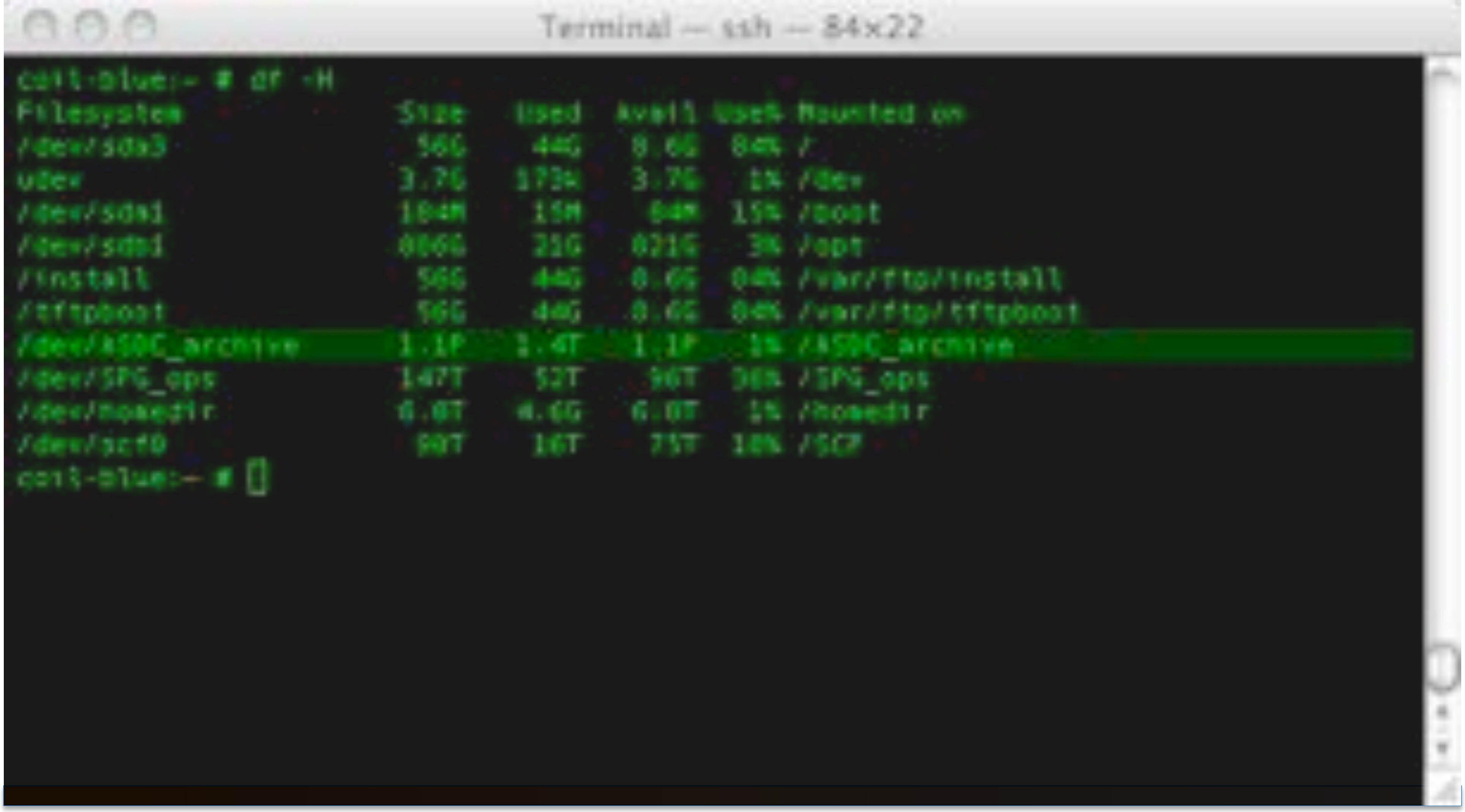


Name	Date Modified
Asgard	Today, 9:27 AM
nohup.out	Aug 20, 2008, 5:28 PM
xen-support	Yesterday, 12:28 PM
iso	Today, 9:31 AM
CentOS-5.2-x86_64-netinstall.iso	Yesterday, 12:29 PM

5 items, 82.43 TB available

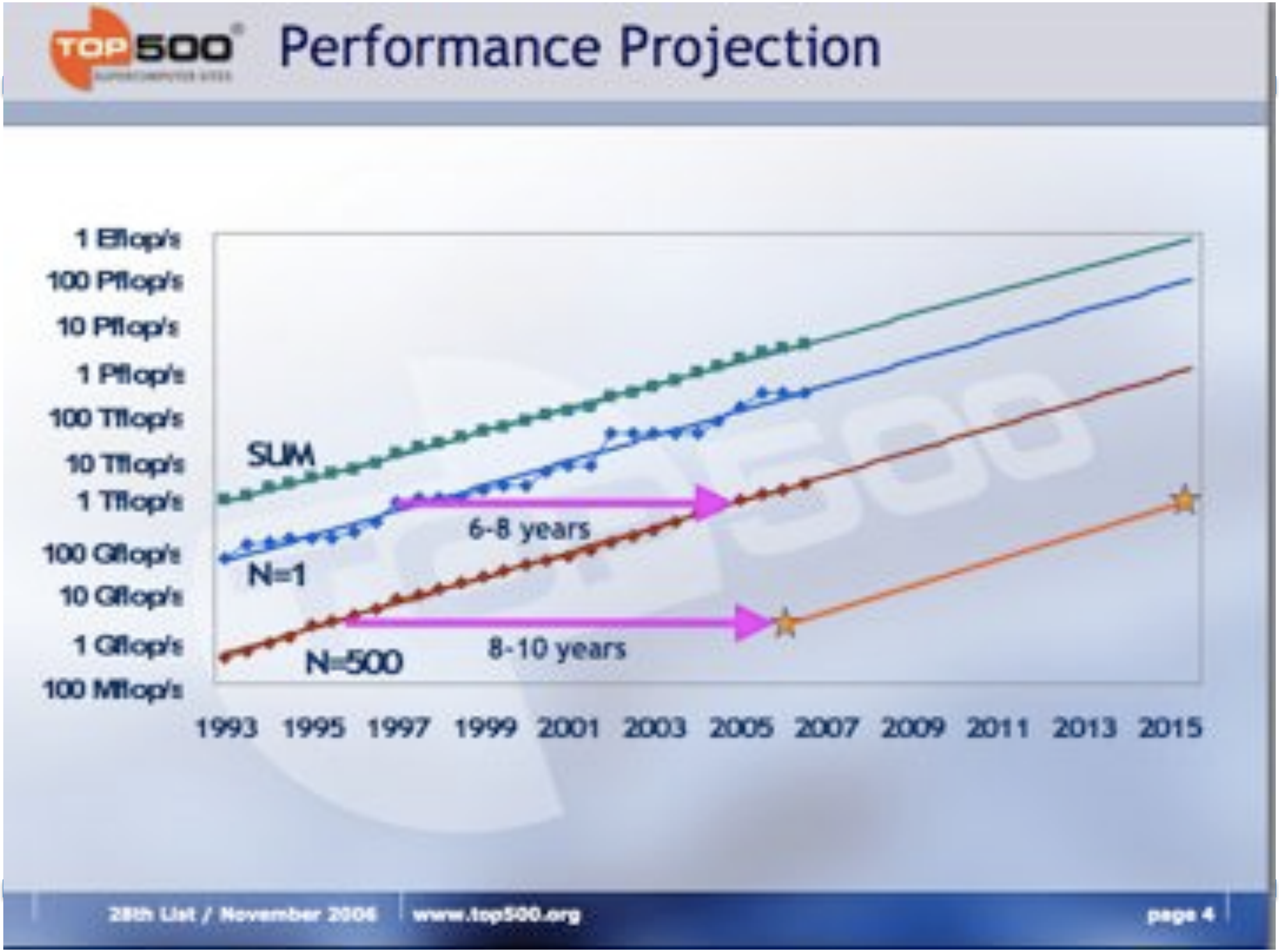
Hardware for the 82TB plus cluster

1.1PB Volume. Even Better



```
coit-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             104M   15M    84M  15% /boot
/dev/sdb1             866G   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  1.4T   1.1P   1% /ASDC_archive
/dev/SPG_ops          147T   52T    96T  36% /SPG_ops
/dev/homedir          6.0T  4.6G   6.0T   1% /homedir
/dev/scf0              90T   16T    75T  18% /SCF
coit-blue:~ # 
```

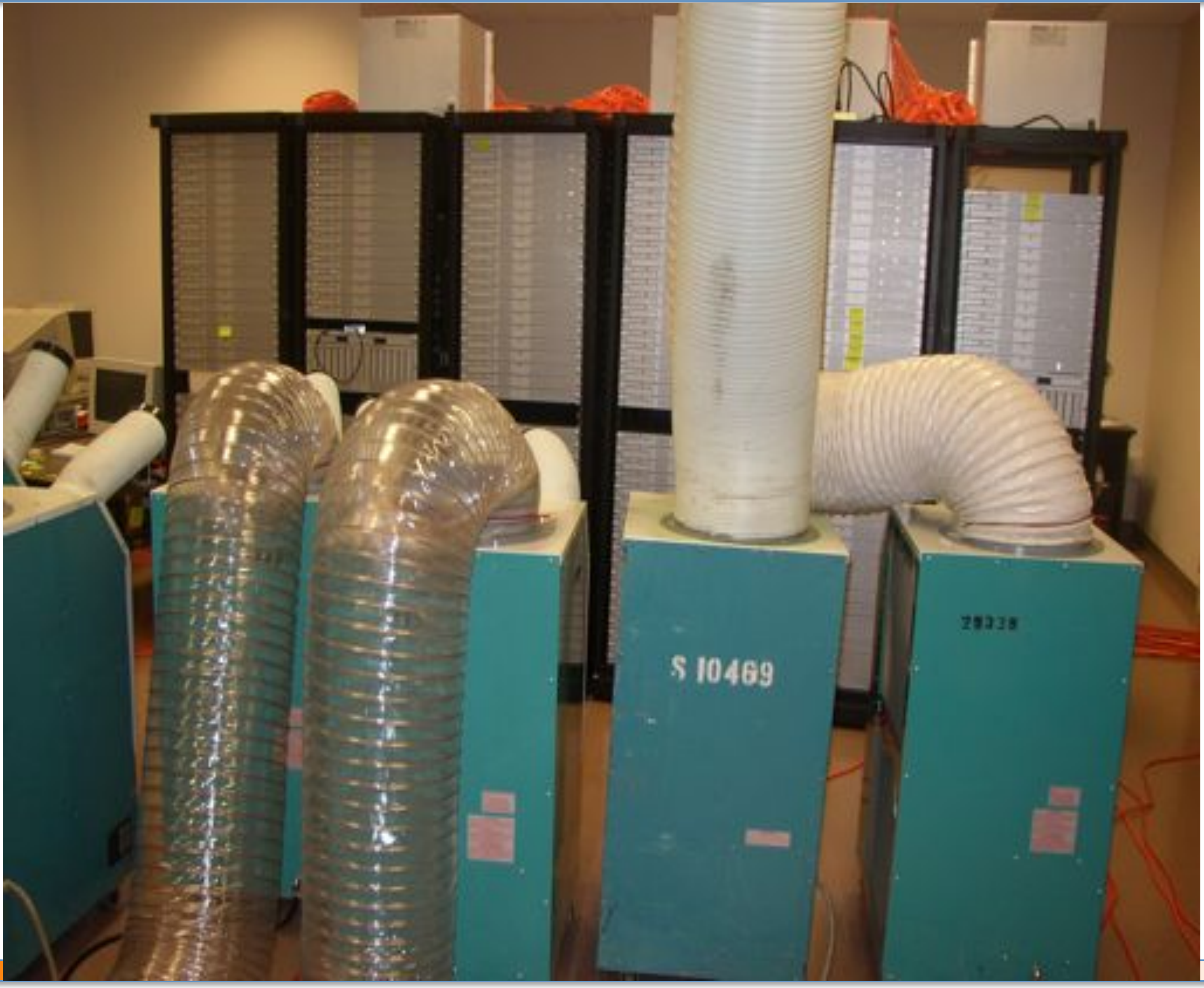
Petabytes are large and heavy



Performance Projection

28th List / November 2006 | www.top500.org | page 4

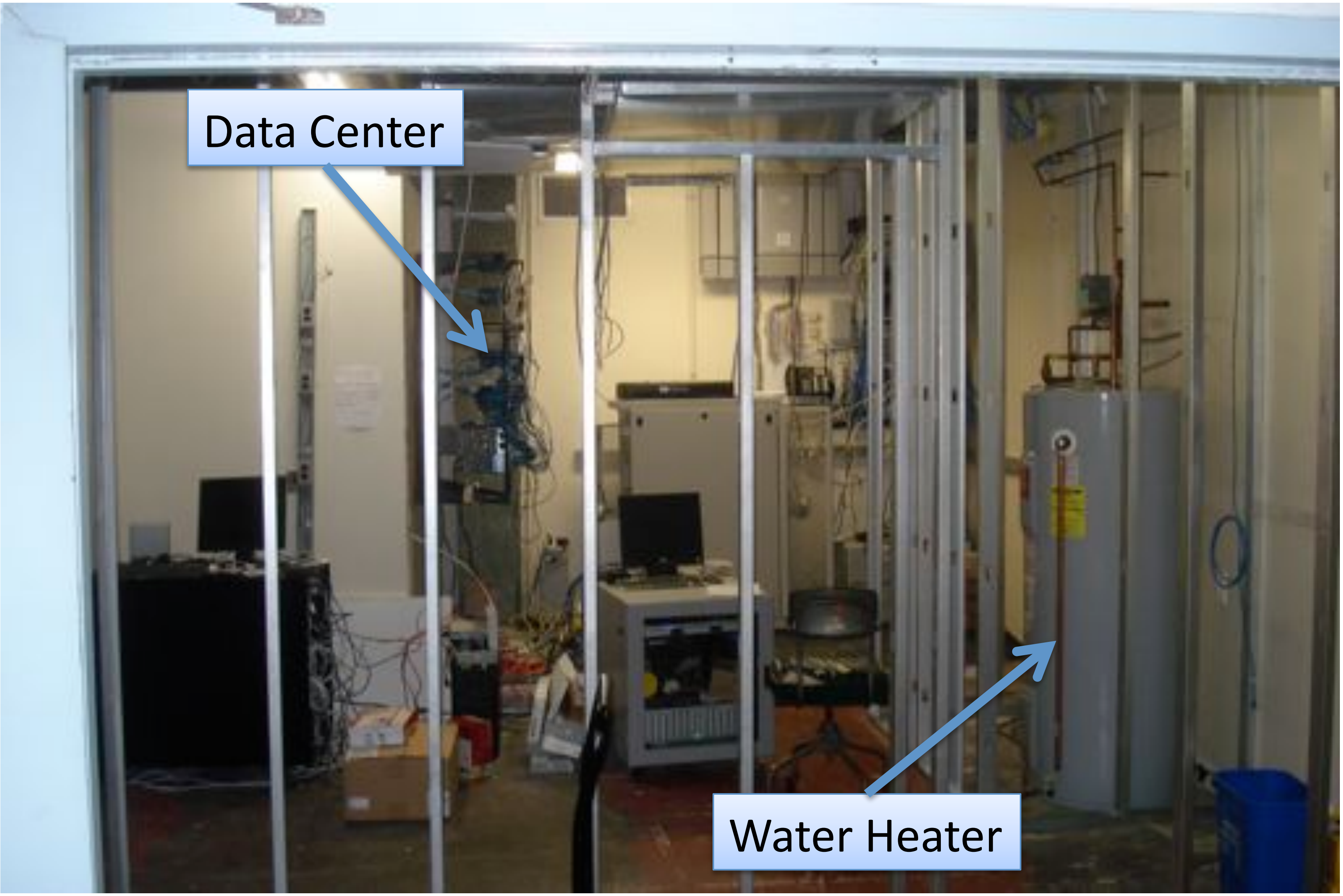
Do not forget facilities planning!



cdwan@bioteam.net

What's the worst that could happen?

All updates happen at once



Data Center

Water Heater

cdwan@bioteam.net

Power for 5PB and 5,000 core cluster

Real World Data Drift

- Volume “Caspian” hosted on server “Odin”
- “Odin” replaced by “Thor”
- “Caspian” migrated to “Asgard”
- “Thor” replaced by “fs01” and “fs02”
- “Asgard” migrated to “/massive”

```
/massive/Asgard/Caspian/blastdb
/massive/Asgard/old_stuff/Caspian/blastdb
/massive/Asgard/can-be-deleted/do-not-delete…
```

cdwan@bioteam.net

Potentially interesting technologies

Technologies you should check out

Amazon EC2: Personal, virtual servers provided by Amazon for pennies on the CPU/hour.

1) Provide credit card
2) Establish DSA key pair
3) Request an “instance”
4) ssh to an IP address as ‘root’

Amazon S3

- Network based storage on Amazon’s servers. Pay by the GB upload / download and GB/month.
- Enables 3rd party apps like ‘dropbox’ and ‘cyberduck’.

Semantic triple datastores (like mediawiki)

Recreational Genomics (23andme et al)

Amazon Compute Cloud

Amazon's EC2 and S3 are "The Grid" as promised in the 90's.

- To my knowledge, all others are academic or vaporware
- This goes double for "private clouds"
- Semantic quibble: "Cloud" != "grid" != "cluster"

Instrument vendors, pharmaceutical companies, and government agencies are all outsourcing *intermittent*, *bursty* computation to the cloud.

Hurdles Cleared:

- Data motion (Fedex-net is up and running as of 2009)
- Security (B2B VPN on a case by case basis)
- Scaling / build out teething issues

Remaining Challenges:

- Social acceptance: Accurate blame for leaks and failures
- Utility / industrial chargeback model (Amex doesn't cut it)
- Quality of service. (Two nines is not sufficient for all use cases)

Cloud Scaling

Storage

- Bioteam contracted to move 40TB of imaging data out of the cloud in early 2009
- Observed high-water-mark of 200TB *into* the cloud in late 2009
- Cloud as long term destination for "permanent" storage
- Still need enterprise level assurance for long term data retention and protection

Clusters

- Bioteam has moved several tools into the cloud for a variety of clients.
- "Burstable" clusters of 100s of nodes are reliable and easy
- 1000s of nodes are more challenging, but possible

Bandwidth remains a challenge

- Fedex remains a not-unreasonable solution

Virtualization (as distinct from ‘cloud’)

Virtualized software delivery is real

- Remote hosting (software as a service)
- VM image (software with no installer)

Server lifecycle management is real

- Old servers don’t die, they become virtual
- Decouple hardware purchases from specific software tools.

Effect on compute clusters

- Have seen virtualized interactive nodes, schedulers, etc.
- ~~Have not seen completely virtual compute nodes~~
  - ~~May be real, I just haven’t seen them in operation yet~~.

Moore's Law Has Changed Character

Code used to get "better" simply by riding the annual increase in clock speed brought on by shrinking electronics.

- Most code has gotten much *less* efficient over time as programmers knew that increased clock speeds and memory sizes would cover their slop.

Very few examples of code that parallelizes out to 1,000s of CPUs.

- That code is already moving to GPUs

Important to revisit architectural assumptions regularly

- Moore's curse: What I used to be proud of doing is now either trivial or actually a bad idea.

23andme spit kit

Closing Thoughts

Career Advice

The "lifestyle entrepreneur" does exist:

- – Neither straightforward nor easy to achieve
- – Requires constant pressure to keep from "mere" business, or bankruptcy.

Know who you work for:

- – Figure out who pays the bills, and what they are buying
- – Be able to explain the high level goals of your project in layman's terms
- – In business, cash is king.
- – In academia, publications are king and PhDs rule.
- – There are not enough tenure track faculty jobs to go around

Avoid morally disreputable projects (your definition will vary)

Never burn any bridges (the community is tiny)

Recommended Annual Exercises

Interview for a job

- Seriously.

Price out a 1,000 core compute cluster and 1PB of storage

- Calculate power and cooling requirements for that system
- Or insert your own benchmark task that you can use to clock technology changes

Develop some tool, from scratch, using whatever technology the cool kids are using that year.

Extra Credit: Consider whether you could live on half your salary, and what you would do with six months of free time per year.

A good example: Your Instructor

Questions?