Essential Scale-Out Computing (CS50)

Dr. James Cuff

Assistant Dean for Research Computing
Harvard University

@jamesdotcuff

first up:

University of Manchester

British Nuclear Fuels Limited

Oxford University

European Bioinformatics Institute

Inpharmatica

Wellcome Trust Sanger Institute

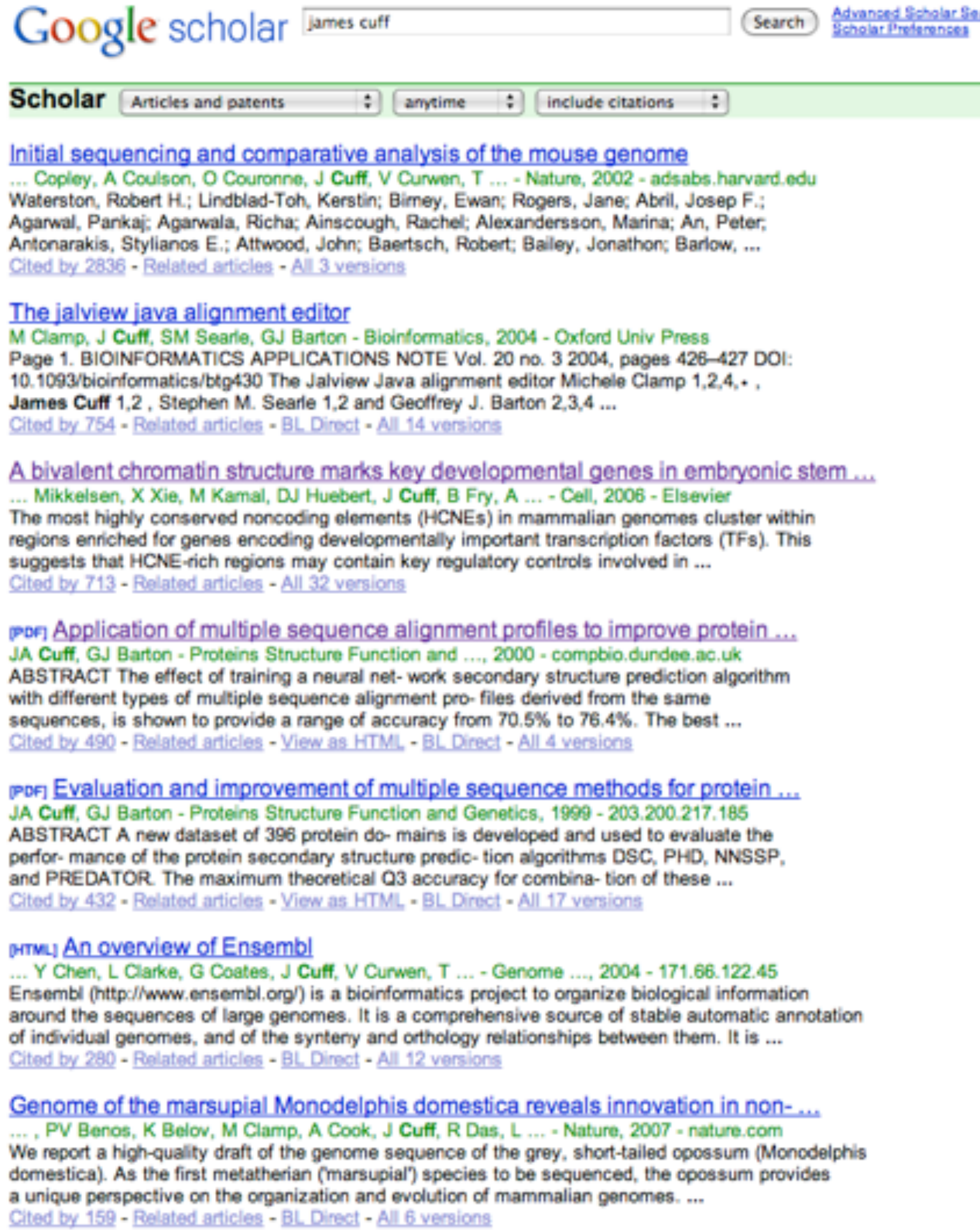
Whitehead Genome Center

Broad Institute of MIT and Harvard

Harvard University

Cycle Computing

Harvard University



basically…

however!





For over eighteen years I've seen research computing scale out…

1996 @ Oxford
1 cpu @ **200Mhz** / **18GB**

2000 @ Sanger / EBI
360 cpu @ **168GHz / 50TB**

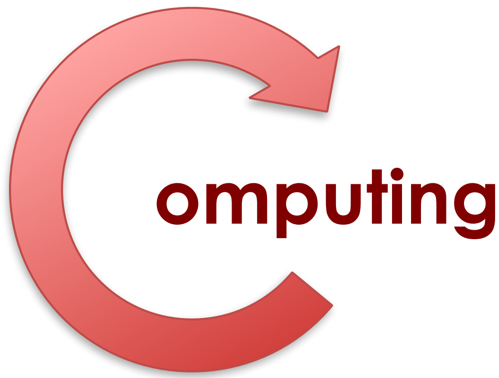
2003-2006 @ Harvard / MIT
200 cpu @ **400GHz / 250-600TB**

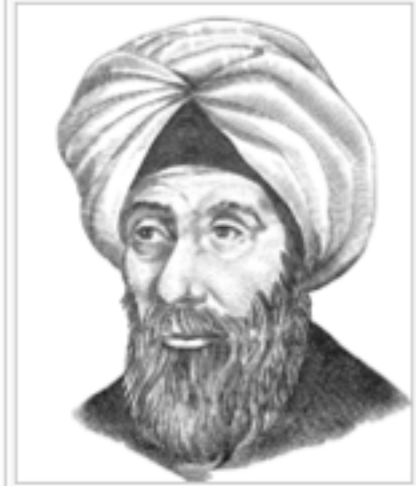
2014 @ Harvard
>59,000 cpu @ **172THz / 15PB**

so let's quickly talk science…

The scientific method

- Formulate your question
- Generate your hypothesis
- **PREDICT** your results
- **TEST** your hypothesis
- Analyze your results





Ibn al-Haytham (Alhazen), 965–1039 Iraq. The Arab scholar who is considered by some to be the father of modern scientific methodology due to his emphasis on experimental data and reproducibility of its results.[4][5]

Computing

PREDICTION and **TESTING**

are the **TWO** cornerstones

of the scientific method

each requires **the** most significant
advances in computation

Remember:

The **TWO PILLARS** Of **SCIENCE**
Are

THEORY and EXPERIMENTATION

And…

COMPUTING is often mentioned as being the **THIRD PILLAR** OF **SCIENCE**…

CS50 Students:

absolutely no pressure for ya'll at all then!

the plan…

History

Harvard

Social Media Things

Some Green Things

Storage

Chaos

Scale Out Hardware

Some Science

so, let's take a moment to look at our history…

The 60's	The 70's	The 80's	The 90's	The 00's	Beyond

*From **centralized** to **decentralized**, **collaborative** to **independent** and right back again!*

Mainframes
~ 0Mbit

VAX
~ 1Mbit

The PC
~ 10Mbit

Beowulf Clusters
~ 1000 Mbit

Central Clusters
~ 10,000 Mbit

Centers provide access to compute

The supercomputing famine, funding gap

Individual computing

Computing is too big to fit under desk, Linux explodes

amazon web services

Clouds/VMware
IaaS, SaaS, PaaS

SHARING

100%	60%	0%	40%	???%

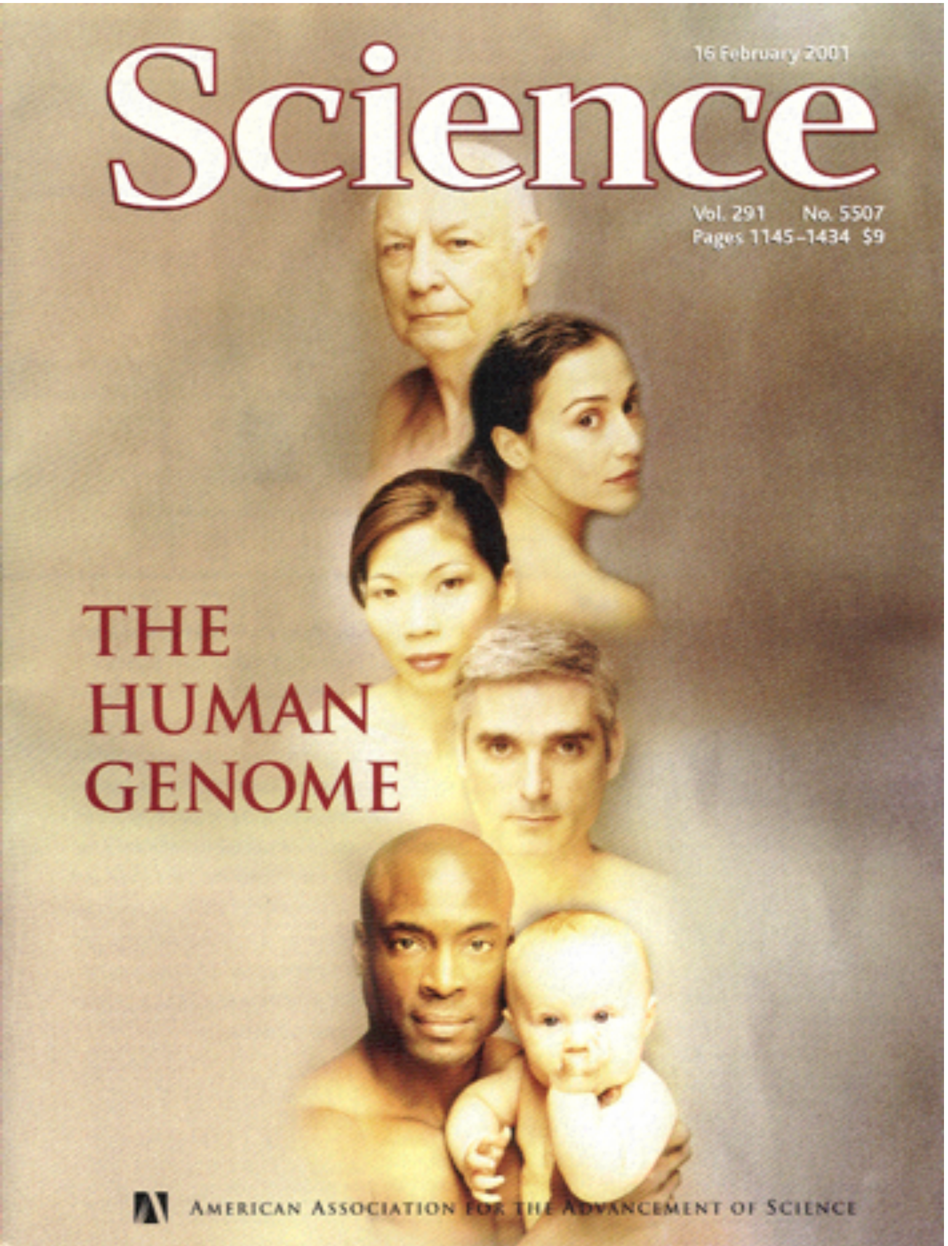
Bigger, better but further and further away from the scientist's laboratory and desktop

The Human Genome Project ca. 2000

aka my first cluster™

- 360 node DEC Alpha DS10L 1U
- Tru64 OS
- 9 racks
- 100KW power
- ATM 622Mb uplink
- 100Mb in rack
- 18 jetstream rs232
- 1,440 cat5 crimps…
- 466MHz x 360 CPU

168GHz


16 February 2001
Science
Vol. 291 No. 5507
Pages 1145–1434 $9
THE HUMAN GENOME
AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE


15 February 2001
nature
www.nature.com
the human genome
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities
naturejobs
genomics special

awesome!

Harvard: my first day in 2006…

not awesome

Incentive

Since human beings are purposeful creatures, the study of incentive structures is central to the study of all economic activity (both in terms of individual decision-making and in terms of co-operation and competition within a larger institutional structure).

Economic analysis, of the differences between societies (and between different organizations within a society) largely amounts to characterizing the differences in incentive structures faced by individuals involved in these collective efforts.

Ultimately, incentives aim to provide value for money and contribute to organizational success.

(wikipedia)

Harvard 2008 - at the start of something!

- 4096 core blade chassis server
- Linux RHEL OS
- 10 racks
- 175KW power
- Ethernet 20,000Mb uplink
- 1000Mb in rack
- No rs232
- No cat5 crimps…
- Non blocking DDR IB
- 2,400MHz x 4096 CPU

9.83THz

Harvard - Today

- ~60,000+ load balanced CPU and climbing!
- ~15.0PB of storage also climbing!
- ca. 200KW/6mnts
- ~600+ virtual machines (KVM)
- ca. 1.8MW of research computing equipment
- 20 dedicated research computing staff
- 0.9PF SP GPGPU, and growing fast!

ok, cool...

history lesson over!

let’s look at some modern scale out compute examples

DISCLAIMER:

I’m more than a little bit obsessed with the scale of social media…

of ounces in an instagram?

- 200 million MAUs(*)
- 20 billion photos
- 60 million new photos / day
- 0.00024GB per photo
- =~ 4,768TB (4.7PB) of disk

[My bee and flower]
(256 kB)

(*) Monthly Active Users

this is tiny spuds!

let's look at the real elephant in the room…

Unique faces (MAUs)

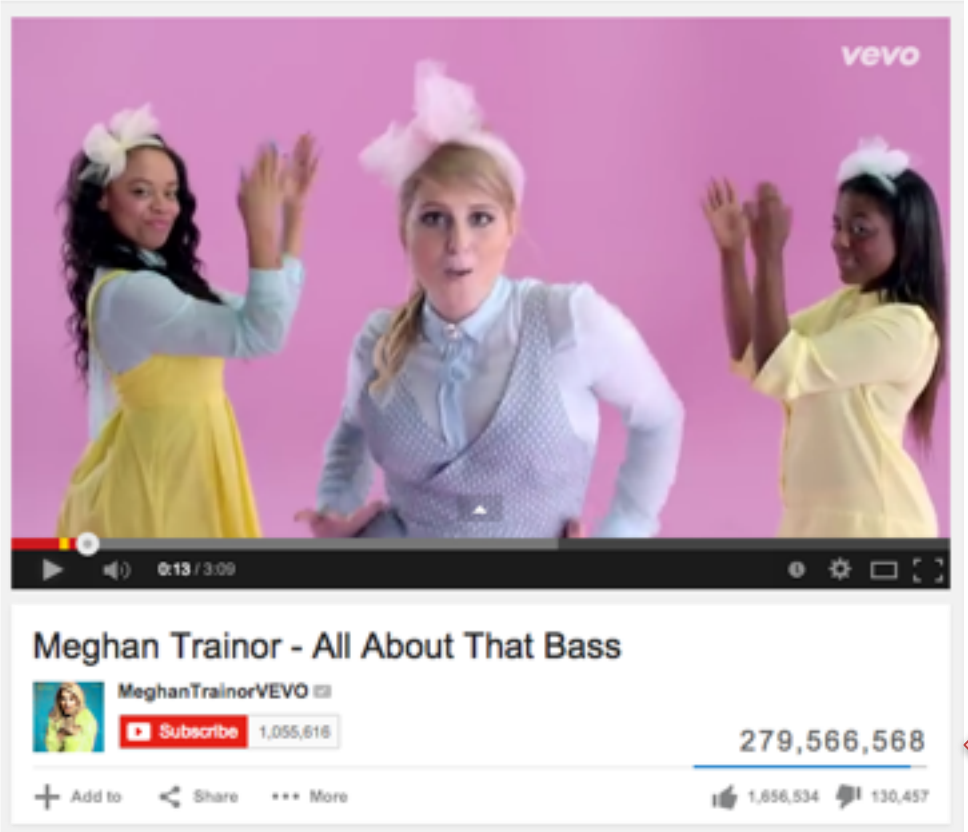
Facebook	=~	1.3 billion
WhatsApp	=~	500 million
Instagram	=~	200 million
Messenger	=~	200 million

2.2 billion total users (1/3rd of the planet)

12 billion msgs/day (7B people on planet)

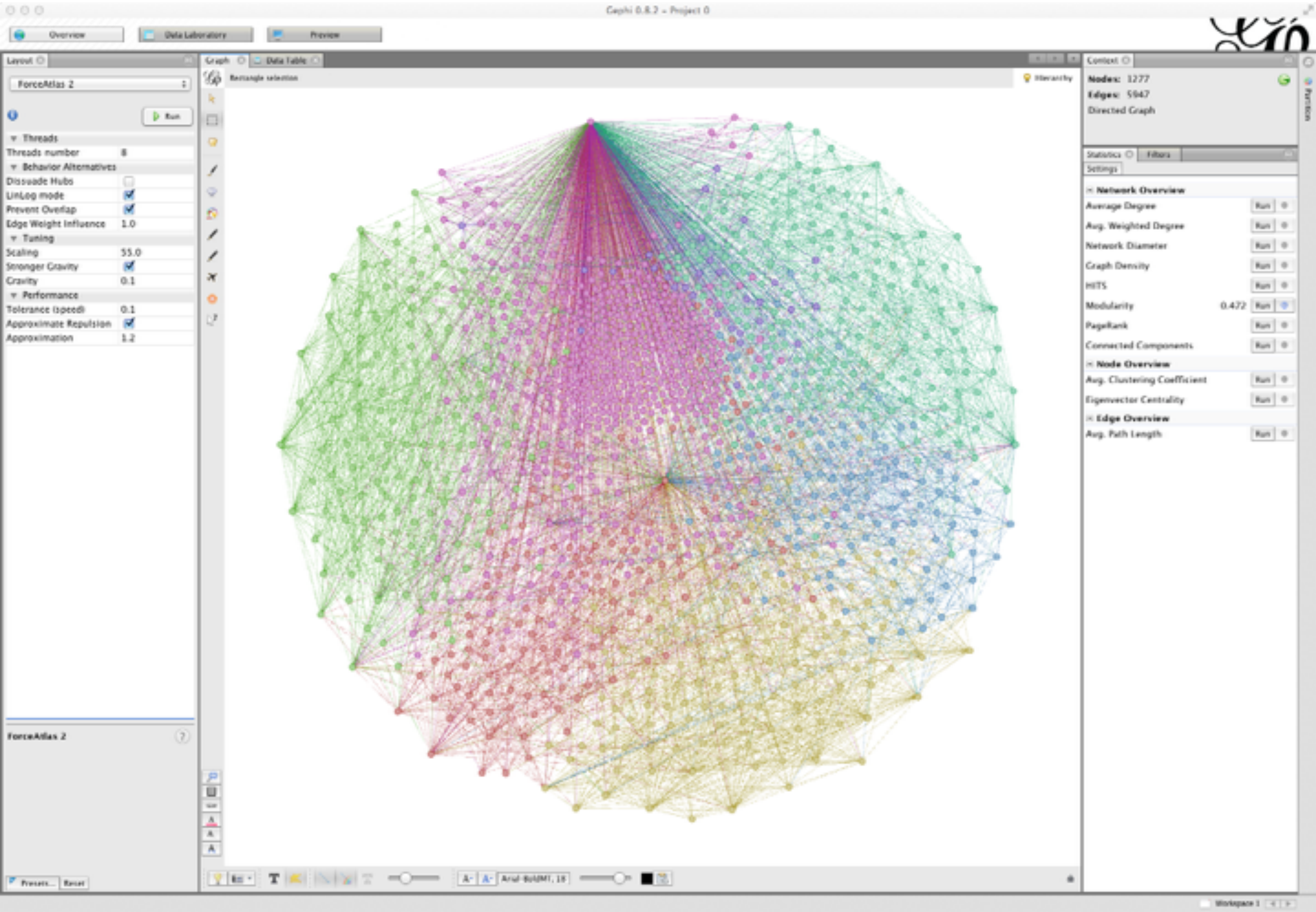
and, it's not even really about the storage or the compute…

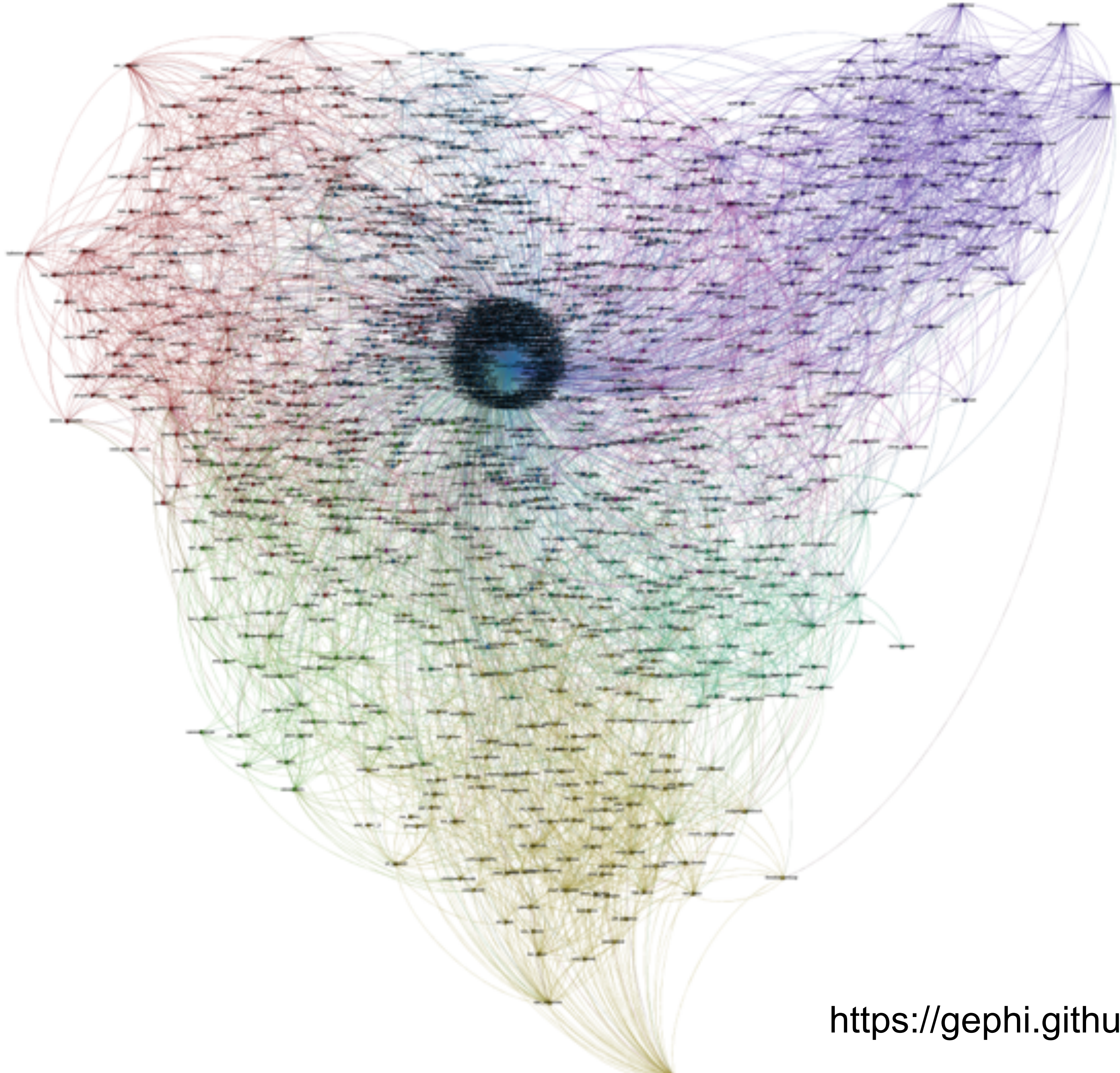
It's All About That GRAPH!



note: very much bass

LinkedIn Data Graph for James Cuff





https://gephi.github.io/

meanwhile back at FriendFace…

April 10 DATA · INFRA

Scaling the Facebook data warehouse to 300 PB

 Pamela Vagata  Kevin Wilfong

At Facebook, we have unique storage scalability challenges when it comes to our data warehouse. Our warehouse stores upwards of 300 PB of Hive data, with an incoming daily rate of about 600 TB. In the last year, the warehouse has seen a 3x growth in the amount of data stored. Given this growth trajectory, storage efficiency is and will continue to be a focus for our warehouse infrastructure.

There are many areas we are innovating in to improve storage efficiency for the warehouse – building cold storage data centers, adopting techniques like RAID in HDFS to reduce replication ratios (while maintaining high availability), and using compression for data reduction before it's written to HDFS. The most widely used system at Facebook for large data transformations on raw logs is Hive, a query engine based on **Corona Map-Reduce** used for processing and creating large tables in our data warehouse. In this post, we will focus primarily on how we evolved the Hive storage format to compress raw data as efficiently as possible into the on-disk data format.

https://code.facebook.com/posts/229861827208629/scaling-the-facebook-data-warehouse-to-300-pb/

well then…

Big Data, Facebook

Facebook Builds Exabyte Data Centers for Cold Storage

BY RICH MILLER ON JANUARY 18, 2013 5 COMMENTS

187 Like | 98 Tweet | 30 in | 15 g+1

Jay Parikh, VP Infrastructure Engineering, Facebook, presents on Facebook's "cold storage" methodology, which the social media giant uses to store user photos. (Photo by Colleen Miller.)

What do you do with an exabyte of digital photos that are rarely accessed? That was the challenge facing Jay Parikh and the storage team at **Facebook**.

The prefix exa indicates multiplication by the sixth power of 1000 or 10^{18} in the International System of Units (SI).

Therefore one exabyte is one quintillion bytes (short scale).

1 EB

= 1000^6 bytes
= 10^{18} bytes
= 1000 petabytes
= 1million terabytes
= 1billion gigabytes.

http://en.wikipedia.org/wiki/Exabyte

http://www.datacenterknowledge.com/archives/2013/01/18/facebook-builds-new-data-centers-for-cold-storage/

which also brings me to being green…

Green computing concepts

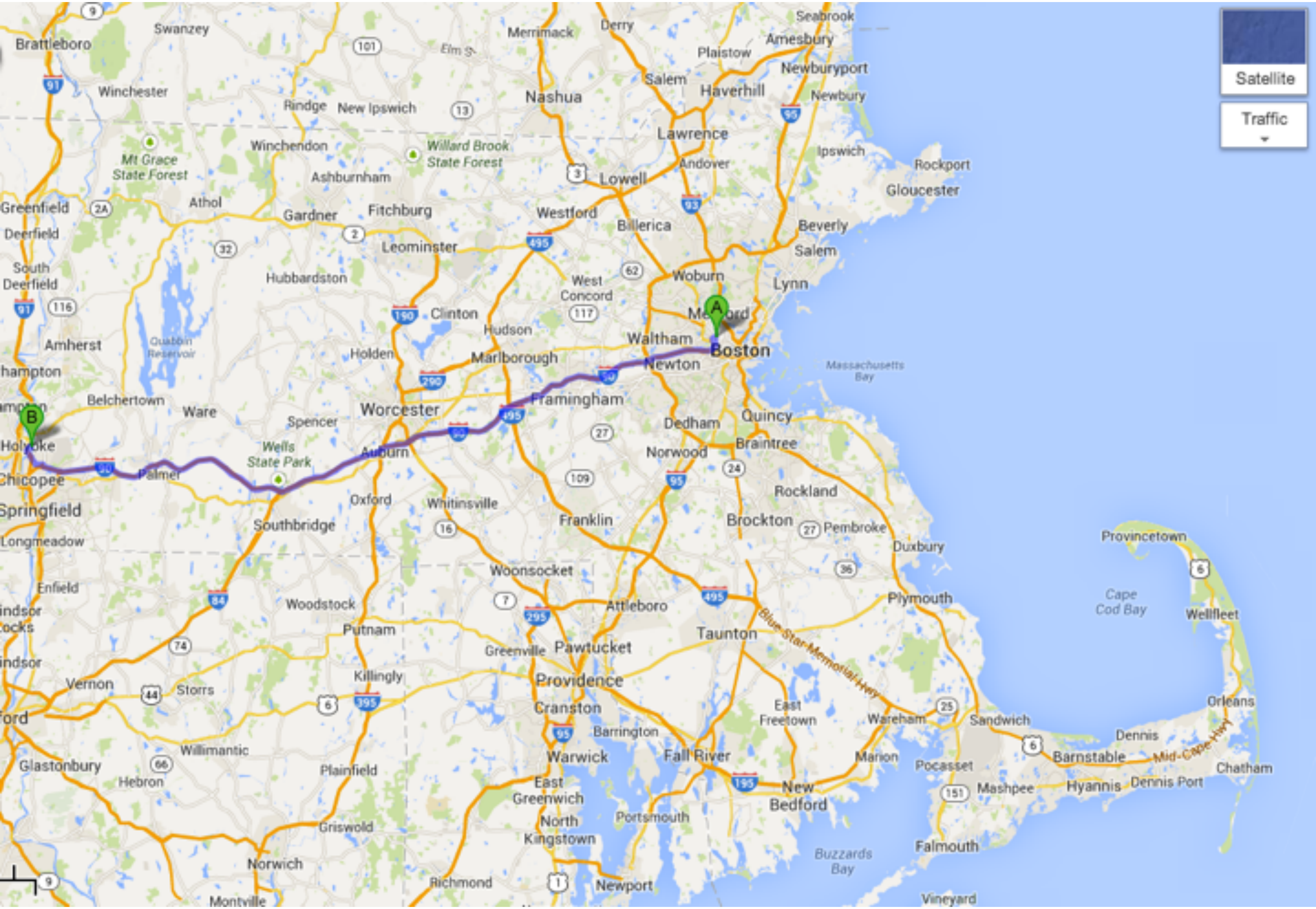
- Product longevity
- Algorithm design
- Datacenter design
- Resource allocation
- Operating systems
- Virtualization

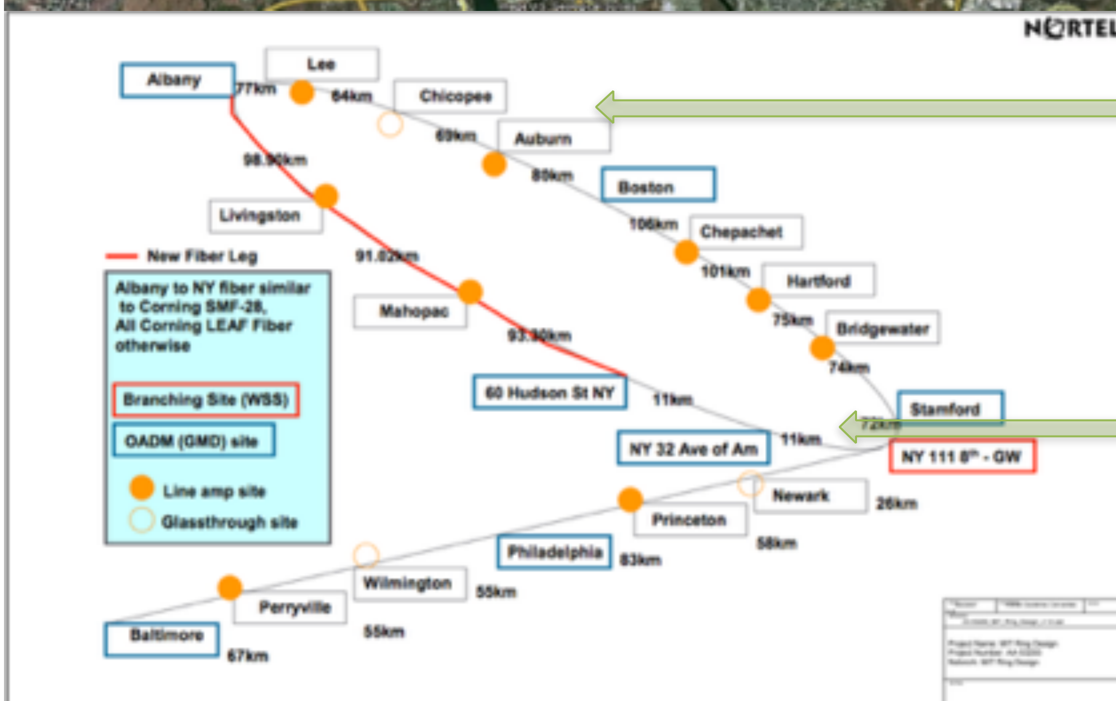
an example from rc

More Ping!
More Power!
More Pipe!

With Less Juice?

#GOWEST!


Satellite
Traffic
Brattleboro
Swanzey
Winchester
Rindge
New Ipswich
Merrimack
Derry
Nashua
Salem
Haverhill
Plaistow
Seabrook
Amesbury
Newburyport
Newbury
Lawrence
Andover
Lowell
Ipswich
Rockport
Gloucester
Winchendon
Willard Brook State Forest
Mt Grace State Forest
Ashburnham
Athol
Gardner
Fitchburg
Westford
Billerica
Beverly
Salem
Greenfield
Deerfield
South Deerfield
Leominster
Hubbardston
West Concord
Woburn
Lynn
Clinton
Hudson
Medford
Amherst
Quabbin Reservoir
Holden
Marlborough
Waltham
Boston
Newton
Massachusetts Bay
Belchertown
Ware
Spencer
Worcester
Framingham
Dedham
Quincy
Holyoke
Wells State Park
Auburn
Norwood
Braintree
Chicopee
Palmer
Springfield
Oxford
Whitinsville
Rockland
Southbridge
Franklin
Brockton
Pembroke
Longmeadow
Duxbury
Provincetown
Enfield
Woonsocket
Woodstock
Attleboro
Plymouth
Cape Cod Bay
Wellfleet
Putnam
Taunton
Blue Star Memorial Hwy
Greenville
Pawtucket
Providence
Vernon
Storrs
Killingly
Cranston
East Freetown
Wareham
Sandwich
Orleans
Barrington
Dennis
Glastonbury
Willimantic
Warwick
Fall River
Marion
Pocasset
Barnstable
Mid-Cape Hwy
Chatham
Hebron
Plainfield
East Greenwich
New Bedford
Mashpee
Hyannis
Dennis Port
North Kingstown
Portsmouth
Griswold
Buzzards Bay
Falmouth
Norwich
Richmond
Newport
Montville
Vineyard



Very large river

Splice!

Route 90

Short path 100 miles

Long path 1000 miles

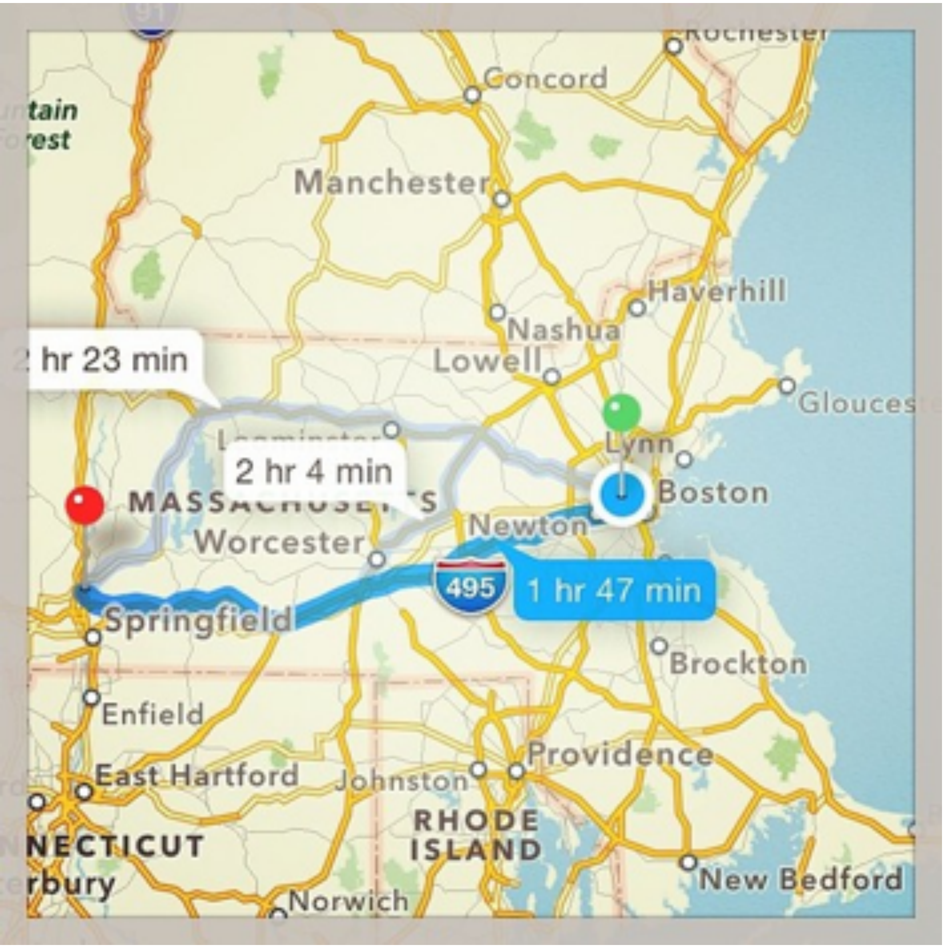
a very large dam

Massachusetts Green High Performance Computing Center

- Hydroelectric power
- MIT, Harvard, UMASS, NEU, BU
- 5MW day one connected load
- Airside economizers (green)
- ca. 640+ racks in "20 rack pods"
- 10% special computing spaces
- Open Feb 2013, first science May (ATLAS)


CAT
CAT


MGHPCC


Concord
Manchester
Haverhill
Nashua
Lowell
Lynn
Boston
2 hr 4 min
Newton
Worcester
495
1 hr 47 min
Springfield
Brockton
Enfield
Providence
East Hartford
Johnston
RHODE ISLAND
Norwich
New Bedford

let's

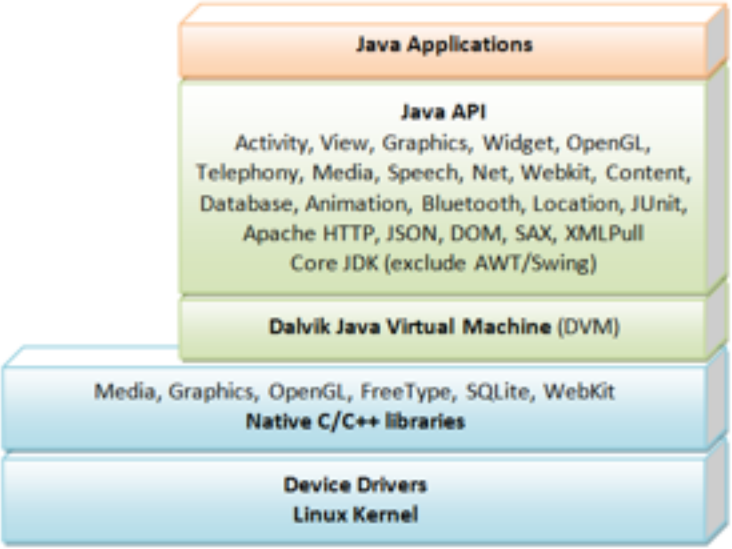
talk

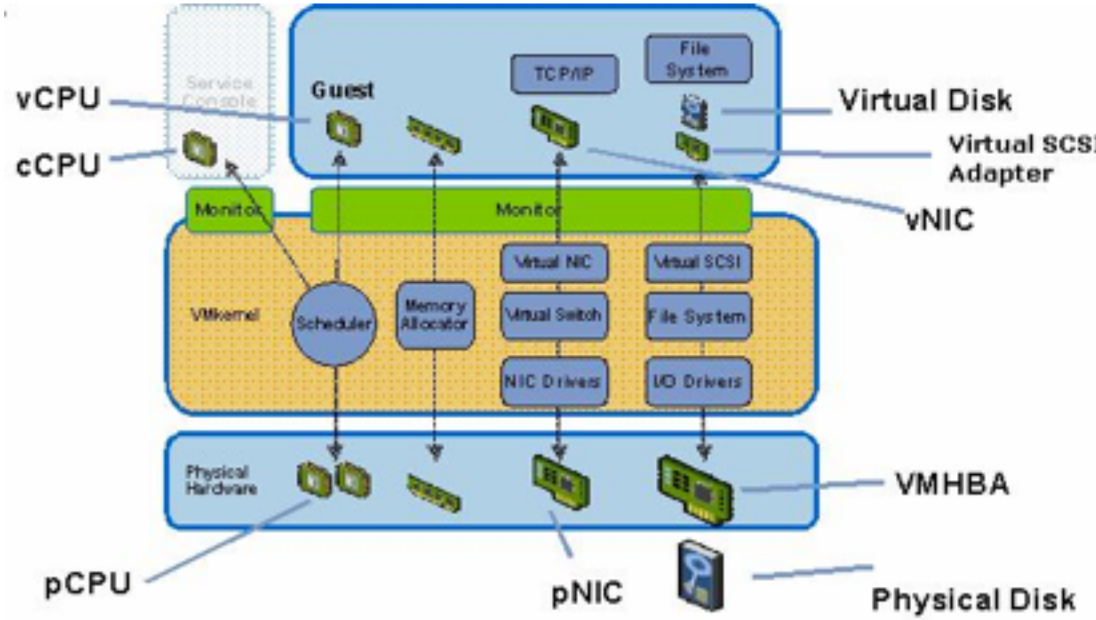
about

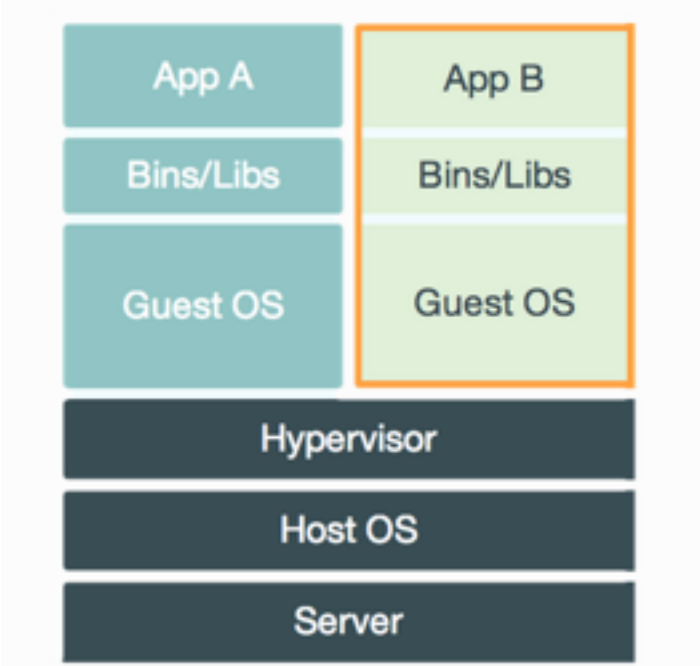
stacks…

OSI (Open Source Interconnection) 7 Layer Model

Layer	Application/Example		Central Device/ Protocols			DOD4 Model
Application (7) Serves as the window for users and application processes to access the network services.	**End User layer** Program that opens what was sent or creates what is to be sent Resource sharing • Remote file access • Remote printer access • Directory services • Network management		**User Applications** SMTP		**GATEWAY** Can be used on all layers	Process
Presentation (6) Formats the data to be presented to the Application layer. It can be viewed as the "Translator" for the network.	**Syntax layer** encrypt & decrypt (if needed) Character code translation • Data conversion • Data compression • Data encryption • **Character Set Translation**		JPEG/ASCII EBDIC/TIFF/GIF PICT			
Session (5) Allows session establishment between processes running on different stations.	**Synch & send to ports** (logical ports) Session establishment, maintenance and termination • Session support - perform security, name recognition, logging, etc.		**Logical Ports** RPC/SQL/NFS NetBIOS names			
Transport (4) Ensures that messages are delivered error-free, in sequence, and with no losses or duplications.	**TCP** Host to Host, Flow Control Message segmentation • Message acknowledgement • Message traffic control • Session multiplexing	PACKET FILTERING	TCP/SPX/UDP			Host to Host
Network (3) Controls the operations of the subnet, deciding which physical path the data takes.	**Packets** ("letter", contains IP address) Routing • Subnet traffic control • Frame fragmentation • Logical-physical address mapping • Subnet usage accounting		**Routers** IP/IPX/ICMP			Internet
Data Link (2) Provides error-free transfer of data frames from one node to another over the Physical layer.	**Frames** ("envelopes", contains MAC address) [NIC card — Switch — NIC card] (end to end) Establishes & terminates the logical link between nodes • Frame traffic control • Frame sequencing • Frame acknowledgment • Frame delimiting • Frame error checking • Media access control		**Switch** **Bridge** **WAP** PPP/SLIP	Land Based Layers		Network
Physical (1) Concerned with the transmission and reception of the unstructured raw bit stream over the physical medium.	**Physical structure** Cables, hubs, etc. Data Encoding • Physical medium attachment • Transmission technique - Baseband or Broadband • Physical medium transmission Bits & Volts		**Hub**			







HOLY STACKS BATMAN!

http://www.reddit.com/r/networking/comments/16tcco/understanding_the_osi_model/

http://www.lifeintech.com/blog/2014/7/19/containerisation-is-the-new-virtualisation

$$\text{FLOPS} = \text{sockets} \times \frac{cores}{socket} \times \text{clock} \times \frac{flops}{cycle}$$

Microprocessors today can do 4 - 16 FLOPs per clock cycle

A single-core 2.5 GHz (clock) has a theoretical performance of 1 MFLOPS

Intel Core 2 and Nehalem:

- 4 DP FLOPs/cycle: 2-wide SSE2 addition + 2-wide SSE2 multiplication
- 8 SP FLOPs/cycle: 4-wide SSE addition + 4-wide SSE multiplication

Intel Sandy Bridge/Ivy Bridge:

- 8 DP FLOPs/cycle: 4-wide AVX addition + 4-wide AVX multiplication
- 16 SP FLOPs/cycle: 8-wide AVX addition + 8-wide AVX multiplication

Intel Haswell:

- 16 DP FLOPs/cycle: two 4-wide FMA (fused multiply-add) instructions
- 32 SP FLOPs/cycle: two 8-wide FMA (fused multiply-add) instructions

AMD K10:

- 4 DP FLOPs/cycle: 2-wide SSE2 addition + 2-wide SSE2 multiplication
- 8 SP FLOPs/cycle: 4-wide SSE addition + 4-wide SSE multiplication

AMD Bulldozer/Piledriver/Steamroller, per module (two cores):

- 8 DP FLOPs/cycle: 4-wide FMA
- 16 SP FLOPs/cycle: 8-wide FMA

Intel Atom (Bonnell/45nm, Saltwell/32nm, Silvermont/22nm):

- 1.5 DP FLOPs/cycle: scalar SSE2 addition + scalar SSE2 multiplication every other cycle
- 6 SP FLOPs/cycle: 4-wide SSE addition + 4-wide SSE multiplication every other cycle

AMD Bobcat:

- 1.5 DP FLOPs/cycle: scalar SSE2 addition + scalar SSE2 multiplication every other cycle
- 4 SP FLOPs/cycle: 4-wide SSE addition every other cycle + 4-wide SSE multiplication every other cycle

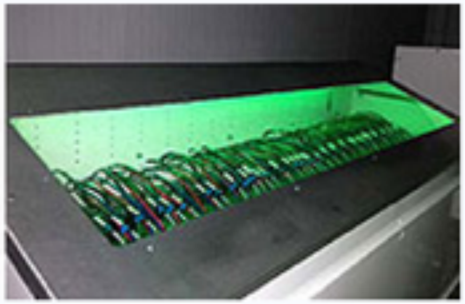
http://stackoverflow.com/questions/15655835/flops-per-cycle-for-sandy-bridge-and-haswell-sse2-avx-avx2

flops per watt

The Green500 List

Listed below are the June 2014 The Green500's energy-efficient supercomputers ranked from 1 to 10.

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	**4,389.82**	GSIC Center, Tokyo Institute of Technology	TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x	34.58
2	**3,631.70**	Cambridge University	Wilkes - Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20	52.62
3	**3,517.84**	Center for Computational Sciences, University of Tsukuba	HA-PACS TCA - Cray 3623G4-SM Cluster, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband QDR, NVIDIA K20x	78.77
4	**3,459.46**	SURFsara	Cartesius Accelerator Island - Bullx B515 cluster, Intel Xeon E5-2450v2 8C 2.5GHz, InfiniBand 4× FDR, Nvidia K40m	44.40
5	**3,185.91**	Swiss National Supercomputing Centre (CSCS)	Piz Daint - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Level 3 measurement data available	1,753.66
6	**3,131.06**	ROMEO HPC Center - Champagne-Ardenne	romeo - Bull R421-E3 Cluster, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR, NVIDIA K20x	81.41
7	**3,019.72**	CSIRO	CSIRO GPU Cluster - Nitro G16 3GPU, Xeon E5-2650 8C 2GHz, Infiniband FDR, Nvidia K20m	86.20
8	**2,951.95**	GSIC Center, Tokyo Institute of Technology	TSUBAME 2.5 - Cluster Platform SL390s G7, Xeon X5670 6C 2.93GHz, Infiniband QDR, NVIDIA K20x	927.86
9	**2,813.14**	Exploration & Production - Eni S.p.A.	HPC2 - iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.8GHz, Infiniband FDR, NVIDIA K20x	1,067.49
10	**2,678.41**	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	54.60


NEC/SMCI 1U Server x 40 Nodes Each node:

- 2x Intel® Ivy-Bridge 2.1GHz 6-Core
- 4x NVIDIA Tesla K20X GPU
- 64GB DDR3 memory
- 120GB SSD
- 4x FDR InfiniBand 56Gbps
- Total Peak: 210TFlops (DP); 630TFlops (SP)

TSUBAME-KFC: Ultra-Green Supercomputer Testbed [2011-2015]

- Fluid Submersion Cooling +
  Outdoor Air Cooling +
  High Density GPU Supercomputing in a 20-feet container (16m^2)
  Cooling Tower: Water 25-35°C >> Outdoor air

- World's top power efficiency (>4.5GFlops/Watt)
- Average PUE 1.05, lower component power
- Field test ULP-HPC results, TSUBAME3.0 Prototype

About

HOW IT BEGAN

A small team of Facebook engineers spent the past two years tackling a big challenge: how to scale our computing infrastructure in the most efficient and economical way possible. Working out of an electronics lab in the basement of our Palo Alto, California headquarters, the team designed our first data center from the ground up; a few months later we started building it in Prineville, Oregon. The project, which started out with three people, resulted in us building our own custom-designed servers, power supplies, server racks and battery backup systems. Because we started with a clean slate, we had total control over every part of the system, from the software to the servers to the data center. This meant we could:

- Use a 480-volt electrical distribution system to reduce energy loss.
- Remove anything in our servers that didn't contribute to efficiency.
- Reuse hot aisle air in winter to both heat the offices and the outside air flowing into the data center.
- Eliminate the need for a central uninterruptible power supply.

The result is that our Prineville data center uses 38 percent less energy to do the same work as Facebook's existing facilities, while costing 24 percent less. Everyone has full access to these specifications. We want you to tell us where we didn't get it right and suggest how we could improve. And opening the technology means the community will make advances that we wouldn't have discovered if we had kept it secret.

WHERE WE GO FROM HERE

The ultimate goal of the Open Compute Project is to spark a collaborative dialogue. We're already talking with our peers about how we can work together on Open Compute Project technology. We want to recruit others to be part of this collaboration -- and we invite you to join us in this mission to collectively develop the most efficient computing infrastructure possible.

real world limits…

R. Landauer

Irreversibility and Heat Generation in the Computing Process

Abstract: It is argued that computing machines inevitably involve devices which perform logical functions that do not have a single-valued inverse. This logical irreversibility is associated with physical irreversibility and requires a minimal heat generation, per machine cycle, typically of the order of kT for each irreversible function. This dissipation serves the purpose of standardizing signals and making them independent of their exact logical history. Two simple, but representative, models of bistable devices are subjected to a more detailed analysis of switching kinetics to yield the relationship between speed and energy dissipation, and to estimate the effects of errors induced by thermal fluctuations.

IBM JOURNAL, JULY **1961**

- At 25 °C the Landauer limit represents an energy of approximately 0.0178 eV

- Theoretically, computer memory operating at the Landauer limit could be changed at a rate of *one billion bits per second* with only 2.85 trillionths of a watt of power expended….

ok so…

~ 65% of the total power @ 199,000,000 watt hours for forty five grand a month…

MGHPCC Energy Use - June, 2013

Period Start (12AM)	6/1/13				
Period Day (12AM)	7/1/13				
Days	30 days				
Hours	720 hours				

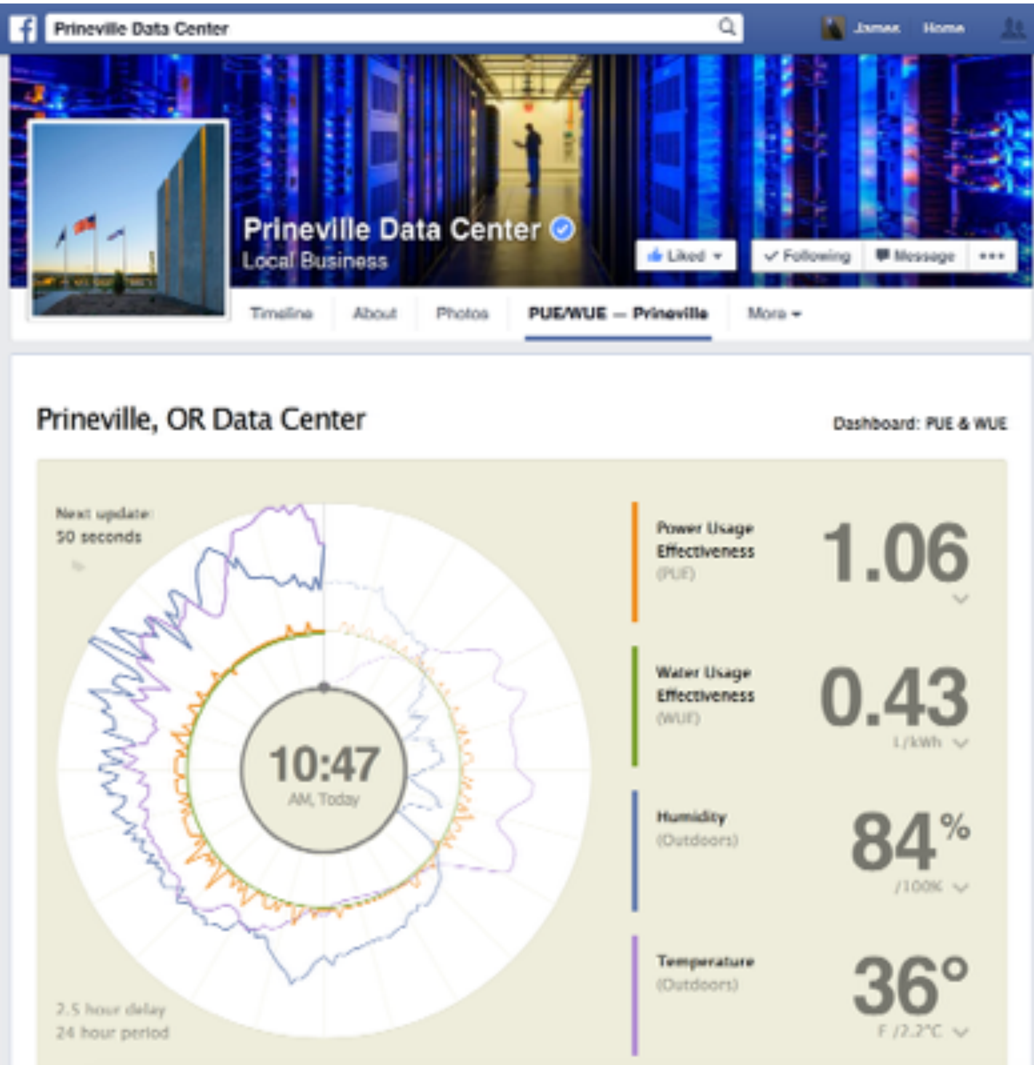
HG&E Invoice

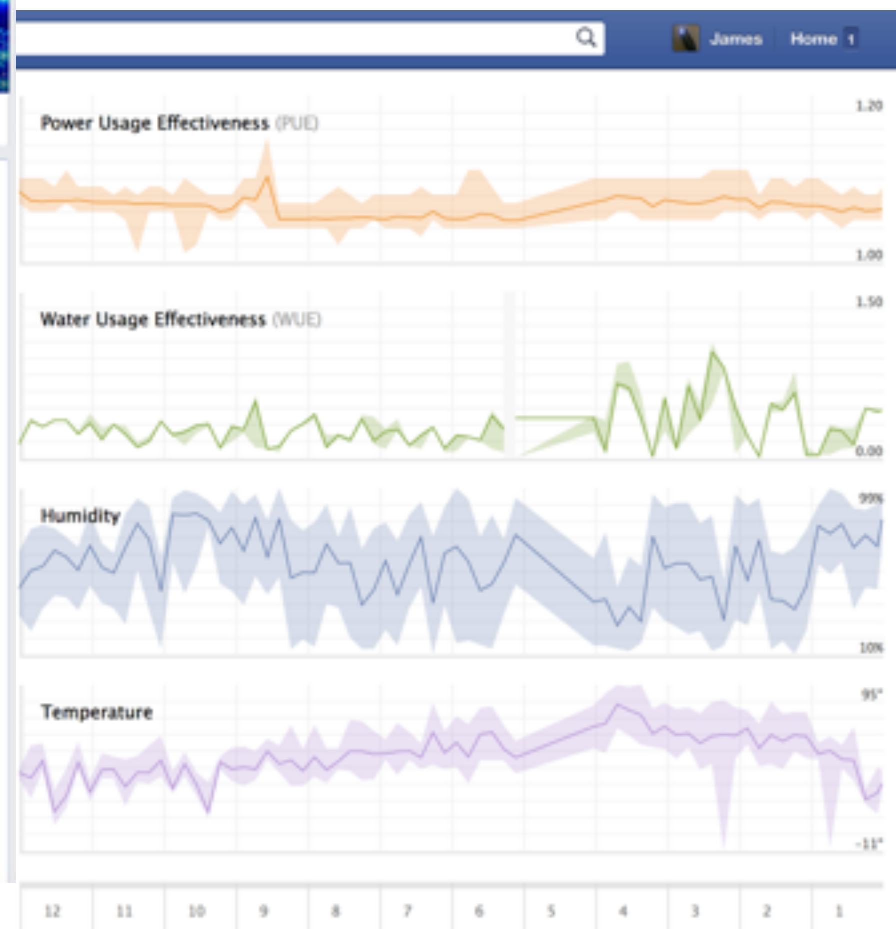
$	78,044.76	Charge
$	(7,769.58)	Discount for prompt payment
$	70,275.18	Net Charge
kWh	698,022	= Energy Use

	Total Consumption			Percent of Total	Member Share	Member Charge ($)
UMass	25,154,168	Wh		8.0%	8.1%	5,722.45
Harvard	199,447,037	Wh		63.7%	64.6%	45,373.20
Northeastern	16,735,392	Wh		5.3%	5.4%	3,807.22
BU	63,619,808	Wh		20.3%	20.6%	14,473.19
MIT	3,952,274	Wh		1.3%	1.3%	899.12
Unassigned	1,501,945	Wh		0.5%		
MGHPCC	2,895,598	Wh		0.9%		
Sandbox	27,703	Wh		0.0%		
Total Computer Room Consumption	313,333,924	44.9%	Wh	100.0%	100.0%	70,275.18
Other Facility Consumption	384,688,076	55.1%	Wh			
Total Energy Consumption	698,022,000	100.0%	Wh			

can you imagine mr zuckerberg's electric bill…

yeah my ‘friends’ are a bit odd…





one more thing on power…

ISO New England @isonewengland 18 Jul

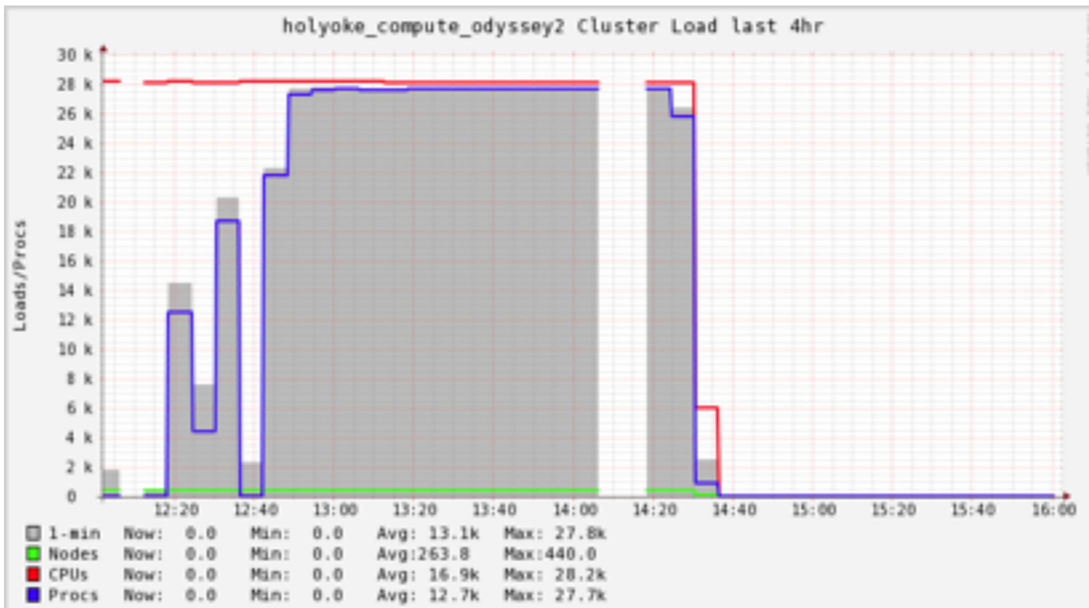
ISO-NE is extending request for consumers to voluntarily conserve--high heat/humidity is pushing up power demand. iso-ne.com/nwsiss/pr/2013...

James Cuff
@jamesdotcuff

Follow

. @isonewengland @GreenHarvard We are doing are part over here in Research Computing! iso-ne.com/nwsiss/pr/2013... pic.twitter.com/e9ZKTze1CQ

12:23 PM - 18 Jul 2013



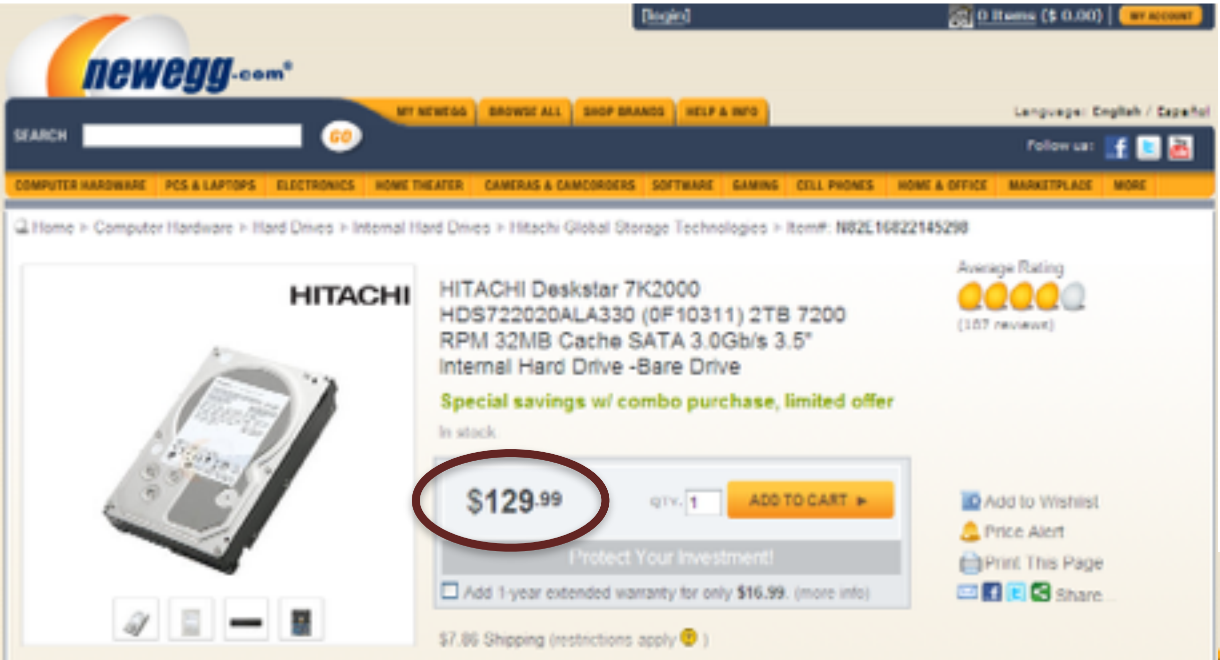
2 RETWEETS

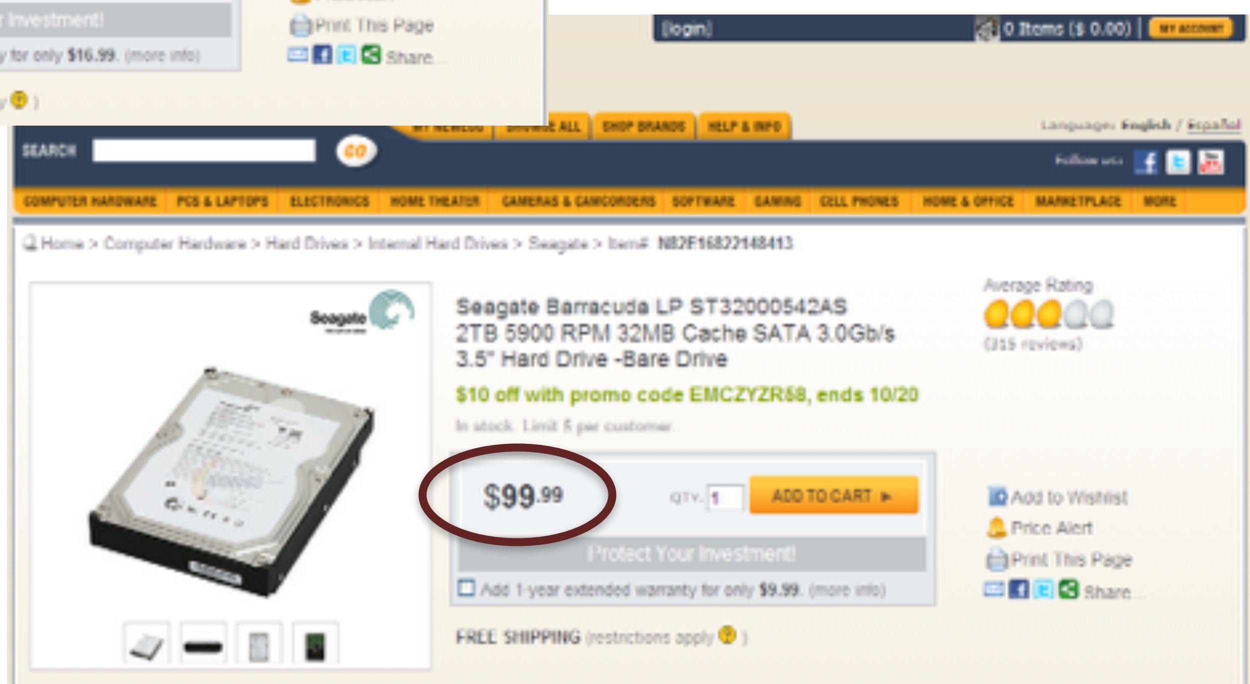
**that time when,
storage leads
to chaos…**

```
[jcuff@hero0101 ~]$ df | wc -l

423

[jcuff@hero0101 ~]$ df -m
| awk '{print sum=sum+$2/1024/1024 " PB"}'
| tail -1

7352.55 PB

[jcuff@hero0101 ~]$
```

If I hear this one more time…





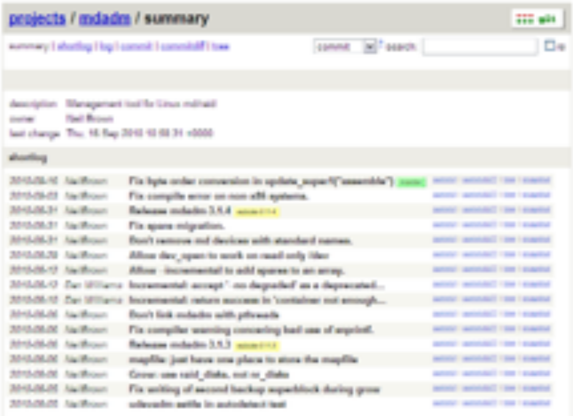
“your storage is too expensive!”
“*…my graduate student can…*”

http://blog.jcuff.net/search/label/STORAGE

… note: way
less bass

So, what could possibly go wrong?

- Magnet, drive head 1 or 0?
- Motors for head and spin
- Drive firmware
- SATA/SAS controller (wires)
- SATA/SAS firmware
- Low level blocks (Perc, Mdadm, LVM)
- File system code (Ext4, XFS, BtrFS)
- OS UBC, VM manager, pages
- DRAM fetch, store
- Algorithms
- Users



Multiplied by 10,000's cpu over 1,000's hosts, with 1,000's individual disk drives and 1000's users

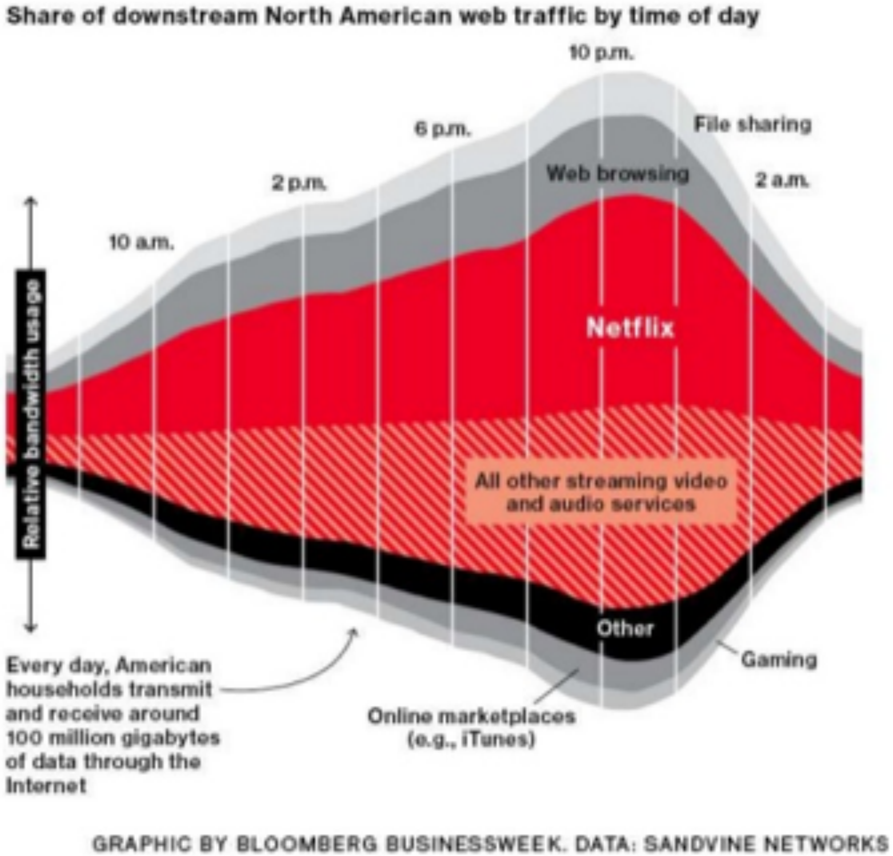
=~ 91,452,464,823,179,310 possible places…

(* enter stage left *)

The Chaos Monkey!

NETFLIX

Netflix Data



- > 37M members
- > 40 countries
- > 1000 device types
- Ratings: > 4M/day
- Searches: > 3M/day
- Plays: > 30M/day
- 1B hours in June 2012
- > 4B hours in Q1 2013
- Log 100B events/day
- 32.25% of peak US downstream traffic

NETFLIX

http://www.slideshare.net/justinbasilico/recommendation-at-netflix-scale

GitHub This repository Search Explore Features Enterprise

Netflix / SimianArmy

Chaos Monkey

Greg Orzell edited this page on Feb 27 · 1 revision

What is Chaos Monkey?

Chaos Monkey is a service which identifies groups of systems and randomly terminates one of the systems in a group. The service operates at a controlled time (does not run on weekends and holidays) and interval (only operates during business hours). In most cases we have designed our applications to continue working when a peer goes offline, but in those special cases we want to make sure there are people around to resolve and learn from any problems. With this in mind Chaos Monkey only runs in business hours with the intent that engineers will be alert and able to respond.

Who in their right mind would willingly chose to work with a Chaos Monkey?

Sometimes you don't get a choice; the Chaos Monkey chooses you

http://blog.codinghorror.com/working-with-the-chaos-monkey/

and remember…

Chaos Monkeys Love Snowflakes!

http://books.google.com/books?id=M4PfbfPnhR4C

some
scale out
hardware…

GPGPU building block

R720

(iDRAC -> 100+ miles away!)

Intel Xeon E5-2650 2.00GHz, 20M Cache, 8.0GT/s QPI, Turbo, 8C, 95W

> 132 x K20 cards

CONNECT-3 VPI ADPT CARD-SGL PT QSFP FDR IB

This rack: >96TFlops @ SP

jcuff — bash — 121×17

```
[root@sa01 ~]# mco facts gpus
Report for fact: gpus

  /dev/nvidia0                                  found 3 times
  /dev/nvidia1,/dev/nvidia0                     found 108 times
  /dev/nvidia3,/dev/nvidia2,/dev/nvidia1,/dev/nvidia0  found 1 times
  /dev/nvidia7,/dev/nvidia6,/dev/nvidia5,/dev/nvidia4,/dev/nvidia3,/dev/nvidia2,/dev/nvidia1,/dev/nvidia0  found 1 times

Finished processing 1359 / 1359 hosts in 1931.15 ms

[root@sa01 ~]# python
Python 2.6.6 (r266:84292, Feb 22 2013, 00:00:18)
[GCC 4.4.7 20120313 (Red Hat 4.4.7-3)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>> (3+(108*2)+4+8)*3.95/1000
0.91245000000000009
>>>
```

We are just short of 1 petaflop of single precision GPGPU!

Because:
Quantum Chemistry @Harvard!

aspuru.chem.harvard.edu

Aspuru-Guzik group

Home | About Alán | Research » | People » | Publications | News

Aspuru-Guzik Research Group

We are a theoretical physical chemistry group in the Department of Chemistry and Chemical Biology at Harvard University.

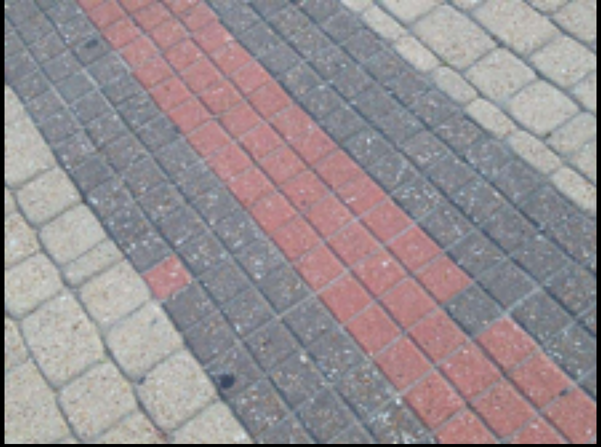
Our research focuses on:

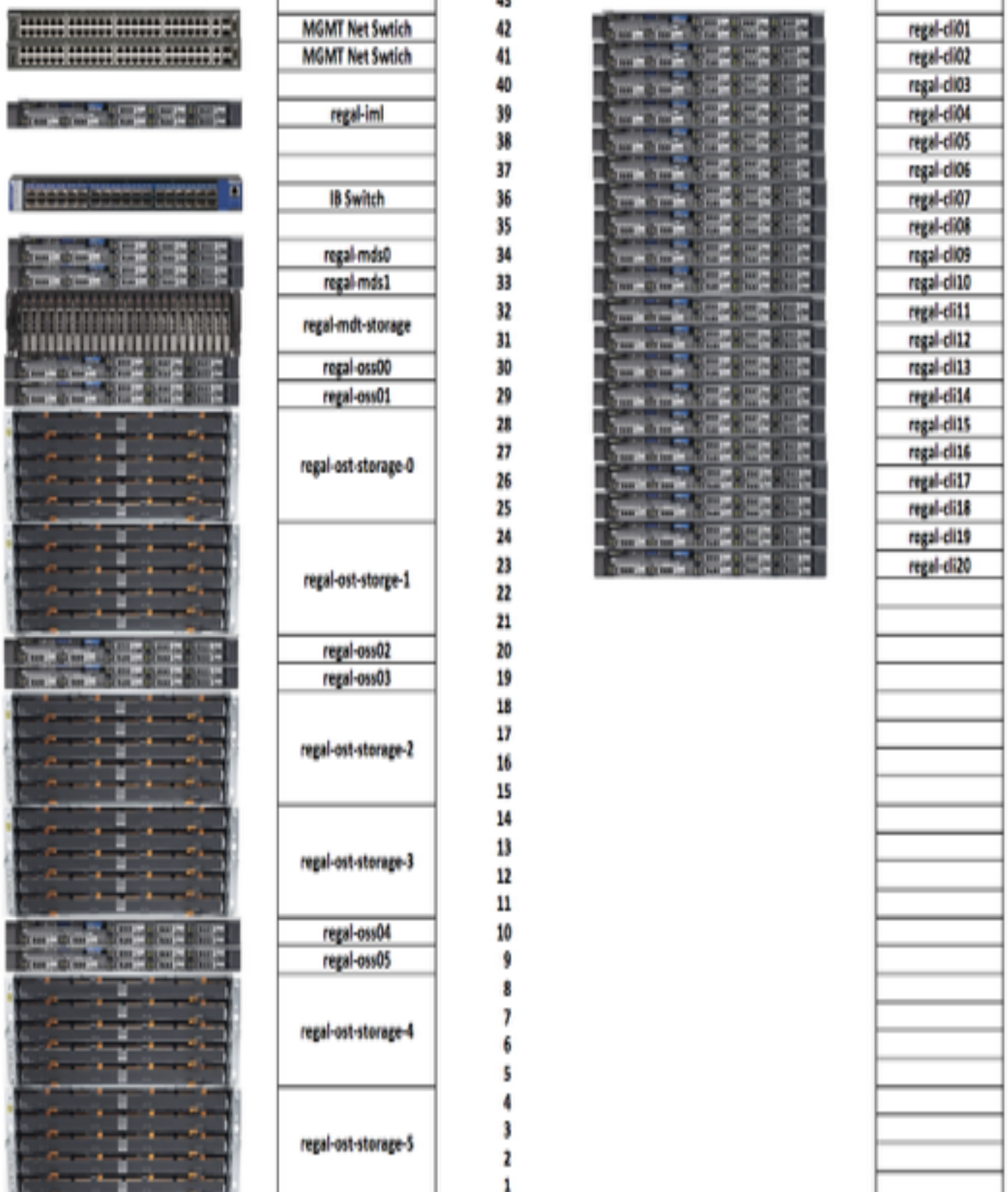
- The connections between quantum computation, quantum information, and chemistry
- Theoretical studies of energy and charge transfer in photosynthetic complexes and renewable energy materials
- Methods development for electronic structure theory: first-principles methods, density functional theory, and quantum Monte Carlo
- Development of the Clean Energy Project, the world's largest distributed computing project for calculating the properties of candidate molecules for organic solar cells

For more information about the group, please use the menu options above.

Powered by WordPress | Designed by Elegant Themes

with scale comes new challenges…



Capacity

- 1050 TB
- 1.1 Billion Files

Peak Throughput

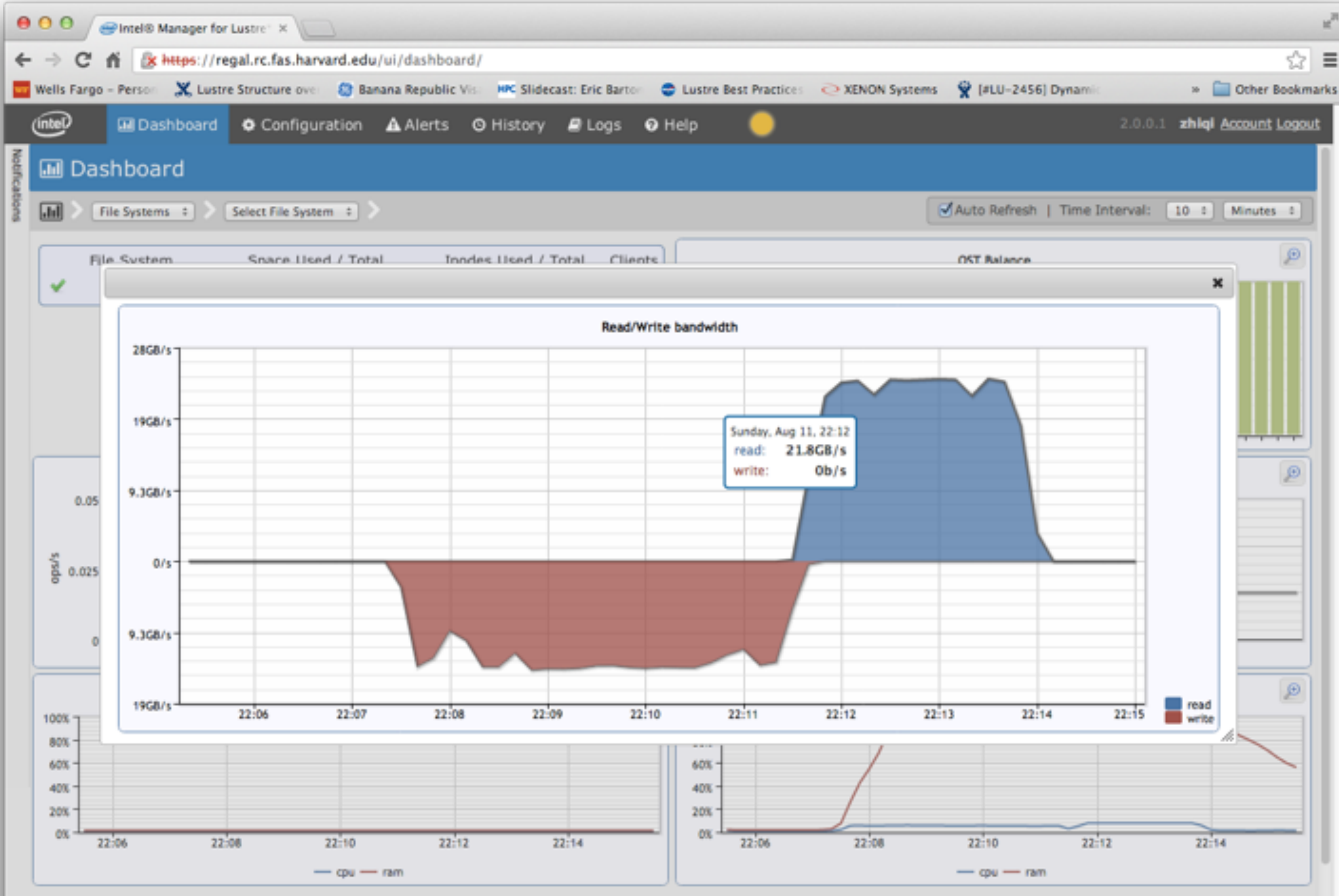
- Write 13GB/s
- Read 20GB/s

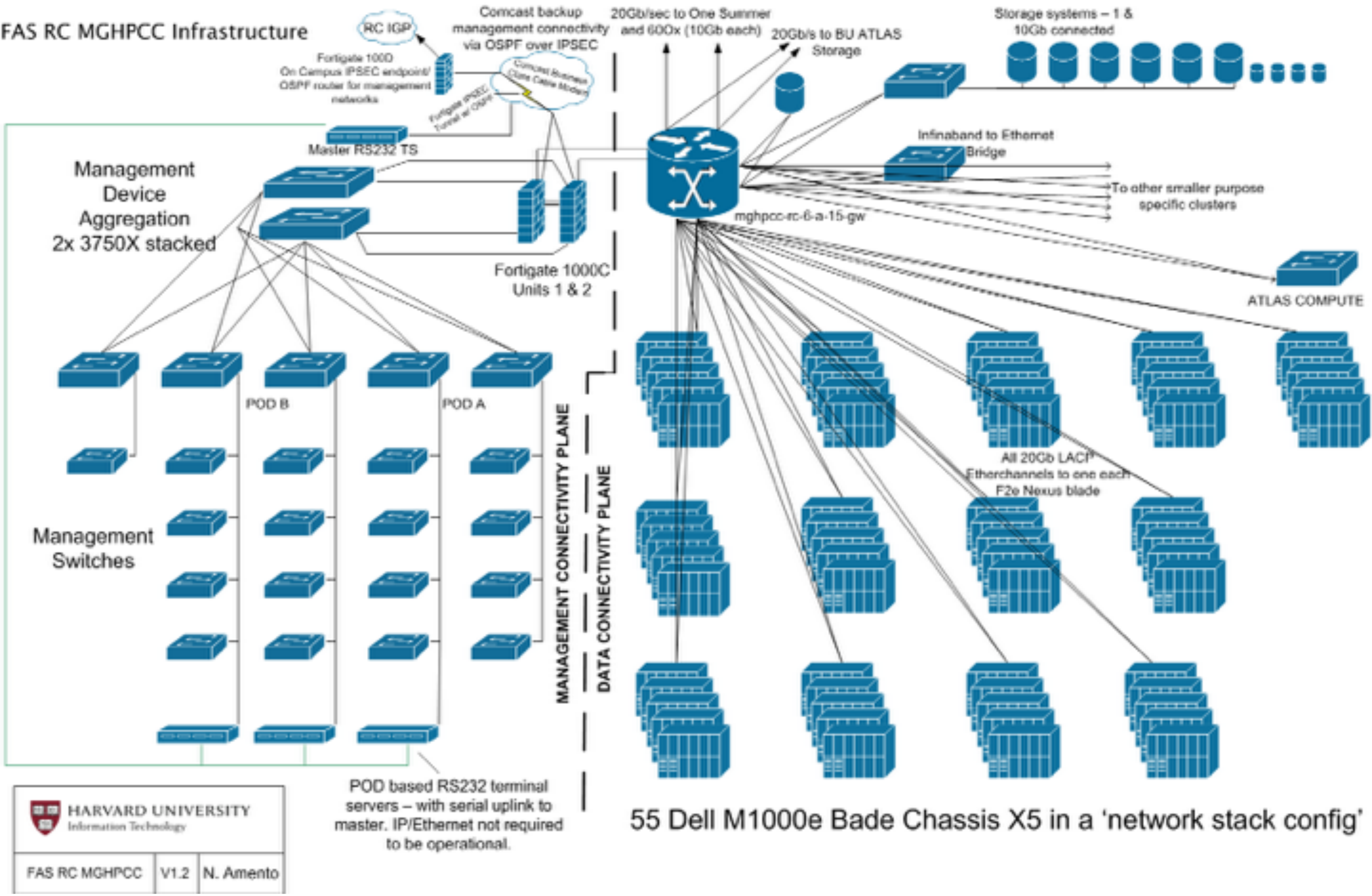
Metadata Rate

- Create 60K
- Lookup 220K
- Getattr 98K
- Setxattr 13K
- Destroy 33K

High Availability

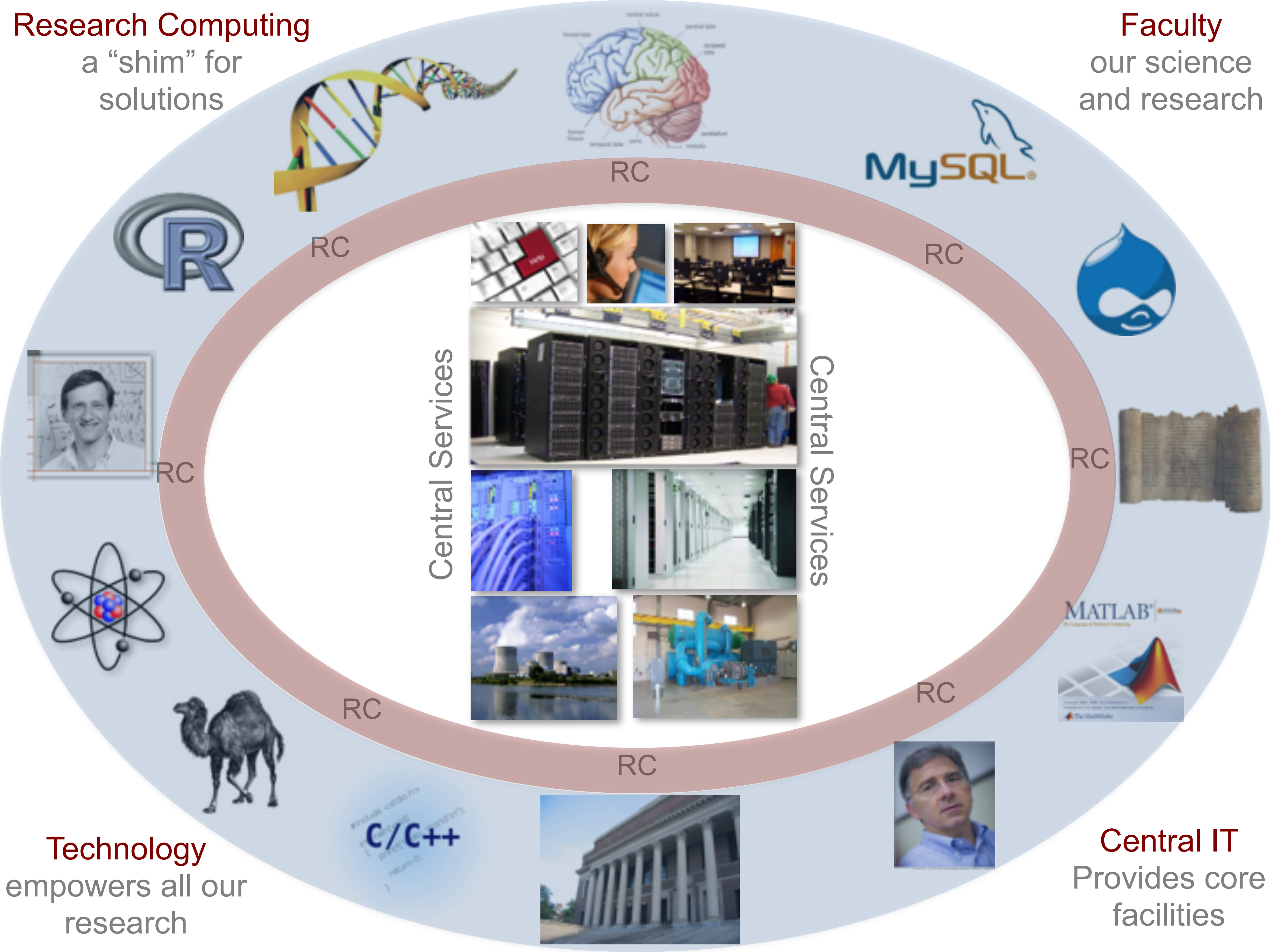
- Active-Standby MDS
- Active-Active OSS
- Active-Active Ctrl
- RAID10/RAID6


Intel® Manager for Lustre
https://regal.rc.fas.harvard.edu/ui/dashboard/
Wells Fargo - Person
Lustre Structure ove
Banana Republic Vis
Slidecast: Eric Barto
Lustre Best Practices
XENON Systems
[#LU-2456] Dynamic
Other Bookmarks
Dashboard
Configuration
Alerts
History
Logs
Help
2.0.0.1
zhlql
Account
Logout
Dashboard
File Systems
Select File System
Auto Refresh | Time Interval:
10
Minutes
File System
Space Used / Total
Inodes Used / Total
Clients
OST Balance
Read/Write bandwidth
28GB/s
19GB/s
9.3GB/s
0/s
9.3GB/s
19GB/s
Sunday, Aug 11, 22:12
read: 21.8GB/s
write: 0b/s
22:06
22:07
22:08
22:09
22:10
22:11
22:12
22:13
22:14
22:15
read
write
0.05
0.025
0
ops/s
100%
80%
60%
40%
20%
0%
22:06
22:08
22:10
22:12
22:14
cpu
ram



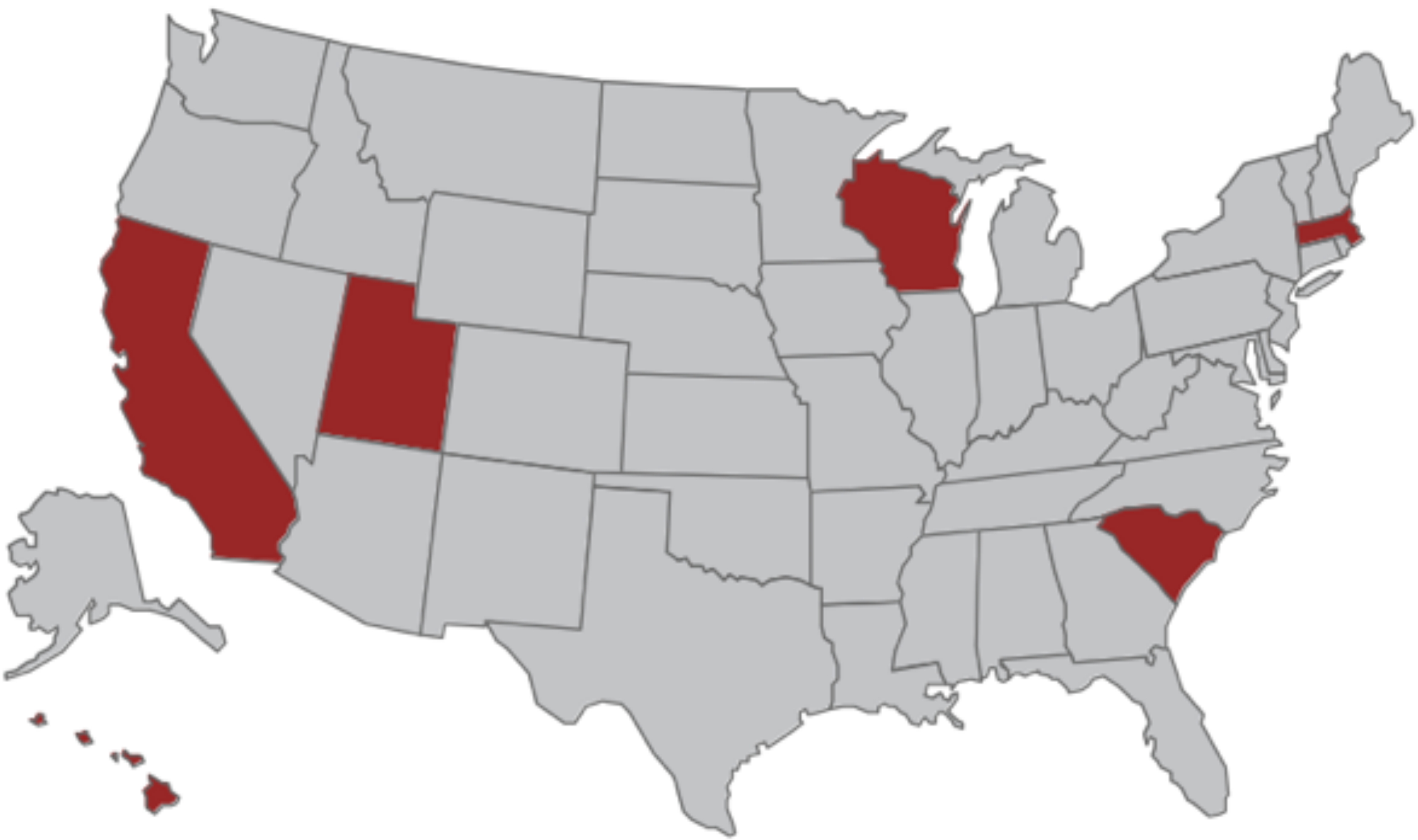
FAS RC MGHPCC Infrastructure
RC IGP
Fortigate 100D
On Campus IPSEC endpoint/
OSPF router for management
networks
Comcast backup
management connectivity
via OSPF over IPSEC
Fortigate IPSEC
Tunnel w/ OSPF
20Gb/sec to One Summer
and 60Ox (10Gb each)
20Gb/s to BU ATLAS
Storage
Storage systems – 1 &
10Gb connected
Master RS232 TS
Management
Device
Aggregation
2x 3750X stacked
Infinaband to Ethernet
Bridge
To other smaller purpose
specific clusters
mghpcc-rc-6-a-15-gw
Fortigate 1000C
Units 1 & 2
ATLAS COMPUTE
POD B
POD A
MANAGEMENT CONNECTIVITY PLANE
DATA CONNECTIVITY PLANE
All 20Gb LACP
Etherchannels to one each
F2e Nexus blade
Management
Switches
POD based RS232 terminal
servers – with serial uplink to
master. IP/Ethernet not required
to be operational.
55 Dell M1000e Bade Chassis X5 in a 'network stack config'
HARVARD UNIVERSITY
Information Technology
FAS RC MGHPCC
V1.2
N. Amento

ok time for a spot of science…

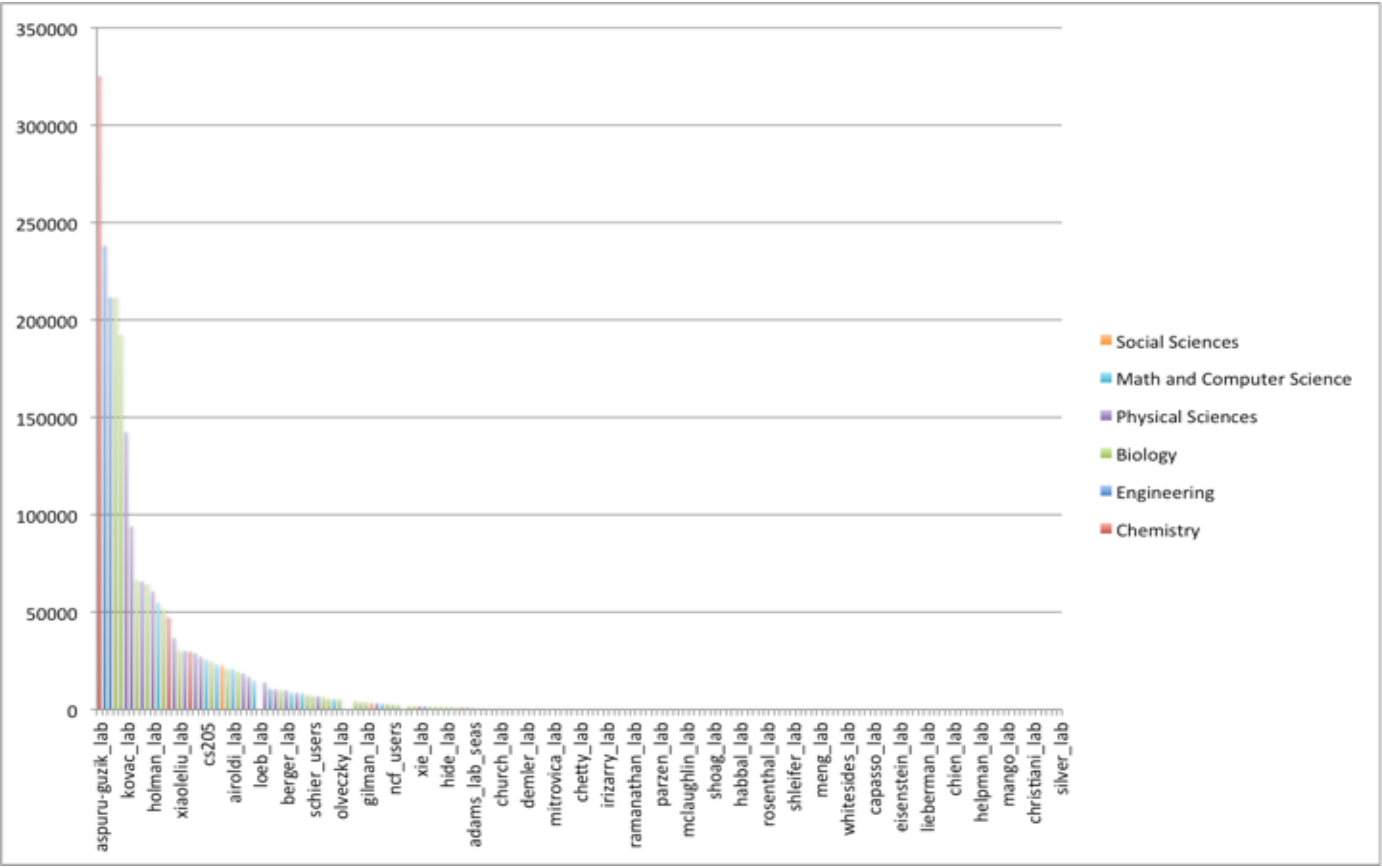

Research Computing
a “shim” for solutions
Faculty
our science and research
Technology
empowers all our research
Central IT
Provides core facilities
RC
RC
RC
RC
RC
RC
RC
RC
Central Services
Central Services
MySQL
MATLAB
C/C++

ACI-REF (condo of condos)

- Goal: **Advance our nation's research & scholarly achievements through the transformation of campus computational capabilities and enhanced coupling to the national infrastructure.**



The “long tail of science”



MGHPCC
MASSACHUSETTS GREEN HIGH PERFORMANCE COMPUTING CENTER

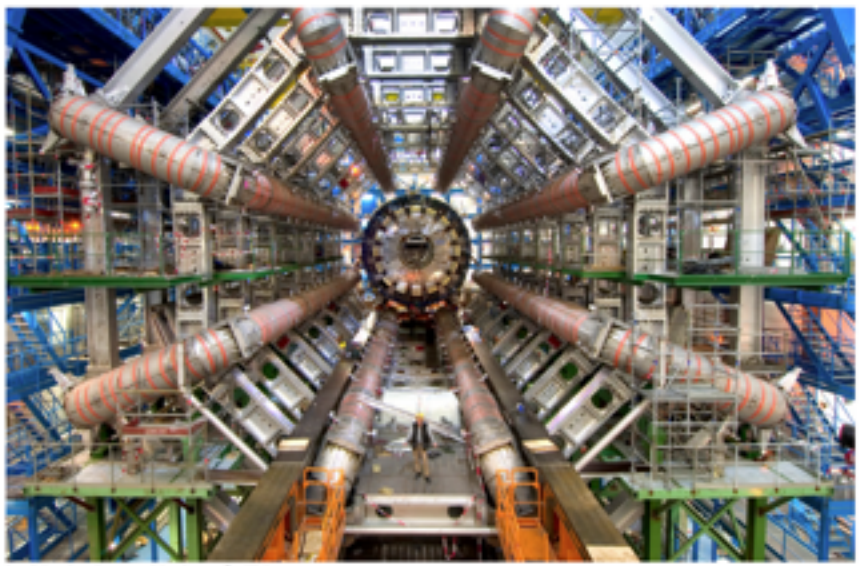
BLOG.

JUL 22

HIGGS BOSON SPOTTED IN HOLYOKE?

by Helen Hill

In one intense week of activity in May, teams from Boston University's Center for Computational Science and Harvard's FAS Research Computing division moved one of the major computing facilities for analysis of data from the Large Hadron Collider (LHC) at CERN from various locations in the Boston area to MGHPCC.



Inside the Large Hadron Collider

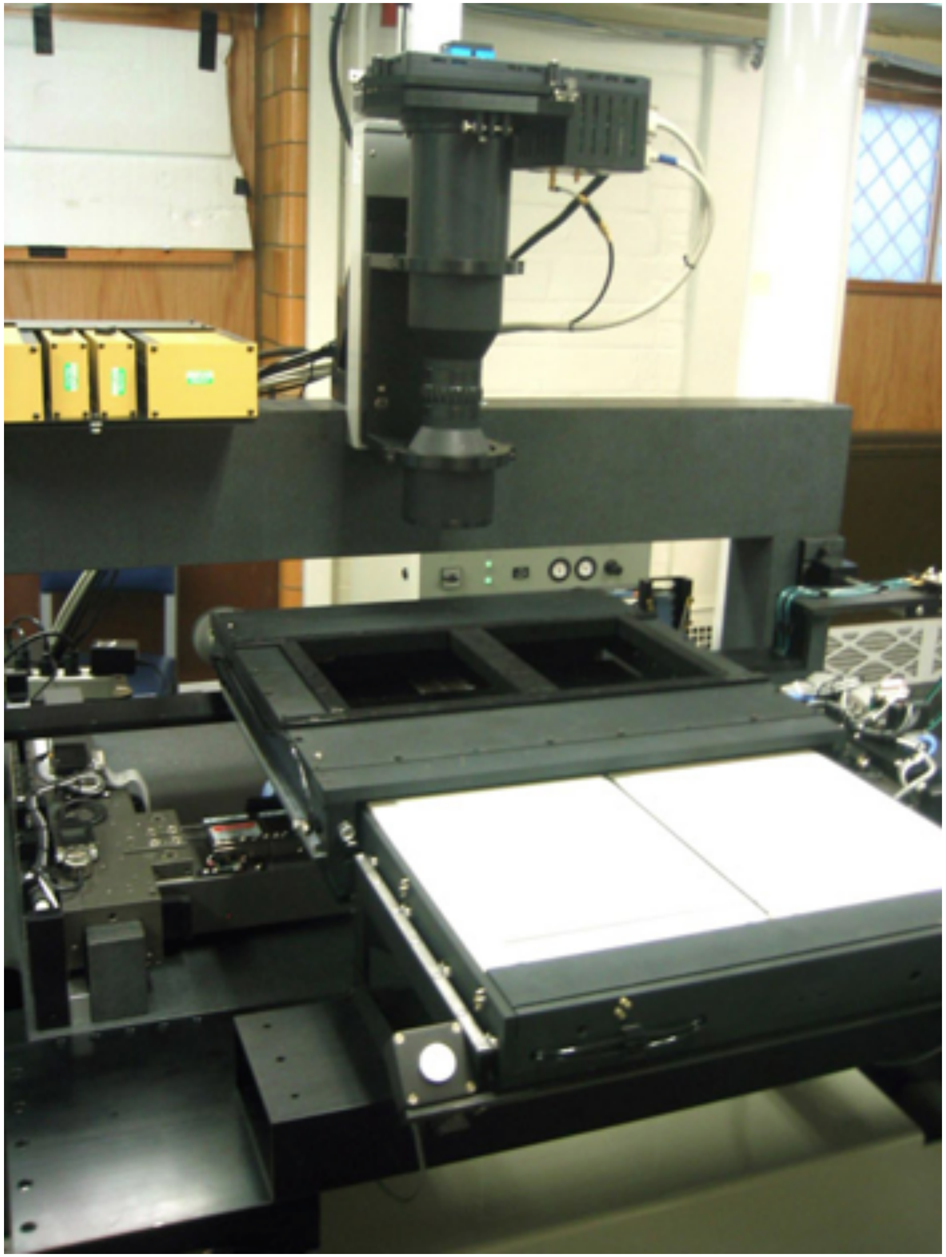
The facility is one of five "Tier 2" centers in the U.S. and is known as the ATLAS Northeast Tier 2 center (NET2). NET2 is part of a worldwide network of computer centers which work together to analyze data from the highest energy proton-proton collisions ever produced in the Laboratory.

CONTACT.

100 Bigelow St. Holyoke, MA 01040
Phone: 413.552.4900
Press: press@mghpcc.org
Research: research@mghpcc.org

QUICK FACTS.

- LOCATION: HOLYOKE, MA USA
- POWER: 10 MEGAWATTS



DASCH

Digital Access
to a Sky Century
@ Harvard

Digitizing the plate archive (450,000 plates) at the Harvard College Observatory

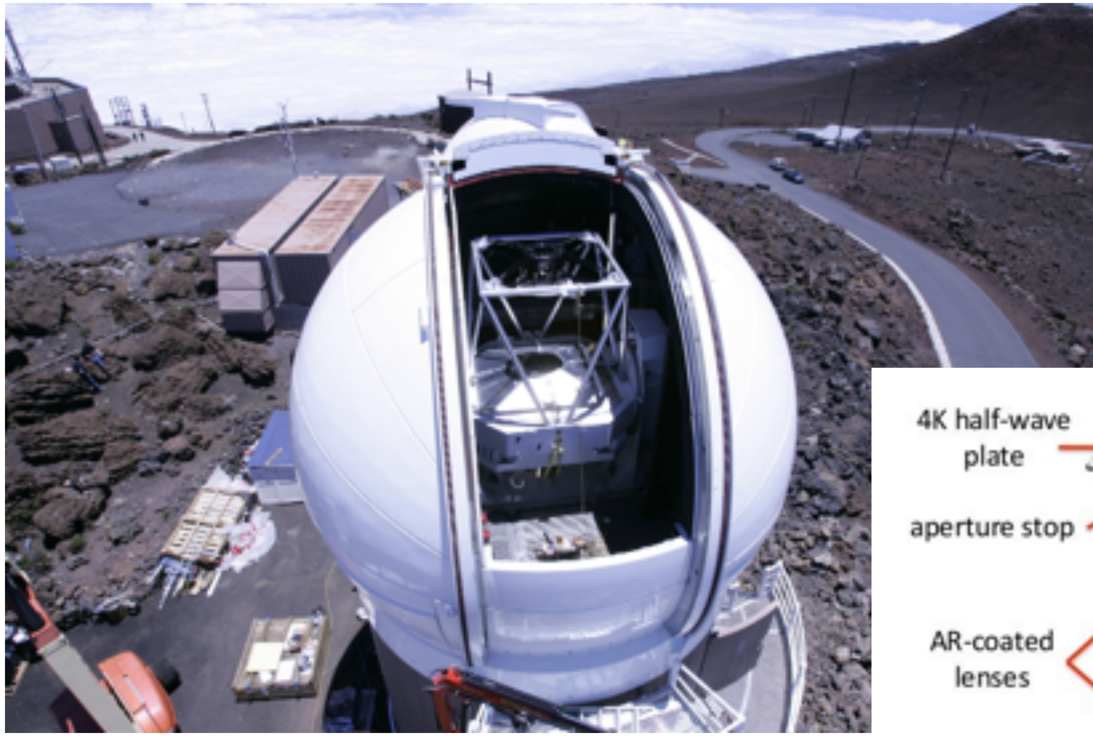
1.5PB!

Pan-STARRS

Panoramic Survey Telescope & Rapid Response System

Detects near Earth asteroids and transient celestial events

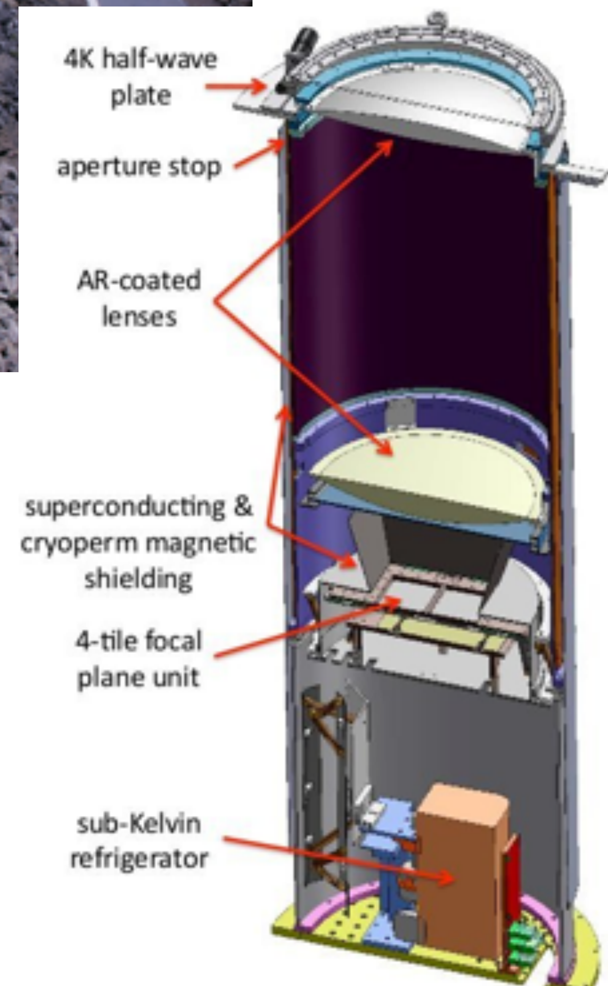
30T / night!



BICEP

Background Imagining of Cosmic Extragalactic Polarization

Studies the polarized light from the Cosmic Microwave Background



BICEP2/KECK INSERT

Research Computing powers computation behind BICEP2 discovery

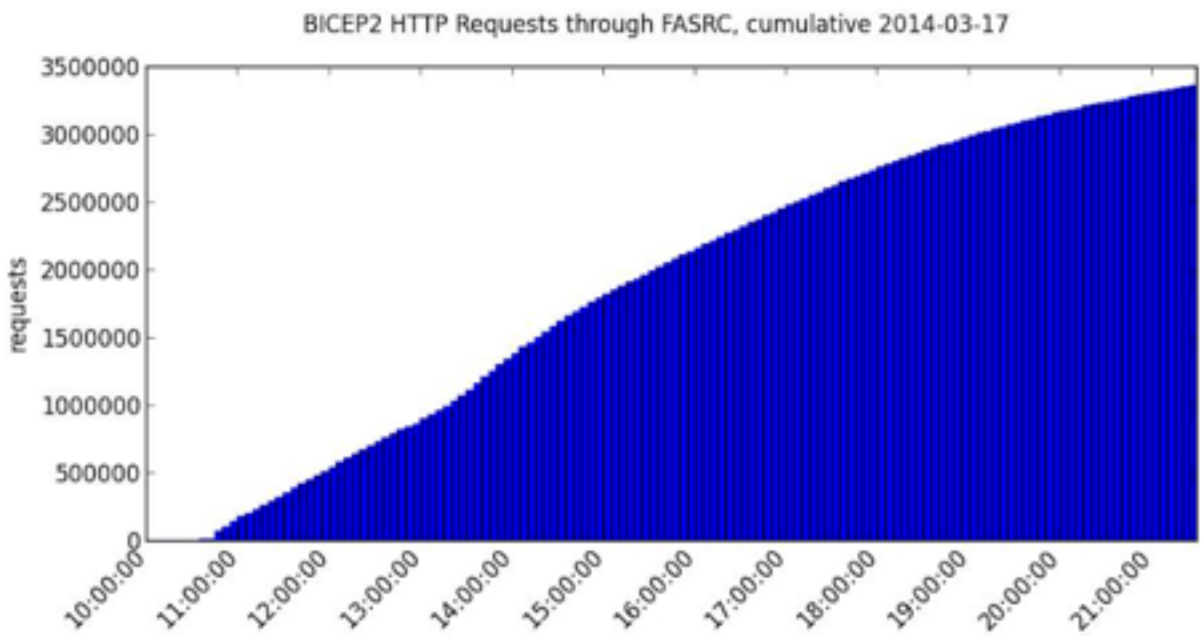
Research Computing has provided computational support and infrastructure for the BICEP2 project. Research Computing provided the BICEP2 project with 400TB of storage space and access to more than 30,000 cores of compute.Research Computing also provided virtual machine support for BICEP2 websites and data. The project consumed 5.1 million CPU hours on Odyssey since 2010.

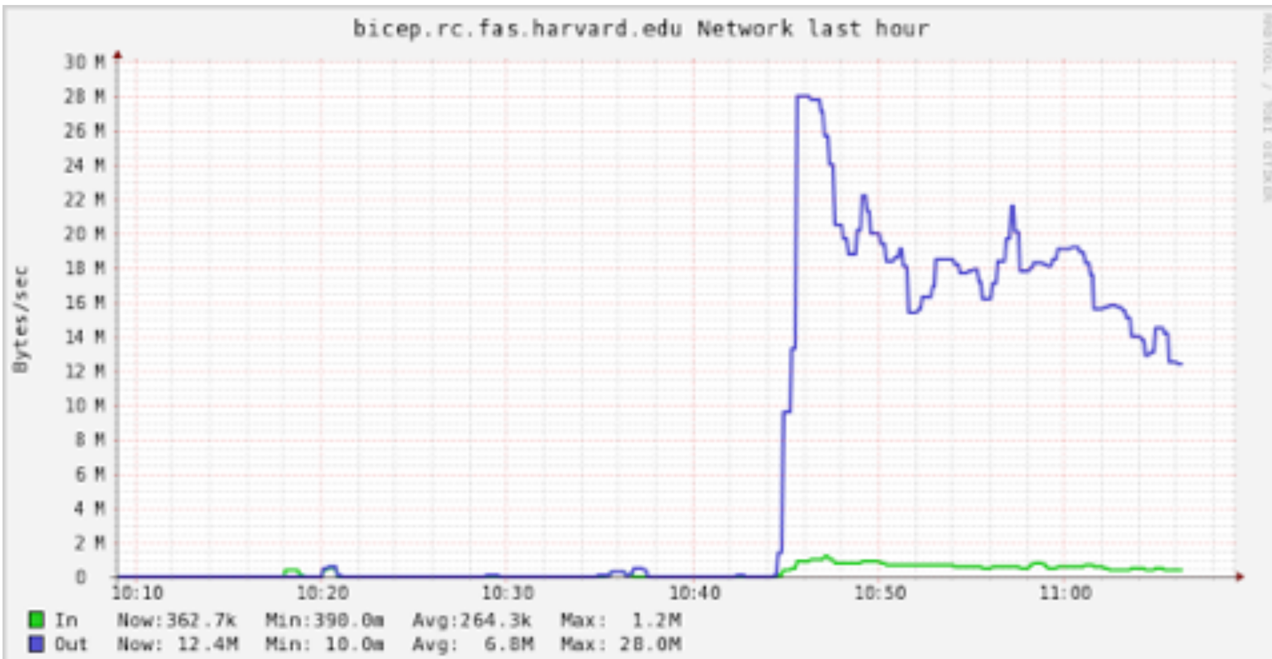
Of those 5.1 million CPU hours, close to 3 million were computed at the new environmentally sustainable MGHPCC facility in Holyoke, MA. MGHPCC compute, powered in part by hydro electricity, came online during the summer of 2013 and was the result of a collaborative effort between Harvard, four other research universities, and the Commonwealth of Massachusetts. The BICEP2 project was one of the first research groups granted access to the Research Computing cores hosted at MGHPCC. As part of the initial test group, the BICEP2 project had sole access to 28,000 cores for four days. Harvard's involvement in MGHPCC is a centerpiece of its commitment to greening IT.

Once testing was complete in August, the project continued accessing Research Computing's MGHPCC compute, closing in on 3 million CPU hours consumed over five months. In that time, the BICEP2 project processed an almost equivalent amount of CPU hours as it had over the previous three years using other Research Computing cores.

The project, led by John Kovac, Associate Professor of Astronomy and Physics and member of the Harvard-Smithsonian Center for Astrophysics, announced detection ofB-mode polarization and gravitational waves providing insight into the universe's first moments after the Big Bang. Albert Einstein's general theory of relativity hypothesized the existence of gravitational waves, but until now they were never physically observed.

The project examined Cosmic Microwave Background (CMB), the oldest light in the universe, to discover signatures of cosmic inflation by detecting Cosmic Gravitational-Wave Background (CGB). The theory of cosmic inflation states that the universe underwent a violent and rapid expansion at only 10^-35 seconds after the Big Bang. Inflation, during the first moments of time, produced CGB, which in turn imprinted a faint but unique signature in the polarization of the CMB. It is this faint signature which led scientists to their discovery and the first solid evidence that inflation had taken place.


BICEP2 HTTP Requests through FASRC, cumulative 2014-03-17
requests
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
10:00:00
11:00:00
12:00:00
13:00:00
14:00:00
15:00:00
16:00:00
17:00:00
18:00:00
19:00:00
20:00:00
21:00:00


bicep.rc.fas.harvard.edu Network last hour
Bytes/sec
30 M
28 M
26 M
24 M
22 M
20 M
18 M
16 M
14 M
12 M
10 M
8 M
6 M
4 M
2 M
0
10:10
10:20
10:30
10:40
10:50
11:00
In Now:362.7k Min:390.0m Avg:264.3k Max: 1.2M
Out Now: 12.4M Min: 10.0m Avg: 6.8M Max: 28.0M

From: James Cuff
Date: Fri, Aug 23, 2013 at 11:32 AM
Subject: Fun friday fact.
To: Mark Vogelsberger, Shy Genel, Dylan Nelson
Cc: Lars Eric Hernquist

Hey team,

Since Tuesday you guys racked up over 28% of the new cluster, which combined is over 78 years of CPU in just three days and it is still only just Friday morning :-)

This is pretty awesome!

Happy friday!

```
--------------------------------------------------------------------------------
Top 10 Users 2013-08-20T00:00:00 - 2013-08-22T23:59:59 (259200 secs)
Time reported in Percentage of Total
--------------------------------------------------------------------------------
  Cluster     Login     Proper Name         Account       Used
--------- --------- --------------- --------------- ----------
  odyssey mvogelsb+ Mark Vogelsber+   cluster_users     16.90%
  odyssey   lsironi  Lorenzo Sironi   cluster_users      9.34%
  odyssey    sgenel       Shy Genel   cluster_users      7.84%
  odyssey    yang12     Darren Yang   cluster_users      5.67%
  odyssey   sstokes    Sarah Stokes   cluster_users      4.55%
  odyssey   mwalker  Matthew Walker   cluster_users      4.15%
  odyssey   dnelson    Dylan Nelson   cluster_users      3.22%
```

NBA drafts Big Data

Harvard researchers have used Odyssey to dig deep into NBA player data, creating a new statistical framework for basketball analytics. The research, led by Kirk Goldsberry, Visiting Scholar at the Center for Geographic Analysis, Luke Bornn, Assistant Professor in the Department of Statistics, Dan Cervone, and Alex D'Amour both PhD students in the Department of Statistics, uses player data from the 2012-2013 NBA season. The dataset, known as SportVU, was collected at 14 NBA arenas and contains 800 million locations of NBA players on the court.

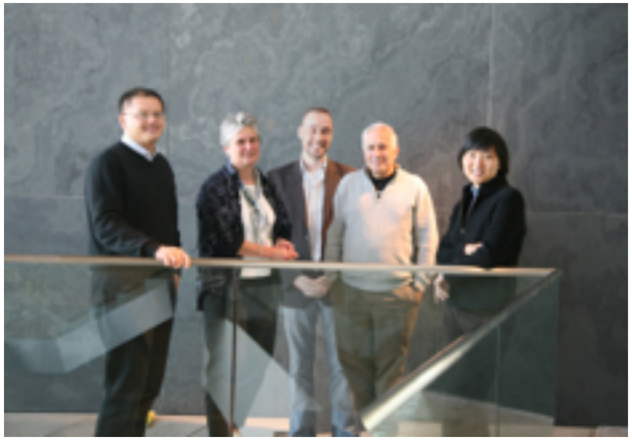
To make sense of this data, Cervone and D'Amour proposed the theory of assigning a value to each basketball possession. If all possessions could be valued, a model could be designed using the SportVU data with metrics such as the locations of players, player scoring abilities, player ball possession, player court position, and player ball handling. Running this type of statistical model would provide analysts with a scientific assessment of "expected possession value" or EPV. Player performance could be statistically quantified at any point in the game. Coaches could use this information to adopt specific strategies for specific players at specific times.

With the model in place, researchers turned to Research Computing's Odyssey cluster for computation. The database researchers built totaled 93 gigabytes. A full analysis of this database required 500 parallel processors and two terabytes of memory. Without the computational power of Odyssey, the analysis of such a large dataset would have been impossible outside of the cluster environment.

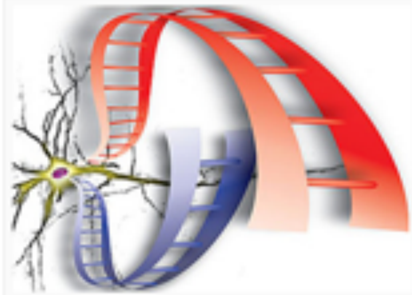
The results from the computational run were what most NBA fans would expect. Chris Paul, point guard for the Los Angeles Clippers, had the highest EPV with 3.48 points added per game. According to the researchers, this meant the Clippers were expected to score 3.48 more points per game because Paul controlled the ball on offense. Ricky Rubio, point guard for the Minnesota Timberwolves, had the lowest EPV with -3.33 points "added" per game. Because Rubio is a poor shooter, each time he takes a shot it would be statistically preferable if a teammate took the shot instead. While Rubio's ball handling skills do add value, his overall EPV is reduced because of shooting weakness.

As datasets grow in size, complexity, and importance, the NBA will not be the only organization looking to high performance computing as a way to measure and model value. What the Harvard researchers essentially revealed is with the right model and numerous useful data points, anything can be scientifically quantified and potentially transform our understanding of the world around us.

The Conte Center



Genomic Imprinting



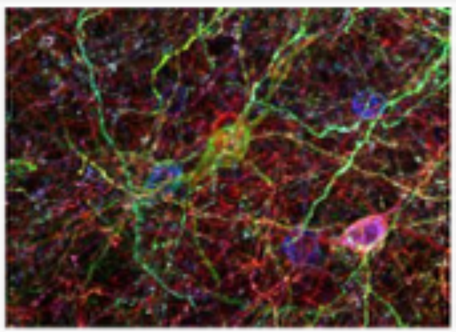
Maternal vs. Paternal DNA
Our genomes consist of chromosomes inherited from our mothers (red DNA helix) and fathers (blue DNA helix). Parent-of-origin specific gene expression, known as "genomic imprinting," is important for the development and function of neurons.
Courtesy of Catherine Dulac

For most genes we inherit two copies, one from mom and one from dad. Often these two copies are expressed similarly—meaning just as much protein or RNA product is made from the mom's copy as the dad's copy. But some genes are expressed differently depending on whether they are maternally or paternally inherited.

This phenomenon, called genomic imprinting, is important in many biological processes, including brain development. In fact, entirely different neurological disorders can arise when one loses the maternal versus paternal copy of the same gene.

Until recently, fewer than 100 imprinted genes were known. In 2010, the Dulac lab identified over 1300 candidate genes that undergo differential expression from either the maternal or the paternal allele in the mouse brain. Many of these appear to be related to brain development and plasticity. As part of the Conte Center, the lab is now gathering data on the "imprintome"—the full collection of imprinted genes—of PV-cells, a class of inhibitory neurons believed to be particularly vulnerable in mental illness. The lab is also investigating how the PV-cell imprintome is altered in mouse models of early life stress and mental illness.

Brain Plasticity



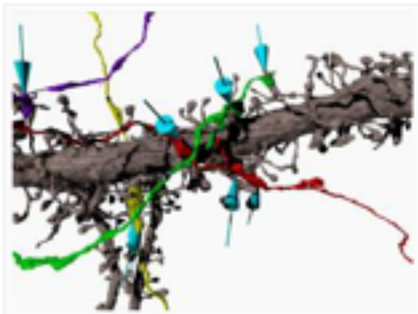
Inhibitory Neurons Controlling Brain Plasticity
A class of inhibitory interneurons called "PV-cells", which orchestrate the timing of critical periods of brain development. Here PV-cells of the mouse cortex have been labeled using Brainbow technology.
Courtesy of Luke Bogart, Dawen Cai, Jeff Lichtman & Takao Hensch

Much of our adult behavior reflects the neural circuits sculpted by experience in infancy and early childhood. At no other time in life does the surrounding environment so potently shape brain function – from basic motor skills, sensation or sleep to higher cognitive processes like language. How this plasticity – or ability of the brain to change – waxes and wanes with age carries an impact far beyond neuroscience, including education policy, therapeutic approaches to developmental disorders, and strategies for recovery from brain injury in adulthood.

Windows of heightened plasticity in the course of brain development are called "critical periods." In 1998, Professor Hensch and colleagues achieved the first direct control over critical periods in the visual system, delaying or accelerating the critical period responsible for balanced representation of left and right eye inputs in the visual cortex. Since then the Hensch lab has identified pivotal brain molecules, cells, and circuits that orchestrate critical periods and rewire neural connections in response to environmental experience—particularly early sensory experience.

As part of the Conte Center, the lab is now using behavioral and electrophysiological readouts to characterize critical periods in the maturation of prefrontal circuits. The focus is on the role of one type of inhibitory neuron within these circuits, the PV-cell—as PV-cells are thought to control the timing of critical periods and exhibit defects in psychiatric disorders such as autism and schizophrenia. The role of PV-cells in prefrontal development and plasticity will be studied both in normal laboratory mice and mouse models of early life stress or mental illness, especially in relation to fear and anxiety behaviors. This functional data will ultimately be integrated with morphologic and genetic data from the Connectome and Imprintome projects carried out by the Lichtman, Zhuang, and Dulac labs.

Connectomics



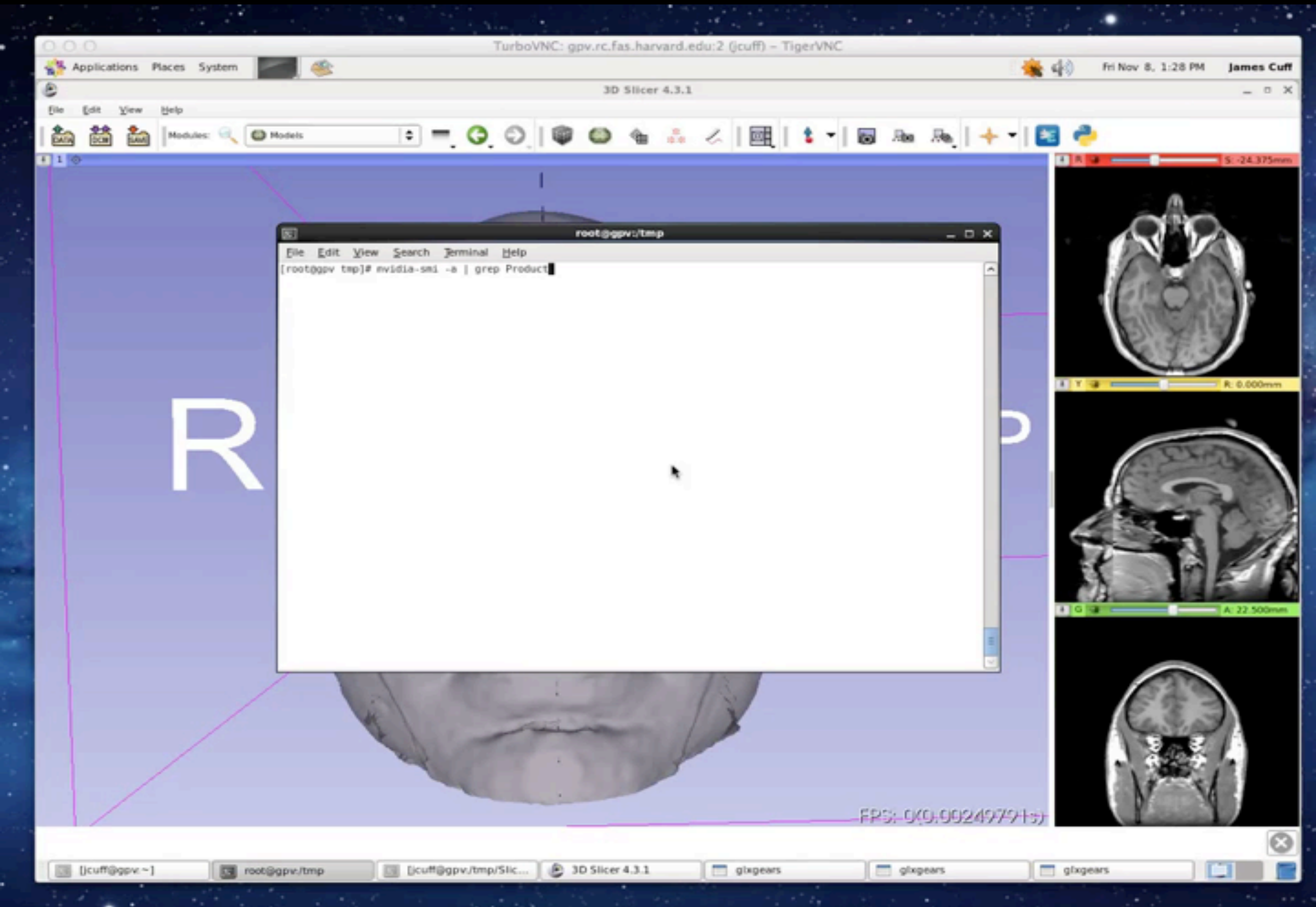
A Tiny Piece of the Connectome
Serial electron microscopy reconstruction of axonal inputs (various colors) onto a small segment of the apical dendrite (grey) of a pyramidal neuron in mouse cerebral cortex. The grey protrusions are dendritic spines. The arrows mark functional synapses, based on the presence of neurotransmitter-containing vesicles.
Courtesy of Bobby Kasthuri and Jeff Lichtman

The cerebral cortex of the human brain contains more than 160 trillion synapses, or sites of communication between neurons. Each neuron receives synaptic inputs from hundreds or even thousands of different neurons, and each sends outputs to a similar number of target neurons, spread out over a large distance. Thus, deciphering the "connectome," or complete connectivity diagram, of even one type of neuron in the cortex poses enormous challenges.

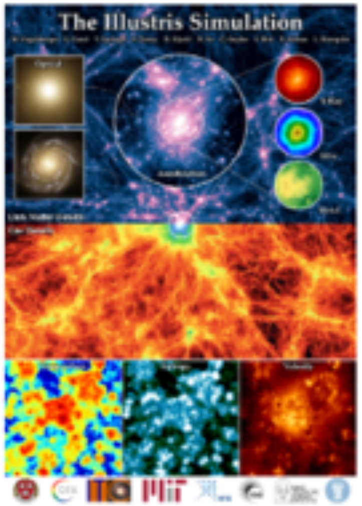
The Lichtman lab has been developing novel imaging methods to overcome these challenges. One new method is the "Brainbow" mouse, in which various combinations of fluorescent proteins are stochastically expressed in neurons—resulting in labeling with over a hundred unique hues. This rainbow labeling strategy allows numerous densely-packed extensions of neurons to be clearly resolved from one another. A complementary approach that the lab has optimized significantly in recent years is serial electron microscopy reconstruction of brain tissue.

As part of the Conte Center, the lab is using these, as well as viral tracing methods, to visualize the connectome of PV-cells in the prefrontal cortex, believed to be particularly vulnerable in mental illness. Of particular interest are changes in the PV- cell connectome across normal cortical development, and alternations in the PV-cell connectome in mouse models of early life stress or mental illness.

Interactive video of an fMRI from a data center in western MA connected to my desktop in Cambridge…


TurboVNC: gpv.rc.fas.harvard.edu:2 (jcuff) - TigerVNC
Applications
Places
System
Fri Nov 8, 1:28 PM
James Cuff
3D Slicer 4.3.1
File
Edit
View
Help
Modules:
Models
root@gpv:/tmp
File
Edit
View
Search
Terminal
Help
[root@gpv tmp]# nvidia-smi -a | grep Product
R
[jcuff@gpv:~]
root@gpv:/tmp
3D Slicer 4.3.1
glxgears
glxgears
glxgears

Simulations run on Odyssey lead to first realistic virtual universe



Astronomers have created the first realistic virtual universe by running large-scale cosmological simulations on supercomputers. The project, known as Illustris, was led by Mark Vogelsberger of MIT and the Harvard-Smithsonian Center for Astrophysics. Illustris produced detailed galaxy simulations by using complex computer models that captured the physical components and processes of the universe from 12 million years after the Big Bang to present day, spanning over 13.8 billion years of cosmic evolution. The simulations contained tens of thousands of galaxies captured in high-detail, covering a wide range of masses, rates of star formation, shapes, and sizes. Researchers also included star-formation-driven galactic winds and black hole thermal energy injection throughout cosmic history to ensure their models were comprehensive enough to produce realistic virtual galaxies.

In order to produce these simulations, the researchers developed their own computer code AREPO. The code was written to simultaneously run on tens of thousands of computer cores. It took three months for the calculations to run, and used a total of 8,000 CPUs running in parallel. Odyssey was used to run several lower-resolution versions of the Illustris virtual universe, as well as hundreds of test simulations from which the computer models were developed.

Once the simulations were complete, the final product was compared to images of the Universe taken by various telescopes, including the Hubble Space Telescope. Researchers found many similarities between the two, though some discrepancies were also discovered. According to Vogelsberger, these discrepancies were ripe for further study and investigation and would potentially lead to new understandings of how the universe evolved.

For further reading see the complete results in Nature or visit the Illustris project website. And check-out the below capstone animation of the simulation.

Video of the evolution of the Universe since a few moments after the big bang…

Is scale out computing essential?

oh YES!

very much YES!

@jamesdotcuff

http://rc.fas.harvard.edu