Welcome Galaxy Czars!!

(and wannabes … like me)

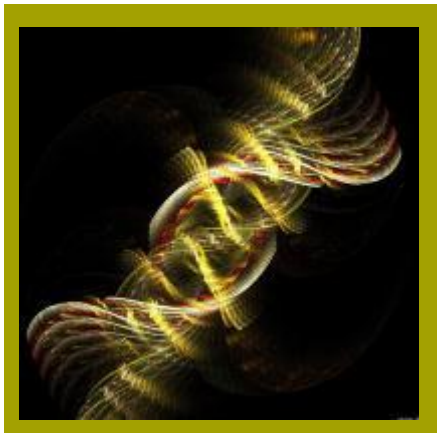
- A few reminders:
  - Technical Problems? please use the chat room for help. We will do our best.
  - To open up your audio line, you must select the "Talk" button on your left hand side. There can only be 6 simultaneous talkers at once, so please leave it unselected unless needed.
  - When talking, please let us know your name & where you are from ☺
  - The call will be recorded & posted for playback later on. This includes all chats!
  - We might use the polling features of Blackboard – check out voting options on your left hand side.
  - Feel free to play around and get used to the interface.
  - Have you checked out the wiki page? Survey results are posted:
    - http://wiki.g2.bx.psu.edu/Community/GalaxyCzars

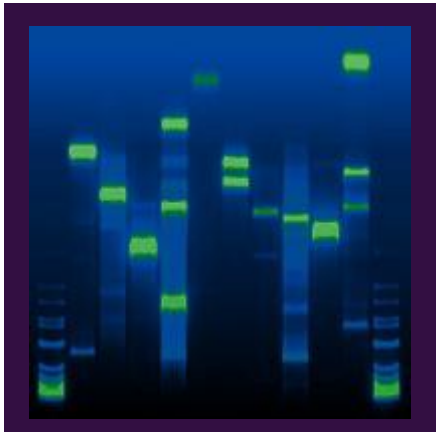
Web Conference Agenda

- Logistics: Address how we want to tackle these calls. Go over generic agenda. Frequency of calls
- Group Goals: what do we want to accomplish with this group beyond discussions/sharing?
- Presentation: Galaxy at Iowa: Discuss our issues with big data, how NGS tools take in/output data and finding the right storage server solution.
- Open Mic & recruit new volunteers for the next call.
- Break out at Galaxy Community Conference

Proposed Generic Agenda for future calls

- 20 min: Galaxy in Our Town. - presentation from a local galaxy institution on what they are doing or a problem they are troubleshooting - or have someone walk through their use cases and pain points.
- 20 min: Galaxy Today/Tomorrow. - presentation on a galaxy coding item. Either from penn state team, or from someone working on a new feature or customization.
- 20 mins: Open Mic Discussion & make point to point connections. IE - organize smaller breakouts if someone wants to host a call specific to an issue.





University of Iowa Custom Galaxy Deployment

And the tale of big data storage …

Ann Black-Ziegelbein
annblack@eng.uiowa.edu
Senior Application Developer
University of Iowa
Iowa Initiative in Human Genetics
Center for Bioinformatics and
Computational Biology

Glenn Johnson
glenn-johnson@uiowa.edu
Senior System Administrator
University of Iowa
High Performance Compute Cluster

Ben Rogers
ben-rogers@uiowa.edu
Director of Research Computing
University of Iowa
High Performance Compute Cluster

Galaxy @ Iowa

Background:

Why, Who, Where, What

Why Bring Galaxy to Iowa

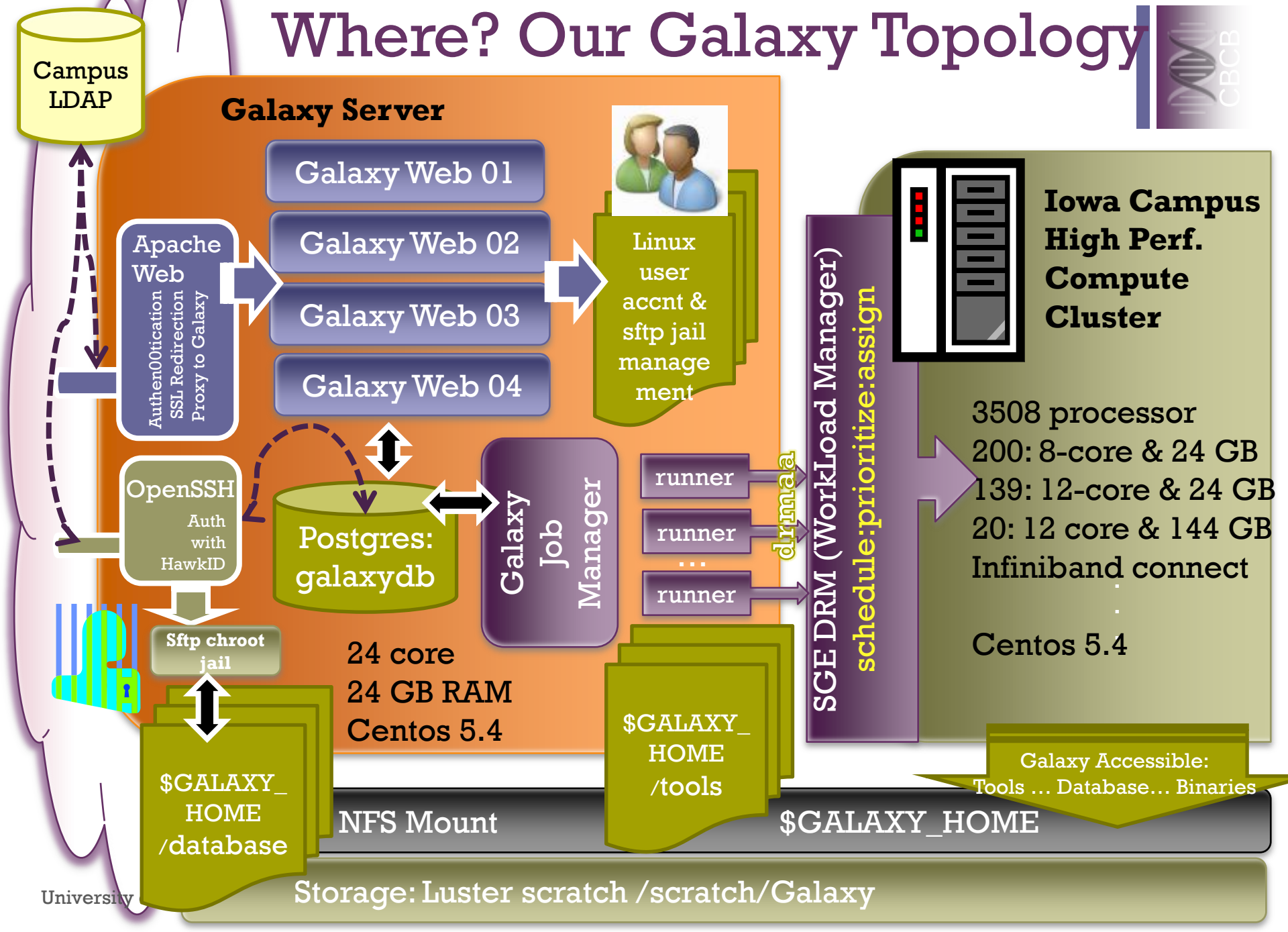
- **Why host a local Galaxy deployment?**
  - Customizable deployment, for example:
    - Control version of tools
    - Tune how tools are exposed
    - Expose additional custom tools and databases
  - No tight data quota caps
  - Local data storage and transfer

- **What stage is our galaxy deployment?**
  - Alpha Deployment Phase, but It's Alive!!!!!!!!
  - Iowa Campus/ University of Iowa Hospital & Clinics access only
  - Hosted with our campus HPC and dispatching jobs to SGE compute cluster
  - Sub-set of tools & reference genomes exposed compared to the public Galaxy deployment
    - Capable of all public Galaxy server functionality
    - Lazy exposing additional tools/reference indices/etc. Upon request.
  - Can view local Galaxy datasets
    - In IGV
    - Iowa Local UCSC browser
  - Published Human exome analysis pipeline is available

Who is using Galaxy @ Iowa

- Have about 50 "alpha user" accounts created.
  - But only 10-12 active regular users
  - Have wiki support as well as a local users listserv for questions/problems
  - JIRA ticket tracking of problems.
- Use case scenarios we expect to support:
  - Core Published Workflows
    - Common, well tested, workflows will be built and published by our bioinformatics core
      - RNA Seq analysis
      - Exome analysis
      - Etc.
  - Experimenters
    - Researchers trying out tools on data and using as a sandbox
  - New Tools
    - Exposing research tools from our bioinformatics research teams
- Future Publicity:
  - Trying to hold off until we can work through some remaining issues
  - Will be holding a bioinformatics short course all on Galaxy @ Iowa

Where? Our Galaxy Topology



What has Iowa Customized for Galaxy

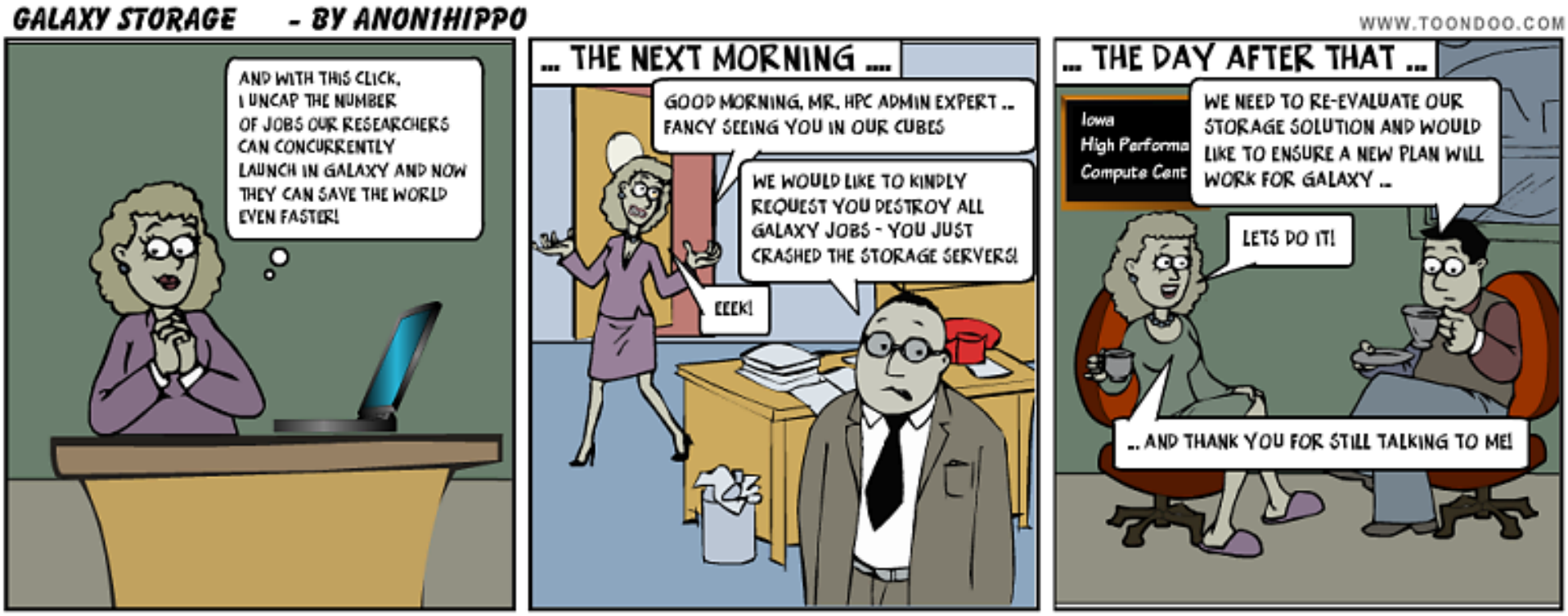
- Linking & autocreating linux accounts from Galaxy user accounts
  - Support sftp transfers (no ftp)
  - Watcher scripts to dynamically change file perms as necessary
- Tool customizations
  - New University of Iowa tools
  - Tweaks to existing tools for our HPC (such as threads, memory, etc)
- Expose a custom to/from configuration for emails
  - Support of auto-ticket opening in JIRA from Galaxy
- Moved workflows to top of tool navigation bar, limited tools exposed to subset

Galaxy @ Iowa

The Problem:

What to do when your storage solution fails.

How we got into this mess …



Iowa & Lustre: Background

- Our existing HPC Storage Solution was Lustre
- What the heck is Lustre?
  - From Wikipedia: "**Lustre** is a parallel distributed file system, generally used for large scale cluster computing"
  - In theory it should scale well and be fast under large load
- How Iowa had Lustre configured
  - Hardware: HP MSA 2312sa P2000 G2
  - MDS (Lustre Meta Data Server)
    - 2 enclosures with 12 300G SAS drives configured as a RAID-50 volume
  - OSS (Lustre Object Storage Servers)
    - 4 sets with 4 enclosures consisting of 48 1T SATA drives
    - Each enclosure configured as a RAID-6 (10 data + 2 parity disks)

Analysis: How Lustre failed us

While performance could be good in spurts Iowa's configuration was not ideal.

- Lustre sends bulk IO in 1024K chunks but the underlying array was not aligned, having a stripe width of 640K. This meant the servers needed to work harder with all types of extra writes.
  - **Typical NGS tooling performed large volume of streaming writes which seemed to exacerbate the situation**
- The OSS servers would ramp to a high load, often >200.
- The OSS machines would become unresponsive
- Job IO would slow and sometimes cause user jobs to die
- Sometimes the lustre servers would evict the clients because the response times were too long. This would kill user jobs as well.

And then what happed.

- With more user load on Iowa's HPC (not just Galaxy) it was decided to look at different storage solutions in addition to a more optimal Lustre configuration
- Galaxy moved off of Lustre storage and onto a ZFS storage box
  - Reconfiguring Lustre meant destroying all data
- Decided to compare the following storage types for Galaxy:
  - Existing ZFS Storage
  - Gluster (not to be confused with Lustre)
    - "is a distributed/parallel/virtual filesystem. It lets you aggregate the capacity and performance of multiple local filesystems on multiple servers and present those files in a single unified view or namespace." – from community. gluster.org
  - Lustre reconfigured

Galaxy @ Iowa
Benchmarking
New Storage Solutions

Not really Apple to Apples….

- What was benchmarked & compared:
  - NFS/ZFS w/ 2 Gb Ethernet
  - NFS/ZFS w/ 4 Gb Ethernet
  - Gluster w/ Infiniband over Ip
  - Sniff tested and aborted:
    - ZFS w/ infiniband
    - Lustre
- Did ZFS really have a chance??
  - We knew ZFS would not scale as well. It is not distributed nor using infiniband.
  - The question was not: "Would ZFS scale for all HPC user access?"
  - The question was: "How quickly would ZFS performance degrade and would it be good enough for our Galaxy goals?"
  - ZFS has cheap expansion, easy to manage, and minimal client CPU, and we already had the box & 50 TB of ZFS Storage!

Benchmarking Storage Solutions

- Benchmarking is always a work in progress … Never quite completed.
  - And I am not claiming to be a performance expert.
- While we are at it … also looked out how to optimize the jobs on each storage architecture
- Our Benchmark Test:
  - BWA sampe paired end read alignment step on Human Exome data.
  - Variations:
    - Input:
      - 1. LI: copy to local disk for local reads
      - 2. RI: leave on remote storage for remote reads
      - 3. SI: have aln input local, fastq input remote
    - Reference indices:
      - 1. LR: copy compressed archive to local disk, extract & reference local
      - 2. RR: leave on remote storage for remote reference
      - 3. 2R: leave on secondary remote storage (ZFS)
    - Output:
      - SO: streaming writes to remote storage
      - CO: streaming writes to local disk, copy to remote upon finish
  - Tested 1,30,60,90 concurrent clients
    - Would liked to have done more … but competed with others for cores, and time investment

Benchmarking Challenges

- Clean consistent runs.
  - Keeping people away from storage during test runs
  - Reproducible numbers …
- Getting cores on the cluster
  - Competed with all users for benchmarking time.
- Understanding failures, analyzing results, babysitting scripts, writing scripts, re-writing scripts …
  - Yeah, it is time consuming.
  - https://bitbucket.org/iihgbiocore/ngssgeloadgenerator/overview
- Not claiming our benchmarking is perfect, but it gives us data to work with.

Galaxy @ Iowa: Our goals

- Our targets:
  - Minimum: Good performance for 60 concurrent clients
  - Great: Good performance for 100 concurrent clients
  - Yowzah!: Good performance for 150 concurrent clients
  - Also: good performance without having to tweak all Galaxy tool wrappers to make things optimal.
  - Oh yeah: it must be stable.

Drum roll … The Results

Taking a closer look …

- The raw data:

Avg. Job Run Time in Seconds				
	Num Clients			
Test Variation	1	30	60	90
zfs_4Gb_RIRRSO	1517	2487	4470	6671
gluster_IB_RIRRSO	3077.00	3640.00	7014.00	7576.00
zfs_4Gb_RILRSO	1567	2631	4538	6362
gluster_IB_RILRSO	1568.00	1979.00	3433.00	5266.00
gluster_IB_RI2RSO	1511.00	2206.00	2442.00	3360.00
gluster_IB_SI2RSO	1495.00	2080.00	2146.00	4447.00
zfs_4Gb_LILRSO	1760	3282	7500	9133
gluster_IB_LILRSO	1980.00	2285.00	2453.00	2642.00
gluster_IB_SILRSO	1607.00	0.00	2580.00	6424.00
zfs_4Gb_LILRCO	3222	4405	6934	10656
gluster_IB_LILRCO	1998.00	3014.00	3390.00	4242.00
gluster_IB_SILRCO	1872.00	2450.00	3601.00	3583.00
gluster_IB_SI2RCO	1843.00	2292.00	3049.00	4002.00

Saved Time By off loading Reference indices (gluster)

Moving input to local or split may not have much effect, hurt zfs?

Writing local and then copying seems to help gluster, hurt zfs ?

*** Remember … the savings adds up over a pipeline of 20+ steps!**

The take aways …

- In general, gluster scaled better

Avg. job run time aggregated over all test variations (seconds)				
	1	30	60	90
gluster	1921	2449	3396	4326
zfs	1860	3626	6473	7810

- Jobs on gluster were more sensitive to test variation
  - Off loading reference Indices made largest impact
  - **With some tuning, got job run time to decrease from ~7500 seconds to ~3500 seconds for 90 clients!!**
- Would like to run ZFS "worst case" (RIRRSO) again as the numbers appear better than expected.
- Are you a details person? You can download a zip of all the reports generated from metrics collected during the test runs:
- https://dl.dropbox.com/u/90475071/bwa-sampe_reports_pdf.zip

What happened to Lustre?

- Lustre & HP MSA might not be a good marriage
  - From discussions in the community Lustre likes certain hardware
- On a 5GB dd test, using a single raid array we saw the following:
  - Gluster – 600MB/s
  - Lustre – 60MB/s (with additional tuning got closer to 600MB/s)
  - NFS – 90MB/s
- Lustre is more expensive and harder to manage
  - Although also more flexible.
- Did not do additional BWA benchmarking since we could not get the raw performance out of it we wanted to make it worth our time

Choosing Storage: Why Gluster?

- Strong performance with scalability
- Cheap expansion. Can add in more hard drives as additional "bricks" as needed.
- Easy to setup and manage.
- Gaining industry traction. Redhat purchased & putting weight behind.
- The Cons:
  - Glusterfsd client process can consume a core on a compute node. This can cause resource contention.
    - Simulated failure by attaching gdb to client glusterfsd process, this affects the client workload, but not the rest of gluster.
  - Failed disks, servers or gluster server processes can cause full gluster file system outages.
- What about ZFS?
  - Would like to benchmark 10Gb ethernet w/ ZFS
  - Current plan: use for reference indices and data libraries
  - Also might be used as a place to hold more "permanent" data which eventually might get archived

Iowa's plan for managing Storage

- Galaxy storage will not be backed up.
- No hard quota caps will be in place … for now
- Auto – cleanup datasets older than 30 days
  - Email will be sent to users in advance allowing them to take action
- Wait, Monitor, and Iterate on this plan.
  - Talk to us in 6 mons and it could be different.

Our local Galaxy: What's next?

- **Our future goals**:
  - Wrap up data storage evaluation & migrate
  - InCommon federated identity management
  - Develop and publish more common workflows
  - Expose additional bioinformatics tools and Galaxy features based on community feedback
  - Integrate with DNA Core sequencers to directly provide sequencer data into Galaxy for analysis
  - Lots of little fit & finish items
  - Multi-version tool support
  - Version tracking of tools and workflows. We need a clear audit trail.

Upcoming Events

- Bioinformatics Short Course August 1-3, 2012
  - **Mutation Detection Using Massively Parallel Sequencing: From Data Generation to Variant Annotation**

  Massively parallel DNA sequencing technologies have ushered in the next wave of the genomics revolution. The clinical application of these technologies will make personalized genomic medicine a reality; in the research laboratory, these technologies are making innovative approaches to genome-wide investigations routine. In the summer of 2012, the Iowa Institute of Human Genetics will offer a Bioinformatics Short Course. The course will focus on most popular next-generation sequencing platforms - the Illumina HiSeq and MiSeq systems. Upon completion of this course, participants will understand the design of a next generation sequencing experiment and the workflow to achieve a particular result. A series of lectures to introduce basic concepts will be interwoven with practical examples of data generation and analysis. Hands-on sessions will enable participants to analyze data from its generation to interpretation. Participants will be required to bring their own laptops. Enrollment is limited.

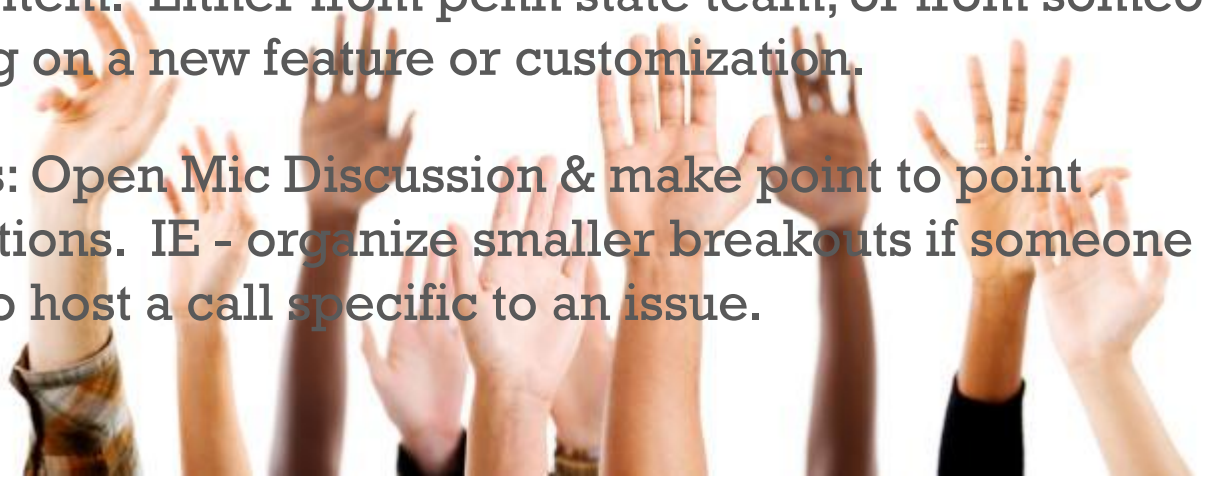
  - **Course will leverage Iowa's Galaxy Deployment**
  - http://www.medicine.uiowa.edu/humangenetics/bioinformaticscourse/
  - Brought to you by:
    - Iowa Initiative in Human Genetics

Please share with us your experiences with storage – we would love to hear about & learn from them!!

Our Next Call – We Need You!

Volunteer, Please!!!

- 20 min: Galaxy in Our Town. - presentation from a local galaxy institution on what they are doing or a problem they are troubleshooting - or have someone walk through their use cases and pain points.
- 20 min: Galaxy Today/Tomorrow. - presentation on a galaxy coding item. Either from penn state team, or from someone working on a new feature or customization.
- 20 mins: Open Mic Discussion & make point to point connections. IE - organize smaller breakouts if someone wants to host a call specific to an issue.



Backup:
More Info on Gluster

GlusterFS

GlusterFS is a distributed file system that works over IP over Infiniband. Same hardware as lustre

- Relies on the backend file system for metadata, allocation, etc.
- Gluster volume metadata is managed by the clients.
- Can achieve higher throughput than lustre with less disks

GlusterFS

- RAID stripe alignment with the file system is managed vis XFS
  - mkfs.xfs –f –d su=64k,sw=10 –l su=64k /dev/dm-0
- This gives consistent write throughput of 500-600 MBps for all file sizes per brick
- The read throughput is about 600-700MBps per brick

GlusterFS

- The aggregate bandwidth from IOR with 32 1G files is

Max Write: 6388.65 MBps

Max Read: 9368.44 MBps