Hall A & C Computing Tips and Tricks

Brad Sawatzky





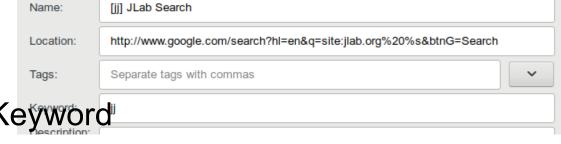
First up: A Couple Quick Tricks to make your Computing Work Suck Less





How to find information

- JLab's web search is ... weak ...
 - →So, do this in Firefox:
 - » Go to www.google.com and search for 'site:jlab.org foo'
 - » Right click on the bookmark and choose 'Properties'
 - Give it a good name
 - Give it a short 'keyword' like 'jj'
 - Clean up the Location as shown, replace 'foo' with %s
 - → Now type 'jj jget' in URL bar
 - » %s in 'Location'
 string is replaced
 with text following Keyword



» 'site:jlab.org' is google-fu to restrict search to jlab.org domain





How to find information

- Trick works great for many things
 - → JLab staff page (https://misportal.jlab.org/mis/staff/staff.cfm)
 - » Keyword: 'page'
 - » Location (can extract from search on 'smith' above):
 - » https://misportal.jlab.org/mis/staff/staff.cfm?field=all&name= %s&Search.x=36&Search.y=11&Search=Search&field=all
 - →ROOT / G4
 - » Keyword: 'gr'
 - » Location:

https://www.google.com/search?hl=en&btnG=Search&q=site:cern.ch%20%s

- →Stackoverflow.com
- →JLab Logbook (a little trickier, but you can work it out)
- \longrightarrow ...





How to work from Offsite

- How to work from offsite without tearing your eyes out because, holy hell, the graphics and menus are just so slow...
- VNC + ssh tunnel to the rescue
 - → VNC: Virtual Network Computing
 - →ssh used to securely move VNC traffic through jlab firewall



- Computer Center How-to
 - → https://cc.jlab.org/accessingvnc
- Old 'howto' I wrote for my collaboration
 - →adapt to machine you use
 - →Search: 'jj vnc session'
 - → https://hallaweb.jlab.org/wiki/index.php/Ho w_to_connect_to_a_d2n_VNC_Session





Offline Analysis Farm Usage / General JLab Computing





Nuts to the Farm, I analyze on my Desktop

- Simple tasks, some analysis OK on the desktop, BUT!!
 - → Thou shalt backup your code!
 - → Thou shalt backup your results!
 - → Who among us has done % rm -rf stuff/
 - » Followed by !@#\$?
- Don't keep only copies on your laptop
- Don't keep only copies on your desktop's hard drive
- Do use git for all code and scripts!
 - → Commit early, commit often
 - → 'git push' often too!
 - » It's a backup!

- Hard drives die and the data are gone.
 - → Drives are large and cheap
 - → But reliability on consumer drives is worse than it used to be!
 - →SSDs are (weirdly) no better!
- IF your hard drive died today, how long would it take to recover?
 - » a day, a week,
 - » a month????





JLab Systems can help!

- /home, /group are automatically backed up
 - →They are snapshotted hourly!

```
% cd .snapshot/
% ls -lrt
```

→Longer term backups are on tape

- /work, /volatile are on heavily redundant filesystems
 - →NOT backed up
 - » Use tape
 - →More on this later...
- NOTE: Your JLab RHEL system on site can mount these directories if needed
 - →Talk to me if this would help





The JLab Farm • Power at your Fingertips

- Farm has many pieces
 - →~14500 compute nodes
 - →~3 PB of online storage
 - →~20+ PB of Tape
- Growing fairly quickly
 - →More disk before Fall
 - » ~500 TB
 - » Focus on Lustre stability
 - →More nodes before Fall
 - » ~7000 more nodes







The JLab Farm • Batch Computing

- The Farm: <u>Batch Computing</u>
 - →No direct access to actual Farm nodes
 - » Use "Interactive" farm nodes for testing
 - ie. ifarm1402
 - →DB and other network access (git, http, etc) generally constrained
 - → Jobs controlled by automated system called "PBS" "Slurm/Auger"
 - →You submit a job to Slurm, and Slurm schedules it to run

- All about trade offs:
 - →"Latency" can be high (hours+ from submission to job execution)
 - » BUT!
 - →Throughput is enormous
 - » 100s (1000s) of your jobs can run at once
 - High bandwidth access to fast storage
 - →A full replay (100s of runs) can be completed in the time it would take 2–3 runs to complete in series on your desktop.





PBS → **Slurm** transition

- PBS (with Auger) managed/scheduled jobs on the farm
 - →Decides when/where jobs can run
 - →Manages staging file from tape, etc
 - →Getting pretty old...
- Replaced by Slurm
 - \rightarrow as of June 4, 2019

- Only a few user-visible changes:
 - →Old swif/jsub/jkill tools still work
 - →stderr/stdout goes to /farm_out/<user>/
 - →Slurm uses real / used memory to schedule now!
 - » You may be able to reduce your memory request!
 - → https://scicomp.jlab.org/docs/auger_slurm





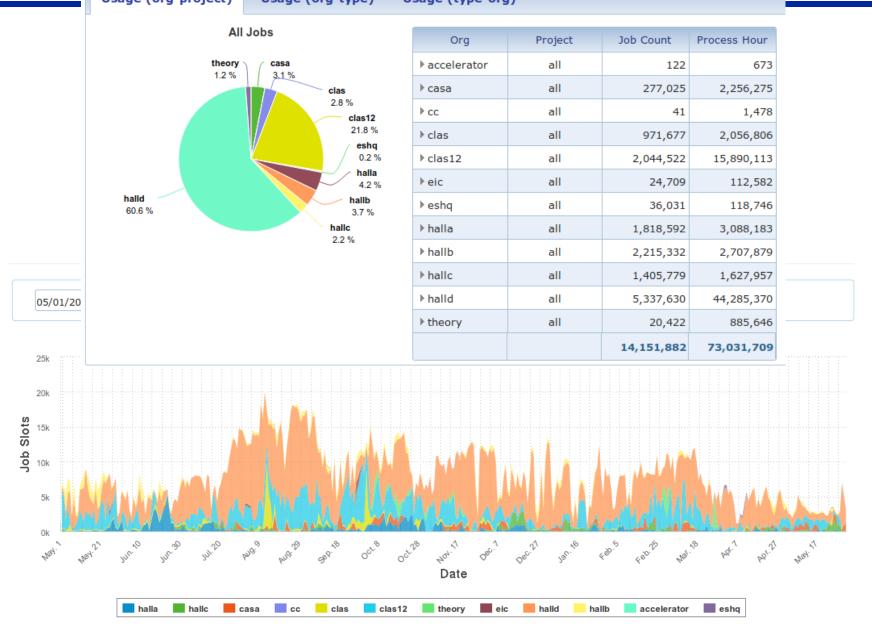
The JLab Farm • Scheduling

- The Farm is a Lab-wide shared resource
 - →Each Hall's budget includes \$\$\$ to support their usage
 - → Rough allocation:
 - » A: 10%, C: 10%
 - » B: 35%, D: 35%
 - » Misc: 10%
- Ruled by "Maui"/<u>Fair Share</u>
 - →Allocations <u>not</u> written in stone and are adjusted based on needs

- The balance is trickier to manage than you may think...
 - →Jobs take time to run (system doesn't know how long beforehand)
 - →Upcoming job load is hard to predict
 - →System balances allocations over a <u>few days</u>, not hours
- More documentation here:
 - → https://scicomp.jlab.org/
 - → https://data.jlab.org/







Do use the Farm!

- The Farm is not your desktop
 - →Need to plan a little and fire off groups of jobs
- Test your job first!
 - →Can it run reliably?
 - » If it doesn't run on ifarm140x, it won't run on the farm!
 - →Is the output what you

want?

Check before firing off 100 jobs

- Simple tasks, some types of analysis can be done on small systems, BUT!!
 - →Thou shalt back up your code!
 - →Thou shalt back up your results!
 - →IF your hard drive died today, how long would it take to recover?
- Don't keep only copies on your laptop
 - Don't keep only copies on your desktop's hard drive





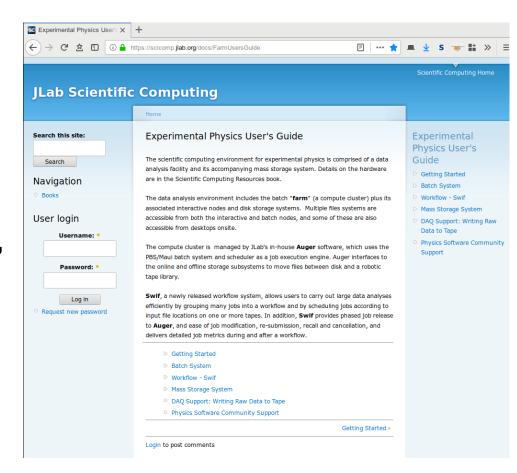
What's a "Job"?

- A 'Job' often maps to a shell script that runs your code
 - →It can do multiple things, but usually it executes a single instance of your software
 - » Analyze one run, or
 - » Simulate "1M" events,
 - » etc...
- NOTE: Output that would normally go to a terminal goes to

/farm out/<user>/

» NEW (Slurm change)

https://scicomp.jlab.org/docs/FarmUsersGuide







Debugging a job

- Generally want a single script that does everything!
 - → Set up full environment
 - → Use full paths
 - » /group/myExp/myscript.sh
 - » ./myscript.sh
- Testing your script:
 - \rightarrow 1st: Run on ifarm140x
 - → 2nd: Submit job to Farm
- Test with the 'debug' Farm track
 - → Max priority, fast sched.
 - → Limited 4 hour runtime
 - → Limited jobs/user

- Test on ifarm140x
 - % ssh you@ifarm1402
 % /group/myExp/myscript.sh
 - →Make sure it worked!
 - » check histos, report files
- Quick Test on Farm

```
% swif add-job -create \
  -track 'debug' \
  <other options> ... \
  /group/myExp/myscript.sh
```

- →Make sure it worked!
 - » check histos, files
 - » check /farm out/<you>/
- Then submit full set!
 - →<u>SWIF!</u>





Make your jobs run faster/better!

- Scheduling jobs takes many things into account
 - →File availability from tape
 - →Memory request
 - →CPU/core request
 - » >1 is useless for podd/hcana
 - →'Fairshare' metric
 - » Average Hall utilization
 - » 'Project' usage
 - ie. experiment
- Details
 - → Fairshare Web Page

- If a Hall / Project is not using 'their' fraction, then those Farm resources are available to anyone on a first-come, first-serve, basis!
 - →If the Farm is idle, you can take advantage!

» Like now!







Make your jobs run faster/better!

- Common Bottlenecks/ Mistakes
 - →CPU count
 - » use 1 core only (for now)
 - →Memory allocation
 - » < 2GB is best!
 - » Smaller → Faster scheduling!
 - →Insufficient debugging/cross checks
 - » Fire off 100s of jobs with bad config, buggy code





Check Job Status



- https://scicomp.jlab.org/scicomp/index.html#/slurmFarmJobs/activeJob
- Job Priority tab disappeared with Slurm...
- Recent Job tab can help you find information how jobs ran (or didn't run...)
 - → ie. Memory usage!
 - → See also: /farm_out/<user>/...





Small I/O Problems

- Small read/write operations are <u>very</u> inefficient
 - → Old/legacy code defaults can be very small (~4kB)
 - → Should be closer to 4MB chunks for decent performance
 - → Buffered IO can bridge the gap if needed
 - » Common errors:
 - 'Debugging' output
 - » stderr << "got here" << endl;</p>
 - » fprintf(stderr, "event %d\n", eventNum);
 - Opening/closing files very frequently
 - Frequent random I/O
 - » ie. searching through a file for a parameter every event
- Workflows / procedures that may work on desktops or older systems don't scale well on modern systems (100s or 1000s of simultaneous jobs)
 - → Can take down / degrade system-wide filesystems
 - → "Lustre" gets a bad rap for this at JLab, but all filesystems would have big problems under many of the observed loads!
 - » (IT/CNI are still "on the hook" to improve reliability though!)





File Systems: Where do I put my stuff?

CNI/IT provides

- →/group a space for groups to put software and some files, backup up by CNI
- →/home your home directory, backed up by CNI
- →Cache write through cache
- →Volatile acts as a scratch space
- →Work unmanaged outside of quotas / reservations



File Systems: Where do I put my stuff?

- CNI/IT provides
 - →/group a space for groups to put software and some files, backup up by CNI
 - →/home your home directory, backed up by CNI
 - →Cache write through cache
 - →Volatile acts as a scratch space
 - →Work unmanaged outside of quotas / reservations
- No really, that doesn't help. Where do I put my stuff so Brad won't hassle me and I can get my work done in peace!





Where do I put my JLab stuff?

- /home/<user>/
 - →hourly snapshots
 - » cd .snapshot/
 - →personal, nonanalysis files
 - » papers, notes, thesis, etc...
 - →analysis scripts: ~OK
 - » use git!
 - →source code: ~OK
 - » /work better
 - →NEVER store ROOT files or CODA files in /home

- Your laptop / desktop
 - →Should really be just a front-end for working on JLab systems
 - →Everybody wants to do backups, but almost no one actually does backups until after they've lost data...







Where do I put my stuff?

- /group
 - →Think "/home" for work groups
 - » papers, thesis, etc
 - →hourly snapshots
 - » cd .snapshot/
 - →analysis scripts: YES
 - » use git!
 - →source code: ~OK
 - » /work is better
 - →papers, thesis, etc in user subdirs is great

- /work
 - →Tuned for speed, small files
 - ie. source code, compilation, etc.
 - →NOT backed up
 - » but is resilient
 - » rm can still bite you
 - →Source code: YES
 - » use git!
 - →ROOT output: ~ick
 - →CODA data: No
 - →YOU must backup to tape
 - » tar + jput (more on this soon)





Where do I put my stuff?

- /volatile
 - →Huge file system
 - » ~ 180 TB
 - →High performance for large files
 - » ie. ROOT output
 - →NOT backed up
 - →Files auto-cleaned based on quota/ reservation/ and filesystem pressure
 - » https://scicomp.jlab.org/scicomp/index.html#/volatile
 - →Analysis output goes here!
 - » Check, then push to tape if good!

- Tape System
 - →Really huge
 - » 43 PB and growing
 - →/mss/hallX/...
 - » Stubs: shows what is in the tape system!
 - » not the actual files
 - →/cache/hallX/...
 - » actual files
 - » auto-clean up in play
 - next slide





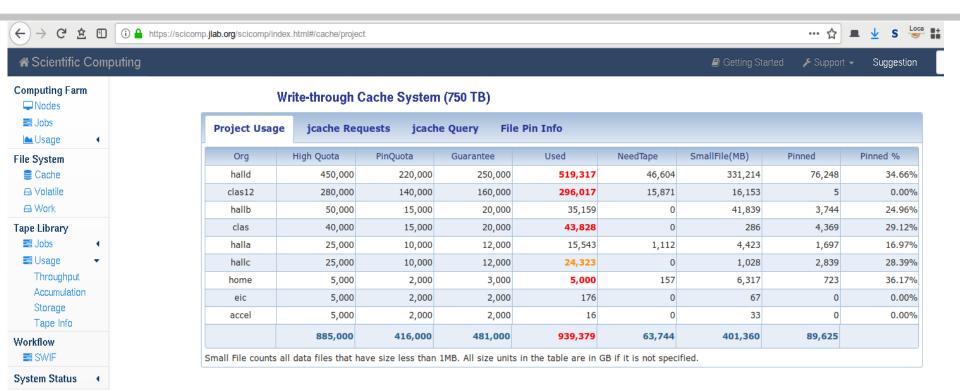
Accessing files from Tape

- Retrieving files from tape
 - →jcache get /mss/.../foo.dat
 - » Manual pull from tape to /cache/.../foo.dat
 - » Never call this (or jget) in a farm script!
 - Let Auger/SWIF do it!
 - » List needed files on <Input> tag
 - » Auger will prestage them for you in advance
 - →jget /mss/.../foo.dat \$PWD/
 - » pull file from tape to any filesystem
 - » generally not the right tool





File duration in /cache



- Files auto-cleaned based on quota and system pressure on /cache
 - →Clean up least-recently-used files first
 - →Can 'pin' files to keep them stable
 - » Shared resource, don't abuse!





Copying files to Tape

- Storing files on tape
 - → jput file /mss/.../
 - » 'jput -h'
 - » Online Docs
 - → 'write-through cache' (Online Docs)
 - » write large file output directly to /cache/hallX/...
 - no 'staging' on /volatile
 - » automagically backed up to tape after a few days
 - guaranteed to be safe on tape before /cache autoremoval kicks in
 - » Gotchas:
 - small files (<1MB) not backed up to tape
 - avoid pathname collisions with files already on tape
 - » ie. 'overwriting' files with same name, etc





Quick Breather

Any Questions on that bit?



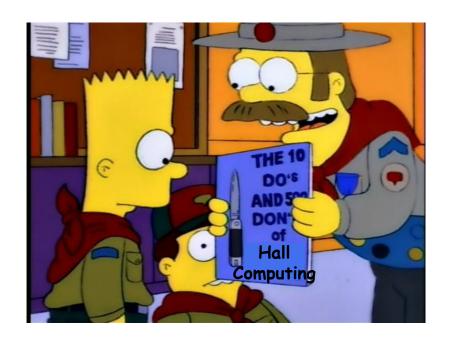


Hall Computing while Running / Online Analysis





Hall Computing Do's and Don'ts!





Rules to live by so your colleagues don't curse your name





- <u>Don't</u> copy or write large files to any "home" directory. This includes
 - →CODA files
 - →ROOT files
 - →tar or zip archives
- When these shared filesystems fill, many things break
 - →processes fail with corrupt output (not just yours!)
 - →machines may require rebooting (disruptive!)
 - →DAQ can die!



- Do use destinations like:
 - →/chafs/work*, /chafs2/work*
 - →/net/cdaq/*
 - →Use symlinks in local dirs
- Do use the tape silo:
 - →/mss/cache/hallX/...





- <u>Don't</u> change global config files in common accounts (adaq, cdaq, a-onl, cvxwrks, coda, etc)
 - » ie .bashrc, .cshrc
- <u>Don't</u> change environment variables (ie. 'setenv ...')
- <u>Don't</u> edit or save files in the global replay directories
- <u>Don't</u> run analysis on the DAQ machines
 - →All can have unexpected, and difficult to debug impacts on processes running behind the scenes



- <u>Do use</u> the appropriate machines and accounts
- Do use the JLab Farm:





- <u>Don't</u> use buggy/untested code in production environments
 - →If your code is:
 - » generating enormous 'log' files
 - » generating core.NNNN dumps
 - » littering hv.1.NNNN.tmp files
 - » pegging a core at 100% without good reason
 - » spewing warnings / debug info
 - → Then it needs fixing <u>before</u> the experiment
- <u>Don't</u> ignore warning in your replay scripts
 - →NEVER disable/hide warnings

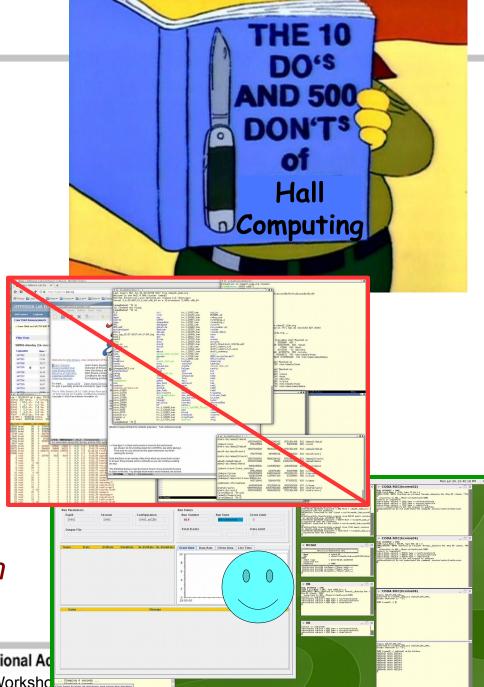


- We've lost way too many beam hours (\$\$\$ + data) to such problems.
- Poor saps have to come in at crazy hours to fix things, and hassle folks to clean up.
 - → Have mercy on us :-)



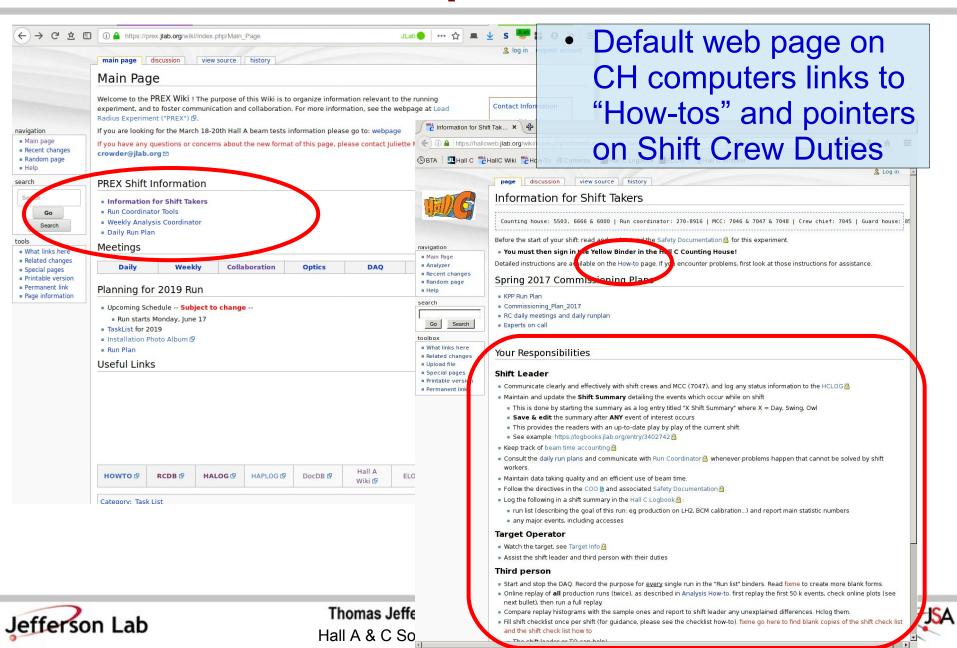


- Remember the Hall Computers are shared machines
 - → Don't clutter CH screens with 100s of windows
 - → Don't shuffle windows around on Shift Crew Machines
 - → Clean up after yourselves
 - » Copy to tape, and/or remove obsolete files
 - » NEVER move or touch CODA data files though
 - talk to Hall expert

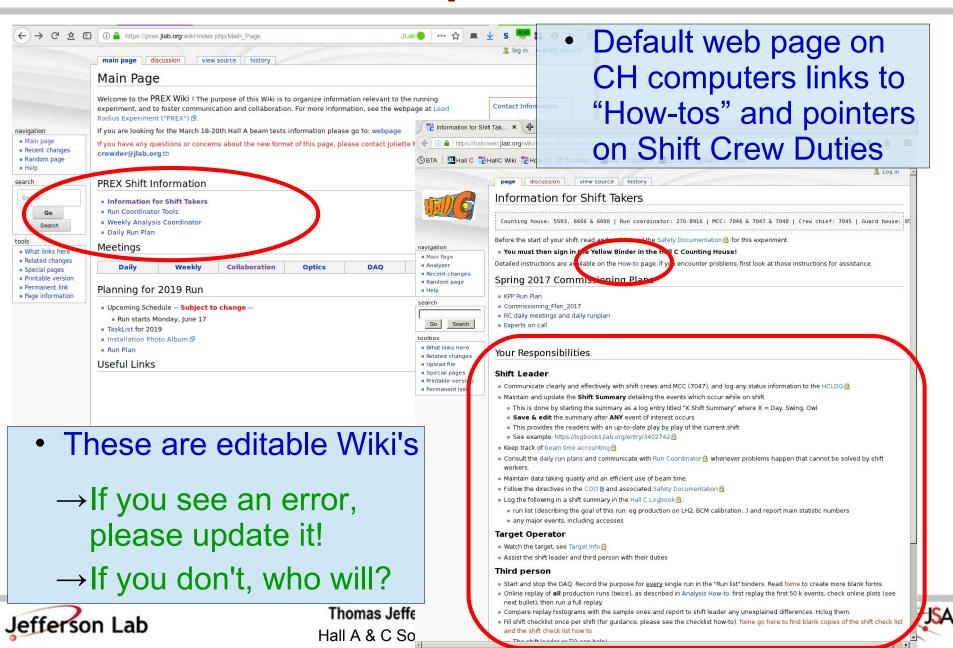




Online Help is Available



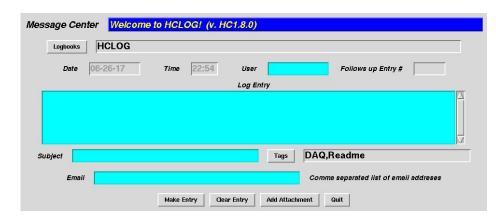
Online Help is Available



Logbook Tips

- Always use concise "Subject"
 - » No log entries titled "problem" please...
 - → Make clear entries that don't rely on being there to make sense
 - » We need to under-stand the entries months/years later
 - → Clean up the "Re: Follow-up:
 Follow-up: Re: Follow-up: ..."
 garbage
 - » it is just noise
 - → Manually link to relevant older entries
 - → Use Tags fields
 - » DAQ, Analysis, Techs, ...

- HALOG, HCLOG are web based:
 - → https://logbooks.jlab.org/book/halog
 - → https://logbooks.jlab.org/book/hclog
- On CH computers, can also make entries via standalone tool:
 - » halog / hclog
 - →Simpler snapshots!

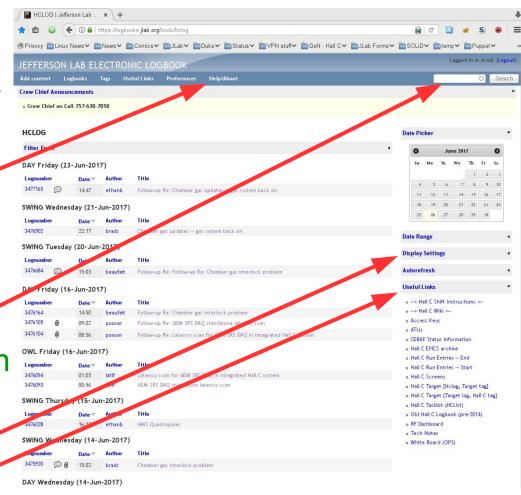






Logbook Tips

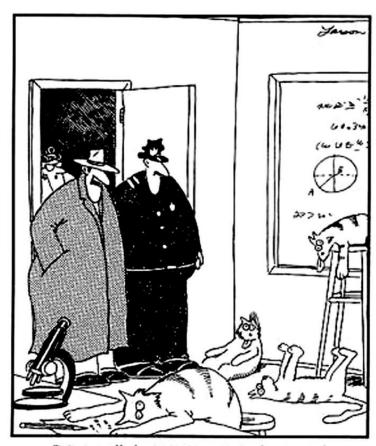
- HALOG, HCLOG are web based:
 - → https://logbooks.jlab.org/book/halog
 - → https://logbooks.jlab.org/book/hclog
- Help is worth reading
- Useful 'tricks'
 - →Use UPPER-CASE boolean logic in search
 - » ie. AND not and
 - → Display Settings Hide Autologs
 - → Useful Links





And, Most Importantly

- If you see something you don't understand...
 - →Ask someone
 - → Make a log entry
 - →Dig in and beat on the problem until it makes sense to you
 - » You're scientists understanding weirdness is literally our business!



"Notice all the computations, theoretical scribblings, and lab equipment, Norm. ... Yes, curiosity killed these cats."





Now ask Questions!







Hall A/C Computer Layout (possibly dated...)



Hall A Counting House Systems

DAQ

adaq1 5.5 TB 12 cores Xeon E5645 RAID-6

5.5 TB

RAID-6

19 TB

RAID-6

adaq2 12 cores Xeon E5645

compton 16 cores Xeon E5620

Blue: RHEL6, 64 bit

Red: RHEL5, 32 bit (Legacy)

Networking: Infiniband 4xSDR=8 Gbps **Online Analysis**

hamoller 4 cores Xeon E5410

aonl1 32 cores Xeon E5-2650v2 aonl2 32 cores Xeon E5-2650v2 aonl3 32 cores Xeon E5-2650v2 aonl4 32 cores Xeon E5-2650v2 Servers

chafs 16-core Xeon E5620 (2011)

19 TB RAID-6

adaqfs 8-core Xeon E5310 (2006)

1.7 TB RAID-5

1 TB each





Hall A Counting House Systems

DAQ

5.5 TB

RAID-6

5.5 TB

RAID-6

19 TB

RAID-6

adaq1 12 cores Xeon E5645

adaq2 12 cores Xeon E5645

compton 16 cores Xeon E5620

Blue: RHEL6, 64 bit

Red: RHEL5, 32 bit (Legacy)

Networking: Infiniband 4xSDR=8 Gbps Online Analysis

hamoller 4 cores Xeon E5410

aonl1
32 cores Xeon E5-2650v2
aonl2
32 cores Xeon E5-2650v2
aonl3
32 cores Xeon E5-2650v2
aonl4
32 cores Xeon E5-2650v2

1 TB each

Offline Expert Machines "Back Room" Servers

chafs 16-core Xeon E5620 (2011)

19 TB RAID-6

adaqfs 8-core Xeon E5310 (2006)

1.7 TB RAID-5

Shift Crew Machines "Front Room"

adaq1 DAQ

hacweb7 EPICS hapc1/hapc2 Scalers / HV

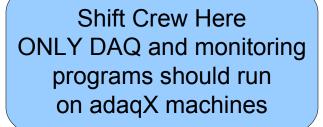
adaq2

DAO

poltarac
Target Control



Hall A Counting House Consoles

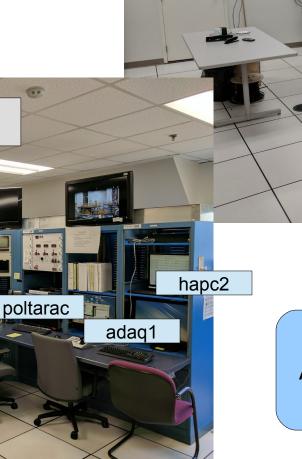


Front Room

adaq2

hapc1

hacweb7

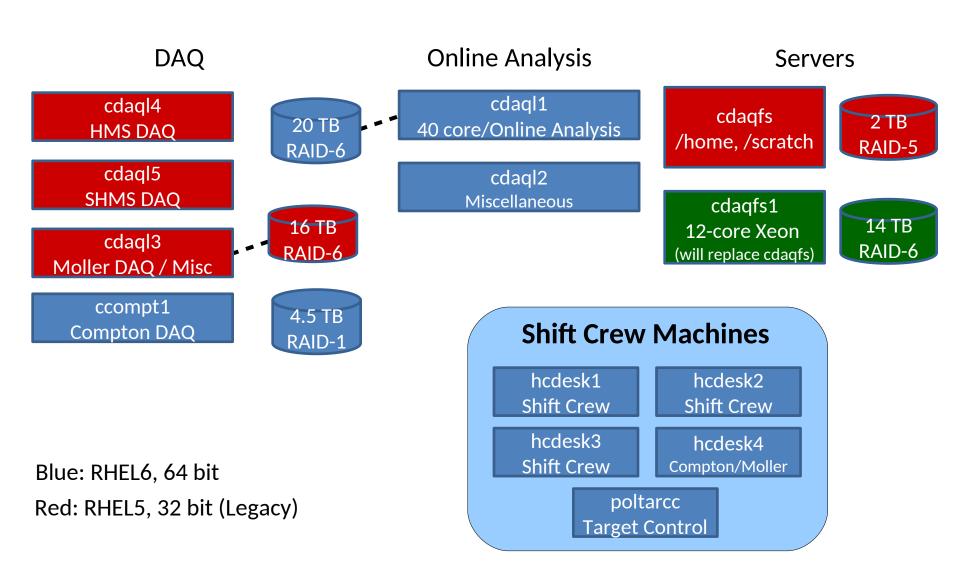


Offline Experts Here
All analysis should be done
on a-onl@aonlX machines

Back Room

aonIX

Hall C Counting House Systems







47

Hall C Counting House Consoles

