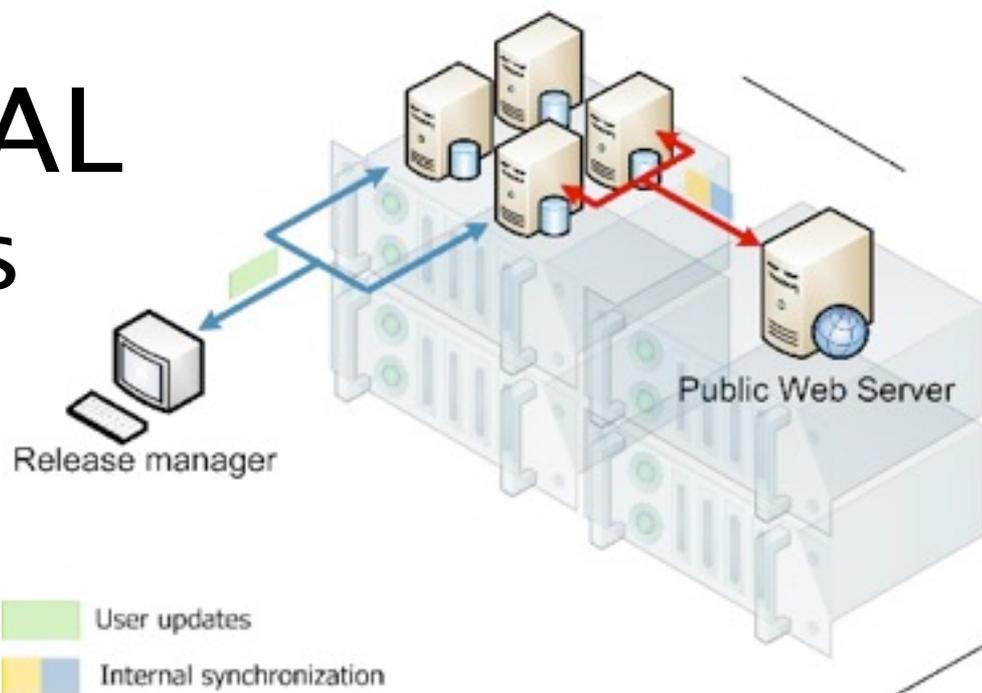




# CVMFS @ FNAL

## Status & Plans



A.Norman (NOvA), A.Lyon (g-2)  
REX Data Handling Group, **qua**  
**Eric Church**

- This is Andrew Norman's talk which I merely have appropriated to point out highlights vis-a-vis LArSoft

- Herb and I think this is a would-be gamechanger for LArSoft
- There would be no need for external installs which have caused non-trivial headaches and which split the codebase far and wide.
- You could run LArSoft from your laptop
- Your university's farm could be availed
- At FNAL, with bluearc disks no longer being a requirement, the available # of grid cores goes from ~2k to ~20k.

- Problem:
  - Need to to distribute LArSoft (170–200 GB) experiment specific software/code distributions to:
    - Worker nodes within FNAL computing clusters
    - Remote computing facilities
    - Individuals doing code development and analysis
  - Must be scalable (1000's of concurrent clients)
  - Must be sustainable/maintainable

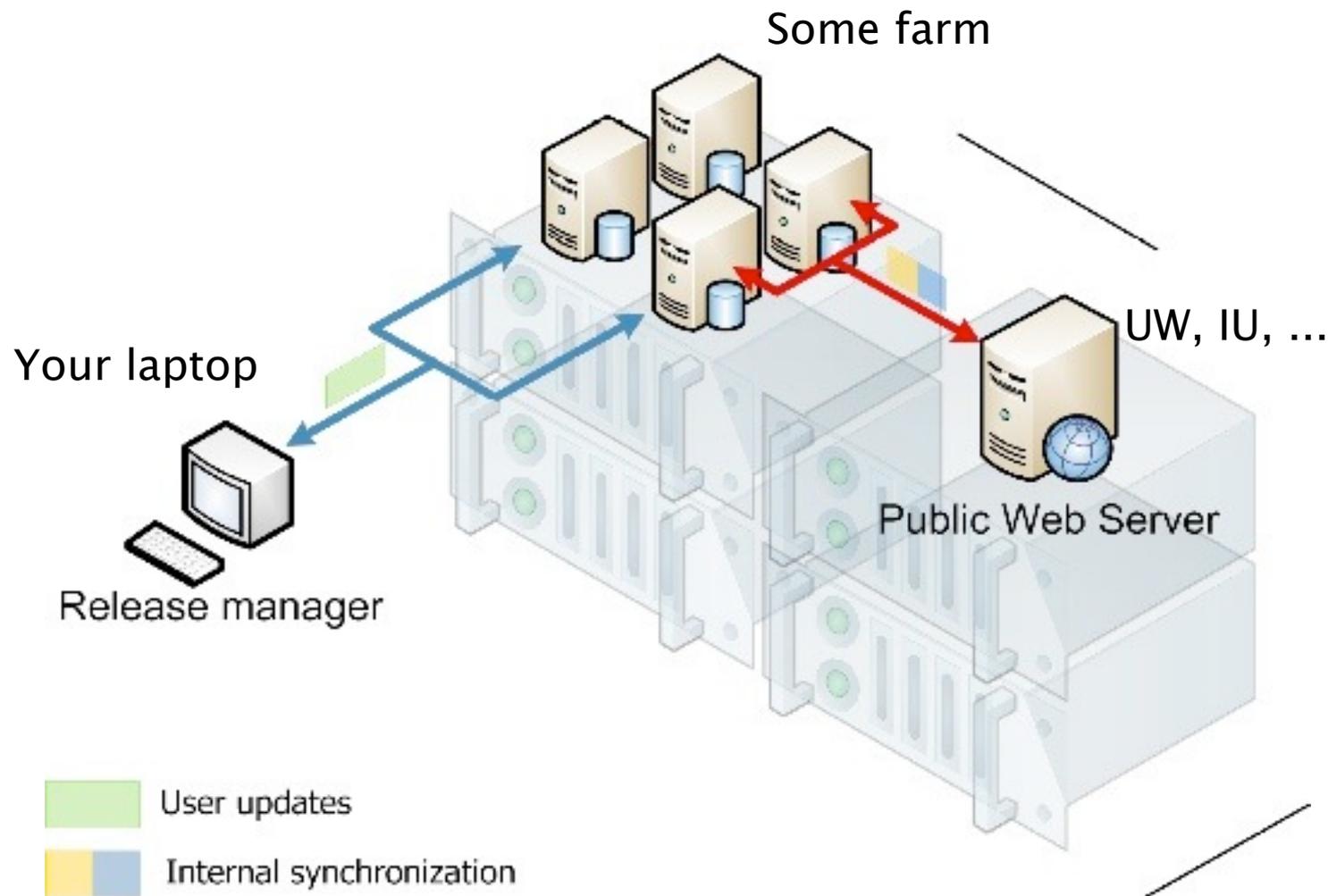
- Current Solution:
  - Bluearc Central Disk Services
    - Experiments deploy “master” code distribution to bluearc
    - Bluearc volumes are mounted (NFS) on computing resources that are local to FNAL
  - Scalability issues
  - Does not work for remote sites
  - Does not work for individual developers (i.e. laptops)

# What is CVMFS?

- The Cern Virtual Machine File System (CVMFS) is:
  - A method of distributing a large software/code distributions over a network
  - Provides **Read-Only** images of a distribution as a pseudo file system
    - Distributions are served via a Web API from a set of servers
    - Clients see a “fuse” based file system
    - Files are transmitted to clients only when they are accessed
    - Transmitted files are cached locally on the client
- Developed & maintained by CERN for Atlas/CMS



# CVMFS @ A glance



# Test Setup

- Test Server
  - Host: cms.hep.wisc.edu
  - Maintained by U. Wisconsin
  - Provides interactive login via gsissh
  - Sufficient disk to host full experiment code distributions
  - Provides automated sync/build of working areas into published CVMFS images
  - Provides web servers for actual distribution

# Clients Deployments

- Currently have operational clients at multiple sites:
  - Fermilab
    - novagpvm10 (gpcf)
  - SMU
    - smuhpc cluster
  - User laptops
    - SLF5, Ubuntu, Mac OSX (snow leopard and lion)
- Fermilab & SMU sites have dedicated squids to protect the Wisc server
- Machine have local cache configurations:
  - FNAL=50 GB
  - SMU=4 GB
  - Laptops (tunable by user, default = 4 GB)



```
anorman@novagpvm10:~/afs/fnal.gov/files/home/room1/anorman — ssh
[anorman@novagpvm10 ~]$ hostname
novagpvm10.fnal.gov
[anorman@novagpvm10 ~]$ df -h
Filesystem                Size      Used Avail Use% Mounted on
/dev/vda2                  15G       7.0G   6.9G  51% /
/dev/vda7                  17G       173M   16G   2% /scratch
/dev/vda6                  2.0G       37M   1.9G   2% /tmp
/dev/vda3                  3.9G      315M   3.4G   9% /var
/dev/vda1                  122M       51M    65M  45% /boot
/dev/vdb1                  40G       2.8G   35G   8% /var/cache/cvmfs2
tmpfs                      5.9G       0      5.9G   0% /dev/shm
gpcf015.fnal.gov:/scratch/nova
                               10T       1.8T   8.3T  18% /scratch/nova
blue3:/nusoft/app          512G      411G  102G  81% /nusoft/app
blue3:/nusoft/data         10T       3.5T   6.6T  35% /nusoft/data
blue3:/nova/app            3.0T      2.7T  361G  89% /nova/app
blue3:/nova/data          140T      100T   41T   72% /nova/data
blue2:/fermigrid-fermiapp  1.1T      835G   266G  76% /grid/fermiapp
blue2:/fermigrid-app       300G      243G   58G   81% /grid/app
blue2:/fermigrid-data      24T       13T    12T   52% /grid/data
blue3:/nova/ana            60T       23T    38T   39% /nova/ana
AFS                        8.6G       0      8.6G   0% /afs
cvmfs2                     9.8G      2.6G   7.3G  26% /cvmfs/cms.hep.wisc.edu
[anorman@novagpvm10 ~]$
```

Cache partition  
(required for each client station)

Bluearc Volumes

CVMFS File system  
(/osg/app/nova)

# Plans

- Starting on larger scale client testing
  - Scaling up with Fermigrid ITB (14–28 node)
    - Mid November using Wisc. server
  - Large scale with new worker nodes in test config (1000 nodes)
    - End November, 1<sup>st</sup> week December with Wisc. server
- Migration to OSG–GOC hosting of distros
  - Server available end Nov/1<sup>st</sup> week Dec.
  - FNAL/NOvA will be beta testers
  - Specification document in internal review for general OSG hosting of server

# Preliminaries

- Someone must build LArSoft for OSX10.N, ubuntu,
- You must take out all weirdo hard-coded paths (/lbnegpvm02/app/users/biff/testarea/files/xyz.root) and request mounts only of a minimal setup: /grid/fermiapp/lbne/lar/code/larsoft/releases/S2012.12.31,  
+ ... /nusoft/

# ... and then ...

- It just works.
- Stay tuned.

# Arcana

# Testing

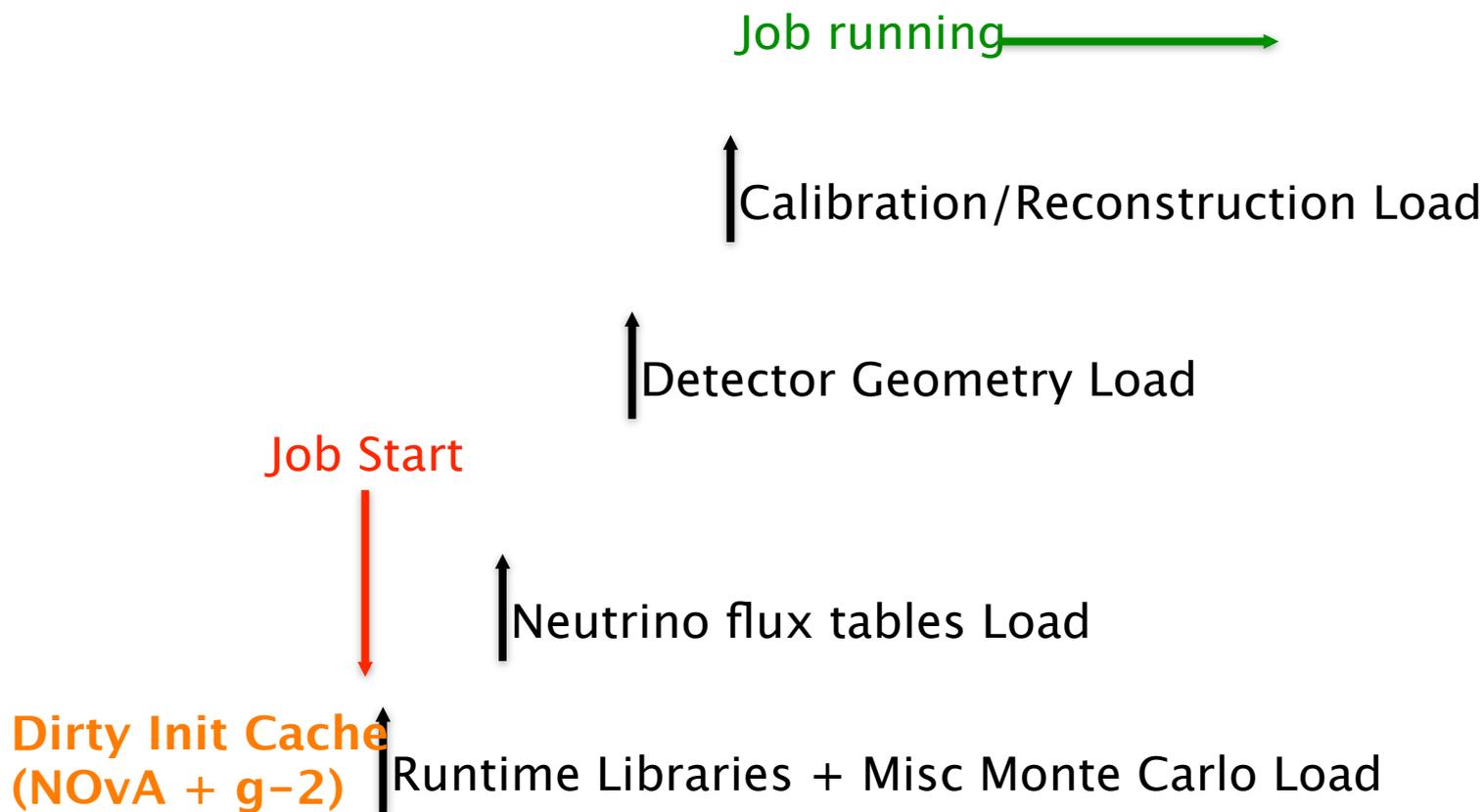
- Initial testing has been with current NOvA offline analysis and production jobs
- Initial tests have been single jobs
  - Required significant work to disentangle all bluearc dependencies (i.e. hard coded paths by users)
  - NOvA @ FNAL
    - ✓baseline **analysis** job on gpcf client
    - ✓baseline **reconstruction** job on gpcf client
    - ✓Full **Monte Carlo** generation job
    - ✓Full **production** job
    - ✓Full **production** job w/ Library Event Matching (LEM)
  - SMU
    - ✓baseline Monte Carlo (single particle gen)
    - ✓baseline unpack (evtdump/evtdisplay)

# Performance & Cache

- CVMFS behaves as expected
  - On GPCF (for baseline analysis)
    - Initial job startup (with clean cache) took < 2 minutes
      - Cache was monitored to grow from < 300 MB to 1.7 GB.
      - Job started and ran normally
    - Subsequent jobs (with cache populated) started immediately
      - Cache is hit essentially 100%
      - Remains populated across multiple NOvA jobs
      - Squid logs indicate minimal misses after initial job run
    - g-2 jobs have cache footprint  $\approx 0.5$  GB
      - Does not conflict with NOvA jobs
  - On SMUHPC
    - Cache foot print  $\approx 1.5$  GB for simple MC
    - Do not have stats on squid or initial population of cache



# Performance & Cache



# Server Side Deployments

- Currently have deployed:
  - Full NOvA code distribution
    - All releases to date (17 separate releases + development)
    - All external packages (packaged as UPS products) to support the releases
    - Total size:  $\approx$  170 GB
  - Full g-2 code distribution
    - 29 GB
- Both installed as:
  - /cvmfs/repository/osg/app/experiment
  - Simple installation procedure (just unwind a tar ball and wait for the server to catch up)