BABAR DATA PRESERVATION STATUS UPDATE

Concetta Cartaro
BABAR Computing Coordinator

FERMILAB
March 28th , 2014

OUTLINE

- *BABAR* quick summaries
- Status of the Long Term Data Access project
- New developments:
  - *BABARTOGO* and the Federated Datasets
- Documentation
- Global status of *BABAR* Computing
- Conclusions

BABAR COLLABORATION

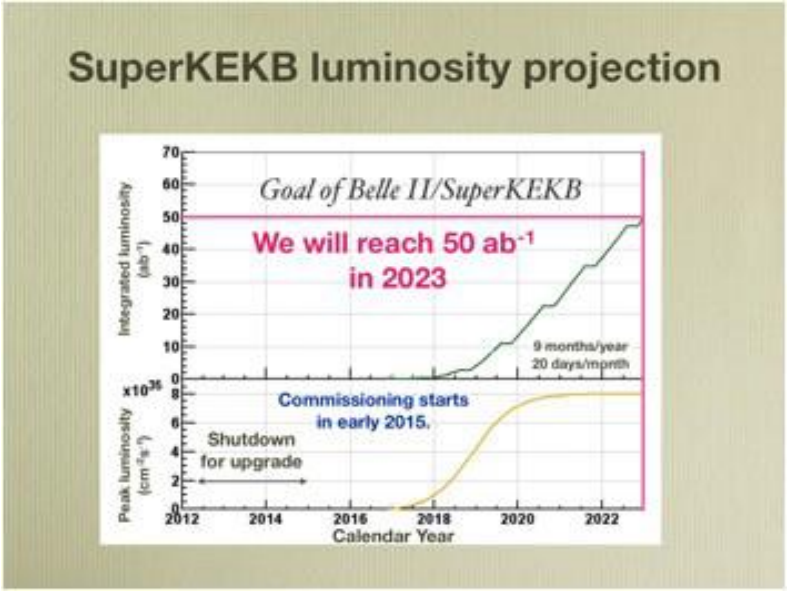
- ~300 members from 72 institutions in 13 countries.

Long Term Task Force

- Prepare the Collaboration for the long term.
  - Consolidate the governance structure.
- Survey of all groups.
  - Projected activity (analysis, review process, …).
- Main roles in the collaboration become permanent.
- Activity still too high to let the various committees merge (executive board, publication board, speakers bureau, …) but it will eventually happen.
- Some analysis working groups merged already.
  - We had 10, now there are 8.

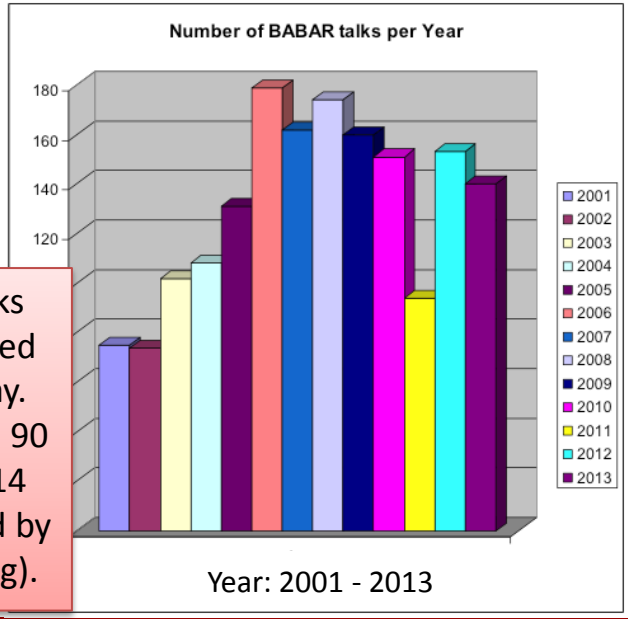
BABAR DATA

- *BABAR* has collected data from Oct 22nd 1999 to Apr 7th 2008.
  - 800TB of raw data, 1.4 PB from the last data reprocessing.
  - 537 papers published/accepted/submitted to date.
  - ~50 on track analyses plus ~30 analyses progressing slower (manpower).
    - Possibilities for new previously unforeseen analyses including discovery analyses.
- *BABAR* (and Belle) data will not be superseded by LHC data.
  - Good match for Belle II data taking schedule.
  - Some datasets expected to remain unique for longer:
    - Υ(3S) dataset.

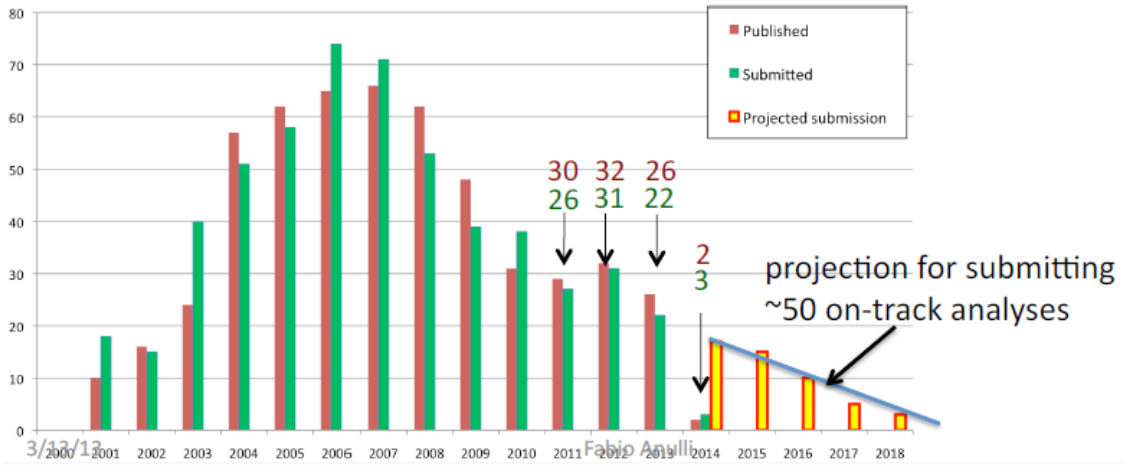


BABAR PUBLICATIONS

- *BABAR*'s recent papers continue to have a high impact:
  - Those from 2011-2013 already have >2400 citations, ~1400 for 2012 and 2013 papers
  - 537 papers published/accepted/submitted to date.
  - ~50 on track analyses plus ~30 analyses progressing slower (manpower).
    - Possibilities for new previously unforeseen analyses including discovery analyses.
      - 10 new analyses started in the past year

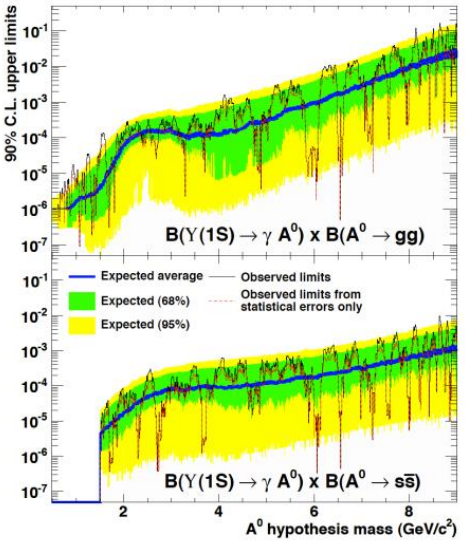


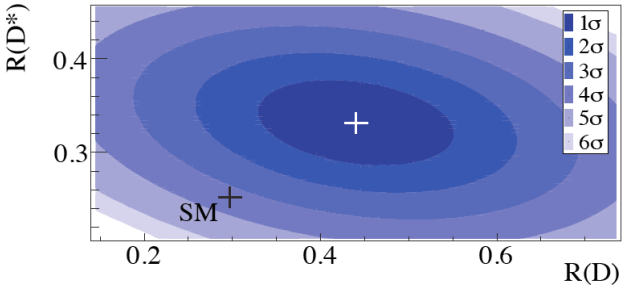
35 talks assigned to May. Expect 90 in 2014 (limited by funding).

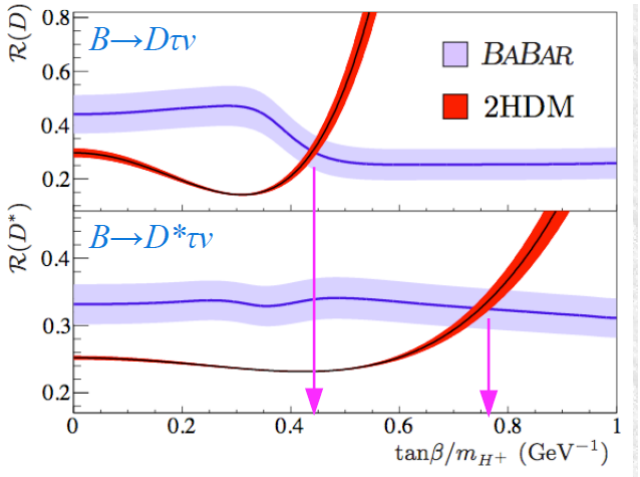


PHYSICS HIGHLIGHTS I

- Search for a light Higgs decaying to two gluons or ss in the radiative decays of ϒ(1S)
  - Phys. Rev. D 88, 031701 (2013)







- B→D(*)τν
- Results incompatible with SM and 2-Higgs Doublet Model Type II, excluded at >3σ on the whole tanβ-m_H plane.
  - PRL 109, 101802 (2012), PRD 88, 072012 (2013)

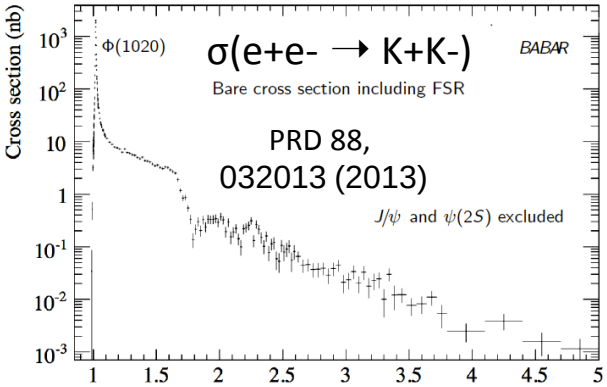
PHYSICS HIGHLIGHTS II

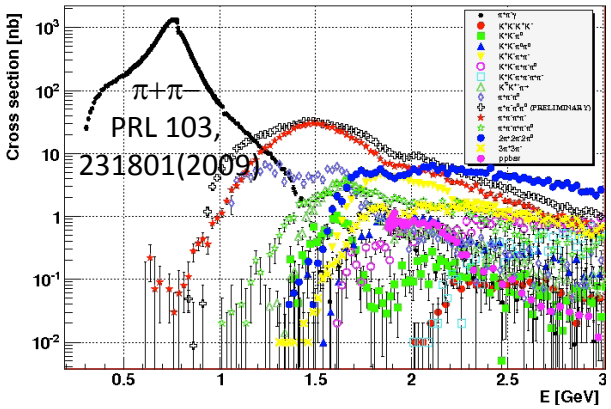
$$a_\mu^{SM} = \left(\frac{g-2}{2}\right)_\mu = a_\mu^{QED} + a_\mu^{had} + a^{weak}$$

BABAR only $a_\mu^{had}(K^+K^-) = (229.5 \pm 1.4 \pm 2.2)\ 10^{-11}$

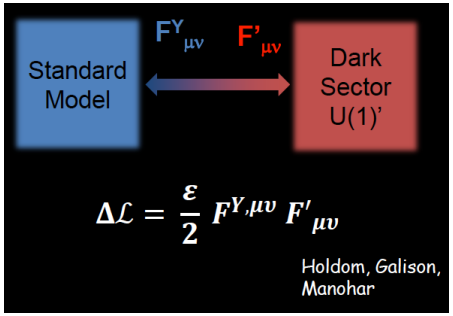
previous average $a_\mu^{had}(K^+K^-) = (216.3 \pm 7.3)\ 10^{-11}$

- BABAR is the only experiment to measure cross sections from threshold up to ~4-5 GeV.
- $e^+e^- \rightarrow \gamma K_s K_L$ preliminary result also available





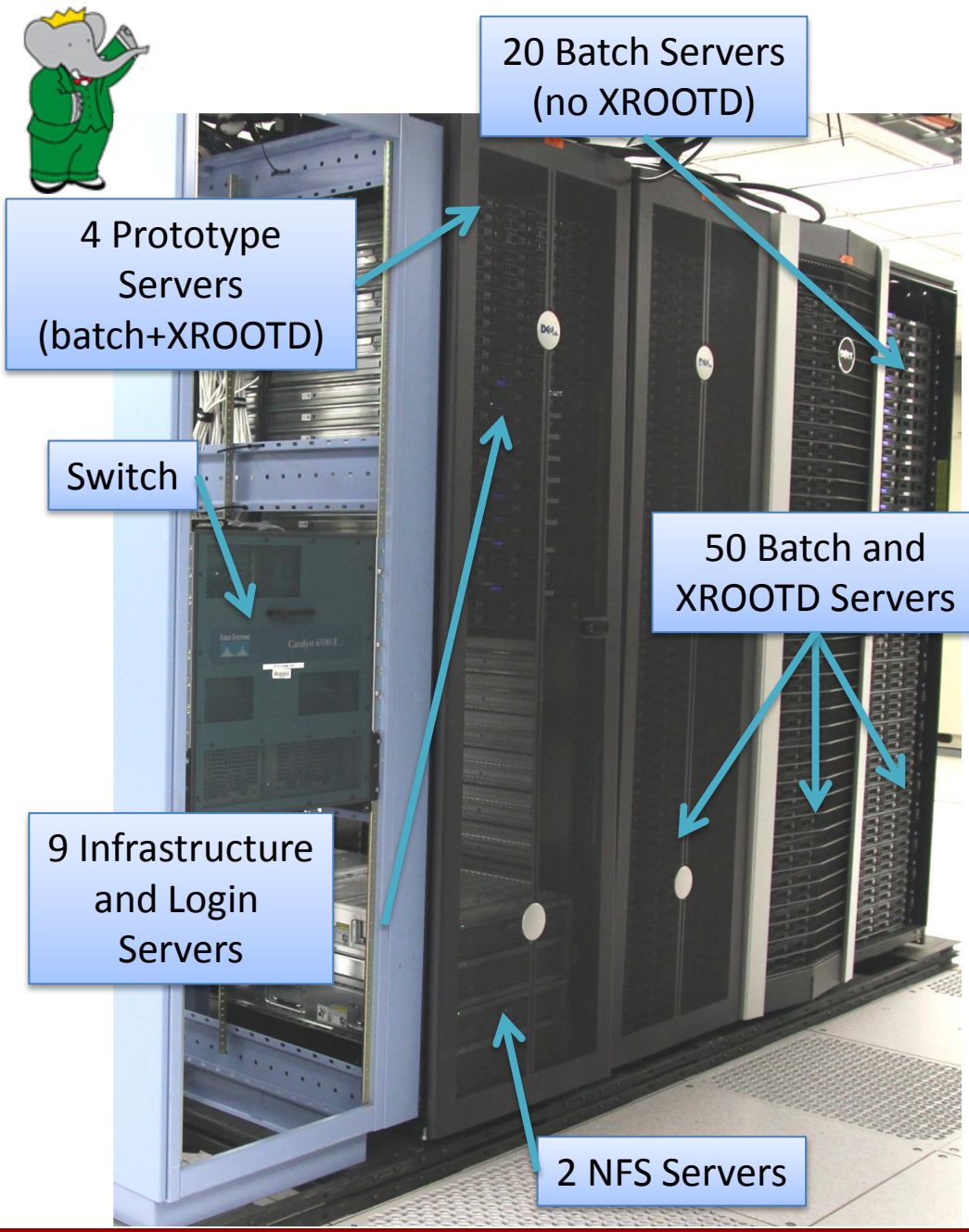
- Plus:
  - Search for dark photons.
  - Searches for new physics effects in B physics and CP violation direct and in mixing.
    - $B \rightarrow X_s ll$, $B \rightarrow \tau\nu$, D and B mixing, …
  - … and many more.



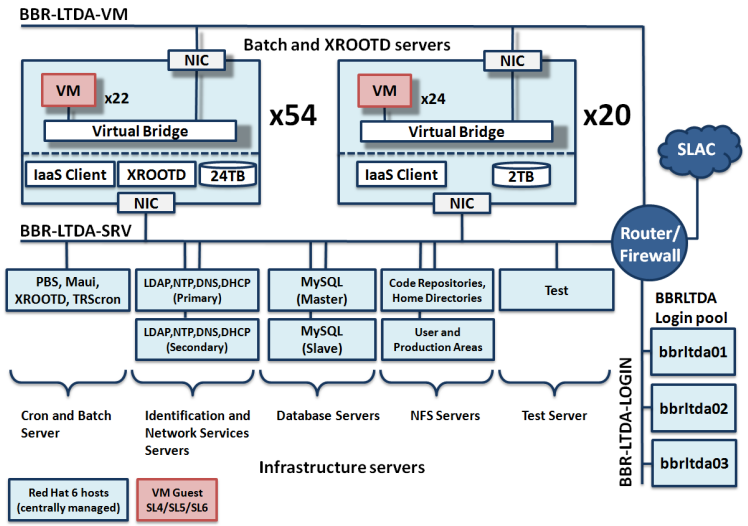
Long Term Data Access

- Insure the ability to support analysis of the *BABAR* data until at least 2018.
  - Preserve data, conditions, calibrations, releases, tools, databases, capability of running production and user jobs including new Monte Carlo models.
  - Accurate documentation to preserve know-how on physics analysis, OSs, Framework, …
- Providing a stable environment.
  - Last validated OSs enclosed in a virtualization layer running the *BABAR* Framework minimizing the effort needed to maintain the system.
    - Releases built on SL4, SL5 and SL6 are all fully functional with ROOT versions from 5.14 to 5.34.
- Use open formats.
- Data storage.
  - 2PB (including raw data) will be stored on tape at SLAC and CC-IN2P3.
  - Most used data sits on disk.
- Central services moved to LTDA.
  - CVS repository (code and analysis documents), nightly builds, …
- LTDA enclave stable since day 1.

March 21st was the 2nd LTDA birthday!



- LTDA Facts
  - 1.33 PB of disk for data and users.
  - SL4, SL5, SL6 platforms available.
    - Security threat associated to a VM running an old OS
  - 1668 job slots (Virtual Machines).
  - 1GB home for each *BABAR* user.
- LTDA user-friendly environment.
  - Interactive VM's for all platforms always available.
- All the R22, R24 and R26 data are available in the LTDA.



BABARTOGO

- A data preservation project for *BABAR* beyond 2018.
  - Portable and versatile, designed mainly for single user machines (laptop,…) .
  - Take advantage of availability of expertise.
- *BABARTOGO* :
  - A SL5 VM with an analysis release fully installed and ready to use.
    - A base image with the OS (2.4GB) and an auxiliary image containing the *BABAR* filesystem with a minimal structure (20GB).
    - A single user, babar, is preconfigured, access as root also possible.
    - Runs with qemu or VirtualBox (open source, multi platform).
  - Start *BABAR* environment, create a test release (local user work/code development area), compile code, mount a directory with data files, and run binary to create analysis ntuples.

SKINNY DATASETS

- Skinny datasets for *BABARTOGO.*
  - Minimal amount of disk needed for the data by a user who wants to use *BABARTOGO.*
- For example consider only a deepCopyMicro (or Mini) skim.
  - Very light weight.

Components	R24g data + MC + skims	R24c data + MC + skims	R24d Y(2S,3S) + MC +Skims	R26b Gen MC (1237,1235,1005)
Micro (TB)	93	196	41	81
Micro+Mini (TB)	223	327	78	161

Single skim examples (some of the largest)

Skim	R24c skim size (TB)
BToDlnu	7.8
BSemiExcl	4.0

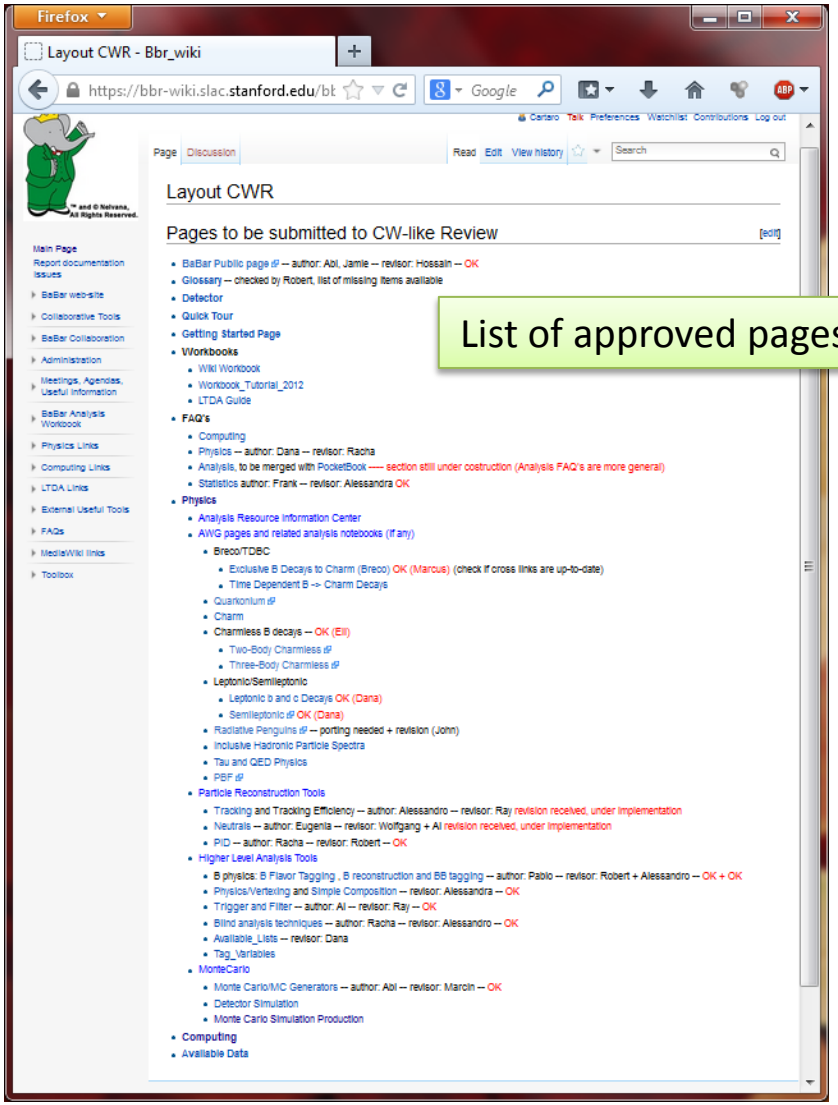
BABARTOGO AND FEDERATED DATASETS

- How to make data available to all instances of *BABARTOGO* ?
- A natural option: XROOTD.
  - The existing import/export tools can be used to transfer files/collections/datasets from site to site (applies to using Xrootd or local file system).
- Basic approach for data and conditions.
  - Fixed NFS mount point in the VM for conditions and data.
  - The user provides a NFS disk with a specific path to the data.
    - /store/SP/R26/001005/200707/26.0.0/… or /store/PRskims/R24/24.3.3b/…
- Using XROOTD we can build a federated xrootd cluster across our TierA sites and participating institutions.
  - First tests going on between SLAC and our German TierA site, GridKa.

Documentation Project

- Project completed.
  - Except for minor ongoing edits on Neutrals and Tracking pages.
    - The wiki, by nature, is always evolving…
- The Documentation Working Group lead by Alessandra Filippi (INFN Torino) is coordinating the effort aided by an advisory committee.
- All the most used and fundamental information have been checked, updated and moved to a Media Wiki server, the *BABAR WIKI.*
  - Old pages clearly marked but kept online for archival purposes.
  - Detector pages and other pages that will supposedly never change again are left in their original location.
  - There are about 10 official members in the DWG but the new wiki pages undergo a Collaboration wide review process and experts sign-off on the content of migrated pages before they are officially released.

DOCUMENTATION



List of approved pages

Wiki main page

A superseded HTML page

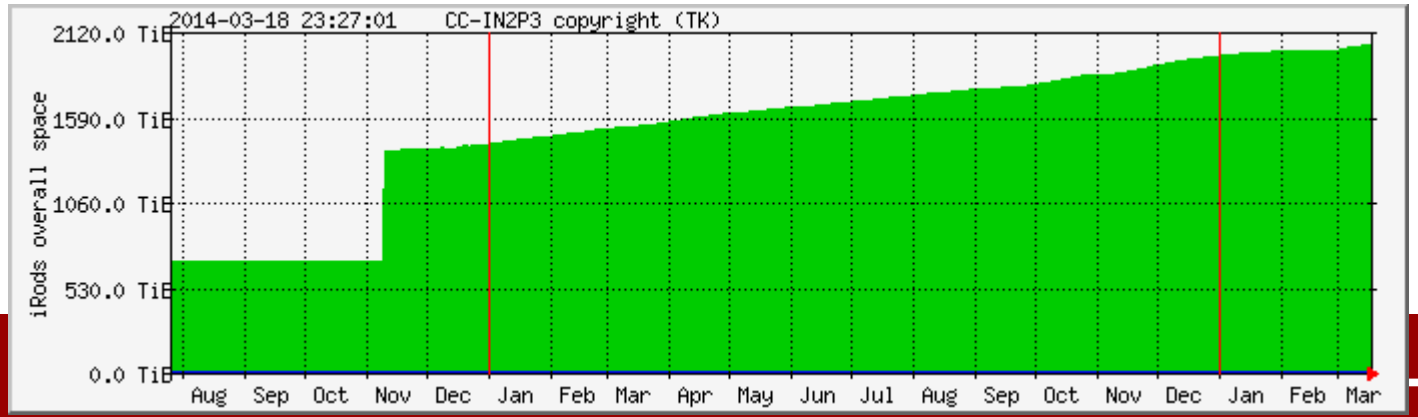
General Status

- LTDA is very stable, dedicated to both users and production.
  - Both simulation production and skimming run on the system at a constant level (200 slots) when needed.
  - One server is dedicated to our OPR (raw data reconstruction) processing.
- Dedicated production resources disappeared (>5 years SUN equipment turned off).
  - Production now runs in the general queues.
    - We have to cope with unpredictable pending times and compete with our own users but it works reasonably well.
- TierA sites are lowering their support to *BABAR*.
- International Finance Committee expectation is that SLAC/DOE will continue to support *BABAR*.
  - Resources needed to accommodate the users moving their work to SLAC.

TierA Sites

- CCIN2P3 will continue to support analysis activity, SP, Skimming and hosts a full copy of *BABAR* data (raw data and the last two reprocessings, namely R22 and R24).
  - Still some data to import but the bulk is done.
  - SP done on best effort.
- CNAF support reduced to the bare minimum (no funds in 2014 for *BABAR*).
  - Active users will continue to receive support but only the needed datasets will remain at CNAF and will be moved to tape. The support may stop at any time.
  - Users that are no longer active will have their data stored at SLAC on tape (15TB) and temporary on LTDA.
- GridKa provides analysis support and skimming and will continue in 2014.
- Uvic provides analysis support on best effort but SP production stopped.

babar iRods instance at CC-IN2P3 at 2014-03-18 23:27:01



Lesson Learned on Manpower

- Loss of data/ intensive skimming and SP production.
  - We have experienced the loss of a quite large amount of simulated data due to broken tapes.
    - Some data recovered from CNAF and CC-IN2P3 but not all.
    - Many datasets and affected, need to reproduce the data.
  - Skimming problems discovered and needed massive debugging.
- To keep a large and complex amount of data alive is somewhat independent from the number of users and we are at our lower limit for data and user support.
  - It is becoming more and more difficult to keep expert people due to funding pressure while all aspects of computing activity needs continue attention.
  - Data curation problem.
- ~2 FTE in 2014.

HARDWARE

- LTDA servers will start to hit the 5 years of lifetime in 2015 and 2016.
- Central servers also will require partial replacement.
  - Disk storage for users, data distribution servers, web servers, …
  - We have servers as old as 12 years!
- Tape migration to new media T10k-C with 5TB of capacity, reformattable as T10k-D (7.5TB).

BABAR COMPUTING GROUP

- TierA site managers, production managers and experts
  - Jean-Yves Nief, Randie Sobie, Vincent Poireau, Paolo Franchini → Sonia Taneja, Dave Brown, Nicolas Arnaud, Alessandro Gaz, Marcus Ebert, Chris Buenger, Alexandre Beaulieu, Rocky So, Doug Johnson, Homer Neal, Douglas Smith, Ray Cowan, Charlotte Hee.
- LTDA developers
  - Tina Cartaro, Homer Neal, Marcus Ebert, Steffen Luitz, Igor Gaponenko, Douglas Smith, Tim Adye, Teela Pulliam.
- Computing Division experts
  - Booker Bense, Lance Nakata, Randall Radmer and all the Unix-Admin team. Wilko Kroeger, Antonio Ceseracciu, Len Moss (retired) →Andrew May.
- SLAC PPA Management and DOE.

THE USUAL SUSPECTS IN RANDOM ORDER!

Conclusion

- LTDA system went into production on March 21st 2012
  - On time and within budget.
  - Extremely stable ("What? No failed skim jobs? It's unnatural!").
- All BaBarians have a LTDA home directory
  - Enough capacity to support about 15 simultaneous analyses
- *BABAR* infrastructure (not LTDA) needs attention
  - Aging hardware (e.g. NFS servers)
  - Tapes
  - Production and data curation

LTDA details

BACKUP

THE LTDA CLUSTER FACTS (I)

- Cisco 6506 network switch with 2x10Gb link card and 192Gb ports
- 9 infrastructure servers (Dell R410/R510)
  - 3 front end machines (**`bbrltda`** load balanced pool), 1 cron server, 1 test server, 2 infrastructure servers (network and identification services), 2 database servers (mirrored)
- 54 batch and storage servers
  - Dell R510: dual 6-core Intel Xeon X5675, 3.07GHz, 48GB RAM, 12x2TB disks
  - 4 were the prototype (dual 6-core Intel Xeon X5670, 2.93GHz, 48GB RAM)
  - 11x2TB disks (no raid) used to stage data through XROOTD
  - 1x2TB used as local scratch
  - 12 physical cores, 24 cores with hyper threading
    - 1 physical core used for the host and the XROOTD services
    - 11 cores (22 w/ hyper-threading) dedicated to batch with one VM per core

THE LTDA CLUSTER FACTS (II)

- 20 batch servers
  - Dell R410: dual 6-core Intel Xeon X5675, 3.07GHz, 48GB RAM, 2x2TB disks mirrored (for OS + local scratch)
  - 12/24 cores used to run batch jobs (VMs)
- 2 NFS servers
  - Sun X4540 Thor server: 12 cores, 32 GB memory and 32TB of effective storage
  - One for local home directories and code repositories and one for user data
- The LTDA cluster is in production mode since March 21st 2012
  - On time and on budget
  - 1.33 PB of disk space for data and users and 1668 job slots
  - SL4, SL5, SL6 platforms available
- All active BaBarians have a 1GB home directory on the LTDA
- Robust backup by using a combination of ZFS filesystem snapshots and tape backup

NFS Backups Details

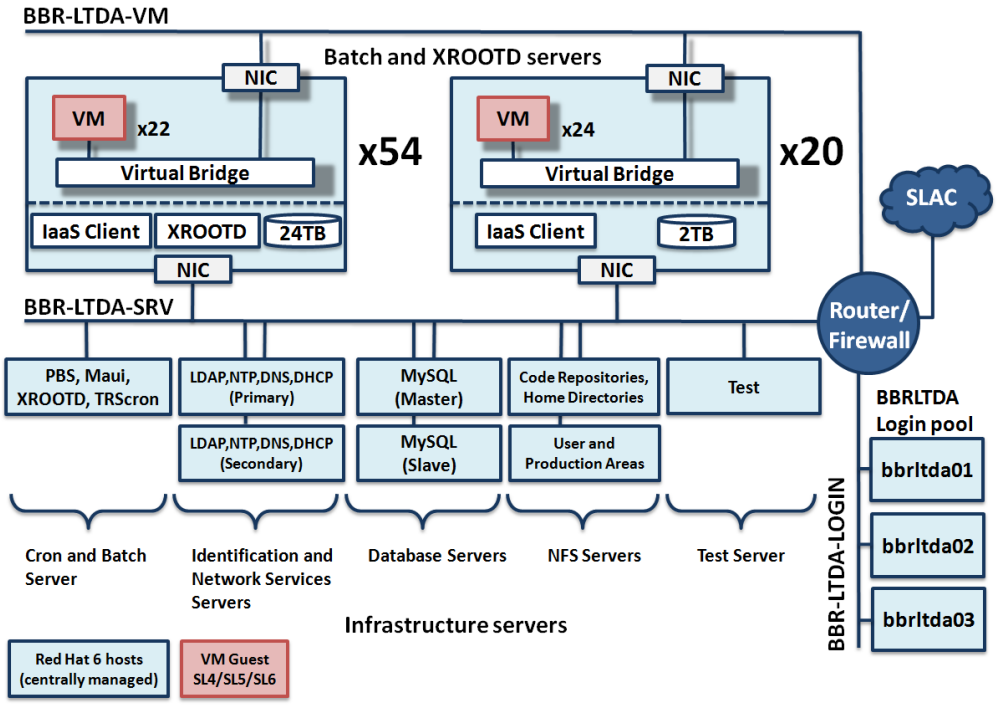
- NFS servers
  - 40TB zpools, 2 hot spares for 32 TB usable space
  - Compression enabled on local home directories, /home, with a factor ~2 gain
- ZFS snapshots implemented for /home and /BFROOT (releases, packages and cvs root directory) for user error recovery
  - snapshots are read-only, so it's protected against user error and are taken every 15min, overwritten every hour. The full hour snapshot is kept until next hour and the midnight snapshot becomes the daily snapshot and is kept for 30 days.
- The second nfs server (working areas only) also stores a snapshot of the first nfs server (home directories and code) so that, in the event of loosing the server, the second one can take the place of the first in just few minutes allowing the cluster to continue working.
- Tape backup for catastrophic events
  - All areas are backed up to tape every day and kept for 30 days
    - Root files are omitted because they are considered reproducible

VIRTUALIZATION & NETWORK

- Security threat associated to a VM connected to a network running old OS

→ Risk based approach assuming that the VMs are compromised

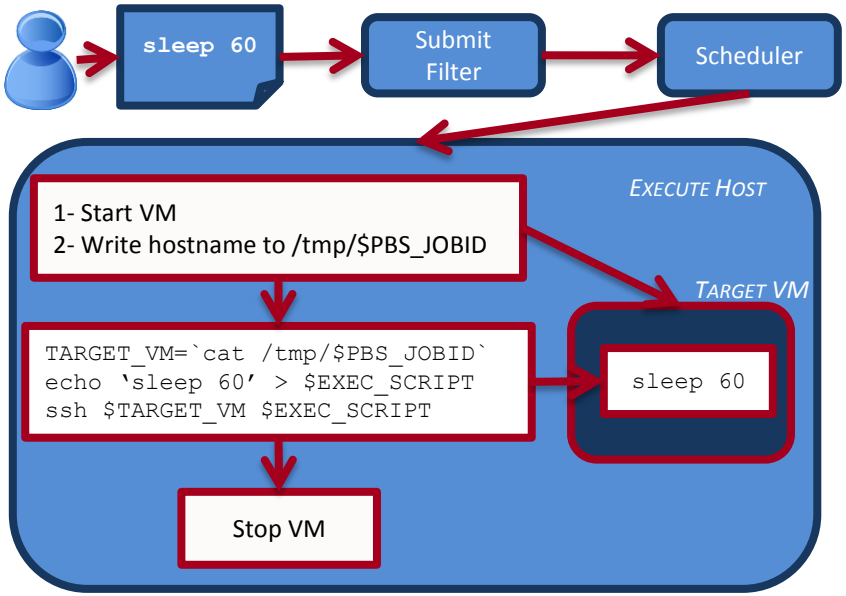
- Isolation of back versioned components with firewall rules
- Physical hosts centrally managed by SLAC CD
- Images are read-only, qcow2 produces a temporary file with changes to OS and scratch area and it is deleted when the VM's shut down



- VMs are not allowed to connect to SLAC network or the world
- The Login network is protected from the VM network
  - Allow one way ssh from Login to VM network
  - VMs are not allowed to write over the Login network
- Well defined services between VM network and SRV network
  - Infrastructure (DNS, LDAP, NTP), file service (Xrootd, nfs), batch scheduling
  - LDAP is a subset of the SLAC Kerberos list mapped on /nfs internal home directories
- Allow SRV and Login networks use SLAC infrastructure

JOB SUBMISSION

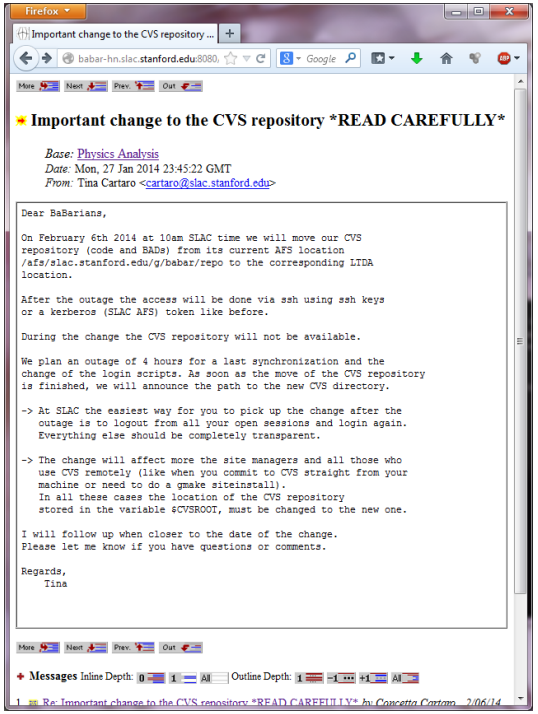
- PBS/Torque is used to manage the batch resources and Maui is the batch scheduler
  - Open source
- PBS Prologues and Epilogues scripts are used to create and destroy the VM's and the needed network environment
  - Home grown system developed by BaBar



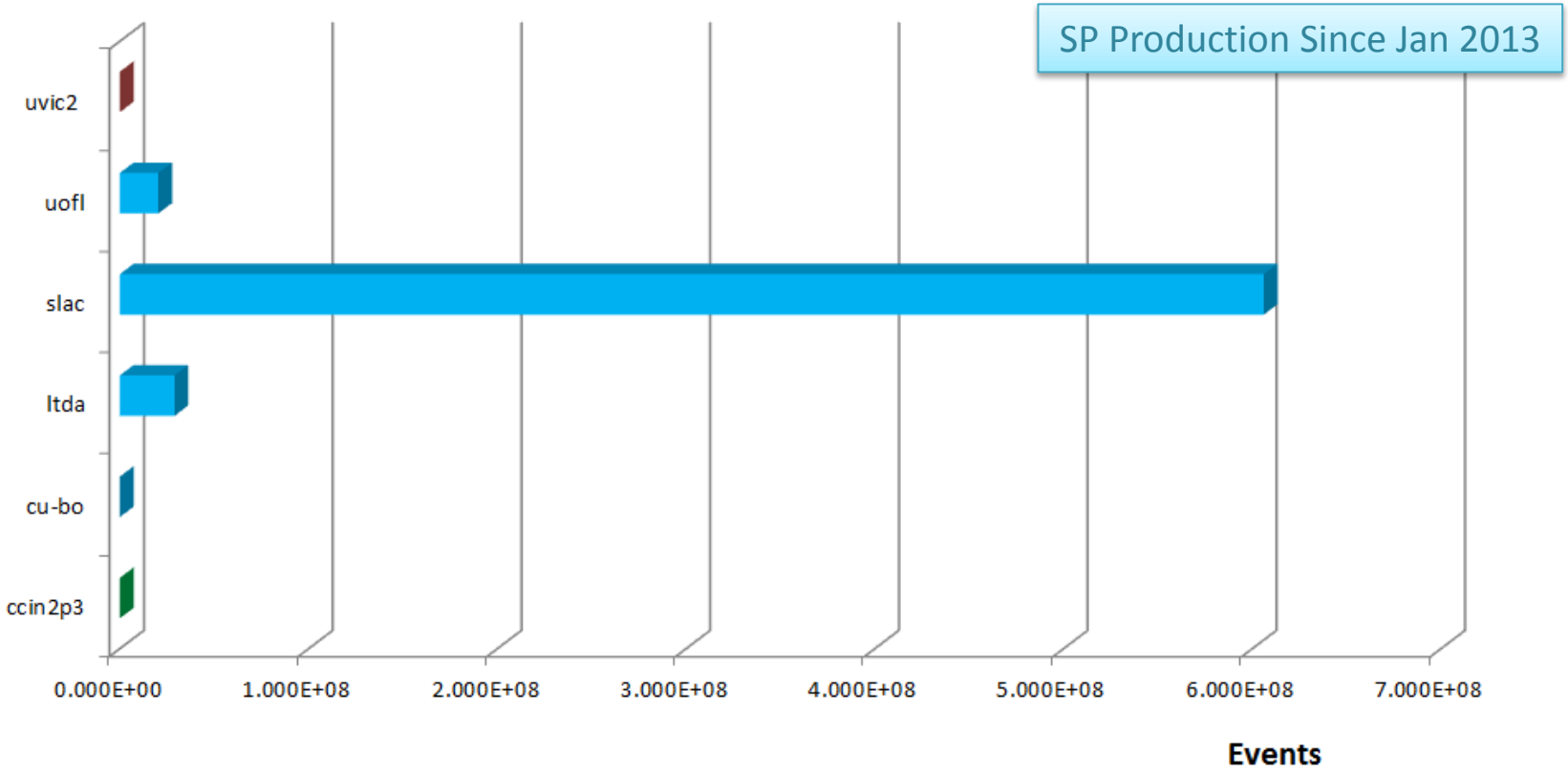
- The virtualization layer uses qemu with kvm support directly
  - Moved away from libvirt due to instability
- Need to create the network interface for the VMs
  - 24 MAC addresses per host and usage status stored in local db

CVS Repository Moved to LTDA

- Code and analysis documents repository moved to LTDA.
  - An AFS instance is kept in synch with the LTDA.
  - For the users at SLAC the change is transparent.
  - The users working from their own machines will need to setup the connection to the new location but it is very straightforward.
- The change makes it easier to develop code and interact with the repository from the virtual machines.



SIMULATION PRODUCTION (SP)



SP PRODUCTION MANAGER
Dave Brown (U. of Louisville)
SP PRODUCTION SUPPORT AND DEVELOPMENT
Douglas Smith

SITE MANAGERS
Dave Brown (SLAC/UofL)
Douglas Smith (LTDA)
Alessandro Gaz (cu-bo)
Nicolas Arnaud (IN2P3)

SKIM PRODUCTION

- Skim cycles
  - R24a1, R24a2, R24a3
  - R24c
  - R24d
  - R24e
  - R24f
  - R24g
  - R24h→ starting
  - R26a
  - R26b
  - R26c → finishing
- Skimming support/development
  - Douglas Smith
- Skimming Sites
  - SLAC: Douglas Smith, Rocky So
  - CCIN2P3: Homer Neal, Alexandre Beaulieu
  - GridKa: Chris Buenger, Marcus Ebert
  - LTDA: Douglas Smith