BaBar Data Preservation
Status Update

Tina Cartaro

BaBar Computing Coordinator

DPHEP Collaboration Workshop

June 9th, 2015

OUTLINE

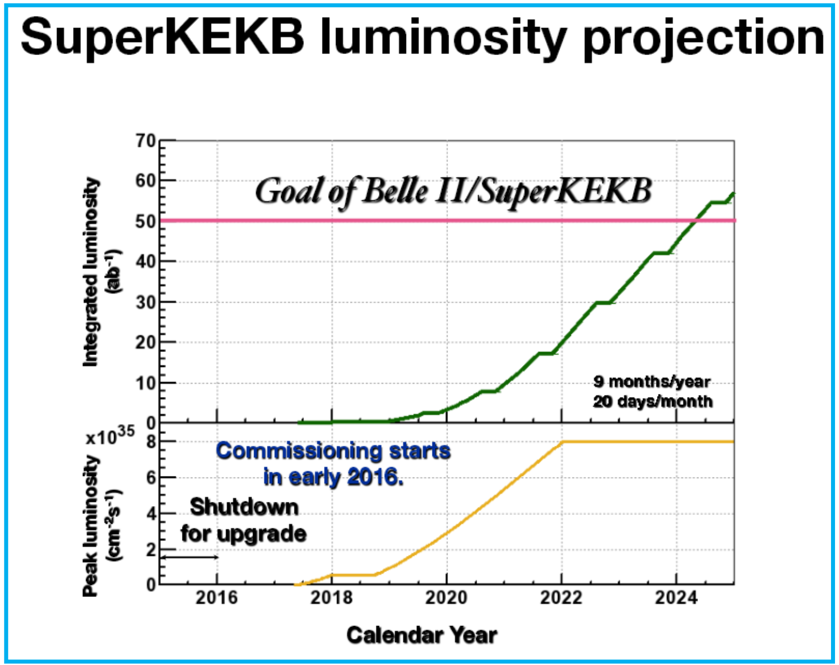
- *BABAR* quick summaries
- Status of the Long Term Data Access project
- New developments
- Documentation
- Global status of *BABAR* Computing
- Conclusions

BaBar Collaboration

- ~270 members from 67 institutions in 13 countries.
  - Plus ~100 associates

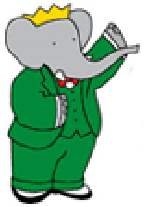
BABAR DATA

- *BABAR* has collected data from Oct 22nd 1999 to Apr 7th 2008
  - 800TB of raw data, 1.4 PB from the last data reprocessing
    - Total data on tape: 2.7 PB
  - 551 papers published to date
  - Over 30 on track analyses and ~20 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- *BABAR* (and Belle) data will not be superseded by LHC data.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 for *BABAR*
    - Υ(5S) dataset for Belle



SuperKEKB luminosity projection

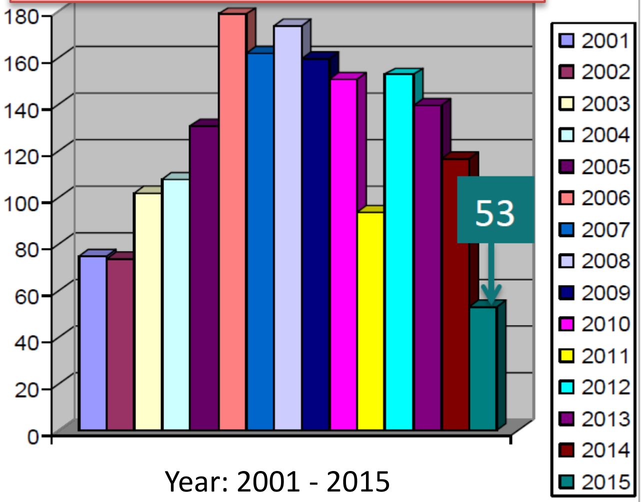
http://www-superkekb.kek.jp/documents/luminosityProjection150319.pdf



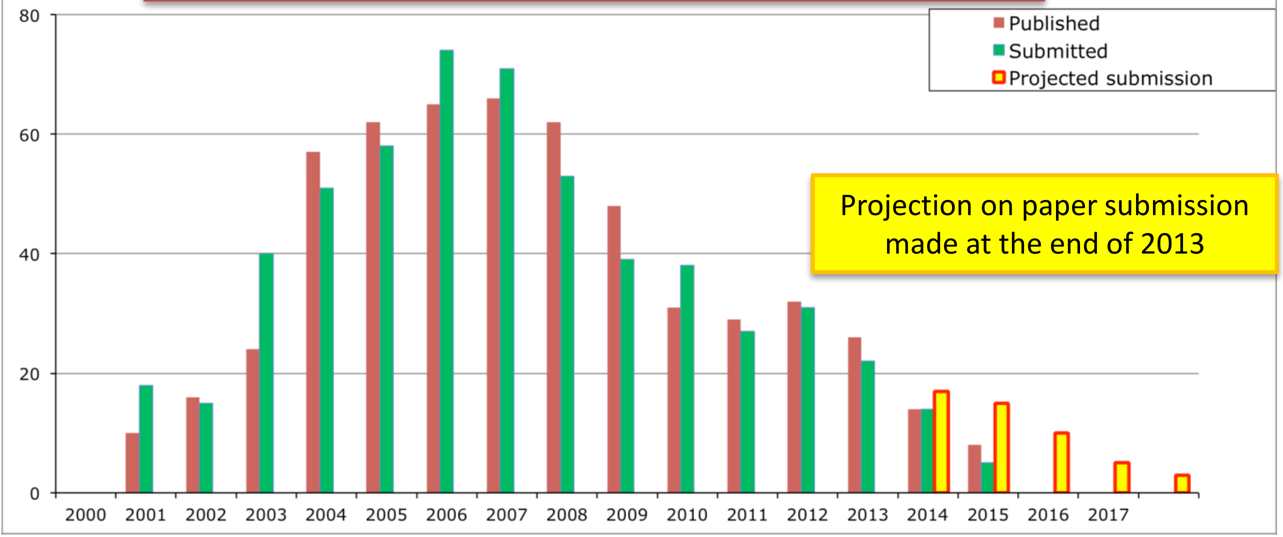
BaBar Publications/Talks

- 30+ analyses on track for publication, ~20 analyses have uncertain future
  - 8 new analyses started in 2014 and 3 more in 2015
  - 14 papers published and 14 submitted in 2014
    - Expect a bit more in 2015
- Accepted conference talks are limited by funding

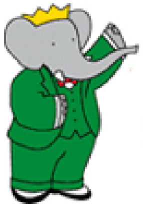
Conference talks by year



Papers published/submitted/projected by year

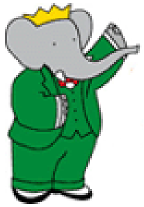


Projection on paper submission made at the end of 2013



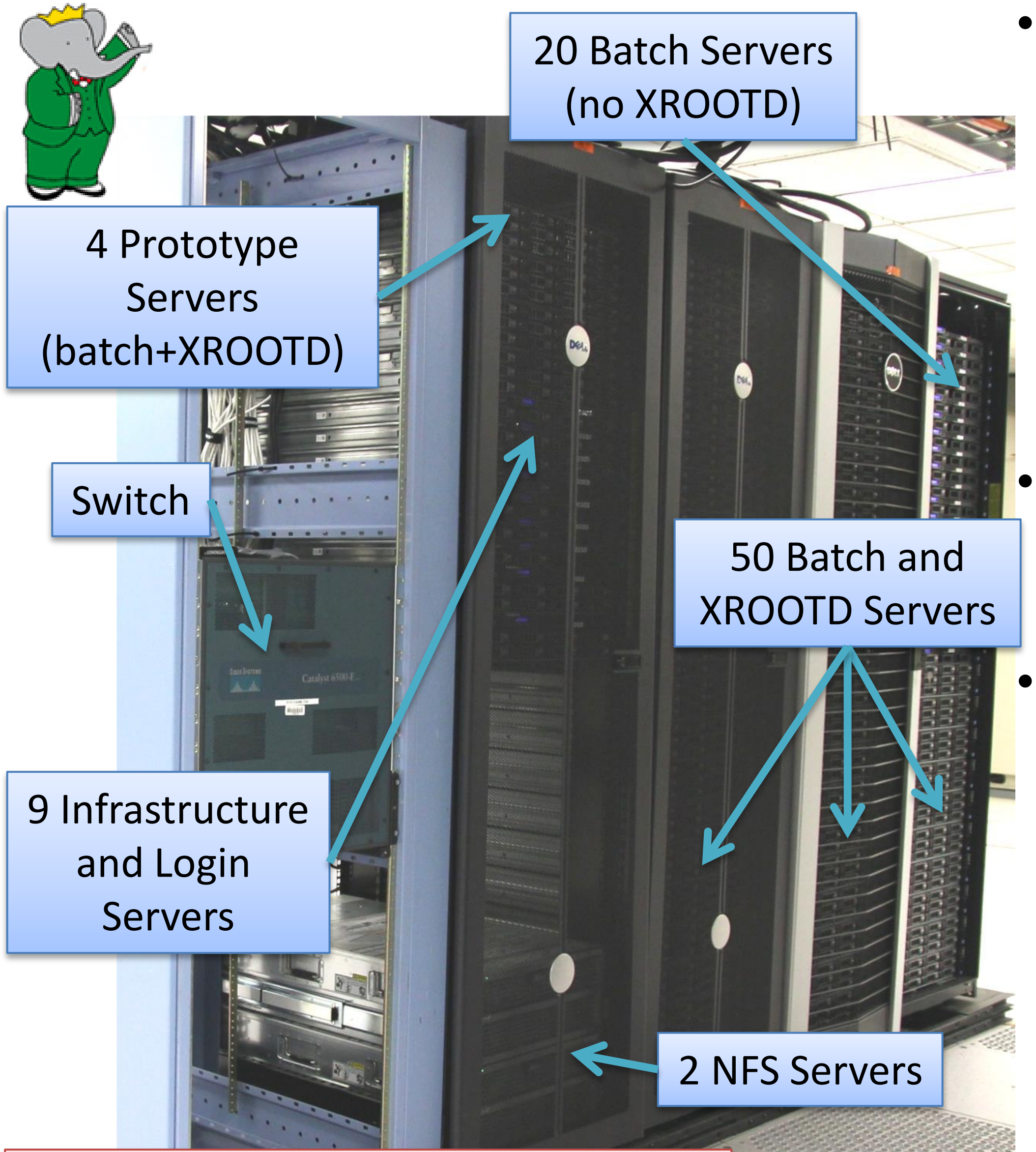
LONG TERM TASK FORCE

- In 2013, the Long Term Task Force surveyed the Collaboration in order to make predictions for the coming years and advice on how to prepare the Collaboration for the long term
  - Consolidate the governance structure
    - Activity still too high to let the various committees merge (executive board, publication board, speakers bureau, …) but it will happen soon
  - Main roles in the collaboration become permanent
  - Merge analysis working groups
    - We had 10, then 8, now 4
  - Freeze author list (easier to maintain)
    - Author list is generated every month from the *BABAR* database, then special adjustments are applied by hand
      - For example: temporary authors
    - Members have to actively sign up



LONG TERM DATA ACCESS

- Insure the ability to support analysis of the *BABAR* data until at least 2018
  - May be adjusted depending on Belle II schedule (aim to 2020?)
  - 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, OS, 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
  - Data is stored on tape at SLAC and CC-IN2P3 (back-up only)
  - Most used data sits on disk (XRootD)



20 Batch Servers (no XROOTD)

4 Prototype Servers (batch+XROOTD)

Switch

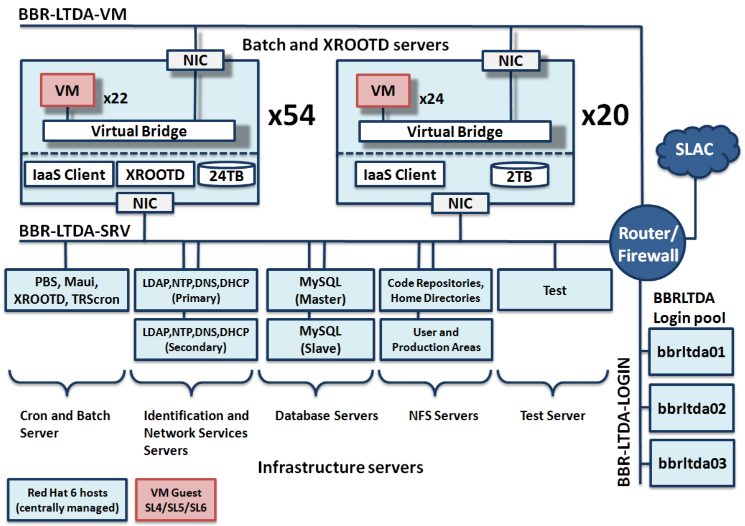
50 Batch and XROOTD Servers

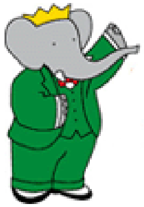
9 Infrastructure and Login Servers

2 NFS Servers

LTDA enclave went in production on March 21st, 2012

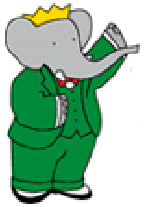
- LTDA Facts
  - 1.3 PB of disk for data and users
  - SL4, SL5, SL6 platforms available
    - Security threat associated to a VM running an old OS
  - 1668 job slots (Virtual Machines)
  - 1GB home for each *BaBar* user
- LTDA user-friendly environment
  - Interactive VM's for all platforms always available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





BABARTOGO

- A project for *BABAR* beyond 2018.
  - Portable and versatile, designed mainly for single user machines (laptop,…) but still possible to run it in batch queues
  - Take advantage of availability of expertise.
- *BABARTOGO* :
  - A 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, VirtualBox, or VMware.
  - Start *BABAR* environment, create a test release (local user code development area), compile code, and run binary to create analysis rootuples accessing data through XRootD.

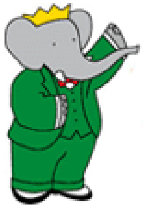


BaBarToGo XRootD Federation (I)

- On demand sharing of *BaBar* data among all collaborating institutions through the creation of a Federated XRootD Cluster accessible by all our VMs in the world.
  - Very similar to what ATLAS and CMS already have

XRootD

- Requirements: a small Linux machine with few 100s GB or few TB of space
  - Easy installation package
  - <u>Little or no manpower needed</u>, just make sure that the machine is up and connected to the network!



BaBarToGo XRootD Federation (II)

- Test installations ready at SLAC, GridKa, Rostock, Mainz, and McGill
  - INFN Ferrara (Italy) and Rutherford Appleton Lab (U.K.) volunteered to join
  - The LTDA enclave itself will be part of it
- Optimization tests for XRootD ongoing
- Developing software packages for data management
  - Installation package ready
    - Basic monitoring tools also included in the package
  - … and it takes two minutes to install and have a XRootD client running
- BEWARE: the XRootD Federation IS NOT data preservation!

Federated Datasets

- All *BaBar* data are distributed among *BaBar* institutions
  - Some data are at all institutions
    - Conditions files, for example
- No whole skims at a single place
  - For performance reasons
- Jobs don't need to know where the data are
  - Jobs can just access data files instantly like always before
- All data files have to be distributed redundantly in the cluster
  - institutions can use simple desktop machines with large disks
    - no need for raid

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)

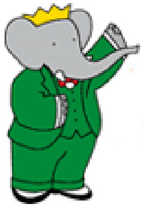
Skim	R24c skim size (TB)
BToDlnu	7.8
BSemiExcl	4.0

HARDWARE STATUS

- LTDA hardware is already over 4 years old
- Web servers are almost 4 years old
- Oracle database servers are 5 years old
- Other NFS servers range from 7 to 12(!!!) years old
- SLAC will stop support of SUN OS by end 2017 and all the Thors/Thumpers (SUN Fire X4540/4500) servers will be decommissioned by then
  - This affects many services
    - XRootD buffer for staging data from tape at SLAC
    - NFS servers that export the majority of user work areas
    - Inside the LTDA two Thors export home directories, *BaBar* file system (releases, reposotories, …) , and LTDA user work areas

NEW HARDWARE

- We are working to acquire new hardware within the next weeks
  - This will likely be the last hardware purchase we can afford
- Few powerful machines (some providing CPU and some providing disk) that will support the computing infrastructure
  - Requirements are not anymore too aggressive given the present and expected level of activity
  - We will continue to use the enormous amount of resources of the LTDA



TOWARD A NEW MODEL

- The technology at the base of our operating model will be virtualization
  - All the services now running on physical hardware will soon run on virtual machines
    - Web services like the wiki or the HyperNews and the production databases for SP, skimming and bookkeeping are some examples
      - this will likely force a transaction from Oracle to MySQL
        - » LTDA and Tier A sites already use exclusively MySQL
  - This will enable us to prolong the availability of the services beyond the lifetime of the hardware itself
    - OpenStack is being adopted at SLAC and this could open a number of possibilities for managing centrally all our virtual machines
  - User files can be stored through the XRootD Federation so that even local disk space will not be needed

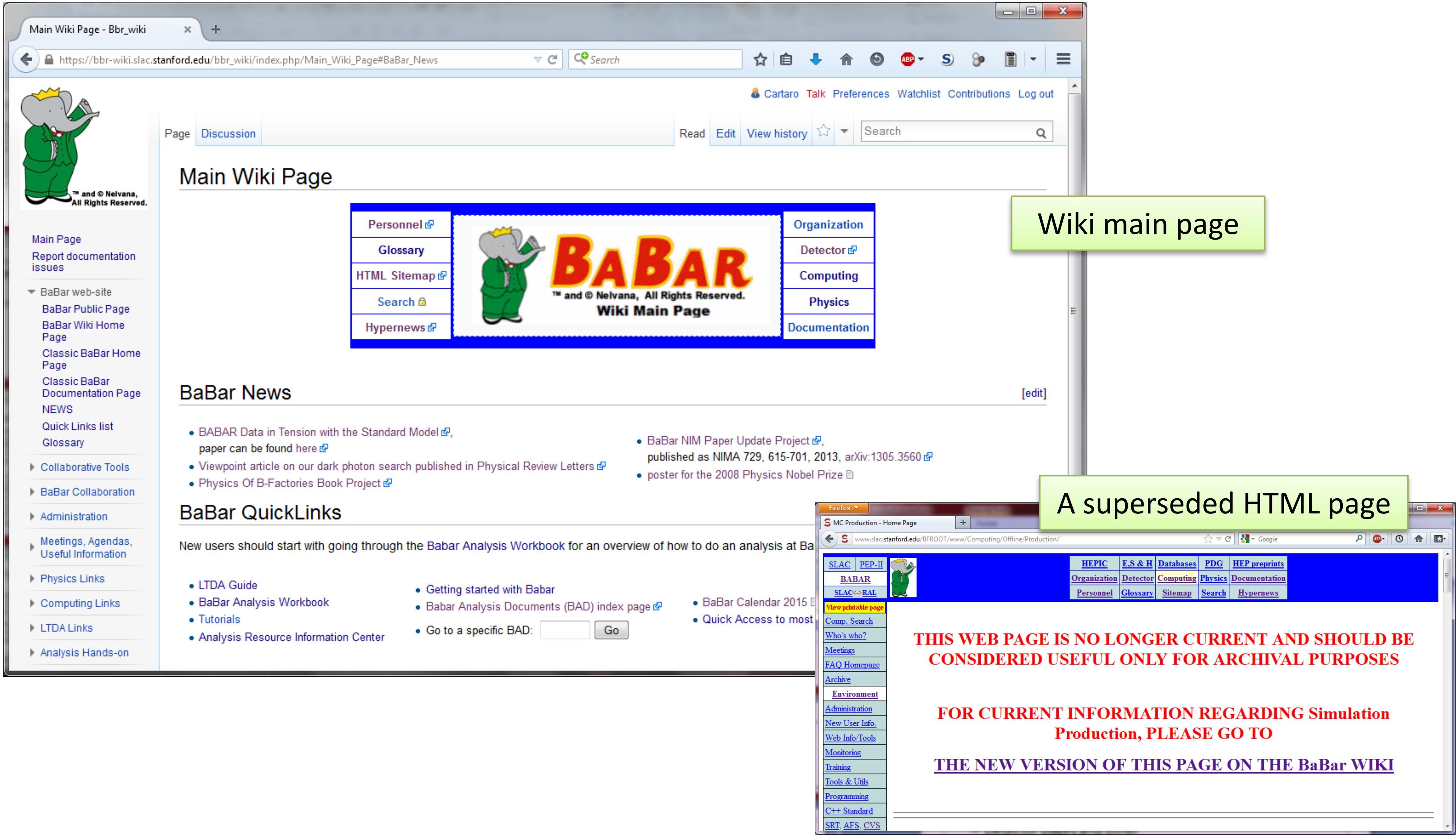
TAPE MIGRATION

- We are currently using 2.7 PB of tape to store our data: XTC (raw data), R22, R24 + R26 (latest reprocessing)
  - We useT10kB (1TB) tapes while SLAC migrated to T10kC (5 TB)
- Current finances will allow us to migrate 2PB to the next generation of media, T10kD (8.5TB)
  - Physically the same media of the T10kC but reformatted to allow for more capacity
  - SLAC has not yet migrated to the D type so we purchase the tapes and wait
  - R22 will not be migrated (D drives can read A/B/C tapes) and will eventually be dropped

BaBar Wiki Documentation Project

- Project completed
- 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
  - The wiki pages undergo a Collaboration wide review process and experts sign-off on the content of migrated/updated pages before they are officially released
  - Pages reviewed periodically
- Solidity of the documentation tested by new users

BaBar Wiki



Wiki main page

Main Wiki Page - Bbr_wiki

https://bbr-wiki.slac.stanford.edu/bbr_wiki/index.php/Main_Wiki_Page#BaBar_News

Cartaro Talk Preferences Watchlist Contributions Log out

Page Discussion Read Edit View history Search

™ and © Nelvana, All Rights Reserved.

- Main Page
- Report documentation issues
- BaBar web-site
  - BaBar Public Page
  - BaBar Wiki Home Page
  - Classic BaBar Home Page
  - Classic BaBar Documentation Page
  - NEWS
  - Quick Links list
  - Glossary
- Collaborative Tools
- BaBar Collaboration
- Administration
- Meetings, Agendas, Useful Information
- Physics Links
- Computing Links
- LTDA Links
- Analysis Hands-on

Main Wiki Page

Personnel	BABAR ™ and © Nelvana, All Rights Reserved. Wiki Main Page	Organization
Glossary		Detector
HTML Sitemap		Computing
Search		Physics
Hypernews		Documentation

BaBar News [edit]

- BABAR Data in Tension with the Standard Model, paper can be found here
- Viewpoint article on our dark photon search published in Physical Review Letters
- Physics Of B-Factories Book Project
- BaBar NIM Paper Update Project, published as NIMA 729, 615-701, 2013, arXiv:1305.3560
- poster for the 2008 Physics Nobel Prize

BaBar QuickLinks

New users should start with going through the Babar Analysis Workbook for an overview of how to do an analysis at Ba

- LTDA Guide
- BaBar Analysis Workbook
- Tutorials
- Analysis Resource Information Center
- Getting started with Babar
- Babar Analysis Documents (BAD) index page
- Go to a specific BAD: Go
- BaBar Calendar 2015
- Quick Access to most

A superseded HTML page

MC Production - Home Page

www.slac.stanford.edu/BFROOT/www/Computing/Offline/Production/

SLAC	PEP-II
BABAR	
SLAC<->RAL	

HEPIC	E,S & H	Databases	PDG	HEP preprints
Organization	Detector	Computing	Physics	Documentation
Personnel	Glossary	Sitemap	Search	Hypernews

- View printable page
- Comp. Search
- Who's who?
- Meetings
- FAQ Homepage
- Archive
- Environment
- Administration
- New User Info.
- Web Info/Tools
- Monitoring
- Training
- Tools & Utils
- Programming
- C++ Standard
- SRT, AFS, CVS

THIS WEB PAGE IS NO LONGER CURRENT AND SHOULD BE CONSIDERED USEFUL ONLY FOR ARCHIVAL PURPOSES

FOR CURRENT INFORMATION REGARDING Simulation Production, PLEASE GO TO

THE NEW VERSION OF THIS PAGE ON THE BaBar WIKI

Manpower

- By the end of CY2015 the official computing manpower supported at SLAC for *BaBar* will be 0.35 FTE
  - Some Collaborators are tirelessly and generously continuing their tasks
  - And many that have changed work or left the field still continue to offer their help when needed
- Already now the capacity of delivering services and support is extremely limited
  - No more data production cycles, only Monte Carlo production
  - *BaBarToGo*, XRootD federation, and the extended level of virtualization will hopefully keep data availability and infrastructure going on the long run
- 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 already now (1.55 FTEs)!
  - It is becoming more and more difficult to keep expert people due to funding pressure while all aspects of computing activity needs continue attention

Tier A Sites

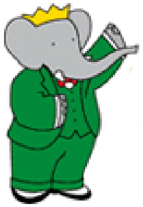
- CCIN2P3 continues to support analysis activity, SP and Skimming if needed and hosts a full copy of *BaBar* data (XTC, R22, R24, R26)
- GridKa provides analysis support and skimming and hosts the main redirector for the *BaBarToGo* XRootD Federation
- UVic and CNAF no longer support *BaBar*

CONCLUSION

- LTDA is extremely stable, dedicated to both users and production
  - LTDA hardware approaching the 5 years
- Tier A sites are lowering their support to *BABAR*
  - Appropriate level of support for the current status
- *BABAR* infrastructure (not LTDA) needs attention
  - Aging hardware
  - Tapes
- Efforts will focus on virtualizing all the services we can and distribute data and resources as much as possible
  - Distributing the data across a XRootD federation is not data preservation!
- International Finance Committee expectation is that SLAC/DOE will continue to support *BABAR*.

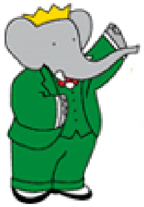
LTDA details

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

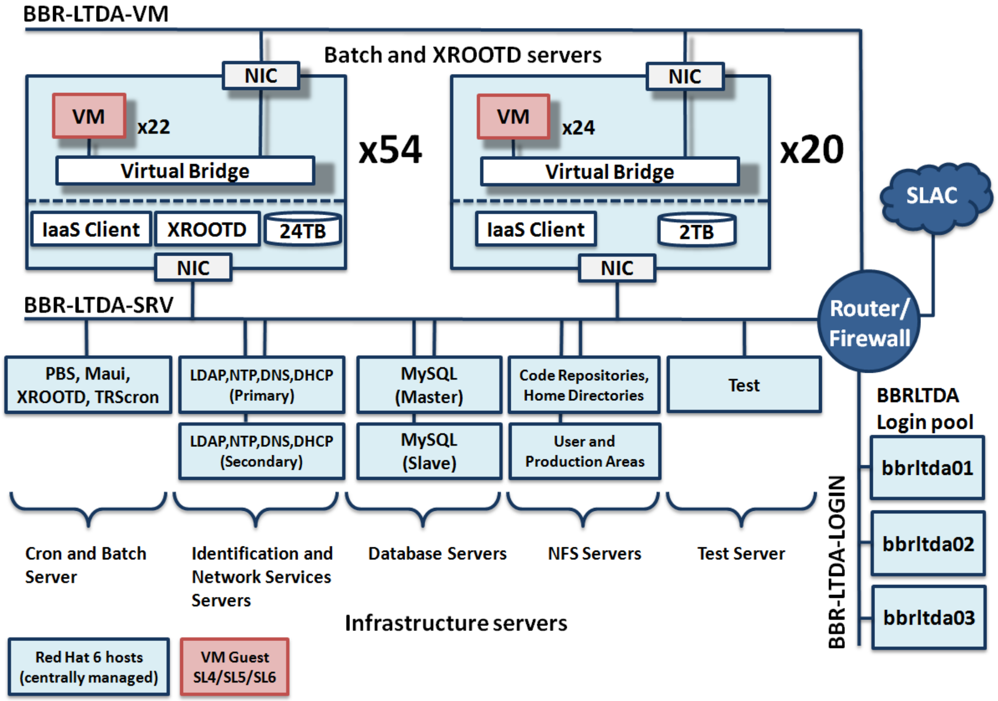
- NFS servers
  - 40TB zpools, 2 hot spares for 32 TB usable space
  - Compression enabled on local home directories, /home, with a factor ~2 gain
- ZFS snapshots implemented for /home and /BFROOT (releases, packages and cvs root directory) for user error recovery
  - snapshots are read-only, so it's 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

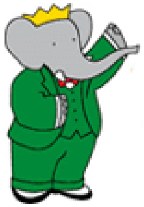
- Security threat associated to a VM connected to a network running old OS

→ 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

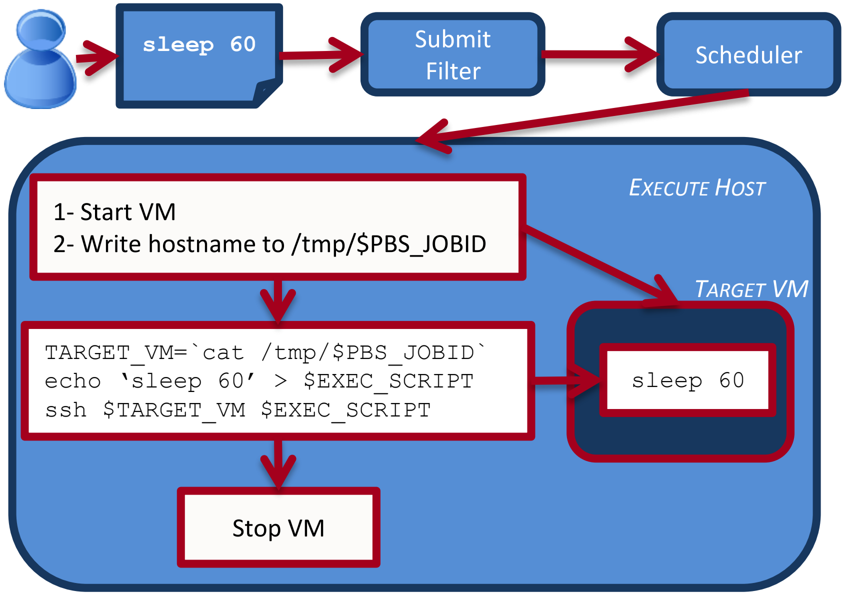


- VMs are not allowed to connect to SLAC network or the world
- The Login network is protected from the VM network
  - Allow one way ssh from Login to VM network
  - VMs are not allowed to write over the Login network
- Well defined services between VM network and SRV network
  - Infrastructure (DNS, 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



1- Start VM
2- Write hostname to /tmp/$PBS_JOBID

```
TARGET_VM=`cat /tmp/$PBS_JOBID`
echo 'sleep 60' > $EXEC_SCRIPT
ssh $TARGET_VM $EXEC_SCRIPT
```

```
sleep 60
```

- 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