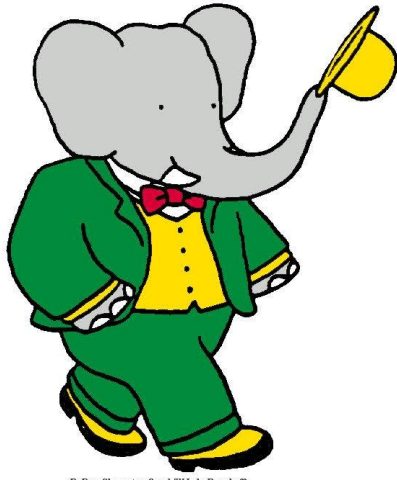


Status on the Preservation of BaBar's Data and Analysis ability

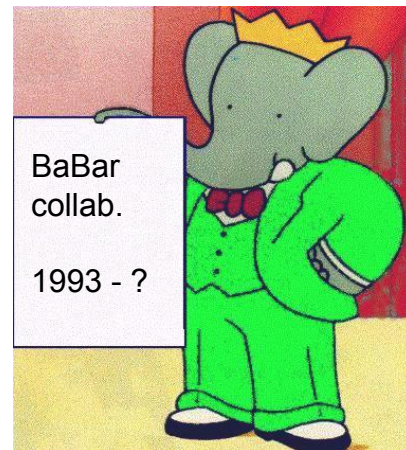
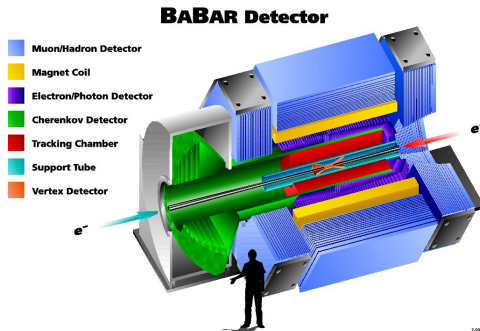
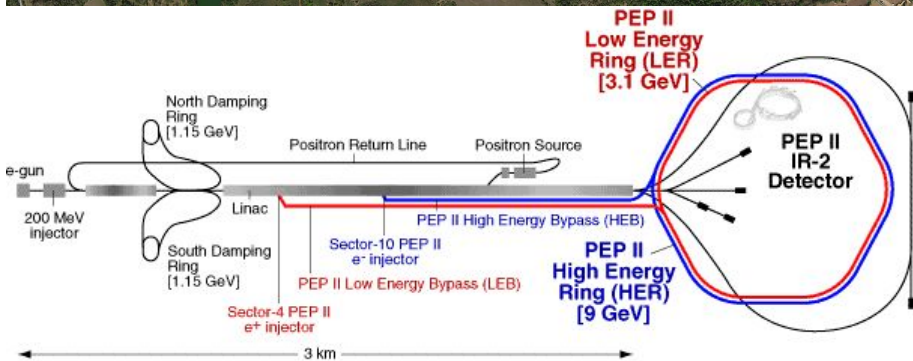


- Status update -

BaBar Experiment



- collider experiment at 
- BaBar founded 1993
- data taking 1999-2008



BaBar Status

- BaBar stopped data taking in 2008, anticipated to do data analyses until 2018
 - but still actively doing analyses (local, no Grid usage)
 - 223 active authors from 14 countries
 - 32 new analyses publications since 2018 (more than 60 incl. conference proceedings)
 - 5 analyses published in 2023 (not incl. conference proceedings)
- Beginning of 2021: support for infrastructure at SLAC finally stopped
 - support extended from 2018 to beginning of 2021
- “Everything” needed to be moved away from SLAC to be able to continue
 - very tightly integration of SLAC services and BaBar services, grown over years
- Moved “everything” to UVic
 - except for data

What is “Everything”?

● Data

- collected collision data and generated MC events (~1.5PB)
 - all in root files
- metadata stored in mysql database
 - number of events per root file, dataset,...

● Analysis environment

- software is 32bit, users usually write C++ code and compile their analysis modules
 - does not compile on 64bit-only systems
- depends on older software releases, e.g. perl, xrootd,...
 - latest verified system: SL6.3, gcc 4.4.x, kernel 2.6,...

● Documentation

- new users still join, sometimes just for a single analysis
- preserving documentation only way to have someone successfully started

● Collaboration tools

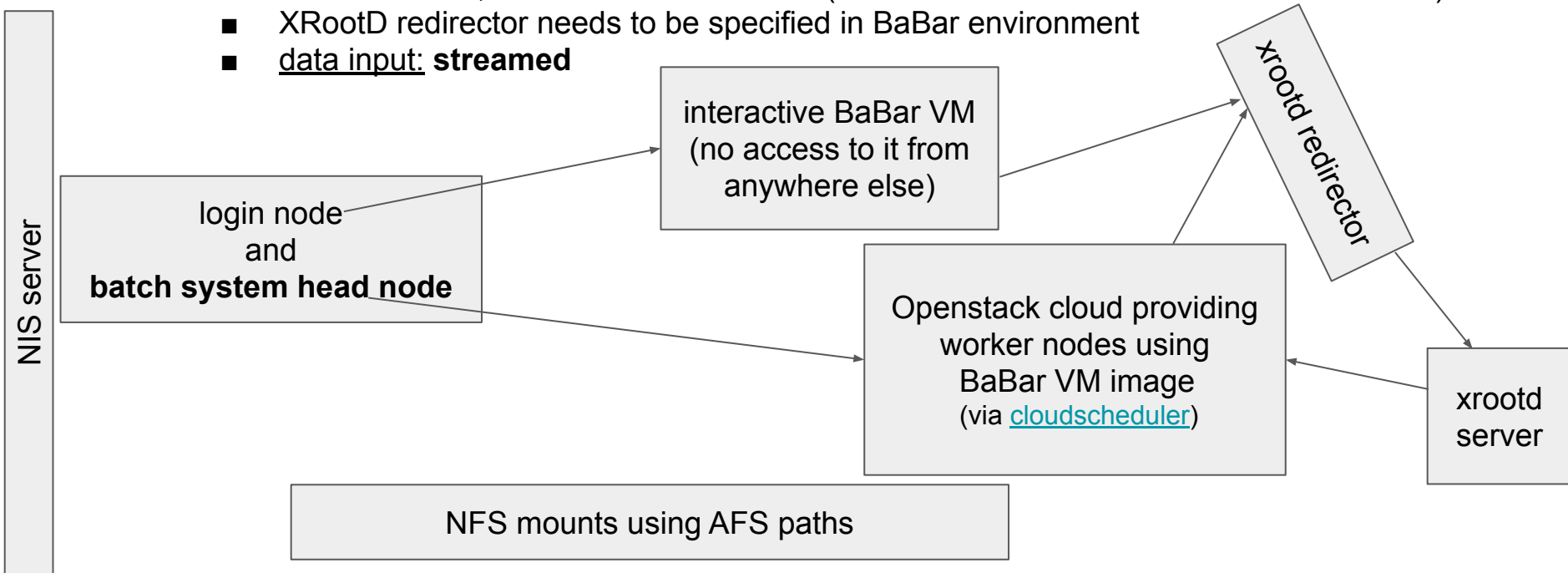
- calendar, analysis review, mailing lists, meeting organizer, ...

Data

- GridKa offered to store data and MC files from the latest processing run (AllEvents, skims, conditions db,...) for active usage
- GridKa also continues to host the metadata db (mariadb)
- IN2P3 hosts since a long time a second copy of all BaBar data, incl. raw data, as backup (not for active usage) and agreed to continue to do that
- CERN offered to also host a copy of all data

Data Access

- access needed from interactive machine and from worker nodes at UVic
 - BaBar uses XRootD, built into the framework (users do not need to care where the data is)
 - XRootD redirector needs to be specified in BaBar environment
 - data input: streamed



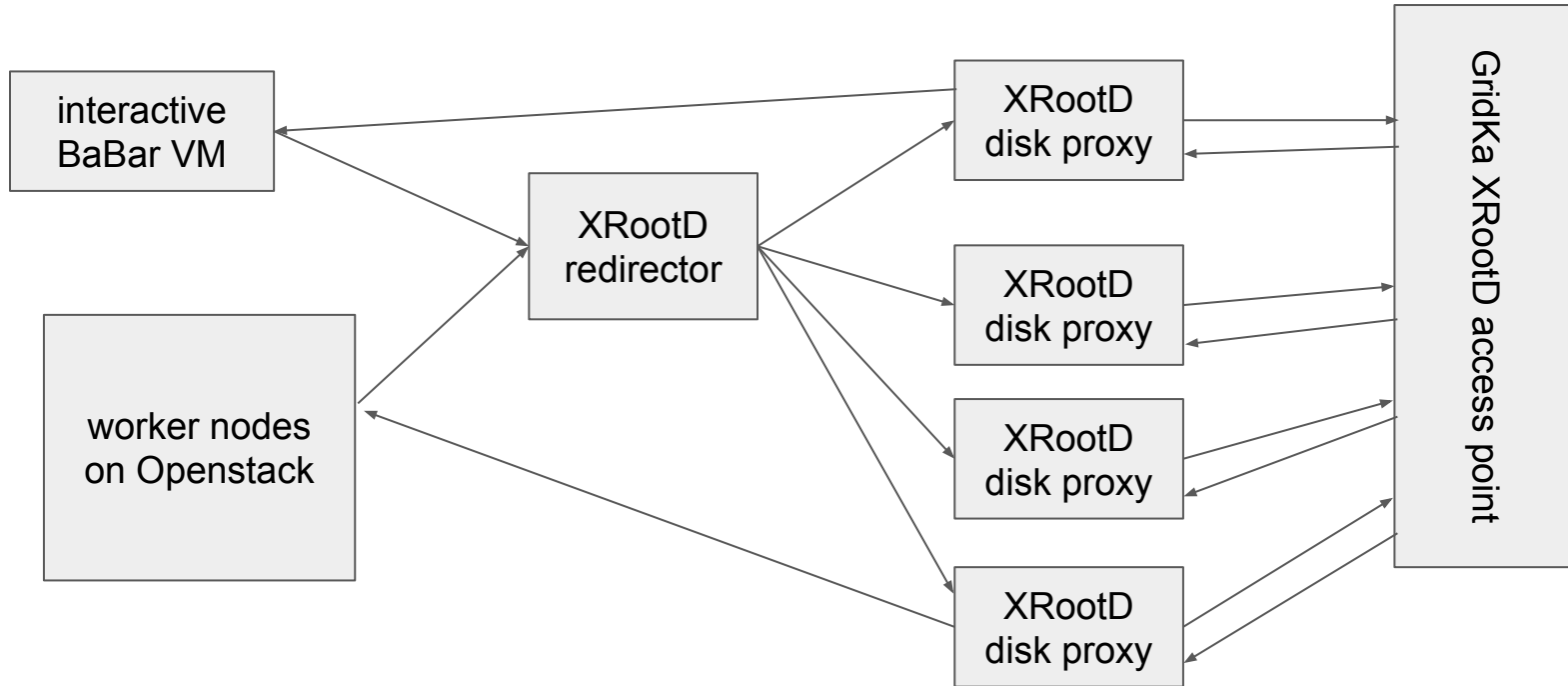
Data Access

- Data available to analyses: ~1.5PB
 - no storage available at UVic for it
- GridKa agreed to host the BaBar data to be used by analyses
 - BaBar site since a long time
 - had already some data on site; anything missing was copied to GridKa
 - also the metadata db to find the data files needed in an analysis was already there
 - only update of content needed
 - BaBar environment configuration specifies XRootD and db access point for the data and db
- Framework at UVic needs to access data at GridKa via streaming...
 - works surprisingly well for normal event data
 - workflow: read event, process, read event, process,...
 - but conditions data is also read via streaming
 - large amount of data each job needs to read



Data Access

- direct access to GridKa ---> access via local cache system



Documentation

- different systems used:
 - [html web pages](#): in AFS within well defined directory structure, r/w rights via ACL, every BaBar user had a SLAC account; edit html files directly in AFS
 - [Wiki](#): added ~2012 to have self contained system editable by anyone in the collaboration via web browser
- html web pages: visible to public or specific groups via .htaccess files, difficult to maintain content
- Wiki: visible only to BaBar members, easy to maintain content
- two new web servers at UVic and a single public web page (rest got access restricted to BaBar members)

Collaboration Tools

- SLAC based mailing lists ---> [Caltech mailing lists](#)
 - only created what is still needed
- old meeting agendas were HTML pages, registration based on SLAC systems
---> switch to use [CERN Indico](#)
- Hypernews was deeply integrated into SLAC
 - sending emails for posts to SLAC emails, notify SLAC systems in case of issues, people joining need SLAC UNIX account,... - but all content of posts in text files
---> moved Hypernews to UVic, made read-only, and removed any mailing feature -> still readable and archive of any communication happened in the past
---> replacement: [CERN egoups](#)
 - also nicely integrated with CERN Indico for accessing BaBar meetings

Collaboration Tools

- Analysis documents, notes, and Analysis metadata
 - old content [archived to INSPIRE](#)
 - new documents will be added too for long term preservation

new system for active analyses and management:

- [Google drive](#) folder for each analysis
 - for documents and other informations
- Google sheets for metadata of each analysis
- review done using CERN egroups (each analysis has its own)
- specific folders for SpeakersBureau, PublicationBoard,...

Open Data

- making data openly available is possible, but not useful by itself
- to make use of the data one also needs
 - Analysis framework
 - Documentation
 - Communication with collaboration members

'BaBar Associates' open-access:

- anyone can join (== data access for anyone)
 - full access to communications and documentation tools and archives
 - analyses for publication to be done within BaBar publication framework
 - e.g. going through the full review process
 - https://babar.heprc.uvic.ca/www/join_BaBar.html

Access to BaBar framework: analysis system at UVic, BaBar-To-Go (VM) at home

Unresolved Issues and Possible Future Issues

- part of the db containing luminosity information as well information about generated MC events are in a SLAC oracle database
 - information from different oracle databases are queried using results of one in the query to another
 - not clear which parts of the oracle databases at SLAC are needed
 - no oracle instance available outside of SLAC; access to the SLAC db not possible from outside of SLAC
 - to use the information: need to dump all BaBar related Oracle databases at SLAC, convert to be used in mariadb, change BaBar code to access mariadb instead (ongoing)
- hardware reliability and replacement
 - very little funding available
 - hardware replacements unsure

Redundancy/Reliability

Hardware overview:

- XRootD proxy server: **old machines**
- XRootD redirector: VM on an **old machine**
- login machine: VM on an **old machine**
- BaBar interactive VM: VM on an **old machine**
- NIS server: VM on an **old machine**
- web server: on VM on an **old machine**
- babar wiki: VM on an **old machine**
- babar Hypernews: VM on an **old machine**
- NFS server: **one new server**, multiple **old machines**

old machine==out of warranty

Redundancy/Reliability:

- protect against disk failure
- protect against server failure

Redundancy/Reliability

login machine VM
NIS Server VM
interactive VM
XRootD redirector VM

hardware raid1 OS
ZFS mirror data disks

- spare server setup the same way
- ZFS send/receive

XRootD proxy server

hardware raid1 for OS
ZFS raidz3 data disks

- multiple servers available
- just cache, loose no data

Web documentation VM
Wiki VM
Hypernews (HN) VM

hardware raid1 for OS
ZFS raidz3 for data disks

- web content on NFS
- HN content on NFS
- images backed up
- daily mysql dump to NFS

4 NFS server:
NFS \$HOME
NFS job output
NFS framework
NFS documentation

all use:
ZFS raidz2/3
hardware raid1 for OS

- spare server setup in the same way
- ZFS send/receive
- extra backup of framework and documentation

Conclusion

- analysis and documentation framework in a well defined directory structure
- analysis environment preserved in VM images
- open data access for anyone via “Associate Program”
- new analysis and documentation infrastructure at UVic with data at GridKa
- data backups at IN2P3 and CERN, documents backed up at INSPIRE
- few unresolved issues related to Oracle db content at SLAC

Conclusion

- analysis and documentation framework in a well defined directory structure
- analysis environment preserved in VM images
- open data access for anyone via “Associate Program”
- new analysis and documentation infrastructure at UVic with data at GridKa
- data backups at IN2P3 and CERN, documents backed up at INSPIRE
- few unresolved issues related to Oracle db content at SLAC

Big Thanks

to

**GridKa, CERN, IN2P3, INSPIRE, Caltech, and UVic HEPRC
groups!**