LTDA Status and Design

Tina Cartaro
SLAC
On behalf of
BaBar LTDA Group

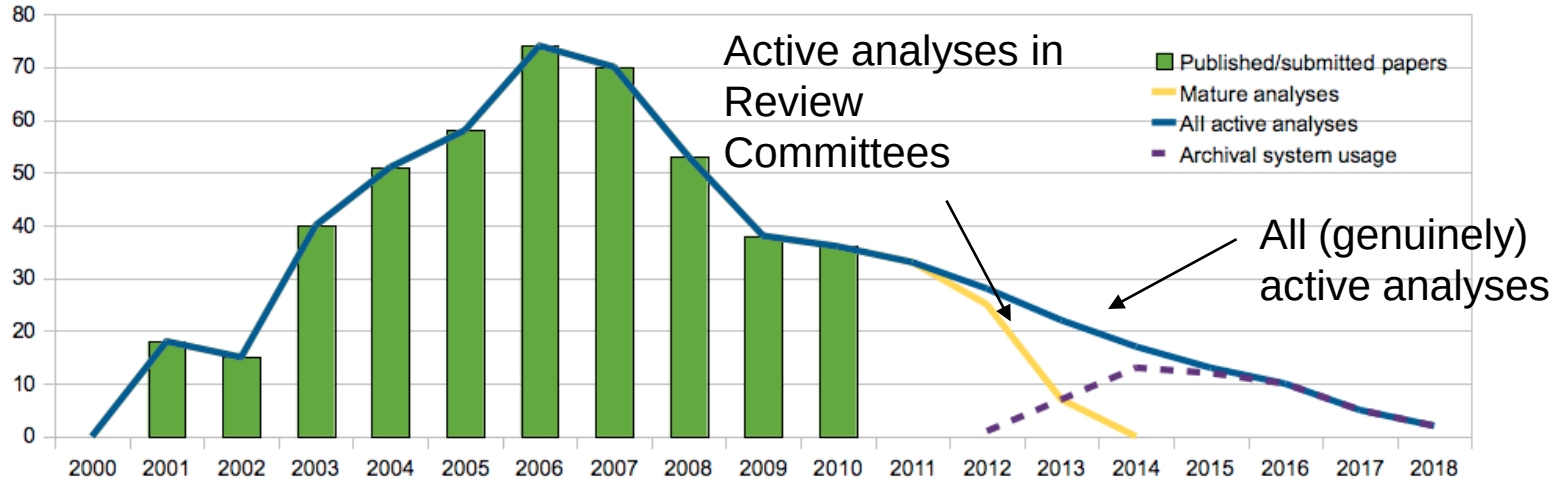
DPHEP V
Fermilab
May 17th, 2011

Outline

- BaBar data and choices for the future
- The Long Term Data Access project
- Requirements and design of the LTDA
- Milestones
- Budget, expertise, and long term planning
- Conclusions

BaBar

- BaBar has collected good data from Oct 22nd 1999 to Apr 7th 2008
  - 800TB of raw data, 1.2 PB from the last data reprocessing
  - 453 published papers to date
  - 126 active analyses
- BaBar (and Belle) data will not be superseded by LHC data
  - Belle II and SuperB will do it in 5-10years



Beyond 2012

- The BaBar Long Term Data Access project aims to preserve both the data and the ability to do analysis until at least 2018 and will provide support for >50 publications foreseen beyond 2012
  - Need to account for the dwindling resources (both manpower and money)
- We need to minimize the efforts needed to maintain the system
  - Validations, upgrades (OS, tools, …) , documentation, hardware maintenance, …
- Close BaBar Framework into a frozen environment
  - Simpler to maintain documentation and provide support
- Still not very easy on the long run
  - Hardware support and lifecycle
  - Maintain out of date OS
    - Potential security risks
    - Need to keep know-how about the old OS and the Framework

Long Term Data Access

- Insure the ability to do analysis on the BaBar data beyond 2012 preserving:
  - Data, conditions and calibrations
  - Releases and tools
  - Databases
  - Capability of doing production
  - Capability of running user jobs
  - Documentation
- Providing a stable environment
  - Last validated OS enclosed in a virtualization layer running the BaBar Framework
- Open formats
  - Data format is based on ROOT which is open and will be part of the system
  - Databases will move away from Oracle and will be stabilized on MySql
  - Code is written in open formats: C/C++, Tcl, Perl, Python.
- Data Storage
  - 2PB will be stored on tape in two Tier A sites (SLAC, CC-IN2P3)
  - Most used data will sit on disk

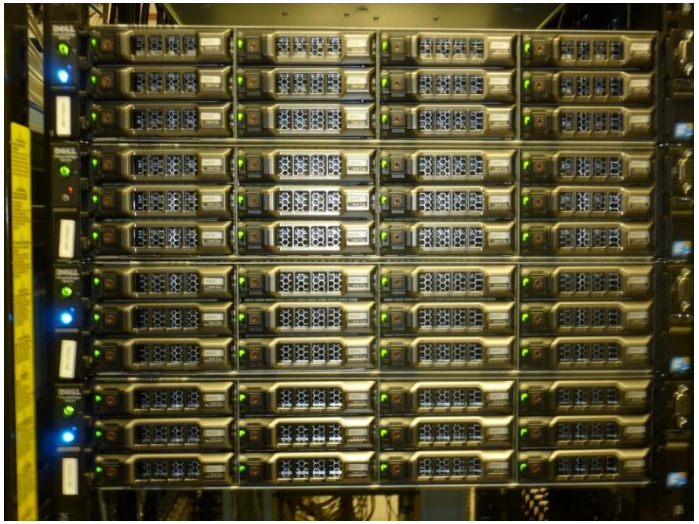
Virtualization

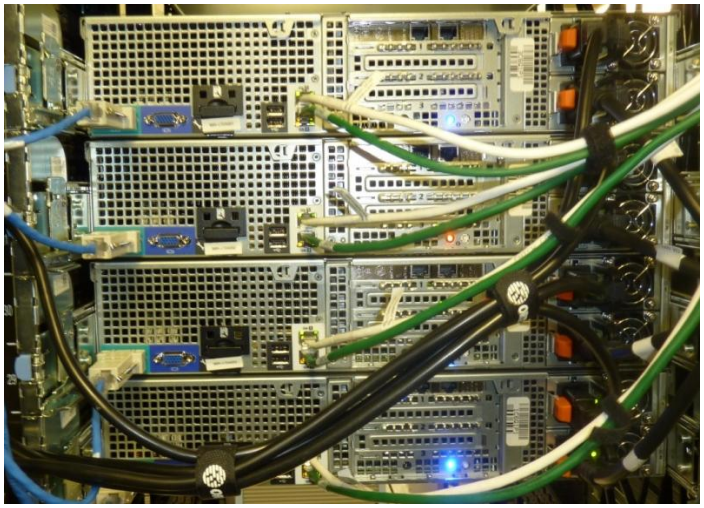
- Virtualization can provide virtual hardware support on which we can run virtual machines (VMs) with the wanted OS
  - Hardware support problem solved for the foreseeable future
  - Managing a small number of images is easier than managing a large physical cluster
- However
  - A VM connected to a network is not different from a physical machine and running old unpatched OS poses a security threat
    - The dynamic creation and destruction of the VMs adds a small amount of security
  - Risk based approach assuming that the VM are compromised

Design Requirements

- Assume the back versioned OS are compromised
  - The LTDA system shall not be able to harm other systems at SLAC or outside
    - Isolation of compromised components
  - The LTDA system shall prevent accidental modification or deletion of data
    - Nearly impossible to protect against intentional acts
    - Maintain user identity for access to old OS; it can be done in simple ways (LRM, ssh, …)
  - Detect all compromised elements
- Directly affects the network architecture
  - Isolation of back versioned components
  - Physical hosts centrally managed by SLAC CD
  - Firewall rules

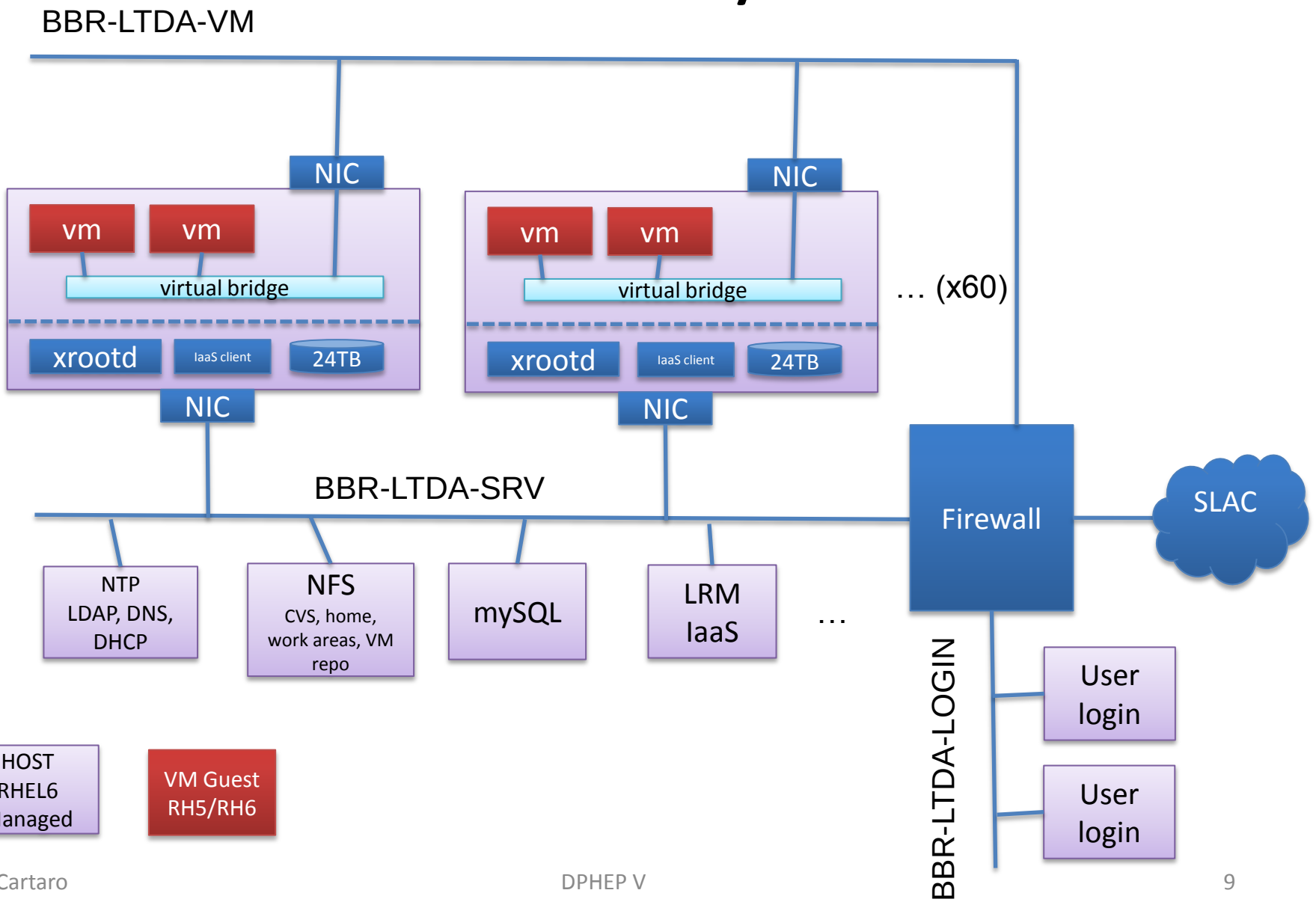
Hardware





- 60 servers Intel dual 6-core, 3GHz, 48GB RAM, 24TB disk
  - 4 existing systems in prototype
  - Purchase remaining systems in 2 stages
- Cisco 6500 network switch (Gbit Ethernet)
- NFS server
- A few dedicated infrastructure & login servers
- Distributed storage and computing resources
  - Each node provides both CPU and storage

Network Layout



Subnets

- Three subnets and one switch to implement the filtering rules between the subnets
  - Login network
    - Open to SLAC network
    - This is as far as the users go
  - Server network
    - File server, DB server, infrastructure (DHCP, DNS, NTP, LDAP)
      - LDAP is a subset of the SLAC Kerberos list mapped on /nfs internal home directories
  - VM network
    - Well defined rules between the VM subnet and the other two

Firewall

- VMs will not be 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
- Allow SRV and Login networks to use SLAC infrastructure

From / To	SLAC	BBR-LTDA-LOGIN	BBR-LTDA-SRV	BBR-LTDA-VM
SLAC	1	ssh, infrastructure	Infrastructure, tape access, backup	X
BBR-LTDA-LOGIN	infrastructure	1	ssh, NFS, infrastructure	ssh (for interactive VM)
BBR-LTDA-SRV	infrastructure, tape access, backup	NFS	1	LRM, DHCP, xroot, NFS, LDAP
BBR-LTDA-VM	X	X	LRM, DHCP, xroot, NFS	1

LTDA Cloud

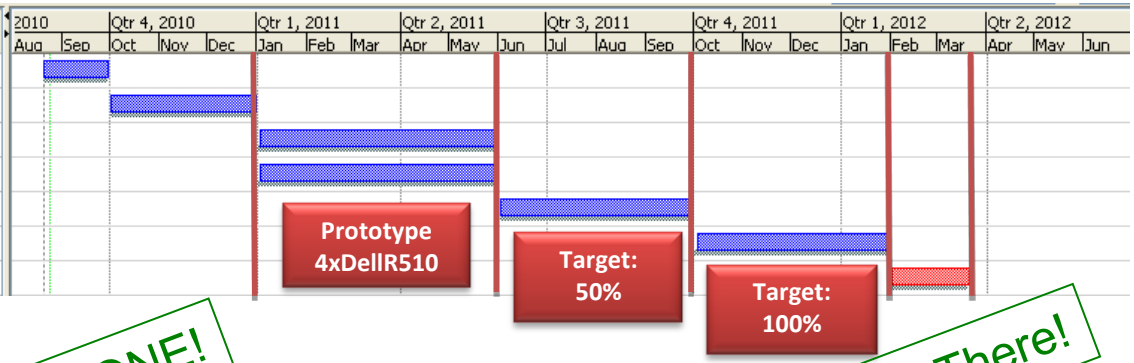
- LTDA will run hundreds of VMs
  - KVM, one VM per core
    - Testing performance with hyperthreading
- Appropriate robust tools to operate a virtual data center
  - Create, configure, assign network address and destroy VMs
- Infrastructure As A Service (IaaS)
- IaaS creates instances of VMs then a batch scheduler (LRM) assign the jobs
  - Right now LTDA uses Nimbus as IaaS and Condor as scheduler
  - Moving to PBS/Torque
    - Easier to configure and manage

From the User Point of View

- Users will login on the bbrltda load balanced pool and will find an environment quite similar to the one they are used
  - Create a job submit script.
  - Submit the jobs and monitor the results.
  - If necessary, starts an “interactive” session for debugging.
- Limitations are due to the restrictions on the VMs
  - The code cannot be edited from a VM if it resides in the home directory
  - Any change done to a VM will disappear with the VM
  - The VM environment is limited and designed to do only one thing: run and debug BaBar code

Main Milestones 2010-2011

	Name	Start	Finish
1	CEP+PO LTDA Prototype	8/20/10 8:00 AM	9/30/10 5:00 PM
2	Installation/configuration complete	10/1/10 8:00 AM	12/31/10 5:00 PM
3	Run7 available for test users	1/1/11 9:00 AM	6/1/11 5:00 PM
4	Fix, test, tune. PO for final design	1/1/11 9:00 AM	6/1/11 5:00 PM
5	Extended system ready (25% to50%)	6/2/11 8:00 AM	9/30/11 5:00 PM
6	100% system acquired	10/1/11 8:00 AM	1/31/12 5:00 PM
7	Deploy LTDA	2/1/12 9:00 AM	3/22/12 9:00 AM



- Sep 2010
  - PO for the prototype
  - Prototype on site by October
- Dec 2010
  - Installation/configuration complete
- Jan 1st 2011
  - System available for test users

DONE!

DONE!

DONE!

DONE!

- Jun 1st 2011
  - Test phase ends, LTDA final design ready
  - PO for first 50%
- Jul 1st 2011
  - First 50% of LTDA available
- Oct 1st 2011
  - PO for 100% of LTDA
- March, 21st, 2012
  - Deployment of the LTDA
  - All new analyses will use LTDA

Getting There!

Beyond the Machines

- Designing and maintaining the LTDA Archival System through the years requires many talents and ahead programming
  - BaBar experts
    - Releases, databases, data management and documentation
    - The Collaboration will have to provide such expertise
  - Computing experts
    - Network architects, security, system and networks administration, virtualization, …
    - 0.5 FTE/year foreseen after 2012
      - Relevant impact on the budget …

Budget

- Hardware including hardware refreshment program
- RH licenses
- Red Hat entitlements for the use of virtualization on the 60 boxes
  - Need to negotiate with RH
    - From the web: 2000$ per machine per year!!!
  - We could use SL5/SL6 but it is not currently supported by SLAC Computing Division (extra cost for 0.5 FTE for SL support)
- Switch Cisco 6509 with 2x10Gb link card and 144Gb ports
  - A dedicated switch is desirable for security reasons, cleanliness of design, and physical mobility of the cluster
  - Negotiate with the Lab for part of the hardware
- We will move soon to Rhel6 that will be supported until Nov 2017. We could buy the extended support to bring us well beyond the end of 2018
- 0.5FTE/year from Computing Division to support the Archival System for its lifetime

LTDA is not just the Archival System

- CHEP
  - Talk on the LTDA by Douglas Smith
- DPHEP
  - Inter collaboration exchanges
  - Outreach
- Ensure survival of our data
  - Copy data offsite (CC-IN2P3)
- Documentation
  - Part of our legacy, not just manuals
- Governance
  - Protect the LTDA and BaBar defining now what our needs are for the years to come

Credits

- The LTDA current status is achieved through the hard work of many experts
  - BaBar software expert and beta tester: Homer Neal
  - Network design: Steffen Luitz
  - Virtualization expert: Kyle Fransham
  - System performance expert: Igor Gaponenko
  - Databases, tools and production: Douglas Smith and Tim Adye
  - Computing Division experts
    - System setup and adminstration: Booker Bense and Lance Nakata
    - Xrootd experts: Wilko Kroeger and Andy Hanuchevsky
    - Network setup: Antonio Ceseracciu
    - BaBar-SLAC liason: Len Moss
  - PPA and BaBar Management

The Advisory Committee

- The LTDA Advisory Committee
  - Justin Albert (BaBar), Fabrizio Bianchi (CSC/BaBar, Committee Chair), Rene Brun (CERN/ROOT), Nobu Katayama (KEK/BELLE), Richard Mount (SLAC/ATLAS), Jean-Yves Nief (CCIN2P3)
  - Cristi Diaconu just joined the AC
    - Thank you!
- Meetings with the AD have always been fruitful
  - Important suggestions at technical and management level

Conclusions

- LTDA is progressing quickly
  - Prototype infrastructure ready and working
  - BaBar Framework running
- DOE in general very supportive for the LTDA project
- Other activities going on as part of the LTDA
  - Documentation and Outreach
- Next big step: finalize the design and get ready to purchase the first half of the LTDA before the end of FY11