Preserving access to ALEPH computing environment via virtual machines

Simone Coscetti¹,², Tommaso Boccali², Marcello Maggi³ and Silvia Arezzini²

¹ Università di Pisa, Lungarno A. Pacinotti, 43, 56126 Pisa, IT
² INFN Pisa, Largo B. Pontecorvo, 3, 56127 Pisa, IT
³ INFN Bari, Via E. Orabona 4, 70125 Bari, IT
E-mail: simone.coscetti@pi.infn.it

Abstract. The ALEPH Collaboration [1] took data at the LEP (CERN) electron-positron collider in the period 1989-2000, producing more than 300 scientific papers. While most of the Collaboration activities stopped in the last years, the data collected still has physics potential, with new theoretical models emerging, which ask checks with data at the Z and WW production energies. An attempt to revive and preserve the ALEPH Computing Environment is presented; the aim is not only the preservation of the data files (usually called bit preservation), but of the full environment a physicist would need to perform brand new analyses. Technically, a Virtual Machine approach has been chosen, using the VirtualBox platform. Concerning simulated events, the full chain from event generators to physics plots is possible, and reprocessing of data events is also functioning. Interactive tools like the DALI event display can be used on both data and simulated events. The Virtual Machine approach is suited for both interactive usage, and for massive computing using Cloud like approaches.

1. Introduction
The datasets collected during many years of detector operations at particle accelerators offer unique opportunities for future scientific studies: new theoretical input, new experimental results and analysis techniques, the quest for high-sensitivity combined analyses, the necessity of cross checks. In many cases high-energy physics (HEP) data sets are unique and cannot superseded by data from newer generations of experiments. The cost of conserving this heritage through a long-term data preservation program would be small, compared to the costs of past experimental projects or to the efforts to re-do the experiments.

The HEP community will benefit from preserved data samples through reanalysis, combination, education and outreach.

HEP experimental collaborations summarize their scientific results through publications in journals or conference proceedings. Long-term preservation and re-use of the primary data beyond the published analyses is generally not contemplated in HEP. In most cases the basic (raw) data disappear in a period of five to ten years after the end of the data taking. The main reasons are rapid changes in storage technologies, and computing and software systems, which are not matched by an effort to migrate the data-analysis infrastructure and to maintain the expertise level needed for new analysis.

The scientific value of long-term analysis of HEP data is difficult to underestimate and well
understood by the HEP community. Long term preservation of HEP data is crucial to preserve the ability of addressing a wide range of scientific challenges and questions at times long after the completion of experiments that collected the data. In many cases, these data are and will continue to be unique in their energy range, process dynamics and experimental techniques. Some scientific opportunities for data preservation are summarized in the following points:

- long-term completion and extension of scientific programs;
- cross-collaboration analysis;
- data re-use;
- education, training and outreach.

1.1. The ALEPH experiment

ALEPH (Apparatus for LEP PHysics at CERN, figure 1) was a particle physics experiment installed at the Large Electron-Positron collider (LEP), at the CERN laboratory in Geneva/Switzerland. It was designed to explore the physics predicted by the Standard Model and to search for physics beyond it.

Figure 1. The ALEPH experiment detectors

ALEPH first measured events in LEP in July 1989. LEP operated at around 91 GeV, the predicted optimum energy for the formation of the Z particle. From 1995 to 2000 (end of the physics program) the accelerator operated at energies up to 209 GeV, above the threshold for producing pairs of W particles.

LEP ran until the year 2000, when it was shut down in order to make way for the Large Hadron Collider (LHC) which has been built in the same tunnel as LEP at CERN. The ALEPH collaboration has completed all its analyses and is essentially at an end.

2. The machine setup

A Scientific Linux CERN 4 (SLC4) distribution has been installed on VirtualBox. A fully functioning ALEPH environment needs the CERNLIB libraries, the ALEPH software and a full access to ALEPH data. The last native environment of the ALEPH experiment was based
on Linux Red Hat 6.2, a direct access to the CERN tape was possible and the software was entirely installed on the AFS file system with heavy dependencies on the CERNLIB. SLC4 has been used few years ago for the last official ALEPH analysis. A Scientific Linux CERN 5 (SLC5) would be suitable for this purpose, but for preservation goals it’s important to have a certified platform for analysis. The CERNLIB rpm is available in the SLC4 repository and no problems have been observed in the coexistence with SLC4 and ALEPH software. At the present time CERN hosts archival ALEPH data on the CASTOR storage system. This entails a slow and complex access to them. A solution that guarantees readiness and ease of use has been investigated, taking into account that nowadays ALEPH data and Monte Carlo sets occupy just a little amount of space (less than 11 TB): all the data have been moved to a current generation disk system at the data center that hosts the CMS Pisa Tier2, and served via the WebDav protocol. SLC4 supports davfs2, a Linux tool for connecting to WebDav shares as though they were local disks.

3. Software Components

Physics Analysis is usually done by comparing data with simulated events. New physics would require simulation and reconstruction to be able to produce new simulated data in the same format of the preserved data, and existing simulated events. For this reason all the components of the entire analysis chain together with the full documentation has been preserved:

- KINGAL - the event generators library;
- GALEPH - the Monte Carlo simulation program;
- JULIA - the reconstruction program;
- ALPHA - the physics analysis package.

With the complete set up of the ALEPH software, the compilation of new analysis and software works, not limited to ancient analysis. The system is suitable for interactive analysis, including all the previous functionality plus DALI (the event display, figure 2) and PAW [5]. More particularly, the entire chain has been tested in two different ways. In the first, small samples of real data have been taken and analyzed by ALPHA, then visualized through DALI. The other test has been made involving the whole analysis chain: a set of Monte Carlo events has been generated with KINGAL, then the simulation and the reconstruction have been made with GALEPH and JULIA, and finally analyzed by ALPHA.

4. Applications

A set of virtual machines like the one described in section 2 is available on a cloud supplied by the INFN-Bari data center (thanks to ReCaS [6] and PRISMA projects) for jobs submission. Such a system is implemented in an OpenStack instance which is able to control large pools of compute, storage and networking resources throughout the data center. The solution includes a complete development environment, where software components at any step can be modified, recompiled and debugged. Moreover, it can be used to reproduce published analysis and completely new studies. The whole software stack is available, for data and Monte Carlo sets, for interactive and batch processing.
5. Conclusions
A SLC4 Linux distribution has been installed in VirtualBox with fully ALEPH environment available: the virtual machine is completely independent from any AFS issues. Cloud computing for job submission is possible via an OpenStack instance hosted by the Bari data center. Data and Monte Carlo sets are available in a current generation disk system and they can be served by a single WebDAV instance.

The whole analysis chain has been reproduced. In this way the machine is ready and available for interactive usage, ancient analysis can be reproduced and new ones are possible. Moreover, software components at any steps can be modified. And finally, interactive and batch processing are possible.

References
[1] The ALEPH Coll. 1990 ALEPH: a detector for electron-positron annihilations at LEP Nucl. Instrum. and Methods A 294 121
[2] http://www.virtualbox.com
[3] https://cernlib.web.cern.ch/cernlib
[4] http://en.wikipedia.org/wiki/WebDAV
[5] http://paw.web.cern.ch/paw
[6] http://recas.ba.infn.it