КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2015 Т. 7 № 3 С. 415–419

Ки&М

ПЛЕНАРНЫЕ ДОКЛАДЫ

УДК: 004.75

Обновления аппаратно-программной базы ALICE перед вторым запуском Большого адронного коллайдера

А. К. Зароченцев[1,a], Г. Г. Стифоров[2,b]

[1] Санкт-Петербургский государственный университет,
Россия, 198504, г. Санкт-Петербург, Петергоф, Университетский просп., д. 35
[2] Лаборатория физики высоких энергий, Объединенный институт ядерных исследований,
Россия, 141980, г. Дубна, ул. Жолио-Кюри, д. 6

E-mail: [a] andrey.zar@gmail.com, [b] gleb.stiforov@cern.ch

Получено 27 октября 2014 г.

В докладе представлен ряд новостей и обновлений ALICE computing к RUN2 и RUN3.
В их числе:
– ввод в работу новой системы EOS;
– переход к файловой системе CVMFS для хранения ПО;
– план решения проблемы Long Term Data Preservation;
– обзор концепции “O square”, совмещающей офлайн- и онлайн-обработку данных;
– обзор существующих моделей использования виртуальных облаков для обработки данных ALICE.
Ряд нововведений показан на примере российских сайтов.

Ключевые слова: GRID, ALICE, CERN, LHC, WLCG, CVMFS, виртуализация

ALICE computing update before start of RUN2

A. K. Zarochentsev[1], G. G. Stiforov[2]

[1] Saint Petersburg State University, 35 University ave., St. Petersburg, Peterhof, 198504, Russia
[2] Laboratory of High Energy Physics, Joint Institute for Nuclear Research, 6 Joliot Curie St., Dubna, 141980, Russia

The report presents a number of news and updates of the ALICE computing for RUN2 and RUN3.
This includes:
– implementation in production of a new system EOS;
– migration to the file system CVMFS to be used for storage of the software;
– the plan for solving the problem of “Long-Term Data Preservation”;
– overview of the concept of “O square”, combining offline and online data processing;
– overview of the existing models to use the virtual clouds for ALICE data processing.
Innovations are shown on the example of the Russian sites.

Keywords: GRID, ALICE, CERN, LHC, WLCG, CVMFS, Virtualisation

Citation: *Computer Research and Modeling*, 2015, vol. 7, no. 3, pp. 415–419 (Russian).

ALICE: RUN1 и RUN2

В 2011 году завершился первый этап работы LHC, или RUN1. Для сохранения и обработки данных в RUN1 использовалась схема, которая полностью удовлетворяла потребностям эксперимента: вычислительная структура основана на GRID-системах WLCG и ARC, в которых VOBox используется для связи с GRID-структурой ALIEN (ALIce ENvironment) [Shabratova, 2010]. Структура хранения данных основана на xrootd-серверах, привязанных к сайтам через VOBOXEs. Структура доступа и обновления программного обеспечения (далее — ПО) основана на торрентах [Shabratova, 2012]. Авторизация организована через x509-аутентификацию и LDAP. Отдельно написан модуль libXrdAliceTokenAcc для авторизации xrootd. Мониторинг работает на основе MonALISA framework [MONitoring Agents…]. Более подробное описание структуры можно найти в докладах [Shabratova, 2010; Shabratova, 2012].

За первый этап эксперимент сохранил более 16 PB данных и постоянно обрабатывалось в среднем до 50 тысяч задач одновременно. На втором этапе работы LHC энергии столкновений возрастут более чем на 60 %, что даст значительный объем данных. В результате потребуется увеличить объемы хранения и производительность обработки более чем в 2 раза.

Решить задачу получится с помощью наращивания ресурсов и улучшения ПО. Например, улучшения систем обработки данных, качества доступа к хранилищам, системы доступа к обновляемому ПО, а так же виртуализации части ресурсов.

Обновления аппаратно-программной базы ALICE

Команда ALICE произвела ряд серьезных изменений в схеме хранения и обработки данных, которые введет к концу 2014 года, на более чем 90 сайтах по всему миру (рис. 1).

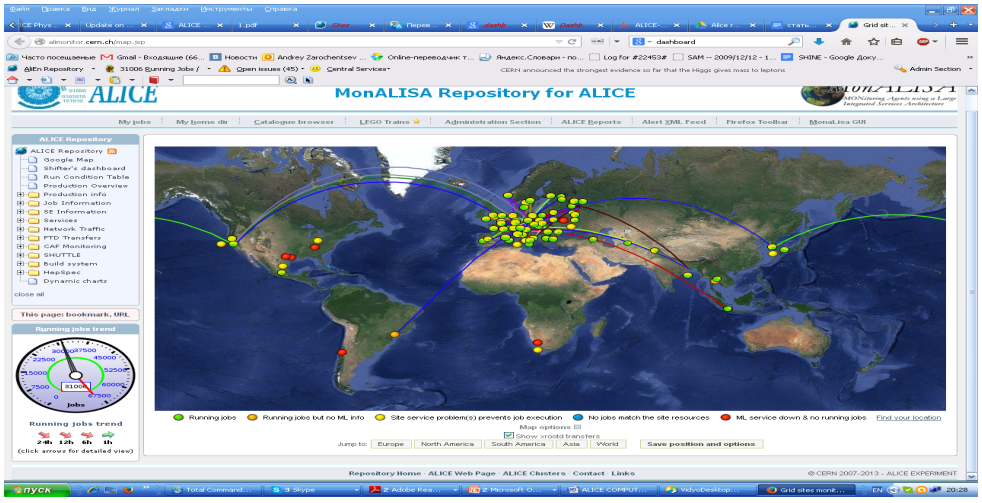

Рис. 1. http://alimonitor.cern.ch/map.jsp xrootd transfers

Например, к концу 2014 года все сайты перейдут на новую версию xrootd. В этой версии исправлены ошибки предыдущих пакетов, добавлена поддержка протокола IPV6 и новых файловых систем, в том числе распределенных, например CEPH [XROOTD project...]. Некоторые сайты планируют перейти на EOS, это система управления xrootd-серверами. EOS позволяет: централизованно администрировать серверы и объединять xrootd-серверы в RAID, что автоматизирует процесс восстановления системы в случае сбоя отдельных компонентов. EOS рекомендуется устанавливать как на сайты уровня Tier-1 (например, сайт RRC-KI-T1) с поддержкой всех компонент, так и на Tier-2 (например, сайт MEPHI) с минимальными требованиями к ресурсам и конфигурации. Подробнее о пакете EOS, для ALICE, можно узнать из докладов Андреаса Петерса [Andreas-Joachim Peters, 2013].

Долгое время для WLCG-сайтов была актуальна проблема с доступом к обновляемому программному обеспечению различных виртуальных организаций (ВО). Стандартным решением проблемы было предоставление менеджеру ВО доступа к некоей директории, доступной на нодах по NFS. На практике данный подход имел недостаток: файловая система NFS замедляла или останавливала работу сайта при 100 и более нодах. Дополнительно к этому обновление ПО зависело от работы VOBox, на котором этот автоматизированный процесс регулярно давал сбои и ошибки.

Альтернативный вариант: засылать весь необходимый обновленный код вместе с данными к задаче, так как самих данных существенно больше в задачах WLCG. В этом случае увеличивался трафик для каждой задачи, что суммарно для всех задач отражалось на производительности. В 2011 и 2012 годах команда ALICE предложила использовать для передачи пакетов ПО p2p-протокол или торренты [Shabratova, 2012].

Данный подход позволял не посылать полный пакет для каждой отдельной задачи. Взамен этого задачи опрашивали ближайшие ноды на наличие пакетов и скачивали необходимые файлы по частям с ближайших источников. В случае отсутствия ближайших источников необходимое ПО скачивалось по http-протоколу с центрального сервера. Использование http-протокола позволяло использовать кэширующий прокси, что экономило трафик, даже если администратор сайта по соображениям безопасности закрывал возможность использования торрентов. Такое решение позволяло экономить трафик по сравнению с вариантом прикрепления пакетов к каждой задаче и решало вопрос с надежностью по сравнению с вариантом использования NFS. В связи с недоверием локальных и сетевых системных администраторов к использованию торрентов зачастую все пакеты всё равно скачивались по http-протоколу. В этом случае трафик экономили только за счет использования прокси.

В 2012 году было найдено новое решение этой проблемы — базируемая на http-протоколе сетевая файловая система CVMFS (CernVM File System (CernVM-FS)) [CernVM File System…, CernVM 3 and…]. CVMFS позволяет использовать кэшируемый прокси для доступа к данным с рабочих нод и централизованно обновлять необходимые данные. В этом случае экономия трафика достигалась, как и в случае с торрентами, за счет использования http- и прокси-сервера. Но CVMFS также позволяет структурировать информацию, создавать отдельные ветви и многое другое**.** До середины 2013 года приняли все виртуальные организации (ВО) WLCG, а к апрелю 2014 года и ALICE планировала перевести все свои сайты на CVMFS, но в итоге все сайты перешли на данную сетевую файловую систему уже к январю 2014. Причем в настоящее время в ALICE computing model ПО с CVMFS используется не только на рабочих нодах, как у остальных ВО, но и на VOBox.

CVMFS изначально была разработана как файловая система для виртуальных машин для минимализации загрузочного образа и гибкости в конфигурации загружаемой системы. На рис. 2 приведена схема mCernVM (micro CERN Virtual machine). mCernVM представляет собой минимальный загрузочный образ, включающий ядро, модуль CVMFS объемом коло 12 МБ и файл contextualization около 64 КБ, включающий настройки подключения к CVMFS-серверу, выбора соответствующей ветки операционной системы и набора программного обеспечения.

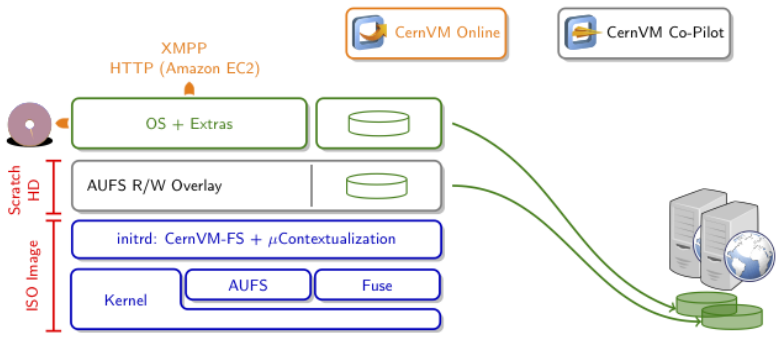

Рис. 2. Структура mCVM

Такой подход позволяет загрузить виртуальную машину с любым доступным набором операционной системы (ОС) и ПО, что дает возможность решить проблему длительного сохранения данных конфигурации для тех или иных вычислений (LTDR — long term data reservation).

В настоящее время довольно много составляющих системы обработки данных CERN переходит в виртуальную среду (HLT-кластер, виртуальные PROOF-кластеры и т. д.) и рассматриваются планы перехода в виртуальную среду и других компонент. В RUN3 планируется перевести вычислительные ресурсы ALICE на облачные системы. Эти системы повысят гибкость использования вычислительных ресурсов, что даст возможность использовать ресурсы поочередно (по требованию) для обработки «сырых» и накопленных данных (онлайн + офлайн). Подробнее о новом подходе к обработке данных “O square” в докладе Предрага Бунчича [Buncic, 2014].

В связи с обновлением структур хранения данных и системы доступа к ПО потребовалось обновление мониторинга. На данный момент у интерфейса http://alimonitor.cern.ch/map.jsp обновили google map до google maps API v3 и добавили возможность наглядно отслеживать xrootd-трафик.

Если ранее взаимодействие сайтов оценивалась только по VOBox, то сейчас отслеживается напрямую обмен файлами по протоколу xrootd (рис.1), как и другая информация по xrootd серверам [Grigoras, 2014]. Отдельно мониторится состояние CVMFS на сайтах [Grigoras, 2014; Grigoras Publishing…].

Страница со статусом отдельных сайтов и описанием проблем сильно облегчила работу администраторов и региональных менеджеров: http://alimonitor.cern.ch/siteinfo/issues.jsp. Значительно расширились возможности личного кабинета, откуда можно запускать собственные расчеты и отслеживать их выполнение в GRID- и AAF- (ALICE Analysis Facility) ресурсах.

Участие Российских сайтов в обновлении аппаратно-программной базы ALICE

Основное достижение российского сектора ALICE GRID — запуск сайта уровня Tier-1 на базе Национального исследовательского центра «Курчатовский институт», RRC-KIAE-T1. В данный момент сайт представляет 150 ТБ дисковых накопителей, более 4000 вычислительных слотов (ядер), планируется внедрять ленточные хранилища. Это сайт — первый из Tier-1, на котором был установлен EOS в качестве системы хранения. Кроме RRC-KIAE-T1 EOS был в 2014 году установлен еще на двух российских сайтах уровня Tier-2, это сайт Санкт-Петербургского государственного университета SPbSU и сайт Национального исследовательского ядерного университета «МИФИ» MEPHI. Про последний сайт стоит сказать отдельно — он был возвращен в активное использование, после двух лет простоя, только в декабре 2014 года.

Российский сегмент в 2013 году пережил спад производительности из-за проблем с сетью GEANT, однако уже к январю 2014 года полностью восстановил старые показатели. В отдельных институтах, таких как СПбГУ и ОИЯИ ведутся работы по адаптации облачных технологий для нужд ALICE computing.

Список литературы

A.-J. Peters EOS CERN Disk Storage, Varna, Bulgaria, 2013. http://nec2013.jinr.ru/files/12/Peters.pdf

CernVM File System. http://cernvm.cern.ch/portal/filesystem

CernVM 3 and μCernVM Beta Release. http://cernvm.cern.ch/portal/ucernvm

Buncic P. ALICE Computing Model RUN2, Tsukuba, Japan, 2014. http://indico.cern.ch/event/274974/contribution/33/material/slides/1.pdf

Grigoras C. News of MonALISA site monitoring, Tsukuba, Japan, 2014. https://indico.cern.ch/event/274974/contribution/87/material/slides/1.pdf

Grigoras C. Publishing ALICE data & CVMFS infrastructure monitoring. 2014. https://indico.cern.ch/event/321470/contribution/4/material/slides/1.pdf

MONitoring Agents using a Large Integrated Services Architecture. http://monalisa.cern.ch/monalisa.html

Shabratova G. The ALICE GRID operation, GRID 2010, Dubna, Russia, 2010. http://grid2010.jinr.ru/files/pdf/grid2010.pdf

Shabratova G. Torrent base of software distribution by ALICE at RDIG, GRID 2012, Dubna, Russia, 2012. http://grid2012.jinr.ru/docs/grid2012.pdf

XROOTD project. http://xrootd.org/