IT инфраструктура обработки данных эксперимента ALICE в Украине

Свистунов С.Я, Пелых В.В, Шадура О.В., Зубатюк Р. И., Баранник С. В., Соболев А.В.

Физики пытаются описать Вселенную помощью физических теорий и моделей. Они изучают результаты наблюдений и экспериментов и формулируют физические законы, которые организуют наше знание в виде ряда компактных утверждений.

Лидирующей лабораторией по исследованию физики частиц является Европейский центр ядерных исследований (ЦЕРН), расположенный близ Женевы (Швейцария) [1]. В ЦЕРНе действует мощный в мире ускоритель частиц, так называемый Большой андронный коллайдер (БАК) [2], который смонтирован в круговом туннеле длиной 27 км на глубине 100 метров. На БАК ускоряются пучки протонов и ионов свинца, соударяясь в 4 точках, где построены основные детекторы для фиксации различных событий касающихся частиц - ATLAS, CMS, ALICE и LHC.

ALICE (A Large Ion Collider Experiment) - эксперимент по изучению физики сильных взаимодействий при сверхвысокой плотности, где ожидается образование новых состояний ядерной материи (в частности, так называемой кварк-глюонной плазмы).

БАК был успешно запущен в ноябре 2009. Когда данные начали поступать с детекторов, распределенная инфраструктура их обработки безупречно заработала в результате многих лет предварительного постепенного развития, обновления и проверки во время предварительного запуска БАК. В результате стабильного функционирования WLCG большое количество ученых оперативно выполняют анализ данных на грид-ресурсах, а научные результаты появляются с беспрецедентной скоростью в пределах недели после получения данных с БАК.

Существующая инфраструктура WLCG постоянно совершенствуется, поглощая новые технологии, основанные на прогрессе в создании сетей, хранилищ, сервисов промежуточного программного обеспечения и интероперабельности грид и облачных вычислений.

Все LHC эксперименты разработали свои собственные вычислительные модели. Они не полагаются только на промежуточное программное обеспечение, предоставляемое в проекте WLCG, и стандартные пакеты обработки данных, а развивают специфические компоненты, которые лучше соответствуют собственным моделям вычислений.

Все LHC эксперименты приложили значительные усилия для предоставления пользователям максимально простого способа использования грид ресурсов. Эти усилия привели к значительному росту числа физиков, которые непосредственно используют грид-инфраструктуру WLCG для анализа данных.

В эксперименте ALICE разработано промежуточное программное обеспечение AliEn (AliCE Environment [3]), что обеспечивает единый интерфейс для прозрачного доступа к вычислительным ресурсам для исследователей сообщества ALICE. На сегодня вычислительная инфраструктура эксперимента ALICE представлена более 45 тысяч вычислительных ресурсов и систем хранения данных объемом более 30 Pb.

В статье представлено текущее состояние вычислительной инфраструктуры ALICE в Украине.

1 Структура программного обеспечения эксперимента Alice

Рассмотрим существующее программное обеспечение, являющееся основой IT инфраструктуры обработки данных эксперимента ALICE. Все программное обеспечение можно разделить на три группы: интерфейсные приложения, центральные компоненты AliEn и сторонние компоненты (Рис. 1).

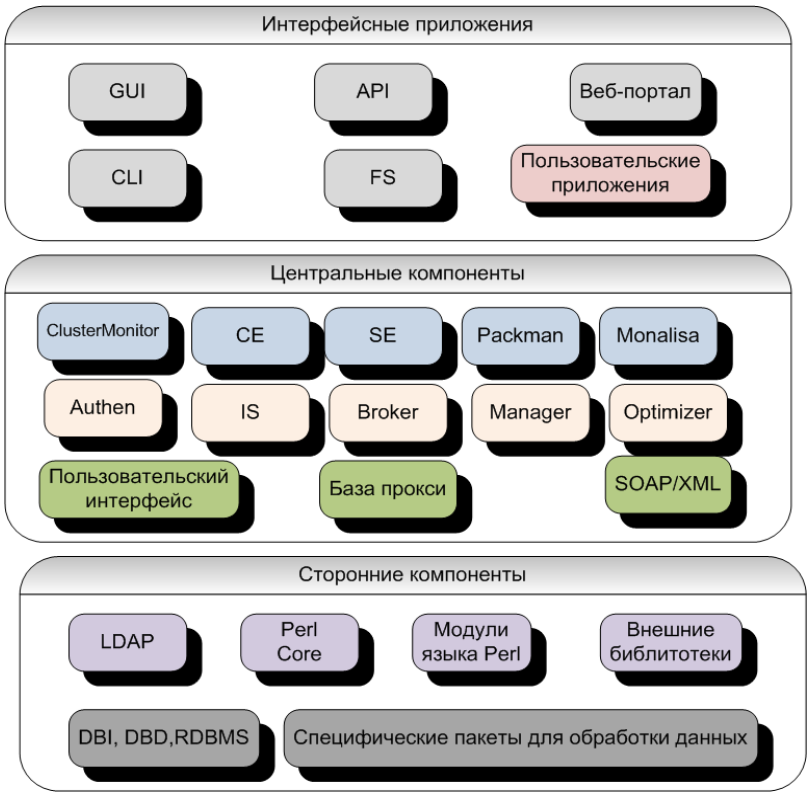

Рис 1. Структура ПО эксперимента Alice

Интерфейсными приложениями является программное обеспечение обеспечивающее контроль и легкий доступ к элементам управления задачами пользователями. Используя AliEn2 клиент, пользователи могут выполнить операции по отправке задачи на выполнение, проверку статуса задачи и управлять файлами.

Центральными компонентами инфраструктуры AliEn являются веб-сервисы, которые общаться с помощью специализированных сообщений построенных на базе структуры XML, используя для этого SOAP протокол. Все центральные сервисы можно разделить по функциональному признаку:

1. Центральные сервисы. Включают: сервис для авторизации (Authen), информационный сервис (IS), центральный брокер (Job Broker), менеджер управления задачами (Job Manager) и пакетами для вычислений (PackMan), менеджер управления передачей данных (Transfer Manager), общие базы данных (DB), сервис протоколирования (Logger) и библиотеку API.
2. Локальные сервисы грид-сайтов — сервис мониторинга состояния сайта (Cluster Monitor), локальный менеджер пакетов (PackMan), вычислительный элемент (CE), элемент хранения данных (SE), модуль системы мониторинга данных (MonaLisa).

Вся информация о конфигурации центральных сервисов определена в центральной базе данных LDAP. Там же хранится информации об основных функциональных единицах грид инфраструктуры: грид-сайтах, пользователях, грид-сервисах.

К сторонним компонентам относятся базовые средства разработки приложений такие как: библиотеки, компиляторы, языки программирования, например Perl, с использованием которого выполнена большая часть AliEn2. В список внешних пакетов следует включить программное обеспечение поддерживающее обработку данных: прикладные пакеты AliRoot, Root, Geant и другие.

2 Процесс управления задачами

Система управления задачами в проекте ALICE основана на использовании центральной очереди, которая содержит в себе задачи, предназначенные для выполнения. Модель работы брокера базируется на схеме использования «предварительной» задачи: на грид-сайте локальный клиент AliEn генерирует задачу-агента, которая запускается в очередь на вычислительном элементе (CE), и после успешного распределения локальной системой управления на

вычислительный нод (WN), задача-агент обращается с запросом в центральную очередь задач для загрузки файлов задачи для дальнейшего выполнения. Благодаря этому обеспечивается гибкое распределение задач по вычислительным кластерам и сводит к минимуму время простоя задачи в очереди.

Управление задачами осуществляется следующими компонентами: менеджером задач (Job Manager), брокером задач (Job Broker), модулем оптимизации (Job Optimizer), информационным сервисом состояния задач, планировщиком очереди (Task Queue) и сервисами аутентификации (Authen), сервисами на грид сайте CE, Cluster Monitor, а также агентами задач (JobAgents).

В состав вычислительного элемента (CE) входят: сервис Grid Gate(GG) отвечающий за приём и подготовку к выполнению грид-задач (присвоение локального пользователя, создание задания для локального менеджера ресурсов), информационная система BDII, а также он взаимодействует с локальным менеджером ресурсов (Local Resource Management System(LRMS)).

Последний необходим для запуска и контроля выполнения задания на вычислительном ноде, он может быть установлен как на том же сервере что и CE. Наиболее распространенными среди грид-сайтов являются такие системы LRMS: PBSPro, LSF, Torque, SGE совместно с Condor или Maui. В процессе обработки статус задачи определяется автоматически и отображается в информационной системе, для этих целей используются провайдеры позволяющие взаимодействовать с LRMS и динамически отображать данные в информационной системе.

Для описания задач в эксперименте Alice используется язык описания Job Description Language(JDL). Описание процесса распределения пользовательских задач рассмотрено ниже с помощью основных сервисов, участвующих запуске задачи, и модели потока данных в ПО AliEn2:

1. TaskQueue представляет собой базу данных с информацией о задачах, которые необходимо выполнить в текущий момент. Главным элементом системы управления является центральная очередь задач, функцией которой является реализация приоритетности выполнения задач.
2. Сервис управления Job Manager отвечает за прием задач, которые были запущены пользователями и за процесс отправления задач в общую очередь, проверку пользовательских лимитов для конкретного пользователя, управления статусами каждого агента и полученных файлов в результате выполнения задачи на грид сайте.
3. Сервис Job Broker отвечает за рассылку задач на грид сайты в зависимости от соответствия требований JDL файла пользовательской задачи и JDL файла грид сайта, предлагающего свои ресурсы для запуска задачи. Критериями выбора конкретного грид-сайта являются: географического положения данных, которые необходимы задаче для обработки, наличие пакетов обработки данных на грид-сайте, лимит времени для выполнения задачи или других специфических пользовательские требований. Информацию о свободных ресурсах и о количестве задач, которые можно выполнить на грид-сайте брокер получает от агента на вычислительном элементе, который периодически опрашивает брокера о том есть ли у него задачи для запуска или нет.
4. Сервис Job Optimizer используется для оптимизации JDL параметров задачи, которая находиться в статусе ожидания запуска в общей очереди. Используя информацию, о текущем состоянии грид инфраструктуры, Job Optimizer может модифицировать содержимое JDL файла задачи для сокращения времени ожидания и оптимизации соответствия задачи JDL параметрам вычислительного элемента. Модификация JDL параметров может быть связана с изменением адресов наборов данных (или реплик) необходимых для обработки данных.

Весь процесс запуска задачи можно разделить на четыре этапа: отправка задачи на выполнение, работа сервиса оптимизации, работа с брокером и агентом, сохранение результатов.

На первом этапе, вычислительный элемент (CE) грид сайта получает JDL файл с описанием задачи, и затем отсылает задачу сервису управления (Job Manager). Job Manager проверяет, использована ли квота запуска задач пользователем, а так же проводит подготовку для запуска задачи и обновляет статус запуска на “INSERTING”.

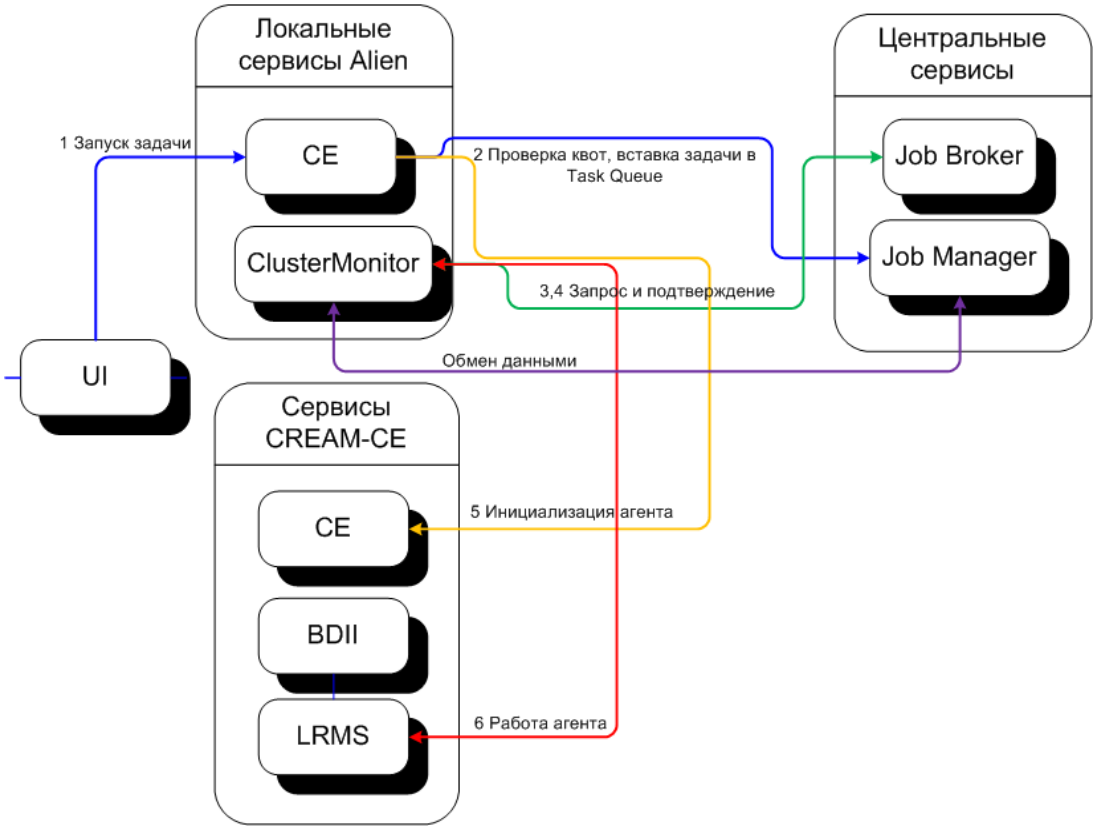

Рис 2. Схема запуска задачи

На втором этапе сервис оптимизации для задачам в статусе «INSERTING» инициирует среду исполнения в AliEn2 клиенте, подготавливает специализированный файл Job Token, помещает задачу в TaskQueue и обновляет статус задачи на «WAITING».

На третьем этапе сервис JobAgent Service опрашивает сервис ClusterMonitor на наличие доступных задач. Параллельно сервис ClusterMonitor опрашивает брокер на наличие в общей очереди задач, которые можно выполнить на данном грид-сайте.

Центральный брокер проверяет доступность очереди грид-сайта, находит подходящие задачи и отсылает результаты сервису ClusterMonitor после чего обновляет статус задач на «ASSIGNED». После получения информации, JobAgent запускает задачи на выполнение. Согласно JDL файлу задачи, JobAgent создает директории, начинает отслеживать процессы для отправки статуса сервису ClusterMonitor, который в свою очередь передаст информацию сервису управления задачами Job Manager.

С момента запуска задачи её статус может быть следующим «STARTED» - начало выполнения, «RUNNING»- процесс выполнения, «SAVING» - передача данных обработки на удаленное хранилище, или «SAVED» - этот статус задача получает в случае удачного сохранения результатов обработки. В то же время JobAgent выполняет следующие операции: отслеживание статуса задачи, отправку статуса JobAgent, запуск задачи, проверку всех процессов, запуск ValidationScript, сохранение лог файлов и другие. Статусы JobAgent могут быть такими: “JOB_STARTED”, “RUNNING_JOB”, “SAVED”.

Сервисы Job Optimizer находят задачу со статусом “SAVED” и “SAVED_WARN” в TaskQueue, затем создают результирующую информацию для пользовательского AliEn клиента, регистрируют файлы в каталоге и обновляют статус задачи на “DONE”.

3 Алиса в Украине

На данный момент в Украине эксперимент Alice представлен двумя грид сайтами – UA-BITP, UA-ISMA в Киеве и в Харькове, которые в совокупности обработали более 100000 задач за период полгода и группой обработки данных эксперимента Алисы.

В состав вычислительных ресурсов ИТФ входит отдельный кластер обслуживающий эксперимент Alice. Это обусловлено тем что на локальном кластере установлены ЦПУ не удовлетворяющие требованиям эксперимента. В состав этого кластера входят 8 вычислительных узлов суммарно имеющих 64 ядра Xeon E5607 и дисковый массив объемом 54Тб.

На этих узлах открыта очередь исключительно для эксперимента ALICE. UA-BITP работает с 2006 года и фактически прошел через все версии Alien, одними с первых перешел на работу с торрентами для распределения пакетов обработки данных для запущенных задач что отразилось эффективно на использовании памяти (проблема использования NFS shared disk для монтирования пакетов на WNs).

UA-ISMA подключился в 2012 году, одним из первых использовав новую версию операционной системы SL6 для работы с пакетами промежуточного программного обеспечения EMI, aLien.

Для интеграции кластера ИСМА НАН Украины в европейскую грид-инфраструктуру EGI и продолжения участия в международном грид-проекте ALICE в настоящее время проведено инсталляцию, базовые настройки и проводится эксплуатация middleware EMI2. Инсталляция EMI2 была проведена без отделения расчетных мощностей кластера в отдельный пул таким образом, чтобы сохранить поддержку общих очередей кластера и одновременную работу двух грид-систем (ARC и EMI2+gLite) с использованием одного общего менеджера задач.

4 Тенденции развития инфраструктуры в эксперименте Alice

В текущем году запланирован переход на новую версию v2-20, отличающейся от текущей управлением и механизмом реализации очереди, реализации каталога хранения данных, коммуникацией между сервисами. Изменения также коснуться баз данных центральных сервисов, их объединением.

Изменения в очереди коснуться параметров максимального времени ожидания, новых статусов ошибок связанных с новыми параметрами, управления перезапуском задач, и задачами в статусе KILLED. В режиме тестирования получены результаты подтверждающие эффективность изменений, уменьшением операционного времени управления задачей в несколько десятков раз.

Тенденцией на будущий год будет реализация облачных вычислений. Этому способствует удобная компьютерная модель ALICE, благодаря использованию унифицированного доступа к данным, по протоколу доступа xrootd, единой очереди Task Queue, отсутствию разницы в структуре между сайтами уровня T1/T2, использованию виртуализированих WNs, возможности использования Cloud API для старта агентов Job Agent. На первом этапе в ЦЕРН будут внедрены облачные вычисления на базе вычислительной фермы высоко уровневого триггера (High Level Trigger), базируясь на решении CernVMFS [3].

Также ключевыми работами на этот год могут быть увеличение эффективности использования ресурсов (CPU, памяти, I/O, оптимизации структуры данных), оптимизации использования алгоритмов, процедур калибрации, пользовательского анализа, качества данных, решения проблемы утечки памяти в коде (memory leaks), анализ доступа к данных (dynamical data placement policies), улучшение качества работы пакетов AliROOT, aLien.

Для украинских сайтов поставлена проблема увеличения эффективности использования CPU, используя технологии виртуализации и облачных вычислений. В текущем году планируется переход на IPv6 маршрутизацию для основных сервисов VOBOX и системы хранения данных SE, и наращивание ресурсов для более эффективных вычислений.

Литература

1. CERN - the European Organization for Nuclear Research; http://public.web.cern.ch/public/.
2. The Large Hadron Collider at CERN; http://lhc.web.cern.ch/lhc/; http://public.web.cern. ch / public / en / LHC / LHC-en.html
3. P. Saiz et al., AliEn-ALICE environment on the GRID, Nucl. Instrum. Meth. A502 (2003) 437; http://alien2.cern.ch/
4. Technical Design Report of the Computing, Printed at CERN / The ALICE Collaboration June 2005 - ISBN 92-9083-247-9.
5. Grid Architecture for ALICE Experiment / [Jianlin Zhu, Pablo Saiz, etc.]
6. Distributing LHC application software and conditions databases using the CernVM file system / J Blomer — 2011.