ОБЪЕДИНЕННЫЙ ИНСТИТУТ ЯДЕРНЫХ ИССЛЕДОВАНИЙ
ЛАБОРАТОРИЯ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

На правах рукописи

Петросян Артем Шмавонович

Методика и программная инфраструктура глобально распределенной обработки данных эксперимента COMPASS

Специальность 05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Автореферат диссертации на соискание ученой степени кандидата технических наук

Дубна – 2021

Работа выполнена в Лаборатории информационных технологий Объединенного института ядерных исследований

Научный руководитель: **Кореньков Владимир Васильевич**
доктор технических наук

Официальные оппоненты:
доктор физико-математических наук — Богданов Александр Владимирович

кандидат физико-математических наук — Крюков Александр Павлович

С электронной версией диссертации можно ознакомиться на официальном сайте Объединенного института ядерных исследований в информационно-телекоммуникационной сети «Интернет» по адресу: https://dissertations.jinr.ru.
С печатной версией диссертации можно ознакомиться в Научно-технической библиотеке ОИЯИ.

Ученый секретарь
диссертационного совета ОИЯИ.05.01.2019.П
д-р физ.-мат. наук — Е.В. Земляная

Общая характеристика диссертации

Работа основана на результатах исследований, выполненных в 2010-2020гг. в Лаборатории информационных технологий (ЛИТ ОИЯИ, Дубна) и Европейской организации по ядерным исследованиям (ЦЕРН, Женева). В диссертации представлены результаты разработки программного комплекса управления процессом обработки данных, получаемых в ходе физических исследований на ускорительном комплексе ЦЕРН.

Актуальность темы

Жизненный цикл современных физических экспериментов нередко длится десятки лет и включает несколько этапов модернизации различных компонентов в связи с их износом или развитием для проведения новых исследований. При этом, если процессы развития элементов физической установки находятся под контролем участников эксперимента, то развитие вычислительной инфраструктуры, версий системного программного обеспечения, использующегося для записи, хранения и обработки собранных экспериментом данных, находится вне зоны контроля участников эксперимента. Любой эксперимент сталкивается с задачей обеспечения сбора, хранения и обработки данных в постоянно изменяющихся условиях. Крайне важно, чтобы система управления обработкой данных эксперимента была спроектирована с учетом неизбежных изменений в качественном и количественном составе компонентов IT-инфраструктуры. Необходимо учитывать и развитие самой физической установки, приводящей к изменениям в процессе обработки данных. В противном случае каждый этап обновления или замены любого из компонентов установки и IT-инфраструктуры будет подразумевать необходимость проведения глубокой модернизации или полного перепроектирования и замены системы управления процессом обработки.

Опыт построения и эксплуатации распределенных вычислительных сред экспериментами на Большом адронном коллайдере (БАК) в ЦЕРН продемонстрировал, что появились решения, способные управлять обработкой данных не только в рамках одного вычислительного центра, но и за его пределами. Появилась возможность организовывать распределен-

ные системы с использованием ресурсов институтов-участников коллаборации. Стали доступны ресурсы разных типов: коммерческие облачные инфраструктуры и высокопроизводительные вычислительные системы. Созданы продвинутые системы управления распределенными данными. Однако, несмотря на то, что сервисы хранения, управления и обработки данных уже существуют, каждый из экспериментов реализует свою систему верхнего уровня, способную обеспечить управление обработкой данных, используя эти сервисы.

Эксперимент COMPASS, начавший работу за несколько лет до введения в эксплуатацию БАК и реализации проекта Worldwide LHC Computing Grid (WLCG), в течение 15 лет хранил и обрабатывал данные только в ЦЕРН, а к 2015 году столкнулся с рядом вызовов, потребовавших проведения масштабной модернизации системы управления обработкой данных.

Цели и задачи исследования

Целью работы является организация глобально распределенной обработки данных физического эксперимента COMPASS.

В рамках диссертационной работы решались следующие задачи:

1. Анализ имеющейся системы обработки данных эксперимента COMPASS и актуальных программных средств, предназначенных для управления обработкой данных физического эксперимента.

2. Разработка методики организации глобально распределенной обработки данных на базе компонентов программной инфраструктуры экспериментов на БАК.

3. Развертывание среды глобально распределенной обработки данных эксперимента COMPASS.

4. Проектирование и разработка программного инструментария, способного обеспечить глобально распределенную обработку данных физического эксперимента COMPASS.

5. Интеграция высокопроизводительных систем для осуществления на них обработки данных эксперимента.

Методы исследования

Диссертационная работа выполнена с применением методов системного анализа, проектирования информационных систем, организации взаимодействия программ, программных систем и глобально распределенной обработки данных, программной инженерии и анализа программного обеспечения.

Научная новизна

1. Впервые осуществлен успешный перенос обработки данных работающего в течение 15 лет эксперимента из среды одного вычислительного центра в глобально распределенную среду.

2. Предложена ориентированная на особенности эксперимента COMPASS методика, позволяющая использовать различные вычислительные ресурсы для выполнения десятков тысяч задач на основе широкого использования групповых операций и параллелизма.

3. Создано реализующее предложенную методику программное обеспечение для эффективной массовой обработки данных эксперимента COMPASS на объединенных в единую глобально распределенную инфраструктуру вычислительных ресурсов различного типа.

Научно-практическая значимость

1. С помощью созданного программных средств обработаны наборы данных разных годов: 2004, 2009-2012, 2015-2018. С 2017 года программный инструментарий является основной платформой для обработки данных эксперимента COMPASS. На текущий момент при помощи созданного программного обеспечения обработано более 150 миллиардов физических событий, оформленных в виде более 13 миллионов задач.

2. Примененные при разработке программного обеспечения средства организации глобально распределённых вычислений позволили использовать в качестве вычислительных ресурсов не только грид-сайты, но и высокопроизводительные вычислительные системы, такие как

Blue Waters (в 2018) и Frontera (в 2019-2020). Подключение высокопроизводительных систем позволяло в два раза увеличить объем доступных эксперименту вычислительных мощностей. Кроме высокопроизводительных ресурсов, в различные периоды использовались вычислительные мощности институтов-участников эксперимента: ОИЯИ и INFN Триест. Внедрение сервисов массовой передачи данных позволило значительно, в некоторых случаях в десятки раз, сократить время, требующееся для записи результатов на носители ленточного хранилища.

Защищаемые положения

1. Методика организации процесса обработки данных физического эксперимента на примере эксперимента COMPASS, позволяющая использовать различные вычислительные ресурсы для управления десятками тысяч одновременно выполняемых задач.

2. Программное обеспечение управления процессом обработки данных эксперимента COMPASS, реализованное на основе предложенной методики, позволяющее обеспечить эффективную обработку данных эксперимента на объединенных в единую глобально распределенную инфраструктуру вычислительных ресурсах различного типа.

3. Внедренная в промышленную эксплуатацию система управления обработкой данных эксперимента COMPASS, подтвердившая за три года эксплуатации свою надежность, способность выдерживать пиковые нагрузки и являющаяся в настоящее время официальной системой обработки данных эксперимента.

Апробация работы

Результаты исследований, положенных в основу диссертации, докладывались автором на научных семинарах Объединенного института ядерных исследований (ОИЯИ), рабочих совещаниях коллаборации COMPASS, проходивших в ЦЕРН и ОИЯИ, на международных рабочих совещаниях, конференциях и симпозиумах, наиболее важные из которых:

- Международные конференции “Распределенные вычисления и Grid-технологии в науке и образовании” (GRID), 4-9 июля 2016, Дубна, ЛИТ ОИЯИ (GRID-2016); 10-14 сентября 2018, Дубна, ЛИТ ОИЯИ (GRID-2018).
- International Symposium on Nuclear Electronics & Computing (NEC), 25-29 сентября 2017, Будва, Черногория (NEC-2017); 30 сентября-4 октября 2019, Будва, Черногория (NEC-2019).
- PanDA Workshop, 22 апреля 2016, ЦЕРН, Швейцария.
- COMPASS Collaboration Meeting, 17 ноября 2017, ЦЕРН, Швейцария.
- Rucio Workshop, 1-2 марта 2018, ЦЕРН, Швейцария.
- 23d International Conference on Computing in High Energy and Nuclear Physics (CHEP-2018), 9-13 июля 2018, София, Болгария.

Публикации и личный вклад

Основные результаты диссертации представлялись автором на научных семинарах ЛИТ ОИЯИ и на международных научных конференциях. По теме диссертации подготовлено 20 научных работ, 18 из которых опубликованы в рецензируемых изданиях, соответствующих требованиям к публикациям Положения о присуждении ученых степеней в ОИЯИ (пр. ОИЯИ от 30.04.2019 № 320).

Представленные в диссертации результаты по созданию программного обеспечения по управлению обработкой данных эксперимента COMPASS выполнена автором. Автор являлся ответственным за проведения исследований, проектирование и реализацию программных модулей, подготовку публикаций. Все представленные в диссертации результаты получены лично автором, либо в соавторстве при определяющем вкладе соискателя.

Соответствие диссертации паспорту специальности

В диссертационной работе присутствуют результаты в трех областях, соответствующих пунктам 3, 8 и 9 паспорта специальности: взаимодействия программ и программных систем, методы создания программ и

программных систем для параллельной и распределенной обработки данных, алгоритмы и программная инфраструктура для организации глобально распределенной обработки данных.

Достоверность результатов

Подтверждается практическим использованием разработанного программного обеспечения в качестве платформы обработки данных эксперимента COMPASS в течении более чем трех лет.

Объем и структура диссертации

Диссертационная работа состоит из введения, трех глав, заключения, перечня наименований и ссокращений, списка цитируемой литературы (82 пункта) и приложения. Работа содержит 100 страниц и включает в себя 22 рисунка и 2 таблицы.

Содержание работы

Во **Введении** раскрывается актуальность темы исследования, описываются основные цели работы, научная новизна, практическая значимость работы, личный вклад автора, приводятся положения, выносимые на защиту и результаты апробации.

В **первой главе** приводится описание эксперимента COMPASS, разбираются процессы сбора, хранения и обработки данных, приводится обзор современных программных комплексов по обработке данных физических экспериментов, описываются средства создания среды распределенной обработки данных. Формулируется постановка решаемой в диссертационной работе задачи.

Эксперимент COMPASS (Common Muon and Proton Apparatus for Structure and Spectroscopy) одобрен в 1997 году и начал набирать данные в 2002 году. В коллаборации участвуют 24 института из 13 стран, около 250 исследователей.

COMPASS производит в среднем 1.5Пб данных в год. Вместе с обработанными данными в настоящее время суммарный объем данных эксперимента составляет около 20Пб.

Упрощенная схема обработки данных по состоянию на 2015-й год представлена на рис. 1. Данная реализация характеризуется рядом особенностей:

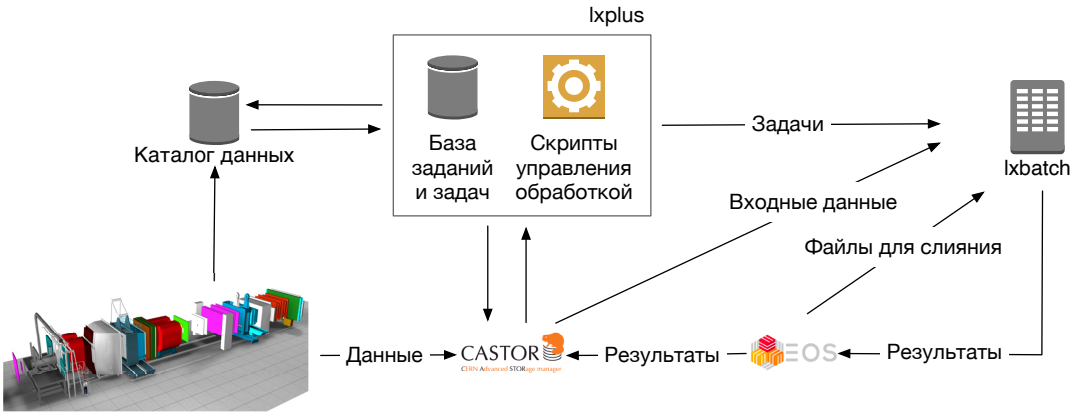

Рис. 1. Схема обработки данных

- реализован только процесс реконструкции физических событий;
- обработка ведется только на вычислительном кластере в ЦЕРН, при этом поддерживается достаточное большое число одновременно выполняемых не параллельных задач: до 9 000;
- прикладное ПО размещено в распределенной сетевой файловой системе AFS;
- используются ленточное хранилище CASTOR для долговременного хранения данных и дисковое хранилище EOS для хранения данных во время обработки;
- после завершения обработки файлов выполняется объединение результатов выходных файлов в файлы большего размера для оптимального хранения на ленточной системе хранения данных;
- система работает в полуавтоматическом режиме: описание заданий организовано в виде структурированных файлов, запуск обработки

и проверка результатов производятся оператором обработки.

Из представленного описания видно, что архитектура системы ориентирована на обработку данных в рамках одного вычислительного центра. Реализованы не все процессы, хорошо поддающиеся автоматизации, как, например, процесс моделирования физических событий. При этом она обеспечивает обработку достаточно большого количества одновременно выполняемых задач.

Ставилась задача проведения модернизации системы управления обработкой данных эксперимента COMPASS в связи со следующими запланированными изменениями в составе IT-инфраструктуры ЦЕРН:

- смена системы управления вычислительными ресурсами: IBM LSF на HTCondor к концу 2018 года;
- планируемый вывод из эксплуатации ленточного хранилища CASTOR, который будет заменен системой CTA (CERN Tape Archive);
- планируемый вывод из эксплуатации файловой системы AFS.

Иначе говоря, требовалось переработать весь интерфейс взаимодействия с вычислительной инфраструктурой. При проведении масштабной модернизации комплекса управления обработкой данных эксперимента COMPASS было необходимо рассмотреть возможность реализации обновленного комплекса с учетом дополнительных требований, удовлетворение которых позволит избежать необходимости повторения процедуры модернизации в будущем во время изменений в составе компонентов IT-инфраструктуры эксперимента. Дополнительные требования можно сформулировать следующим образом:

- возможность смены типа использующегося вычислительного ресурса без необходимости изменения архитектуры системы для обеспечения плавного перехода с одного типа вычислительного ресурса на другой в будущем;
- расширение набора автоматизированных процессов обработки данных;

- возможность организации обработки на вычислительных ресурсах участников коллаборации;
- возможность обеспечить обработку данных на высокопроизводительных вычислительных ресурсах.

Ключевыми характеристиками современных комплексов для обработки данных физических экспериментов являются:

- обеспечение обработки данных в географически распределенной среде на всех доступных ресурсах любого типа;
- обеспечение управления распределенными данными;
- предоставление единого интерфейса для любого типа задач;
- максимальная автоматизация всех процессов обработки данных.

На момент проведения исследования наиболее полно описанные выше характеристики были реализованы в комплексах управления обработкой данных экспериментов на Большом адронном коллайдере в ЦЕРН: в условиях дефицита вычислительных мощностей и возможностей обеспечить хранение набираемых данных экспериментами были созданы высокоавтоматизированные системы, эффективно управляющие хранением и обработкой данных в глобально распределенной гетерогенной вычислительной среде.

Многие компоненты программных комплексов, разработанных экспериментами на БАК, реализованы в виде независимых программных продуктов с открытым кодом, доступных для использования сторонними пользователями. При этом, если компоненты можно подобрать, исходя из ожидаемых объемов данных, имеющихся вычислительных ресурсов и ресурсов хранения, то систему, управляющую этими компонентами и обеспечивающую их согласованное взаимодействие в рамках процесса обработки данных, каждый эксперимент вынужден разрабатывать для реализации собственной логики обработки задач.

Для оценки возможности использования существующих систем и сервисов был проведен анализ программных комплексов экспериментов на

Табл. 1. Функциональные особенности программных комплексов экспериментов на БАК

	ATLAS	CMS	ALICE	LHCb
Работает с различными системами хранения	да	да	нет	да
Работает с различными типами систем управления локальными вычислительными ресурсами	да	да	нет	да
Поддерживает различные протоколы передачи данных	да	нет	нет	да
Интеграция со сторонними поставщиками вычислительных ресурсов (“облачные” ресурсы)	да	нет	нет	да
Интеграция с высокопроизводительными вычислительными системами (суперкомьютерами)	да	нет	нет	нет
Компоненты доступны в виде независимых пакетов, можно подобрать их набор под задачу	да	нет	нет	нет

БАК, построенных в ЦЕРН и делящих с экспериментом COMPASS многие элементы как физической, так и IT-инфраструктуры. Ключевыми характеристиками при оценке являлись подтвержденное использование программного компонента за пределами эксперимента, для которого он был разработан, универсальность, надежность, масштабируемость, наличие поддержки и перспективы развития.

Результаты сравнения комплексов обработки данных экспериментов на БАК представлены в табл. 1.

Программное обеспечение, используемое экспериментами на БАК для связи элементов географически распределенной вычислительной инфраструктуры в единое вычислительное пространство, можно отнести к связующему программному обеспечению. Связующее программное обеспечение разрабатывалось экспериментами на БАК для использования максимального доступного объема географически распределенных вычислительных ресурсов различного типа. Однако возможность работы с различными типами вычислительных ресурсов может быть использована и в рамках одного вычислительного центра в качестве инструмента, облегчающего миграцию с одного вычислительного ресурса на другой. Этот же подход можно применить и к системам управления распреде-

ленными данными: вместо того, чтобы разрабатывать подключение к системе хранения другого типа, можно использовать систему управления данными, которая поддерживает оба типа систем хранения. Такой подход позволяет избежать необходимости перестраивать программный комплекс при смене какого-то компонента на другой, но и накладывает обязательства по установке и поддержке связующего программного обеспечения.

Кроме того, использование связующего программного обеспечения позволяет расширить доступные вычислительные мощности или же интегрировать добровольные вычислительные ресурсы, доступные не на постоянной основе, а в рамках гранта, проекта и т.п. Примерами таких ресурсов являются облачные вычислительные инфраструктуры, в том числе коммерческие, и суперкомпьютеры.

Для организации распределенных вычислений, подразумевающих использование сторонних вычислительных ресурсов и элементов хранения, требуется использование следующих систем и сервисов:

- система аутентификации/авторизации;
- информационная система для описания ресурсов;
- система управления нагрузками, распределяющая задачи по вычислительным ресурсам и контролирующая их выполнение;
- система управления данными в распределенной среде.

Для упрощения и синхронизации развертывания прикладного программного обеспечения на вычислительных ресурсах может использоваться сервис кэширования программного обеспечения.

Необходимость использования каждой из этих систем определяется исходя из модели сбора, хранения и обработки данных каждого эксперимента. Например, эксперименты на БАК используют частично пересекающийся набор сервисов, подобранный для решения задач каждого из экспериментов, и разработанную каждым экспериментом систему управления процессами обработки данных.

Таким образом, использование описанных выше систем и сервисов позволяет организовать среду вычислений эксперимента, но для каждого

конкретного эксперимента требуется разработать еще и систему верхнего уровня, которая и будет отвечать за управление процессом обработки данных.

Во **второй главе** формулируется методика организации среды глобально распределенной обработки данных физического эксперимента, подбираются компоненты среды распределенных вычислений эксперимента COMPASS, описывается архитектура системы управления процессами обработки данных эксперимента.

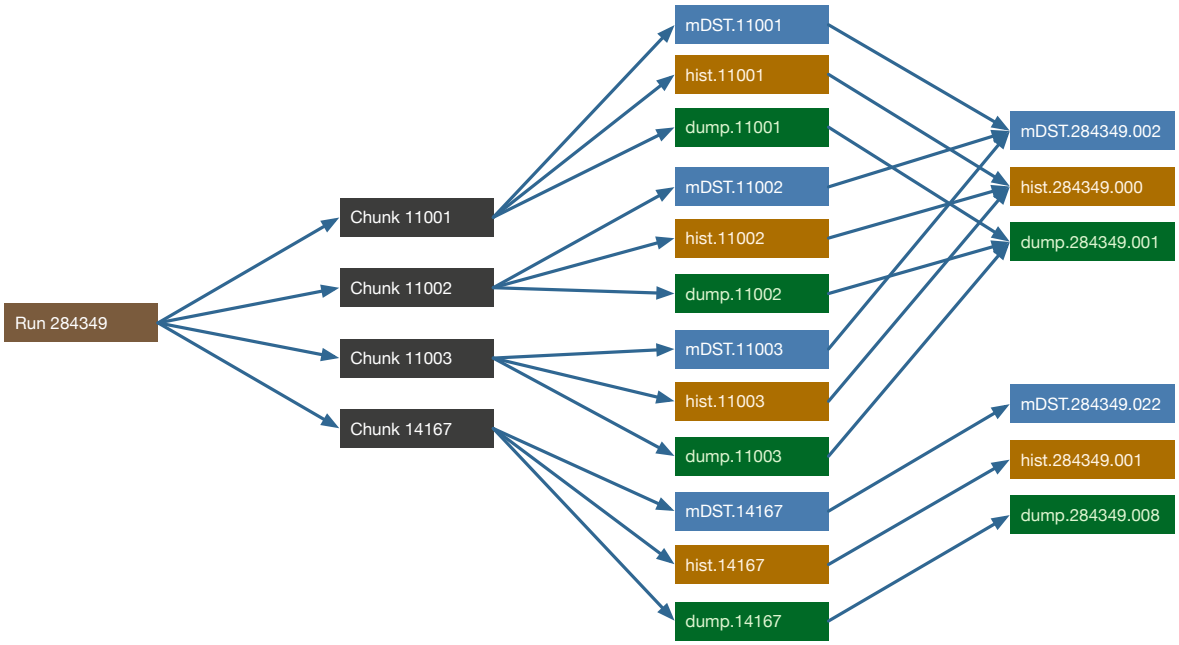

Рис. 2. Процесс обработки данных в ходе реконструкции физических событий

Процесс обработки физических данных обычно состоит из нескольких этапов: описания задания, подготовки данных, обработки, выгрузки результатов, удаления временных файлов. Каждый из этих этапов разбивается на более мелкие шаги. Например, цепочка обработки задач реконструкции, представленная на рис. 2, состоит из нескольких шагов, выполняемых различным прикладным программным обеспечением, процедур слияния и проверок по результатам выполнения каждого шага. Данные для обработки объединяются в задания, обычно состоящие

из десятков тысяч файлов, для каждого из которых обычно создается отдельная задача. Очевидно, что наборы данных таких масштабов невозможно обработать без автоматизации. Так как каждый эксперимент уникален и процессы обработки данных одного эксперимента отличаются от других, каждый эксперимент разрабатывает свою реализацию системы управления обработкой данных.

Проектируемая система управления обработкой эксперимента COMPASS должна производить подготовку данных, обработку ошибок и проверку результатов для не менее чем 9 000 одновременно выполняемых задач. Должна обеспечиваться обработка параллельно выполняющихся заданий разного типа: моделирования, реконструкции физических событий, отбора событий по характеристикам.

Изучение опыта организации распределенной вычислительной среды WLCG и развития методов управления обработкой данных экспериментов на БАК позволяет сформулировать методику организации обработки данных физического эксперимента в глобально распределенной среде:

1. Развернуть инфраструктуру распределенных вычислений:
    - создать виртуальную организацию на собственном или стороннем сервисе аутентификации и авторизации;
    - зарегистрировать пользователей виртуальной организации;
    - установить вычислительные элементы (CE) в вычислительных центрах институтов-участников эксперимента;
    - подключить виртуальную организацию к вычислительным элементам;
    - установить элементы хранения (SE) в вычислительных центрах институтов-участников эксперимента;
    - подключить виртуальную организацию к элементам хранения эксперимента;
    - установить или зарегистрировать сервис кэширования прикладного программного обеспечения;
    - разместить прикладное ПО эксперимента на серверах сервиса кэширования;

- подключить сервис кэширования прикладного ПО к вычислительным центрам институтов-участников эксперимента.

2. Исходя из модели обработки данных, оценить целесообразность использования системы управления распределенными данными:

   - при хранении всех набираемых данных в одном хранилище использование системы управления распределенными данными возможно для упрощения работы с хранилищами за счет предоставления более удобного пользовательского интерфейса и соглашений о наименовании;
   - при хранении всех набираемых данных в рамках одного хранилища но с использованием нескольких типов хранилищ, например дискового и ленточного, следует рассмотреть использование системы управления распределенными данными для управления данными;
   - если планируется доставка наборов данных для обработки на удаленных вычислительных ресурсах, то предпочительно организовать доставку с использованием системы управления распределенными данными;
   - при хранении не пересекающихся наборов данных на нескольких удаленных друг от друга хранилищах использование системы управления распределенными данными обязательно.

3. Основываясь на ожидаемых потоках данных и доступных вычислительных ресурсах произвести выбор системы управления нагрузкой:

   - если объемы обрабатываемых задач не превышают возможности одного вычислительного центра и не предполагается использование облаков и высокопроизводительных систем, то возможно использование любой из упомянутых в таблице 1 систем управления нагрузкой;
   - при использовании сторонних ресурсов различного типа, в том числе научных или коммерческих облаков, возможно примене-

ние в качестве системы управления нагрузкой систем DIRAC и PanDA;

- если планируется использование крупных высокопроизводительных систем, то нагрузку лучше организовать при помощи системы PanDA.

4. Разработать систему управления обработкой данных в соответствии с моделью обработки данных конкретного эксперимента:

   - спроектировать и создать базу данных заданий и задач для учета их состояний во время выполнения процессов обработки данных;
   - разработать веб-интерфейс управления обработкой данных;
   - разработать модули сопровождения процессов обработки:
     - для работы с каталогом данных;
     - для взаимодействия с системой управления нагрузкой;
     - для управления данными во время обработки;
     - для взаимодействия с сервисом передачи файлов;
     - принятия решений.

Сервисы, обеспечивающие аутентификацию и авторизацию, поддерживаются в рамках инфраструктуры WLCG (World Wide LHC Computing Grid). При этом эксперименты на БАК, как это уже обсуждалось в главе 1, развивают свои реализации систем описания ресурсов, управления обработкой, распределения задач и управления распределенными данными.

Для эксперимента COMPASS среда распределенной обработки формируется на базе компонентов, поддерживаемых в рамках WLCG: виртуальная организация (VO), вычислительные элементы (CE), элементы хранения (SE), кэшируемая файловая система для распространения ПО эксперимента (CVMFS). Сервисы доставки задач на вычислительные ресурсы из набора систем и сервисов, реализованных коллаборацией ATLAS как удовлетворяющие наиболее полно требованиям к системе

управления обработкой данных, озвученным в первой главе. Список сервисов включает в себя систему управления нагрузкой PanDA. Необходимо адаптировать для COMPASS приложения Pilot и PanDA Monitoring. Так как эксперимент хранит данные только в ЦЕРН, можно отказаться от системы управления распределенными данными. Этот же аспект позволяет отказаться от сложной информационной системы. Необходимо разработать систему управления процессом обработки данных, состоящую из следующих компонент: базы данных, веб-интерфейса описания заданий и задач, сервисов сопровождения процесса обработки данных, сервисов управления данными в процессе их обработки.

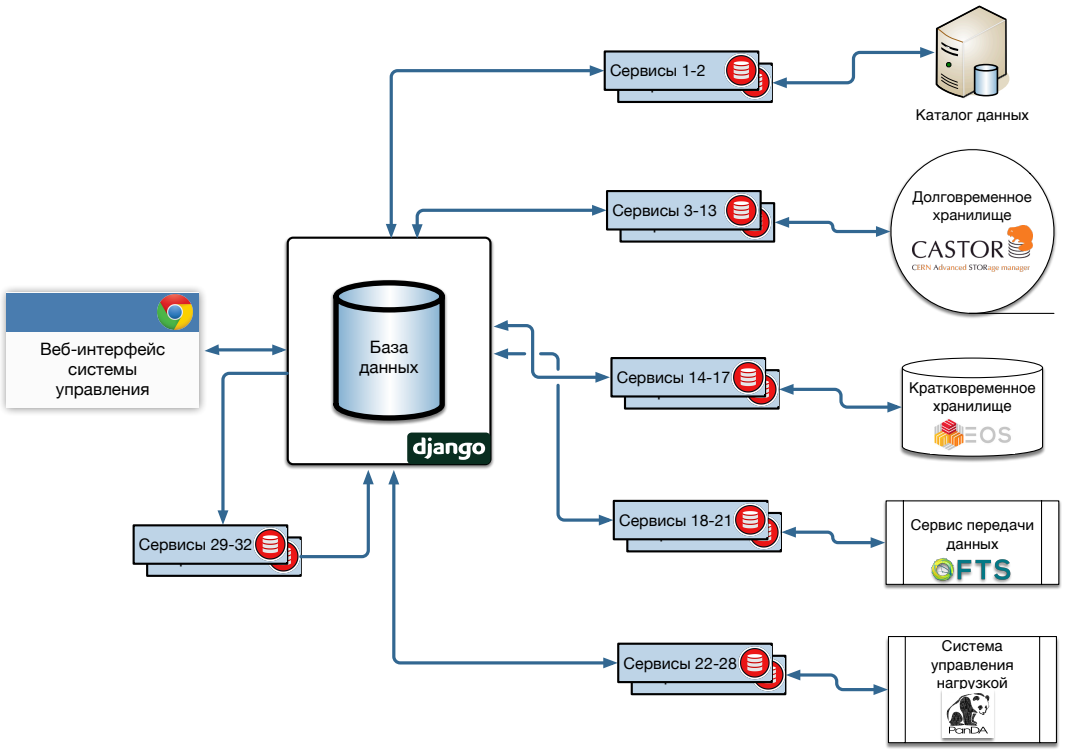

Рис. 3. Архитектура системы управления обработкой данных

Архитектура системы управления обработкой данных представлена на рис. 3. Учитывая характер вычислений: обеспечение выполнения множества независимых задач, большое количество шагов обработки, четкая последовательность принятий решений, различные сценарии обработки, который должна обеспечивать проектируемая система управления обработкой данных, автором предложено использовать микросервисную ар-

хитектуру. Микросервисная архитектура позволяет обеспечить гибкость и масштабируемость при максимальной независимости компонент друг от друга: экземпляры сервисов можно размещать на различных физических серверах. Также такая архитектура позволяет не допускать чрезмерного разрастания исходного кода и, таким образом, проста в поддержке. Полный список и описание микросервисов можно найти в тексте диссертации.

Третья глава посвящена результатам работы, особенностям организации обработки данных на вычислительных кластерах и высокопроизводительных вычислительных системах, текущему состоянию, сравнению с предыдущей реализацией системы управления обработкой данных и перспективам развития программного комплекса.

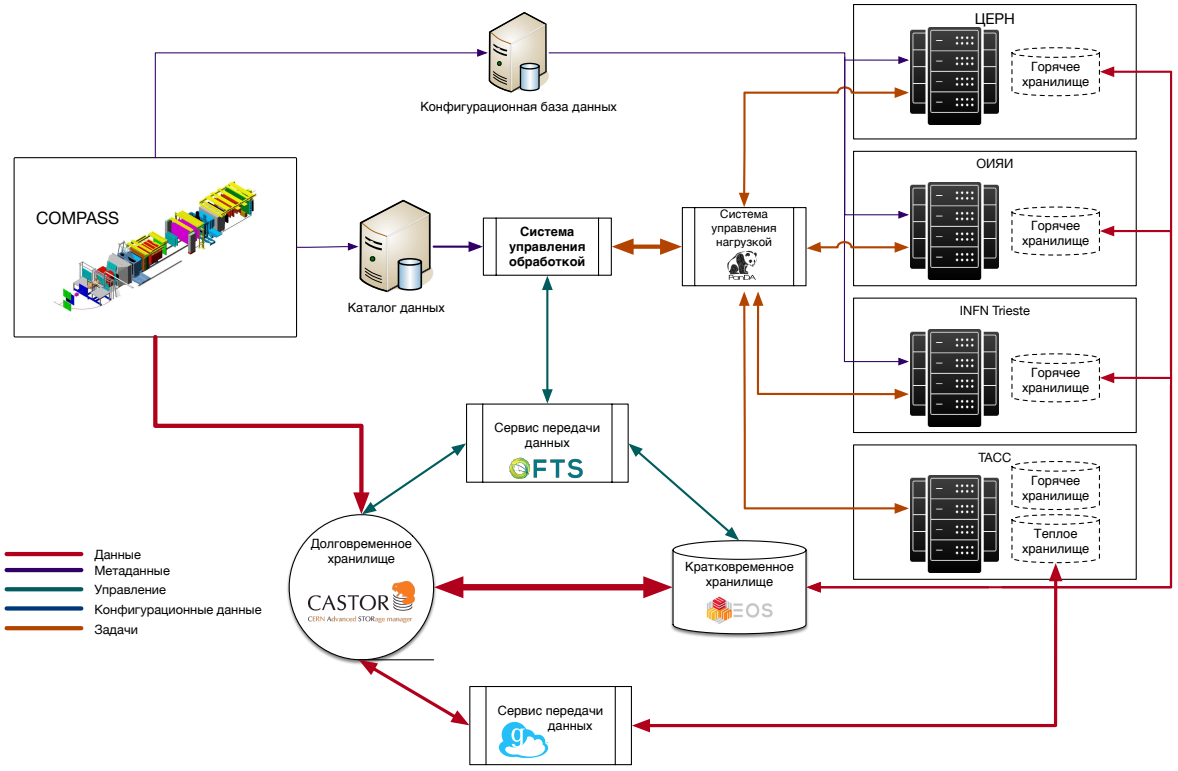

Рис. 4. Обработка данных в 2020 году

Разработаны более 30 управляющих модулей, обеспечивающих сопровождение процессов обработки данных: генерацию задач из параметров задания, подготовку файлов на ленточном хранилище, отсылку задач, проверку статуса завершения задачи, отправку задач агрегации, провер-

ку результатов агрегации, отправку файлов на ленточное хранилище, архивацию журналов задач и отправку архивов на ленточное хранилище, удаление логов задач с временного хранилища. Каждый из сервисов выполняет только одно действие с минимально возможным набором входных данных.

Использование возможностей систем управления нагрузками по интеграции вычислительных ресурсов различного типа позволило увеличить количество доступных для эксперимента вычислительных ресурсов, а также обеспечить обработку на высокопроизводительных вычислительных ресурсах. На рис. 4 представлена схема обработки данных эксперимента COMPASS с использованием как вычислительных кластеров, так и высокопроизводительных систем: суперкомпьютера Frontera Техасского суперкомпьютерного центра.

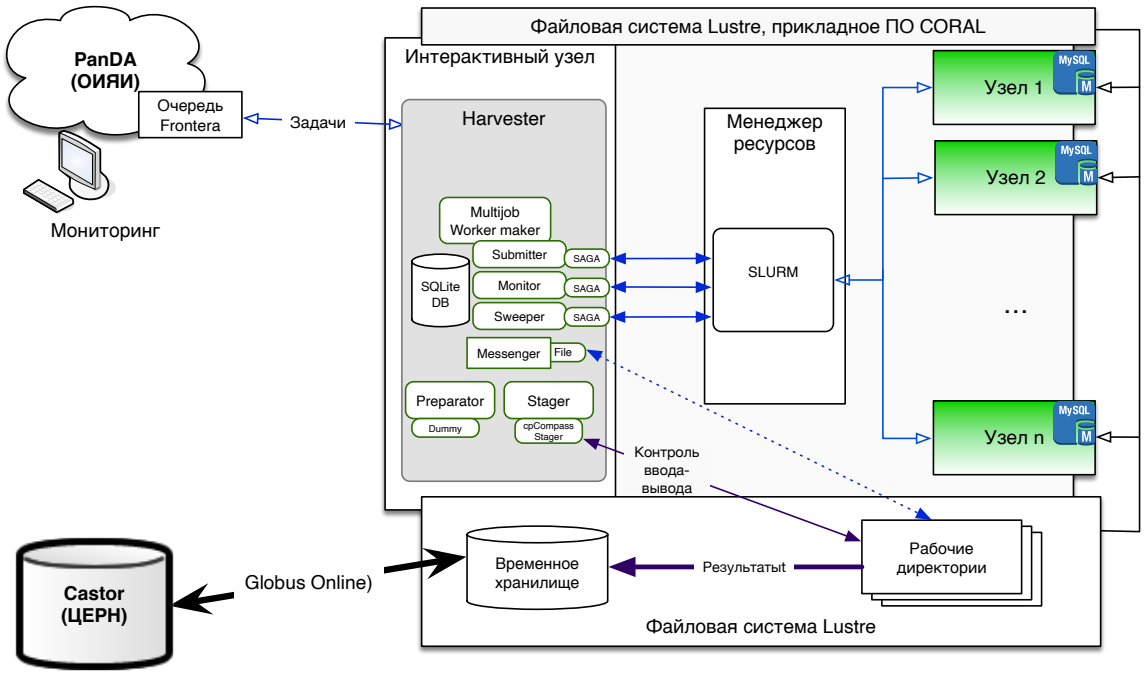

Рис. 5. Организация обработки данных на суперкомпьютере Frontera

Следует подробнее остановиться на организации обработки данных

на высокопроизводительных системах. Использование подобных систем предъявляет повышенные требования ко всем компонентам комплекса управления обработкой данных, как к программным, так и к физическим. Связано это с характерным для вычислений на высокопроизводительных системах скачкообразным профилем изменения нагрузок: в случае работы с обычной фермой задачи доставляются на вычислительный узел и запускаются по одной, в случае же работы с суперкомпьютером несколько тысяч задач группируются в одну отсылку и, после ожидания в очереди, начинают выполняться практически одновременно. Кроме этого, задачи на большее количество узлов на таких системах получают более высокий приоритет в очереди, поэтому необходимо поддерживать как можно большее количество как можно более крупных задач в очереди. Была разработана реализация пилотного приложения для выполнения на высокопроизводительных системах. Особенностью работы задачи реконструкции является поддержание постоянного соединения с конфигурационной базой данных. В случае выполнения задачи на грид-кластере каждый рабочий узел осуществляет сетевое подключение к удаленной базе данных. В случае же выполнения задачи на рабочем узле высокопроизводительной системы подключение к удаленной базе данных невозможно, потому что обычно на подобных системах сетевое подключение на рабочих узлах отсутствует. Таким образом, для обеспечения корректной работы прикладного ПО необходимо перед запуском задач реконструкции на каждом узле запускать СУБД. Схема организации обработки данных на высокопроизводительном ресурсе Frontera представлена на рис. 5. При обработке используются MPI-задачи на 50 вычислительных узлах, каждая из которых состоит из 2 800 индивидуальных задач. На суперкомпьютере выполняется вся цепочка процесса обработки задач реконструкции данных, представленная ранее на рис. 2, включая агрегацию результатов. Управление заданиями и задачами на данном ресурсе полностью интегрировано в систему управления обработкой данных эксперимента COMPASS и с точки зрения пользователя производится таким же образом, как и управление обработкой данных на обычных грид-кластерах.

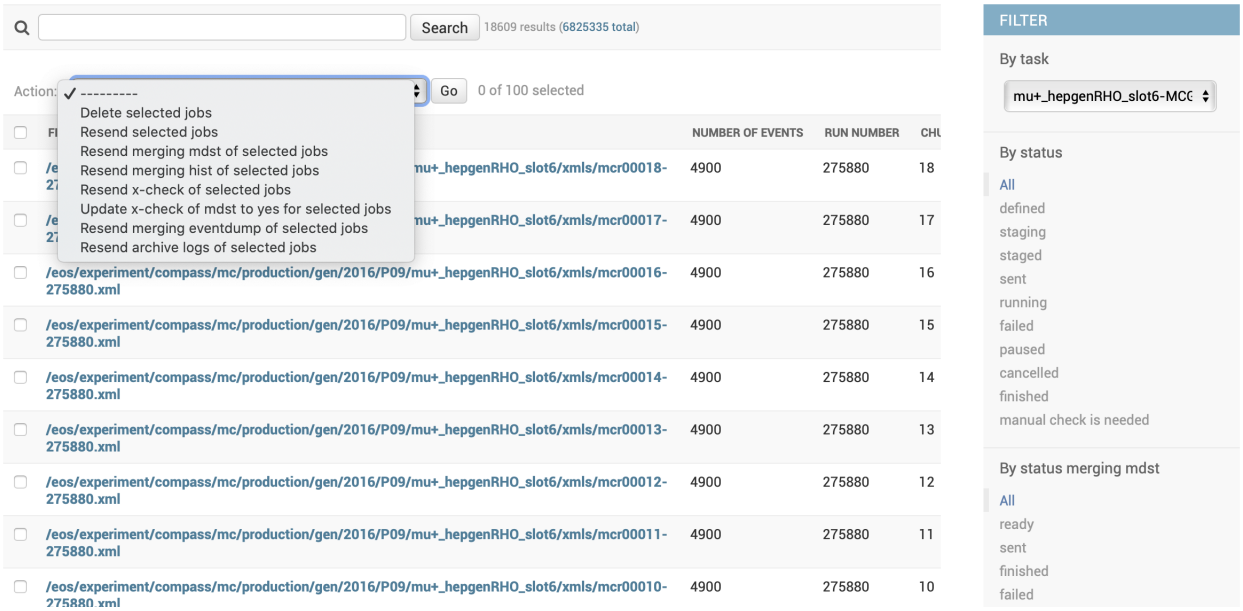

Рис. 6. Элементы web-интерфейса системы управления обработкой данных

Для упрощения управления обработкой данных разработан web-интерфейс, сочетающий в себе как пульт управления, так и большое количество средств отслеживания состояния заданий и задач. Элементы web-интерфейса системы управления представлены на рис. 6.

Сравнение систем управления обработкой по ключевым показателям по состоянию до и после модернизации кратко приведено в таблице 2.

Табл. 2. Ключевые отличия системы управления обработкой до и после модернизации

2015	2020
Описание заданий в виде файлов данных.	Описание заданий и управление обработкой через web-интерфейс.
Обработка ведется на сервисе lxbatch в ЦЕРНе.	Обработка данных ведется на нескольких сайтах: ЦЕРН, ОИЯИ, INFN Триест, Техасский суперкомпьютерный центр.
Реализован процесс реконструкции данных.	Реализованы процессы реконструкции данных и фильтрации событий, моделирования Монте-Карло.
Обработка ведется в контексте ответственного пользователя: система взаимодействует с сервисом lxbatch и системами хранения используя команды bsub, cp, mv и тп.	Обработка ведется в распределенной среде: система взаимодействует с вычислительными центрами посредством сервиса PanDA, использует аутентификацию по X509-сертификату, протокол передачи данных XRootD и сервис FTS.
Мониторинг в виде сводной web-страницы и набора проверочных скриптов для проверки результатов обработки.	Мониторинг заданий и задач в интерфейсе ProdSys, мониторинг сервисов PanDA, Auto Pilot Factory, HTCondor, EOS, FTS. Проверка результатов производится автоматически.

Результаты работы можно сформулировать следующим образом:

1. Предложена методика организации обработки данных в распределенной среде. Данная методика была успешно апробирована при создании программного комплекса для обработки данных эксперимента COMPASS и может быть использована для построения инфраструктуры распределенных вычислений и других, существующих или проектируемых, экспериментов.

2. Разработан и внедрен в эксплуатацию комплекс управления процессом обработки данных эксперимента COMPASS.

3. Расширилось количество физических процессов обработки данных: комплекс поддерживает все процессы, подразумевающие массовую обработку данных.

4. Подавляющее число операций системы управления процессами обработки данных выполняется в фоновом режиме набором независимых друг от друга сервисов, широко используются групповые операции и распараллеливание процессов.

5. Внедрение веб-интерфейса управления и большого набора средств мониторинга позволило значительно снизить нагрузку управляющих обработкой данных сотрудников.

6. Построенная инфраструктура показала способность выдерживать пиковые нагрузки, возникающие при использовании высокопроизводительных вычислительных систем, более чем в десять раз превышающие нагрузки при работе с вычислительными центрами среды грид.

7. За три года эксплуатации система без значительных доработок уже несколько раз прошла циклы адаптации к появляющимся вычислительным ресурсам различных типов: ресурсам среды грид и высокопроизводительным системам.

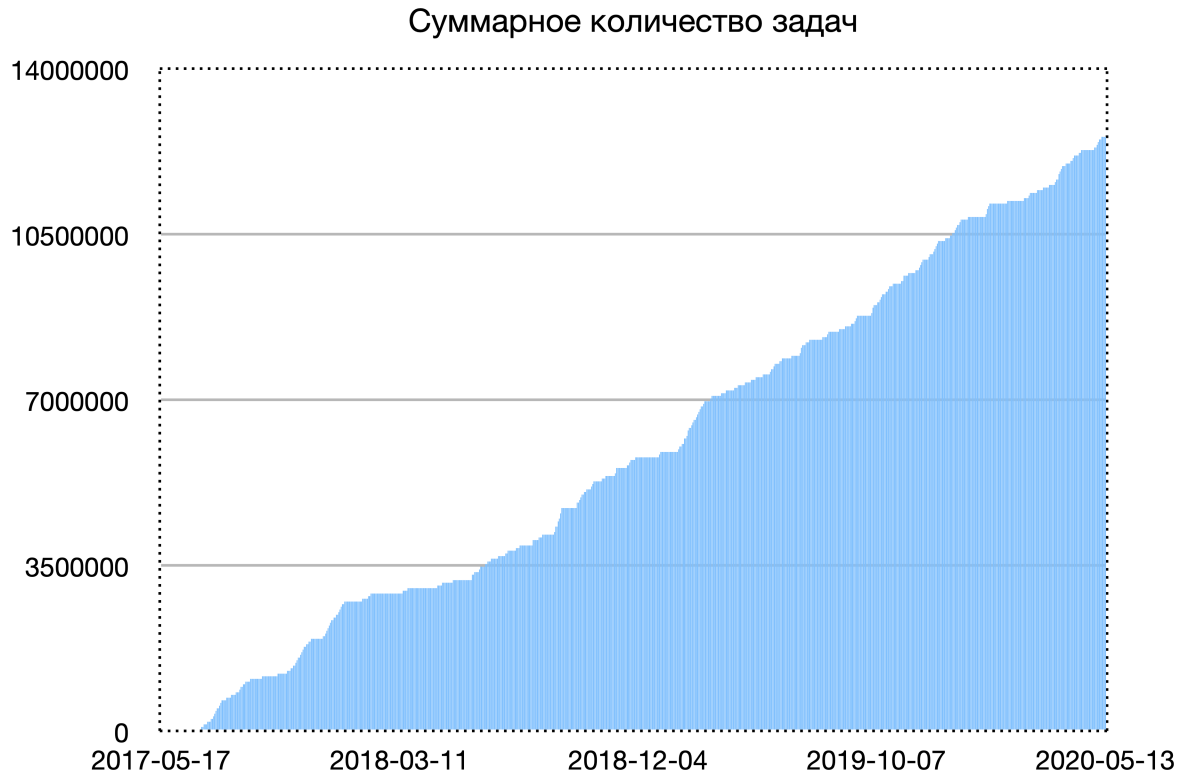

Рис. 7. Суммарное количество обработанных задач

Таким образом, спроектирован, разработан и введен в эксплуатацию программный комплекс, обеспечивающий эффективное использование

доступных эксперименту COMPASS географически распределенных вычислительных ресурсов. За время эксплуатации разработанного программного комплекса на вычислительных ресурсах ЦЕРН, ОИЯИ, Национального института ядерной физики в Триесте и Техасского суперкомпьютерного центра выполнено более 13 миллионов задач, при этом обработано около 150 миллиардов физических событий эксперимента COMPASS (рис. 7). Значимость созданного программного обеспечения для успешной реализации экспериментальной программы проекта COMPASS подтверждается письмом от координатора этого проекта Вансана Андрё.

Публикации по теме диссертации

1. *Oleynik D., Petrosyan A., Garonne V., Campana S.* ATLAS DQ2 deletion service // Journal of Physics: Conference Series. – 2012. – Vol. 396. – DOI: – 10.1088/1742-6596/396/3/032083. – URL: https://iopscience.iop.org/article/10.1088/1742-6596/396/3/032083.

2. *Stewart G.A., Petrosyan A. et al.* Advances in service and operations for ATLAS data management // Journal of Physics: Conference Series. – 2012. – Vol. 368. – DOI: 10.1088/1742-6596/368/1/012005. – URL: https://iopscience.iop.org/article/10.1088/1742-6596/368/1/012005.

3. *Petrosyan A., Oleynik D.* DDM DQ2 deletion service. Implementation of central deletion service for ATLAS experiment // Proceeding of The 5th International Conference "Distributed Computing and Grid-technologies in Science and Education" (GRID 2012). – 2012. – P. 189.

4. *Petrosyan A., Oleynik D., Belov S., Andreeva J., Kadochnikov I.* ATLAS Off-GRID sites (TIER-3) monitoring // Proceeding of The 5th International Conference "Distributed Computing and Grid-technologies in Science and Education" (GRID 2012). – 2012. – P. 195.

5. *Belov S., Kadochnikov I., Korenkov .V., Kutouski M., Oleynik D., Petrosyan A.* VM-based infrastructure for simulating different cluster and storage solutions used on ATLAS Tier-3 sites // Journal of Physics: Conference Series. – 2012. – Vol. 396.

– DOI: 10.1088/1742-6596/396/4/042036. – URL: https://iopscience.iop.org/article/10.1088/1742-6596/396/4/042036.

6. *Korenkov V., Petrosyan A. et al.* JINR (Dubna) and Prague Tier2 sites: Common experience in the WLCG grid infrastructure // Physics of Particles and Nuclei Letters. – 2013. – Vol. 10. – P. 288—294. – DOI: 10.1134/S1547477113030023. – URL: https://link.springer.com/article/10.1134/S1547477113030023.

7. *Anisenkov A., Di Girolamo A., Klimentov A., Oleynik D., Petrosyan A.* AGIS: The ATLAS Grid Information System // Journal of Physics Conference Series. – 2014. – Vol. 513. – DOI: 10.1088/1742-6596/513/3/032001. – URL: https://iopscience.iop.org/article/10.1088/1742-6596/513/3/032001.

8. JINR Participation in the WLCG Project / Korenkov V., Petrosyan A. [et al.] // JINR LIT Scientific Report 2012-2013/ ed. by G. Adam [et al.]. – Dubna. – 2014. – Chap. Networking, Computing, Information and Grid Technologies. – URL: https://lit.jinr.ru/sites/default/files/LIT_Report_2014_r.pdf.

9. *Andreeva J., Petrosyan A. et al.* Monitoring of large-scale federated data storage: XRootD and beyond // Journal of Physics: Conference Series. – 2014. – Vol. 513. – DOI: 10.1088/1742-6596/513/3/032004. – URL: https://iopscience.iop.org/article/10.1088/1742-6596/513/3/032004.

10. *Maeno T., Petrosyan A. et al.* Evolution of the ATLAS PanDA workload management system for exascale computational science // Journal of Physics Conference Series. – 2014. – Vol. 513. – https://iopscience.iop.org/article/10.1088/1742-6596/513/3/032062.

11. *Петросян А.Ш.* Современное использование сетевой инфраструктуры в системе обработки задач коллаборации ATLAS // Компьютерные исследования и моделирование. – 2015. – №6. – С. 1343–1349. – DOI: 10.20537/2076-7633-2015-7-6-1343-1349. – URL: http://crm.ics.org.ru/journal/article/2406/.

12. *Klimentov A., Petrosyan A. et al.* Next generation workload management system for big data on heterogeneous distributed computing // Journal of Physics Conference Series. — 2015. — Vol. 608. — https://iopscience.iop.org/article/10.1088/1742-6596/608/1/012040.

13. *De K., Petrosyan A. et al.* The future of PanDA in ATLAS distributed computing // Journal of Physics: Conference Series. – 2015. – Vol. 664. – DOI: 10.1088/1742-6596/664/6/062035. – URL. https://iopscience.iop.org/article/10.1088/1742-6596/664/6/062035.

14. *Megino F.B., Petrosyan A. et al. on behalf of the ATLAS collaboration* PanDA: Exascale Federation of Resources for the ATLAS Experiment at the LHC // EPJ Web Conferences. – 2016. – Vol. 108. – DOI: 10.1051/epjconf/201610801001. – URL: https://www.epj-conferences.org/articles/epjconf/abs/2016/03/epjconf_mmcp2016_01001/epjconf_mmcp2016_01001.html.

15. *Petrosyan A.* PanDA for COMPASS at JINR // Physics of Particles and Nuclei Letters. – 2016. – №13. P. 708–710. – DOI: 10.1134/S1547477116050393.

16. *Petrosyan A., Zemlyanichkina E.* PanDA for COMPASS: processing data via Grid // CEUR Workshop Proceedings. – 2017. – Vol. 1787. – P. 385–388. – URL: http://ceur-ws.org/Vol-1787/385-388-paper-67.pdf.

17. *Petrosyan A.* COMPASS Grid Production System // CEUR Workshop Proceedings. – 2018. – Vol. 2023. – P. 234–238. – URL: http://ceur-ws.org/Vol-2023/234-238-paper-37.pdf.

18. *Petrosyan A.* COMPASS Production System: Processing on HPC // CEUR Workshop Proceedings. – 2018. – Vol. 2267. – P. 139–144. – URL: http://ceur-ws.org/Vol-2267/139-144-paper-25.pdf.

19. *Petrosyan A.* COMPASS Production System Overview // EPJ Web of Conferences. – 2019. – Vol. 214. – DOI: 10.1051/epjconf/201921403039. – URL: https://www.epj-conferences.org/articles/epjconf/abs/2019/19/epjconf_chep2018_03039/epjconf_chep2018_03039.html.

20. *Petrosyan A., Malevanniy D.* Distributed data processing of the COMPASS experiment // CEUR Workshop Proceedings. – 2019. – Vol. 2507. P. 94–98. – URL: http://ceur-ws.org/Vol-2507/94-98-paper-15.pdf.