Федеральное государственное бюджетное учреждение
Национальный исследовательский центр
«Курчатовский институт»

На правах рукописи

Климентов Алексей Анатольевич

Методы обработки сверхбольших объемов данных в распределенной гетерогенной компьютерной среде для приложений в ядерной физике и физике высоких энергий

Специальность 05.13.11 — математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Автореферат

диссертации на соискание ученой степени
доктора физико-математических наук

Москва — 2017

Работа выполнена в Федеральном государственном бюджетном учреждении Национальный исследовательский центр «Курчатовский институт»

Научный консультант: Кореньков Владимир Васильевич, доктор технических наук, директор Лаборатории информационных технологий ОИЯИ (г. Дубна)

Официальные оппоненты: Аветисян Арутюн Ишханович, доктор физико-математических наук, профессор, член-корреспондент РАН, директор Института системного программирования РАН (г. Москва)

Воеводин Владимир Валентинович, доктор физико-математических наук, профессор, член-корреспондент РАН, заместитель директора НИВЦ МГУ, заведующий кафедрой суперкомпьютеров и квантовой информатики Московского государственного университета им. М.В. Ломоносова (г. Москва)

Оныкий Борис Николаевич, доктор технических наук, профессор, заведующий кафедрой анализа конкурентных систем НИЯУ МИФИ (г. Москва)

Ведущая организация: Институт ядерной физики им. Г.И. Будкера СО РАН (г. Новосибирск)

Защита состоится «____»_____ 2018 г. в _____ часов на заседании диссертационного совета Д 720.001.04 в Лаборатории информационных технологий Объединенного института ядерных исследований, г. Дубна Московской области.

С диссертацией можно ознакомиться в научной библиотеке ОИЯИ и на сайте http://wwwinfo.jinr.ru/announce_disser.htm

Автореферат разослан «_____» _____ г.

Учёный секретарь диссертационного совета
доктор физико-математических наук, профессор Иванченко И.М.

Общая характеристика работы

Актуальность темы. Исследования в области физики высоких энергий (ФВЭ) и ядерной физики (ЯФ) невозможны без использования значительных вычислительных мощностей и программного обеспечения для обработки, моделирования и анализа данных. Это определяется рядом факторов:

- большими объемами информации, получаемыми с установок на современных ускорителях;
- сложностью алгоритмов обработки данных;
- статистической природой анализа данных;
- необходимостью (пере)обрабатывать данные после уточнения условий работы детекторов и ускорителя и/или проведения калибровки каналов считывания;
- необходимостью моделирования условий работы современных установок и физических процессов одновременно с набором и обработкой «реальных» данных.

Введение в строй Большого адронного коллайдера (БАК, LHC) [1], создание и запуск установок такого масштаба, как ATLAS, CMS, ALICE [2–4], новые и будущие проекты класса мегасайенс (FAIR [5], XFEL [6], NICA [7]), характеризующиеся сверхбольшими объемами информации, потребовали новых подходов, методов и решений в области информационных технологий. Во многом это связано:

- со сложностью современных детекторов и количеством каналов считывания, например, размеры детектора ATLAS составляют 44 х 25 м, при весе 7000 т, детектор имеет 150 млн датчиков для считывания первичной информации;
- со скоростью набора данных (до 1 Пбайт/с);
- с международным характером современных научных сообществ и требованием доступа к информации для тысяч ученых из десятков стран (в научные коллаборации на LHC входят более восьми тысяч ученых из более чем 60 стран, сравнимое количество ученых будет работать в проектах FAIR и NICA);
- с высокими требованиями к обработке данных и получению физических результатов в относительно короткие сроки.

Научный прорыв 2012 года — открытие бозона Хиггса [8] — стал триумфом научного мегапроекта Большого адронного коллайдера, в последующие годы эксперименты на LHC исследовали свойства новой частицы, одновременно были увеличены светимость и энергия коллайдера. Современные эксперименты работают с данными в эксабайтном диапазоне и являются заметными «поставщиками» так называемых «Больших данных» и методов работы с ними. Как и в случае со Всемирной паутиной (WWW) — технологией, созданной в Европейском центре ядерных исследований (ЦЕРН) для удовлетворения растущих потребностей со стороны ФВЭ к обмену информацией между учеными и совместному доступу к ней, вызвавшей бурное развитие информационных технологий и систем связи в конце XX в., технологии Больших данных начинают влиять на исследования в других научных областях, включая нанотехнологии, астрофизику, биологию и медицину. Большие данные часто является связующим звеном, которое объединяет разработки в различных областях науки в единый мегапроект [9].

Стратегия научно-технологического развития России [10] определяет цель и основные задачи, а также приоритеты научных исследований и технологических разработок. Российские информационноемкие программы исследований, поддерживаемые Правительством РФ, такие как физика высоких энергий и ядерная физика, астрофизика, науки о Земле, биоинформатика и материаловедение будут производить эксабайты данных в ближайшем будущем. Проблемы, которые ставит развитие областей науки с большими объемами данных, многочисленны. Данные эксабайтного масштаба, как правило, распределены и должны быть доступны для больших международных сообществ. Для управления и обработки больших массивов данных необходимы многоуровневые интеллектуальные системы, системы управления потоками данных и заданий, контроля и мониторирования, а также системы хранения информации.

Вопросы разработки компьютерной модели, архитектуры распределенных и параллельных вычислительных систем для обработки данных, рассмотрение основополагающих принципов и моделей таких систем, анализ алгоритмов параллельных вычислений обсуждаются в классических работах начала XXI в. Э. Таненбаума и М. ван Стеена [11], а также В.В. Воеводина и Вл.В. Воеводина [12]. Следует отметить, что во второй половине XX в. классические работы Н.Н. Говоруна [13] о применении ЭВМ для обработки и анализа данных в области физики частиц, совпавшие по времени с запуском новых ускорителей в СССР (У10, У70), ЦЕРН (PS, SPS) и США (AGS, SLAC), оказали большое влияние на развитие методов обработки данных в ФВЭ и ЯФ и во многом заложили основу будущих компьютерных моделей обработки данных.

Уже на этапе создания архитектуры и компьютерной модели для экспериментов на LHC (1998–2001) стало очевидным, что хранение и обработка данных не могут быть выполнены в одном центре, даже таком крупном как (ЦЕРН). Следует отметить, что это понимание было вызвано техническими, финансовыми и социологическими причинами, в том числе и отсутствием в начале XXI в. решений, предложенных десятилетием позже ведущими коммерческими ИТ компаниями.

LHC — уникальный ускоритель, в котором каждые 50 нс происходит столкновение протонов при энергии 13 ТэВ с рождением около 1600 заряженных частиц, каждая из них регистрируется и анализируется триггером высокого уровня. В результате работы триггера около 1000 событий ежесекундно отбираются для дальнейшей обработки и анализа. Статистика, набранная за время работы LHC в 2010–2017 гг., составляет более 60 Пбайт «сырых» (неприведенных) данных. Управляемый объем данных современного физического эксперимента близок к 300 Пбайт. В 2014 и 2016 гг. физиками международного сотрудничества ATLAS было обработано и проанализировано 1.4 Эбайта данных. Беспрецедентный объем информации, поступающей в течение второй фазы работы LHC (2015–2019), и ожидаемое возрастание объема информации на следующих этапах работы коллайдера, как и требования к вычислительным комплексам на современных и будущих установках (FAIR, XFEL, NICA), потребовали разработки новой компьютерной модели, методики и методов управления загрузкой, созданию новых систем для обработки данных. Необходимым условием для своевременной обработки данных и получения физического результата в короткие сроки (в течение года) стал переход от использования гомогенной вычислительной среды (грид) к гетерогенной вычислительной инфраструктуре с использованием суперкомпьютеров (СК),

академических и коммерческих центров облачных вычислений, «волонтерских» компьютеров и отдельных вычислительных кластеров.

Еще на раннем этапе развития компьютерной модели LHC (2000-е годы) было принято решение объединить существующие и вновь создаваемые вычислительные центры (более 200) в распределенный центр обработки данных, и сделать это таким образом, чтобы физики университетов и научных организаций участвующих стран имели равные возможности для анализа информации. В результате работы физиков, ученых и инженеров в области ИТ была создана система, известная сегодня как WLCG (Worldwide LHC Computing Grid) [14]. На сегодняшний день WLCG — самая большая академическая распределенная вычислительная сеть в мире, состоящая из более чем 300 вычислительных центров в 70 странах. Более 8000 ученых использовали эти мощности для анализа данных в поисках новых физических явлений.

Грид-технологии были предложены в конце прошлого века Я. Фостером и К. Кессельманом. Основная концепция грид изложена в книге «The Grid: a Blueprint to the New Computing Infrastructure» [15]. Именно задачи ФВЭ и ЯФ привели к широкому использованию грид-технологий и потребовали существенных изменений и развития информационно-вычислительных комплексов (ИВК) в составе физических центров (в работе В.В. Коренькова [16] рассмотрена эволюция ИВК ОИЯИ в составе грид-инфраструктуры и приведено обоснование этого развития).

В WLCG ежедневно выполняется до трех миллионов физических задач, общее дисковое пространство превышает 400 Пбайт, результаты обработки данных архивируются, распределяются между центрами обработки и анализа данных и поступают непосредственно на «рабочее место» физика. Подобную систему можно сравнить с огромным вычислительным комплексом, узлы которого соединены высокоскоростным интернетом. Объемы передачи данных между центрами составляют до 10 Гбайт/с (среднее значение в течение дня). Создание системы заняло около 10 лет и потребовало вложений не только в инфраструктуру вычислительных центров во многих странах мира, но и развития сетевых ресурсов. Для обмена данными между центрами WLCG были созданы две компьютерные сети, ориентированные на задачи LHC: LHCOPN (LHC Optical Private Network) [17] и LHCONE (LHC Open Network Environment) [18]. Создание WLCG стало возможным в результате совместной работы тысяч ученых и специалистов и больших финансовых вложений.

Д-р Фабиола Джианотти (руководитель эксперимента ATLAS в 2008–2013 гг., директор ЦЕРН с 2014 г.) на семинаре, посвященном открытию новой частицы, сказала: «Мы наблюдаем новую частицу с массой около 126 ГэВ. Мы не смогли бы провести обработку и анализ данных так быстро, если бы не использовали грид. Центры во всех странах, участницах эксперимента, были задействованы в обработке данных LHC, практически это был стресс-тест для вычислительных мощностей, и грид показал себя высокоэффективной и надежной системой».

Роль распределенных компьютерных инфраструктур при обработке данных на первом этапе работы LHC подробно рассмотрена в работах автора, в том числе в соавторстве с В.В. Кореньковым и А.В. Ваняшиным [19, 20], опубликованных в 2012–2014 гг. Тогда же автором были сформулированы основополагающие принципы развития компьютерной модели для экспериментов в области физики частиц, новые требования к федерированию географически распределенных вычислительных

ресурсов, требования к глобальным системам для распределенной обработки данных и методам управления загрузкой в гетерогенной компьютерной среде [21].

Можем ли мы сказать, что LHC и WLCG выполнили поставленную задачу? Если говорить об открытии новой частицы, то да. Ни ускоритель Теватрон (в Лаборатории им. Э. Ферми, США), ни Большой электрон-позитронный коллайдер ЛЭП (LEP) в ЦЕРН за десятилетия работы не смогли зарегистрировать предсказанную в 1964 г. частицу. Однако более важно получить ответ на следующие вопросы. Достаточно ли классическое решение грид, реализованное в рамках проекта WLCG, для решения задач следующих этапов работы коллайдера? Как должна развиваться компьютерная модель для этапа superLHC (2022–2028), а также для новых комплексов, таких как FAIR, XFEL, NICA? Ответить на эти вопросы невозможно без понимания логики создания проекта WLCG и тех условий, в которых была разработана и реализована первая компьютерная модель распределенных вычислений для LHC. Необходимо проанализировать ограничения компьютерной модели и понять**,** насколько они носят фундаментальный характер, почему потребовалось создание новой компьютерной модели и распределенной системы обработки данных для второго и последующих этапов работы LHC. Применима ли новая компьютерная модель для экспериментов на установках класса мегасайенс в «эпоху Больших данных».

Работы по созданию концепции и архитектуры систем для распределенной обработки данных экспериментов в области ФВЭ, ЯФ и астрофизики были начаты в конце XX в. [22]. Тогда же были разработаны и реализованы первые сервисы для обнаружения ошибок и защиты информации, сервисы управления данными и ресурсами, сформулированы требования по взаимодействию сервисов внутри грид-систем. Следует отметить пионерские работы по развитию и созданию грид в России, в первую очередь в ЛИТ ОИЯИ (В.В. Кореньков), НИИЯФ МГУ (В.А. Ильин) [23–25], разработки ИПМ им**.** М.В. Келдыша [26]. Многие идеи по концепции вычислительных сред, определившие нынешние подходы, были предложены в работах Института системного анализа РАН (А.П. Афанасьев) [27, 28], а в работах НИВЦ МГУ рассмотрены вопросы эффективности работы суперкомпьютерных центров и проблемы их интеграции (Вл.В. Воеводин) [29, 30].

Важным этапом развития систем для обработки данных явилось обоснование принципов построения и архитектуры системы, разработка методов планирования выполнения заданий. Это позволило создать принципиально новое программное обеспечение, необходимое для управления данными и заданиями в распределенной среде, разработать методы оценки эффективности функционирования систем управления загрузкой, методы оценки эффективности работы ВЦ (в рамках грид-инфраструктуры) и методы распределения задач обработки и данных с целью оптимального использования вычислительного ресурса [31].

Компьютерная модель обработки данных физического эксперимента прошла в своем развитии много этапов: от модели централизованной обработки данных, когда все вычислительные ресурсы были расположены в одном месте (как правило там же, где находилась экспериментальная установка), к разделению обработки и анализа, которые по-прежнему велись централизованно, и моделирования данных, проводившегося в удаленных центрах. В эпоху LHC была предложена и реализована иерархическая компьютерная модель MONARC [32]. Следующим этапом стала модель равноправных центров внутри однородной грид-инфраструктуры — «смешанная модель» [33, 34]. В настоящее время компьютерная модель,

предложенная и реализованная автором [35], предполагает равноправное использование центров грид и интегрированных с грид ресурсов облачных вычислений и суперкомпьютерных центров в рамках единой гетерогенной среды. Дальнейшее развитие компьютерной модели для этапа superLHC и комплексов FAIR, XFEL, NICA потребовало разработки концепции и архитектуры единой федеративной киберинфраструктуры в гетерогенной вычислительной среде [36].

Для обработки и управления большими массивами данных необходимы многоуровневые интеллектуальные системы и системы управления потоками заданий. Создание таких систем имеет свою эволюцию, сравнимую по количеству этапов с развитием компьютерной модели физических экспериментов. От набора программ, написанных на скриптовых языках и имитирующих работу планировщика в рамках одного компьютера, до систем пакетной обработки, таких как LSF [37] или PBS [38], с последующей разработкой пакетов программ управления загрузкой промежуточного уровня грид (HTCondor [39]), и на последнем этапе развития — разработка и создание высокоинтеллектуальных систем управления загрузкой (AliEN, Dirac, PanDA [40–42]). Эти системы способны управлять загрузкой и позволяют обрабатывать данные одновременно в сотнях вычислительных центров. Практическое использование систем управления загрузкой показало их ограничения по параметрам масштабируемости, стабильности, возможности использования компьютерных ресурсов вне грид. Выявились трудности при интегрировании информации глобальных вычислительных сетей с информацией об имеющемся вычислительном ресурсе, быстроте «захвата» этого ресурса (что стало особенно заметно при переходе от модели MONARC к смешанной компьютерной модели, а также при использовании СК и коммерческих ресурсов облачных вычислений). Другой существенной проблемой стала реализация способа разделения вычислительного ресурса между различными потоками заданий: обработки данных, моделирования, анализа, а также предоставления вычислительного ресурса для задач эксперимента («виртуальной организации»), отдельных научных групп и ученых в рамках установленных квот использования вычислительного ресурса.

Таким образом, запуск Большого адронного коллайдера и создание новых ускорительных комплексов класса мегасайенс, характеризующихся сверхбольшими объемами информации и многотысячными коллективами ученых, обусловили новые требования к информационным технологиям и программному обеспечению. В эти же годы произошло качественное развитие информационных технологий, появление коммерческих вычислительных мощностей, превышающих возможности крупнейших ВЦ в области ФВЭ и ЯФ, развитие и резкое повышение пропускной способности глобальных вычислительных сетей. Требования по обработке данных на LHC и развитие ИТ привели к необходимости решения фундаментальной проблемы — разработки систем нового поколения для глобально распределенной обработки данных, разработки новой компьютерной модели физического эксперимента, позволяющей объединять различные вычислительные ресурсы и включать новые ресурсы (например, интегрировать ресурсы грид и суперкомпьютеры в единую вычислительную среду) [43].

Цель и задачи работы. Разработка и развитие методов, архитектур, компьютерных моделей и программных систем, реализация соответствующих программных и

инструментальных средств для приложений ФВЭ и ЯФ при обработке сверхбольших объемов данных.

Для достижения поставленной цели в диссертационной работе необходимо решить следующие основные задачи:

- Разработать компьютерную модель для экспериментов в области ФВЭ и ЯФ, позволяющую объединять высокопропускные вычислительные мощности (грид), высокоскоростные вычислительные мощности (суперкомпьютеры), ресурсы облачных вычислений и университетские кластеры в единую вычислительную среду.
- Разработать принципы построения и архитектуру системы для глобальной обработки данных эксабайтного масштаба для тысяч пользователей в гетерогенной вычислительной среде.
- Разработать методы управления потоками заданий в гетерогенной вычислительной среде, позволяющие учитывать неоднородность потоков заданий и запросов пользователей, с целью оптимального использования вычислительных ресурсов, доступных в современном физическом эксперименте.
- На основе разработанных принципов и архитектуры создать масштабируемую (обработка данных эксабайтного диапазона в O(100) центрах O(1000) пользователями O(10^6) научных заданий/день) систему для обработки данных современного физического эксперимента.
- Разработать систему мониторирования и оценки эффективности работы глобальной системы для обработки данных в распределенной гетерогенной компьютерной среде.

Научная новизна работы

- Разработана компьютерная модель современного физического эксперимента для управления, обработки и анализа данных эксабайтного диапазона в гетерогенной вычислительной среде.
- Реализация разработанной модели для приложений в области физики частиц впервые позволила использовать различные архитектуры: грид, суперкомпьютеры и ресурсы облачных вычислений для обработки данных физического эксперимента через единую систему управления потоками заданий, сделав это «прозрачно» для пользователя.
- Разработаны принципы построения, методы, архитектура и программная инфраструктура системы для глобальной распределенной обработки данных. На этой основе создана система управления потоками заданий, не имеющая мирового аналога по производительности и масштабируемости (более $2*10^6$ задач, выполняемых ежедневно в 250 вычислительных центрах по всему миру).
- Решена проблема разделения вычислительного ресурса между различными потоками научных заданий (обработка данных, Монте-Карло моделирование, физический анализ данных, приложения для триггера высшего уровня) и группами пользователей (эксперимент, научная группа, университетская группа, ученый).
- Разработаны новые методы управления научными приложения ФВЭ и ЯФ для суперкомпьютеров, с использованием информации о временно

свободных ресурсах, позволяющие повысить эффективность использования суперкомпьютеров (СК), в частности для Titan, Anselm, СК НИЦ КИ.

Защищаемые положения

- Новая компьютерная модель современного физического эксперимента позволяет использовать гетерогенные вычислительные мощности, включая грид, облачные ресурсы и суперкомпьютеры, в рамках единой вычислительной среды.
- Новые принципы построения и архитектура глобальной системы для обработки данных в гетерогенной вычислительной среде позволяют эффективно использовать вычислительные ресурсы и снимают противоречие по доступу к ресурсу между физическим экспериментом, группами пользователей и отдельными учеными.
- Разработанный комплекс методик, методов и система для управления потоками заданий, созданная на их основе, повышают эффективность обработки данных физических экспериментов и обеспечивают обработку данных в эксабайтном диапазоне в масштабе более $2*10^6$ задач в день в 200 вычислительных центрах для 1000 пользователей.
- Новые методы предсказания популярности (востребованности) классов и наборов данных, а также модель динамического управления данными в распределенной среде для сверхбольших объемов данных повышают эффективность использования распределенного вычислительного ресурса.
- Подсистема мониторирования и оценки эффективности работы глобальной системы для обработки данных обеспечивает высокий уровень автоматизации при анализе работы системы и сбоев в работе распределенной вычислительной инфраструктуры и ее аппаратно-программных компонент.

Практическая значимость. Основные результаты данной работы являются пионерскими и используются в действующих экспериментах в области ФВЭ и ЯФ и в других областях науки. В том числе результаты работ, положенных в основу диссертации, используются в двух крупнейших экспериментах в области ФВЭ и ЯФ — ATLAS и ALICE на LHC, эксперименте COMPASS на SPS, а также для приложений биоинформатики на суперкомпьютерах НИЦ КИ:

- вычислительные модели экспериментов ATLAS и AMS опираются на результаты работ, положенных в основу диссертации;
- разработанная и созданная система управления потоками заданий в гетерогенной компьютерной среде используется в экспериментах на ускорителях LHC и SPS и принята в качестве базовой для будущего коллайдера NICA;
- разработанная система для обработки данных была также применена для исследований ДНК мамонта на суперкомпьютере НИЦ КИ и в европейском проекте BlueBrain.

Разработанная система управления загрузкой не имеет мировых аналогов по масштабируемости и отказоустойчивости. До $2*10^6$ задач выполняются ежедневно, в

2016 г. физиками ATLAS было обработано 1.4 Эбайта данных. Таким образом, система уже сейчас работает в эксабайтном диапазоне.

Реализация результатов работы. Результаты диссертации были получены под руководством и при личном участии соискателя в следующих международных проектах: WLCG — проект грид для LHC, megaPanDA — проект по созданию нового поколения системы управления заданиями в гетерогенной компьютерной среде, проект ATLAS на LHC, проекты AMS-01 и AMS-02 на Международной космической станции (МКС), проект metaMiner — по созданию системы поиска аномалий и предсказания поведения комплексных распределенных вычислительных систем, проект Federated Storage — по созданию прототипа распределенной компьютерной среды.

Автор диссертации внес определяющий вклад при выполнении ряда национальных российских и международных проектов, в том числе L3, AMS, ATLAS, megaPanDA, в которых он являлся одним из руководителей (или руководителем) компьютерной и программной частями проекта и одновременно основным архитектором создаваемых систем и программного обеспечения.

Работы в 2013–2016 гг. были поддержаны грантом Министерства образования и науки РФ по привлечению ведущих ученых, тремя грантами РФФИ и грантом РНФ. В настоящее время автор является руководителем мегагранта и руководителем двух международных проектов совместно с ЦЕРН и DESY — «Создание федеративного распределенного дискового пространства» и «Использование алгоритмов машинного обучения для приложений ФВЭ».

Базовая вычислительная модель реализуется в проекте ATLAS на LHC и рассматривается как основная для ускорительного комплекса NICA (ОИЯИ).

Созданы системы управления загрузкой для распределенной обработки данных в НИЦ КИ (для приложений биоинформатики), ОИЯИ (для эксперимента COMPASS в ЦЕРН), ЦЕРН (эксперименты ATLAS), ORNL (для высокоинтенсивных научных приложений), EPFL (проект BlueBrain, Лозанна, Швейцария), ASGC (эксперимент AMS-02, Тайпей, Тайвань).

Апробация диссертации. Результаты работы являются итогом более чем 20-летней научной и организационной деятельности соискателя. Основные результаты диссертации докладывались и обсуждались на научных семинарах НИЦ «Курчатовский институт», ОИЯИ, ЦЕРН, БНЛ, НИЯУ МИФИ, ТПУ, докладывались на конференциях, рабочих совещаниях и научных семинарах экспериментов COMPASS, AMS, L3. Результаты работ регулярно обсуждались международными научными коллаборациями ATLAS и ALICE, в том числе на пленарных заседаниях во время конференций и на симпозиумах консорциума WLCG. Результаты, представленные в диссертации, докладывались на международных и российских конференциях, в том числе:

- международных конференциях «Computing in High Energy Physics» (CHEP): 2002 (Пекин, КНР), 2004 (Интерлакен, Швейцария), 2007 (Ванкувер, Канада), 2009 (Прага, Чехия), 2011 (Тайпей, Тайвань), 2012 (Нью Йорк, США), 2015 (Окинава, Япония), 2016 (Сан-Франциско, США);
- международных конференциях «Advanced computing and analysis techniques in physics research» (ACAT): 1994 (Комо, Италия), 2002 (Москва, Россия),

2008 (Эричи, Италия), 2014 (Прага, Чехия — пленарный доклад), 2016 (Вальпараисо, Чили), 2017 (Сиэттл, США);

- международных конференциях по физике высоких энергий ICHEP: 2012 (Мельбурн, Австралия), 2014 (Валенсия, Испания), 2016 (Чикаго, США);
- международной конференции Real-Time Computer Applications in Nuclear and Plasma Physics, 1994 (Дубна, Россия);
- международной конференции «Calorimetry in High Energy Physics», 1999 (Лиссабон, Португалия);
- международном симпозиуме IEEE «Nuclear Science Symposium and medical imaging conference» (IEEE NCC/MIC), 2003 (Портланд, США);
- международном совещании «Physics and Computing at ATLAS», 2008 (Дубна, Россия);
- международном симпозиуме «Grid and Clouds Computing», 2010 (Тайпей, Тайвань);
- международных симпозиумах «Nuclear Electronics and Computing» (NEC): 2011 (Варна, Болгария — приглашенный доклад), 2015 (Будва, Черногория), 2017 (Будва, Черногория);
- международных конференциях «Распределенные вычисления и грид технологии в науке и образовании» (GRID): 2012 (Дубна, Россия), 2016 (Дубна, Россия — приглашенный доклад);
- международных конференциях «Наука Будущего» (Science of the Future): 2014 (Санкт-Петербург, Россия), 2016 (Казань, Россия — приглашенный доклад), 2017 (Нижний Новгород, Россия);
- международных конференциях «Smoky Mountains Computational Science and Engineering», июль 2014, (Ноксвилл, США — приглашенный доклад), сентябрь 2017 (Гатлинбург, США — приглашенный доклад);
- международной конференции «Data Analytics and Management in Data Intensive Domain», 2015 (Обнинск, Россия);
- VI Московском суперкомпьютерном форуме, 2015 (Москва, Россия);
- международной конференции «Supercomputing 2016», 2016 (Солт Лейк Сити, США);
- конференции консорциума World LHC Computing Grid, 2016 (Лиссабон, Португалия);
- международной конференции «Instrumentation for Colliding Beam Physics». 2017 (Новосибирск, Россия — приглашенный доклад).

Соискатель являлся членом программных и международных комитетов конференций CHEP, NEC, GRID, DAMDID, а также (со)руководителем международных симпозиумов по обработке данных LHC (Дубна, 2008, 2014), суперкомпьютерам (Нью-Йорк, США, 2013), методам машинного обучения для научных приложений в физике высоких энергий (Москва, 2016) и «Программное обеспечение для будущих экспериментов» (Петергоф, 2017), где также были представлены результаты работ, положенных в основу данной диссертации.

Публикации и личный вклад автора. Изложенные в диссертации результаты получены соискателем в результате его многолетней научной и организационной

деятельности по разработке и созданию программного обеспечения, систем для обработки и анализа данных и компьютерных моделей для экспериментов ФВЭ, ЯФ и астрофизики (L3, AMS, ATLAS), в частности системы управления загрузкой в гетерогенной компьютерной среде для этих экспериментов, а также выполненных им работ для экспериментов класса мегасайенс в Лаборатории «Технологии Больших данных» НИЦ «Курчатовский институт», созданной и руководимой соискателем.

Все исследовательские работы и разработки по теме диссертации, от постановки задачи и выбора методики до получения результатов, выполнены соискателем и/или под его непосредственным руководством, его вклад в эти работы является определяющим. Все выносимые на защиту результаты получены соискателем лично.

По теме диссертации автором опубликовано свыше 150 печатных работ, в том числе по основным результатам — 68 (из них 47 из перечня ведущих рецензируемых научных изданий). Результаты работы также опубликованы в отчетах по руководимым автором инфраструктурным и научным проектам в рамках мегагранта Правительства РФ и проектам, поддержанным РНФ и РФФИ.

Структура и объем диссертации. Диссертация состоит из введения, 4 глав, заключения, списка литературы из 115 наименований**,** полный объем работы составляет 238 страниц.

Краткое содержание работы

Во **введении** обосновывается актуальность темы, приводятся цель и задачи работы, формулируются научная новизна и практическая значимость полученных результатов, приводятся выносимые на защиту положения и информация об апробации работы.

Первая глава диссертации посвящена развитию компьютерной модели экспериментов в области физики элементарных частиц, физики высоких энергий, ядерной физики и для эксперимента AMS-02 на МКС. Глава состоит из трех частей.

В первой части приведен краткий обзор компьютерных моделей для наиболее значимых этапов развития экспериментов в области ФВЭ и ЯФ за последние 20 лет, а также для экспериментов AMS и AMS-02 (в них автором была предложена и реализована распределенная компьютерная модель обработки данных). Подробно рассмотрена роль компьютинга и программного обеспечения (ПО) для экспериментов в области частиц.

Рассмотрено как менялись требования к ПО, информационным технологиям (ИТ) и состав научных коллабораций за последние десятилетия (в таблице 1 приведен сравнительный анализ экспериментов в области физики частиц на ведущих мировых ускорителях во второй половине XX и начале XXI в.).

Таблица 1. Характеристики экспериментов в области физики частиц в последние 60 лет

Годы	Число сотрудников эксперимента	Объем данных, технология хранения и обработки данных и информации
Конец 1950	2–3	Кбиты, записи в рабочих журналах
1960 (У7)	10–15	Кбайты, перфокарты, бумажные носители
1970 (У10,У70, PS, AGS)	~35	Мбайты, магнитные ленты Онлайн обработка: PDP 8, оффлайн обработка: ЕС, IBM 360
1980 (SPS, У70)	~100	Гбайты, магнитные ленты и диски Онлайн обработка: Caviar, PDP 70, VAX, СМ4 Оффлайн обработка: ЕС, IBM 370, БЭСМ 6, VAX 8800
1990 (LEP, SLAC, Теватрон, RHIC)	700–800	Тбайты, магнитные ленты, диски Онлайн обработка: VAX, спец.процессоры; оффлайн Оффлайн обработка: ЕС, IBM 370, VAX 8800, Appollo, SGI, Sun
2010 (LHC)	~3000	Пбайты, магнитные ленты, диски Онлайн обработка: кластеры, графические процессоры Оффлайн обработка: грид (200 ВЦ)

Как следует из таблицы, по мере создания новых ускорителей менялся не только объем информации и технологии хранения и обработки данных, но и произошел качественный рост числа участников физического эксперимента, что потребовало организации обработки и анализа данных для тысяч ученых в сотнях ВЦ.

Далее в разделе анализируется, как менялись требования к мощности вычислительного ресурса по мере создания и ввода в эксплуатацию экспериментов на LHC. Обоснована необходимость создания глобальной распределенной системы для обработки данных, рассмотрены причины и факторы, повлиявшие на выбор иерархической модели организации центров при распределенной обработке данных, подробно рассмотрена компьютерная модель, предложенная проектом MONARC. В этой части также рассматривается концепция грид, перечислены основные компоненты грид-инфраструктуры и взаимодействие между ними, рассмотрен вопрос, как принятие парадигмы грид повлияло на управление потоками данных для экспериментов в области физики частиц.

Вторая часть первой главы посвящена реализации модели MONARC на первом этапе работы коллайдера LHC. Распределение центров по уровням и функции центров для каждого из предложенных в модели уровней (T0:T1:T2).

В главе подробно проанализированы ограничения модели, в том числе, связанные с иерархией центров и статическим характером связки 1:T1–n:T2, когда любой сбой в работе центра первого уровня (T1) практически останавливал работу всех связанных с ним центров второго уровня (T2), в результате чего эксперименты лишались мощностей до 10 центров одновременно. Другими существенными ограничениями иерархической модели были:

- Определение вычислительного ресурса как совокупности вычислительных узлов, дискового пространства и систем архивирования информации без учета пропускной способности WAN и качества линий связи.
- Статическая методика распределения данных между центрами: а) было определено изначально, какой объем данных (реальных и моделируемых)

будет находиться в каждом центре; б) было определено изначально, сколько копий данных каждого типа («сырых» и приведенных) будет распределено между центрами обработки.

- Отсутствие понятия «популярности» (востребованности) для данных и групп данных.
- Предложенная методика обработки данных при статическом характере организации вычислительного ресурса и распределения данных между центрами грид-инфраструктуры. Наиболее точно ее можно определить слоганом «задачи обработки идут к данным». Такой подход привел к задержке при обработке и моделировании данных, так как требовал одновременного наличия данных и свободного вычислительного ресурса в одном и том же ВЦ.
- Вычислительный ресурс центров был ориентирован на среднюю загрузку. В результате это вело к недостатку вычислительного ресурса в периоды пиковой нагрузки (работа коллайдера с повышенной светимостью), анализ данных в период, предшествующий основным научным конференциям, при сверке гипотез между несколькими научными группами и/или экспериментами и к неоптимальному использованию вычислительного ресурса во время плановых остановок коллайдера, праздников и т.д.
- Ограничения самой модели, предполагающей гомогенность используемого ресурса, наличие ПО промежуточного уровня («middleware») во всех центрах обработки данных.

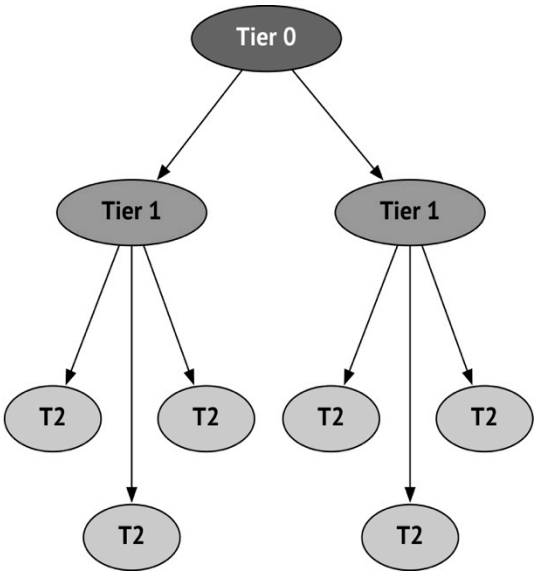

Рисунок 1. Иерархическая компьютерная модель (проект MONARC)

При всех ограничениях реализация модели распределенных вычислений, предложенная проектом MONARC (рис. 1), стала значительным шагом в развитии компьютинга в области физики частиц. Более 200 центров в 60 странах мира вошли в консорциум WLCG, был получен первый опыт по распределенной обработке данных. Вычислительный ресурс распределялся следующим образом: 15% находилось в ЦЕРН (уровень T0), 40% распределялось между 11 центрами уровня T1 (данное распределение было крайне неравномерным, так, для экспериментов ALICE, ATLAS и CMS вклад центров уровня T1 варьировался от 5 до 40%), 45% ресурса распределялось между центрами уровня T2 (рис. 2).

Реализованная модель успешно работала на первом этапе работы коллайдера, но поддержание ее в рабочем состоянии требовало больших человеческих затрат, как со стороны персонала ВЦ (инфраструктура), так и со стороны научных коллабораций для поддержания работы сервисов управления и обработки данных.

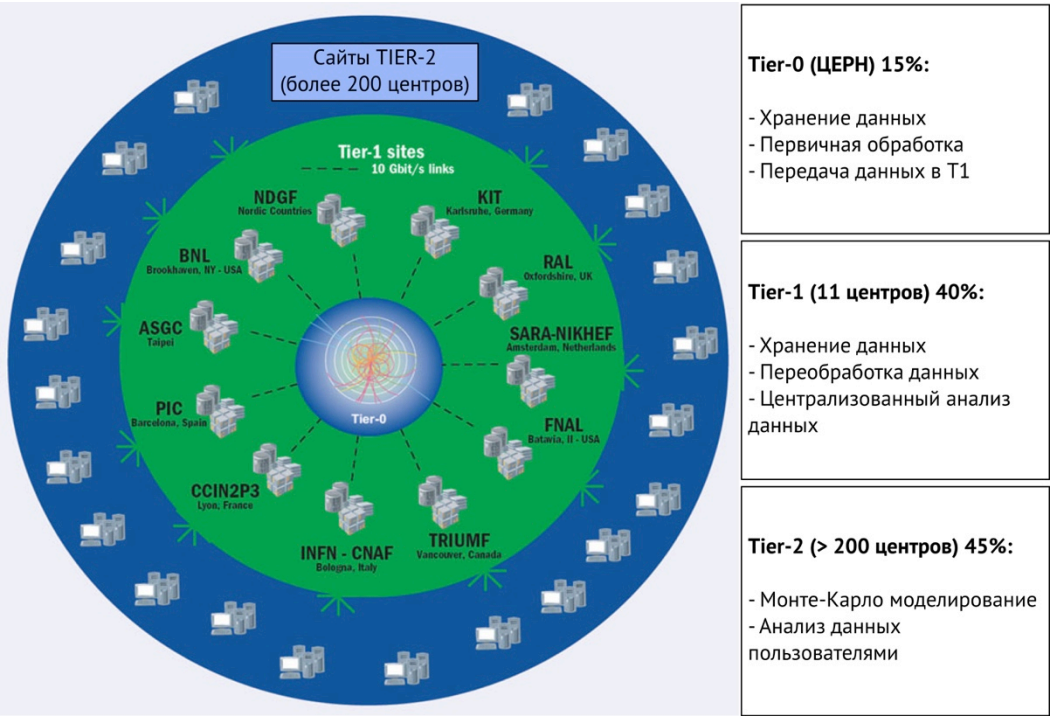

Рисунок 2. Организация грид-сайтов WLCG на момент запуска LHC

Схематично компьютерная модель для первого этапа работы экспериментов на LHC представлена на рисунке 3.

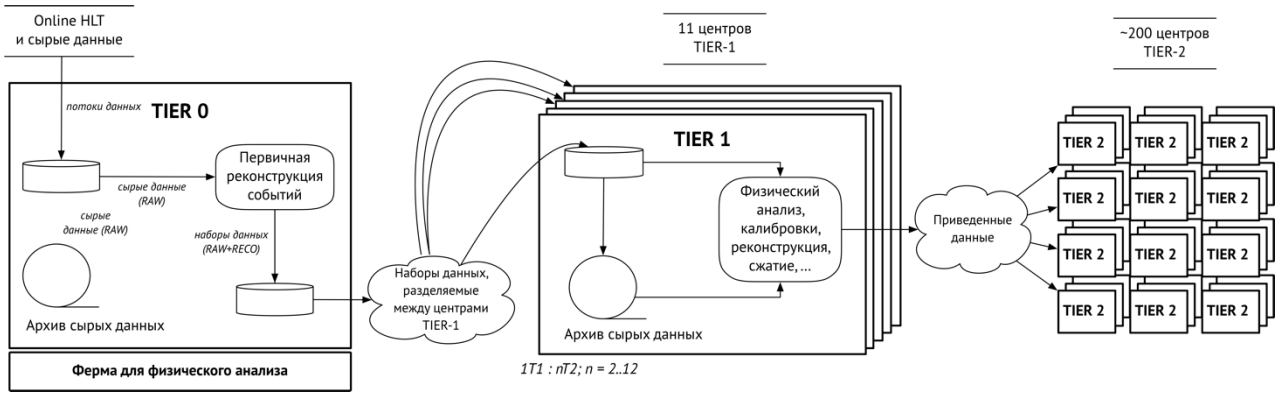

Рисунок 3. Компьютерная модель во время первого этапа работы LHC

В третьей части обоснован переход от модели MONARC к «смешанной» компьютерной модели и определены базовые принципы новой модели в рамках грид-инфраструктуры.

Рассмотрены классы данных и классы потоков заданий физического эксперимента. Описаны предложенная автором иерархическая модель данных и введенные понятия «класс данных», а также иерархия данных: *событие, файл, датасет, контейнер*. Описаны предложенные методы определения значимости и популярности классов данных на основе их востребованности и научных приоритетов. Дано обоснование и описание термодинамической модели управления данными, реализация которой привела к оптимизации распределения данных между центрами и использования вычислительного ресурса в целом. Результатом введения

термодинамической модели и динамической системы распределения данных стала возможность увеличения набора данных в 4 раза (со 100 событий/с после триггера последнего уровня до 400 событий/с) без увеличения имеющегося дискового ресурса для хранения информации. Рассмотрена разработанная методика оценки стабильности работы центров грид и роль глобальной вычислительной сети как ресурса, который необходимо учитывать наряду с дисковым и вычислительным ресурсом при выборе ВЦ для обработки и хранении данных. Новые методы для определения стабильности работы грид центров позволили отказаться от двухшаговой передачи данных между центрами уровня Т2 (что являлось необходимым условием модели MONARC) и использовать их дисковый ресурс для хранения популярных данных. Рассмотрен метод динамического распределения данных между центрами обработки в зависимости от популярности данных. На примере эксперимента ATLAS показано, как увеличилось количество центров, используемых для постоянного хранения данных (на 100%), и как изменилось количество копий наборов данных за счет динамического изменения количества копий для (не)востребованных наборов данных. В результате стала возможной реализация «смешанной» компьютерной модели (рис. 4) и эффективное использование ресурса центров всех уровней в рамках грид-инфраструктуры.

Основным выводом данной главы является: разработка приведенных в главе методик и методов, расширение понятия «вычислительный ресурс» за счет включения в его состав ресурса WAN (что дало возможность отказаться от иерархической компьютерной модели и реализовать «смешанную модель» для распределенной обработки и управления данными в грид-среде).

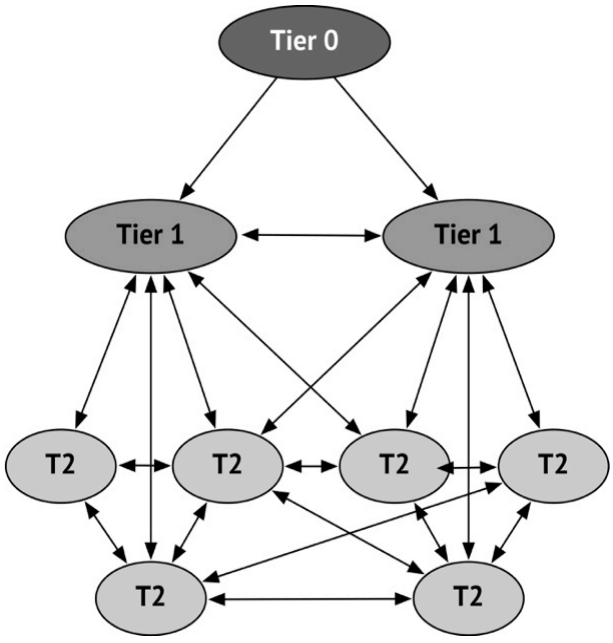

Рисунок 4. «Смешанная» компьютерная модель для экспериментов LHC

Результаты исследований, изложенные в первой главе, подтверждают следующее защищаемое положение: разработаны методы определения популярности классов данных и наборов данных, а также разработана модель динамического управления данными в распределенной среде для сверхбольших объемов данных.

Вторая глава посвящена анализу требований к вычислительной инфраструктуре для обработки, моделирования и анализа данных, а также роли суперкомпьютеров для приложений ФВЭ и ЯФ.

В первой части второй главы рассмотрены требования к вычислительному ресурсу на втором и последующих этапах работы LHC. На рисунке 5 показано количество задач, ожидавших свободного вычислительного ресурса на первом этапе работы LHC (2010–2013). Хорошо видно, что в случае пиковых нагрузок, как правило предшествующих основным научным конференциям и этапам массовой переобработки данных, очередь на выполнение могла достигать более миллиона задач в день. Увеличение энергии и светимости ускорителя на втором и последующих

этапах работы LHC приведет к более чем двукратному увеличению размера события и, соответственно, времени обработки, что в свою очередь потребует большего ресурса для обработки и хранения данных. Объемы данных к этапу superLHC (2023–2028) возрастут более чем в 100 раз (рис. 6). Это ведет к необходимости увеличения существующего вычислительного ресурса за счет суперкомпьютеров, ресурсов облачных вычислений, университетских кластеров и создания единой среды (киберинфраструктуры) для обработки и анализа данных экспериментов на LHC.

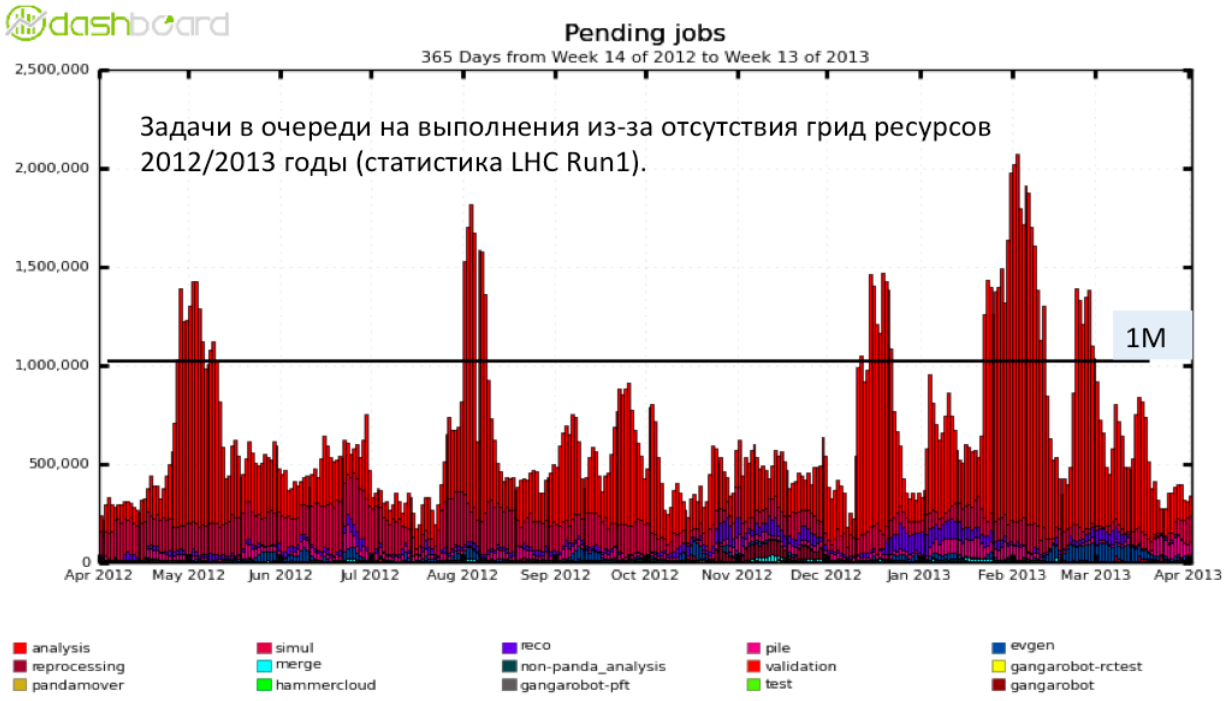

Рисунок 5. Количество задач в очереди на выполнение из-за отсутствия вычислительного ресурса

В разделах 2.1 и 2.2 рассмотрены существующие проблемы при создании федеративной киберинфраструктуры и принципиальные подходы, которым необходимо следовать:

- единый метод и уровень абстракции управления ресурсами;
- общая система управления загрузкой и данными в гетерогенной компьютерной среде;
- интегрируемые и развиваемые средства для управления программным обеспечением федеративной распределенной киберинфраструктуры.

Рассмотрены вопросы конвергенции высокопропускных HTC (High Throughput Computing) и высокопроизводительных HPC (High Performance Computing) вычислений. Это необходимо для понимания того, как научные приложения ФВЭ и ЯФ (это типичный пример HTC приложений, в которых практически не используются параллельно выполняемые и связанные между собой задачи) могут выполняться на суперкомпьютерах и как при этом будет меняться модель обработки данных.

Рост запросов на вычислительные мощности в экспериментах на LHC

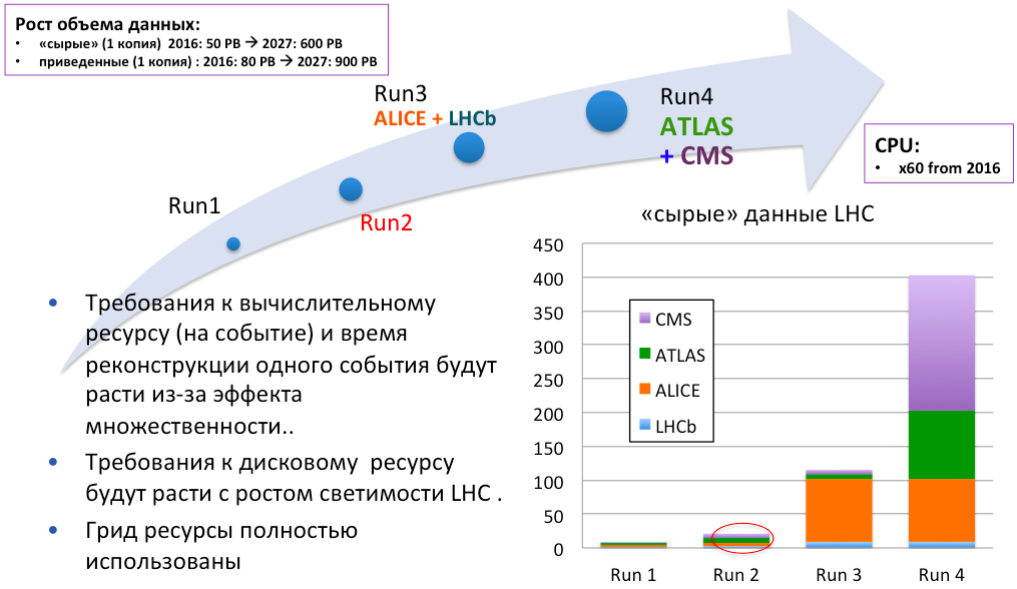

Рисунок 6. Рост запросов на вычислительные мощности экспериментов LHC и ожидаемый объем данных для этапа высокой светимости LHC

Подробно обоснована роль системы управления потоками заданий и систем для распределенной обработки данных как необходимого слоя при интеграции разнородных вычислительных систем. На рисунке 7 представлена иерархия базовых уровней системы управления загрузкой и вычислительных мощностей распределенной системы обработки данных в гетерогенной компьютерной среде.

Далее (**раздел 2.3**) рассматривается роль суперкомпьютеров для выполнения научных программ в ФВЭ и ЯФ в целом и для LHC экспериментов в частности, а также возможная роль приложений физики частиц для суперкомпьютеров.

Глава содержит изложение понятий, методов и программных средств, лежащих в основе разработок, которые описаны в последующих главах диссертационной работы.

Выводы ко второй главе формулируются следующим образом. Обоснована значимость разработки новой архитектуры системы управления потоками заданий и программного обеспечения для ее реализации. Обоснованы принципы распределенной системы для обработки данных, которая могла бы работать с динамически изменяющимися вычислительными ресурсами и использовать мощности, доступные в течение относительно коротких временных интервалов. Также показана необходимость расширения модели компьютинга и введения понятия ВЦ без «дискового элемента», поскольку ни суперкомпьютерные центры, ни центры облачных вычислений не предоставляют дисковый ресурс для постоянного хранения данных (речь идет о дисковом ресурсе масштаба сотен Пбайт, необходимых для экспериментов на LHC).

В главе подтверждено следующее защищаемое положение: разработанная компьютерная модель современного физического эксперимента позволяет использовать гетерогенные вычислительные мощности в рамках единой вычислительной инфраструктуры.

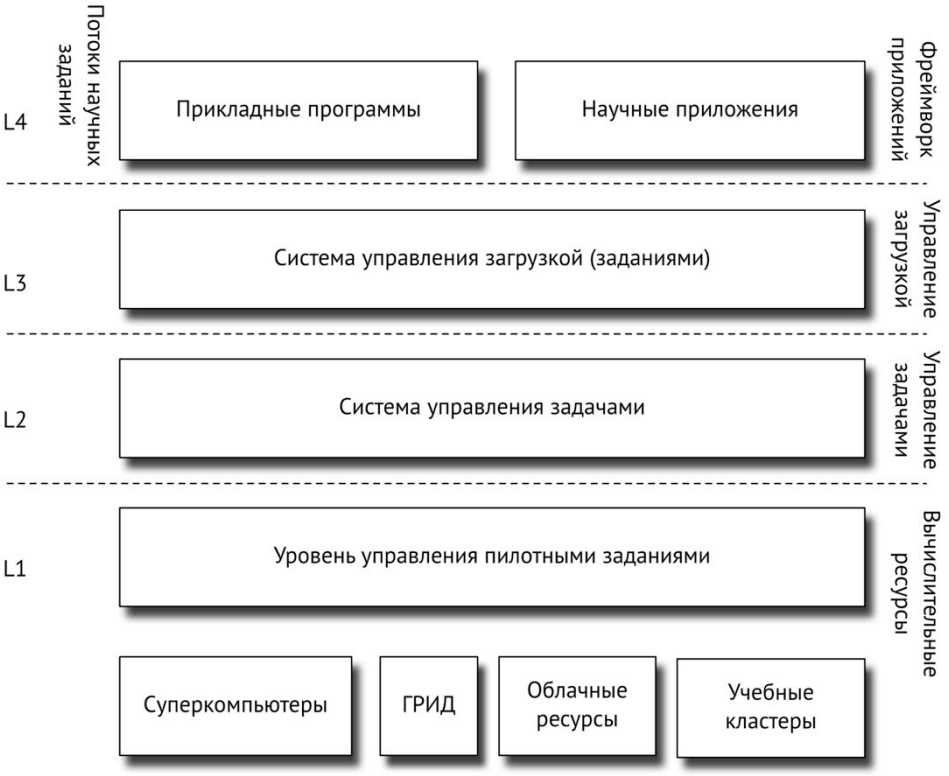

Рисунок 7. Иерархия базовых уровней системы управления загрузкой и вычислительных мощностей для обработки данных в гетерогенной компьютерной среде

Третья глава диссертации посвящена разработке концепции, методов и архитектуры системы управления заданиями в распределенной гетерогенной компьютерной среде. В главе проведен анализ классов научных приложений в ФВЭ и ЯФ, предложена модель данных системы распределенной обработки. Основными типами потоков заданий являются:

- потоки заданий, выполняемые системой для распределенной обработки данных эксперимента (VO — «виртуальной организации»): а) обработка и (пере)обработка данных, б) Монте-Карло моделирование, в) создание приведенных данных для физического анализа, г) обработка данных для триггера высокого уровня, д) специальные случаи, например, «обработка поездом» для всех или нескольких физических групп;
- потоки заданий, выполняемые системой по требованию физических групп эксперимента: а) создание данных для физического анализа, проводимого физической группой, б) анализ данных;
- потоки заданий, выполняемые отдельными пользователями (физический анализ данных).

Определены требования к системе управления загрузкой в гетерогенной среде. Одним из основных требований является создание модели выполнения заданий для динамически определяемых федеративных гетерогенных ресурсов, независящей от типа инфраструктуры, ее неоднородности с учетом параметров, определяющих динамику возможного изменения ресурса. В целом предлагаемая модель выполнения заданий и управления загрузкой имеет следующие основные особенности:

- интеграция информации о выполняемом задании и ресурсах;

- стратегия выполнения основывается на последовательности решений, используемой для выполнения данного задания, которая может измениться при условии изменения инфраструктуры и/или типа задания.
- такой подход позволяет интегрировать рабочую нагрузку и ресурсы: а) оценить потребности, необходимые для выполнения задания (рабочей нагрузки), б) оценить возможности ресурса, выработать стратегию выполнения и начать выполнение задания.

В **разделах 3.2–3.4** описывается разработанная логическая модель данных для системы глобальной распределенной обработки данных. Определены основные сущности модели: *запрос, список входных параметров, шаг обработки, задание, задача, пилотная задача*. И далее описывается архитектура системы глобальной обработки данных физического эксперимента. Сформулированы требования к функциональности, масштабируемости и отказоустойчивости такой системы. Подробно рассмотрены этапы выполнения научного задания и основные компоненты системы для распределенной обработки данных. Система должна отвечать следующим основным требованиям:

- Сотни вычислительных центров, распределенных по всему миру, для конечного пользователя должны «выглядеть» как единый ВЦ.
- Система должна обеспечивать доступ и выполнение заданий на O(100) ВЦ для $O(10^3)$ пользователей и $O(10^6)$ задач в день.
- Вычислительные центры могут быть представлены не только центрами грид, но и суперкомпьютерными центрами и центрами облачных вычислений, при этом все центры рассматриваются как равноправные участники. Весь набор центров составляет единый пул вычислительных ресурсов. Отметим, что этот подход стал возможен после перехода к описанной в первой главе диссертации «смешанной» компьютерной модели и обеспечил создание гетерогенной киберинфраструктуры, интегрировав дополнительные вычислительные ресурсы с грид-ресурсами.
- Очередь на выполнение пользовательских заданий в распределенной среде должна быть единой, сравнимой по функциям с очередью пакетной обработки на локальном ВЦ. Все участники эксперимента («виртуальной организации») должны иметь доступ к ресурсам VO через единую систему запуска заданий или, на более высоком уровне, через систему *«запросов»*.
- Ошибки в работе вычислительных центров и задержки, связанные с распределенным характером обработки, должны быть минимизированы. Для этого необходимо использовать «позднюю привязку» реально выполняемой задачи к вычислительному ресурсу, используя концепцию «пилотных задач».
- Сложность и разнообразие промежуточного программного обеспечения (ППО) грид должны быть «скрыты» от пользователя. Для этого необходимо выполнение следующих условий: а) система управления загрузкой «знает» о ППО и взаимодействует с ним (в обоих направлениях), конечный пользователь взаимодействует только с системой управления загрузкой; б) механизмы автоматизации управления загрузкой «скрыты» от пользователя.
- Изменения и эволюция ППО не должны менять пользовательский интерфейс управления заданиями.

- Система должна быть адаптируема к изменениям в аппаратном и программном обеспечении вычислительных центров, и эти изменения не должны быть видимы для пользователя.
- Единая система управления загрузкой должна использоваться для всех классов задач физического эксперимента, таких как моделирование, реконструкция, физический анализ, а также для потоков заданий, генерируемых экспериментом, физическими группами, отдельными учеными.
- Система должна обладать высокой степенью автоматизации в части обнаружения и «исправления» ошибок, связанных со сбоями в работе распределенной инфраструктуры.
- Мониторирование и контроль должны быть частью системы управления загрузкой.
- Задания должны использовать ресурс, выделенный для работы виртуальной организации, согласно единой системе приоритетов, пользовательских квот, квот для классов задач.

Далее в разделе определены принципы построения архитектуры системы, которая должна быть разработана таким образом, чтобы обеспечить непрерывный и оптимальный доступ научного сообщества к вычислительными ресурсам. Что, в свою очередь, должно быть достигнуто за счет использования расширяемой многоуровневой архитектуры системы. В этом же разделе рассмотрены принципы построения системы, ее уровни, функции и взаимодействие компонент системы между собой, а также взаимодействие системы для обработки данных с внешними системами, такими как система управления данными или информационная система. На основе требований к архитектуре и анализа ее основных функций предложена архитектура системы для обработки данных в распределенной среде, способная обрабатывать данные эксабайтного диапазона. На рисунке 8 приведена архитектура и детальная схема взаимодействия компонент системы для глобальной распределенной обработки данных megaPanDA.

В **разделе 3.5** описывается методика управления потоками задач и заданий. Детально рассмотрена предложенная методика разделения вычислительного ресурса между различными классами заданий и подробно описаны предложенные методы управления потоками заданий. Изложена возможность организации распределенной обработки данных нового типа, таких как «обработка поездом» и «постоянная обработка».

Описан принципиально новый метод использования вычислительного ресурса, когда для выполнения заданий обработки или моделирования данных динамически формируется группа из нескольких (от одного до общего числа доступных центров) ВЦ — «всемирное облако», позволяющее эффективно использовать имеющийся вычислительный ресурс, в отличие от статических связок T1:nT2 в модели MONARC. Выбор центров основан на оценке стабильности их работы, определяемой по предложенной во второй главе методике, с учетом качества и загрузки WAN. «Всемирное облако» имеет основой сайт (сайт-ядро) и вспомогательные сайты (сайты-спутники).

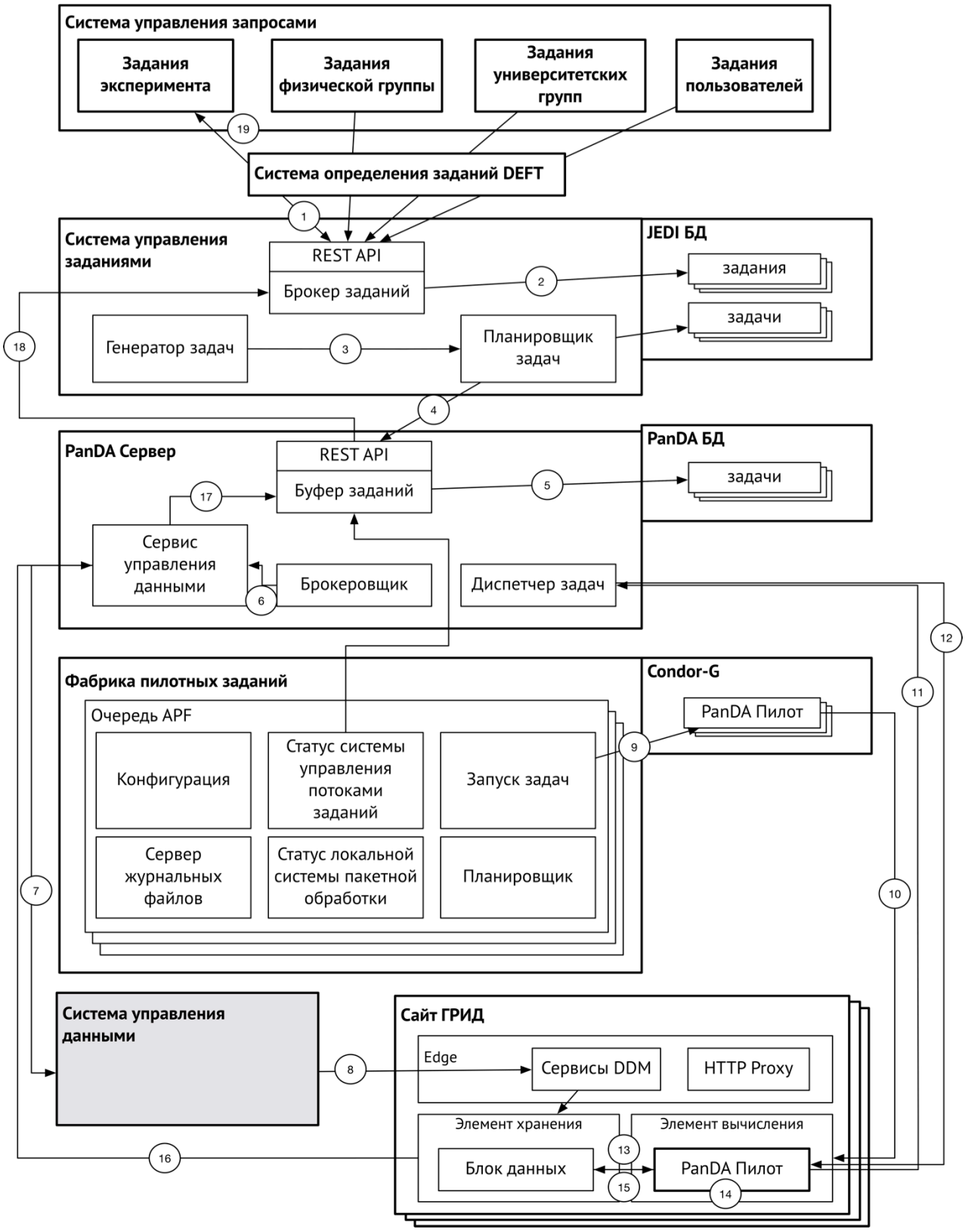

Рисунок 8. Архитектура системы для распределенной обработки данных megaPanDA

При выборе центрального сайта помимо характеристик, связанных с его стабильностью и аппаратными возможностями, рассчитываются следующие величины: выполняемая работа $RW = (nEvents - nEventsUsed) \cdot cpuTime$,

выполняемая работа для всех центров обработки: $totalRW = \sum RW(i = 0,...,n)$, а также вес каждого сайта (при расчете веса учитываются объемы данных на сайте — *sizeA*, общий объем данных — *sizeT*, свободное и общее дисковое пространство — соответственно *spaceF* и *spaceT* и наличие входных данных только на ленте (*TW*)):

$$weight = \frac{1}{totalRW} \cdot \frac{sizeA}{sizeT} \cdot TW \cdot \frac{spaceF}{spaceT}.$$

В результате выбирается сайт с наибольшим весом. На следующем этапе происходит выбор сайтов-спутников, формирующих вместе с сайтом-ядром «всемирное облако», используемое для выполнения задания. При выборе сайтов-спутников также учитываются как характеристики самих ВЦ, например, свободный дисковый и вычислительный ресурсы, количество запросов на передачу данных от/к данного сайта и пропускная способность между центральным ВЦ (ядро) и «спутником».

Раздел 3.6 посвящен проблеме распределения вычислительного ресурса между потоками различных заданий физического эксперимента. Введена иерархия из трех уровней, начиная с самого высшего («анализ» и «производство данных»), с последующим разделением «производства данных» на обработку и моделирование (уровень 2), что в конечном итоге позволяет разделять ресурсы для отдельных кампаний по обработке или моделированию данных (уровень 3). Описана модель разделения ресурса и сформулированы базовые определения модели, рассмотрена методика иерархического распределения долей. Применение метода динамического распределения вычислительного ресурса позволяет более эффективно разделять ресурс между потоками заданий и сокращает время ожидания для задач анализа данных на 40%.

Последний раздел главы (**раздел 3.7**) посвящен реализации системы управления потоками заданий эксперимента ATLAS на основе разработанных принципов и методов. Подробно рассмотрены характеристики созданной системы. Система не имеет мировых аналогов по производительности (1.4 Эбайта данных было обработано в 2016 г.) и масштабируемости (система выполняет до $2*10^6$ задач в день в более чем 250 центрах, см. рисунок 9). В разделе обоснованы принципы подсистемы мониторинга и ее компоненты.

Рассмотрены архитектура, функции и реализация подсистемы мониторинга и применение методов «машинного обучения» для предсказания времени выполнения заданий в распределенной вычислительной среде.

Рассмотрено, как подходы и создание системы для обработки и анализа данных в ФВЭ и ЯФ могут быть использованы для приложений биоинформатики.

Выводы к третьей главе обосновывают возможность дальнейшего развития компьютерной модели и перехода к гетерогенной вычислительной среде после создания системы для обработки данных нового поколения и подтверждают следующие защищаемые положения:

- новые принципы построения и архитектура глобальной системы для обработки данных в гетерогенной компьютерной среде, которые позволяют эффективно использовать вычислительные ресурсы и снимают противоречие по доступу к ресурсу между экспериментами, группами пользователей и отдельными учеными;

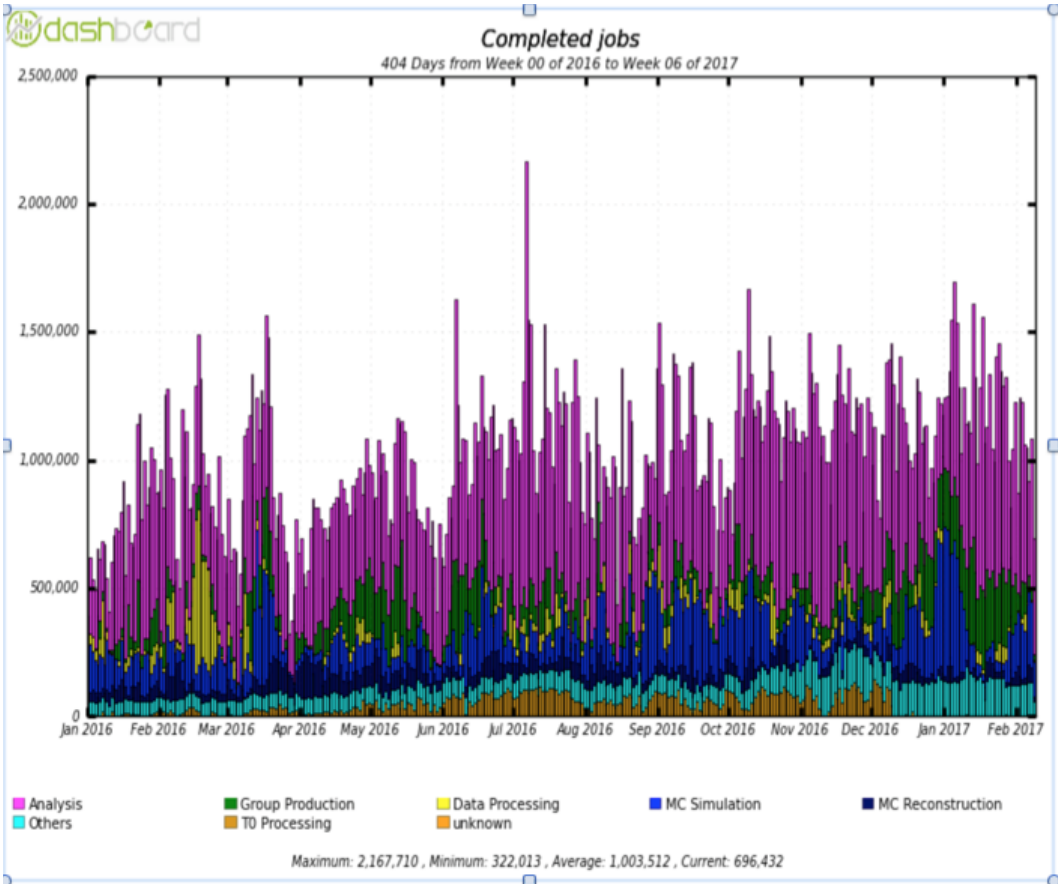

Рисунок 9. Количество задач для различных потоков заданий, выполненных в 2014–2017 гг.

- разработанный комплекс методик, методов и созданная на их основе система управления потоками заданий, повышают эффективность обработки данных экспериментов и обеспечивают обработку данных в эксабайтном диапазоне в масштабе более $2*10^6$ задач в день в более чем 200 ВЦ для более чем 1000 пользователей;
- подсистема мониторирования и оценки эффективности работы глобальной системы для обработки данных обеспечивает высокий уровень автоматизации при анализе работы системы и сбоев в работе распределенной вычислительной инфраструктуры и ее аппаратно-программных компонент.

Четвертая глава посвящена дальнейшему развитию компьютерной модели, интеграции суперкомпьютеров и ресурсов облачных вычислений с распределенными вычислительными ресурсами грид.

В главе обоснованы новые принципы и подходы к использованию вычислительного ресурса, в частности переход от прогнозирования мощности вычислительного ресурса при средних нагрузках к прогнозированию необходимой вычислительной мощности для пиковых нагрузок. Обосновано введение новых функций, необходимых в системе обработки данных. Фундаментальным вопросом для развития компьютерной модели в области физики частиц является следующий вопрос — как данные будут обрабатываться, анализироваться и моделироваться через 7–10 лет? В главе дается ответ на этот вопрос, а также обосновывается возможность использования созданной системы для обработки данных на этапе superLHC.

Рассмотрены требования к распределенным системам для обработки данных на последующих этапах работы LHC, а также требования, предъявляемые к таким системам на новых ускорительных комплексах (FAIR, NICA). Обосновано, что предложенная компьютерная модель и система для глобальной распределенной обработки данных могут быть применены за пределами экспериментов на LHC, в частности для коллайдера NICA и установки XFEL.

Обсуждаются принципы и архитектура при интеграции грид, «облачных» вычислительных ресурсов и суперкомпьютеров, рассмотрены конкретные примеры реализации и изменение компьютерной модели физических экспериментов.

Подробно обсуждаются проблемы и вызовы при интеграции и использовании суперкомпьютеров в созданной системе для обработки данных в приложениях физики частиц. Проанализированы возможные архитектурные решения для выполнения заданий ФВЭ и ЯФ в фоновом режиме на суперкомпьютерах, а также конкретная реализация подходов интеграции и разработанных методов в рамках работ, выполненных в НИЦ КИ, и в работах для экспериментов на LHC. На рисунке 10 приведена архитектура суперкомпьютерной компоненты созданной системы для распределенной обработки и анализа данных megaPanDA, описанной в главе 3.

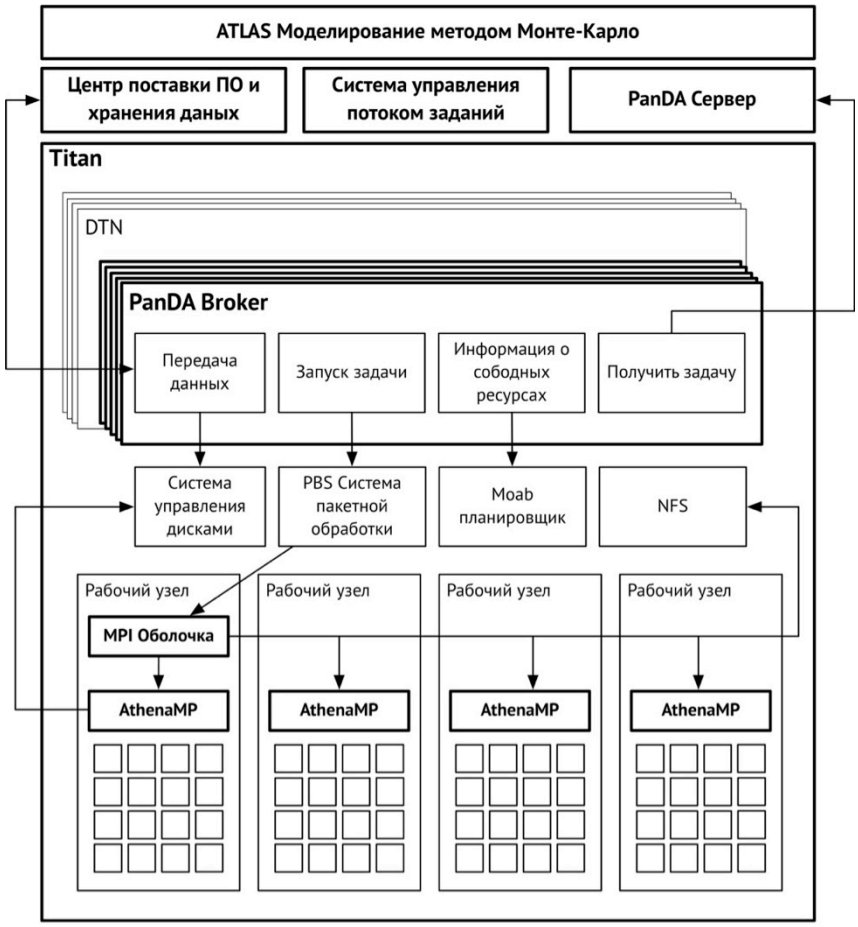

Рисунок 10. Архитектура системы управления заданиями для суперкомпьютера

На рисунке 11 представлен двумерный график, показывающий корреляцию между количеством свободных узлов и длительностью временного интервала, в течение которого они были свободны (измерения проводились для OLCF Titan в течение месяца, было сделано более 62.5 тысячи измерений).

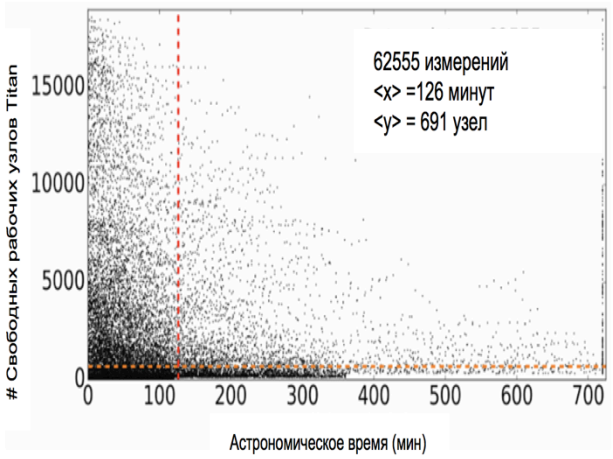

Рисунок 11. Корреляция количества свободных узлов суперкомпьютера Titan и времени, когда узлы были свободны

Из графика видно, что в среднем свободен 691 рабочий узел в течение 126 минут (красная и оранжевая линии на графике соответственно) и до 15 тысяч узлов в течение 30–100 минут. Это дает брокеру задач дополнительные возможности в их распределении по вычислительным ресурсам, а варьируя число генерируемых или моделируемых событий — «создавать» задачи различной длительности. Среднее время выполнения задач приведено на рисунке 12 (задание №9235668, ноябрь 2016).

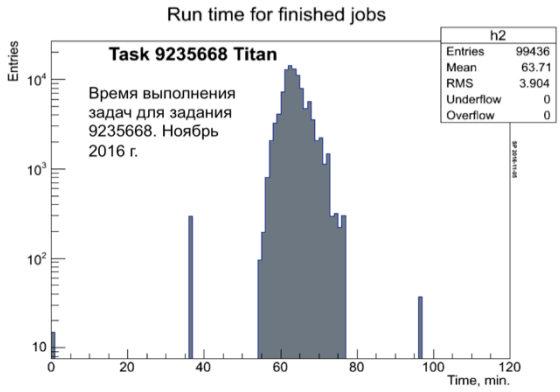

Рисунок 12. Среднее время выполнения задач моделирования эксперимента ATLAS на суперкомпьютере Titan

Из приведенного распределения времени выполнения 99.4 тысяч задач видно, что среднее время выполнения составляет около 64 минут. Таким образом, варьирование числа обрабатываемых/генерируемых событий, а значит и времени выполнения задания в зависимости от количества и длительности свободных узлов СК, позволит более эффективно использовать «свободные узлы» СК для выполнения задач. На рисунке 13 схематично представлена компьютерная модель, принятая после интеграции ресурсов грид с ресурсами суперкомпьютеров, облачными и университетских кластеров.

Далее в **подразделах 4.2.1 и 4.2.2** рассмотрено применение созданной системы для глобальной обработки данных (megaPanDA) приложений биоинформатики (решение задачи анализа данных геномного секвенирования древней ДНК шерстистого мамонта), а также использование суперкомпьютера НИЦ КИ для приложений экспериментов ATLAS и ALICE на LHC.

Заключительный раздел главы (**раздел 4.4)** посвящен рассмотрению архитектурных принципов и методов, а также анализу существующих технологий для создания федеративного дискового пространства в рамках гетерогенной киберинфраструктуры. В разделе проанализированы этапы развития российского грид сегмента WLCG (RDIG), обоснована мотивация создания единого дискового пространства, приведены требования, предъявляемые к такой системе, рассмотрен выбор возможной технологии и методы реализации федерации на примере российских центров.

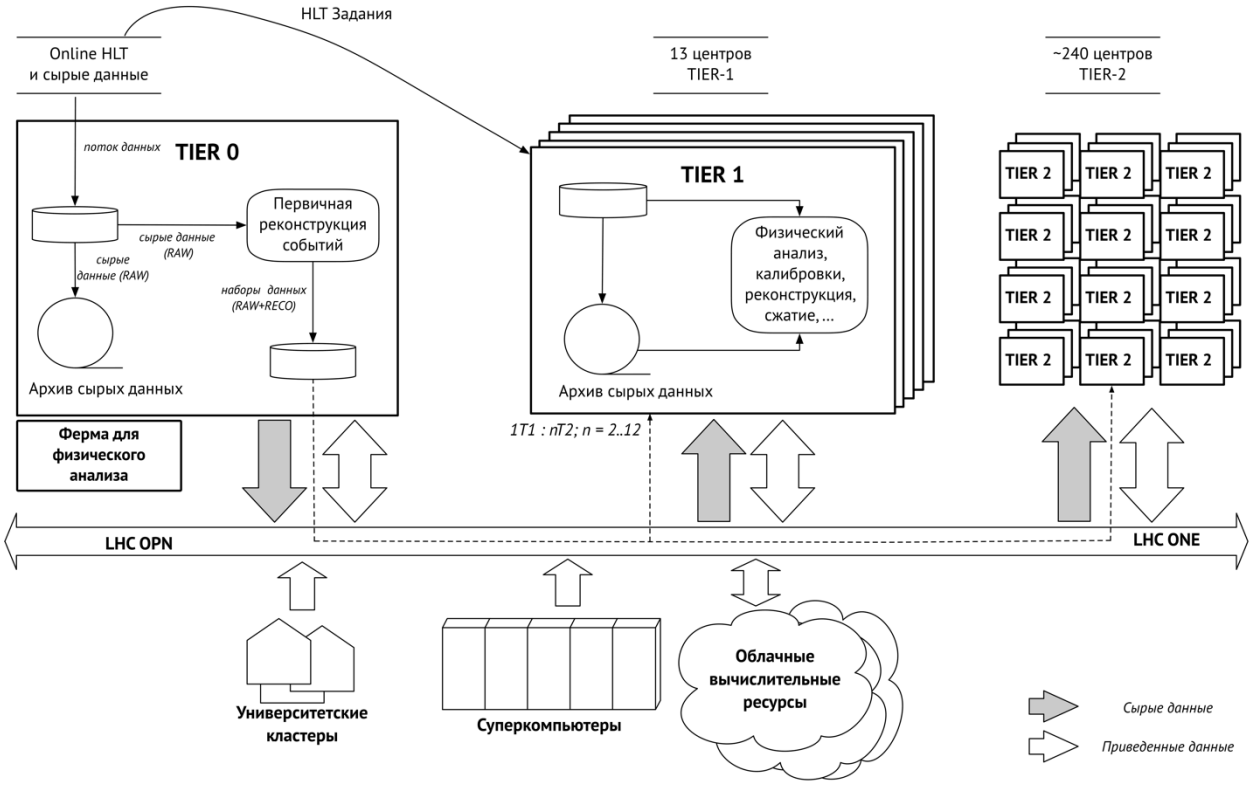

Рисунок 13. Новая компьютерная модель, реализованная для второго и последующих этапов работы LHC

Отдельно рассмотрена роль WAN при создании таких федераций. Рассмотрена модель и прототип федеративного дискового пространства в рамках RDIG (рис. 14). Рассмотрен потенциал данной работы, по сравнению с подобными исследованиями, ведущимися в ЦЕРН и других центрах, приведены примеры по проверке работоспособности федерации и характеристики выполнения реальных научных приложений ФВЭ и ЯФ. Для оценки эффективности и стабильности работы прототипа федерации использовались программы экспериментов ATLAS и ALICE.

Первый класс программ требовал значительного вычислительного ресурса при сравнительно небольших требованиях к чтению/записи данных (программы восстановления треков детектора переходного излучения в условиях высокой загрузки). Второй класс программ требовал высокой эффективности при чтении/записи данных (программы фильтрации событий на основе физической информации и создания наборов событий для последующего физического анализа). В данном разделе также обсуждаются возможные методы распределения данных внутри федерации для более эффективного доступа к ним из центров, входящих в федерацию, обсуждаются вопросы надежности и применимости данного решения для будущих экспериментов и приложений за пределами ФВЭ и ЯФ.

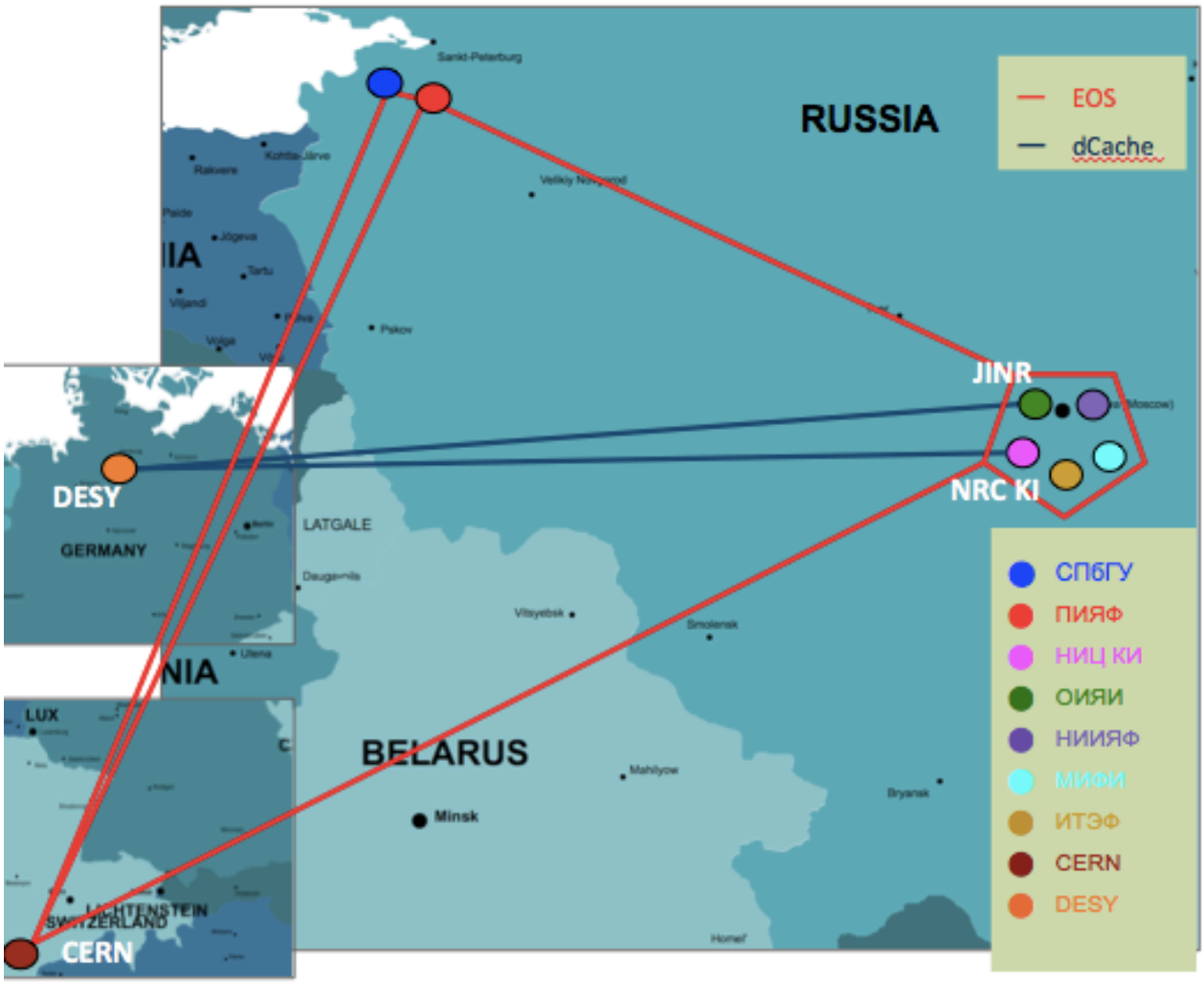

Рисунок 14. Прототип федерации в рамках российского сегмента грид RDIG

В заключении сформулированы основные выводы и результаты диссертации.

1. На основе методики, методов и архитектуры, разработанных в диссертации, создана глобальная распределенная система для обработки данных на основе динамического управления потоками заданий и динамическим распределением данных с учетом пропускной способности WAN. Реализация такой системы стала ключевым этапом для дальнейшего развития компьютерной модели и сделала возможным создание гетерогенной киберинфраструктуры, позволив использовать ресурсы суперкомпьютеров и ресурсы «облачных вычислений» наряду с существующей инфраструктурой грид, нивелировав архитектурные различия вычислительных мощностей. Таким образом, разнородные вычислительные ресурсы доступны пользователям в виде единой киберинфраструктуры. Созданная система имеет уникальные характеристики и позволяет:
    a. динамически организовывать группы ресурсов («всемирное облако») для выполнения научных заданий и динамически разделять вычислительный ресурс между различными классами заданий: Монте-Карло моделирование, обработка данных, анализ данных, потоки заданий отдельных научных групп;

   b. выполнять более $2*10^6$ задач в день в 250 ВЦ (1.4 Эбайта данных было обработано только в 2016 г.);
   c. использовать систему для приложений других научных областей, таких как астрофизика и биоинформатика, что подтверждает универсальность созданного ПО.
2. Исследованы, разработаны и реализованы базовые принципы подсистемы мониторирования, включая предсказание времени выполнения заданий, и аккаунтинга (учета работы отдельных центров и заданий).
3. Исследованы и реализованы методы определения популярности (востребованности) классов данных и отдельных наборов данных физического эксперимента.
4. Разработана методика управления научными приложения ФВЭ и ЯФ для суперкомпьютеров с использованием информации о временно свободных ресурсах, повышающая эффективность использования суперкомпьютеров (реализация методики для суперкомпьютера Титан позволила повысить эффективность использования с 89 до 94%).

Основные результаты, выводы, рекомендации и архитектурные решения, изложенные в диссертации, использовались при реализации следующих национальных и международных проектов:

- эксперименты ATLAS и ALICE на LHC в ЦЕРН (распределенная обработка, моделирование и анализ данных);
- эксперимент COMPASS на ускорителе SPS в ЦЕРН;
- проект развития глобальной грид-инфраструктуры для LHC (WLCG);
- эксперименты AMS и AMS-02 на МКС;
- проект «Federated data storage» WLCG;
- проект по развитию и применению методов «машинного обучения» для обнаружения аномалий в работе сложных распределенных систем и исследования их работы;
- проект «Создание Лаборатории Технологий больших данных для проектов в области мегасайенс» в рамках реализации постановления 220 Правительства РФ.
- проект Российского научного фонда «MetaMiner for BigData: Создание гетерогенной системы хранения метаинформации для научных экспериментов эксабайтного масштаба и применение методов “машинного обучения” для выявления нарушений при функционировании распределенных систем обработки и анализа Больших данных».
- проект Российского фонда фундаментальных исследований «Исследование гетерогенных киберинфраструктур, разработка и создание прототипа компьютерной федерации на основе высокоскоростных вычислений, облачных вычислений и суперкомпьютеров для хранения, обработки и анализа Больших данных».

Основные результаты данной работы являются пионерскими и используются в действующих научных экспериментах. Уже сейчас результаты диссертации используются в двух крупнейших экспериментах в области ФВЭ и ЯФ: ATLAS и

ALICE на LHC, эксперименте COMPASS на SPS, а также в приложениях биоинформатики на суперкомпьютерах НИЦ КИ.

Результаты диссертационной работы могут быть использованы при создании компьютерной модели для этапа высокой светимости LHC (этап superLHC), а также для новых комплексов, таких как FAIR, NICA, ErIC, и проектов класса мегасайенс: XFEL, LSST.

Другими перспективными направлениями данной работы являются созданная система для глобальной распределенной обработки данных и федерирование географически распределенных дисковых ресурсов. Первое направление может быть использовано как метауровень для планирования потоков заданий на современных суперкомпьютерах, что, как показано в диссертации, повышает эффективность использования суперкомпьтерных мощностей, второе направление интересно при рассмотрении эволюции российского сегмента грид (RDIG) и предоставляет широкие возможности по оптимизации его инфраструктуры и повышению надежности его работы.

По материалам диссертации подготовлены и читаются лекционные курсы в НИЯУ МИФИ, ТПУ, МФТИ. Подготовлена магистерская программа в ТПУ и Университете «Дубна».

Список основных публикаций автора по теме работы:

Работы из перечня ведущих рецензируемых научных изданий, рекомендованных для публикации основных результатов диссертаций:

1. А. Климентов, Р. Машинистов, А. Пойда. От PanDA до мамонта. Открытые системы, 2017, №3.
2. A. Klimentov, M. Grigorieva, A. Kiryanov, A. Zarochentsev. BigData and Computing Challenges in High Energy and Nuclear Physics, JINST Volume 12, June 2017, DOI:10.1088/1748-0221/12/06/C06044.
3. A. Klimentov et al. (ATLAS collaboration). Performance of the ATLAS Transition Radiation Tracker in Run 1 of the LHC: tracker properties. Feb 21, 2017. 45 pp. JINST 12 (2017) no.05, P05002 CERN-EP-2016-311, DOI: 10.1088/1748-0221/12/05/P05002.
4. K. De, S. Jha, A. Klimentov, T. Maeno, R. Mashinistov, P. Nilsson, A. Novikov, D. Oleynik, S. Panitkin, A. Poyda, K.F. Read, E. Ryabinkin, A. Teslyuk, V. Velikhov, J.C. Wells, and T. Wenaus. Integration of Panda Workload Management System with Supercomputers, ISSN 1547-4771, Physics of Particles and Nuclei Letters, 2016, Vol. 13, No. 5, c. 647–653.
5. В.Е. Велихов, А.А. Климентов, Р.Ю. Машинистов, А.А. Пойда, Е.А. Рябинкин. Интеграция гетерогенных вычислительных мощностей НИЦ «Курчатовский институт» для проведения масштабных научных вычислений. Известия ЮФУ. Технические науки. №11 (184), 2016. Тематический выпуск: СУПЕРКОМПЬЮТЕРНЫЕ ТЕХНОЛОГИИ (2016) 88-100 DOI:10.18522/2311-3103-2016-11-88100.
6. A. Aad, A. Klimentov et al. (ATLAS collaboration). “Search for charged Higgs bosons in the $H^{\pm} \rightarrow tb$ decay channel in pp collisions at $\sqrt{s}$=8 TeV using the ATLAS detector”. JHEP 1603 (2016) 127.
7. K. De, A. Klimentov, T. Maeno, P. Nilsson, T. Wenaus. Accelerating Science Impact through Big Data Workflow Management and Supercomputing, EPJ Web Conf. 108 (2016) 01003.

8. F.B. Megino, A. Klimentov et al. on behalf of ATLAS collaboration. PanDA: Exascale Federation of Resources for the ATLAS Experiment at the LHC, EPJ Web Conf. 108 (2016) 01001.
9. Аулов В.А., Климентов А.А., Машинистов Р.Ю., Недолужко А.В., Новиков А.М., Пойда А.А., Тертычный И.С., Теслюк А.Б., Шарко Ф.С. Интеграция гетерогенных вычислительных инфраструктур для анализа данных геномного секвенирования. Математическая биология и биоинформатика, 2016, т. 11, вып. 2, с. 205–213. doi: 10.17537/2016.11.205.
10. A. Zarochentsev, A. Kiryanov, A. Klimentov, D. Krasnopevtsev, P. Hristov. Federated data storage and management infrastructure. Journal of Physics: Conference Series, 2016, Vol. 762, Number 1.
11. A. Klimentov, K. De, S. Jha, T. Maeno, P. Nilsson, D. Oleynik, S. Panitkin, J. Wells, T. Wenaus. Integration of PanDA Workload Management System with Supercomputers for ATLAS and Data Intensive Science. Journal of Physics: Conference Series, 2016, Vol. 762, No. 1.
12. M.A. Grigorieva, M.V. Golosova, M.Y. Gubin, A.A. Klimentov, V.V. Osipova and E.A. Ryabinkin. Evaluating non-relational storage technology for HEP metadata and meta-data catalog. Journal of Physics: Conference Series, 2016, Vol. 762, No. 1.
13. D.V. Krasnopevtsev, A.A. Klimentov, R.Yu. Mashinistov, N.L. Belyaev, and E.A. Ryabinkin on behalf of the ATLAS collaboration, Study of ATLAS TRT Performance with GRID and Supercomputers, ISSN 1547-4771, Physics of Particles and Nuclei Letters, 2016, Vol. 13, No. 5, pp. 659–664.
14. G. Aad, A. Klimentov et al. (ATLAS collaboration). Searches for Higgs boson pair production in the $hh \rightarrow bb\tau\tau$, $\gamma\gamma WW^*$, $\gamma\gamma bb$, $bbbb$ channels with the ATLAS detector. Phys.Rev.D 92 (2015) 092004.
15. K. De, A. Klimentov, T. Maeno, T. Wenaus. The future of PanDA in ATLAS distributed computing,, J.Phys.Conf.Ser. 664 (2015) no.6, 062035.
16. K. De, A. Klimentov, D. Oleynik, S. Panitkin, A. Petrosyan, J. Schovancova, A.Vaniachine and T. Wenaus, Integration of PanDA workload management system with Titan supercomputer at OLCF, J. Phys.Conf.Ser. 664 (2015) no.9, 092020.
17. M. Borodin, K. De, J. Garcia Navarro, D. Golubkov, A. Klimentov, T. Maeno and A. Vaniachine, Scaling up ATLAS production system for the LHC Run 2 and beyond : project ProdSys2. J.Phys.Conf.Ser. 664 (2015) no.6, 062005.
18. M.V. Golosova, M.A. Grigorieva, A.A. Klimentov, E.A. Ryabinkin, G. Dimitrov and M. Potekhin, Studies of Big Data metadata segmentation between relational and non-relational databases, J.Phys.Conf.Ser. 664 (2015) no.4, 042023.
19. М. Григорьева, М. Голосова, А. Климентов, Е. Рябинкин. Научное хранилище метаданных для экзабайтных экспериментов. Открытые Системы, 2015, №4. ISSN 1028-7493.
20. A. Klimentov, P. Buncic, K. De, T. Maeno, T. Wenaus. Next Generation Workload Management System For Big Data on Heterogeneous Distributed Computing, J. Phys.Conf. Ser. 608 (2015) no.1, 012040.
21. M. Borodin, D. Golubkov, A. Klimentov, A. Vaniachine. Multilevel Workflow System in the ATLAS Experiment, J.Phys.Conf.Ser. 608 (2015) no.1, 2015.
22. G. Aad, A. Klimentov et al. (ATLAS collaboration). Search for a Charged Higgs Boson Produced in the Vector-Boson Fusion Mode with Decay $H(\pm)\rightarrow W(\pm)Z$ using pp Collisions at $\sqrt{s}=8$ TeV with the ATLAS Experiment. Phys Rev Lett. 2015 Jun 12;114(23):231801. Epub 2015 Jun 9.
23. A. Belyaev, A. Berezhnaya, L. Betev, P. Buncic, K. De, D. Drizhuk, A. Klimentov, Y. Lazin, I. Lyalin, R. Mashinistov, A. Novikov, D. Oleynik, A. Polyakov, A. Poyda, E. Ryabinkin, A. Teslyuk, I. Tkachenko and L. Yasnopolskiy. Integration of Russian Tier-1

Grid Center with High Performance Computers at NRC-KI for LHC experiments and beyond HENP. Journal of Physics: Conference Series, Volume 664, Article Number: 092018, 2015, DOI: 10.1088/1742-6596/664/9/092018.
24. V. Aulov, K. De, D. Drizhuk, A. Klimentov, D. Krasnopevtsev, R. Mashinistov, A. Novikov, D. Oleynik, A. Poyda, E. Ryabinkin, I. Tertychnyy, A. Teslyuk. Workload Management Portal for High Energy Physics Applications and Compute Intensive Science. Procedia Computer Science, Volume 66, стр. 564-573, doi: 10.1016/j.procs.2015.11.064.
25. Fernando Barreiro Megino, Kaushik De, Jose Caballero, John Hover, Alexei Klimentov, Tadashi Maeno, Paul Nilsson, Danila Oleynik, Sergey Padolski, Sergey Panitkin, Artem Petrosyan, Torre Wenaus, on behalf of the ATLAS collaboration. PanDA: Evolution and Recent Trends in LHC Computing. Procedia Computer Science, Volume 66, стр. 439-447.
26. K. De, A. Klimentov, T. Maeno, T. Wenaus. Evolution of the ATLAS PanDA workload management system for exascale computational science. J.Phys.Conf.Ser. 513 (2014) 032062. DOI:10.1088/1742-6596/513/3/032062.
27. K. De, D. Golubkov, A. Klimentov, M. Potekhin and A. Vaniachine. Task Management in the New ATLAS Production System, 10.1088/1742-6596/513/2/032078. J.Phys.Conf.Ser 513 (2014) 032078.
28. G. Aad, A. Klimentov et al. (ATLAS collaboration). Search for supersymmetry in events with large missing transverse momentum, jets, and at least one tau lepton in 20 fb(1) of root s=8 TeV proton-proton collision data with the ATLAS detector. JOURNAL OF HIGH ENERGY PHYSICS, Issue: 9 Article Number: 103 Published: SEP 18 2014, ISSN 1029-8479.
29. A. Klimentov et al. (ATLAS collaboration). `Triggers for displaced decays of long-lived neutral particles in the ATLAS detector. JINST 8 (2013) P07015.
30. G. Aad, A. Klimentov et al. (ATLAS collaboration). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC, Physics Letters B, 716, 2012, p. 1–29.
31. D. Golubkov, B. Kersevan, A. Klimentov, A. Minaenko, P. Nevski, A. Vaniachine et al. ATLAS Grid Data Processing: system evolution and scalability, Journal of Physics: 2012, Vol. 396, 032049.
32. A. Klimentov, P. Nevski, M. Potekhin and T. Wenaus, The ATLAS PanDA monitoring system and its evolution, J. Phys. Conf. Ser. 331 (2011) 072058.
33. G. Aad, A. Klimentov et al. (ATLAS collaboration), Performance of the ATLAS Trigger System in 2010, Eur.Phys.J. C 72 (2012) 1849.
34. M. Titov, G. Zaruba, A. Klimentov, and K. De, “A probabilistic analysis of data popularity in ATLAS data caching,” Journal of Physics: Conference Series, vol. 396, no. 3, 2012.
35. A. Anisenkov, A. Klimentov, R. Kuskov and T. Wenaus, ATLAS Grid information system, J.Phys. Conf. Ser. 331 (2011) 072002.
36. J. Catmore, A. Klimentov et al. ATLAS data re-processing. 17th International Conference on Computing in High Energy and Nuclear Physics (CHEP09). March 2009. Journal of Physics volume 219, 2010.
37. A. Klimentov et al. PanDA. Production and Analysis backend. 17th International Conference on Computing in High Energy and Nuclear Physics (CHEP09). March 2009. Journal of Physics volume 219, 2010.
38. A. Klimentov et al. (ATLAS collaboration). The ATLAS Simulation Infrastructure.2010. 53 pp. Eur.Phys.J. C70 (2010) 823-874 DOI: 10.1140/epjc/s10052-010-1429-9.
39. A. Klimentov, M. Titov et al. ATLAS Data Transfer Request Package (DaTRI) J. Phys.: Conf. Series. Proc. 18th Int. Conf. on Computing in High Energy and Nuclear Physics (CHEP2010).
40. A. Klimentov et al. (ATLAS collaboration). The ATLAS Experiment at the CERN Large Hadron Collider, JINST 3 S08003 DOI: 10.1088/1748-0221/3/08/S08003 (2008).

41. A. Klimentov et al. (ATLAS collaboration). Expected Performance of the ATLAS Experiment - Detector, Trigger and Physics. arXiv:0901.0512 [hep-ex]. 2008.
42. A. Klimentov et al. (AMS Collaboration). The Alpha Magnetic Spectrometer (AMS) on the International Space Station, Part I, Results from the test flight on the Space Shuttle, Physics reports, vol.366/6, 331-404.
43. A. Klimentov, V. Choutko, M. Pohl. Computing Strategy of Alpha-Magnetic Spectrometer Experiment, NIM (2003) 502.
44. A. Klimentov et al. (L3 collaboration). Results from the L3 experiment at LEP. Feb 1993. 202 pp. Phys.Rept. 236 (1993) 1-146 CERN-PPE-93-31 DOI: 10.1016/0370-1573(93)90027-B.
45. A. Klimentov et al. (L3 collaboration). The Construction of the L3 Experiment, Nucl. Instrum. Methods A 289 (1990) 35.
46. A. Klimentov et al. Hadron calorimetry in the L3 detector. Nucl. Instrum. Methods A302 (1990) 53.
47. S. Burov, Yu. Galaktionov, A. Klimentov et al. A test and calibration setup for mass produced proportional chambers, preprint CERN, EP/88-84 (1988).

Публикации в других научных изданиях:

1. Alessio Angius, Danila Oleynik, Sergey Panitkin, Matteo Turilli, Kaushik De, Alexei Klimentov, Sarp H. Oral, Jack C. Wells, Shantenu Jha. Converging High-Throughput and High-Performance Computing: A Case Study. arXiv:1704.00978 [cs.DC].
2. Alessio Angius, Danila Oleynik, Sergey Panitkin, Matteo Turilli, Kaushik De, Alexei Klimentov, Sarp H. Oral, Jack C. Wells, Shantenu Jha. High-Throughput Computing on High Performance Platforms : A Case Study; published in: e-Science (e-Science), 2017 IEEE 13th International Conference proceedings, DOI: 10.1109/eScience.2017.43.
3. А.А. Климентов, В.В. Кореньков, В.Е. Велихов. Интеграция параллельных и распределенных вычислений для решения масштабных задач . Сборник трудов Московского пятого суперкомпьютерного форума, Октябрь 29/30 2015.
4. M. Borodin, K. De, J. Garcia Navarro, D. Golubkov, A. Klimentov. Unified System for Processing Real and Simulated Data in the ATLAS Experiment. Proceedings of the XVII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2015), Obninsk, Russia, October 13–16, 2015.
5. A. Klimentov et al. ATLAS Distributed Computing. International Symposium on Nuclear Electronics and Computing. Varna, Bulgaria, Sep. 2009.
6. А.А. Климентов, Р.Ю. Машинистов, А.М. Новиков, А.А. Пойда, И.С. Тертычный. Комплексная система управления данными и задачами в гетерогенной компьютерной среде. Труды XVII международной конференции DAMDID/RCDL'2015, "Аналитика и управление данными в областях с интенсивным использованием данных", Обнинск, Россия, октябрь 13/16, 2015,
7. K. De, A. Klimentov, T. Maeno and T. Wenaus. PanDA: A New Paradigm for Distributed Computing in HEP Through the Lens of ATLAS and other Experiments. 37я Международная конференция по Физике Высоких Энергий, Валенсия, Испания, 2014.
8. K. De, A. Klimentov, J. Schovancova, T. Wenaus. The new Generation of the ATLAS PanDA Monitoring System, PoS ISGC2014 (2014) 035.
9. А. Климентов. К вопросу о федеративной организации распределенной ЦЕРН. Суперкомпьютеры. 2015, 20, с. 26–28.
10. K. De, A. Klimentov, P. Nilsson, S. Panitkin, Extending ATLAS Computing to Commercial Clouds and Supercomputers, By ATLAS Collaboration, PoS ISGC2014 (2014) 034.
11. А. Ваняшин, А. Климентов, В. Кореньков. За большими данными следит PanDA. Суперкомпьютеры, 2013, №3 (11), с. 56–61.
12. I. Fisk, M. Girone, A. Klimentov. The Common Analysis Framework Project. Труды международной конференции Computing in High Energy and Nuclear Physics, 2013.

13. А. Климентов, В. Кореньков. Распределенные вычислительные системы и их роль в открытии новой частицы . Суперкомпьютеры, 2012, №3 (11), с. 7–11.
14. S. Campana, A. DiGirolamo, J. Elmsheuser, S. Jezequel, A. Klimentov, J. Schovancova, C. Serfon, G. Stewart, D. van der Ster, I. Ueda and A. Vaniachine. ATLAS Distributed Computing Operations : Experience and improvements after 2 full years of data-taking. May 2012, 19th International Conference on Computing in High Energy and Nuclear Physics (CHEP12). May 2012. New-York, USA.
15. J. Andreeva, D. Benjamin, S. Campana, A. Klimentov, V. Korenkov, D. Oleynik, S. Panitkin, A. Petrosyan. Tier-3 Monitoring Software Suite (T3MON) proposal. Препринт ЦЕРН, ATL-SOFT-PUB-2011-001, 2011, p. 7.
16. G. Negri, J. Shank, D. Barberis, K. Bos, A. Klimentov and M. Lamanna. Distributed computing in ATLAS, PoS ACAT08 (2008) 035.
17. S. Albraid, D. Costanzo, F. Giannotti, A. Klimentov et al. Metadata for ATLAS, препринт ЦЕРН ATL-COM-GEN-2007-001. (2007).
18. A. Klimentov, et al. AMS-02 Computing and Ground Data Handling. Computing in High Energy Physics Conference Proceedings, Sep 2004, Interlaken, Switzerland.
19. S. Bracci, S. Falciano, A. Klimentov, C. Luci et al. The Upgrade of the L3 Third Level Trigger for High Luminosity Runs at LEP. Proceedings of the Fourth International Conference on Advanced Technology and Particle Physics, Como, Italy, 3-7 October 1994.
20. S. Burov, A. Klimentov, V. Koutsenko et al. The distributed DAQ system of hadron calorimeter prototype. Preprint ITEP-181 (1989).
21. С. Буров, А. Климентов, В. Куценко и др. Тестовая система для исследования пропорциональных камер. Препринт ИТЭФ 104 (1985).
22. С. Буров, А. Климентов, В. Куценко и др. Система сбора и анализа данных прототипа адронного калориметра установки Л3. Препринт ИТЭФ 155 (1985).

Список литературы

1. LHC – The Large Hadron Collider, http://lhc.web.cern.ch/lhc
2. The ATLAS Collaboration, G. Aad et al. The ATLAS Experiment at the CERN Large Hadron Collider // Journal of Instrumentation, Vol. 3, S08003, 2008.
3. The CMS Collaboration, S. Chatrchyan et al. The CMS experiment at the CERN LHC // Journal of Instrumentation, Vol. 3, S08004, 2008.
4. ALICE Collaboration, K. Aamond et al. The ALICE experiment at the CERN LHC, JINST 3 (2008) S08002.
5. H.H. Gutbrod et al. (Eds.) FAIR Baseline Technical Report, ISBN 3-9811298-0-6. Nov. 2006.
6. M. Altarelli et al. (Eds). “XFEL: The European X-ray Free-Electron Laser Technical Design Report”, DESY 2006-097 (DESY, 2007).
7. G.V. Trubnikov et al. “Project of the Nuclotron-based Ion Collider Facility (NICA) at JINR”, Proceedings of EPAC 08 (Genoa, 2008), pp. 2581–2583.
8. The ATLAS Collaboration, G. Aad, A. Klimentov et al “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”, Physics Letters B, 716, 2012, pp 1–29.
9. J. Ratchford, U. Colombo, “Megascience,” UNESCO World Science Report, 1996.
10. “О Стратегии научно-технологического развития Российской Федерации”, Указ Президента Российской Федерации от 01.12.2016 г. № 641.
11. Э. Таненбаум, М. ван Стен. Распределенные системы. Принципы и парадигмы. СПб.: Питер, 2003. С. 876.

12. В.В. Воеводин, Вл.В. Воеводин. Параллельные вычисления. СПб.: БХВ-Петербург, 2002, С. 608.
13. Н.Н. Говорун. Некоторые вопросы применения электронных вычислительных машин в физических исследованиях. Автореферат диссертации на соискание ученой степени доктора физико-математических наук, ОИЯИ, 10-4437. Дубна, 1969.
14. WLCG: Worldwide LHC Computing Grid, http://wlcg.web.cern.ch
15. I. Foster, K. Kesselman. The Grid: a Blueprint to the New Computing Infrastructure Morgan Kaufman Publishers, 1999, p. 690.1.
16. В.В. Кореньков. Методология развития научного информационно – вычислительного комплекса в составе глобальной грид-инфраструктуры. Диссертация на соискания ученой степени доктора технических наук. Дубна, 2012.
17. LHCOPN: LHC Optical Private Network, http://wlcg.web.cern.ch
18. LHCONE: LHC Open Network Environment, http://lhcone.cern.ch
19. А. Климентов, В. Кореньков. Распределенные вычислительные системы и их роль в открытии новой частицы // Суперкомпьютеры, 2012, №3 (11), с. 7–11.
20. А. Ваняшин, А. Климентов, В. Кореньков. “За большими данными следит PanDA” // Суперкомпьютеры, 2013, №3 (11), pp. 56–61.
21. А. Климентов. К вопросу о федеративной организации распределенной ЦЕРН // Суперкомпьютеры, 2015, №20, с. 26–28.
22. http://toolkit.globus.org/toolkit/about.html
23. В. Бедняков, В. Кореньков. Перспективы Грид-технологий в промышленности и бизнесе // Знание–сила, 2010, №10, с. 97–103.
24. V. Ilyin, V. Korenkov, A. Soldatov: RDIG (Russian Data Intensive Grid) e- Infrastructure. Proc. of XXI Int. Symposium of Nuclear Electronics&Computing ((NEC`2007, Varna, Bulgaria), ISBN 5-9530-0171-1, Dubna, 2008, p. 233–238.
25. V. Ilyin, V. Korenkov, A. Kryukov, Yu. Ryabov, A. Soldatov: Russian Date intensive Grid (RDIG): current status and perspectives towards national Grid initiative. Proc. of Int. Conf. "Distributed computing and Grid-Technologies in Science and Education, GRID-2008", ISBN 978-5-9530-0198-4, Dubna, 2008, p.100-108.
26. В.Н. Коваленко, Д.А. Корягин. Распределенный компьютинг и грид. // Технологии грид. Т.1, — М.: ИПМ им. М.В. Келдыша, 2006. — С. 7–28.
27. A.P. Afanasiev, S.V. Emelyanov, Y.R. Grinberg, V.E. Krivtsov, B.V. Peltsverger, O.V. Sukhoroslov, R.G. Taylor, V.V. Voloshinov: Distributed Computing and Its Applications. Felicity Press, Bristol, USA, 2005, 298 p.
28. А.П. Афанасьев, В.В. Волошинов, С.В. Рогов, О.В. Сухорослов. Развитие концепции распределенных вычислительных сред. Проблемы вычислений в распределенной среде // Сб. трудов ИСА РАН / Под ред. С.В. Емельянова, А.П. Афанасьева. — М.: Эдиториал УРСС, 2004.
29. Вл.В. Воеводин, С.А. Жуматий. Вычислительное дело и кластерные системы. — М.: Изд-во МГУ, 2007, 150 с.
30. Вл.В. Воеводин. Top500: числом или уменьем? // Открытые системы, 2005, №10, с. 12–15.
31. A. Klimentov et al. PanDA. Production and Analysis backend. 17th International Conference on Computing in High Energy and Nuclear Physics (CHEP09). March 2009. Journal of Physics volume 219, 2010.
32. M. Aderholz et al. Models of Networked Analysis at Regional Centers for LHC Experiments (MONARC) - Phase 2 Report CERN/LCB, 2000-001 http://monarc.web.cern.ch/MONARC.
33. A. Klimentov, M. Pohl. “AMS-02 Computing and Ground Data Handling”, Computing in High Energy Physics Conference Proceedings, Sep 2004, Interlaken, Switzerland.(2000).
34. S. Campana, A. DiGirolamo, J. Elmsheuser, S. Jezequel, A. Klimentov, J. Schovancova, C. Serfon, G. Stewart, D. van der Ster, I. Ueda and A. Vaniachine, “ATLAS Distributed

Computing Operations : Experience and improvements after 2 full years of data-taking", May 2012, 19th International Conference on Computing in High Energy and Nuclear Physics (CHEP12). May 2012.
35. A. Klimentov et al. "Extending ATLAS Computing to Commercial Clouds and Supercomputers", PoS ISGC2014 (2014) 034.
36. A. Zarochentsev, A. Kiryanov, A. Klimentov, D. Krasnopevtsev and P. Hristov, "Federated data storage and management infrastructure", Journal of Physics: Conference Series, Volume 762, Number 1.
37. Load Sharing Facility. https://www-03.ibm.com/systems/spectrum-computing/products/lsf/index.html
38. Portable Batch System. http://www.pbspro.org/
39. HTCondor. Official site : https://research.cs.wisc.edu/htcondor/
40. S. Bagnaso, L. Betev, P. Buncic et al., "The ALICE Workload Management System: Status before the real data taking", Journal of Physics: Conference Series 219 (2010) 062004.
41. S.K. Paterson and A. Tsaregorodtsev, "DIRAC optimized workload management", Journal of Physics: Conference Series. Volume 119 part 6 (2008).
42. А. Ваняшин, А. Климентов, В. Кореньков. За большими данными следит PanDA // Суперкомпьютеры. 2013, №3 (11), с. 56–61.
43. A. Klimentov, P. Buncic, K. De, T. Wenaus et al. "Next Generation Workload Management System For Big Data on Heterogeneous Distributed Computing", J. Phys.Conf. Ser. 608 (2015) no.1, 012040.