Отзыв официального оппонента

на докторскую диссертацию Климентова Алексея Анатольевича «Методы обработки сверхбольших объемов данных в распределенной гетерогенной компьютерной среде для приложений в ядерной физике и физике высоких энергий», представленной на соискание ученой степени доктора физико-математических наук по специальности 05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Диссертационная работа А.А. Климентова посвящена современной и актуальной проблеме - обработке сверхбольших объемов данных (мультипетабайтный и эксабайтный диапазон) в ядерной физике (ЯФ) и физике высоких энергий (ФВЭ). Исследования в этих областях науки невозможны без использования вычислительных систем, а также программного обеспечения для обработки и анализа данных. Развитие экспериментальной программы в области физики элементарных частиц потребовало разработки новых подходов, методов, компьютерной модели и программного обеспечения для решения этих задач. Требования к информационным технологиям естественным образом следуют из необходимости управлять огромными объемами информации, возникающими в экспериментах на современных ускорителях. Задачи обработки сверхбольших объемов данных стали особенно актуальны во втором десятилетии нашего века после запуска Большого адронного коллайдера (БАК) и начала обработки данных для таких экспериментов как ATLAS, ALICE, CMS. Было необходимо предложить принципиально новые подходы для обработки данных и новую компьютерную модель (модель обработки данных) для экспериментов в области ФВЭ и ЯФ. Диссертация А.А. Климентова является несомненно актуальной работой в данном направлении, посвященной созданию глобальной системы для распределенной обработки и анализа сверхбольших объемов данных на основе динамического управления потоками заданий и динамическим распределением компьютерных ресурсов. Реализация подобной системы стала ключевым этапом для создания и развития гетерогенной вычислительной среды, обеспечив возможность использовать суперкомпьютеры, университетские кластеры и ресурсы «облачных вычислений», наряду с созданной на первом этапе работы БАК инфраструктурой грид, нивелировав архитектурные различия вычислительных ресурсов для конечного пользователя и физических экспериментов.

Диссертация состоит из введения, четырех глав, заключения, списка литературы и списка используемых сокращений.

Во **введении** автор обсуждает актуальность обработки сверхбольших объемов данных для приложений в области физики элементарных частиц с использованием вычислительных мощностей в распределенной гетерогенной компьютерной среде. Автор формулирует предмет и цели исследования, обосновывает научную новизну и практическую значимость полученных результатов.

Первая глава посвящена развитию компьютерной модели экспериментов в области физики высоких энергий, ядерной физики и астрофизики (на примере

экспериментов на ускорителях и коллайдерах, а также эксперимента AMS на международной космической станции). Рассматривается, как менялись требования к информационным технологиям и программному обеспечению по мере развития ускорительных и астрофизических экспериментов. Важным выводом является, что за последние 15-20 лет возросли не только объемы данных, но и произошло качественное изменение в составе научных коллабораций. Так в международное сообщество ATLAS входит более 3000 ученых из десятков стран мира, что накладывает дополнительные требования на создание вычислительной инфраструктуры и организацию процесса хранения, обработки и анализа данных. В данной главе автор обосновывает необходимость эволюции иерархической модели компьютинга, реализованного на первом этапе работы БАК и перехода к «смешанной модели». В этой же главе приводится описание предложенных методов определения значимости и популярности классов и наборов данных. Обосновывается создание термодинамической модели управления данными в распределенной компьютерной среде и реализация предложенной автором модели для сверхбольших объемов данных.

Во **второй главе** автор обосновывает требования к вычислительной инфраструктуре для обработки, моделирования и анализа данных современных физических экспериментов, а также рассматривается роль суперкомпьютеров для научных приложений в области физики высоких энергий и ядерной физики. Одним из основных выводов данной главы является вывод о расширении созданной системы грид за счет других архитектур, таких как суперкомпьютеры, ресурсы «облачных вычислений», университетские кластеры, что в свою очередь ведет к созданию новой компьютерной модели современного физического эксперимента. Принципиально отметить, что переход от однородной вычислительной инфраструктуры, основанной на центрах высопропопускных вычислений (грид) к гетерогенной вычислительной инфраструктуре был предложен автором диссертации. В последующих главах (глава 4) показано, что это в свою очередь позволило перейти к подходу использования дополнительных вычислительных ресурсов в момент пиковых нагрузок, что было принципиально невозможно при реализации иерархической и однородной компьютерной модели.

Третья глава диссертации посвящена разработке концепции и архитектуры системы управления заданиями в распределенной гетерогенной компьютерной среде. Автор проводит анализ классов научных приложений в современных экспериментах в области ФВЭ и ЯФ и предлагает модель данных для системы распределенной обработки. В данной главе автор обосновывает, что реализация новой модели обработки данных в неоднородной компьютерной среде возможна лишь при создании единой системы для обработки и анализа данных, которая позволит нивелировать архитектурные различия для конечного пользователя. Автор описывает созданную систему обработки данных на примере эксперимента ATLAS и ее масшабируемость и функциональность. Система выполняет более 2 миллионов заданий в день в более чем 250 ВЦ по весму миру. Завершающая часть главы содержит описание системы мониторирования.

В **главе четыре** автор рассматривает дальнейшее развитие компьютерной модели, в том числе для этапа высокой светимости на БАК и для новых установок (NICA, ОИЯИ, Россия; FAIR, GSI, Германия). Обосновывается рост роли

суперкомпьютеров для моделирования и обработки данных современных физических экспериментов, а также применение созданной системы обработки данных для решения других научных задач, на примере задач по секвенированию генома мамонта, выполненных на суперкомпьютере в НИЦ "Курчатовский институт". Особый интерес представляет описание созданного прототипа распределенной федеративной инфраструктуры на базе вычислительных центров НИЦ КИ, ОИЯИ, СПбГУ и других Российских университетов, что может существенно изменить ландшафт компьютерных центров для обработки данных ФВЭ и ЯФ.

В **заключении** диссертации сформулированы основные полученные результаты.

Научная новизна диссертации заключается в разработке методов, моделей и программных средств, позволивших обрабатывать данные эксабайтного диапазона в распределенной неоднородной компьютерной среде. Реализация, разработанных автором, компьютерной модели и системы для обработки данных позволили впервые использовать одновременно различные архитектуры для приложений ФВЭ и ЯФ, и создать уникальную по своим характеристикам систему управления потоками заданий. Следует отметить, что начальной мотивацией для данного исследования явились потребности экспериментов ФВЭ и ЯФ, но созданные средства и методы во многом универсальны, и могут быть востребованы в других областях, требующих распределенной обработки данных.

Диссертация и ее результаты, безусловно, соответствует квалификации доктора физико-математических наук по специальности 05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей.

Теоретическое значение работы состоит в решении новых оригинальных задач в области программного обеспечения и информационных технологий для научных приложений в области физики элементарных частиц. Практическая значимость, полученных результатов, состоит в создании глобальной системы для распределенной обработки данных, используемой в экспериментах в области ядерной физики и физики высоких энергий.

В диссертации обобщены результаты 20-летней работы автора. Основные результаты диссертации опубликованы в российских и международных журналах (более 150 печатных работ).

Достоверность проведенных исследований обоснована их практическим применением для ряда физических экспериментов и подтверждается апробацией результатов на семинарах и рабочих совещаниях экспериментов L3, AMS, ATLAS, ALICE, COMPASS, а также в докладах, представленных на российских и международных конференциях.
По диссертационной работе могут быть сделаны следующие замечания :

1. Не хватает сравнения отдельных компонент разработанной системы по хранению данных и управлению заданиями с известными открытыми решениями работы с большими данными (Apache Spark, Apache Ignite, Slurm).

2. Не приведены данные, позволяющие судить об эффективности масштабируемости полученной распределенной системы для обработки данных, что несколько нивелируются приведенными абсолютными цифрами количества обработанных заданий.

Приведенные замечания ни в коем случае не влияют на общую положительную оценку работы.

Автореферат правильно и полно отражает содержание диссертации.

Считаю, что рассматриваемая диссертация является законченным научным исследованием, в котором получены новые фундаментальные результаты, и, безусловно, отвечает критериям «Положения о присуждении ученых степеней», предъявляемым к докторским диссертациям, а сам автор, Климентов Алексей Анатольевич, несомненно, заслуживает присуждения ему ученой степени доктора физико-математических наук по специальности 05.13.11 — математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей.

Официальный оппонент

Доктор физико-математических наук,

член-корреспондент РАН, директор Федерального

государственного бюджетного учреждения науки

Институт системного программирования

им. В.П.Иванникова

А.И.Аветисян

«26» февраля 2018 г.

109004 г. Москва, ул. Александра Солженицына, д. 25

Телефон : +7 (495) 912-46-14,

Адрес электронной почты : arut@ispras.ru