ОБЪЕДИНЁННЫЙ ИНСТИТУТ ЯДЕРНЫХ ИССЛЕДОВАНИЙ
ЛАБОРАТОРИЯ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

На правах рукописи

Нечаевский Андрей Васильевич

Методы и средства моделирования распределенных систем хранения и обработки данных на основе результатов их мониторинга

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

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Дубна – 2021

Работа выполнена в Лаборатории информационных технологий Объединенного института ядерных исследований.

Научный руководитель: *Ососков Геннадий Алексеевич, профессор, доктор физико-математических наук, главный научный сотрудник ЛИТ ОИЯИ*

Официальные оппоненты: *Дегтярев Александр Борисович, доктор технических наук, профессор кафедры компьютерного моделирования и многопроцессорных систем, факультет прикладной математики - процессов управления, Федеральное государственное бюджетное образовательное учреждение высшего образования «Санкт-Петербургский государственный университет»*

Царегородцев Андрей Юрьевич, кандидат физико-математических наук, Институт Ядерной Физики и Физики Частиц Национального Центра Научных Исследований (Марсель, Франция), инженер-исследователь высшей квалификации

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

Ученый секретарь диссертационного
совета ОИЯИ.05.01.2019.П,
доктор физ.-мат. наук Е. В. Земляная

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. В настоящее время в России создаются современные научные установки, которые генерируют большие объемы экспериментальных данных. Например, интенсивность потока после запуска ускорительного комплекса NICA (ОИЯИ, г. Дубна) будет составлять десятки петабайт в год. Современные системы хранения и обработки данных – это сложные распределенные программно-аппаратные комплексы, построенные с применением грид и облачных технологий, требующие определенного режима работы, меняющегося как при увеличении объемов поступающих данных, так и при изменении качества и состава оборудования. Прежде, чем приступить к созданию распределенной IT-инфраструктуры, необходимо решить, какова будет ее архитектура в зависимости от стоимостных факторов и интенсивности потоков данных. С этой целью необходимо выполнить моделирование рассматриваемой вычислительной структуры, чтобы на полученной динамической модели, учитывающей реальную специфику системы и поступающих потоков данных, выбрать оптимальную архитектуру. Как показал проведенный анализ доступных аналитических методов моделирования, в силу ограниченных теоретических предпосылок они не могут быть применены для моделирования сложных компьютерных комплексов многоуровневой архитектуры с реальными распределениями входных потоков заданий, сложной многоприоритетной дисциплиной их обслуживания и динамическим распределением. Системы хранения и обработки данных являются сложными и многокомпонентными установками, включающими кластеры, а также узлы, реализованные в облачной архитектуре, при их создании и изменении необходимо использовать имитационное моделирование. Под имитационной моделью понимается универсальное средство исследования сложных систем, представляющее собой логико-алгоритмическое описание поведения отдельных элементов системы и правил их взаимодействия, отображающих последовательность событий, возникающих в моделируемой системе.

Имитационное моделирование грид и облачных систем позволяет обнаружить узкие места в архитектуре центров обработки данных, проводить эксперименты с изменением топологии и заменой ресурсов для проверки предлагаемых решений без непосредственного вмешательства в функционирование вычислительного центра, тестировать алгоритмы управления задачами и распределения ресурсов по группам пользователей. Зачастую моделирование применяют только на этапе проектирования грид и облачных систем, однако, эксперименты продолжаются годами и десятилетиями, при этом

объемы обрабатываемой информации имеют тенденции роста, поэтому одновременно с эксплуатацией системы происходит ее развитие, не только качественное, но и количественное. Очевидно, что для достижения оптимальных результатов моделирование должно носить постоянный характер на протяжении всего жизненного цикла экспериментов. В настоящее время процессы моделирования и мониторинга рассматриваются, как независимые задачи, не связанные между собой. Чтобы повысить точность получаемых результатов, необходимо в качестве входных данных для моделирования использовать статистику, накопленную во время работы подобных вычислительных инфраструктур. Для этого требуется разработка программных средств, объединяющих процессы моделирования и мониторинга систем хранения и обработки данных больших научных экспериментов.

Цели и задачи исследования

Создание методов и средств для моделирования распределенных систем хранения и обработки данных с учетом результатов их мониторинга.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Исследование различных подходов к моделированию распределенных систем хранения и обработки данных больших объемов;
2. Разработка подхода для моделирования систем хранения и обработки данных с использованием результатов мониторинга в качестве входных параметров для динамической коррекции параметров модели;
3. Создание программного инструментария для моделирования систем хранения и обработки данных, реализующего идею синтеза процессов мониторинга и моделирования;
4. Применение разработанного программного инструментария для конкретных прикладных задач.

Методы исследования

Для решения поставленных задач в диссертации использованы методы теории вероятностей и математической статистики, теории системного анализа и проектирования сложных систем, методы Монте-Карло.

Научная новизна диссертационной работы

1. Разработан оригинальный подход для моделирования вычислительных систем распределенной обработки, основанный на учете данных мониторинга, используемых для корректировки параметров модели, что выгодно отличает новый подход от других известных программ

моделирования, поскольку позволяет оперативно учитывать изменения как архитектуры, так и динамику загрузки системы.

2. Разработан программный инструментарий, реализующий предложенные методы и алгоритмы, позволяющий провести предварительные исследования по анализу и выбору вариантов инфраструктур с учетом их загрузки и определить наиболее эффективное решение для построения IT-инфраструктуры обработки и хранения данных.
3. Впервые получены данные моделирования для вычислительных центров экспериментов BM@N и MPD ускорительного комплекса NICA, а также вычислительного центра ИФВЭ[1] в Пекине. Результаты моделирования позволили проверить различные варианты организации инфраструктуры и дать рекомендации по составу оборудования.

Положения, выносимые на защиту

1. Новый подход к моделированию систем хранения и обработки данных на основе использования результатов мониторинга для динамической коррекции параметров модели.
2. Программный инструментарий для моделирования систем хранения и обработки данных, реализующий синтез процессов моделирования и мониторинга, позволяющий провести предварительные исследования различных вариантов организации IT-инфраструктуры, оценить возможности существующей архитектуры при решении задач хранения и обработки данных и дать рекомендации по ее оптимизации.
3. Научные результаты по прогнозированию IT-инфраструктуры экспериментов BM@N и MPD ускорительного комплекса NICA, производительности MPI-приложений на облачной инфраструктуре ЛИТ, а также инфраструктуры вычислительного центра ИФВЭ в Пекине, полученные в результате моделирования и обеспечившие получение информации, необходимой для оптимизации режима работы и повышения эффективности указанных систем.

[1] Научно-исследовательский институт физики высоких энергий китайской академии наук

Практическая значимость

1. Разработано программное обеспечение для моделирования распределенных вычислительных центров - комплекс SyMSim[2] (Свидетельство о регистрации программы для ЭВМ №2017618100 "Программный комплекс для моделирования распределенных систем хранения и обработки данных на основе результатов их мониторинга", дата регистрации 21.07.2017).
2. Предложенный в исследовании подход позволяет провести исследования IT-инфраструктуры для принятия управленческих решений по ее модернизации с целью сохранения скорости получения результатов экспериментов при постоянном повышении потока данных, а также сократить срок создания технического проекта. Это достигается за счет обоснованного вычислительным экспериментом прогноза уровня загрузки оборудования и уточнения необходимого запаса ресурсов.
3. Разработанный программный инструментарий был использован для моделирования вычислительных центров экспериментов BM@N и MPD ускорительного комплекса NICA, в результате которого определены основные характеристики этих систем и даны практические рекомендации по выбору оборудования. Проведенное моделирование вычислительного центра ИФВЭ в Пекине позволило дать рекомендации по развитию центра.

Достоверность представленных в диссертации результатов, подтверждается сопоставлением реальных статистических данных и результатов, полученных с использованием имитационной модели. Все представленные результаты прошли апробацию в научных изданиях, на международных конференциях. На программный инструментарий получено свидетельство о регистрации авторских прав.

Апробация работы

Результаты работы неоднократно докладывались и обсуждались на семинарах ЛИТ ОИЯИ, а также на российских и международных научных конференциях:

– The International Conference «Distributed computing and Grid technologies in science and education» 2014, 2016 (Dubna, Russia);

[2] Synthesis of Monitoring and Simulation - синтез мониторинга и моделирования

- XVII международная конференция DAMDID/RCDL “Аналитика и управление данными в областях с интенсивным использованием данных“, 2015 (Обнинск, Россия);
- The International Conference on Mathematical Modeling and Computational Physics, 2015 (Stara Lesna, Slovakia);
- The 4th International Young Scientists Conference and Summer School, 2015 (Athens, Greece);
- 44th meeting of the PAC for Particle Physics, JINR, 2015 (Russia, Dubna, poster session)
- RO-LCG 2016 Conference “Grid, Cloud, and High-Performance Computing in Science”, Magurele, Romania, 26-28 October, 2016
- 23d International Conference on Computing in High Energy and Nuclear Physics, CHEP2018, Sofia, Bulgaria, 2018.

Диссертационное исследование выполнялось в рамках Проблемно-тематического плана научно-исследовательских работ и международного сотрудничества ОИЯИ по теме № 05-6-1118-2014/2023 “Информационно-вычислительная инфраструктура ОИЯИ”, а также при поддержке:

- гранта РФФИ в рамках научного проекта № 14-07-00215 “Разработка средств планирования построения и развития систем хранения и обработки больших объёмов данных на основе синтеза процессов моделирования и мониторинга” (руководитель Ососков Г.А.);

- гранта РФФИ в рамках научного проекта № 15-29-01217 “Разработка программно-аппаратного комплекса для численных исследований джозефсоновских наноструктур на базе облачного центра ЛИТ ОИЯИ с использованием параллельных вычислений” (руководитель Ососков Г.А.);

- грантов ОИЯИ для молодых ученых и специалистов (№14-603-02, №15-603-03) и стипендии имени Говоруна (ЛИТ ОИЯИ).

В 2015 году результаты, представленные в диссертационном исследовании, удостоены премии Губернатора Московской области в сфере науки и инноваций для молодых ученых.

В диссертационной работе использованы результаты двух магистерских работ, где автор был консультантом:

1. *Пряхина Д.И.* "Моделирование процессов управления распределенными данными для крупных проектов", 2015
2. *Айриян В.С.* "Создание и исследование имитационной модели облачной инфраструктуры на примере LIT Cloud", 2017

Личный вклад

Содержание диссертации, а также основные результаты и положения, выносимые на защиту, отражают персональный вклад автора.

Подготовка к публикациям полученных результатов проводилась совместно с соавторами, при этом вклад соискателя был определяющим. Все представленные в диссертации результаты получены лично автором, либо в соавторстве при определяющем вкладе соискателя.

Соответствие диссертации паспорту специальности

Данная диссертационная работа соответствует формуле специальности, поскольку содержит в себе разработку программного инструментария для моделирования процессов хранения и обработки данных в сложных многокомпонентных системах, которыми являются распределенные IT-инфраструктуры.

Проблематика диссертации соответствует областям исследований:

- пункт 3 формулы специальности - модели, методы, алгоритмы, языки и программные инструменты для организации взаимодействия программ и программных систем;
- пункт 4 формулы специальности - системы управления базами данных и знаний;
- пункт 9 формулы специальности - модели, методы, алгоритмы и программная инфраструктура для организации глобально распределенной обработки данных.

Публикация результатов

По результатам исследований, составивших основу диссертации, опубликовано 17 работ, выполненных в течение 2013-2019 гг. в соответствии с требованиями к публикациям Положения о присуждении ученых степеней в ОИЯИ (пр. ОИЯИ от 30.04.2019 № 320). Получено 1 свидетельство о регистрации программного продукта.

Структура и объем работы

Диссертация состоит из 3 глав, введения, заключения и приложения, содержит 134 страницы, включает 34 рисунка, 21 таблицу и библиографию из 71 наименования.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность выбранной темы, определены цель и задачи, решаемые в работе. Отражена практическая ценность полученных результатов и приведены сведения об апробации результатов диссертационной работы. Изложена структура диссертационной работы.

В первой главе дается анализ существующих средств имитационного моделирования грид и облачных систем, указаны причины неэффективности их использования для решения поставленных задач моделирования систем хранения и обработки данных крупных научных экспериментов.

В силу своей сложности и высокой стоимости разработка систем сбора, передачи и распределенной обработки больших объемов данных (десятки петабайт) требует предварительных исследований по выбору их оптимальной структуры с учетом стоимости предполагаемых ресурсов и загрузки. Подобные исследования должны основываться на тщательном моделировании как потока заданий с учетом их типов и статистических данных о распределении времени их поступления и требуемых компьютерных ресурсов для их выполнения, так и состава моделируемой IT-инфраструктуры.

Существуют различные программные инструменты имитационного моделирования грид и облачных вычислительных систем. На момент проведения исследования большинство этих систем моделирования были рассчитаны на решение своих узкоспециализированных задач и не обладали набором функций для полноценного моделирования грид и облачных вычислительных центров для обработки данных физических экспериментов (Таблица 1).

Таблица 1. Сравнение систем моделирования

Функция	GridSim	OptorSim	SimGrid	CloudSim	iCanCloud
Репликация данных	Да	Да	Нет	Да	Нет
Планировщик задач	Да	Нет	Да	Да	Да
Резервирование ЦПУ	Да	Нет	Нет	Да	Да
Генерация фонового сетевого трафика	Да	Да	Да	Да	Да
Графический интерфейс	Нет	Да	Нет	Нет	Да
Моделирование гибридных архитектур	Нет	Нет	Нет	Нет	Нет
Моделирование ленточного робота	Нет	Нет	Нет	Нет	Нет
Использование данных мониторинга	Нет	Нет	Нет	Нет	Нет

На основании анализа, проведенного в диссертации, делается вывод, что систем моделирования грид и облачных вычислений, учитывающих данные мониторинга, не существовало.

В диссертации предложен подход к моделированию распределенных вычислительных систем, основанный на учете данных мониторинга, используемых для динамической коррекции параметров модели (рис. 1): задания через систему управления нагрузкой (1) поступают на обработку в вычислительную систему (2), информация о статусе выполнения заданий поступает в БД (3).

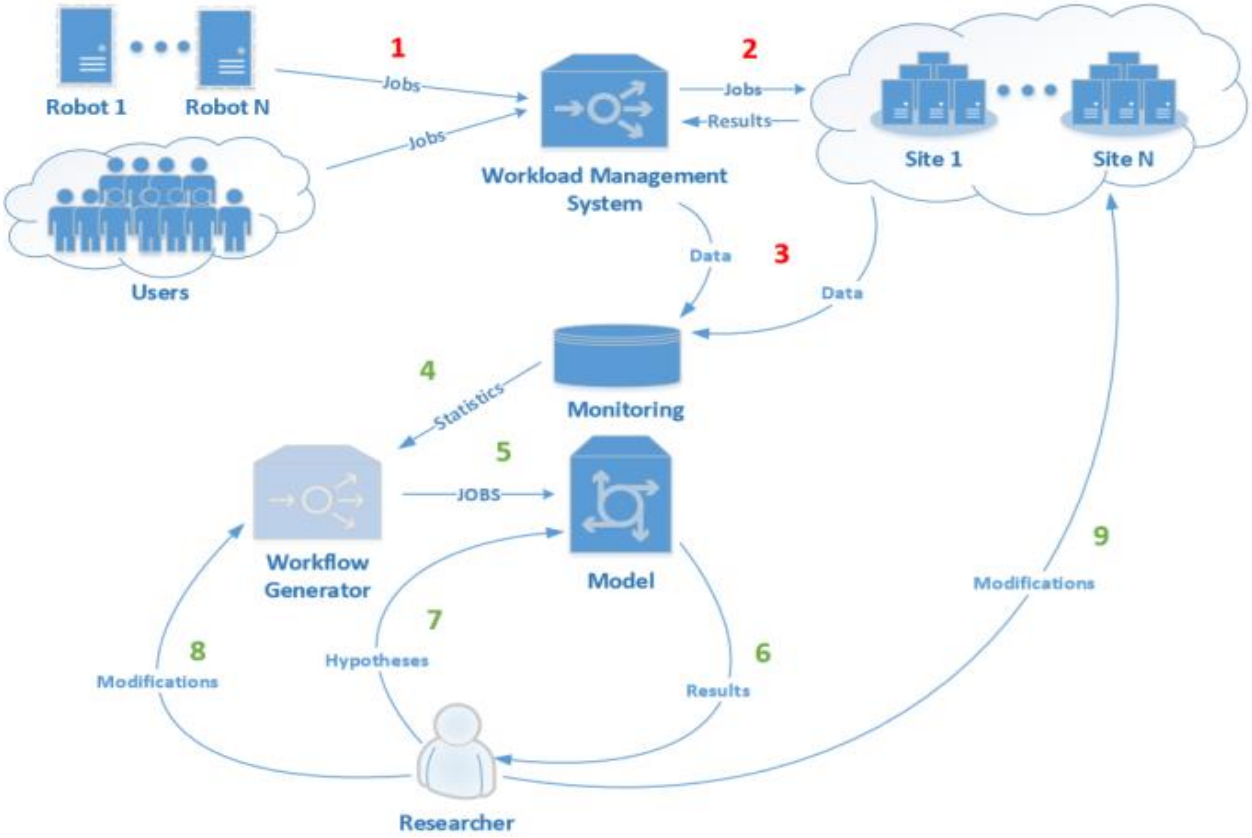


Рис.1. Моделирование распределенной системы с учетом данных мониторинга

Статистические данные мониторинга используются в качестве входного потока для имитационной модели. Также на базе статистических данных о задачах может быть сгенерирован новый поток входных данных для модели (4, 5). Исследователь получает результаты моделирования и анализирует их (6), далее он может изменить параметры и проверять новые гипотезы (7,8). Результаты моделирования могут быть использованы для инициализации процедуры изменения конфигурации ресурсного центра с целью улучшения его характеристик (9).

В результате проведенного анализа существующих систем моделирования грид и облачных вычислительных систем обоснована необходимость объединения процессов моделирования и мониторинга в рамках одного программного продукта для проектирования и развития систем хранения и обработки данных больших научных экспериментов. Для решения этой задачи была сформулирована базовая концепция моделирования:

1. Целью моделирования современного вычислительного центра является удовлетворение некоторого критерия оптимальности, минимизирующего стоимость оборудования при безусловном выполнении SLA (Service Level Agreement);
2. Лучший способ динамически оценить качество работы системы - использовать средства мониторинга;
3. Программа моделирования должна быть совмещена с реальной системой мониторинга распределенной системы обработки данных через базу данных (БД);
4. Целесообразно принять двоякую структуру модели, которая состоит из ядра - его стабильной основной части, независимой от моделируемого объекта, и декларативного модуля для ввода параметров модели, определяющих конкретный распределенный вычислительный центр (его настройки и параметры, полученные из информации мониторинга, как поток данных, поток заданий и т.д.);
5. БД содержит описание инфраструктуры, каждого ее узла, связей между ними, информацию о запущенных заданиях, временах исполнения, результаты мониторинга работы различных подсистем, а также результаты моделирования;
6. Веб-портал необходим для связи имитационной модели и БД, выбора параметров моделирования и сохранения результатов в БД.

Для реализации системы моделирования требовалось описать основные объекты и события грид и облачных систем. Также требовалось разработать структуру базы данных, которая будет содержать описание IT-инфраструктуры, каждого ее узла, связей между ними, данные мониторинга работы различных подсистем и результаты моделирования. Кроме этого необходимо разработать инструментарий, позволяющий автоматически генерировать входные параметры модели, реализовать интерфейсы для редактирования параметров модели системы обработки данных и отображения полученных результатов.

Во второй главе описана реализация программного инструментария для моделирования систем хранения и обработки данных SyMSim, выполняющего синтез процессов моделирования и мониторинга.

Разработанное программное обеспечение SyMSim позволяет моделировать обработку потока заданий IT-инфраструктурой, обладающей заданными ресурсами и правилами их резервирования и использования. Базовый функционал имитационной модели реализован путем доработки и расширения классов системы моделирования GridSim. Для решения поставленных задач разработаны дополнительные классы, объекты, генераторы, интерфейсы и модули. Объекты реализуются в виде java-классов, при этом пользователь задает только внешнее описание их количества и связей между ними, но не изменяет текст программы. Схема работы программного комплекса представлена на рисунке 2.

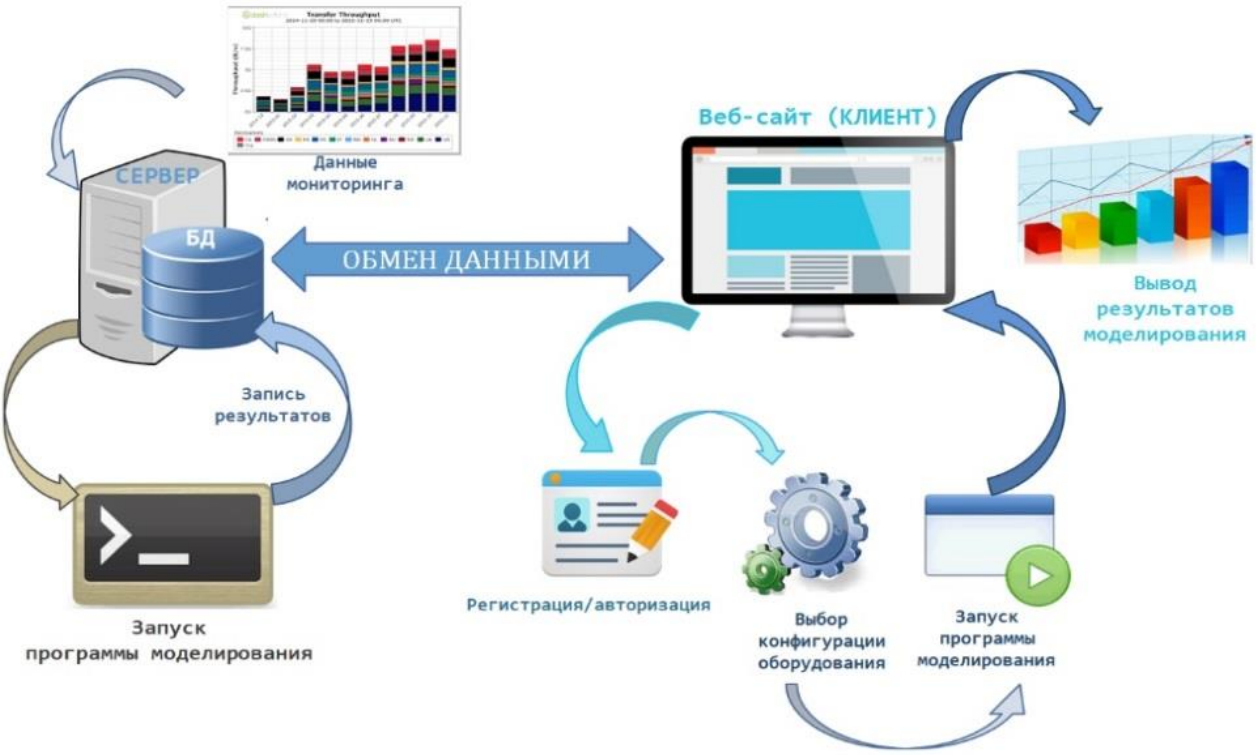


Рис.2. Схема работы разработанного программного инструментария

Процесс имитационного моделирования состоит в прохождении набора заданий через заданную IT-инфраструктуру. Объектами модели являются задачи, процессоры, файлы, ленты, дисковые хранилища, линии передачи данных, роботизированные библиотеки. Список событий, происходящих с объектами, включает: поступление задачи в очередь, извлечение задачи из очереди, занятие или освобождение вычислительного ресурса, передачу файла, извлечение ленты из хранилища, монтирование, чтение – запись на ленту и др. Набор реализованных классов позволяет имитировать все процессы, происходящие в системе.

Опыт мировых физических экспериментальных центров показывает, что для долговременного и архивного хранения больших объемов данных производимых детекторами физических экспериментов наиболее целесообразно использование роботизированных ленточных библиотек. Для моделирования роботизированной библиотеки были разработаны классы, представленные на рисунке 3. Набор этих классов позволяет моделировать процессы, происходящие с копией файла на лентах: загрузку и выгрузку ленты манипулятором, монтирование на драйве, поиск файла на ленте и его чтение/запись.

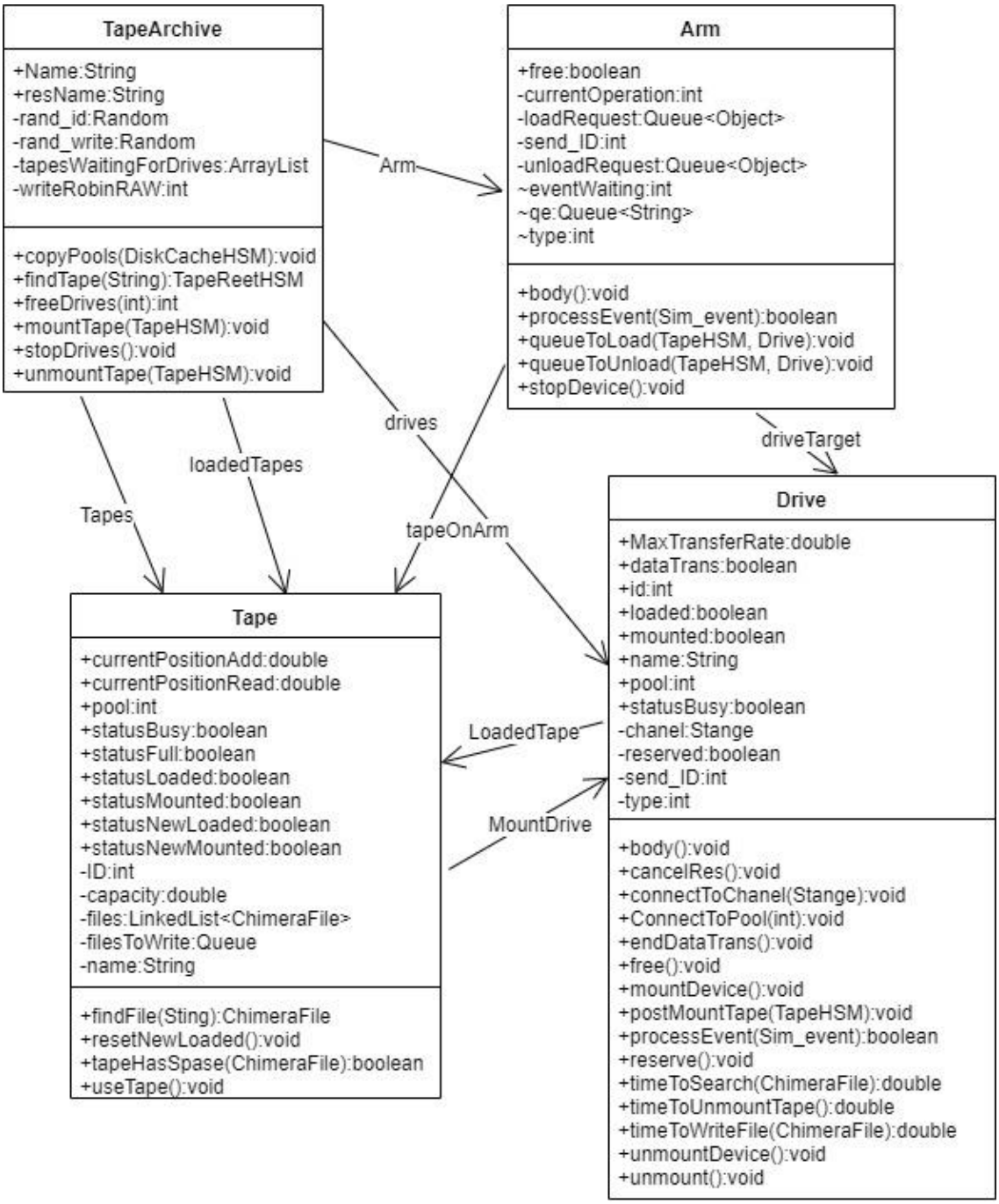


Рис. 3. Реализация классов роботизированной библиотеки

Процесс моделирования инфраструктуры с роботизированной библиотекой заключается в следующем. Файлы должны быть записаны на ленты библиотеки. Параллельно этому процессу выполняются задания обработки. Каждое задание требует единственный файл, копия которого может находиться на дисковом массиве, или на ленте. Задание начинает выполняться, если есть свободный слот и все файлы доступны на дисковом хранилище. Если файл хранится в роботизированной библиотеке, задание резервирует слот, но выполнение задерживается до момента его загрузки на диск. Процесс перемещения файла из библиотеки в дисковое хранилище включает в себя операцию помещения ленточного картриджа (tape) на драйв (drive), которую выполняет рука робота (arm), монтирования файловой системы картриджа на драйве и записи файла на диск. Копии файлов на дисковых накопителях могут стираться сборщиком мусора, если к ним долгое время нет обращений. Такой алгоритм работы с файлами соответствует алгоритму, реализованному в системе dCache (www.dcache.org), имеющей широкое распространение в центрах обработки данных.

Входной поток заданий для моделирования формируется через БД. Для этого реализована возможность получения и хранения данных мониторинга работы вычислительного центра, чтобы использовать их в качестве входных данных для моделирования. При отсутствии необходимой статистики по вычислительному центру или моделированию нового последовательность выполнения вычислительного эксперимента будет следующая: анализируется статистика заданий по аналогичным вычислительным центрам; строятся распределения заданий по времени исполнения и размеру входного файла; далее выдвигается и проверяется гипотеза о количестве типов заданий в потоке.

Генерация потока заданий осуществляется следующим образом (рис. 4):

— задаются необходимые параметры потока заданий: период; список сайтов-источников; количество файлов в системе; максимальное и минимальное количество задач в день для каждого из трех типов (моделирование, реконструкция, анализ); максимальное и минимальное процессорное время, количество событий и объем памяти);

— для каждого периода случайно определяется количество задач;

— вычисляется количество процессорного времени, памяти и событий для задачи, время старта, а также генерируются дополнительные параметры.

Рис. 4. Веб-интерфейс генератора потока заданий

Случайные последовательности с различными дискретными и непрерывными распределениями, имитирующие вышеуказанные случайные величины, а также такие параметры работы системы как, например, количество заявок в очереди, время ожидания загрузки, длительность выполнения задачи генерируются по стандартным алгоритмам моделирования случайных величин с соответствующими распределениями с использованием генератора случайных чисел с равномерным распределением в интервале (0,1), основанным на «Вихре Мерсенна»[3].

Сгенерированный таким образом поток заданий сохраняется в БД. После чего пользователь задает необходимую конфигурацию системы хранения и запускает программу моделирования.

Запуск программы моделирования осуществляется на сервере. Параметры и результаты моделирования сохраняются в БД после каждого запуска программы, чтобы предоставить возможность пользователю сравнить результаты работы систем с использованием различных конфигураций оборудования. Результаты моделирования доступны пользователю в виде диаграмм и графиков, что упрощает дальнейший анализ.

Результатом работы модели является найденная величина времени обработки потока заданий для разных вариантов структуры вычислительной установки и производительности отдельных ее частей, что позволяет оценить, как эти факторы влияют на время обработки. Также пользователь получает

[3] https://www.boost.org/doc/libs/1_75_0/boost/random/mersenne_twister.hpp

данные о нагрузке на ресурсы системы (CPU, RAM, дисковый буфер), время ожидания задач в очереди, пропускную способность сети, данные по нагрузке на робота ленточной библиотеки. Кроме того в результате моделирования можно уточнить какими резервами обладает вычислительная установка, то есть какова верхняя граница интенсивности потока задач (данных) и какие компоненты установки оказывают на нее существенное влияние.

В третьей главе приведены результаты моделирования систем, служащие доказательством успешной верификации предложенных методов. В качестве апробации разработанного программного обеспечения приведены результаты решения конкретных прикладных задач.

Для проверки адекватности модели проведен вычислительный эксперимент по моделированию файловой загрузки в центрах распределенных вычислений с двухуровневой структурой типа Tier-0 - Tier-1 (рис. 5). Показана возможность моделирования вариантов эксплуатации IT-инфраструктуры для экспериментов физики высоких энергий. В ходе исследований создана модель системы обработки данных эксперимента NICA-MPD с возможностью предсказания измерений параметров на основе собранной в ходе мониторинга статистики поступающих потоков данных и использования ресурсов системы.

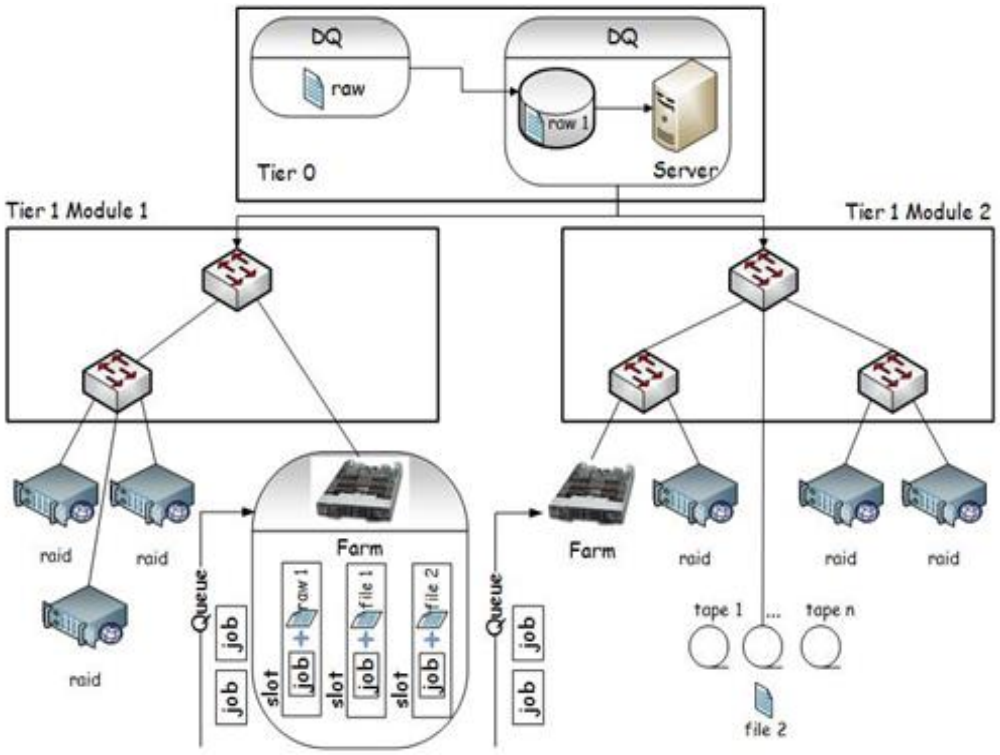


Рис. 5. Схема моделируемой структуры Tier0-Tier1

Моделируемая структура состоит из вычислительной фермы, дискового массива – пула, каналов связи, ленточного робота и лент. Характеристики взяты

из системы мониторинга CMS Tier1 и TDR DAQ MPD. Пропускная способность каналов связи от 10 до 100 ГБ/с. Tier0 обозначает центр сбора данных из планируемого эксперимента MPD-NICA. Полученные данные должны храниться на дисках. Одна из задач – получение рекомендации по выбору объема дискового хранилища. Исходными данными для начала моделирования являются параметры оборудования, статистика о потоках данных и статистика о потоках задач.

На рис.6 представлены данные о реальных и смоделированных характеристиках загрузки центра Tier1. Показаны данные мониторинга за один месяц (06.2015) и модельные данные. Сравнение реальных и смоделированных данных по завершенным задачам выполнено по сопоставимости средних значений с учетом коридора среднеквадратичного разброса значений. Результаты сравнения, показанные на рис. 6, показали сопоставимые характеристики.

Для проверки адекватности модели проведены различные вычислительные эксперименты по моделированию файловой загрузки в центрах распределенных вычислений. Показана возможность моделирования нагрузки на дисковый буфер в зависимости от скорости поступления данных. Все экспериментальные данные должны быть переданы в центр хранения и обработки. В случае проекта NICA-MPD это петабайты данных за один месяц работы детектора. Моделирование позволяет прогнозировать состояние системы обработки экспериментальных данных при изменении характеристик потока данных (например, увеличении или уменьшении). Одним из ограничений модели является то, что одно задание может требовать только один файл. Однако другие задания также могут требовать один из тех файлов, которые уже подгружены.

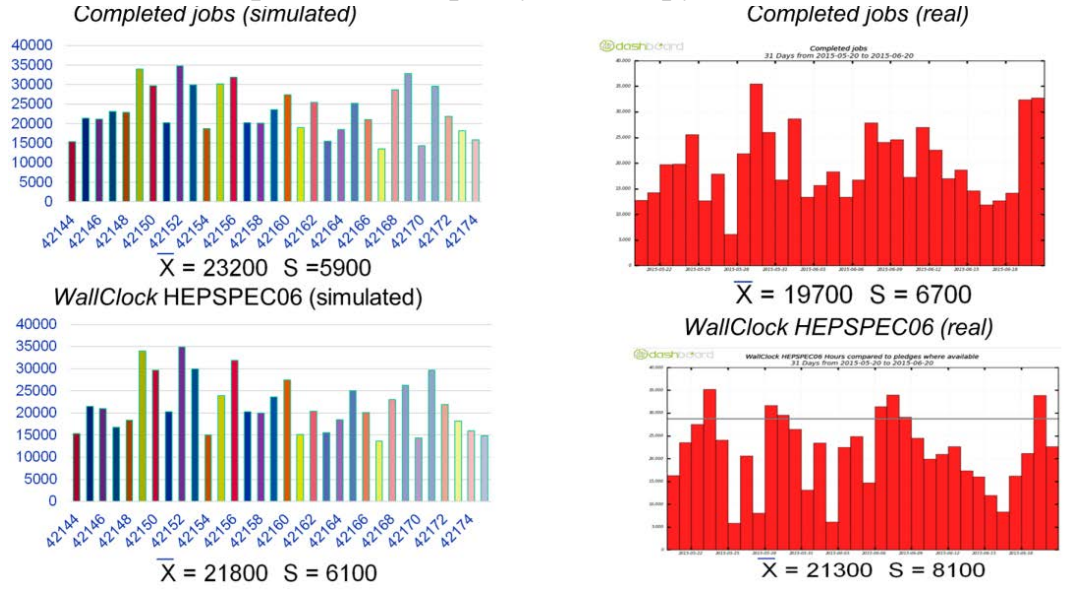
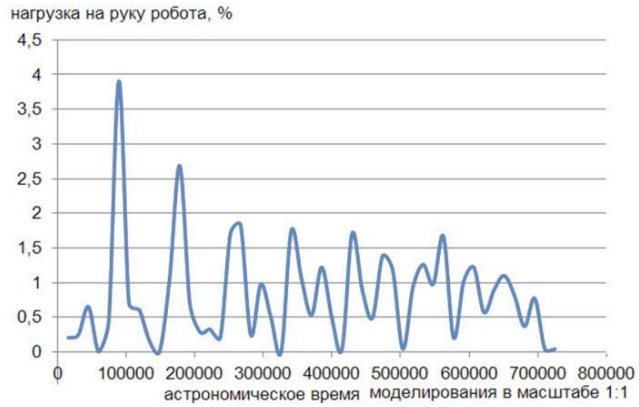
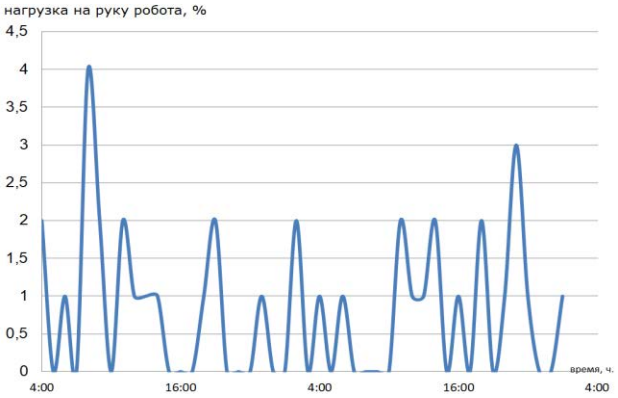


Рис. 6. Реальные и сгенерированные данные по числу завершенных задач

Ниже представлены результаты по моделированию нагрузки на ленточный робот (рис. 7а). Нагрузка определяется следующим образом: исходя из среднего времени движения руки робота 6 с., вычисляется максимальное количество движений за временной промежуток. Загрузка – отношение количества движений при моделировании к максимально возможному количеству движений. Оказалось, что рука робота будет загружена не более, чем на 4%. Причем вначале нагрузка на руку возрастает, потому что идет массовая загрузка файлов, которые требуются для выполнения задач, а потом нагрузка снижается, потому что часть файлов уже есть в буфере. Такая же ситуация наблюдается при работе реальной системы (рис. 7б).



а) результаты модели



б) реальные данные

Рис.7. Нагрузка на роботизированное захватное устройство ленточной библиотеки

Проведенные вычислительные эксперименты позволили адаптировать параметры модели к реальным процессам приема, хранения и обработки данных и подготовить программу для моделирования системы хранения и обработки данных с экспериментов, которые будут выполняться на вычислительном комплексе NICA.

Моделирование системы обработки и хранения данных эксперимента BM@N.

В состав комплекса NICA входят различные экспериментальные установки: BM@N, SPD, MPD. Для сбора данных эксперимента BM@N, на момент проведения исследований, предполагалось создание распределенной системы. Данные эксперимента при этом должны записываться на дисковые накопители для последующей передачи по коммуникационным линиям связи в центр уровня Tier-1, расположенный в ЛИТ (ОИЯИ). Предполагалось, что количество данных, полученных с установки, будет на уровне 1 ПБ за запуск. Данные должны записываться на диски и одновременно обрабатываться на компьютерной инфраструктуре ЛИТ. В диссертации приведены предполагаемые

характеристики потока данных с детекторов эксперимента BM@N, а также результаты моделирования.

Моделировалось количество свободных ядер на ферме ЛИТ во время выполнения задач реконструкции в зависимости от частоты их запуска, при этом полная обработка осуществлялась одновременно с поступлением событий. Предполагается, что эксперимент BM@N будет продолжаться 1500 часов. Первые 120 часов работы эксперимента будет осуществляться настройка оборудования, поэтому информация записываться не будет. Затем в течение 1000 часов будут накапливаться данные с эксперимента при интенсивности потока 70% от максимальной и записываться на диски, при этом полная обработка заданий будет осуществляться на 880 ядрах, экспресс обработка – на 120 ядрах. Оставшиеся 380 часов данные будут накапливаться с эффективностью 100%. После завершения эксперимента, т.е. окончания набора данных, полная обработка заданий будет осуществляться на 1000 ядрах.

Моделировались две стратегии пофайловой обработки событий: немедленная обработка 100% событий каждого файла, что приведет к накоплению большой очереди необработанных событий, которые придется обрабатывать потом после окончания сеанса, или обработка только 50%, т.е. каждого второго события, что существенно уменьшит очередь (рис. 8). Одно событие в среднем обрабатывается 1 с., за один запуск эксперимента получено 2500000 файлов и обработка событий осуществляется одновременно с их поступлением. Это значит, что потребуется некоторое количество времени после окончания эксперимента, чтобы полностью обработать все поступившие события. Моделирование этих двух стратегий показало, что при обработке каждого события в файле (100% событий) после окончания эксперимента потребуется 2000 часов, чтобы закончить обработку всех событий. Если же обрабатывать каждое второе событие (50%), то для завершения обработки потребуется только 500 часов.

На базе разработанного программного обеспечения SyMSim проведено проектирование вычислительного комплекса по приему и обработке данных с эксперимента BM@N. Моделирование позволило определить предполагаемую загрузку канала связи, количество ядер для полной и экспресс обработки заданий, требуемый объем ресурсов хранения данных.

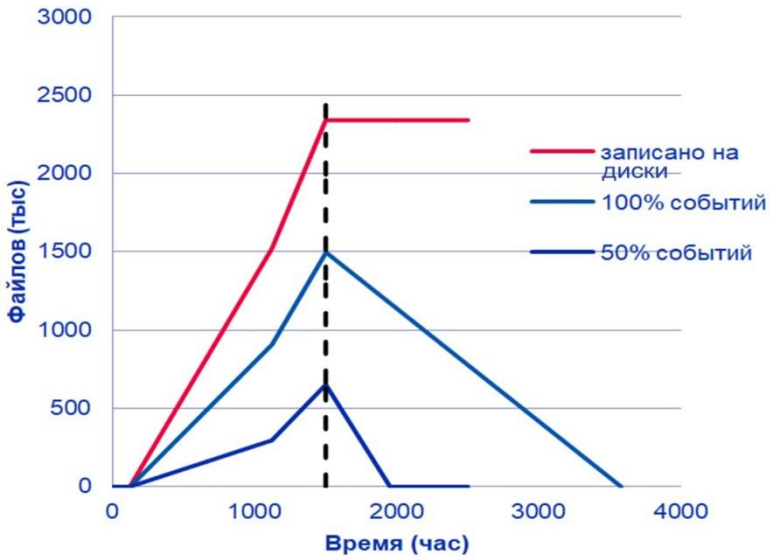


Рис. 8. Два сценария обработки событий

Моделирование вычислений MPI-приложений на облачной инфраструктуре

На базе вычислительного центра ОИЯИ разработан параллельный алгоритм и соответствующий комплекс программ для моделирования сверхпроводящих процессов в системе длинных джозефсоновских переходов с использованием технологии MPI (далее ДДП-моделирование). Наличие в ЛИТ облачной инфраструктуры позволило перенести выполнение ДДП-моделирования в облачную среду. Для оптимизации схемы параллельных вычислений необходимо протестировать работу алгоритма при различных сочетаниях параметров оборудования, количества процессоров и уровней распараллеливания. Таким образом, возникает проблема оценки влияния различных факторов гетерогенной среды (частоты процессора, пропускной способности коммуникационной сети, её латентности) на скорость вычислений конкретной задачи. Эту проблему предлагается решать методом имитационного моделирования.

Для моделирования вычислительных процессов, использующих интерфейс MPI, был применен комплекс SyMSim. Оригинальность предлагаемого подхода заключается в том, что применяется дискретное моделирование событий, что позволяет в рамках единого подхода описать программный комплекс, использующий интерфейсы MPI как на нескольких ядрах, в рамках одного сервера, так и виртуальные машины, взаимодействующие между собой в облачной архитектуре. Полигон представлял из себя корзину с 8 лезвиями, каждое из которых – это сервер Dell PowerEdge FC430 с 48 ядрами (два процессора Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz) и 256 ГБ ОЗУ, а также два SSD диска по 256 ГБ). Сервера внутри этой корзины взаимодействуют между

собой через Ethernet сеть с пропускной способностью 10 Гбит/с. Был осуществлен запуск 10 прогонов, при фиксированном количестве контактов 10 и изменяющемся количестве узлов в параллельном облачном кластере от 1 до 10 с шагом 1 (т.е. 10 контактов на 1 узле, 10 контактов на 2 узлах и т.д. 10 контактов на 10 узлах). Десятикратные повторения позволяли получить среднее время и его разброс для каждого количества рабочих узлов параллельного кластера при ДДП-моделировании с 10 контактами.

При построении модели шаги алгоритма ДДП-моделирования интерпретируются как несколько потоков заданий. Количество потоков совпадает с количеством процессоров, и каждый поток обрабатывается назначенным ему процессором. Задание (шаг алгоритма) может выполняться только после получения информации от предыдущих шагов.

Пусть параллельные процессы пронумерованы от 1 до N, число итераций T. На первом шаге –все процессы запускаются одновременно. Процесс m на текущей итерации t может быть запущен, если он получил данные от процесса $m-1$, выполненного на итерации $t-1$. Кроме того, существует процесс, который должен получить данные от всех процессов по окончании последней итерации. Время расчета одной итерации определяется случайным числом, распределенным по нормальному закону с известным средним значением. При таких упрощениях вычисления можно представить в терминах модели следующим образом. Процесс находится в состоянии ожидания до тех пор, пока не получает сигнал о готовности данных. После окончания работы процесса через случайный промежуток времени процесс посылает сигнал о наличии данных всем остальным процессам. После этого имитируется передача данных от одного процесса к другому, и алгоритм продолжается. Для моделирования такой структуры потребовалась незначительная модификация базовой версии SyMSim, в которой задания рассматриваются как независимые. Для учёта задержек, связанных с подготовкой буферов информации к передаче, в программу вводились дополнительные величины задержек, которые определялись по разнице времени выполнения на одном и двух процессорах.

Для сравнения результатов имитационного моделирования с экспериментом был использован аналитический подход при следующих упрощающих предположениях:

- Сумма количества операций, выполняемых для полного расчета, постоянна и не зависит от количества процессоров;
- Пропускная способность коммуникационной среды такова, что время обмена информацией не зависит от количества процессоров;

– Размер буфера обмена постоянный и не зависит от количества процессоров;
– Количество итераций постоянно и не зависит от количества процессоров;
– Время, затраченное программой до начала итераций и после их завершения, мало и им можно пренебречь.

Время расчета прямо пропорционально временным затратам на выполнение всех итераций и обратно пропорционально числу процессоров, а с учетом затрат на буферизацию данных получаем простую формулу:

$$T = \frac{T_v \cdot I}{n} + I \cdot t \qquad (1)$$

для $n > 1$, где n — количество процессоров, T_v — время, которое затратит один процессор на одну итерацию без учета обмена, I — количество итераций, t — время передачи буферов между процессорами за итерацию. Данные, полученные при моделировании, аналитически и в результате тестовых прогонов задачи ДДП-моделирования, представлены ниже (рис. 9).

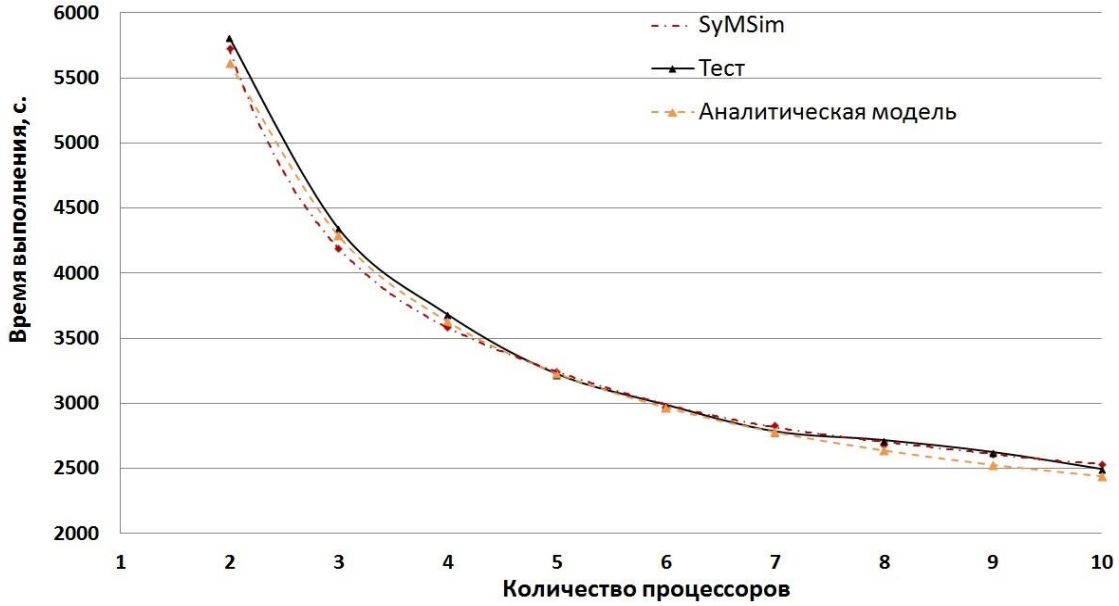


Рис. 9. Сравнение имитационной и аналитической модели с результатами теста

Сравнение результатов, полученных эмпирическим путем, с результатами имитационного моделирования показало, что имитационная модель корректно моделирует параллельные расчеты, выполненные с использованием технологии MPI, и подтвердило рекомендацию, что для быстрого счета задач такого класса необходимо одновременно с увеличением числа процессоров увеличивать пропускную способность сети. В случае алгоритма ДДП-моделирования удалось получить формулу, выражающую зависимость времени расчета от числа процессоров при фиксированной конфигурации системы. Приведённые

результаты демонстрируют, что программное обеспечение SyMSim можно успешно использовать для оценки времени выполнения MPI алгоритмов в облачной среде с учетом межпроцессорных соединений. Это позволит без проведения серии тестовых запусков в реальной компьютерной обстановке определить целесообразность использования облачной структуры, оптимальное количество процессоров при известном типе сети, характеризуемой пропускной способностью и латентностью. Совпадение результатов эксперимента, аналитической формулы и имитационного моделирования демонстрирует перспективность предлагаемого подхода.

Моделирование вычислительного центра ИФВЭ (Китай)

Научно-исследовательский институт физики высоких энергий китайской академии наук является крупнейшей китайской лабораторией по изучению физики элементарных частиц. IT-инфраструктура ИФВЭ обеспечивает поддержку крупномасштабных научных проектов. Моделирование процессов прохождения потоков данных и заданий при работе этого вычислительного центра имело целью выявить проблемы, возникающие в ходе обработки данных. Упрощенная схема IT-инфраструктуры ИФВЭ представлена на рис. 10.

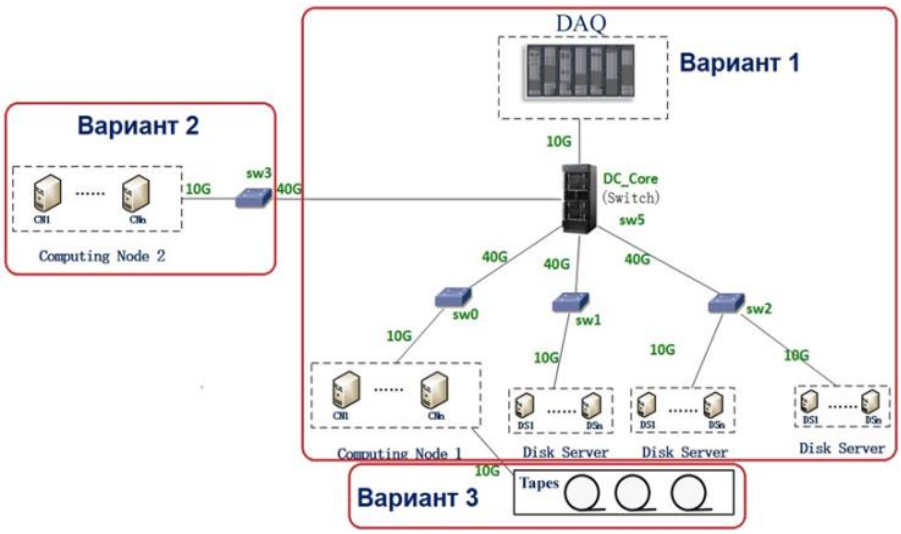


Рис. 10. Упрощенная схема возможных вариантов развития вычислительного центра ИФВЭ

Для выбора путей развития вычислительной мощности ВЦ исследовались различные варианты обработки данных: на ферме с 500 CPUs (вариант 1) и добавление еще одной фермы с 500 CPUs (вариант 2). Несколько заданий могут требовать один и тот же файл. Для того чтобы задание начало выполняться, файл должен быть доступен в локальном пуле дисков.

При добавлении еще одной фермы общее время ожидания увеличилось на 75%, поскольку требуется загрузка файлов на обе фермы, и скорость загрузки файлов постоянна, что доказывает нерациональность этого варианта (Таблица 2).

Для ускорения процесса прохождения заданий и повышение эффективности использования процессоров были предложены различные решения, в том числе был смоделирован вариант добавления роботизированной библиотеки хранения данных (вариант 3, рис. 10).

Таблица 2. Сравнение времени ожидания различных вариантов

	Общее вр. счета (мин.)	**Кол-во задач без ожидания**	**Кол-во задач с ожиданием выполнения до 60 мин.**	**Кол-во задач с ожиданием выполнения от 60 мин.**	**Общее время ожидания (мин.)**
Вариант 1(500 CPU)	1934,1	8567	872	561	1205
Вариант 2(2*500CPU)	1374,2	7598	1369	1006	2104

Система DAQ принимает данные о событиях с детектора с определенной частотой. Моделирование позволило определить зависимость общего количества переданных файлов от этой частоты. Поскольку количество передаваемых файлов также зависит от количества драйвов в роботизированной ленточной библиотеке, была смоделирована эта двойная зависимость. Показано количество файлов, которые сохраняются в буфер из системы DAQ без копии на ленты (мы не можем удалить такие файлы) и количество файлов с копией на ленте. Моделирование показало, что при достаточном количестве драйвов мы можем избежать очереди файлов и записать все файлы на ленты.

Программный инструментарий SyMSim был специально адаптирован для моделирования инфраструктуры вычислительного центра ИФВЭ в Пекине. Получены количественные характеристики процессов запуска задач и обработки потока данных, необходимых для оптимизации IT-инфраструктуры. Моделирование передачи данных с DAQ на ленточные ресурсы хранения позволило оценить необходимое количество лент для записи всех данных. Результаты моделирования показывают, что предложенный подход позволяет дать рекомендации по выбору топологии сети и характеристикам необходимого оборудования.

В силу общности реализации разработанный программный инструментарий SyMSim может быть применен для решения широкого класса задач при проектировании и поддержке центров обработки и хранения данных в различных областях.

В заключении сформулированы основные результаты диссертационной работы.

1. Предложен подход к моделированию систем хранения и обработки данных, который позволяет использовать результаты мониторинга для динамической коррекции параметров модели.

2. С использованием предложенного в диссертации подхода создан программный инструментарий для проведения анализа вариантов IT-инфраструктур с учетом их загрузки с целью выработки на этой основе эффективных решений для построения распределенной системы обработки и хранения данных.

3. На базе разработанного программного обеспечения проведено моделирование вычислительных центров для экспериментов MPD и BM@N комплекса NICA и вычислительного центра ИФВЭ (Пекин) и даны рекомендации по оптимизации этих IT-инфраструктур с учетом характеристик оборудования, что подтверждено отзывами разработчиков программного обеспечения эксперимента BM@N проекта NICA и руководством ВЦ ИФВЭ Пекина.

4. Разработанные методы и программные средства использованы для моделирования выполнения MPI-приложений на облачной инфраструктуре ЛИТ ОИЯИ, в результате чего определены условия эффективности использования облачных инфраструктур для проведения MPI-расчетов.

Успешно проведенное моделирование IT-инфраструктур различных типов подтверждает перспективность дальнейшего использования разработанных методов, алгоритмов и программных инструментов для решения задач проектирования и прогнозирования работы распределенных гетерогенных центров.

СПИСОК ОСНОВНЫХ ПУБЛИКАЦИЙ АВТОРА ПО ТЕМЕ ДИССЕРТАЦИИ

1. Andrey Nechaevskiy, Gennady Ososkov, Darya Pryahina, Vladimir Trofimov, Weidong Li. *Simulation approach for improving the computing network topology and performance of the China IHEP Data Center* // EPJ Web of Conferences, Vol. 214, 2019, 08018
2. Кутовский Н.А., Нечаевский А.В., Ососков Г. А., Пряхина Д.И., Трофимов В.В. *Моделирование межпроцессорного взаимодействия при выполнении MPI-приложений в облаке* // Компьютерные исследования и моделирование, Т.9, № 6, 2017, с.955-963.
3. Korenkov V.V., Nechaevskiy A.V., Ososkov G.A., Pryahina D.I., Trofimov V.V., Uzhinskiy A.V. *Simulation concept of NICA-MPD-SPD Tier0-Tier1 computing facilities* // Particles and Nuclei Letters, Vol. 13, № 5, 2016, p.1074-1083.
4. Korenkov V.V., Nechaevskiy A.V., Ososkov G.A., Potrebenikov Y.K., Pryahina D.I., Trofimov V.V., Uzhinskiy A.V. *Simulation of distributed data processing system for BM@N experiment of T0-T1 NICA project* // Selected Papers, 7th International Conference Distributed Computing and Grid-technologies in Science and Education, CEUR Workshop Proceedings, ISSN 1613-0073, Vol.1787, 2016, p.307–311.
5. Korenkov V.V., Nechaevskiy A.V., Ososkov G.A., Pryahina D.I., Trofimov V.V., Uzhinskiy A.V., Voytishin N.N. *The JINR Tier1 Site Simulation for Research and Development Purposes* // European Physical Journal (EPJ) – Web of Conferences, Vol.108, 02033, 2016.
6. Korenkov V.V., Nechaevskiy A.V., Ososkov G.A., Potrebenikov Y.K., Pryahina D.I., Trofimov V.V., Uzhinskiy A.V. *Optimization of Distributed Data Processing System for NICA BM@N Experiment by Using Simulation* // Procedia Computer Science, Vol.101, 2016, p.333-340.
7. Кореньков В.В., Нечаевский А.В., Ососков Г.А., Пряхина Д.И., Трофимов В.В., Ужинский А.В. *Моделирование грид и облачных сервисов как важный этап их разработки* // Системы и средства информатики, Т.25, вып.1, 2015, с.3-19.
8. Кореньков В.В., Нечаевский А.В., Ососков Г.А., Пряхина Д.И., Трофимов В.В., Ужинский А.В. *Синтез процессов моделирования и мониторинга для развития систем хранения и обработки больших массивов данных в физических экспериментах* // Компьютерные исследования и моделирование, Т.7, № 3, 2015, с.691-698.

9. Ososkov G.A., Korenkov V.V., Nechaevskiy A.V., Pryahina D.I., Trofimov V.V., Uzhinskiy A.V., Balashov N.A. *Web-Service Development of the Grid-Cloud Simulation Tools* // Procedia Computer Science, Vol.66, 2015, p.533-539.
10. Nechaevskiy, A.V., Pryahina, D.I., Trofimov, V.V. *Usage of data of a Tier1 site monitoring for simulation of the file distribution strategies* // CEUR Workshop Proceedings, ISSN 1613-0073, Vol.1536, 2015, p.173-178.
11. Korenkov V.V., Nechaevskiy A.V., Ososkov G.A Pryahina D.I., Trofimov V.V., Uzhinskiy A.V. *Simulation of Grid and Cloud Services as the Means of Improvement of Their Development Efficiency* // CEUR Workshop Proceedings, ISSN 1613-0073, Vol.1297, 2014, p.13–19.
12. Кореньков В.В., Муравьев А.Н., Нечаевский А.В. *Пакеты моделирования облачных инфраструктур* // Системный анализ в науке и образовании, №2, 2014.
13. Кореньков В.В., Нечаевский А.В., Ососков Г.А., Пряхина Д.И., Трофимов В.В., Ужинский А.В. *Моделирование грид-облачных сервисов проекта NICA, как средство повышения эффективности их разработки* // Компьютерные исследования и моделирование, Т.6, № 5, 2014, с.635-642.
14. Кореньков В.В, Нечаевский А.В., Трофимов В.В. *Разработка имитационной модели сбора и обработки данных экспериментов на ускорительном комплексе НИКА* // Информатика и ее применения, Т.7, вып. 3, 2013, с.130-137.
15. Кореньков В.В., Нечаевский А.В., Трофимов В.В. *Моделирование распределенной системы сбора, передачи и обработки данных для крупных научных проектов (мегапроект НИКА)* // Информационные технологии и вычислительные системы, №4, 2013, с.37-44.
16. Нечаевский А.В. *История развития компьютерного имитационного моделирования* // Системный анализ в науке и образовании, № .2, 2013.
17. Кореньков В.В., Нечаевский А.В., Трофимов В.В. *Модель системы offline обработки данных эксперимента НИКА* // Системный анализ в науке и образовании, № 4, 2012.