УДК 004.414.23, 519.876.5

МОДЕЛЬ СИСТЕМЫ OFF-LINE ОБРАБОТКИ ДАННЫХ ЭКСПЕРИМЕНТА НИКА

Кореньков Владимир Васильевич[1], Нечаевский Андрей Васильевич[2], Трофимов Владимир Валентинович[3]

[1]Кандидат физико-математических наук, профессор;
ГБОУ ВПО «Международный Университет природы, общества и человека «Дубна»,
Институт системного анализа и управления;
141980, Московская обл., г. Дубна, ул. Университетская, 19.
Заместитель директора лаборатории;
Объединенный институт ядерных исследований,
Лаборатория информационных технологий;
141980, Московская обл., г. Дубна, ул. Жолио-Кюри, 6;
e-mail: korenkov@cv.jinr.ru.

[2]Инженер-программист;
Объединенный институт ядерных исследований,
Лаборатория информационных технологий;
141980, Московская обл., г. Дубна, ул. Жолио-Кюри, 6;
e-mail: nechav@mail.ru.

[3]Ведущий программист;
Объединенный институт ядерных исследований,
Лаборатория информационных технологий;
141980, Московская обл., г. Дубна, ул. Жолио-Кюри, 6;
e-mail: tvv@jinr.ru.

В работе обоснована необходимость создания имитационной модели системы хранения и обработки данных ускорительного комплекса НИКА. На данном этапе работ в качестве платформы для создания модели выбрана GridSim. Для моделирования предложен ряд задач. В статье приведены результаты работы модели, а также сформулированы параметры оценки эффективности модели. Представлены интерфейсы для работы пользователя и графического отображения результатов.

Ключевые слова: GridSim, грид, имитационная модель, моделирование, обработка данных, НИКА.

OFF-LINE DATA PROCESSING SIMULATION FOR NICA EXPERIMENT

Korenkov Vldimir[1], Nechaevsky Andrey[2], Trofimov Vladimir[3]

[1]Candidate of Science in Physics and Mathematics, professor;
Dubna International University of Nature, Society, and Man,
Institute of system analysis and management;
141980, Dubna, Moscow reg., Universitetskaya str., 19.
Deputy Director of the Laboratory;
Joint institute for nuclear researches,
Laboratory of Information Technologies;
141980, Moscow reg., Dubna, Joliot-Curie, 6;
e-mail: korenkov@cv.jinr.ru.

[2]Software Engineer;
Joint institute for nuclear researches,
Laboratory of Information Technologies;
141980, Moscow reg., Dubna, Joliot-Curie, 6;
e-mail: nechav@mail.ru.

[3]*Lead Programmer;*
Joint institute for nuclear researches,
Laboratory of Information Technologies;
141980, Moscow reg., Dubna, Joliot-Curie, 6;
e-mail: tvv@jinr.ru.

The need of simulation grid systems for NICA experiment is proved. The platform GridSim for simulation grid systems is chosen. A number of tasks is offered for making simulation. Results of simulation are given in this work, and also some parameters of the model efficiency are formulated. The interfaces for the users work and results display presented.

Keywords: GridSim, grid, simulation, NICA.

Введение

В настоящее время в Объединённом институте ядерных исследований (ОИЯИ) создаётся ускорительный комплекс НИКА. Комплекс НИКА представляет собой ускоритель тяжёлых ионов НИКА и установку МПД, объединяющую детекторы для изучения ядерной материи в горячем и плотном состоянии, которое возникает при столкновении ускоренных тяжёлых ионов. МПД является источником информации с интенсивностью потока десятки петабайт в год.

Ожидаемая интенсивность потока информации настолько велика, что массивы данных характеризуются как сверхбольшие. Для обработки таких потоков информации используются распределённые системы коллективного пользования, построенные на грид технологиях.

Для оптимизации структуры будущего комплекса обработки информации необходимо определить его основные параметры, структуру и проверить предлагаемые технические решения с помощью моделирования.

Система обработки информации ускорительного комплекса НИКА

Хранение и использование экспериментальных данных в современных экспериментах физики высоких энергий является актуальной проблемой. Объем получаемых и обрабатываемых данных исключает возможность хранения и использования информации не только на одном кластере, но и в пределах одной организации, поэтому на первый план выходит создание распределённой системы хранения и обработки данных для эксперимента.

В случае НИКИ поток данных имеет следующие параметры:

- высокая скорость набора событий (до 6 КГц),
- в центральном столкновении Au-Au при энергиях НИКА образуется до 1000 заряженных частиц,
- размер файла с первоначальной моделируемой информацией с детекторов для 100000 событий занимает сейчас порядка 5 ТБ.

Схема получения и обработки данных представлена на рис. 1. Данные, идущие от персональных компьютеров поддетекторов МПД (Multi Purpose Detector), накапливаются специально предназначенными для сборки событий программами (Event Builder) компьютерной фермы в режиме онлайн и записываются на диск в режиме офлайн после формирования события через специально предназначенную для этой цели 10 Гб/с волоконно-оптическую линию связи. Каждая ЕВ записывает один «рабочий файл» каждую минуту сбора данных.

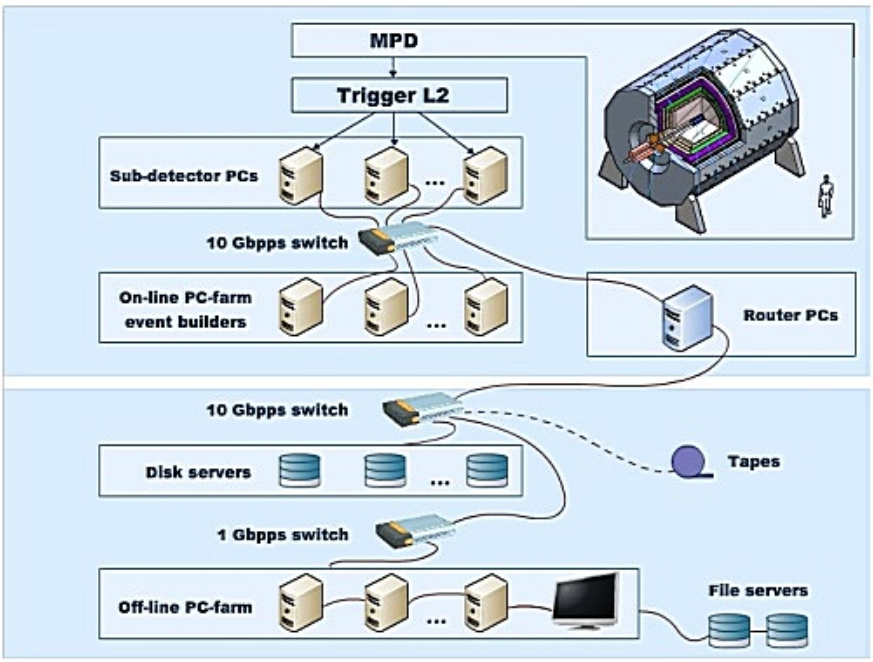

Рис. 1. Схема обработки физических данных ускорительного комплекса НИКА

После триггера высокого уровня отобранные события записываются в RAW файлы (скорость записи – один файл в 1 минуту сбора данных), и затем полностью восстанавливаются.

Прогнозируемое количество обрабатываемых событий при этом приблизительно 19 миллиардов. Принимая скорость передачи данных от датчиков 4.7 GB/s, общий объем исходных данных может быть оценен в 30 PB ежегодно, или 8.4 PB после обработки. Эти оценки основаны на особенностях DAQ, предыдущем опыте, и подобных оценках, выполненных для эксперимента ALICE [1].

В качестве системы обработки физической информации ускорительного комплекса НИКА планируется грид структура. Существующие решения по созданию распределенных систем для сбора, передачи и обработки сверхбольших объемов информации базируются на общих принципах построения грид инфраструктур. Так компьютерная инфраструктура для эксперимента ALICE представляет собой иерархическую структуру с компьютерными центрами класса Tier 0/1/2 (таблица 1). Для хранения и обработки данных с эксперимента ПАНДА предполагается использование инфраструктуры построенной на принципах грид по образу и подобию ALICE.

Суть иерархической модели состоит в том, что весь объем информации с детекторов после обработки в реальном времени и первичной реконструкции на вычислительных мощностях должен направляться для дальнейшей обработки и анализа в региональные центры. Иерархический принцип организации информационно-вычислительной системы предполагает создание центров разных уровней (Tier's). Уровни различаются как по масштабу вычислительных и архивных ресурсов, так и по выполняемым функциям [2]:

Таблица 1. Уровни иерархической модели и их функции

Tier0	первичная реконструкция событий, калибровка, хранение копий полных баз данных
Tier1	полная реконструкция событий, хранение актуальных баз данных по событиям, создание и хранение наборов анализируемых событий, моделирование, анализ
Tier2	репликация и хранение наборов анализируемых событий, моделирование, анализ
Tier3	кластеры отдельных исследовательских групп

При создании распределенной системы требуется принять решения по архитектуре инфраструктуры, количеству ресурсных центров, объему требуемых ресурсов. Кроме того, необходимо обеспечить достаточную пропускную способность, решить проблемы сохранности данных (устойчивость к повреждениям и удалениям) на протяжении всего жизненного цикла проекта, обеспечить распределение ресурсов между различными группами пользователей, выбрать алгоритмы обработки и запуска задач и многое другое. Для решения этих вопросов, а также обоснования решений, требуется создание имитационной модели обработки данных эксперимента. Возникает необходимость создания имитационной модели, которая бы удовлетворяла всем условиям.

Актуальность темы обуславливается тем, что на основе этой модели в дальнейшем может быть сформулировано конкретное техническое задание на разработку грид инфраструктуры.

Исходя из вышеизложенного, для проектирования грид структуры центров обработки и параметров системы off-line обработки данных ускорительного комплекса НИКА необходимо создать имитационную модель. Согласно планируемой процедуре использования, модель должна включать в себя интерфейс пользователя, собственно ядро моделирования, которое обрабатывает описания структуры обработки и потока заданий и определяет параметры прохождения заданий, систему визуализации результатов моделирования. Ядро моделирования будет включать имитаторы ограниченного набора функций грид, которые наиболее существенно влияют на прохождение заданий.

Платформа моделирования GridSim

Изучив предлагаемый на сегодняшний день инструментарий моделирования грид систем [3], мы решили разрабатывать систему моделирования off-line обработки данных ускорительного комплекса НИКА на базе платформы GridSim.

Проект GridSim [4] разрабатывается группой исследователей в лаборатории по изучению облачных и распределенных вычислений отдела информатики и компьютерных вычислений в университете Мельбурна, Австралия.

GridSim это набор библиотек предназначенных для построения модели грид системы. Она в свою очередь построена на стандартной библиотеке java SimJava, с помощью которой можно моделировать поток дискретных событий во времени. Приложение создаётся расширением классов GridSim и объединением их в программу, которая моделирует обработку потока заданий грид структурой, обладающей определёнными ресурсами и с заданной дисциплиной их резервирования и использования. В сравнении с другими пакетами моделирования грид GridSim обладает рядом преимуществ (таблица 2). Основные принципы, на которых построено описание ресурсов и их использование следующие:

а) позволяет моделировать гетерогенные типы ресурсов;

б) приложения с различными параллельными прикладными моделями могут быть смоделированы;

в) нет никаких ограничений на количество задач, которые могут быть отправлены на определенный ресурс;

г) пропускная способность сети между ресурсами может быть задана;

д) система поддерживает моделирование статистических и динамических планировщиков заданий;

е) статистика всех или выбранных операций может быть зарегистрирована.

Таблица 2. Функциии и свойства симуляторов грид [5]

Функция	*GridSim*	*OptorSim*	*Monarc*	*ChicSim*	*SimGrid*	*MicroGrid*
Репликация данных	Да	Да	Да	Да	Нет	Нет
Издержки записи/чтения диска	Да	Нет	Да	Нет	Нет	Да
Комплексное фильтрование или запросы данных	Да	Нет	Нет	Нет	Нет	Нет

Функция	*GridSim*	*OptorSim*	*Monarc*	*ChicSim*	*SimGrid*	*MicroGrid*
Планировка пользовательских задач	Да	Нет	Да	Да	Да	Да
Резервирование ЦПУ	Да	Нет	Нет	Нет	Нет	Нет
Симуляция нагрузки	Да	Нет	Нет	Да	Нет	Нет
Диференцированное QoS сети	Да	Нет	Нет	Нет	Нет	Нет
Генерация фонового сетевого трафика	Да	Да	Нет	Нет	Да	Да

Эта платформа позволяет пользователям моделировать работу грид системы с возможностью симулирования характеристик ресурсов и вычислительных сетей при различных конфигурациях. С помощью GridSim можно проводить воспроизводимые эксперименты, которые сложно реализовать в настоящем окружении динамических грид систем.

Таким образом, в качестве инструмента для моделирования выбрана система GridSim. В [6] показано, что GridSim недостаточно эффективен для моделирования больших систем, но в нашем случае количество центров обработки не превысит 20.

Задачи моделирования

В рамках выполняемой работы мы рассматриваем ряд задач, которые можно моделировать.

Первая задача подразумевает моделирование распределения данных на «нулевом» уровне (Tier0) обработки данных ускорительного комплекса НИКА. Зная характеристики событий и предполагаемый объем данных, необходимо ответить на следующие вопросы: сколько понадобится устройств для записи/чтения данных, что произойдет, если пользователь запросит файл с ленты, как будет работать при этом вся система и т.п.

На данный момент предполагается, что распределённая обработка и анализ моделированных данных с установки МПД о событиях столкновения тяжелых ионов на коллайдере НИКА будет возможна на вычислительных ресурсах следующих научных центров:

а) Объединенный институт ядерных исследований, Дубна, Россия.

б) Институт ISS (Institute of Space Sciences), Румыния, город Бухарест.

в) Кейптаунский университет (Университет Кейптауна) — одно из ведущих высших учебных заведений в ЮАР, расположенное в городе Кейптаун.

г) Санкт-Петербургский государственный университет – высшее учебное заведение города Санкт-Петербург, входящий в группу национальных исследовательских университетов России.

В дальнейшем список будет существенно расширяться как за счёт российских, так и международных партнёров. Из этого следует вторая задача – разработка инструмента для моделирования грид системы. На основании концепции дизайн-проекта ускорительного комплекса НИКА [7] можно построить модель, отражающую общие принципы построения систем в грид архитектуре, с возможно более широкой возможностью вариации параметров и возможностью дальнейшего их уточнения.

Структура модели

Имитационная модель моделирует прохождение набора заданий заданными пользователем параметрами, через грид структуру с заданной пользователем топологией и параметрами центров обработки. Модель позволяет получить оценку временных параметров обработки потока заданий при заданной пользователем дисциплине распределения ресурсов между заданиями и структурой очередей к центрам обработки.

Моделирование даёт ответы на вопросы:

а) какие вычислительные ресурсы требуются для обработки данных, чтобы получить результат в заданное время;

б) как должны быть связаны между собой центры обработки;

в) какое должно быть разделение функций между центрами;

г) какая стратегия запуска заданий должна применяться;

д) сколько памяти необходимо выделить для хранения информации.

Модель рассчитывает 8 параметров, определяемых как процент, или абсолютное значение:

а) средняя загрузка сети по дням [%];

б) количество активных /ожидающих заданий;

в) количество запрошенных и используемых ЦПУ;

г) использование грид структуры по часам [%] в день;

д) использование ресурсов хранения данных [%];

е) процент отказавших ЦПУ по дням [%];

ж) объем переданных данных в час;

з) использование кластеров [%].

Данный набор параметров достаточен для оценки эффективности топологии структуры, оценки её технического оснащения, эффективности алгоритмов распределения задач по узлам обработки.

С технической точки зрения система моделирования состоит из трех функциональных компонентов:

а) веб-сервер;

б) веб-клиент;

в) комплекс моделирования грид структуры.

Входные данные для имитации хранятся в каталоге файлов, и изменяются через веб-интерфейс (рис. 2). Это даёт пользователю возможность описывать и менять моделируемую грид структуру и параметры её загрузки заданиями, хранить варианты структур, выбирать вариант структуры.

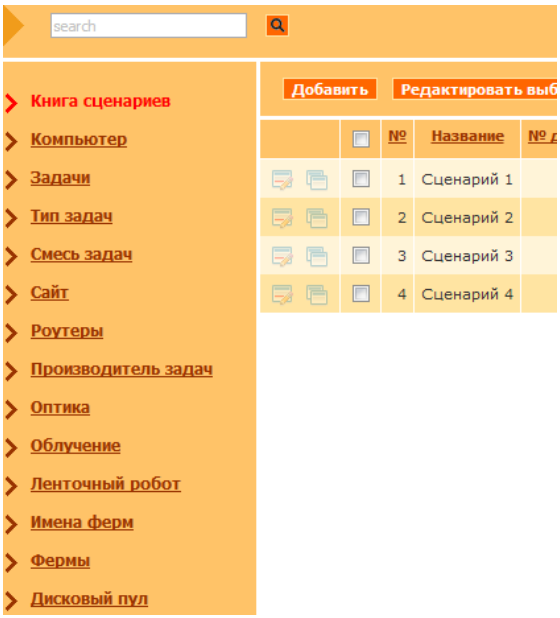

Рис. 2. Веб-интерфейс системы моделирования

На данном этапе выполнены следующие работы: создан веб-интерфейс редактирования модели с одним тестовым сценарием работы грид, выделены ключевые параметры оценки модели, созданы средства визуализации результатов, имитационная модель прошла отладку и верификацию.

Построение модели

Смоделируем ситуацию, приближенную к реальным условиям. При построении потока будем исходить из того, что каждое задание обрабатывает примерно 3000 событий, за сутки – 190080 заданий, за час – 7200. При таких нагрузках имеет смысл применить масштабирование. Выберем масштаб 1:40, т.е. моделируется прохождение 180 заданий, для обработки событий набранных в течении 1,5 мин, которые будут обрабатываться системой из 900 одинаковых процессоров, без учёта времени на передачу файлов. Временное распределение выбирается через равные промежутки времени для заданий реконструкции и равномерно распределённым для задач моделирования и физического анализа.

Рассмотрим пример моделирования двухуровневой системы обработки, состоящей из сайтов Tier0 (T0) и Tier2 (T2), как показано на рис. 3.

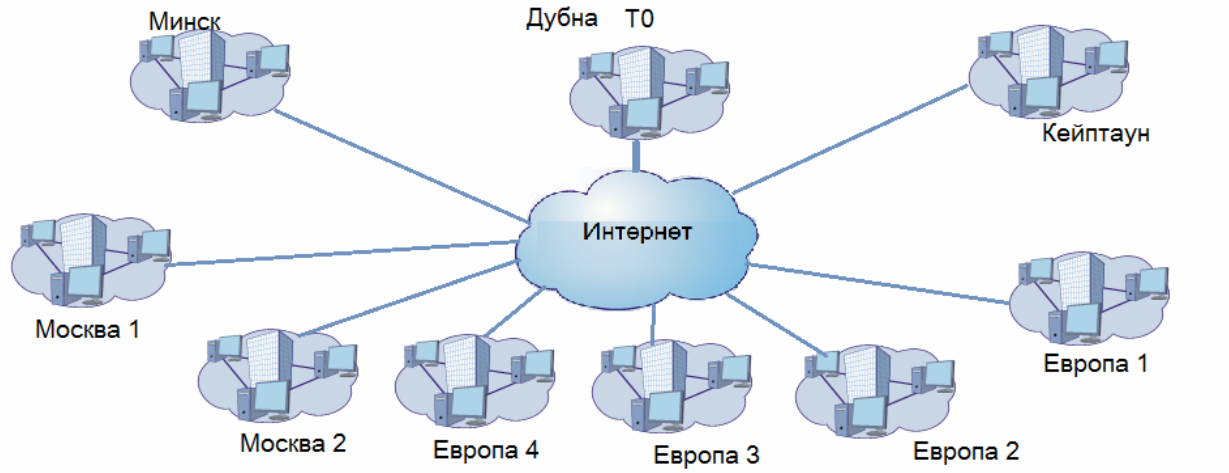

Рис. 3. Двухуровневая система обработки данных

Параметры таких систем приведены в таблице 3. Мощность процессора дана в единицах относительных 2.56 kSI2k, условно принято, что все процессоры сайта одинаковые. Под количеством процессоров понимается – какие вычислительные мощности сайт выделяет под задачу НИКА.

Таблица 3. Параметры сайтов

№	*Наименование / Tn*	*Доступ с Tier0 Гб/сек*	*MTU*	*Задержка (мс.)*	*Мощность проц. отн. условного*	*Кол-во процентов*
1	Dubna / Tier0	50	1500	1	1	500
2	Minsk / Tier2	5	1500	4	1	100
3	Москва1/ Tier2	5	1500	2	1	100
4	Москва2/ Tier2	5	1500	2	1	100
5	Кейптаун/ Tier2	1	1500	20	2	50
6	Европа1/ Tier2	5	1500	3	1	150
7	Европа2/ Tier2	5	1500	3	1	100
8	Европа3/ Tier2	5	1500	3	2	200
9	Европа4/ Tier2	5	1500	3	3	100

Рассмотрим 2 способа организации обработки информации.

В первом случае вся информация хранится на сайте уровня Tier0, а запуск заданий производится независимыми пользователями (рисунок 4 и 5).

Во втором случае информация также хранится на Tier0, но предпочтение на выполнение заданий отдаётся сайту Tier0. (рисунок 6 и 7).

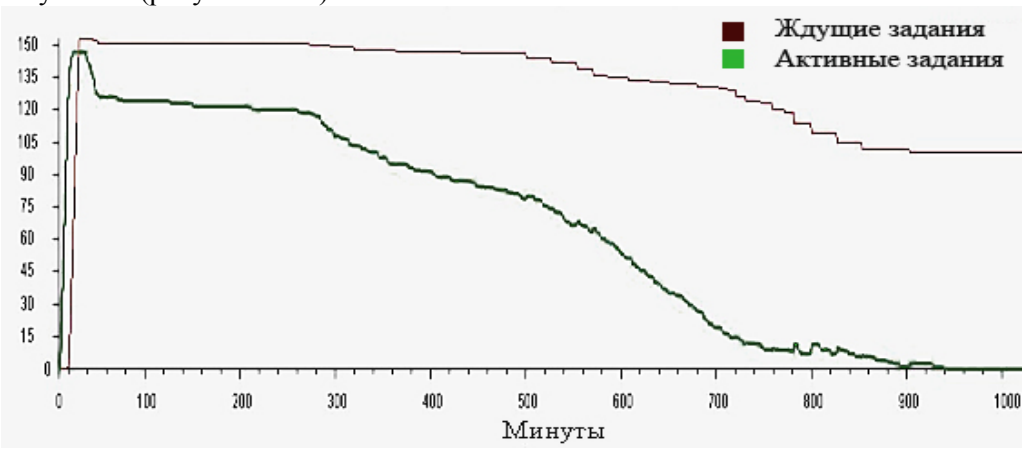

Рис. 4. Количество активных/ожидающих заданий.
Информация хранится на T0, запуск производится независимыми пользователями

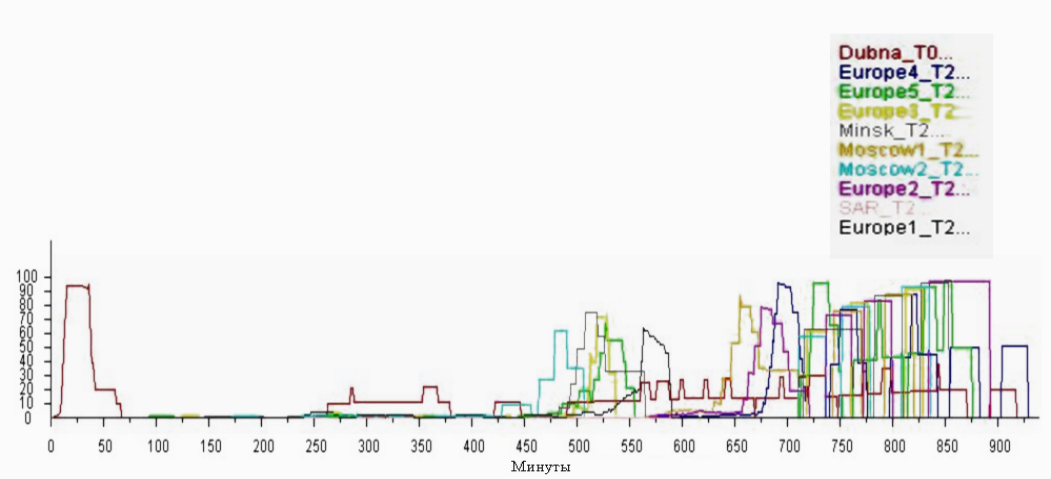

Рис. 5. Загрузка кластеров в %.
Информация хранится на T0, запуск производится независимыми пользователями

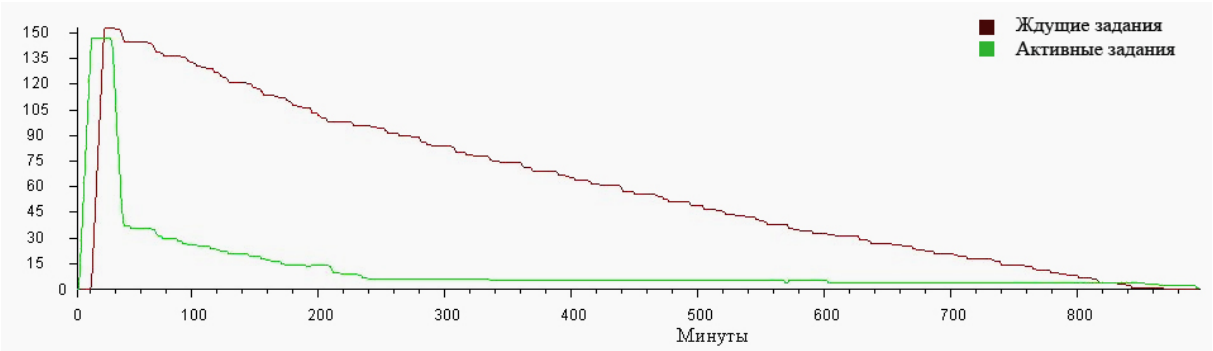

Рис. 6. Количество активных/ожидающих заданий.
Информация хранится на T0, предпочтения запуска – с T0

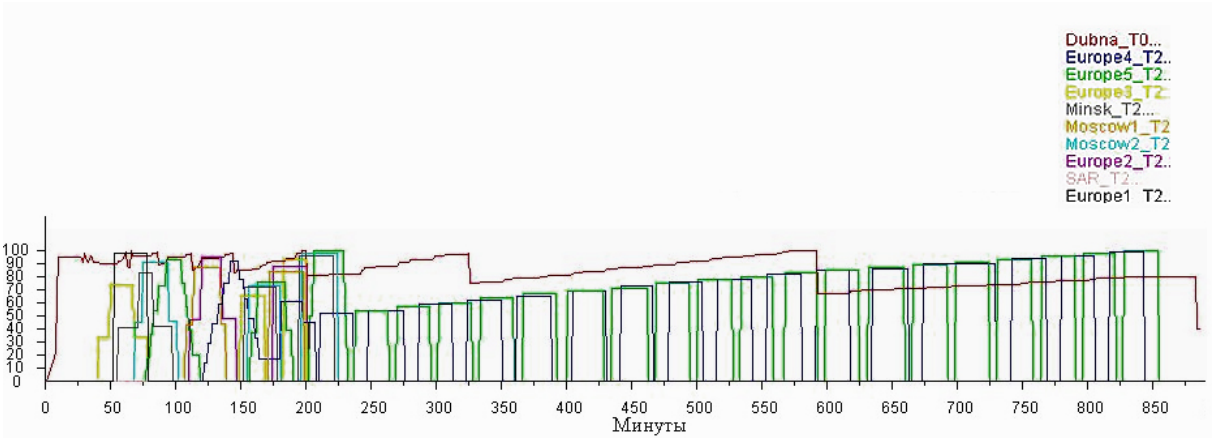

Рис. 7. Загрузка кластеров в %.
Информация хранится на T0, предпочтения запуска – с T0

Анализ полученных результатов

Поскольку точных данных нет ни по составу сайтов, ни по их конфигурации, анализ можно провести только на качественном уровне, получая соотношения времён обработки при различных архитектурах и организации процесса. Разница между первым и вторым случаем возникает из-за того, что задания требуют загрузки данных. Во втором случае эти загрузки производятся реже, поскольку основной поток заданий направляется на сайт T0.

Наихудший вариант архитектуры – это распределение информации по структуре обработки и отсутствие координации в запуске заданий (первый вариант). Возникающие при этом встречные потоки передачи практически блокируют обработку.

Наилучший вариант, это распределение информации по системе совместно с централизованным запуском заданий (второй вариант). Сайты, имеющие максимальное количество информации получают максимальные загрузки. Таким образом, централизованный запуск и мониторинг заданий необходим.

Предлагаемая архитектура аналогична используемой в проекте ПАНДА. Это даёт возможность использовать накопленный опыт при создании системы обработки эксперимента НИКА.

Заключение

Созданная система моделирования позволяет проводить разнообразные эксперименты с исследуемым объектом, не прибегая к физической реализации, что позволяет предсказать и предотвратить большое число неожиданных ситуаций в процессе эксплуатации, которые могли бы привести к неоправданным затратам, потере информации, а, возможно, и к повреждению дорогостоящего оборудования. В процессе моделирования можно определить минимально необходимое оборудование, обеспечивающее потребности передачи, обработки и хранения информации, оценить необходимый запас производительности оборудования, обеспечивающего возможное увеличение производственных потребностей, выбрать несколько вариантов оборудования с учетом текущих потребностей и перспективы развития в будущем, провести проверку работы системы, выявить ее «узкие» места и т.д.

Методика применения системы моделирования позволит определить параметры системы обработки информации ускорительного комплекса НИКА на этапе технического проектирования. Исходя из результатов моделирования, формулируется требование к архитектуре системы обработки – распределённое хранение накопленных данных и централизованный запуск заданий обработки.

Результаты работы могут быть рекомендованы для использования при проектировании грид системы для сбора, передачи, обработки и хранения данных с мегаустановок или других аналогичных установок, генерирующих большие объемы данных.

Список литературы

1. ALICE Collaboration (P. Cortese et al.), ALICE Technical Design Reort of the Computing. – CERN/LHCC 2005-018, ALICE TDR 12, 2005
2. Ильин В.А., Кореньков В.В., Солдатов А.А. Российский сегмент глобальной инфраструктуры LCG // Открытые системы. – 2003. – №1.
3. Нечаевский А.В., Кореньков В.В. Пакеты моделирования DataGrid // Системный анализ в науке и образовании: сетевое научное издание. – 2009. – №1. − [Электронный ресурс]. URL: http://www.sanse.ru/archive/12.
4. GridSim http://www.gridbus.org/gridsim/, 2012.
5. Sulistio A., Cibej U., Venugopal S., Robic B., Buyya R. A Toolkit for Modelling and Simulating Data Grids: An Extension to GridSim, Concurrency and Computation: Practice and Experience (CCPE), On-line ISSN: 1532-0634, Printed ISSN: 1532-0626, 20(13): 1591-1609, Wiley Press, New York, USA, Sep. 2008.
6. Aida K., Takefusa A., Nakada H., Matsuoka S., Sekiguchi S., Nagashima U. Performance Evaluation Model for Scheduling in a Global Computing System. //The International Journal of High Performance Computing Applications, Sage Publications, USA (2000). – Vol. 14. – №3.
7. Sissakian A., Sorin A. Многоцелевой Детектор – MPD для изучения столкновений тяжелых ионов на ускорителе NICA (Концептуальный дизайн-проект), версия 1.4. – 2011. –[Электронный ресурс]. URL: http://nica.jinr.ru/files/CDR_MPD/MPD_CDR_ru.pdf.