# Использование данных мониторинга сайта Tier1 для моделирования стратегий распределения файлов

© А.В. Нечаевский © Д.И. Пряхина © В.В. Трофимов

Объединенный институт ядерных исследований,
Дубна
nechav@jinr.ru pry-darya@yandex.ru tvv@jinr.ru

## Аннотация

В статье представлены результаты реализации проекта, который посвящен разработке и созданию средств моделирования грид и облачных сервисов для развития систем хранения и обработки больших массивов данных в физических экспериментах.
На базе проведенных исследований в ЛИТ ОИЯИ разработана новая система моделирования грид и облачных сервисов, ориентированная на повышение эффективности разработки путем учета качества работы уже функционирующей системы при проектировании ее дальнейшего развития за счет объединения самой программы моделирования с системой мониторинга реального (или модельного) грид-облачного сервиса через специальную базу данных.
На примере продемонстрировано применение программы для моделирования системы хранения сайта уровня Tier 1.
Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов № 14-07-00215 и № 15-29-01217.

## 1 Введение

Объемы обрабатываемой информации имеют тенденции экспоненциального роста, что ведет к развитию современных центров обработки и хранения данных с применением грид-облачных систем обработки данных и необходимости моделирования и прогнозирования их работы. Такие системы используются, например, для обработки и хранения данных с экспериментов физики высоких энергий, где ускорители частиц производят объем данных до сотен петабайт в год. Наиболее известные эксперименты LHC-CMS [1], LHC-Atlas [2], и находящиеся в процессе создания или проектирования FAIR-PANDA [3], BES-III [4], NICA-MPD [5]. Суть распределенных вычислений заключается в том, что вся информация от датчиков экспериментов должна быть направлена к огромному количеству других центров обработки данных.

**Труды XVII Международной конференции DAMDID/RCDL'2015 «Аналитика и управление данными в областях с интенсивным использованием данных», Обнинск, 13-16 октября 2015**

В настоящее время при проектировании грид систем используется подход, когда задача создания модели и формулировки рекомендаций по построению выполняется однократно при проектировании системы. Однако эксперименты продолжаются годами и десятилетиями, одновременно с эксплуатацией системы происходит ее развитие, не только качественное, но и количественное. При эволюции WLCG произошли качественные изменения систем хранения информации, а вместо планируемых трех уровней обработки данных появилось четыре. Таким образом, даже при значительных усилиях, вложенных на этапе проектирования в понимание конфигурации систем и их количественных характеристик, невозможно развивать систему без дополнительных исследований. Разработчики и эксплуатирующие организации сталкиваются с проблемой оптимизации и прогнозирования поведения системы. Оптимизация — это широкая стратегия, включающая минимизации всех возможных рисков (не только случаев критических клинчей, переполнений и т.д., но и финансовых потерь от медленной работы, резервации излишних ресурсов или простоя узлов) и одновременно максимизации всех полезных свойств системы - скорости работы, быстроте реакции на сигналы о неполадках и полной занятости оборудования.

В предыдущих работах авторов [6,7] описана программа моделирования, основанная на использовании языка GridSim [8] и алгоритмов планирования потока заданий ALEA [9]. Для запуска программы требуется задать состав и топологию центров обработки моделируемой грид-структуры, а также распределение ресурсов между заданиями. После этого программа выполняет имитационное моделирование процессов прохождения сгенерированного набора заданий через грид-структуру. В качестве результатов вычисляются временные оценки искомых параметров потока заданий.

Моделирование системы позволяет ответить на ряд вопросов. 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.

Таким образом, возникает необходимость создания методологии и программного окружения, позволяющего моделировать системы на постоянной основе, прогнозировать поведение системы при значительных изменениях.

## 2 Подход к моделированию

Одним из необходимых этапов проектирования физической установки является создание модели обработки данных. Любой проект сопровождается документом, озаглавленным “Data processing model”. Содержание документа отвечает на следующие вопросы:

- Как выглядит выбранная архитектура системы обработки вообще, будет ли создаваться структура в парадигме грид или хранение и обработка сосредоточится в единственном центре;
- Какой состав и параметры оборудования потребуются для организации обработки данных;
- Какие протоколы обмена данными будут использованы;
- Как оптимальным образом распределить данные, если они хранятся в географически распределённых местах;
- Какие времена обработки следует ожидать, при заданной архитектуре и выбранном оборудовании;
- Как на работу системы повлияют отказы её компонентов;
- На каких уровнях находятся пределы производительности.

Говоря о том, какую технологию моделирования применить, следует учесть, что возможность применения аналитических моделей для рассматриваемых задач ограничена по следующим соображениям. Существует несколько подходов при аналитическом моделировании грид и облачных систем, которые можно сгруппировать в два типа:

- система рассматривается как многоканальная система массового обслуживания, с состояниями, управляемыми марковским процессом, с ограничениями на распределения входных потоков и на дисциплины обслуживания, вызванными теоретическими предпосылками;

- и второй тип, когда система рассматривается, как динамическая стохастическая сеть, описываемая системами уравнений, позволяющими учитывать, как маршрутизацию, так и распределение ресурсов в сети, причем изучению подлежат равновесные и неравновесные состояния сети [10].

Оба подхода выдают результат моделирования, как правило, в виде асимптотических распределений и в силу ограниченных теоретических предпосылок не могут быть применены для моделирования конкретных сложных компьютерных сетей многоуровневой архитектуры с реальными распределениями входных потоков заданий, сложной многоприоритетной дисциплиной их обслуживания и динамическим распределением ресурсов.

Поэтому авторы считают правильным использовать имитационное моделирование.

Предлагаемый авторами подход состоит в имитационном моделировании, объединенным с мониторингом процессов прохождения задач и передачи данных. Эффективность этой концепции состоит в том, что техническое решение проверяется на модели прежде, чем обсуждается его фактическая реализация. Развернутое обоснование предлагаемого подхода, а также описание функций мониторинга, используемых при моделировании, дано в предыдущей работе [11], поэтому отметим здесь, что модель должна рассматриваться как неотъемлемая часть системы обработки данных, а данные мониторинга – как входные для моделирования. Это позволит принимать обоснованные проектные решения при развитии системы. Объединив моделирование и мониторинг в рамках одного программного пакета, можно добиться существенного снижения эксплуатационных затрат и вложений в увеличение мощности моделируемой системы с целью сохранения скорости получения результата экспериментов, при постоянном повышении интенсивности потока данных.

Центральным компонентом такой процедуры принятия решения по развитию вычислительной установки является имитационная модель вычислительной структуры, в которую в качестве входной поступает информация, накапливаемая в ходе мониторинга существующей установки в специальной базе данных (БД) , где она модифицируется в соответствии с планами развития моделируемой грид-облачной структуры.

Схема программы SyMSim (Synthesis of Monitoring and Simulation – Синтез Мониторинга и Моделирования), реализующей идею синтеза процессов мониторинга и моделирования, представлена на рис. 1.

Данные мониторинга реальной грид-облачной системы поступают в БД следующим образом: задания через систему управления нагрузкой (1) поступают на обработку в вычислительную систему (2), информация о статусе выполнения заданий поступает в БД (3). Статистические данные используются в качестве входного потока для модели. Также на базе статистических данных о задачах можно сгенерировать новый поток

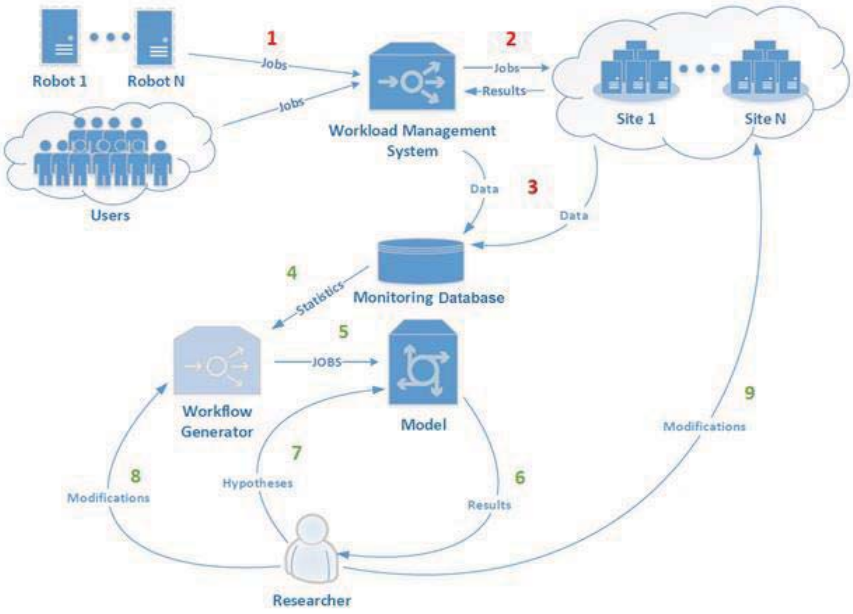


Рис.1. Информационная модель программы моделирования SyMSim

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 сайта для улучшения его характеристик (9).

После объединения моделирования и мониторинга в рамках одного программного пакета, возникает потребность в организации хранилища для данных мониторинга грид-систем, различных параметров моделирования и результатов работы программы. Поэтому важно правильно спроектировать и разработать БД, а также создать программное окружение и удобный интерфейс взаимодействия с ней.

## 3 Проектирование и разработка базы данных

БД содержит описание грид-структуры, каждого ее узла, связей между узлами, информацию о запущенных заданиях, временах исполнения, результаты мониторинга работы различных подсистем грид, а также результаты моделирования. Разработан веб-интерфейс, с помощью которого осуществляется описание вычислительной структуры, что включает задание параметров сайтов и связей между ними. Описанию присваивается идентификатор, который указывается в параметрах запуска модели. Модель считывает информацию из БД и строит описание вычислительной структуры. Характеристики потока задач, которые подлежат обработке, также задаются через веб-интерфейс. На основе заданных характеристик запускается набор утилит, который позволяет статистически проанализировать результаты мониторинга и сформировать поток заданий, аналогичный, или отличающийся от проанализированного на управляемое воздействие пользователя, в виде упорядоченной по времени последовательности записей в БД. Результатом работы программы моделирования служит последовательность записей в БД, отражающая все события, происходящие в системе. К ним относится, например, поступление задания, начало и конец обработки задания, начало и конец передачи файла, манипуляции с лентами и т.д. Все события описываются в едином формате. Запись привязывается к внутреннему времени.

Для обеспечения совместимости с системами мониторинга, формат БД был выбран совпадающим с системой управления потоком задач эксперимента ATLAS [12]. Такая совместимость дает возможность использовать результаты мониторинга потока без изменения его параметров. В качестве системы управления базами данных (СУБД) используется PostgreSQL.

Разработка процесса взаимодействия системы моделирования с БД осуществлялась на языке программирования Java в среде NetBeans. Был создан Java-проект, включающий класс, который реализует подключение к БД, создание курсоров, отправку запросов типа SELECT и INSERT, вывод результатов запросов. Взаимодействие приложения с СУБД PostgreSQL реализовано с помощью платформенно-независимого промышленного стандарта JDBC [13].

JDBC основан на концепции драйверов, позволяющих получать соединение с БД по специально описанному url. Драйверы могут загружаться динамически (во время работы программы). Загрузившись, драйвер регистрирует сам себя и вызывается автоматически, когда программа требует url, содержащий протокол, за который отвечает драйвер.

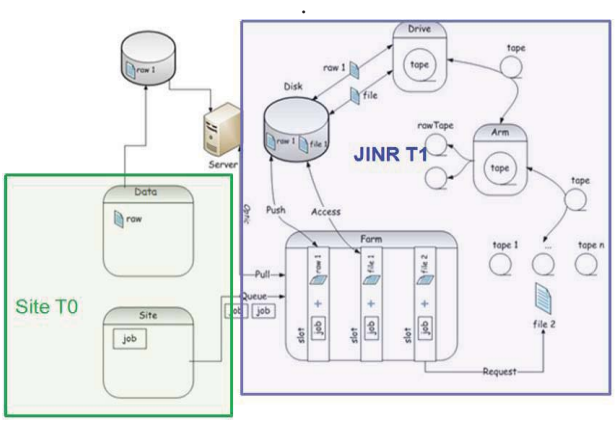


Рис.2. Схема прохождения заданий через SyMSim

## 4 Пример применения программы моделирования SyMSim

В настоящей работе на основе данных мониторинга одной из грид-систем WLCG [14], сохраняемых в специально разработанной БД, через веб-интерфейс выполняется их статистический анализ, результаты которого позволяют затем генерировать адекватный поток заданий для изменения параметров моделирования.

При разработке программы SyMSim удалось расширить ее сферу применения на ставшие весьма актуальными в последнее время системы облачных вычислений.

Для иллюстрации возможностей разработанной программы SyMSim по имитационному моделированию облачных вычислений ниже приведен пример ее применения для оптимизации простой облачной структуры. Эта структура предназначена для обработки данных физического эксперимента, но в принципе в такую схему укладываются и другие структуры связанные с хранением и обновлением больших массивов цифровой информации, в том числе и возможный вариант реализации электронной библиотеки.

### 4.1 Описание модели

Объектами моделирования являются вычислительные установки, предназначенные для обработки информации объёмом до десятков петабайтов в год. Как показал многолетний опыт работы центров разных уровней для распределенных вычислений и хранения данных, объединенных в систему WLCG, единственным способом хранения объёмов информации производимых такими детекторами является использование роботизированных библиотек. Данные затем обрабатываются на фермах, включающих тысячи процессоров. Предполагается, что моделируемая структура предназначена для обработки данных физического эксперимента, но другие структуры, связанные с хранением и обновлением больших массивов цифровой информации, также могут быть смоделированы.

Конструктивными элементами системы хранения и обработки данных являются оборудование передачи данных, дисковые хранилища, счётные фермы, ленточное хранилище. Информация с физического оборудования поступает на дисковые пулы, которые используются как хранилище данных и результатов работы программ. Для дальнейшей обработки данные с дисковых пулов переносятся на локальные диски компьютеров фермы. Для хранения больших объёмов информации используется роботизированная библиотека. При такой архитектуре можно говорить о двухуровневой структуре Tier0-Tier1. Её существенной особенностью является повышенная надёжность и пропускная способность каналов связи. Все объекты вычислительного центра связываются между собой локальной сетью, в состав которой входят линии передачи данных и коммутационные устройства. Исходя из этого, можно не принимать во внимание ограниченную надёжность каналов связи и их пропускную способность.

Альтернативный способ построение структуры – разделение функций хранения и обработки информации между уровнями Tier1 и Tier2, объединёнными в грид-структуру. В этом случае модель должна учитывать качество каналов связи и способы логической организации хранения файлов на географически распределённых сайтах Tier1.

При построении потока заданий предполагается, что вычисления выполняются на процессоре, взятом в качестве образца. При моделировании для конкретного оборудования вводится коэффициент приведения производительности оборудования к этому образцовому процессору.

Итак, рассматривается модель структуры, предназначенной для хранения данных в роботизированной библиотеке с набором кассет с магнитными лентами, из загрузчиков-драйвов которых робот автоматически достаёт требуемые ленты и устанавливает в одно или несколько устройств чтения-записи. Схема прохождения задания через систему моделирования SyMSim представлена на рис. 2. Задание начинает выполняться, если есть свободный слот-процессор и все файлы доступны на дисковом хранилище облака. 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 на драйв, которую выполняет рука робота, монтирования файловой системы картриджа на драйве и записи файла на диск.

Объект моделирования представляет собой несколько сайтов WLCG, соединённых каналами связи. При этом только один сайт моделируется детально. В его состав входят счётные узлы, дисковые пулы, роботизированное ленточное

хранилище. Остальные сайты рассматриваются, как хранилища данных, и от дисковых пулов отличаются только свойствами линий связи.

Линии связи разделяются на внешние и внутренние. Внутренние имеют постоянную пропускную способность. Пропускная способность связей между сайтами является случайной величиной, распределённой по экспоненциальному закону.

На сайт поступают данные и поток заданий. Смесь заданий состоит из трёх частей: моделирование, реконструкция и анализ. Части распределены между собой в заданной пропорции. Требования к данным для разных заданий различные. Моделирование не требует входных данных, но генерирует выходные. Анализ не требует входных данных и не генерирует выходные. Восстановление требует входные данные, но не генерирует выходных.

Моделируемая структура состоит из вычислительной фермы, дискового пула, каналов связи, ленточного робота IBM 35683 и лент. Параметры сайта Tier 0 следующие: 500 слотов, 1 пул дисков, средняя скорость сбора данных – один файл каждые 7 секунд. Сайт Tier 1 принимает файлы с сайта Tier 0. Tier 1 имеет следующие параметры: 2400 слотов, 8 дисковых пулов, 9 драйвов, 600 лент, задачи поступают на сайт в среднем каждые 5 секунд. Технические параметры устройств соответствуют реальным. Пропускная способность каналов связи от 10 до 100 ГБ/с.

При моделировании объектов приняты некоторые упрощения: количество активных сайтов ограничено; берётся единый поток заданий для разных стратегий; каждое задание требует единственный файл; несколько заданий могут требовать один и тот же файл; в начальный момент времени файлы статически распределены между сайтами, дисками и лентами; файлы, записанные на дисковые пулы сайта остаются там до конца эксперимента; каждый файл имеет единственную копию.

### 4.2 Проверка функционирования модели

Поток заданий взят из статистики CSM по завершению заданий на T1_JINR_RU за сутки начиная с 2015-07-09 08:00:03. На рис. 3 приведено сравнение гистограмм по совпадению реальных и модельных почасовых данных о числе завершенных заданий. Видно совпадение с учётом сдвига модельных данных на величину среднего потребляемого времени процессора.

### 4.3 Вычислительный эксперимент

Анализируются следующие стратегии распределения файлов по линиям связи по количеству завершённых заданий в течение часа после двух часов работы: (1) все файлы на локальных дисках; (2) файлы на локальных дисках и сайтах в соотношении 80% на 20%; (3) файлы на локальных дисках и лентах в соотношении 80% - 20%; (4) файлы на локальных дисках, лента и сайтах в соотношении 80% - 10% - 10%.

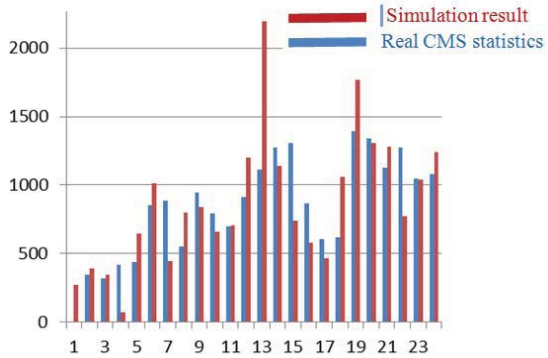


Рис.3. Реальные и сгенерированные данные по числу завершенных задач за 24 часа

На рис. 4 и рис. 5 показано количество заданий, завершившихся между 8 и 16 часами системного времени вычислительного эксперимента и частота загрузки лент. Показано, что хранение на лентах, когда нет группировки файлов, существенно замедляет счёт. С другой стороны, качество линии связи большого влияния не оказывает.

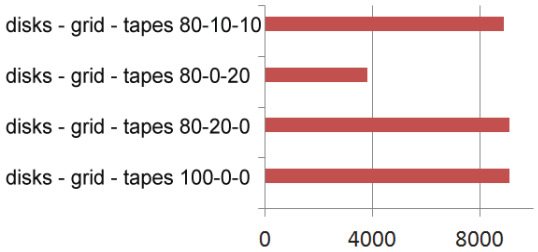


Рис.4. Сравнение стратегий распределения файлов

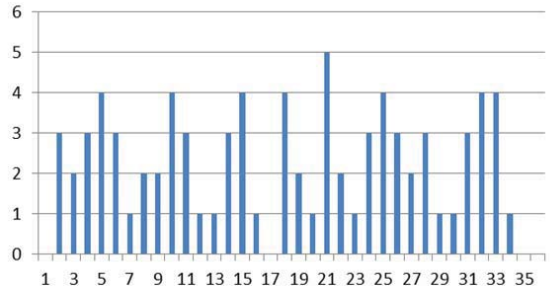


Рис.5. Распределение нагрузки на ленты

## Заключение

Предложенный подход к моделированию и анализу вычислительных грид-облачных структур в экспериментальной физике высоких энергий основан на учете данных их мониторинга, используемых затем для динамической коррекции параметров моделирования. Новизна работы обоснована объединением в рамках одного программного продукта процессов моделирования и мониторинга реального (или модельного) грид-облачного сервиса через специальную БД. Объединив моделирование и мониторинг в рамках одного программного пакета, можно добиться существенного снижения эксплуатационных затрат и вложений в увеличение мощности моделируемой

системы с целью сохранения скорости получения результата экспериментов при постоянном повышении интенсивности потока данных.

Проведенный вычислительный эксперимент по моделированию файловой загрузки в центрах распределенных вычислений с двухуровневой структурой типа Tier0-Tier1 показал хорошее совпадение результатов работы программы SyMSim с данными мониторинга реальной компьютерной системы, послужившей прототипом для модели.

В силу общности своей реализации разработанная программа моделирования SyMSim может быть также применена для решения более широкого класса задач проектирования виртуальных центров обработки и хранения больших массивов данных. В частности, программу можно применять для проектирования и последующего развития хранилищ информации общего доступа, не ограниченных областью физического эксперимента.


### Благодарности

Авторы выражают свою признательность профессору В.В. Кореньков, профессору Г.А. Ососков, А.В. Ужинскому за ценные консультации и помощь в работе.


## Литература


[1] CMS detector web page http://home.web.cern.ch/about/experiments/cms

[2] ATLAS detector web page home.web.cern.ch/about/experiments/atlas

[3] PANDA - веб-портал проекта www-panda.gsi.de/

[4] BESIII - веб-портал проекта http://bes3.ihep.ac.cn/

[5] Сисакян А.Н., Сорин А.С. Многоцелевой Детектор – MPD для изучения столкновений тяжелых ионов на ускорителе NICA (Концептуальный дизайн-проект), версия 1.4. – 2011. – nica.jinr.ru/files/CDR_MPD/MPD_CDR_ru.pdf.

[6] В.В Кореньков, А.В. Нечаевский, В.В. Трофимов Разработка имитационной модели сбора и обработки данных экспериментов на ускорительном комплексе НИКА // Информационные технологии и вычислительные системы, 4, 2013, Стр. 37-44.

[7] В.В Кореньков, А.В. Нечаевский, Г.А. Ососков, Д.И. Пряхина, В.В. Трофимов, А.В. Ужинский Моделирование грид и облачных сервисов как важный этап их разработки // Системы и средства информатики, Т. 25, 1, 2015, Стр. 4-19.

[8] GridSim web-portal: URL:http://www.gridbus.org/gridsim/, 2012

[9] D. Klusacek, L. Matyska, and H. Rudova. Alea - Grid scheduling simulation environment// In 7th International Conference on Parallel Processing and Applied Mathematics (PPAM 2007), volume 4967 of LNCS, pages 1029-1038. Springer, 2008.

[10] Ю.С. Попков Макросистемы и grid-технологии: моделирование динамических стохастических сетей // Проблемы управления, № 3, 2003.

[11] 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, 2014. Т. 6. № 5. Стр. 635–642.

[12] Maeno T., De K., Panitkin S. PD2P: PanDA Dynamic Data Placement for ATLAS // For the ATLAS Collaboration. – 7 c

[13] JDBC Overview http://www.oracle.com/technetwork/java/overview-141217.html

[14] The Worldwide LHC Computing Grid. http://home.web.cern.ch


## Usage of Data of a Tier1 Site Monitoring for Simulation of the File Distribution Strategies


Andrey V. Nechaevskiy, Dariya I. Pryahina, Vladimir V. Trofimov



The results of the project which is focused on the development and design of tools for grid and cloud services simulation are shown in the paper. The simulation is needed for efficient optimization of complex distributed data processing systems of scientific experiments.

A new grid and cloud services simulation system has been developed in LIT JINR. This system is focused on improving the efficiency of the grid-cloud systems development by using work quality indicators of some real system to design and predict its evolution. For these purposes the simulation program is combined with real monitoring system of the grid-cloud service through a special database.

An example of the program usage to simulate a Tier 1 data storage system is given.