Московский ордена Трудового Красного Знамени
физико-технический институт
(государственный университет)
Факультет общей и прикладной физики
Кафедра фундаментальных и прикладных проблем физики микромира

Объединенный институт ядерных исследований
Учебно-научный центр

Кадочников И.С.

Архитектура мониторинга сайтов Tier-3.

Магистерская диссертация

Научный руководитель:
к.ф.-м.н.,КОРЕНЬКОВ В.В.

Рецензент:
д.ф.-м.н., БЕДНЯКОВ В.А.

Дубна, Июнь 2011

Содержание

Введение

Модель обработки данных

Эксперименты, проводимые на Большом адронном коллайдере, беспрецедентны по своим масштабам, не только с точки зрения физических размеров установки и энергии сталкивающихся частиц, но и с точки зрения объёма производимого потока данных. Для его обработки было бы недостаточно вычислительных мощностей одного, даже очень крупного, суперкомпьютерного центра, поэтому одновременно со строительством ускорителя подготавливалась система распределённых вычислений для хранения и обработки результатов экспериментов. Было принято решение, что для обработки полученных данных необходимо объединить мощности множества географически распределенных вычислительных систем.

В качестве программной среды, объединяющей географически распределённые вычислительные центры, используется технология грид, а объединённая инфраструктура носит название WLCG (Worldwide LHC Computing Grid). Структура WLCG – иерархическая.

У начала информационного потока находится Tier0 центр, расположенный в ЦЕРН, получающий данные непосредственно с экспериментальных установок. Целью Tier 0 является первоначальный отбор и реконструкция событий, хранение сырых данных, а также распределение данных по Tier 1.

Данные с Tier 0 получают 11 центров первого уровня, расположенных в различных странах по всему миру. Tier 1 сайты предназначены для постоянного хранения данных, реконструкции и обработки событий и анализа. С каждым Tier 1 центром связано облако Tier 2 центров, целью которых является моделирование событий и физический анализ данных.

Распределением данных и задач по Tier 1 и Tier 2 центрам в рамках виртуальной организации осуществляется в автоматическом режиме. Для

решения этих задач разработаны специализированные программные комплексы, называемые центральными сервисами. Для управления данными коллаборации ATLAS разработан фреймворк DQ2, для распределения задач между сайтами используется система PANDA. Согласованная работа сервисов обеспечивает работоспособность всех процессов обработки данных в виртуальной организации.

Центральные сервисы включают в себя средства мониторинга запускаемых задач и передачи файлов, а также используют получаемую из них информацию о популярности наборов данных для управления их размещением и оптимизации количества хранимых копий данных.

Необходимость создания Tier 3

Несмотря на значительный объём вычислений, необходимых для репроцессинга и калибровки, объём работ можно планировать, а процессы хорошо отлажены. Физический анализ, однако, гораздо менее предсказуем. Прежде, чем будет получен новый, выдающийся результат, неизбежен перебор множества различных вариантов.

Конкуренция в стремлении к открытиям заставляют учёных и коллаборации использовать технические возможности вычислительной техники на пределе, а администраторы должны поддерживать баланс между долговременными и хорошо продуманными планами и новыми, часто радикальными, подходами к расчётам.

Объёмы данных и количество людей, вовлечённых в эксперименты на БАК, нагружают вычислительные ресурсы, хранилища, сети и персонал гораздо больше, чем эксперименты на Тэватроне. Выработка и исполнение долговременных стратегий продакшена и анализа для CDF и DØ было сложной задачей даже после многих лет работы ускорителя из-за необходимости постоянно реагировать на скачкообразно возрастающую светимость, новые и развивающиеся методы анализа, неожиданные

революции в технологиях. Несмотря на хорошо проработанное долгосрочное планирование, в результате накопленного опыта работы ускорителя и экспериментальной установки, необходимо вносить корректировки в планы. Такие корректировки накладывают свой отпечаток на все подсистемы и вычислительная – не исключение. На данный момент пересматривается модель вычислительной структуры коллаборации ATLAS для удовлетворения потребностей на ближайшие два-три года.

В начале процесса получения научных результатов лежит поток сырых данных от триггеров высокого уровня (High Level Triggers, HLT) по 1,8 Мб с частотой 200 Гц (в текущем году частота будет увеличена до 400 Гц, а в 2012 до 600Гц), что составляет 30 Тб в день. В конце же – уменьшенный набор данных, подходящий для конечного анализа, на настольном компьютере участника ATLAS. Размер редуцированных данных должен удовлетворять важнейшему требованию: время проведения анализа на всём наборе данных должно быть приемлемо для запускающего.

Простая оценка показывает: на большинстве систем максимальная скорость обращения к ROOTtuple ограничена вводом-выводом на уровне 10 Мб/с. Время терпения человека можно оценить в один час, ожидая, что чтение данных и построение по ним гистограммы должно укладываться в этот интервал. Если предположить, что за год редкий сигнал плюс фон накапливают только миллион событий, то в данном оценочном примере получаем необходимый размер конечных данных: 40Кб на событие. Сырые данные должны быть уменьшены до 2% изначального размера, а выборка событий с триггеров высокого уровня должна быть уменьшена в 300 000 без потери важной информации.

При утверждении схемы обработки данных с БАК этот процесс многократно рассматривался, но полного понимания ситуации с финальным анализом, результатом которого будут научные статьи, не существует. Одной из проблем является нереалистичность простого примера, приведённого

выше: размера в миллион событий для анализируемой выборки недостаточно. То есть, в большинстве случаев, простой анализ на рабочей станции не возможен, а размер набора данных составляет несколько терабайт. При этом оценка приемлемого времени ожидания в один час остаётся верной, что означает необходимость использовать для анализа многопроцессорные системы. Однако в предыдущих крупномасштабных экспериментах замечена закономерность, противодействующая росту размера наборов данных, а именно: физики стремятся уменьшить используемые наборы, чтобы вести анализ как можно ближе к своему настольному компьютеру.

Стремление получить данные как можно ближе к человеку, ведущему анализ, предсказуемо. Это позволяет минимизировать влияние неизбежных факторов, задерживающих работу с системой, обслуживающей многих пользователей. Проблемы быстрого запуска и перезапуска задач, передачи данных и удалённого мониторинга удобнее всего решать при локальном контроле над вычислительным комплексом. Для решения этих проблем коллборация предлагает задействовать вычислительные ресурсы малых физических групп (институтов или университетов)[1]. В терминах коллаборации такой ресурс называется сайтом Tier3.

Определение Tier 3

Точное определение Tier 3 сайту дать сложно, поскольку при построении структуры WLCG добавление третьего уровня иерархии не предусматривалось. Главной отличительной чертой Tier 3 является локальное управление, то есть отсутствие обязательств по их структуре, размеру и предоставляемым грид-сервисам. Обычно Tier 3 сайт используется для физического анализа и расположен географически близко к группе, которая его использует.

Существующие Tier 3 сайты можно разделить на несколько типов [2].

Tier3af – крупный вычислительный центр, используемый для анализа несколькими группами страны или региона. Примером может являться немецкий центр DESY.Tier3af может являться или не являться грид сайтом. Под грид сайтом мы понимаем сайт с поддержкой необходимых сервисов для получения задач и данных из грид.

Co-located Tier 3 – сайт, расположенный рядом с существующим центром первого или второго уровня. Такие Tier 3 обычно используют ту же инфраструктуру, но ресурсы отделены и находятся в локальном или региональном подчинении.

Tier3gs – сайт, обладающий полным набором сервисов полноценного Tier2 сайта. Может, так же как и Tier 2, принимать участие в репроцессинге и использоваться для анализа удалённо через грид.

Tier3g – сайт с частичной поддержкой грид-сервисов. Например, элемент хранения, использующий только протокол Gridftp, или сайт, на котором есть лишь клиенты грид-сервисов.

Tier3w – рабочая станция без кластеризации, но с набором программ для доступа к грид в качестве потребителя. Обладает ограниченными возможностями для хранения данных и вычислительной мощностью.

Предпосылки создания системы мониторинга сайтовTier 3

Мониторинг – неотъемлемая часть обслуживания вычислительных комплексов и систем. Для поддержки сайта очень важно иметь данные о состоянии инфраструктуры и нагрузке на её элементы. Однако помимо общей потребности в мониторинге, система мониторинга для Tier 3 должна удовлетворять некоторым специфическим требованиям [3].

Поскольку Tier3 сайт может не поддерживать всех грид-сервисов, мониторинг средствами центральных сервисов управления задачами и данными невозможен. Однако для корректной работы центральных сервисов

им требуется информация о задачах, запускаемых на Tier 3 сайтах, и в первую очередь о популярности используемых наборов данных.

В случае, когда центры второго и третьего уровня использую общую инфраструктуру, необходимо также учесть возможное уменьшение ресурсов, предоставленных коллаборации, по сравнению с договорённостями, за счёт их использования для нужд Tier 3.

Важной задачей является мониторинг запускаемых задач. Это как поможет локальным администраторам и пользователям отслеживать их выполнение, так и предоставлять информацию о загрузке системы для своевременного планирования ее расширения.

Для получения информации о популярности наборов данных для анализа недостаточно отслеживания доступа к информации через грид, необходим мониторинг доступа к файлам, находящимся в системе хранения Tier 3 сайтов.

Одной из особенностей сайтов Tier 3 является существенное ограничение в специализированных кадровых ресурсах для системного администрирования сайтов. Поэтому одним из требований к системе мониторинга Tier 3 сайтов является лёгкость установки и настройки, а также наличие полной документации для развертывания программных средств.

Tier 3 мониторинг

Потребители

Система мониторинга Tier 3 сайтов должна удовлетворять требованиям двух групп пользователей. Локальным системным администраторам необходима подробная информация о работе подсистем сайта. В то же время для менеджмента виртуальной организации нужен глобальный мониторинг качества предоставленных Tier 3 сайтами услуг, в том числе передачи данных между сайтами и объёмов проведённой обработки данных, а для центральных сервисов важен глобальный мониторинг задач и наборов данных на Tier 3.

Ограничения

Мониторинг не должен требовать затрат большого количества ресурсов для запуска и поддержки, как вычислительных, так и кадровых.

Компоненты системы мониторинга

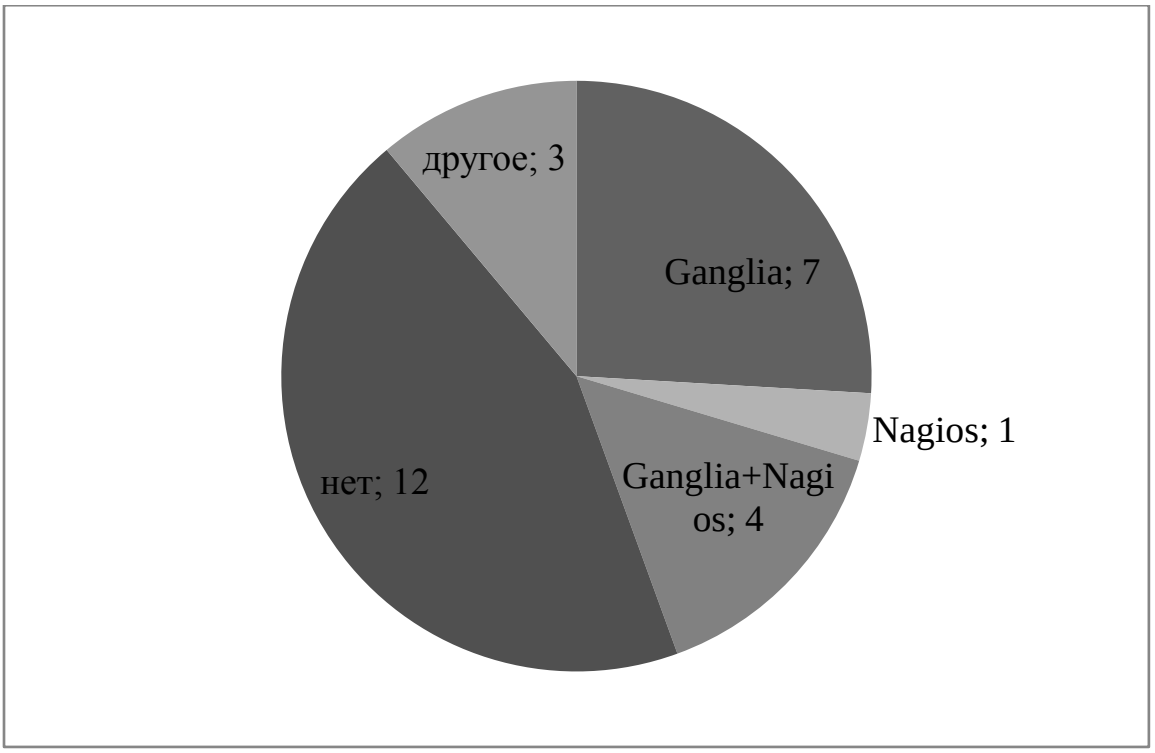
Система мониторинга состоит из двух компонентов: набора программного обеспечения для контроля локальной инфраструктуры и центрального сервиса мониторинга активности виртуальной организации на облаке Tier 3 сайтов. Локальный мониторинг должен отслеживать нагрузку на узлы кластера и сеть, состояние batch-системы, системы запуска задач (PROOF или Athena), системы хранения данных. Система глобального мониторинга предназначена для сбора суммарной информации с систем локального мониторинга и их представления аналогично существующим средствам мониторинга в рамках виртуальной организации.

Архитектура

Архитектура локального мониторинга

Для мониторинга локальной вычислительной инфраструктуры может быть использован ряд пакетов. По результатам опроса[4], наиболее популярным среди Tier 3 сайтов, уже имеющих систему мониторинга, оказался пакет Ganglia. Распределение популярности систем мониторинга показано на диаграмме 1.

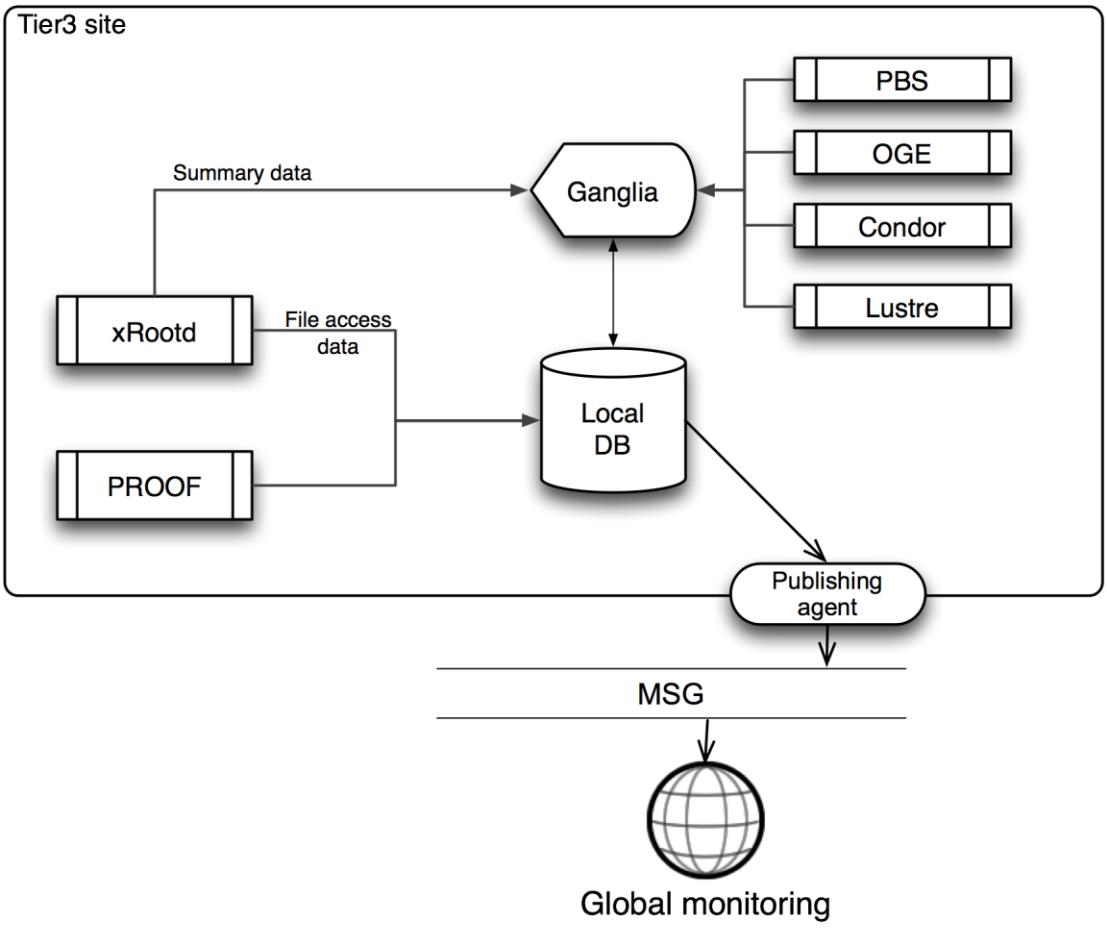
Диаграмма 1. Системы мониторинга Tier 3 сайтов.



В связи с большей распространённостью, система локального мониторинга основана на пакете Ganglia, используемом для сбора и представления отслеживаемых метрик для нужд локального администрирования. Схема потоков данных в системе локального мониторинга представлена на диаграмме 2. Помимо метрик по умолчанию, в Ganglia собирается информация о работе batch-системы и системы хранения. В связи с тем, что на некоторых Tier 3 сайтах для мониторинга уже используется Nagios,

вариант локального мониторинга на его основе для наиболее популярных конфигураций Tier 3 сайтов также должен быть разработан.

Диаграмма 2. Архитектура локального мониторинга



Помимо сбора подробных метрик для представления администраторам, компонента локального мониторинга также отправляет необходимые данные глобальной системе мониторинга. В число этих метрик входит информация о задачах анализа, запускаемых на PROOF и данные о доступе к файлам из системы хранения для определения популярности наборов данных.

Ganglia

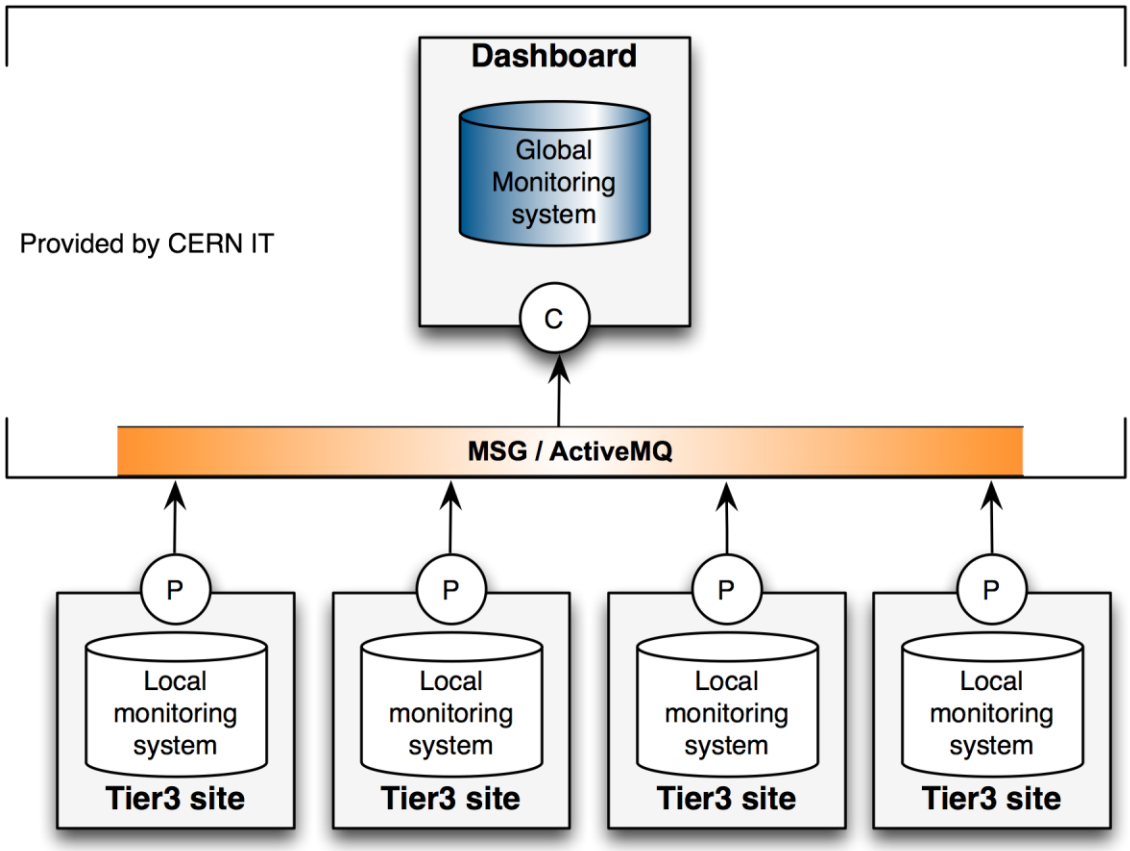
Ganglia - масштабируемая распределенная система мониторинга кластеров и облачных систем. Позволяет отслеживать в реальном времени широкий набор метрик, в том числе нагрузку на процессор, сетевые соединения, память машины, степень заполнения жесткого диска, и просматривать

историю изменения данных параметров. Также предусмотрена возможность добавления пользовательских метрик с помощью подключаемых модулей, либо утилиты gmetric. [5]

Архитектура глобального мониторинга

Система глобального мониторинга представляет собранные данные с помощью Dashboard – средства мониторинга, используемого всеми виртуальными организациями WLCG для отслеживания большинства глобальных процессов хранения, обработки и анализа данных. Для доставки информации, отправляемой локальным мониторингом, используется система доставки сообщений MSG (Message system for GRID) построенная на технологии ActiveMQ. Схема работы глобального мониторинга изображена на диаграмме 3.

Диаграмма 3. Архитектура глобального мониторинга.



Dashboard

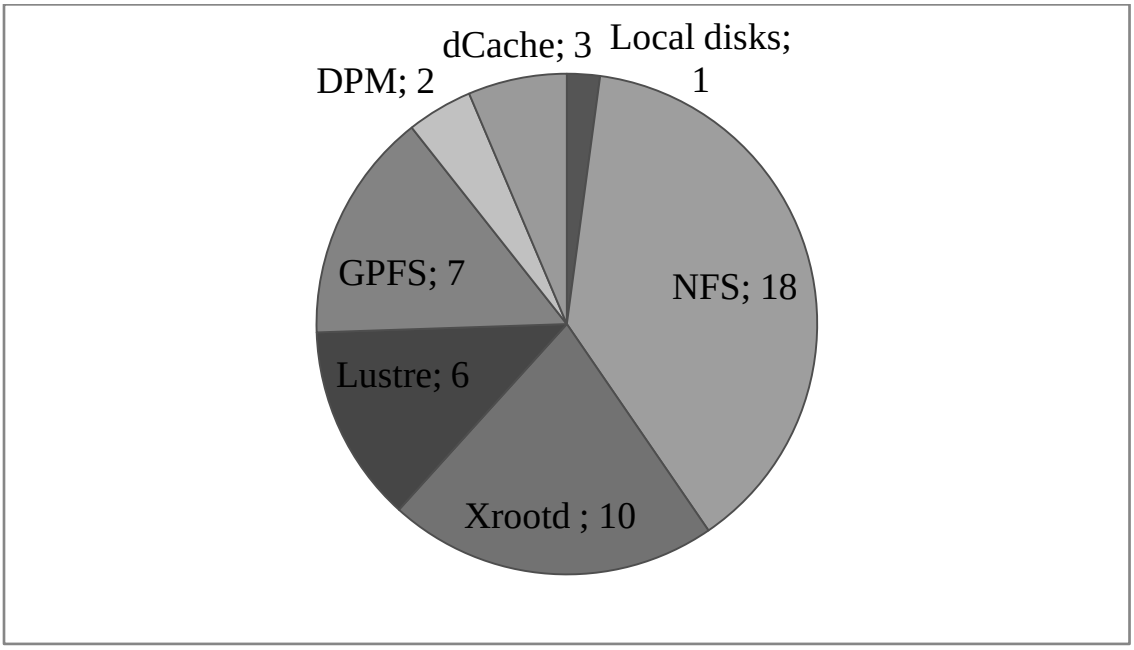
Единая система мониторинга распределённых систем виртуальных организаций WLCG. Собирает информацию о различных областях деятельности эксперимента: выполнение задач, управление данными, тесты передачи данных, мониторинг эффективности работы сайтов. Dashboard поддерживает различные варианты промежуточного программного обеспечения грид: OSG, LCG, gLite. Четыре эксперимента на БАК используют Dashboard: ATLAS, ALICE, CMS, LHCb.

Реализация

Тестовая инфраструктура

Одним из направляющих факторов данной работы были результаты опроса [4], проведённого среди Tier 3 сайтов ATLAS. Распределение используемых систем хранения данных показано на диаграмме 4.

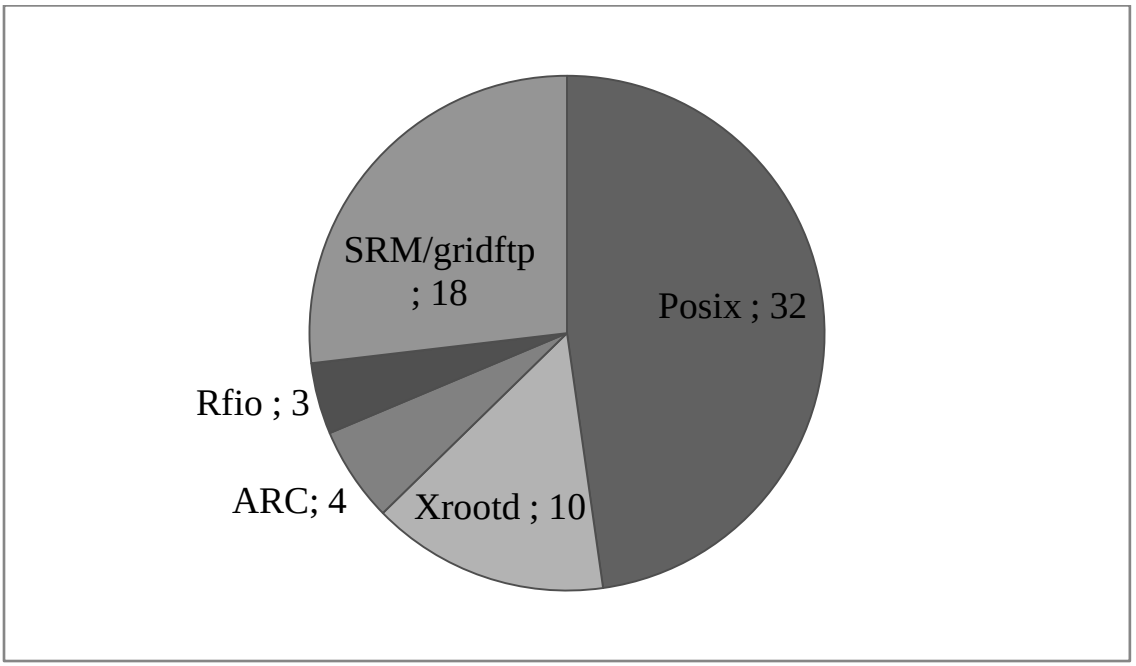
Диаграмма 4. Системы хранения.



NFS (Network File System, сетевая файловая система) не требует, в связи с простым устройством, мониторинга. При работе над системой локального мониторинга приоритетом была наиболее популярная XRootD.

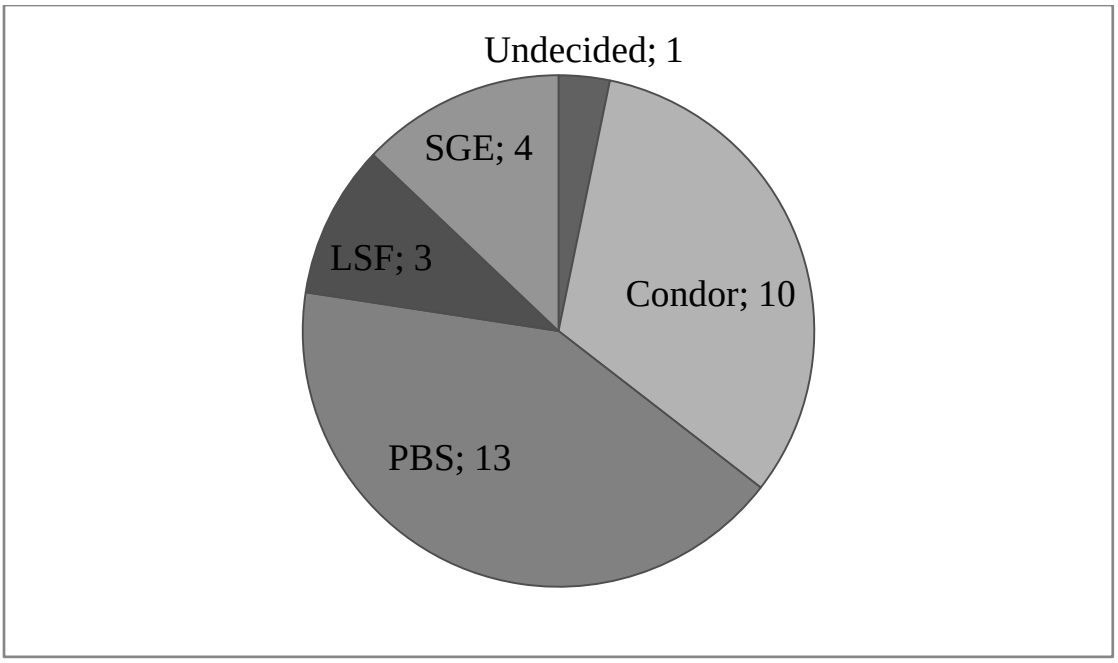
Распределение используемых протоколов доступа к системе хранения показано на диаграмме 5.

Диаграмма 5. Доступ к системам хранения.



Распределение используемых batch-систем показано на диаграмме 6.

Диаграмма 6. Batch-системы.



Из результатов опроса можно сделать вывод, что система локального мониторинга должна работать с разнообразными вариантами конфигурациями наблюдаемого кластера.

Важной частью работы над локальным мониторингом Tier 3 являлось тестирование существующих средств мониторинга различных подсистем

сайта. Эта задача включает в себя создание тестового кластера для каждой batch-системы и системы хранения, установку на него Gangliaили Nagios, настройку мониторинга рассматриваемой системы, установку необходимых для этого расширений системы мониторинга. При этом проверяется в первую очередь работоспособность системы, а производительность тестового кластера не является критическим параметром.

Принимая вышесказанное во внимание, тестовая инфраструктура может быть основана на виртуальных машинах. Это позволяет легко создавать тестовые кластеры по мере надобности и максимально эффективно использовать имеющиеся вычислительные ресурсы. В качестве технологии виртуализации была выбрана OpenVZ.

Тестовая инфраструктура, созданная для разработки Tier 3 мониторинга, состоит из 5 виртуальных кластеров по 3 сервера в каждом. На кластеры установлены PBS, PROOF, Condor, OGE и XRootD. Отдельно создан виртуальный сервер Nagios и машина для разработки. Вся тестовая инфраструктура размещается на одном физическом сервере.

Разработка элементов локального мониторинга

В рамках участия в разработке системы мониторинга Tier 3 сайтов T3Mon, были изучены средства локального мониторинга инфраструктуры и нескольких возможных систем хранения и управления кластером. На основе полученного опыта были составлены рекомендации по установке и настройке данных средств мониторинга. Для интеграции пакета мониторинга и наблюдаемых систем была разработана необходимая программная компонента.

Были изучены средства настройки Ganglia и проведена установка системы Ganglia на два виртуальных кластера по три сервера. Процесс инсталляции был тщательно документирован, в дальнейшем эта информация была

оформлена в виде инструкции по настройке Ganglia для администраторов Tier3 сайтов.

Также изучалась система управления кластером PBS. На основе выбранной схемы организации кластера на один из виртуальных кластеров была установлена система пакетной обработки данных PBS из трёх рабочих узлов, один из которых также исполнял роль сервера и планировщика. В качестве локального средства мониторинга batch-системы было проведено внедрение пакета jobmonarch. В процессе была реализована система запуска тестовых задач для проверки работоспособности полученной программной среды.

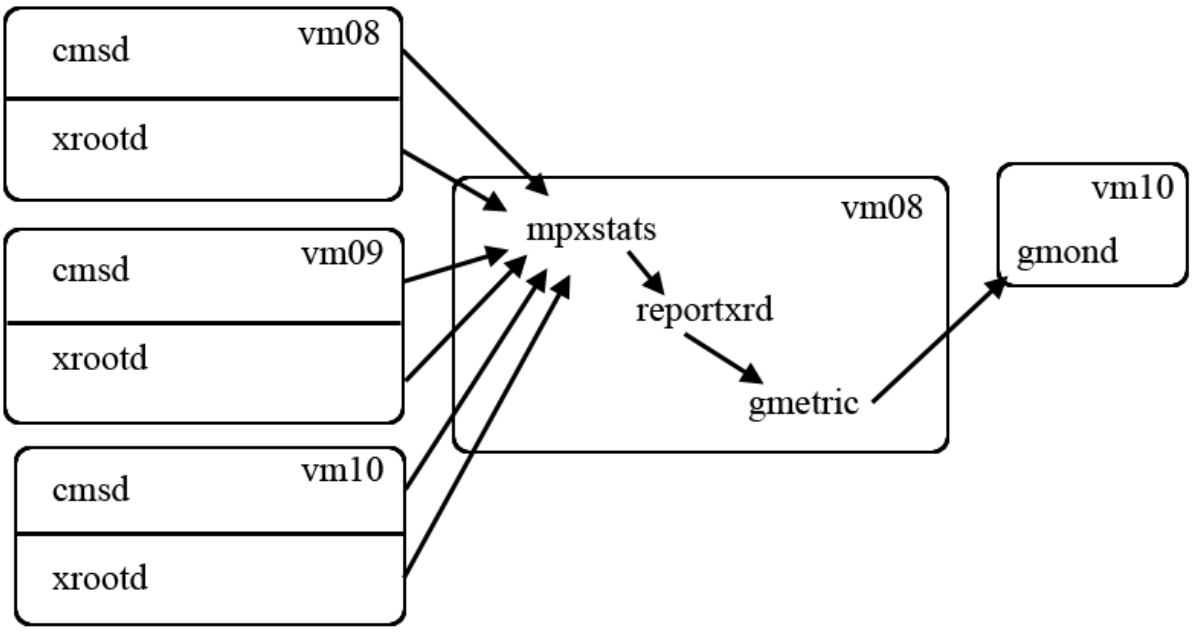
На второй виртуальный кластер была установлена система хранения XRootD. Один из узлов выступал в роли xrootd менеджера, два – в роли серверов с данными. При настройке серверов была запущена отправка детального потока данных и кратких отчётов системы мониторинга. Детальный поток в дальнейшем будет использоваться для отслеживания популярности хранимых наборов данных. Кроме того, была проведена работа по интеграции потока кратких отчётов в систему Ganglia с представлением полученных данных в виде графиков.

На основе полученного опыта была создана инструкция для локальных администраторов по настройке мониторинга XRootD с помощью Ganglia. Её применимость на практике проверили администраторы ИФВЭ в Протвино, настроив с её помощью мониторинг XRootD сервера.

МониторингXRootD

Для интеграции информации о статусе серверов XRootD в Ganglia была реализована схема формирования метрик, показанная на диаграмме 7. В связи с количеством метрик, публикуемых XRootD, было принято решение исследовать возможность их разделения на логические группы. Для приведения данных к соответствующему формату, разделения их на группы и отправки в Ganglia самостоятельно создана программа reportxrd.

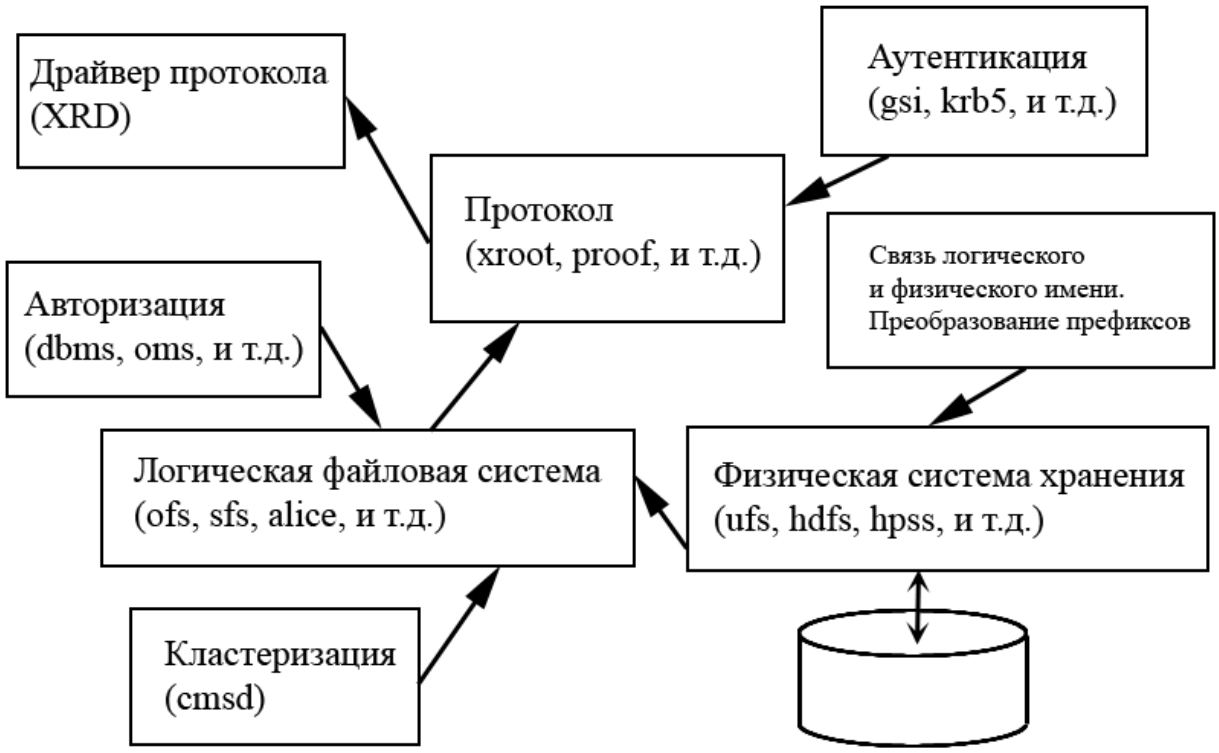
Диаграмма 7. Потоки данных в системе мониторинга XRootD



XRootD – сетевой протокол доступа к данным. Благодаря модульной структуре, XRootD поддерживает аутентификацию и авторизацию, распределение нагрузки по кластеру любого размера, различные типы хранилищ данных, интеграцию с другими системами хранения и протоколами доступа к данным.

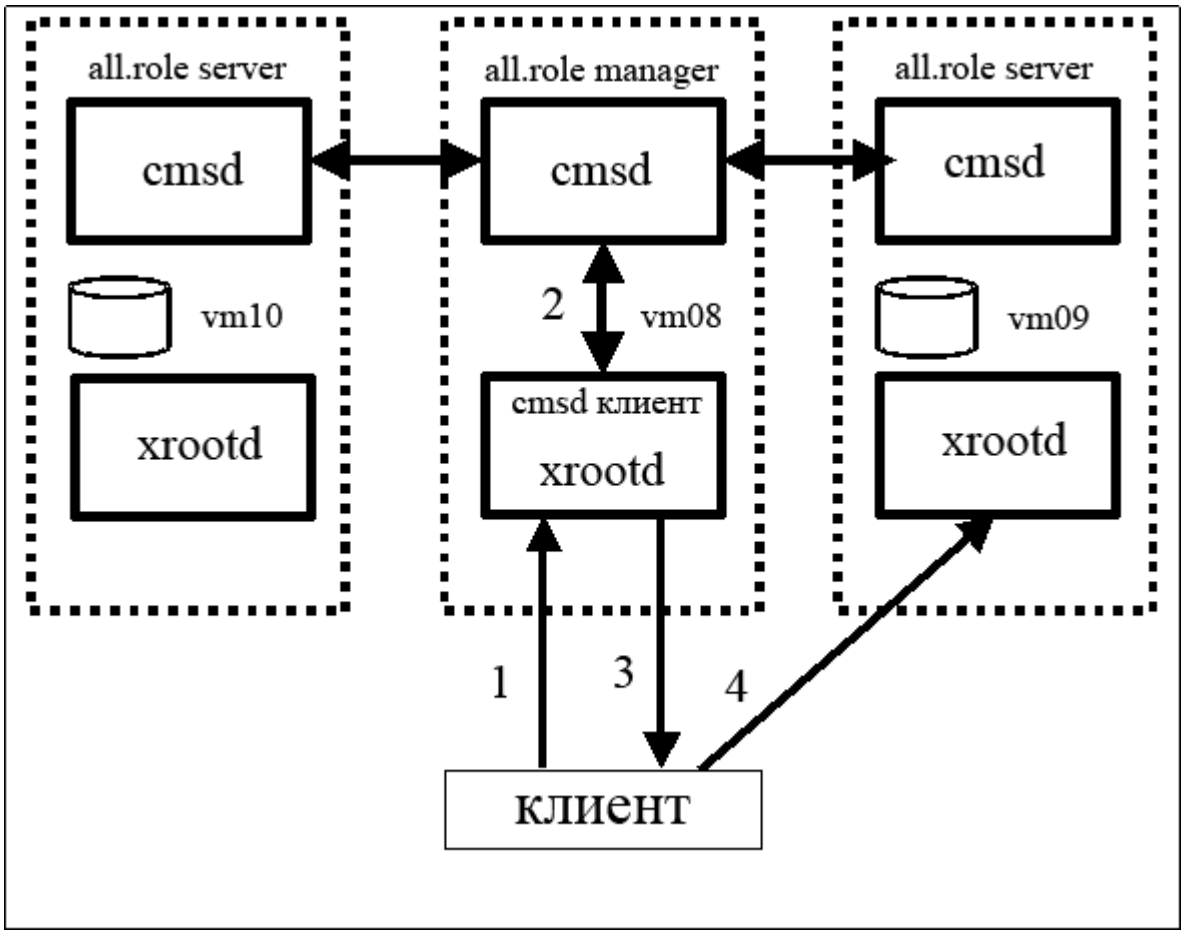
Благодаря широким возможностям настройки и расширения, xrootd используется системами root и proof, поддерживает доступ в формате POSIX, монтирование с помощью FUSE, доступ из Грид с помощью SRM, globus-url-copy, gridFTP и другие варианты доступа к данным. Схема соединения модулей XRootD показана на диаграмме 8.

Диаграмма 8. Модульная структура XRootD



При работе была изучена работа схемы кластеризации xrootd серверов. Принцип работы xrootd кластера следующий: клиент запрашивает у xrootd менеджера файл, запрос передаётся cmsd менеджеру. Cmsd менеджер опрашивает подчинённые ему cmsd сервера и через xrootd менеджер перенаправляет клиента на xrootd сервер, где файл имеется в наличии. Диаграмма 9 изображает процесс обращения клиента к XRootD кластеру.

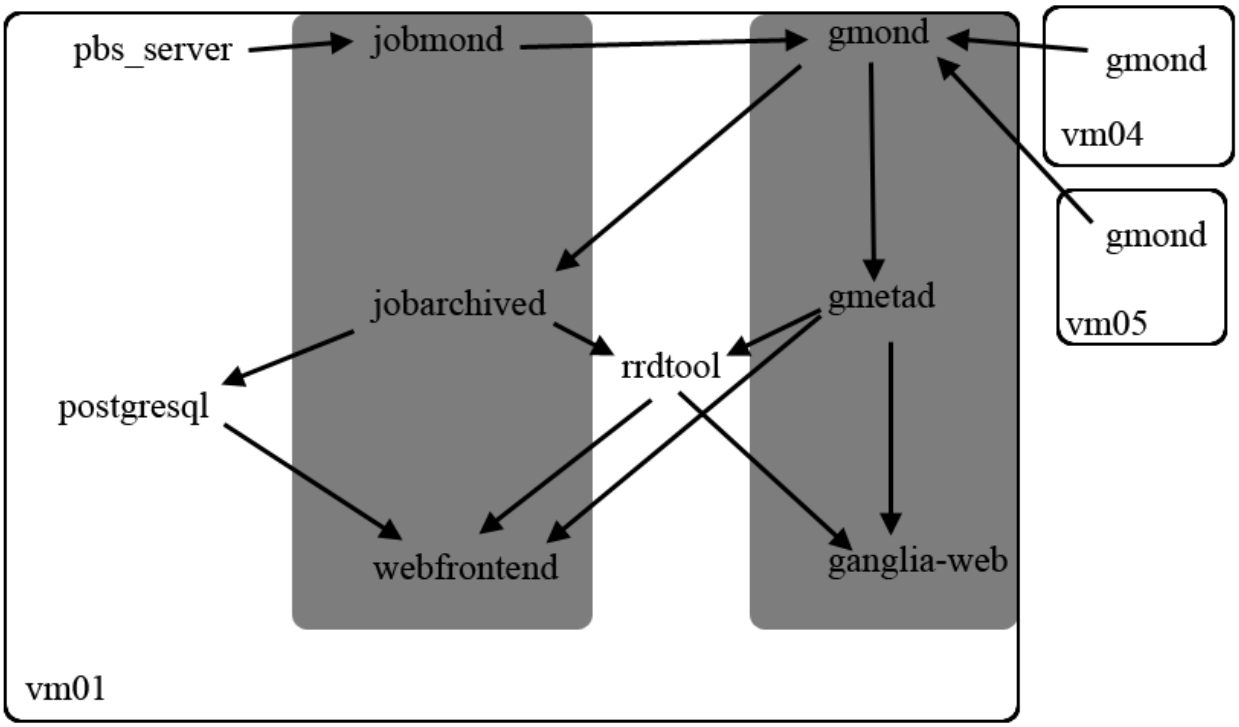
Диаграмма 9. Принцип работы XRootD кластера



Мониторинг PBS

Для мониторинга системы управления кластером PBS, было изучено и внедрено дополнение к Ganglia для работы с batch-системами jobmonarch. Схема потоков данных при интеграции Ganglia с Jobmonarch и PBS изображена на диаграмме 10.

Диаграмма 10. Схема работы Jobmonarch



Jobmonarch – это дополнение к системе мониторинга Ganglia, которое позволяет отслеживать параметры задач и очередей в системах управления локальными ресурсами и обеспечивает графическое представление нагрузки на кластер. Он также предоставляет возможность архивировать информацию о задачах для дальнейшего анализа возможных проблем. [6]

Реализация других компонентов

В 2011 году, мониторинг сайтов Tier3 выделен в специальное направление деятельности в ATLAS Distributed Computing. Данные работы проводятся в тесной коллаборации с ЛИТ ОИЯИ. Над реализацией данного проекта со стороны ОИЯИ работает команда специалистов. Разработкой программного обеспечения занимаются Артём Петросян, Данила Олейник, Сергей Белов и Владимир Васильев. За изучение возможности применения существующих решений отвечают Николай Кутовский, Иван Кадочников и Анатолий Якшов. В качестве специалиста по системе PROOF участвует Люция Валова. Разработкой и администрированием локального мониторинга с помощью Nagios занимается Павел Дмитриенко.

Со стороны CERN работы курируются Юлией Андреевой – руководителем группы разработки и поддержки Dashboard в CERN IT.

Выводы

В ходе работы на тестовом кластере была отработана методика настройки Ganglia для мониторинга локальной инфраструктуры, PBS и XRootD. По результатам была составлена документация, позволяющая применить этот опыт для настройки мониторинга реальных Tier 3 сайтов. Результаты работы востребованы и в ближайшее время будут проходить дальнейшее тестирование в ЛЯП ОИЯИ, BNL и на других сайтах. На рабочем совещании ATLAS Technical Interchange Meeting был сделан доклад о результатах проведённой работы.

Работы над проектом мониторинга Tier 3 сайтов продолжаются. Прототип комплекса средств локального мониторинга будет закончен в ближайшем будущем. Продолжается разработка компонентов системы глобального мониторинга. Результат данной работы имеет большую важность для коллаборации ATLAS.

Список литературы

[1] *Raymond Brock, Doug Benjamin, Gustaaf Brooijmans, Sergei Chekanov, Jim Cochran, Michael Ernst, Amir Farbin, Marco Mambelli, Bruce Mellado, Mark Neubauer, Flera Rizatdinova, Paul Tipton, Gordon Watts* U.S. ATLAS Tier 3 Task Force. – 2009

[2] *Benjamin D.* AGIS: Tier 3 needs [презентация] // ATLAS Computing Technical Interchange Meeting. – 2011

[3] *Andreeva J.* Tier-3 Monitoring Software Suite (T3MON) proposal . – http://cdsweb.cern.ch/record/1336119

[4] *Benjamin D.* Atlas Tier 3 survey [презентация]// ATLAS Software & Computing Workshop. – 2010. – https://indico.cern.ch/getFile.py/access?contribId=44&sessionId=3&resId=1&materialId=slides&confId=76896

[5] Ganglia [сайт]. – http://ganglia.info/

[6] Jobmonarch [сайт]. – https://subtrac.sara.nl/oss/jobmonarch/