**КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2015 Т. 7 № 6 С. 1343−1349**

Ки&М

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ

УДК: 004.75

Современное использование сетевой инфраструктуры в системе обработки задач коллаборации ATLAS

А. Ш. Петросян

Лаборатория информационных технологий, Объединенный институт ядерных исследований,
Россия, 141980, г. Дубна, ул. Жолио-Кюри, д. 6

E-mail: artem.petrosyan@jinr.ru

Получено 30 апреля 2015 г.

Важнейшим компонентом распределенной вычислительной системы является сетевая инфраструктура. Несмотря на то что сеть составляет основу такого рода систем, она часто является незаметным партнером для систем хранения и вычислительных ресурсов. Мы предлагаем интегрировать сетевой элемент напрямую в распределенные системы через уровень управления нагрузками. Для такого подхода имеется достаточно предпосылок. Так как сложность и требования к распределенным системам растут, очень важно использовать имеющуюся инфраструктуру эффективно. Например, одни могут использовать измерения качества сетевых соединений в механизмах принятия решений в системе управления задачами. Кроме того, новейшие технологии позволяют другим задавать сетевую конфигурацию программно, например используя ПКС — программно-конфигурируемые сети. Мы опишем, как эти методы используются в системе управления задачами PanDA, применяемой коллаборацией ATLAS.

Ключевые слова: ATLAS, PanDA, распределенные вычисления, системы управления задачами, механизмы принятия решений, сеть, измерения сетевой производительности, программно-конфигурируемые сети

The New Use of Network Element in ATLAS Workload Management System

A. Sh. Petrosyan

*Laboratory of Information Technologies, Joint Institute for Nuclear Research, 6 Joliot Curie,
Dubna, 141980, Russia*

A crucial component of distributed computing systems is network infrastructure. While networking forms the backbone of such systems, it is often the invisible partner to storage and computing resources. We propose to integrate Network Elements directly into distributed systems through the workload management layer. There are many reasons for this approach. As the complexity and demand for distributed systems grow, it is important to use existing infrastructure efficiently. For example, one could use network performance measurements in the decision making mechanisms of workload management systems. New advanced technologies allow one to programmatically define network configuration, for example SDN — Software Defined Networks. We will describe how these methods are being used within the PanDA workload management system of the ATLAS collaboration.

Keywords: ATLAS, PanDA, distributed computing, workload management system, decision making mechanisms, network, network performance measurements, software defined networks

Citation: *Computer Research and Modeling*, 2015, vol. 7, no. 6, pp. 1343–1349 (Russian).

Введение

PanDA — система управления заданиями обработки данных, реализующая единую очередь заданий как для пользовательского анализа, так и для симуляции и массовой обработки заданий [PanDA, 2015; Evolution of the ATLAS…, 2015]. Все задания отправляются в обработку через один и тот же интерфейс. Это позволяет скрыть всю сложность процессов и инфраструктуры, которая стоит за отправкой задания в обработку, и сделать работу распределенной средой обработки простой для пользователя. В дальнейшем задание, отправленное в систему, обрабатывается в зависимости от условий, необходимых данных и ресурсов, на одном из 170 вычислительных сайтов, подключенных к обработке заданий эксперимента ATLAS [ATLAS, 2015].

PanDA разрабатывалась как система, предназначенная для функционировании в режиме больших нагрузок, обеспечивающая обработку данных в распределенной среде, и используется коллаборацией ATLAS с 2005 года. С 2007 года система стала единственной средой запуска и массовой обработки данных коллаборации. С 2008 года система адаптирована к запуску задания пользовательского анализа. В настоящее время через систему проходит 200–300 тысяч заданий ежедневно.

PanDA является системой, автоматически распределяющей задания по сайтам. Как любая автоматическая система, обрабатывающая большое количество заданий одновременно, PanDA спроектирована так, чтобы вмешательство человека в работу системы было минимальным. Человек задает общие условия, в которых будет функционировать система, а далее процесс обработки идет в автоматическом режиме. Так как PanDA работает в режиме высокой нагрузки годами, то, при анализе истории обработки заданий и условий, которые приводили к возникновению или исчезновению различных проблем, становится возможным, изменяя параметры функционирования системы, отслеживать то, как они влияют на производительность системы.

Важнейшим элементом инфраструктуры, обеспечивающим пересылку больших объемов данных, является сеть. В течение долгого времени сеть воспринималась как элемент, функционирующий «как есть», без попыток активного управления и без использования текущего состояния соединений между сайтами при принятии решений на уровне системы обработки заданий. Тем не менее очевидно, что в автоматизированных распределенных системах необходимо учитывать состояние всех компонентов. В PanDA учитывать состояние сети чрезвычайно важно потому, что PanDA автоматически выбирает сайт, на котором будут выполняться задания, используя при этом данные о наличии на сайте достаточного количества доступных процессоров и объема дискового пространства. При этом параметры сетевых соединений указываются при конфигурации системы, но, будучи указанными в конфигурации, используются без учета текущего состояния связей между сайтами.

Источники сетевой информации

Для включения текущего состояния сетевых соединений в логику принятия решений системы PanDA необходимо обеспечить хранение актуальных значений в информационной системе SchedConfig. Как и любая большая распределенная система, система управления данными коллаборации ATLAS оснащена средствами мониторинга инфраструктуры. Сразу несколько сервисов постоянно тестируют состояние сетевых соединений между сайтами, на которых происходит обработка заданий. Так как при передаче данных используется несколько протоколов, две системы измеряют скорость передачи файлов различного размера с каждого на каждый сайт инфраструктуры. Таким образом производят измерения сервисы XRootD и Sonar. По протоколу XrootD тестируются соединения между сайтами, входящими в XRootD-федерацию (FAX) [Federated ATLAS …, 2015], а сервис Sonar тестирует состояние соединений между всеми сайтами, входящими в инфраструктуру коллаборации ATLAS. Набор сайтов, входящих в FAX, и набор сайтов, входящих в ATLAS, могут не совпадать. Сервис perfSONAR представляет собой сервис, отслеживающий входящий и исходящий трафик [perfSONAR, 2015]. Сервис

perfSonar размещается на специально выделенной для этого машине, устанавливаемой на сайте. Этот сервис тестирует как сайты, входящие в ATLAS, так и сайты, входящие в FAX.

Хранилища сетевой информации

Каждый из сервисов, производящих оценку качества сетевых соединений, FAX, Sonar и perfSONAR, имеет свое хранилище. Однако выборочные данные, предоставляемые этими сервисами, собираются также и в единое хранилище: Site Status Board (SSB), отображающее различную информацию о состоянии сервисов сайтов, входящих в инфраструктуру коллаборации [Site Status Board, 2015]. В SSB хранится информация, необходимая для оперативного управления и оценки состояния сервисов. Также в SSB хранится история изменений значений метрик о работе сервисов и сайтов. Данные в SSB обновляются постоянно и независимо каждым из сервисом, поставляющим информацию в систему. Из SSB самые свежие данные о сетевых соединениях раз в час переносятся в AGIS (ATLAS Grid Information System), где хранится вся информация об инфраструктуре коллаборации ATLAS, а оттуда — во внутреннюю базу данных PanDA SchedConfig, что позволяет использовать их в дальнейшем в качестве параметров при запуске заданий и в других процессах принятия решений. Интерфейс Site Information Status Board представлен на рис. 1.

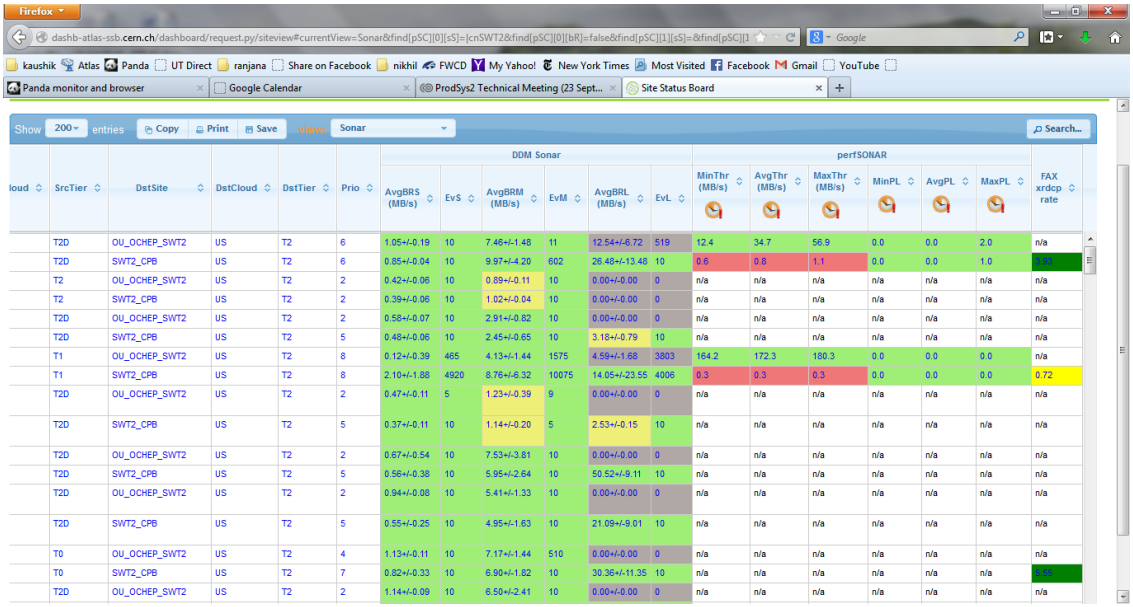

Рис.1. Интерфейс Site Status Board

Использование сетевых метрик при планировании заданий

Работа, направленная на включение и использование актуального состояния сетевой инфраструктуры в процессах автоматического принятия решений, была начата во второй половине 2013 года в рамках проекта BigPanDA [Next Generation …, 2015], основными целями которого были расширение списка проектов, в которых применяется система PanDA, и дополнение системы модулями дистрибуции и более тесной работы с сетевой инфраструктурой. Таким образом, в наборе систем PanDA появилась новая подсистема: Intelligent Networking [Intelligent Networking, 2015], в свою очередь состоящая из нескольких компонентов:

1) сервиса, регулярно доставляющего необходимые для принятия решений данные из внешних источников, таких как тесты FAX, Sonar и perfSONAR, в конфигурационную базу данных SchedConfig;

2) модулей, использующих данные о состоянии сетевых соединений между сайтами в принятии решений о распределении данных и заданий по сайтам инфраструктуры;
3) веб-интерфейса, отображающего параметры сетевых соединений между вычислительными сайтами инфраструктуры, и историю изменений сетевой конфигурации, используемой в процессе принятия решений.

В качестве сценариев использования сетевой информации выбраны наиболее очевидные.

1. В случае если на сайте, куда отправлено задание, время ожидания обработки превышает время, которое понадобится, чтобы переместить эти данные на ближайший (лучший с точки зрения сетевого соединения) и не загруженный сайт, то задание пересылается на такой сайт. В данном механизме используются данные измерения качества соединений, производимые сервисом Sonar.
2. В случае если сайт, куда отправлены задание, сильно загружен и близко (в терминах сети) находится сайт, поддерживающий удаленное чтение по протоколу XRootD, то задание будет перемещена на сайт, а необходимые для выполнения задания данные будут читаться с первого сайта. В данном механизме используются данные измерения качества соединений по протоколу XRootD.
3. Выбор сайтов уровня Tier 1, откуда сайт T2D получает данные для обработки, изменяется на основании качества соединений между ними. Этот механизм использует данные измерений, производимые сервисом Sonar.
4. Изменение статуса сайта уровня Tier 2, назначение его сайтом T2D, участвующим в получении данных с различных облаков на основании оценки качества сетевых соединений. Если сайт уровня Tier 2 в течение длительного времени демонстрирует хорошие показатели при получении данных с сайтов уровня Tier 1, то ему присваивается статус T2D. Также реализован и обратный механизм автоматического лишения сайта статуса сайта, участвующего в обработке данных, поступающих с нескольких сайтов уровня Tier 1: с T2D на T2. В данном механизме используются данные измерений, производимые сервисом Sonar и perfSONAR.
5. При запуске задания набор сайтов, откуда будут получены данные, определяется динамически (вводится понятие динамических облаков). В данном механизме используются данные измерений, производимые сервисом Sonar.
6. Использование программно-конфигурируемых сетей для запроса канала определенной пропускной способности, что позволяет уменьшить время копирования при передаче большого объема данных.

В настоящее время реализованы первые четыре механизма. Каждый из них выполнен в виде независимого сервиса, запускаемого автоматически, выполняющего свою часть работы и не связанного с остальными. В каждом определяется и учитывается период времени, в течение которого полученные измерения сетевых соединений сохраняют актуальность и могут быть использованы, ведь в случае системы, работающей в режиме реального времени, возможны ситуации, когда данные, на основании которых необходимо принять решение, могут оказаться просроченными. Значения скорости передачи, частоты обновления, количества сайтов для обновления также являются входными параметрами системы.

Первые результаты

Включение сетевых метрик в процессы принятия решений PanDA показали достаточно позитивные результаты.

Различные части цепочки доставки и работы с сетевыми метриками были реализованы в различное время. Сервис доставки метрик в базу данных SchedConfig работает с осени 2013 года. Сервис выбора лучших сайтов для перемещения задания и для удаленного чтения работает с весны 2014 года. Сервис рекомендации и автоматического обновления связей T1- и T2D-сайтов работает с осени 2014 года. Веб-интерфейс постоянно развивается и работает с осени 2013 года.

Задания, использующие протокол XRootD для чтения данных с удаленных сайтов, в разное время составляют до 17 % от общего числа выполняемых заданий пользовательского анализа и значительно снизили среднее время ожидания задания на сайте в очереди до начала выполнения. Кроме этого, данный механизм, перемещая задания с перегруженных сайтов на недогруженные, позволяет более равномерно загружать имеющиеся вычислительные ресурсы. Статистика выполнения заданий и сравнение количества заданий, используемых и не использующих механизмы чтения с удаленных сайтов, представлены на рис. 2 и рис. 3.

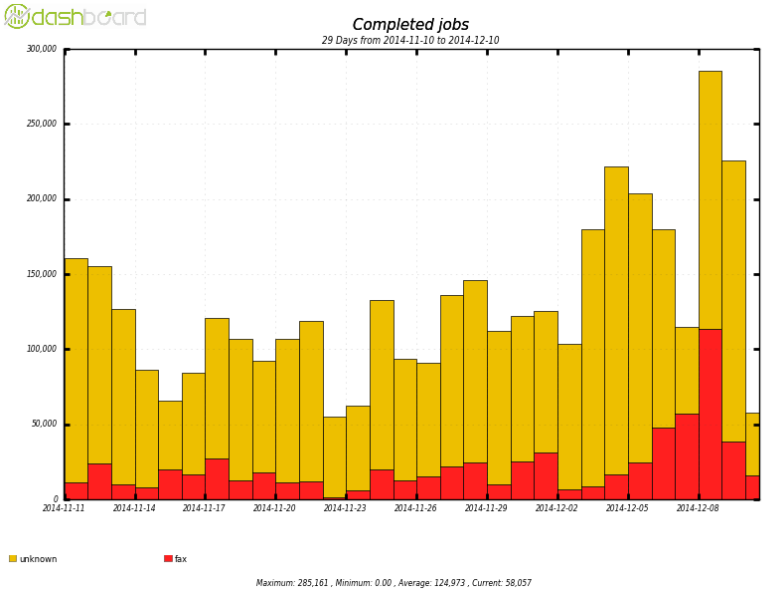

Рис. 2. Количество завершенных заданий, работающих с локальными данными или использующих механизмы чтения с удаленных сайтов

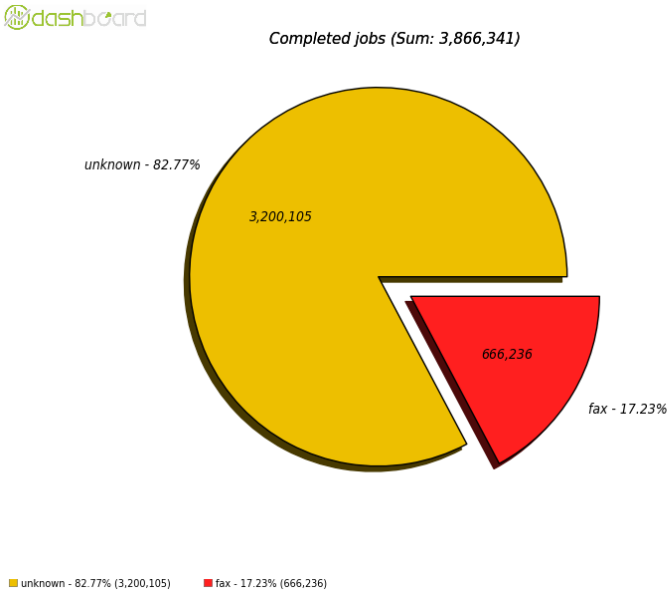

Рис. 3. Количество завершенных пользовательских заданий, использующих локальные и удаленные данные

Механизм автоматического выбора сайтов T1 для T2D-сайтов находится в режиме тестовой эксплуатации и активирован для ограниченного списка сайтов, расположенных в США. Обновление происходит раз в неделю. Для остальных сайтов построен механизм, предлагающий изменить значение поля, задающего, с каких сайтов уровня Tier 1 T2D-сайт получает дан-

ные, но не производящий непосредственного обновления полей в базе данных. Включение данного механизма рекомендации позволило освободить от оперативной работы, связанной с наблюдением за качеством сетевых соединений между сайтами, менеджеров, занимающихся выбором и обновлением связей T2D-сайтов.

Для облегчения принятия решений построен графический интерфейс, на котором отображается история изменений связей сайтов (реальная и рекомендуемая). Пример информации, представляемой данным интерфейсом, показан на рис. 4.

Multicloud statistics for queues on SWT2_CPB

	Auto Multicloud Update	Multicloud Append	Current	History of suggested
ANALY_SWT2_CPB-pbs	OFF		None	
SWT2_CPB-pbs	ON		NL,CA,FR	2014-12-08: NL,CA,IT 2014-12-01: NL,IT,FR 2014-11-24: ,ND,CERN,DE 2014-11-17: CA,FR,NL 2014-11-10: FR,NL,IT
SWT2_CPB_Install	OFF		None	
SWT2_CPB_MCORE-pbs	OFF		CA,FR,NL,TW,CA,DE,ND,IT,US,UK,ES	

Рис. 4. Отображение истории обновления поля multicloud production

В настоящее время реализовано три интерфейса мониторинга:

1) предоставляющий данные о метриках, присутствующих в базе данных SchedConfig и использующихся при принятии решений [Матрица … двумя сайтами, 2015];
2) отображающий матрицу связей между сайтами инфраструктуры, лучшие сайты для каждого из сайтов и историю обновлений T2D-сайтов [Матрица … лучших связей, 2015];
3) отображающий историю выполнения заданий, использующих механизмы удаленного чтения [Мониторинг заданий…, 2015].

Заключение

Десятилетие накопления опыта управления процессами в распределенной среде из более чем 150 сайтов, объединенных для обработки данных большого физического эксперимента, привело к появлению уникальной системы, способной эффективно справляться с огромными потоками данных, функционируя в автоматическом режиме. Один из важнейших компонентов системы заключается в том, что сеть стала применяться как элемент, который используется другими подсистемами «как есть» и играет важнейшую роль в процессе принятия решений самой системы. Важную роль в данном процессе сыграла реализация и работа сервисов, оценивающих актуальное состояние связей между сайтами и построение интерфейсов мониторинга результатов этих измерений. В дальнейшем, после оснащения вычислительных центров компонентами активного управления сетевыми соединениями, такими как резервирование каналов и программно-управляемые сети и включение возможностей управления этими компонентами в соответствующие модули системы, процесс интеграции только усилится, что позволит еще более эффективно использовать возможности сети при распределении заданий.

Список литературы

Матрица сайтов инфраструктуры с указанием лучших связей между каждыми двумя сайтами. http://aipanda021.cern.ch/networking/t1tot2d_matrix/, 2015.

Матрица сайтов инфраструктуры с указанием протоколов, связей и времени обновления. http://aipanda021.cern.ch/networking/network_links_between_sites/, 2015.

Мониторинг заданий, фокусирующийся на локальном или удаленном чтении при помощи протокола XRootD. http://goo.gl/zJy09t, 2015.

Эксперимент ATLAS. http://atlas.ch/, 2015.

Federated ATLAS storage systems using XRootD.

https://twiki.cern.ch/twiki/bin/view/AtlasComputing/AtlasXrootdSystems, 2015.

Intelligent Networking. https://twiki.cern.ch/twiki/bin/view/PanDA/IntelligentNetworking, 2015.

Klimentov A. et al. Next generation workload management system for big data on heterogeneous distributed computing // Journal of Physics Conference Series. — 2015. — Vol. 608. — http://inspirehep. net/record/1372988/.

Maeno T. et al. Evolution of the ATLAS PanDA workload management system for exascale computational science // Journal of Physics Conference Series. — 2014. — Vol. 513. — http://inspirehep.net/record/1302031/.

PanDA. https://twiki.cern.ch/twiki/bin/view/PanDA/PanDA, 2015.

perfSONAR, performance focused Service Oriented Network monitoring ARchitecture. http://www.perfsonar.net/, 2015.

Site Status Board. http://dashb-atlas-ssb.cern.ch/dashboard/request.py/siteview?view=Network%20measurements#currentView=Network%2520measurements&highlight=false, 2015.