Интеграция суперкомпьютера НИЦ «Курчатовский институт» с центром Грид первого уровня[1]

А.А. Климентов, Р.Ю. Машинистов, А.М. Новиков, А.А. Пойда, Е.А. Рябинкин, И.С. Тертычный

Национальный Исследовательский Центр “Курчатовский Институт”

Эксперименты на Большом Адронном Коллайдере (БАК) находятся в ожидании новых открытий, которые возможно будут получены в 2015 году, и связанным с ними вычислениями. Потребность в вычислительных ресурсах для таких процессов как моделирование, обработка и анализ данных, по-видимому, превзойдет возможности, предоставляемые Грид консорциумом БАК. Одновременно с этим объем научных вычислений будут расти. В связи с этим крайне важной задачей становится интеграция дополнительных вычислительных ресурсов, изначально не используемых в модели организации вычислений БАК. Грид-центр первого уровня в НИЦ «Курчатовский институт» является частью инфраструктуры WLCG и будет обрабатывать и хранить до 10% данных экспериментов ATLAS, ALICE, LHCb. Кроме того, Курчатовский институт располагает многофункциональным вычислительным комплексом, включающим суперкомпьютер HPC2 с пиковой производительностью 0.12 ПФлопс. Предоставление даже небольшой части ресурсов суперкомпьютера для вычислений БАК позволит повысить общую производительность Грид-инфраструктуры. В 2014 году мы начали инновационную работу по созданию единого портала, интегрирующего центр Грид первого уровня и суперкомпьютер в НИЦ «Курчатовский институт». Портал предоставит интерфейс для запуска пользовательских задач обработки данных эксперимента ATLAS с использованием общего хранилища данных.

1. Введение

В ходе экспериментов ATLAS [1] и ALICE, проводимых при первом запуске БАК [2], было получено более 30 ПБайт данных, что существенно превышает объемы данных, получаемых в любых других областях науки, даже таких, как геномика и анализ климата. Чтобы решить беспрецедентную проблему обработки такого большого объема данных, в ходе проекта WLCG (Worldwide LHC Computing Grid) [3] была развернута вычислительная Грид-среда.

В 2015 году планируется повысить точность экспериментов, проводимых на БАК, за счет повышения скорости столкновения элементарных частиц. Ожидается, что вычислительные мощности, необходимые для моделирования, обработки и анализа данных превысят текущие мощности инфраструктуры WLCG. Чтобы этого не произошло, требуется привлечение новых вычислительных ресурсов.

Центр Грид первого уровня (Tier-1 центр) в НИЦ “Курчатовский институт” в Москве является частью WLCG и будет обрабатывать и хранить до 10% всех данных, полученных в экспериментах ALICE, ATLAS и LHCb. Кроме того, в Курчатовском институте установлен суперкомпьютер с пиковой производительностью 0,12 ПФлопс. Предоставление даже части ресурсов суперкомпьютера под задачи БАК существенно повысит суммарную производительность Грид-инфраструктуры.

В 2014 году в Курчатовском институте была начата инновационная работа по созданию единого портала, интегрирующего Грид-центр первого уровня и суперкомпьютер в НИЦ «Курчатовский институт». Портал предназначен для запуска пользовательских задач обработки данных эксперимента ATLAS на ресурсах Tier-1 и суперкомпьютера НИЦ “Курчатовский

[1] Данная работа выполнена в рамках мега-гранта правительства РФ, контракт No 14.Z50.31.0024. Мы благодарны нашим коллегам из НИЦ “Курчатовский институт” и ЦЕРН за обсуждение результатов работы, особая благодарность группе разработчиков WMS PanDA, группам МИФИ. Суперкомпьютер в НИЦ “Курчатовский институт” поддерживается как часть центра коллективного пользования (проект RFMEFI62114X0006 при финансовой поддержке Министерства Образования и Науки РФ).

институт” с единой точкой входа и общим хранилищем данных. В качестве базовой технологии мы выбрали систему управления задачами PanDA (Production and Distributed Analysis), хорошо зарекомендовавшую себя в эксперименте ATLAS.

В статье описан разработанный портал, представлены архитектурные и технические решения по его развертыванию в НИЦ “Курчатовский институт” и приведены результаты тестирования и апробации.

2. Система управления потоком заданий PanDA

В качестве основы для разрабатываемой технологии и реализующей его программной системы был выбран подход, используемый в системе управления заданиями PanDA, с 2007 года успешно используемой для обработки, анализа и моделирования данных эксперимента ATLAS, проводимого на БАК.

Программный комплекс PanDA обеспечивает “прозрачность” обработки данных в распределенной вычислительной инфраструктуре. Он предоставляет среду выполнения для широкого диапазона экспериментальных приложений, автоматизирует централизованную обработку данных, обеспечивает анализ данных для десятков групп физиков, поддерживает пользовательский поток операций, обеспечивает единый доступ к распределенным глобальным ресурсам, предоставляет состояние и историю выполняемых операций через интегрированную систему контроля и управляет распределением данных.

Масштабируемость PanDA была продемонстрирована в процессе работы коллаборации ATLAS при быстром росте числа выполняемых заданий во время первого рабочего запуска БАК (2010-2013 гг.). Программный комплекс PanDA был разработан достаточно гибким для адаптации новых технологий обработки и хранения данных и сетевых технологий. Общая архитектура системы PanDA представлена на рисунке 1.

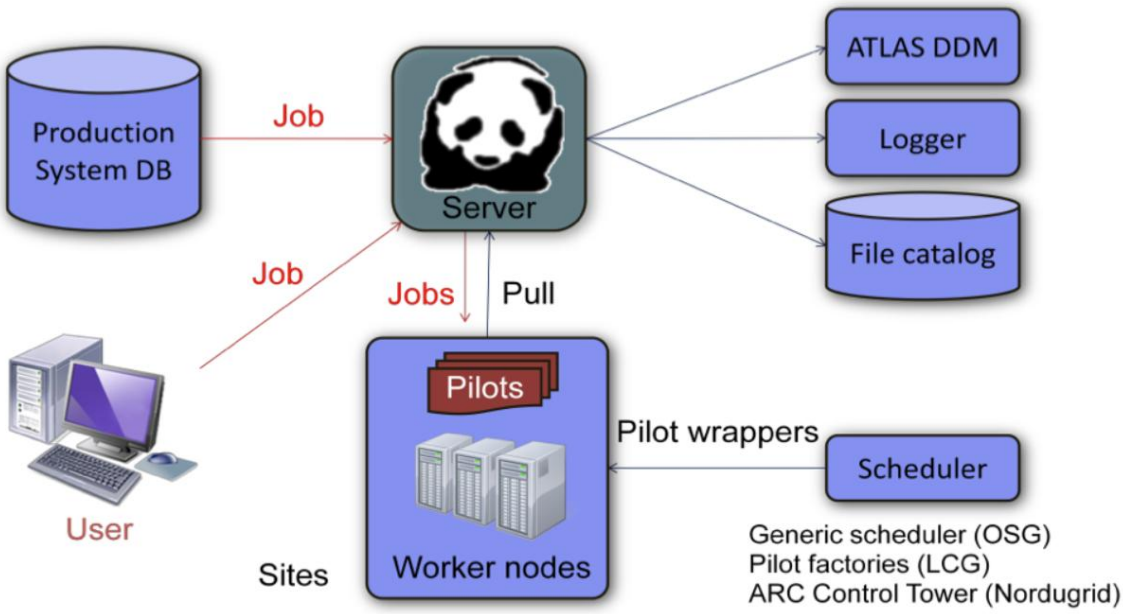

Рис. 1. Архитектура системы вычислений

Основными компонентами системы являются:

- Сервер. Сервер принимает от пользователей задачи и управляет их выполнением: сортирует задачи по разным очередям в зависимости от требований, распределяет их на доступные вычислительные ресурсы (при этом сервер отслеживает, чтобы объемы и состав ресурсов удовлетворял требованиям задачи), отслеживает состояние задачи и ее жизненный цикл, осуществляет дополнительные проверки успешности выполнения задачи, поддерживает метаинформацию о всех активных и завершенных задачах и

т.п.[4]

- Подсистема пилотных заданий. Пилотные задания осуществляют механизм поздней привязки задачи к среде вычислений. Они запускаются на ресурсах, предназначенных для выполнения пользовательских задач, проверяют их состояние, резервируют, собирают информацию и передают на сервер, который в ответ присылает задачу. Пилотное задание инициирует запуск присланной задачи непосредственно на ресурсах и контролирует процесс ее выполнения.
  Система пилотных задач позволяет не только осуществлять позднюю привязку, но и скрыть неоднородность различных вычислительных сред с помощью их унифицированного интерфейса “сервер-пилот”. Таким образом, можно интегрировать в единую вычислительную среду различные вычислительные инфраструктуры, например, Грид-инфраструктуру, инфраструктуру облачных вычислений, относительно обособленные машинные кластеры и даже суперкомпьютеры.
  Еще одной функцией пилотных заданий является подготовка входных данных для задачи и выгрузка результатов. И здесь как и в случае вычислительной среды наблюдается большая неоднородность систем хранения (разные протоколы, разный функционал), которую пилотные задания могут унифицировать для вышестоящих слоев.

3. Интеграция суперкомпьютера и Tier-1 в НИЦ “Курчатовский институт”

Суперкомпьютер НИЦ “Курчатовский институт” это высокопроизводительный вычислительный кластер второго поколения с пиковой производительностью 122,9 ТФлопс сдан в эксплуатацию с сентября 2011 года. Кластер состоит из 1280 счётных двухпроцессорных узлов, объединенных высокопроизводительной сетью передачи данных и сообщений InfiniBand DDR, имеет суммарную оперативную память 20,5 Тбайт и систему хранения данных на 144 Тбайт. На счётных узлах кластера установлена операционная система Linux (CentOS). Система хранения данных построена на параллельной файловой системе Lustre 2.0. Для управления распределением ресурсов и выполнением счетных заданий используется менеджер ресурсов SLURM. Для интеграции суперкомпьютера НИЦ “Курчатовский институт” и Tier-1 потребовалось немного расширить классическую схему PanDA . Разработанная схема представлена на рисунке 2.

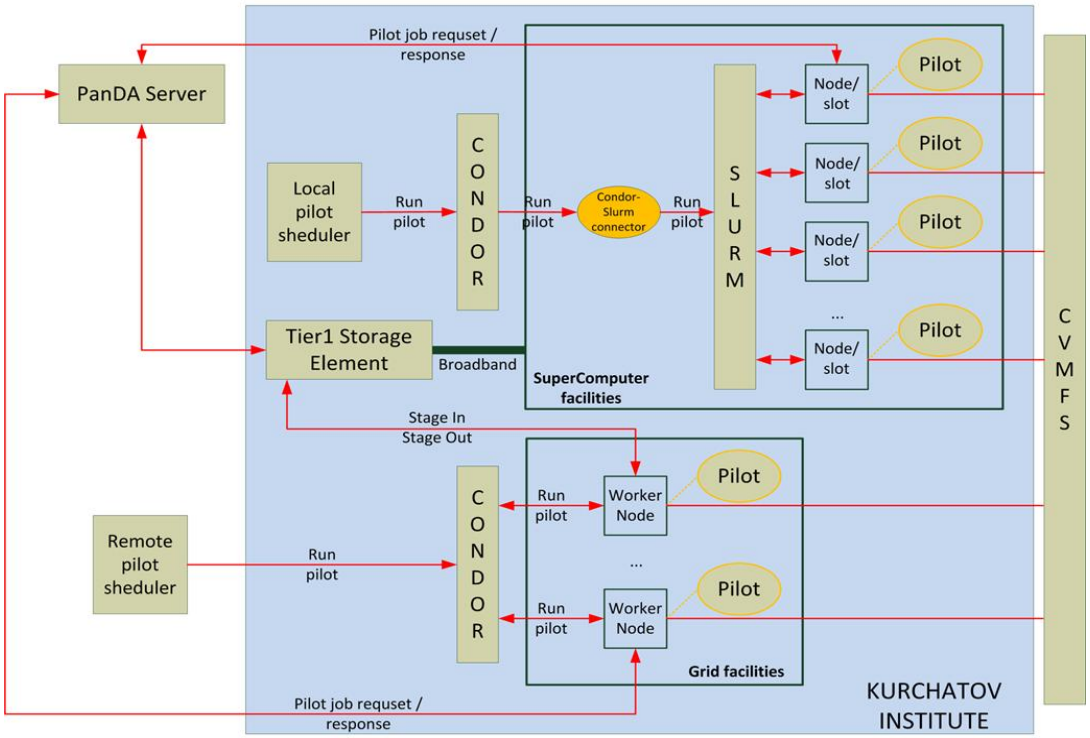

Рис. 2. Схема интеграции интеграции суперкомпьютера НИЦ “Курчатовский институт” и Tier-1

Центральным компонентом на схеме как и в классическом варианте является PanDA сервер. Мы развернули свой сервер в Курчатовском институте, куда можем посылать простые задачи. Но так как интерфейсы взаимодействия сервера и клиентов не были модифицированы, то в качестве сервера мы можем использовать как свой локальный экземпляр, так и центральный PanDA сервер, установленный в CERN.

Сервер распределяет пользовательские задачи между Tier-1 центром и суперкомпьютером в НИЦ “Курчатовский институт”. Для этого используется механизм пилотных заданий (см. раздел 2). Для Грид-инфраструктуры пилотные задания запускаются на рабочих узлах кластера специальной программной компонентой autopilot factory (APF). Это стандартная компонента системы PanDA. В нашем случае, мы используем централизованный autopilot factory. APF запускает пилотные задания через менеджер ресурсов CONDOR. В качестве хранилища данных используется Tier-1 хранилище, размещенное в НИЦ “Курчатовский институт”.

Для суперкомпьютера установлен локальный APF, который также при помощи CONDOR запускает пилотные задачи на рабочие узлы через менеджер ресурсов SLURM. Для интеграции CONDOR и SLURM мы реализовали программный медиатор, транслирующий задачи запускаемые CONDOR в SLURM и сопрягающий команды управления двух систем.

Рабочие узлы суперкомпьютера НИЦ “Курчатовский институт” имеют доступ во внешнюю сеть, что позволяет пилотным заданиям, запущенным на рабочих узлах функционировать аналогично тому, как это происходит на обычном кластере Грид-инфраструктуры. Для этого потребовалось произвести ряд дополнительных настроек. В частности, на узлах была смонтирована удаленная файловая система CVMFS и открыт доступ к заданному набору внешних адресов. В качестве хранилища данных используется то же Tier-1 хранилище, связанное с рабочими узлами суперкомпьютера высокоскоростным каналом.

Таким образом, сервер PanDA может инициировать и поставить на выполнение задачу по обработке данных, выбрав в качестве вычислительной инфраструктуры либо суперкомпьютер, либо Грид-сайт НИЦ “Курчатовский институт”. Для пользователя действия будут одинаковыми как в одном, так и в другом случае. Система автоматически отправит задачу на выполнение на требуемый ресурс, а результаты будут одинаковым образом сохранены и зарегистрированы в Tier-1 хранилище.

4. Экспериментальные исследования

Начиная с ноября 2014 года развернутый в НИЦ “Курчатовский институт” портал принимал и успешно обрабатывал задачи эксперимента ATLAS, для чего он был интегрирован с сервером системы управления задачами PanDA эксперимента ATLAS, установленным в CERN. Для этого он был прописан в центральной информационной системе PanDA, что позволило серверу PanDA посылать задачи моделирования эксперимента ATLAS как в Грид-инфраструктуру, так и на суперкомпьютер НИЦ “Курчатовский институт”.

Интеграция осуществлена путем создания на центральном сервере PanDA в CERN отдельной очереди задач, к которой обращались пилотные задания, запускаемые на ресурсах НИЦ “Курчатовский институт”. В ходе эксперимента более 200 пользовательских задач запускалось и успешно выполнялось ежедневно. Одним из наиболее важных исследований, решаемых на суперкомпьютере является реконструкция событий протон-протоного взаимодействия с высоким числом взаимодействия для изучения производительности Трекового Детектора Переходного Излучения (TRT) [5].

На рисунках 3-4 приведена статистика выполнения задач эксперимента ATLAS на ресурсах НИЦ “Курчатовский институт” (служебные задачи - синий, пользовательские - красный).

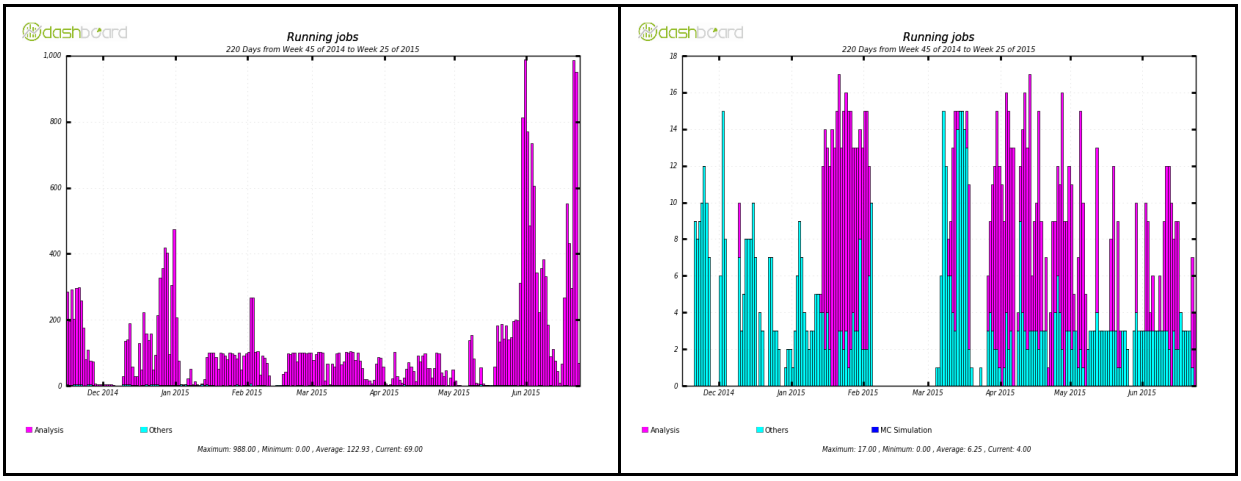

Рис. 3. Статистика выполняемых задач эксперимента ATLAS на Грид-сайте первого уровня и суперкомпьютере НИЦ “Курчатовский институт”

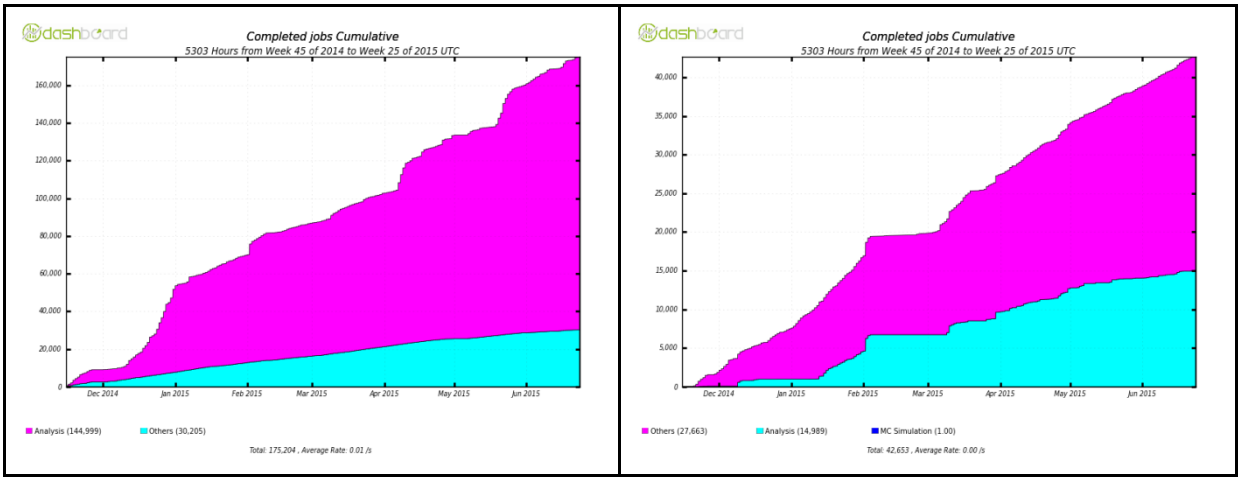

Рис. 4. Статистика успешно завершенных задач эксперимента ATLAS на Грид-сайте первого уровня и суперкомпьютере НИЦ “Курчатовский институт”

5. Заключение

В НИЦ “Курчатовский институт” был разработан и развернут единый портал, интегрирующий Грид-центр первого уровня и суперкомпьютер в НИЦ «Курчатовский институт». Портал предназначен для запуска пользовательских задач обработки данных эксперимента ATLAS на ресурсах Tier-1 и суперкомпьютера НИЦ “Курчатовский институт” с единой точкой входа и общим хранилищем данных, позволяя тем самым привлечь суперкомпьютер НИЦ “Курчатовский институт” в качестве дополнительного вычислительного ресурса для Грид-инфраструктуры, изначально не учтенного в модели организации вычислений БАК. В качестве базовой технологии была использована система управления задачами PanDA. Ежедневно на суперкомпьютере выполняется более 200 пользовательских задач.

Литература

1. Collaboration, ATLAS, and G. Aad. "The ATLAS experiment at the CERN large hadron collider." J. Instrum 3 (2008): S08003.
2. Worldwide LHC Computing Grid. Сайт проекта. Электронный ресурс. URL: http://wlcg.web.cern.ch/
3. Maeno, Tadashi. "PanDA: distributed production and distributed analysis system for ATLAS." Journal of Physics: Conference Series. Vol. 119. No. 6. IOP Publishing, 2008.
4. T. Maeno et al. “Evolution of the ATLAS PanDA workload management system for exascale computational science” 2014 J. Phys.: Conf. Ser. 513 032062.
5. ATLAS collaboration, ATLAS Inner Detector Technical Design Report vol. 1, ATLAS TDR 4, CERN/LHCC 97-16 (1997) and vol. 2, ATLAS TDR 5, CERN/LHCC 97-17 (1997)

Integration of Russian Tier 1 center with HPC at NRC "Kurchatov institute"

Ivan Tertychnyy, Ruslan Mashinistov, Alexander Novikov, Alexey Poyda, Alexei Klimentov and Eygene Ryabinkin

Keywords: HPC, Grid, Distributed computing

LHC experiments preparing for the precision measurements and further discoveries that will be made possible by much higher LHC collision rates from 2015 (Run2). The need for simulation, data processing and analysis and would overwhelm the expected capacity of WLCG computing facilities unless the range and precision of physics studies were to be curtailed. To meet this challenge the integration of the opportunistic resources into LHC computing model is highly important.

Tier-1 facility at Kurchatov Institute (NRC-KI) in Moscow is a part of WLCG and it will process and store up to 10% of total data obtained from ALICE, ATLAS and LHCb experiments. In addition Kurchatov Institute has supercomputers with peak performance 0.12 PFLOPS. Delegation of even a fraction of super-computing resources to the LHC Computing will notably increase total capacity.

In 2014, we have started a pioneer work to develop a portal combining a Tier-1 and a supercomputer in Kurchatov Institute. This portal is aimed to provide interfaces to run Monte-Carlo simulation at the Tier-1 Grid and supercomputer, using common storage. PanDA (Production and Distributed Analysis) workload management system having great results at the ATLAS was chosen as underlying technology.