Портал для организации обработки научных данных на гетерогенных вычислительных ресурсах в НИЦ «Курчатовский институт»

В. А. Аулов, Д. Д. Дрижук, А. А. Климентов, Р. Ю. Машинистов, А. М. Новиков, А. А. Пойда[a], Е. А. Рябинкин, И. С. Тертычный

Национальный исследовательский центр «Курчатовский институт»,
123098, Москва, пл. Академика Курчатова, 1

E-mail: [a] poyda@wdcb.ru

Центр обработки данных в Национальном Исследовательском Центре "Курчатовский Институт" (НИЦ КИ) обеспечивает своим пользователям доступ к новейшим гетерогенным вычислительным ресурсам, таким как суперкомпьютеры, облачные и грид вычисления. Необходимость интегрировать вычислительные мощности в единую систему для пользователей сподвигла нас на разработку и реализацию информационного портала.

Система управления потоком рабочих заданий Production and Distributed Analysis system (PanDA) была выбрана в качестве базы для портала. PanDA продемонстрировала превосходные показатели по производительности и масштабируемости при обработке данных и управления загрузкой для эксперимента ATLAS на Большом Адронном Коллайдере в ЦЕРН. Платформа PanDA была установлена в НИЦ КИ, адаптирована к конкретным нуждам научного сообщества и интегрирована с вычислительными ресурсами НИЦ КИ. Для осуществления этой интеграции потребовалась разработка пользовательского интерфейса для предоставления ученым единой точки доступа к вычислительным ресурсам, а также системы управления файлами для перемещения входных/выходных данных между файловыми хранилищами и рабочими узлами.

Портал продемонстрировал свою эффективность в работе с биоинформатическими научными приложениями, в частности, по анализу генома мамонта. Успех применения в биологии вызвал интерес ученых из других наук, в которых востребованы высокопроизводительные вычисления, поэтому мы планируем расширить сферу применения портала в ближайшем будущем. В статье представлены наши достижения, а также потенциальные возможности применения портала.

Ключевые слова: распределенные вычисления, суперкомпьютеры, большие данные, системы управления потоком задач.

Работа выполнена в рамках мега-гранта правительства РФ, контракт № 14.Z50.31.0024 и гранта РФФИ № 16-37-00249 мол_а. Данная работа была выполнена с использованием высокопроизводительных вычислительных ресурсов федерального центра коллективного пользования в НИЦ “Курчатовскии институт”.

1. Введение

Цель изложенной в данной статье работы – разработка портала для обработки данных, объединяющего гетерогенные вычислительные ресурсы в общую вычислительную инфраструктуру. Актуальность данной работы для НИЦ “Курчатовский институт” обусловлена наличием в институте целого ряда вычислительных ресурсов с различной архитектурой, в том числе:

- суперкомпьютерный кластер с пиковой производительностью 122,9 ТФЛОПС, состоящий из 1280 узлов, каждый из которых имеет 2 процессора по 4 ядра (итого 8 ядер на узел);
- один из самых крупных Грид-центров в России: Tier-1 центр, который хранит до 10% всех данных трех экспериментов ALICE, ATALS и LHCB, проводимых на Большом адронном коллайдере (БАК);
- академическое облако, построенное на базе облачной платформы OpenStack и имеющее пиковую производительность 1,5 Терафлопс.

Это вычислительные ресурсы, которые мы хотим объединить в первую очередь. Но помимо этого в институте есть и другие вычислительные платформы, например, SMP-кластер, кластер на узлах с графическими ускорителями и др., которые также могут быть интегрированы в общее вычислительное пространство посредством разработанного портала. Объединение гетерогенных ресурсов в общую вычислительную среду позволит, во-первых, организовать общую очередь задач и более эффективно распределять задачи между ресурсами в зависимости от загруженности последних, а во-вторых, облегчит для конечных пользователей процесс организации их совместного использования.

2. Архитектура портала

Чтобы достичь результата мы воспользовались методами и подходами, применяемыми в физике высоких энергий для обработки больших данных экспериментов, проводимых на БАК. В частности, в качестве программной основы мы взяли систему управления заданиями PanDA [Maeno, 2008], созданную и успешно применяемую для управления тысячами одновременно выполняющихся задач на вычислительных ресурсах, распределенных по всему миру.

На рисунке 1 представлена архитектура разработанного портала. Основными компонентами портала являются:

- Сервер, управляющий распределением задач на ресурсы.
- Агрегируемые вычислительные ресурсы и ресурсы хранения данных. Мы используем модульный подход с унифицированными интерфейсами, чтобы можно было расширять список агрегируемых ресурсов.
- Веб-интерфейс, позволяющий пользователям формировать, ставить на выполнение и отслеживать состояния их задач. Интерфейс реализован на языке python и поддерживает авторизацию по протоколу OAuth 2.
- Файловый каталог и система транспортировки данных для регистрации файлов в распределенной среде хранилищ и асинхронного перемещения входных и результирующих данных между вычислителями и хранилищем.
- HTTP API, реализующий функции для взаимодействия портала с внешними системами.
- Системное персонифицированное FTP-хранилище, используемое для подготовки входных данных для задач.

Использование системы PanDA в качестве программной основы позволило поддержать совместимость с системой PanDA, используемой в ЦЕРН для обработки данных эксперимента

ATLAS [The ATLAS Collaboration, Aad et al., 2008]. В частности, мы настроили запуск пилотных заданий на Курчатовском суперкомпьютере таким образом, что они принимают задачи не только с нашего сервера, но и с центрального сервера PanDA, установленного в ЦЕРН.

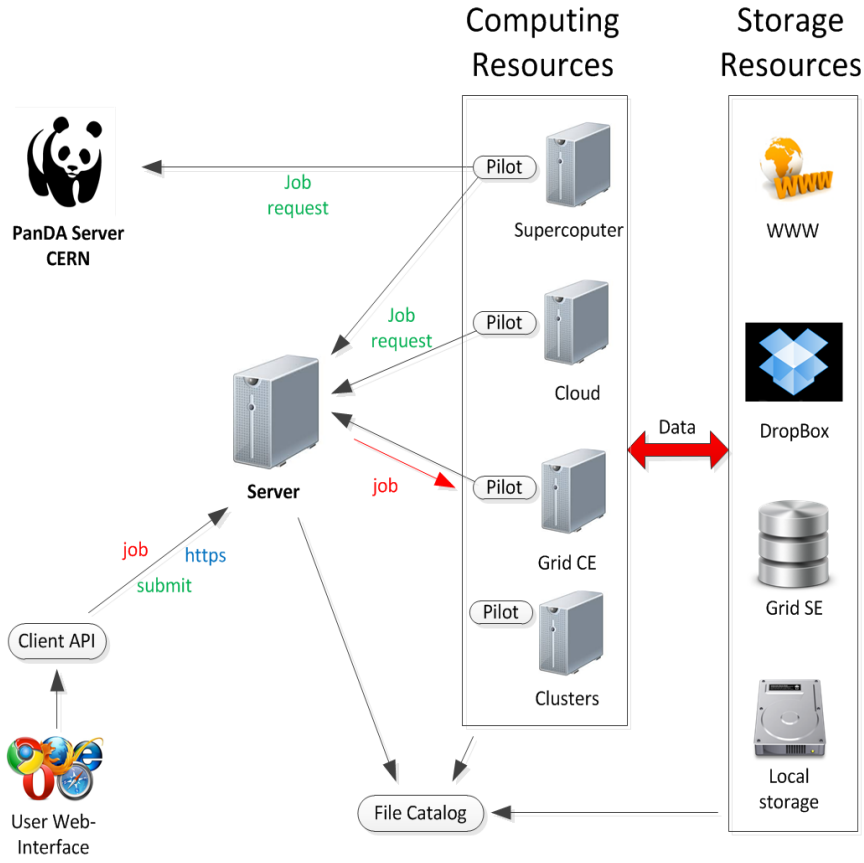

Рис. 1. Архитектура разработанного портала

Разработанный портал реализует вычислительную схему, заложенную в PanDA, и показывает наибольшую эффективность для вычислительно-емких задач, которые могут быть разбиты на множество более мелких задач, запускаемых распределенно в параллельном режиме. Чтобы пользователю не пришлось делать это вручную, мы разработали и интегрировали в портал программный модуль, реализующий рабочий поток, поддерживающий автоматическое разбиение входных данных на блоки меньшего размера, подготовку и запуск задач по обработке этих блоков, сбор результатов. Схема рабочего потока приведена на рисунке 2.

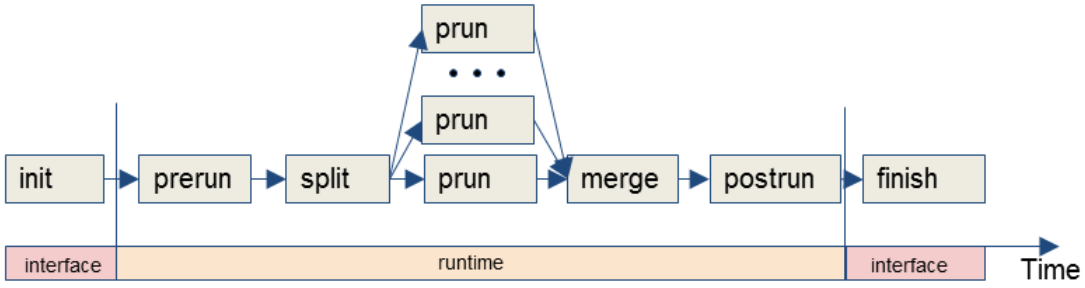

Рис. 2. Схема рабочего потока, поддерживающего автоматическое разбиение входных данных на блоки меньшего размера, подготовку и запуск задач по обработке этих блоков, сбор результатов

Рабочий поток состоит из 7 этапов:

- На первом шаге (init) производится инициализация рабочего потока, проверка входных файлов и файлов конфигурации.
- На втором шаге (prerun) производится предварительная обработка всего блока данных целиком, а также выполняется дополнительная обработка, общая для всех последующих параллельных задач.
- На третьем шаге (split) выходные данные предыдущего этапа разбиваются на отдельные блоки, для каждого из которых формируется независимая задача. Число блоков, на которые разделяются данные, зависит от производительности ресурсов, доступных для параллельной обработки, специфики программного обеспечения и пожелания пользователя.
- На четвертом шаге (prun) производится параллельное выполнение полученных на предыдущем шаге задач. При этом задачи могут выполняться на разных вычислительных ресурсах.
- На пятом шаге (merge) производится сборка результатов выполненных параллельных задач в цельный пакет для обработки. Сборка производится с помощью специальной задачи.
- На шестом шаге (postrun) производится постобработка собранных результатов как единого объекта данных.
- На последнем шаге (finish) производится завершение конвейера, проверка выходных файлов и возврат результатов пользователю.

Представленная схема существенно сокращает время работы всего рабочего потока за счет массового параллелизма на четвертом шаге и автоматизации настройки, запуска и перезапуска обработки всех этапов.

3. Апробация портала на задачах анализа данных геномного секвенирования

Мы провели апробацию разработанного портала на задаче из области биоинформатики, а именно на задаче анализа данных геномного секвенирования, в которой требуется провести многоэтапный анализ порядка 350 ГБ данных, содержащих результаты секвенирования ДНК мамонта и сравнить их с ДНК современного африканского слона. Для проведения подобного анализа может быть использован популярный фреймворк PALEOMIX [Schubert et al., 2014], который включает специализированные программные компоненты для обработки данных секвенирования. Сам PALEOMIX достаточно требователен к ресурсам, особенно к объему оперативной памяти, и обработка данных с его помощью – достаточно длительный процесс. Так, например, обработка 350 ГБ входных данных в описанной задаче на одном 80-ядерном сервере занимает порядка месяца.

Мы адаптировали PALEOMIX под описанную ранее схему рабочего потока разработанного портала, что было нетривиально, так как сам PALEOMIX как программный продукт не поддерживает схему распределенных вычислений между несколькими машинами. Однако, некоторые из компонент PALEOMIX могут обрабатывать отдельные фрагменты входных данных независимо от остальных фрагментов. Мы воспользовались этим обстоятельством и распараллелили работу PALEOMIX путем разбиения входных данных на большее число файлов меньшего размера и параллельной обработки этих файлов с последующей сборкой выходных данных из результатов обработки отдельных файлов. Мы разделили входные данные на 135 блоков, для каждого блока сформировали отдельную задачу, и полученные 135 задач запустили параллельно на разных узлах суперкомпьютера. В результате мы добились существенного сокращения времени выполнения анализа – с месяца до нескольких дней.

4. Заключение

В результате проделанной работы мы разработали и развернули портал для обработки данных, объединяющий гетерогенные вычислительные ресурсы Курчатовского института в общую вычислительную инфраструктуру. Для этого мы установили PanDA в Курчатовском институте, интегрировали его с вычислительными ресурсами НИЦ "Курчатовский институт", разработали дополнительные компоненты, такие как пользовательский веб-интерфейс, систему управления данными и HTTP API для взаимодействия с внешними приложениями. Мы также реализовали и интегрировали в портал рабочий поток, поддерживающий автоматическую нарезку входных данных, подготовку и запуск задач по обработке нарезанных блоков, сбор результатов.

Портал показал высокую эффективность для вычислительно-емких задач, которые могут быть разбиты на более мелкие подзадачи, выполняемые независимо, например, для задач анализа данных геномного секвенирования. В дальнейшем мы планируем расширить использование портала на другие предметные области.

Список литературы

Maeno T. on behalf of PANDA team and ATLAS collaboration. PanDA: distributed production and distributed analysis system for ATLAS // Journal of Physics: Conference Series. — 2008. — Vol. 119, No. 6.

Schubert M. et al. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX // Nat Protoc. 2014. Vol. 9(5). — P. 1056−82. doi: 10.1038/nprot.2014.063. Epub 2014 Apr 10. PubMed PMID: 24722405.

The ATLAS Collaboration, Aad G. et al. The ATLAS Experiment at the CERN Large Hadron Collider // Journal of Instrumentation. — 2008. — Vol. 3. — P. S08003.

Portal for organization of scientific data processing on heterogeneous computing resources at NRCKI

V. Aulov, D. Drizhuk, A. Klimentov, R. Mashinistov, A. Novikov, A. Poyda[a], E. Ryabinkin, I. Tertychnyi

National Research Center «Kurchatov Institute», 1, Akademika Kurchatova pl., Moscow, 123182, Russia

E mail: [a] poyda@wdcb.ru

National Research Center "Kurchatov Institute" Data Processing Center provides its user community with state of art large scale heterogeneous computing resources, such as supercomputers, cloud and grid computing. The requirement to integrate computing facilities as a single computing entity for the end-user has motivated us to develop and to implement the data processing portal.

The ATLAS Production and Distributed Analysis workload management system (PanDA) has been chosen as a base technology for the portal. PanDA has demonstrated excellent capabilities to manage various workflows at scale in the ATLAS experiment at LHC. PanDA instance was installed at NRC KI, adapted to the scientific needs of its user community and integrated with NRC KI computing resources. This integration has required the development of user interface to provide scientists with a single point of access to the computing resources, and file handling system to transfer input/output data between file storages and working nodes.

The portal has demonstrated its efficiency in running bioinformatical scientific applications, in particular for genome analysis. Success of biological application has attracted interest from other compute-intensive sciences, and we plan to expand portal's usage in the nearest future. In this report our accomplishments will be reviewed and then we'll discuss the portal's potential usage.

Keywords: distributed computing, supercomputers, Big data, workflow management systems.

NRC "Kurchatov Institute" laboratory team work is conducted with support from Ministry of Education and Science Russian Federation, contract № 14.Z50.31.0024, and with partial support from RFBR project № 16-37-00249 mol_a.