УДК 004.272.23(591)

ТЕХНОЛОГИИ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ В СПЕЦИАЛИЗИРОВАННОМ ИНФОРМАЦИОННО-РЕСУРСНОМ ЦЕНТРЕ (СИРЦ) СОЮЗА МЬЯНМЫ

¹Тхуреин К.Л., ¹Богданов А.В., ²Пья Сон Ко Ко, ²Чжо За

¹Санкт-Петербургский государственный университет, Санкт-Петербург,
e-mail: trkl.mm@mail.ru, bogdanov@csa.ru;

²Санкт-Петербургский государственный морской технический университет, Санкт-Петербург,
e-mail: pyaesonekoko@gmail.com, kyawzaya4436@gmail.com

В данной статье рассмотрены ключевые принципы правительства Мьянмы по развитию информационно-коммуникационных технологий (ИКТ), в которых ключевую роль играет создание Ресурсного центра для государственного управления, субъектов экономики, науки и образования. Его основная задача – предоставление качественных информационных и вычислительных услуг, а в период кризиса и в условиях недостаточности диверсификации прогнозирования и планирования социальных и экономических процессов и в альтернативной организации всесторонней информационно-вычислительной поддержки систем государственного управления Союза Мьянмы. В данной статье показаны методы консолидации данных с использованием Распределенной Федеративной СУБД. Определены ЭКО системы Больших Данных на основании инструментария работы с ними. Проект суперкомпьютерного центра для Правительства Республики Мьянмы Больших данных (Big Data) поддерживает файловые системы, обеспечивая работу иерархической системы запоминающих устройств. Для создания многопользовательских прикладных информационных систем применяются общепринятые системы управления базами данных (СУБД), поддерживающие реляционную модель представления данных и ориентированные на распределенную обработку данных. В данной статье рассматриваются технологии, которые в дальнейшем будут тестироваться. В основном описание посвящено модели данных и модели запросов. Знать модель данных крайне важно, так как данные будут моделироваться для тестов, и впоследствии будут оптимизироваться. Рассматриваются тестируемые базы данных No SQL, Cassandra и Riak.

Ключевые слова: распределенные базы данных, распределенная система, централизованные базы данных, федеративные базы данных, СУБД, SVDMS, Vmware

BIG DATA PROCESSING TECHNOLOGIES IN THE SPECIALIZED INFORMATION RESOURCE CENTER (SIRC) OF THE UNION OF MYANMAR

¹Tkhurein K.L., ¹Bogdanov A.V., ²Pyae Sone Ko Ko, ²Kyaw Zaya

¹Saint Petersburg State University, St.Petersburg, e-mail: trkl.mm@mail.ru, bogdanov@csa.ru;

²Saint-Petersburg State Marine Technical University, St.Petersburg,
e-mail: pyaesonekoko@gmail.com, kyawzaya4436@gmail.com

This article is providing the principles of the Government of Myanmar for the development of information and communication technologies (ICT), which can create of a Resource Center for public administration, subjects of the economy, science and education section with a play key role. Its main task is to provide quality information and computing services, in times of crisis and in conditions of insufficient diversification of forecasting and planning of social and economic processes and in the alternative organization of comprehensive information and computing support for public administration systems of the Union of Myanmar. This article shows methods of data consolidation using Distributed Federated DBMS. Defined the system of Big Data and based on the tools for working with them. Big Data supports file systems, providing a hierarchical system of storage devices. To create multi-user application information systems, generally accepted database management systems (DBMSs) are used that support the relational model of data representation and are focused on distributed data processing. In this article discusses about the technologies will be tested in the future. The main description is devoted to the data model and the query model. Knowing the data model is extremely important, as the data will be modeled for tests, and subsequently will be optimized. We tested about No SQL, Cassandra and Riak and their test results for BigDATA.

Keywords: distributed databases, distributed system, centralized databases, federated databases, DBMS, SVDMS, Vmware

Правительство Мьянмы работает над достижением цифрового видения Мьянмы на будущее путем создания национальной политики в области информационно-коммуникационных технологий (ИКТ). Правительство Мьянмы признает, что развитие информационно-коммуникационных технологий является незаменимым фактором для среднесрочного и долгосрочного роста Мьянмы и представляет собой потенциал

данного роста, а также является важным условием для повышения эффективности исполнения бюджета и административной эффективности посредством «электронного правительства».

Создание сети высокопроизводительных вычислительных центров в Союзе Мьянмы является сложной и важной задачей. Назначением ВВЦ (Ресурсного центра) является предоставление органам государ-

ственного управления, субъектам экономики, науки и образования качественных информационных и вычислительных услуг, а в период кризиса и в условиях недостаточности диверсификации прогнозирования и планирования социальных и экономических процессов также заключается и в альтернативной организации всесторонней информационно-вычислительной поддержки систем государственного управления Союза Мьянмы на базе высокопроизводительных вычислений особой результативности.

Ресурсный центр несёт также функции обеспечения бесперебойной работы всей иерархической корпоративной системы высокотехнологичных отраслей промышленности в целом, а также функции обеспечения связи между отдельными отраслями и их предприятиями. В функции Ресурсного центра входит, что информационно-алгоритмическое обеспечение решения ресурсоёмких научно-технологических задач [1].

Ресурсный центр должен обеспечивать информационное обслуживание пользователей. Он должен быть укомплектован экспертами и специалистами в различных прикладных областях, способных оптимизировать использование вычислительных ресурсов и предлагать новые решения преодоления возникающих проблем.

Цель работы состоит в исследовании и разработке технических решений для анализа построения и применения высокопроизводительных вычислительных систем как основы суперкомпьютерной инфраструктуры для научных исследований. Для дости-

жения поставленной цели в работе необходимо решить следующие задачи:

- Разработка архитектуры, повышения производительности и параллельному управлению распределёнными системами баз данных для создания многопользовательских прикладных информационных систем.

- Создание суперкомпьютеров с различным набором характеристик, сетевых и программных решений, воплотивших в себя разработанные архитектурные, сформировать на их основе интегрированную вычислительную информационно-коммуникационную среду проведения научных исследований и решения прикладных задач.

Суперкомпьютерные технологии и создание центров высокопроизводительных вычислений

В данной статье мы приведём наши предложения по насыщению создаваемого вычислительного центра компьютерными системами. На рис. 1 приведена общая, укрупнённая схема вычислительного центра. К основным компонентам центра относятся вычислительный комплекс, сети и комплекс хранения данных. Поскольку стоимость систем вычислительного кластера и необходимого программного обеспечения составляет внушительную сумму, то развёртывание центра и насыщение его необходимым оборудованием будет происходить последовательно, в течение всего срока создания Специализированного информационно-ресурсного центра СИРЦ.

Рис. 1. Общая схема сбалансированной архитектуры суперкомпьютерного центра

Структурно-технические решения приводят к тому, что суперкомпьютерные системы быстро обновляются. Чтобы работа суперкомпьютерных центров была максимально эффективной, ориентируются на наиболее современные и быстродействующие системы.

Современный суперкомпьютер строится путем объединения большого количества серийно выпускаемых быстродействующих микропроцессоров, которые используются также и для построения обычных рабочих станций и серверов [2, 3].

В условиях современного развития телекоммуникационных технологий отпадают жёсткие ограничения географического размещения Ресурсного центра. Основопологающим условием выбора является наличие научного коллектива, способного разработать, организовать и эксплуатировать данный Центр. Наиболее целесообразным вариантом является его создание на базе научного центра (высокопроизводительных вычислений и информационных систем) в городе Най Пей До, который имеет необходимые вычислительные средства, коммуникационное оборудование и научный коллектив, способный выполнить работы по созданию Ресурсного центра в кратчайшие сроки [4].

Разместить Центр целесообразно на площадях, имеющих необходимую инфраструктуру для функционирования, как открытой, так и закрытой части Центра, прежде всего энергообеспеченность. Размещение на одних и тех же площадях открытой и закрытой части желательно, но некритично. Площадка под центр должна иметь возможности как для размещения и функционирования открытой части центра, в статусе международного центра GRID (отдельное здание с необходимыми телекоммуникационными каналами связи с международными и национальными информационными сетями), так и для закрытой части (с реализацией условий обработки и передачи закрытой информации).

Вычислительные и телекоммуникационные ресурсы создаваемого Центра позволяют обеспечить в первую очередь не только наиболее изученные потребности большой строительной промышленности, но и потребности других стратегических отраслей промышленности, в частности ТЭК, что резко повышает эффективность функционирования Центра [5].

Анализ и развитие технологий обработки больших данных в научных исследованиях

Одной из важных областей использования высокопроизводительных вычисли-

тельных центров является создание и обработка функций базы данных. Создание вычислительных средств в вычислительно-информационные узлы и организация позволит создать единый вычислительно-информационный ресурс, БД в естественных и гуманитарных науках [6].

Под большими данными понимается широкое разнообразие массивов данных, которые не могут быть обработаны традиционными приложениями из-за своего огромного объема. Анализ больших данных – сложная работа: разделение, контроль, передача, визуализация, хранение и сохранение информации [7]. Большие данные анализа часто подразумевают применение прогнозной аналитики или других передовых методов с целью извлечения из разных данных определенной информации.

В области математики на научном потенциале больших данных системы очень важна оптимизация технологии решения задач на высокопроизводительных вычислительных системах на использовании крупномасштабного моделирования [8].

Рис. 2. Высокопроизводительная серверная архитектура центра обработки данных в Союзе Мьянмы

Анализ Системы управления Большими данными и их результаты

В этой статье представлены результаты тестирования производительности баз

данных. Результаты представлены для трех различных баз данных: Cassandra, Riak и MySQL. В распределенной системе базы данных должны быть установлены в кластере, который содержит несколько узлов. Производительность операций чтения, как ожидается, увеличится с увеличением фактора репликации, в то время как производительность операции записи, как ожидается, уменьшится, так как система имеет дополнительную нагрузку из-за необходимости выполнять те же операции записи в различных узлах. Кроме того, во фрагментированной MySQL не предусмотрен никакой механизм репликации. Но это выходит за рамки этой работы по оценке производительности и изучению влияния репликации на базы данных NoSQL [9].

Результаты деятельности представлены в этой статье в соответствии с двумя типами рабочих нагрузок: «в основном чтение» (read heavy) и «чтение и запись» (write heavy).

В этом тесте измерялась производительность баз данных при большом количестве операций обновления данных. В данном случае 30 процентов составляли операции обновления, а остальные 70 процентов – операции чтения. На рис. 3 можно наблюдать среднее время ожидания чтения и обновления, когда нагрузка на систему возрастает. Как показано на рисунке, для всех систем задержка операций увеличивалась при увеличении нагрузки.

Cassandra, которая оптимизирована для большого числа операций записи,

имела наименьшую задержку при высокой нагрузке, как для чтения, так и для операций обновления. Riak и MySQL масштабируются в геометрической прогрессии, достигая аналогичной задержки в соответствии с нагрузкой, особенно при низких нагрузках. С другой стороны, Cassandra масштабируется линейно, хотя для операций чтения сначала и были более высокие задержки, чем у других систем. На Riak и MySQL операции обновления всегда имеют большие задержки, чем операции чтения при той же нагрузке. Тем не менее на Cassandra наблюдается обратное: операции обновления всегда были быстрее [10, с. 31–32].

Что касается операций чтения, существует не такая большая разница в результатах в отношении трех систем. Тем не менее на операциях обновления существует высокая дисперсия по трем системам, особенно при более высоких нагрузках (более 7000 операций/с). При 10000 операциях/с Cassandra достигла средней задержки 9,85 мс для операций обновления, что на 80% меньше, чем Riak (48,54), и на 71% меньше, чем MySQL (33,01).

Ожидается, что за счет увеличения доли операций обновления производительности Cassandra будет продолжать улучшаться, так как она выполняет операции записи более эффективно, чем операции чтения, в то время как другие системы будут работать хуже, увеличивая разницу в производительности между Cassandra и MySQL и Riak [11, с. 42–43].

Рис. 3. Результаты тестирования с типом нагрузки вида «чтение и запись»

Выводы

Проведенные тесты показывают, что Cassandra и Riak обеспечивают производительность на уровне с реляционной базой данных. Cassandra обеспечивает гораздо лучшую производительность, чем остальные системы, если речь идет об операциях записи, но ее общая производительность снижается из-за операций записи. Производительность Riak очень похожа на шардинг MySQL.

Поддержание высокого уровня науки поможет не только сохранить научный потенциал Союза Мьянма, но и окажет позитивное влияние на дальнейшее развитие науки. В данной статье приведены наши предложения по насыщению создаваемого вычислительного Специализированного информационно-ресурсного центра (СИРЦ) для Мьянманских ученых, сохраняя и возрождая Мьянманскую научную элиту.

В Центре предусматриваются подразделения и инфраструктурные мощности, обеспечивающие следующие направления: аналитико-прогнозное, экономики жизнедеятельности, системного моделирования, ситуационного анализа, социологическое, исследовательско-конструкторское по отраслям, информационно-образовательное, развития прикладного программирования, интерфейсного обеспечения, статистического структурирования, высокоскоростной сети, удаленного обмена данными, математического моделирования и алгоритмов, использования микропроцессорной базы,

конструирования микроэлектроники, архитектуры высокопроизводительных вычислений, инфраструктуры высокопроизводительных вычислений, распределенных вычислений, программирования суперкомпьютеров и др.

Список литературы

1. Irina Ulitina, Thurein Kyaw Lwin, Alexander Bogdanov. The Impact Of Big Data On The Choice Of Used Storages And Modern Trends In Managing Big Data. Computer Science and Information Technology, CSIT 2019. P. 57–59.
2. Сергей Абрамов. Суперкомпьютеры: Обратные Рекорды // Наука и Жизнь. 2019. № 1. С. 42–45.
3. Богданов А.В., Тхуреин Киав Лин. Использование Технологий БД Для Системной Интеграции Гетерогенных Комплексов // Известия СПбГЭТУ «ЛЭТИ». 2012. № 4. С. 21–24.
4. Леонид Черняк. Смутное время СУБД // Открытые системы. 2012. № 2. С. 13–17.
5. Min Chen, Shiwen Mao, Yin Zhang, Victor C.M. Leung. Big Data: Related Technologies, Challenges, and Future Prospects. Springer Briefs in Computer Science. 2014. P. 51–57.
6. Mike Hogan. Shared-Disk vs. Shared-Nothing. Comparing Architectures For Clustered Databases. [Electronic resource]. URL: http://www.scaledb.com/pdfs/WP_SdvsSN.pdf (date of access: 22.03.2020).
7. Charu C Aggarwal. An introduction to social network data analytics. Springer Science and Business Media. 2011. P. 56–58.
8. James Manyika. Big data: The Next Frontier For Innovation, Competition, And Productivity. McKinsey Global Institute. 2011. June. P. 27–31.
9. Мартин Фаулер, Прамодкумар Дж. Садаладж. NoSQL: новая методология разработки нереляционных баз данных. М.: «Вильямс», 2013. С. 154–158.
10. Васвани В. MySQL: использование и администрирование. СПб.: «Питер», 2011. 368 с.
11. Гольцман В. MySQL Библиотека программиста. СПб.: «Питер», 2010. 253 с.