Новый алгоритм хранения и передачи РСДБ-данных на программный коррелятор РАН

© И. А. Безруков[1], А. В. Вылегжанин[1,2,3], Я. Л. Курдубова[1], В. Ю. Мишин[1], А. И. Сальников[1], В. А. Яковлев[1]

[1] ИПА РАН, г. Санкт-Петербург, Россия
[2]ФТИ им. А. Ф. Иоффе, г. Санкт-Петербург, Россия
[3] Санкт-Петербургское отделение МСЦ РАН — филиал ФГУ ФНЦ НИИСИ РАН, г. Санкт-Петербург, Россия

Ежедневно порядка 20 Тбайт данных, полученных в ходе РСДБ-наблюдений на радиотелескопах РТ-13 (обсерватории ИПА РАН «Бадары», «Зеленчукская»), предварительно буферизуются в системе хранения данных (СХД) в Санкт-Петербурге. По локальной вычислительной сети эти данные передаются на коррелятор в центр корреляционной обработки РАН. Основные требования к обмену данными между СХД и коррелятором — целостность и оперативность доставки данных. Для выполнения этих требований серверы СХД и коррелятор подключены к локальной вычислительной сети центра корреляционной обработки РАН через коммутатор с портами 10 GbE.

В статье предложены рекомендации по программной модернизации СХД, сделанные с учетом опыта эксплуатации СХД. Модернизация СХД позволит увеличить используемый объем дисковой памяти, скорость обмена данными и отказоустойчивость.

Ключевые слова: РСДБ, РСДБ-сеть «Квазар-КВО», РТ-13, VGOS, система хранения данных (СХД), LACP, программный коррелятор, дисковые массивы.

https://doi.org/10.32876/ApplAstron.46.3-9

Введение

С 2016 г. в центре корреляционной обработки (ЦКО) РАН происходит интенсивный обмен РСДБ-данными между новыми 13-метровыми [1, 2] радиотелескопами РТ-13 сети «Квазар-КВО», системой хранения данных (СХД) [3] и программным коррелятором [4]. СХД обеспечивает предварительную буферизацию, проверку целостности и кратковременное хранение данных. Последующий обмен данными между СХД и программным коррелятором осуществляется по высокоскоростной локальной вычислительной сети (ЛВС).

Анализируя опыт эксплуатации СХД в ИПА РАН и опыт коллег, работающих с дисковыми накопителями в сети EVN (The European VLBI Network) [5], можно перечислить основные тенденции в развитии СХД для РСДБ-данных:

— растет дисковый объем передаваемых и хранимых данных;

— растет скорость обмена данными между участниками, как локальными так и внешними;

— система усложняется с ростом числа компонентов (узлов) в эксплуатируемых СХД, соответственно, растет и число точек отказа от отдельных дисков до серверов и дисковых полок в целом.

Существенное увеличение объема РСДБ-данных, в связи с вводом в эксплуатацию радиотелескопа РТ-13 в обсерватории «Светлое» и увеличением объема РСДБ-данных по международным программам, потребует аппаратной модернизации СХД: добавления новых серверов, замены контроллеров, дисков и дисковых полок на более современные аналоги.

Для использования всех ресурсов СХД, дисковой и сетевой подсистем, а также для обеспечения их бесперебойной работы, помимо аппаратного обновления необходимо изменить принцип взаимодействия участников обмена трафиком с СХД. Понимание алгоритмов хранения/размещения данных в СХД и способов получения этих данных программным коррелятором, а также анализ работы текущей версии СХД позволит изменить характеристики системы так, чтобы обеспечить оптимальный вариант модернизации СХД.

Этапы модернизации СХД

Аппаратная модернизация

Первая версия СХД в ЦКО РАН включала 2 сервера Dell R730 с дисковыми полками Dell MD1220 на 48 дисков 2.5 дюйма с суммарным объемом дисковой памяти порядка 50 Тбайт. В 2017 г. дисковые полки были заменены на полки Supermicro с дисками форм-фактора 3.5 дюйма [6], что позволило увеличить суммарную дисковую емкость до 300 Тбайт. В 2018 г. для хранения данных с радиотелескопа РТ-13 в обсерватории «Светлое» в состав СХД вошло еще 2 сервера Dell R430 и дисковая полка Supermicro с дисками форм-фактора 3.5 дюйма, что позволило достичь схемы *n+1*, где *n* — количество обсерваторий или источников данных, а *+1* — дополнительный резервный сервер, обеспечивающий отказоустойчивость системы.

Модернизация алгоритма распределения данных

На рис. 1 представлена структурная схема процесса обмена РСДБ-данными между источниками данных, серверами СХД и узлами коррелятора. На схеме: C01, C02 — узлы коррелятора, S1, S2 — серверы СХД, цветом различаются данные двух обсерваторий.

Отличительная особенность такой схемы — статическая привязка размещения данных конкретной обсерватории к определенному серверу.

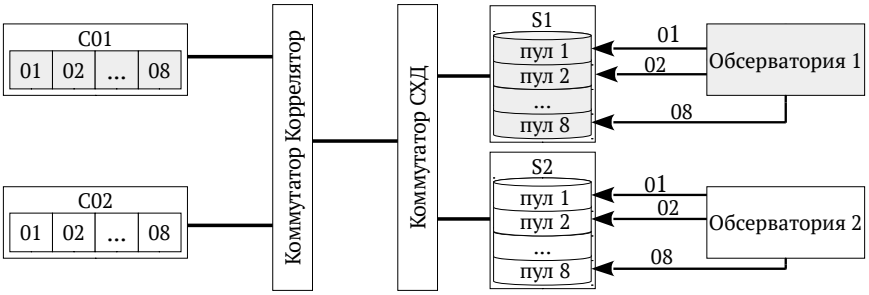

Рис. 1. Структурная схема обмена РСДБ-данными

РСДБ-данные на уровне скана (наблюдения одного источника) представляют собой набор физически независимых файлов по числу каналов широкополосной системы преобразования сигналов [7]. Количество и выбор конкретных каналов задаются программой наблюдения. Файлы каждого канала хранятся на закрепленной за этим каналом дисковой группе. Таких дисковых групп 8 (по числу каналов широкополосной системы преобразования сигналов). Схема деления дисковой емкости на группы обусловлена необходимостью обеспечения заданной производительности операций ввода-вывода и отказоустойчивости СХД.

В ходе эксплуатации СХД были выявлены недостатки действующей схемы размещения данных. Во-первых, если в наблюдении задействовано меньше 8 каналов (часто ведутся наблюдения по 4–6 каналам), то дисковая емкость СХД используется на 50–70 %. Во-вторых, из-за статического закрепления данных как на дисковых группах, так и на серверах, в случае необходимости проведения ремонтных или профилактических работ требуется остановка работы СХД на время обновления конфигурации на всех участниках обмена данными.

В новом аглоритме хранения данных предлагается перейти от статической схемы размещения РСДБ-данных к динамической. При таком алгоритме (рис. 2) данные из обсерваторий будут равномерно распределяться по всем доступным серверам и дисковым группам СХД.

Для координации участников обмена данными и хранения информации требуется сервер для метаданных, в качестве которого может быть использован один из серверов СХД. Переход к такому способу распределения данных позволяет:

— использовать всю дисковую емкость всех серверов СХД;

— выводить часть СХД на обслуживание;

— реализовать в дальнейшем схему резервирования данных на уровне избыточного массива независимых узлов RAIN (Redundant Array of Independent Nodes) вместо избыточного массива независимых дисков (RAID).

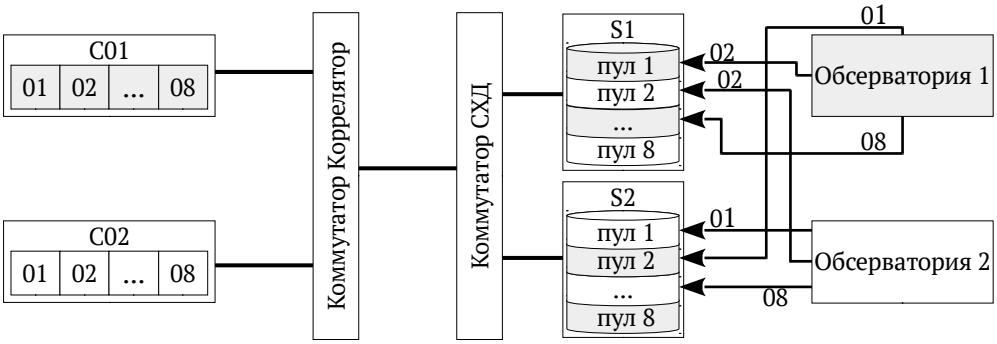

Рис. 2. Структурная схема динамического расспределения РСДБ-данных в СХД

Сетевая структура СХД

Рассмотрим особенности обмена данными между узлами программного коррелятора с РСДБ-данными, хранящимися в СХД. Структурная схема коммутации сети ЦКО РАН представлена на рис. 3.

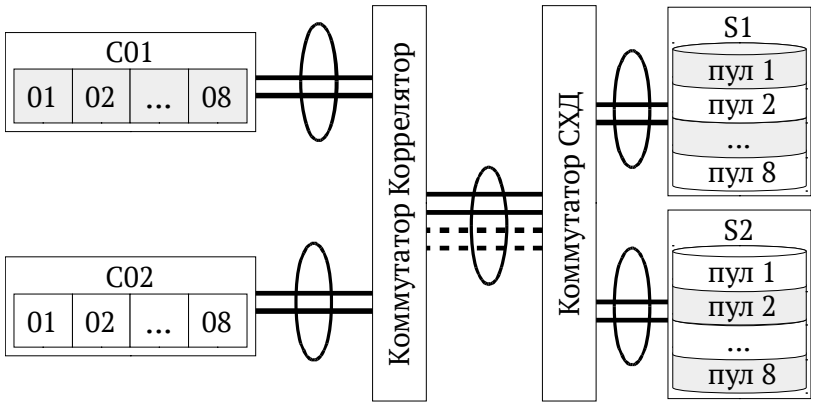

Рис. 3. Схема коммутации сети ЦКО РАН

Каждый сервер сети ЦКО РАН подключен к двум портам 10 GbE соответствующего коммутатора. На рисунке сплошной линией обозначено подключение в стандарте 10 GbE, прерывистой – планируемое дополнительное подключение. Каждая такая пара каналов объединена в один логический канал передачи данных по технологии LACP (Link Aggregation Control Protocol) [8] (на рис. 3 такое объединение отображено эллипсом). Алгоритм балансировки, т. е. выбора канала, по которыму будут передаваться данные, работает на основе расчета хеш-функции [9] от нескольких входных параметров. В качестве таких параметров используются MAC- и IP-адреса отправителя/получателя и номера TCP/UDP портов. Математически операцию выбора канала можно представить следующей формулой:

$$Port = \big((\, SrcIP \oplus DestIP) \oplus (SrcMac \oplus DestMac)\big)\ mod\ PortsInAggregateGroup.$$

В ходе анализа сетевого обмена данными между СХД и коррелятором выяснилось, что текущая схема хранения данных на СХД не позволяет полноценно использовать преимущества объединенного канала и получить скорость передачи данных близкую к 20 Гбит/с. Это объясняется особенностью работы узла коррелятора с данными, расположенными на сервере СХД. При установлении соединения и обмене данными входным параметром алгоритма балансировки является только одна пара MAC-, IP-адресов и портов NFS (Network File System) соединения для всех файлов каналов. Все вышеперечисленное приводит к тому, что в процессе передачи данных в агрегированном канале между сервером и вычислительным узлом задействован только один канал 10 GbE.

На рис. 4 представлена зависимость скорости передачи данных от времени между СХД и коррелятором при одновременной обработке нескольких часовых сессий за 15 часов.

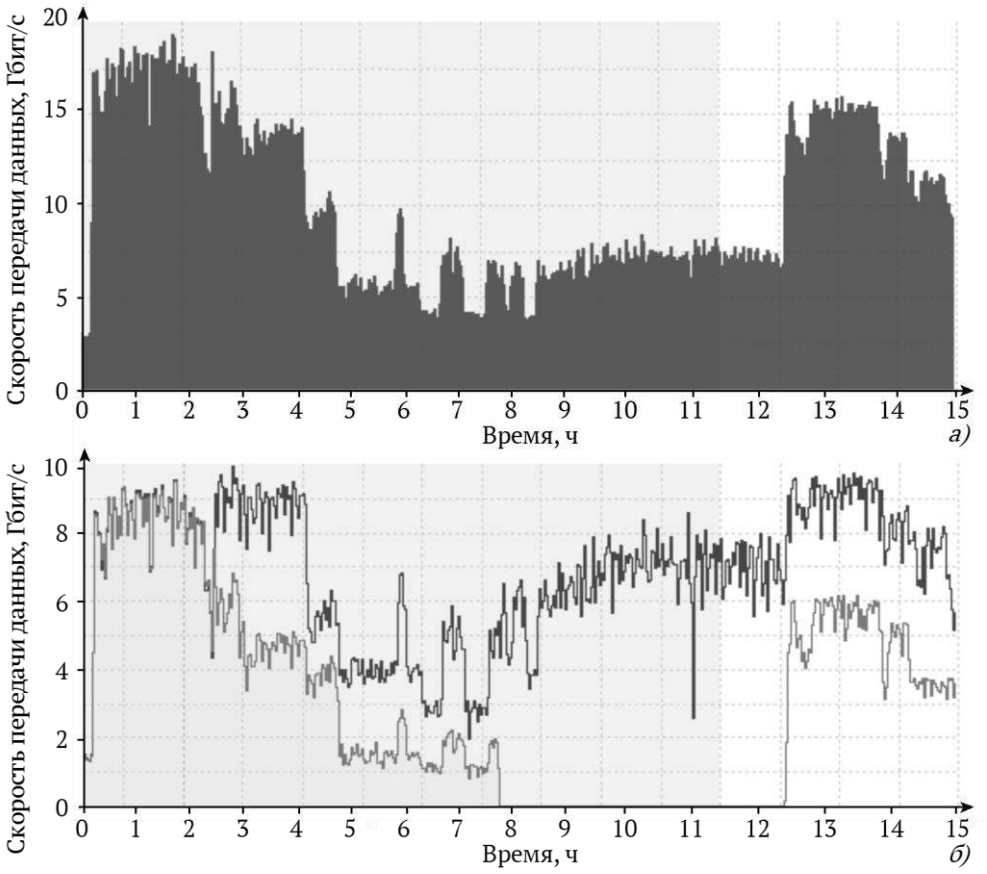

Рис. 4. Зависимость скорости передачи данных между СХД и программным коррелятором от времени: *а* — суммарная скорость передачи данных, *б* — распределение потока данных в агрегированном канале

Анализ данных, представленных на рис. 4, показывает, что величина максимальной скорости передачи данных составляет порядка 20 Гбит/с (при одновременной работе 6 узлов программного коррелятора). В момент обработки одной сессии (середина временной шкалы) скорость передачи данных ограничена величиной 10 Гбит/с, причем весь поток данных передается по одной физической линии агрегированного канала на 2 узла программного коррелятора.

Переход на новую динамическую схему размещения РСДБ-данных на N серверах приведет к тому, что у алгоритма балансировки появится больше входных параметров и возрастет вероятность использования всех физических линий связи во всех агрегированных каналах, следовательно, возрастет скорость обмена данными между СХД и программным коррелятором.

Проведенные согласно новому алгоритму эксперименты по обработке данных, размещенных в СХД, помогли выявить периодическое падение скорости передачи данных, которое вызвано, скорее всего, некорректной настройкой сетевого оборудования, и разработать рекомендации по реконфигурации сетевого оборудования.

Заключение

Переход на новый алгоритм хранения данных позволит:

— использовать все доступное место СХД независимо от числа каналов при наблюдении;

— обеспечить обслуживание/ремонт части оборудования СХД без потери управления и отказоустойчивости всей системы хранения данных;

— увеличить эффективность балансировки трафика на агрегированном интерфейсе.

Совместно со специалистами, обслуживающими кластер программного коррелятора, планируется проверить алгоритм настройки балансировки сетевого коммутатора, к которому подключены узлы коррелятора, расширить агрегированный канал между коммутаторами коррелятора и СХД, а также включить узлы коррелятора в систему мониторинга информационной сети ИПА РАН.

Работа выполнена с использованием оборудования ЦКП/УНУ «Радиоинтерферометрический комплекс «Квазар-КВО» и ЦКП «Суперкомпьютерный центр «Высокопроизводительные вычисления в радиоастрометрии и космической геодезии».

Литература

1. *Ипатов А. В., Гаязов И. С. и др.* РСДБ-система нового поколения // Труды ИПА РАН. — СПб.: ИПА РАН, 2013. — Вып. 10. — С. 216–222.

2. *Ipatov A. V., Ivanov D. V.* New Generation VLBI: Intraday UT1 Estimations / International VLBI Service for Geodesy and Astrometry 2016 General Meeting Proceedings “New Horizons with VGOS” / ed. by Behrend D., Baver K. D., and Armstrong K. L. — NASA/CP–2016–219016. — Hanover: NASA Center for AeroSpace Information, 2016 — P. 106–110.

3. *Яковлев В. А., Безруков И. А., Сальников А. И., Вылегжанин А. В.* Передача РСДБ-данных в широкополосном канале связи // Труды ИПА РАН. — СПб.: ИПА РАН, 2016. — Вып. 39. — С. 127–133.

4. *Суркис И. Ф., Зимовский В. Ф., Кен В. О., Курдубова Я. Л., Мишин В. Ю., Мишина Н. А., Шантырь В. А.* Программный РСДБ-коррелятор на гибридном процессорном кластере // Труды ИПА РАН. — СПб.: ИПА РАН, 2015. — Вып. 33. — С. 64–71.

5. EVN (The European VLBI Network) [Электронный ресурс]. — URL: https://deki.mpifr-bonn.mpg.de/Working_Groups/EVN_TOG/Disk_Inventory (дата обращения: 15.07.2018).

6. *Безруков И. А., Сальникова А. И., Яковлева В. А., Вылегжанин А. В.* Исследование производительности дисковой подсистемы системы буферизации и передачи данных // Приборы и техника эксперимента. — М.: Наука, 2018 — Вып. 4. — С. 5–10.

7. *Маршалов Д. А., Бердников А. С. и др.* Результаты предварительных испытаний широкополосной цифровой системы преобразования сигналов для радиотелескопов // Труды ИПА РАН. — СПб.: ИПА РАН, 2015. — Вып. 32. — С. 27–33.

8. LAG and LACP Command Reference — URL: https://www.cisco.com/c/en/us/td/docs/optical/cpt/r9_3/command/reference/cpt93_cr/cpt93_cr_chapter_01000.pdf (accessed: 05.07.2018).

9. *Яковлев В. А., Безруков И. А., Сальников А. И., Вылегжанин А. В.* Система мониторинга информационной сети ИПА РАН // Труды ИПА РАН. — СПб.: ИПА РАН, 2018. — Вып. 46. — С. 126–131.

A New Algorithm to Store and Transfer VLBI Data to the Software Correlator of the Russian Academy of Sciences

I. A. Bezrukov, A. V. Vylegzhanin, Y. L. Kurdubova, V. Y. Mishin, A. I. Salnikov, V. A. Yakovlev

Every day about 20 Tbytes of observational VLBI data which are received from the RT-13 radio telescopes (IAA RAS “Badary” and “Zelenchukskaya” observatories) are buffered preliminarily in the IAA RAS data storage system in St. Petersburg. Then these data are transferred to the software correlator in the IAA RAS Correlation Processing Center. Data integrity and high speed of its delivery are the main requirements for the data exchange process between the data storage system and the correlator. The data storage system servers and the correlator are connected to the RAS Correlation Processing Center local network through a switchboard with 10 GbE ports to meet these requirements. Our article contains recommendations to improve the software of the storage system which are based on the data storage system operating experience. The upgraded software of the data storage systems will increase the disc memory, the speed of data exchange, and fault-tolerance.

Keywords: VLBI, “Quasar” VLBI network, RT-13, VLBI Global Observing System (VGOS), data storage system, Link Aggregation Control Protocol (LACP), software correlator, HDD arrays.