СОЗДАНИЕ КЛАСТЕРНОГО ГРИДА НА БАЗЕ ВЫСОКОПРОИЗВОДИТЕЛЬНЫХ РЕСУРСОВ СФУ

С.В. Маколов, Д.А. Кузьмин, А.П. Бугай

Введение

В настоящее время Сибирский федеральный университет (СФУ) является одним из крупнейших научно-образовательным центром в Сибирском Федеральном округе. В СФУ сосредоточено большое количество вычислительных ресурсов, основными из которых являются три кластерные системы «IBM 1350» общей производительностью более 10 Тфлопс, предназначенные для ведения расчетов с большим объемом входных данных, с которыми обычные персональные компьютеры не справляются.[1] Кластера представляют собой отдельные вычислительные системы с отдельными точками входа, расположенные географически в разных помещениях, и обслуживают они разные группы пользователей.[2]

Имея несколько вычислительных систем, организации сталкиваются с проблемой оптимального их использования. Это связано с неравномерной загрузкой вычислительных ресурсов, с применением в определенных сферах деятельности организации, со спецификой задач решаемых на них. Очень часто можно наблюдать ситуации, когда одна из систем загружена уже на 100% и еще стоит очередь из желающих запустить задачу на счет, хотя другие вычислительные системы могут не использоваться вовсе. Любая перегрузка сказывается отрицательно на всем, в том числе и на вычислительной системе – суммарная производительность системы снижается, а, следовательно, и снижается производительность всего комплекса, состоящего из них.

Разрозненность систем так же создает большие неудобства в процессе их обслуживания. Системы периодически приходится обновлять, переконфигурировать. Одни и те же действия приходится проделывать на трех системах из-за того, что точки управления у систем свои. Тем самым, администратор тратит в 3 раза больше времени на одну итерацию, которую необходимо выполнить на 3-х системах.

Мониторинг использования этих систем показывает, что достаточно часто возникают ситуации, когда один ресурс загружен на 70%, а другой не используется вовсе, либо загружен менее чем на 5%. Если рассмотреть данную ситуацию со стороны износа оборудования, рассматривать необходимо каждую систему в отдельности и с углублением до каждого вычислительного узла кластера, т.к. одна система работает с большей нагрузкой и, следовательно, вероятность поломки нагруженной системы в разы больше менее загруженной. Это подтверждается и собранными статистическими данными по используемым в комплексе высокопроизводительным кластерам.

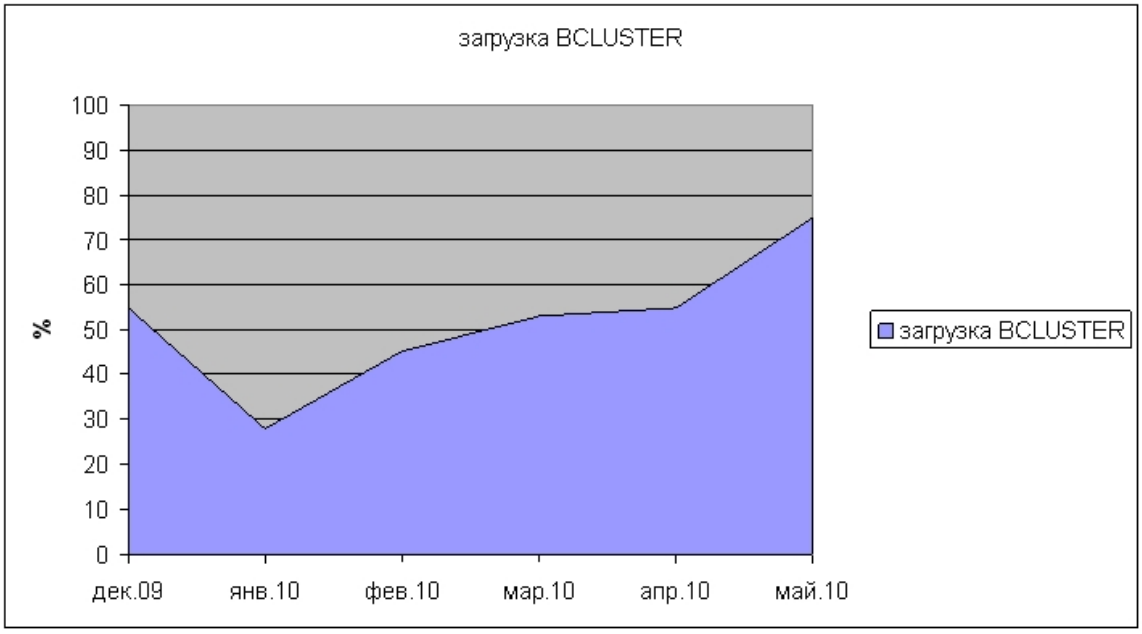

Рис.1 Загрузка кластерной системы BCLUSTER

На рисунке 1 показана полугодовая загрузка кластерной системы «BCLUSTER». Загрузка двух учебных кластерных систем «SCLUSTER1» и «SCLUSTER2» значительно ниже. Это связано, прежде всего, со спецификой их использования и уровнем ресурсоемкости студенческих задач.

Если проанализировать статистические данные по поломкам за тот же год, то видна закономерность: «Чем выше уровень загрузки вычислительных ресурсов, тем больший процент поломок приходится на них»

(таблица 1). Причем поломки происходят на тех вычислительных ресурсах, которые загружены на 100%. Вычислительные узлы с загрузкой менее 70% не нашли место в статистике по поломкам.

К тому же пользователи, использующие вычислительные ресурсы Института космических и информационных технологий Сибирского федерального университета получают доступ в виде учетных записей пользователей операционной системе семейства Linux, что не удобно для значительной части, пользователей, привыкших к повседневно используемой операционной системе семейства Windows.

Можно с уверенностью сказать, что для более эффективного и правильного использования данных ресурсов необходимо создать систему, которая позволила организовать:

- единую систему управления всеми ресурсами;
- единую, прозрачную систему постановки задач и получения результатов пользователям;
- систему, эффективно использующую вычислительные ресурсы ИКИТ СФУ.

Таблица 1

Соотношение загрузки кластера и процент поломок оборудования

Кластерная система	Средняя загрузка за 6 месяцев года*,%	Процент поломок**,%
BCLUSTER	51,8	18
SCLUSTER1	3,2	3
SCLUSTER2	4,6	5

Проблема управления распределенными вычислительными ресурсами не является уникальной только для СФУ - с этими проблемами сталкиваются организации, имеющие в своем составе мощные вычислительные ресурсы.

В середине 90-х годов XIX века двое американских учёных Ян Фостер (Ian Foster) и Карл Кессельман (Karl Kesselman) своей книгой "The Grid: Blueprint for a New Computing Infrastructure" впервые широко озвучили новую распределенную вычислительную систему – Грид, дали определение Грид, – это гибкое, защищённое, координированное совместное использование ресурсов группами пользователей, организаций и других ресурсов.[3] Идеи книги приобрели четкие очертания после выхода в свет двух статей, «The Anatomy of the Grid» и «The Physiology of the Grid», в которых описывается архитектура и требования к инфраструктуре Грид-сети. В 2001 г. было переопределено понятие Грид, как «скоординированное разделение ресурсов и решение проблем в динамической, многокомпонентной виртуальной организации», где виртуальная организация - это группа предприятий, объединяющих свои вычислительные ресурсы в единую Грид-систему и совместно их использующая.[4,5]

Грид-системы включают в себя службы по управлению и администрированию вычислительных ресурсов, службы, предоставляющие прозрачный и удобный способ постановки задач на счет, а так же осуществляют контроль исправности ресурсов и используют системы управления задачами, равномерно распределяющими нагрузку на эти ресурсы.

Опираясь на существующий мировой опыт построения грид и имеющиеся технологии в данной сфере распределенных вычислительных систем необходимо было рассмотреть возможность создания грид на базе вычислительных ресурсов ИКИТ СФУ, предложить архитектура грид-системы СФУ, проработать возможный путь её создания и внедрения.

Структура грид-системы СФУ

Для реализации грид-системы на базе вычислительных ресурсов ИКИТ СФУ необходимо было решить, какие изменения необходимо провести в структуре вычислительной сети, соединяющей вычислительные ресурсы, и какое промежуточное программное обеспечение нужно использовать для получения эффективного создаваемого вычислительного пространства.

Первым этапом создания грид-системы ИКИТ СФУ стало изучение особенностей построения корпоративной сети СФУ для представления проекта по реализации локальной грид-системы. В сети СФУ сосредоточено около 300 активных коммутирующих и маршрутизирующих устройств, обеспечивающих комфортную работу персоналу и студентам СФУ. Архитектура ЛВС СФУ реализована в виде иерархических уровней: уровень доступа, уровень распределения и уровень ядра, каждый из которых решает свои задачи. Это позволяет безболезненно для сети добавлять различные уровни, расширяющие функциональные возможности и решаемые сетью задачи; минимизировать ресурсные затраты для поиска и устранения неисправностей в сети.

К уровню Доступа на сегодняшний день подключено около 10 тысяч пользовательских устройств. Это и компьютеры в компьютерных классах, и пользовательские станции сотрудников СФУ, и мобильные

устройства, подключенные по беспроводным каналам, и серверные станции. В это число устройств входит и три кластерные системы ИКИТ СФУ, доступ к которым возможен, как и из внутренней сети СФУ, так и с глобальной сети Интернет. (Рис.2)

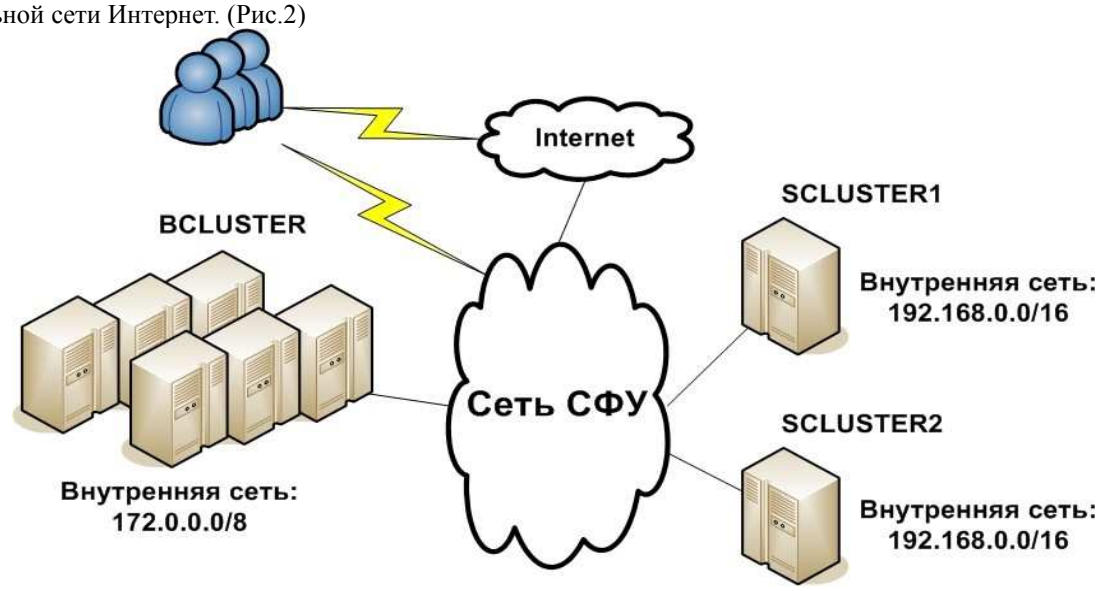

Рис.2 Подключение кластеров к ЛВС СФУ

Т.к. данные кластерные системы расположены в одном корпусе (здании), то была возможность изменить сетевую структуру комплекса без значительных затрат. Для организации функциональной единой системы управления необходимо было организовать единую сеть управления комплекса, независящую от сети СФУ. Это, прежде всего, связано с большим количеством активных и пассивных сетевых узлов, а также протяженными кабельными системами между кластерами, входящими в создаваемую грид-систему. При поломке сетевого оборудования, либо при повреждении кабеля, произойдет нарушение целостности системы управления, которое может привести к серьезным последствиям, поэтому должно быть сведено к минимуму влияние внешних факторов, приводящими к отказу системы управления.

Для создания функциональной единой системы управления ресурсами, взаимодействуя с Информационно-телекоммуникационным комплексом СФУ и учитывая их требования, была спроектирована архитектура грид-системы СФУ (Рис.3), в том числе была учтено создание сети хранения данных (SAN), предоставляющая более широкий и стабильный канал передачи данных пользователей между кластерами.

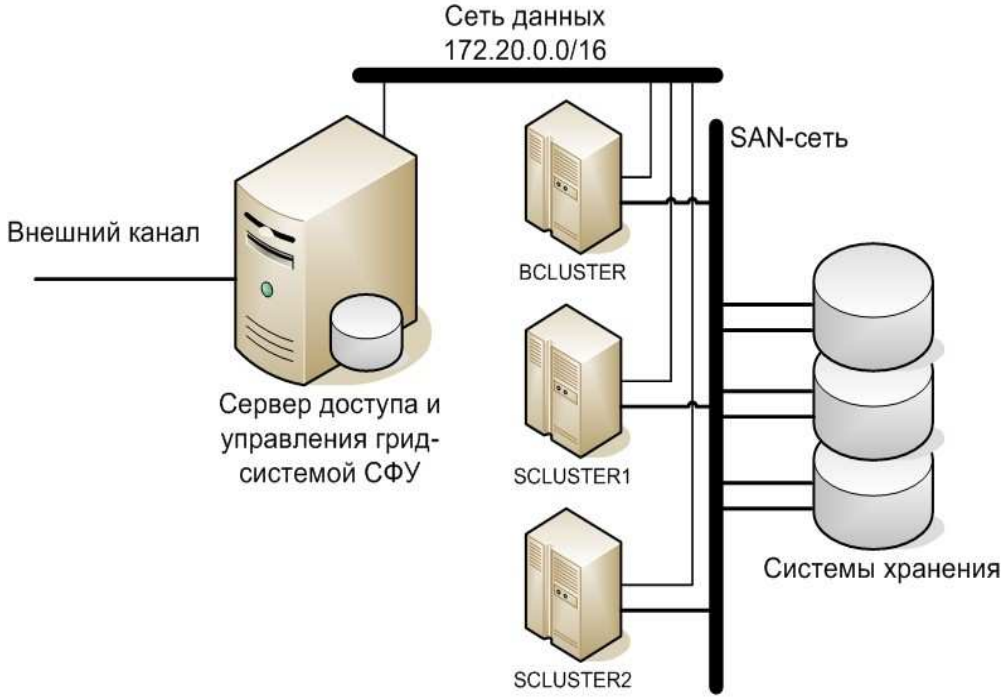

Рис.3 Архитектура грид-системы СФУ

Сервер доступа и управления грид-системой СФУ

В грид-системе основную роль ведет сервер доступа и управления грид-системой СФУ (Грид-шлюз), отвечающий за управление, мониторинг грид-системы, за равномерное распределение загружаемых на счет задач пользователей. Для пользователей и администраторов Грид-шлюз представлен в виде web-портала. Он предоставит пользователям интуитивно-понятный графический интерфейс для постановки задач, просмотра статистики поставленной задачи, скачивания готовых результатов. Для администраторов он будет являться единой точкой управления и мониторинга грид-системы СФУ в целом. Так же на Грид-шлюзе необходимо присутствие системы оповещения, которая будет оповещать по электронной почте и/или SMS-сообщениями пользователей о произошедших изменениях с поставленными задачами на счет (например, аварийная остановка задачи, успешное завершение задачи и т.д.), а администраторов о проблемах с оборудованием системы (например, выход из строя вычислительного узла, потеря соединения с удаленным ресурсом и т.д.).

Грид-шлюз в системе является элементом управления и мониторинга, две эти функциональные возможности накладывают требование высокой надежности и готовности. Обеспечить данные требования может создание грид шлюза на основе кластера серверов. Это система высокой надежности и готовности (high-availability (HA) system), отказоустойчивая компьютерная система, в которой в случае отказа гарантируются автоматическое восстановление работоспособности в течение нескольких секунд и сохранность данных. В такой системе сервисы не принадлежат какому-то конкретному серверу в кластере, а принадлежат кластеру целиком (кластеризованные сервисы). И в случае выхода одного сервера из строя, его сервисы быстро и автоматически начинают предоставляться другим сервером кластера.[6]

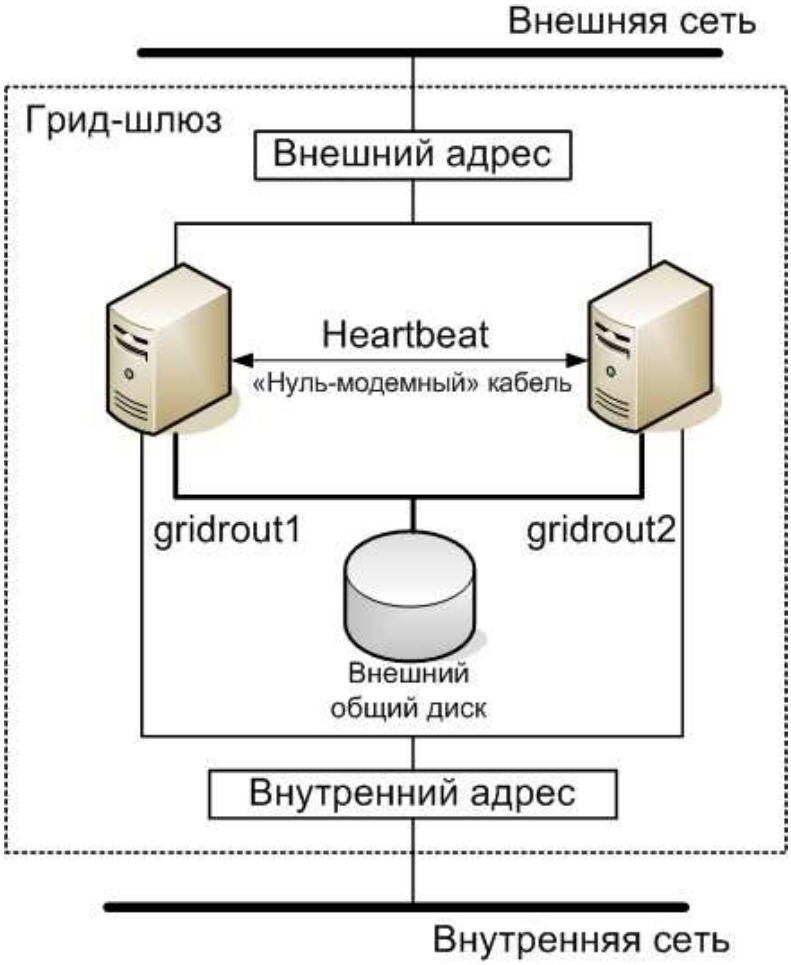

Рис.4 Грид-шлюз

Данный кластер представляет собой два сервера, объединённых логически, способных обрабатывать идентичные запросы и использующихся как единый ресурс. Тем самым при выходе из строя одного сервера из кластера серверов управления доступность ресурса (грид-система СФУ) не будет утеряна (Рис.4).

Требования для реализации HA-кластера:

- Аппаратная часть: 2 сервера с установленной на них операционной системой Linux, 1 общий внешний жесткий диск, 1 «нуль-модемный» последовательный кабель;
- Программная часть: Утилита Heartbeat.

Утилита Heartbeat, входит в состав любой операционной системы семейства Linux и предоставляет основные функции, требующиеся для любой HA-системы, например, запуск и останов ресурсов, мониторинг

доступности системы в кластере и передача прав владения общим IP-адресом между узлами кластера. Она следит за состоянием конкретной службы (или служб) по последовательному кабелю, интерфейсу Ethernet, либо по обоим. Для проверки состояния и доступности службы используется специальный квитирующий монитор heartbeat. Heartbeat предоставляет фундамент для сложных сценариев обработки не исправности узлов HA-систем.[7]

Сеть управления

Для осуществления единой системы мониторинга и управления грид-системой необходимо создать общую сеть управления для трех кластеров. Необходимо проложить отдельный от Сети СФУ канал по технологии Gigabit EtherChannel с пропускной возможностью 4 Гбит/сек. Это позволит уйти от зависимости от работоспособности сети СФУ, уйти от многочисленных коммутирующих устройств, от больших кабельных трасс, уменьшив вероятность недоступности узлов грид системы между собой, тем самым, увеличив надежность корпоративной грид-системы СФУ. Это позволит наблюдать за функционированием каждого вычислительного узла во всей грид-системе и уменьшит время на локализацию поломки.

SAN-сеть

Кластерные системы ИКИТ СФУ имеют в своем составе системы хранения IBM DS3400: BCLUSTER – 20 Тбайт, SCLUSTER1 и SCLUSTER2 – по 2,4 Тбайт. К системе хранения кластера BCLUSTER доступ осуществляется через оптические коммутаторы IBM SAN 16B2, т.к. доступ необходим нескольким серверам управления дисковыми массивами. Кластерные системы SCLUSTER1 и SCLUSTER2 имеют по одному серверу управления дисковыми массивами и поэтому подключение осуществлено напрямую (Рис.5). Чтобы любой узел грид-системы СФУ мог оперативно получить доступ к данным необходимо их объединить в одну SAN-сеть.

Для обеспечения доступа к системам хранения SCLUSTER1 и SCLUSTER2 необходимо предварительно подключить их к коммутаторам IBM SAN B24, располагающиеся в УЛК1-01, с уже подключенной к ним системой хранения IBM DS4700 с общим объемом памяти 50 ТБ, тем самым, получив возможность к расширению дискового пространства для размещения данных пользователей (Рис.6).

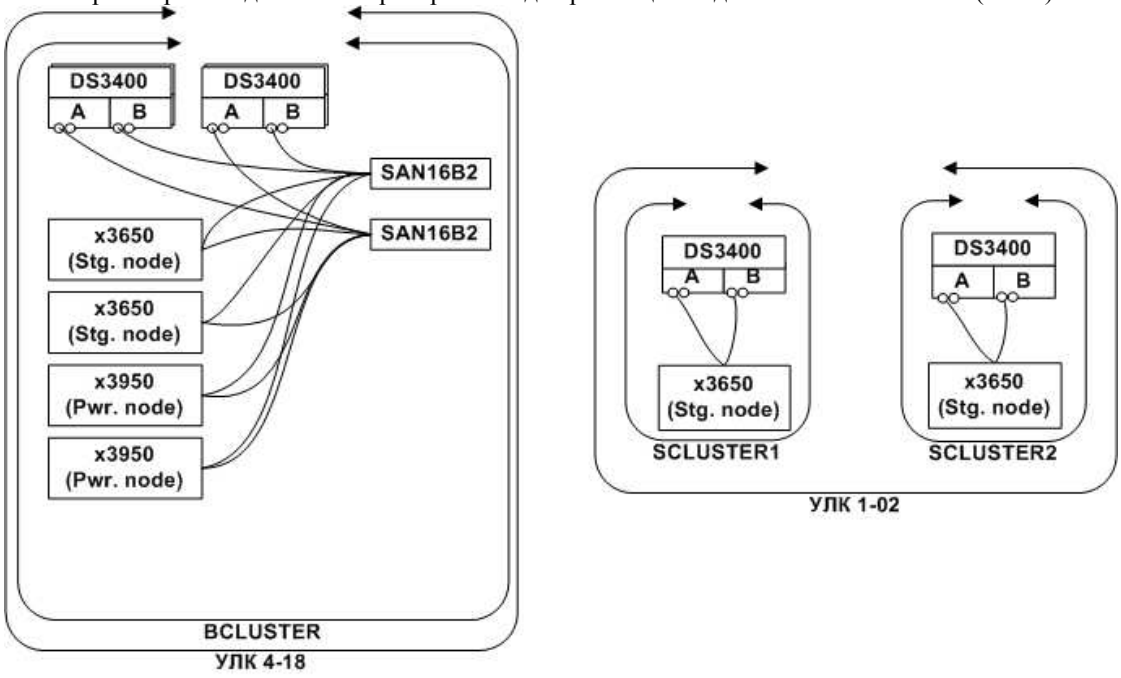

Рис.5 SAN-сеть до изменения в пределах ИКИТ СФУ

Для обеспечения доступа к системам хранения SCLUSTER1 и SCLUSTER2 необходимо предварительно подключить их к коммутаторам IBM SAN B24, располагающиеся в УЛК1-01, с уже подключенной к ним системой хранения IBM DS4700 с общим объемом памяти 50 ТБ, тем самым, получив возможность к расширению дискового пространства для размещения данных пользователей (Рис.6). Затем необходимо соединить пары оптических коммутаторов IBM SAN 16B2 и IBM SAN B24, находящиеся в BCLUSTER и УЛК1-01, соответственно. Так как данные пользователей находятся на системах хранения IBM DS3400, входящая в кластерные системы BCLUSTER, SCLUSTER1 и SCLUSTER2, интенсивный обмен данными будет происходить между системами хранения IBM DS3400, расположенных в УЛК1-02 и УЛК4-18. Канал связи, соединяющий эти 2 коммутатора, будет активно использоваться, поэтому он должен обеспечивать максимальную скорость передачи информации между ними.

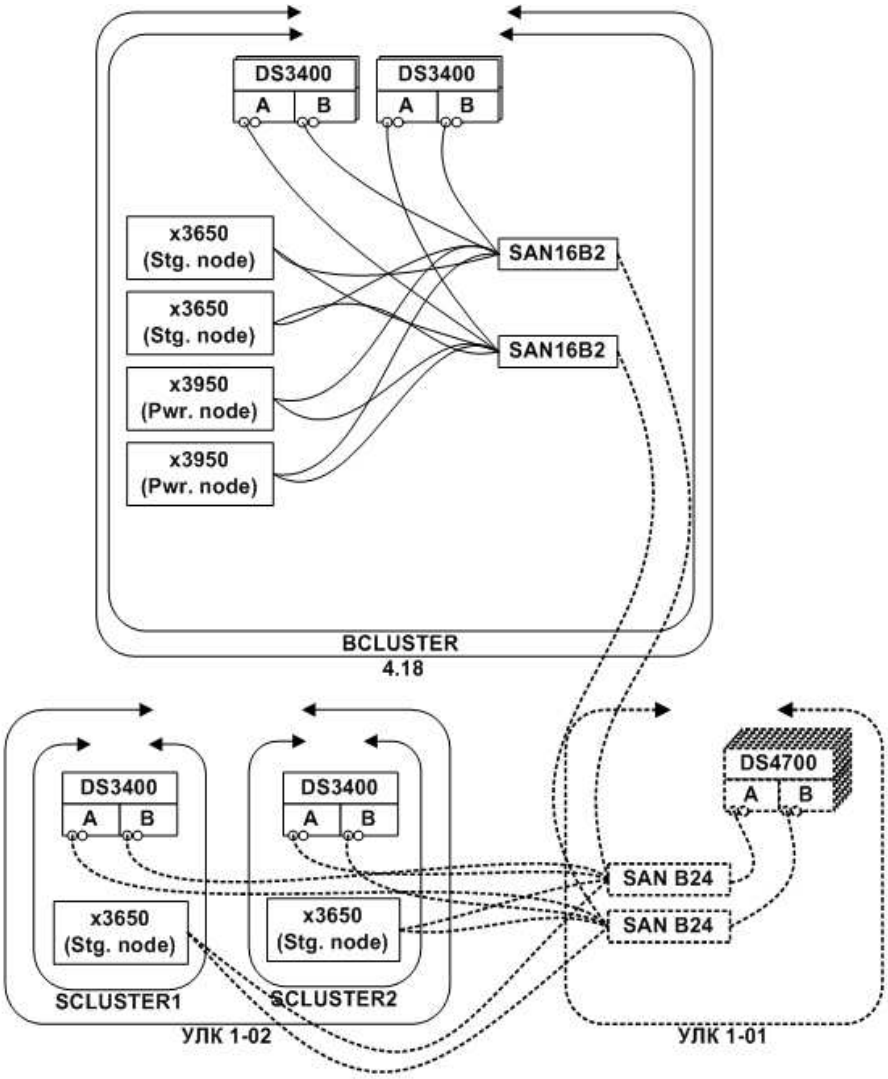

Рис.6 SAN-сеть после изменения в пределах ИКИТ СФУ

Программное обеспечение грид-системы СФУ

Функционирование ни одной из существующих вычислительных систем не возможно без программного обеспечения (операционная система, драйвера, приложения). Так и грид-система, являясь распределенной вычислительной системой, не может функционировать без специализированного программного обеспечения.

Важнейшим компонентом грид-инфраструктуры является промежуточное программное обеспечение (ППО), которое предназначено управлять заданиями, обеспечивать безопасный доступ к данным большого объема в универсальном пространстве имен, перемещать и тиражировать данные с высокой скоростью из одного географически удаленного узла на другой и организовывать синхронизацию удаленных копий.

Не менее важным компонентом грид-инфраструктуры являются и системы пакетной обработки задач (Portable Batch System, PBS). Основное назначение системы пакетной обработки заданий состоит в запуске программы на исполнение на тех узлах кластера, которые в данный момент не заняты обработкой других заданий, и в буферизации задания, если в данный момент отсутствуют свободные ресурсы для его выполнения. Большинство подобных систем предоставляют и множество других полезных услуг.

Заключение

Создание грид-системы на базе распределенных вычислительных ресурсов СФУ, позволило организовать единую систему управления всеми ресурсами; единую, прозрачную систему постановки задач и получения результатов пользователям; систему, эффективно использующую ресурсы ИКИТ СФУ.

Создана общая сеть управления Gigabit Ethernet (отдельная от Сети СФУ), которая обеспечила реализацию единой системы мониторинга и управления грид-системой, что позволило уйти от зависимости от работоспособности сети СФУ, уйти от многочисленных коммутирующих устройств, от больших кабельных трасс, уменьшив вероятность недоступности узлов грид системы между собой, тем самым, увеличив

надежность корпоративной грид-системы СФУ. Так же дало возможность наблюдать за функционированием каждого вычислительного узла во всей грид-системе и уменьшило время на локализацию поломок.

Реализована единая сеть хранения данных, что в перспективе можно развить в создание вычислительного корпоративного «облака» СФУ, предоставляющего стандартные сервисы, необходимые для обеспечения учебного процесса. Проработан и реализован путь вхождения создаваемой локальной грид-системы СФУ в российское грид-сообщество «ГридННС»[8].

Вступив в данное сообщество, СФУ будет иметь возможность использовать, не только вычислительные, но и программные ресурсы всего вычислительного потенциала общероссийского грид-проекта «ГридННС».

ЛИТЕРАТУРА:

1. Комплекс высокопроизводительных вычислений ИКИТ СФУ. [Электронный ресурс]. – Режим доступа: http://cluster.sfu-kras.ru/ свободный.
2. Бугай А.П., Маколов С.В. "Создание комплекса высокопроизводительных вычислений Сибирского федерального университета". /Высокопроизводительные параллельные вычисления на кластерных системах: материалы Девятой международной конференции-семинара/ Владимирский государственный университет, г. Владимир, 2009 г с.64-65.
3. I.Foster, K.Kesselman "The Grid: Blueprint for a New Computing Infrastructure". Morgan Kaufmann Puublishers, Inc., San Franscisco, California,1999
4. I.Foster, C.Kesselman, St.Tuecke "The Anatomy of the Grid", International Journal of High Performance Computing Applications, 2001
5. I.Foster, C.Kesselman, J.Nick, St.Tuecke "The Physiology of the Grid", Global Grid Forum, 2002
6. Программное обеспечение повышенной готовности промежуточного уровня в Linux. [Электронный ресурс]. – Режим доступа: http://www.ibm.com/developerworks/ru/linux/library/l-halinux свободный.
7. Linux-HA. [Электронный ресурс]. – Режим доступа: http://www.linux-ha.org/wiki/Main_Page свободный.
8. ГридННС. [Электронный ресурс]. – Режим доступа: http://www.ngrid.ru/trac/ свободный.