На правах рукописи

Петров Дмитрий Леонидович

Алгоритмы миграции данных в высокомасштабируемых облачных системах хранения

05.13.11 — Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Санкт-Петербург — 2011

Работа выполнена в Санкт-Петербургском государственном электротехническом университете «ЛЭТИ» им. В.И. Ульянова (Ленина) (СПбГЭТУ)

Кафедра математического обеспечения и применения ЭВМ (МО ЭВМ)

Научный руководитель:	к.т.н., доцент кафедры МО ЭВМ СПбГЭТУ Татаринов Юрий Станиславович
Официальные оппоненты:	д.т.н., профессор кафедры «Вычислительной техники» СПбГЭТУ Водяхо Александр Иванович
	к.т.н., старший научный сотрудник «Центр перспективных исследований» Санкт-Петербургского государственного политехнического университета Писарев Андрей Сергеевич
Ведущая организация:	ОАО «Информационные телекоммуникационные технологии» (ОАО «Интелтех»)

Защита состоится «02» ноября 2011 г. в 15:00 на заседании совета по защите докторских и кандидатских диссертаций Д 212.238.01 СПбГЭТУ по адресу: 197376, Санкт-Петербург, ул. Профессора Попова, дом 5.

С диссертацией можно ознакомиться в библиотеке СПбГЭТУ.

Автореферат диссертации разослан «30» сентября 2011 г.

Ученый секретарь совета по защите
докторских и кандидатских диссертаций

Д 212.238.01, к.т.н. Щеголева Н.Л.

Общая характеристика работы

Актуальность работы

В последние годы произошло сильное изменение в практике и теории организации распределенных вычислительных систем связанное с появлением концепции cloud computing, или «облачных вычислений». Согласно этой концепции, вычислительные ресурсы арендуются по требованию через Интернет, а вычислительные системы лишь временно используют их для выполнения своих функций. Уже сегодня имеется возможность аренды вычислительных ресурсов с поминутной, и даже посекундной оплатой. Это позволяет создавать новые типы вычислительных систем с уникальными технико-экономическими характеристиками за счет гибкости в оплате и возможности арендовать потенциально бесконечное количество ресурсов.

Одними из самых востребованных современным бизнесом типов вычислительных систем являются системы хранения данных (СХД), которые представляют собой множество распределенных устройств хранения, объединенных вычислительной сетью и представленных пользователям в виде одного логического устройства большой емкости. Концепция облачных вычислений оказывает сильное влияние на современные СХД, что привело к появлению нового класса систем хранения - высокомасштабируемые облачные системы хранения данных (ВО-СХД, Scalable Storage Cloud). ВО-СХД, в отличие от традиционных СХД, используют в своем составе не фиксированное количество устройств хранения, а арендуют устройства по мере необходимости, или же высвобождают ряд устройств, когда необходимость в них отпадает. Для эффективного использования арендованных ресурсов ВО-СХД вынуждены регулярно производить процедуру масштабирования, т.е. изменения количества устройств хранения, входящих в систему.

Масштабирование выполняется операционной системой (ОС) ВО-СХД,

оно связано с переконфигурацией хранилища, т.е. с перемещением огромного количества элементов данных (блоков данных) между устройствами хранения. Масштабирование и переконфигурация неразрывно связаны с алгоритмами миграции данных, которые строят план миграции (перемещения элементов данных) на основе текущего и целевого распределения элементов данных по устройствам хранения. Выполнение миграции данных не должно приводить к снижению качества обслуживания клиентов системы хранения, для чего в алгоритмах необходимо учитывать пропускную способность сети и максимальный объем данных, который можно передавать в один момент времени с одного устройства хранения на другое.

Существующие алгоритмы миграции данных не учитывают особенности ВО-СХД, и высвобождение лишних устройств хранения (добавление новых устройств) возможно производить лишь после полного завершения процедуры переконфигурации. Во время выполнения этого длительного этапа лишние устройства остаются задействованными, а новые устройства - не до конца использованными. Разработка специализированных алгоритмов миграции данных для управления арендованными ресурсами в ВО-СХД, способных сократить время масштабирования, позволит повысить эффективность использования ресурсов и понизить затраты на аренду устройств хранения, что представляет собой **важную научную задачу**, имеющую большое практическое значение.

Целью диссертационной работы является разработка алгоритмов миграции данных в ВО-СХД, способных уменьшить время масштабирования, по сравнению с традиционными алгоритмами миграции. В соответствии с указанной целью, в работе сформулированы и решены следующие **задачи**:

1. анализ существующих моделей СХД и алгоритмов миграции данных;

2. анализ особенностей ВО-СХД;

3. разработка модели ВО-СХД;

4. разработка алгоритмов миграции данных для ОС ВО-СХД;

5. анализ свойств разработанных алгоритмов.

Методы исследования. В исследовании формализация моделей производилась с помощью методов теории вычислительных систем и теории графов. Для описания и анализа алгоритмов в работе были использованы методы теории графов и теории алгоритмов.

Научная новизна работы заключается в следующем:

1. предложена математическая модель задачи миграции данных ВО-СХД, способная менять состав устройств хранения, и на основе модели сформулирована задача миграции данных в ВО-СХД в виде многокритериальной оптимизационной задачи;

2. предложены переборный и полиномиальный аппроксимационный алгоритмы миграции данных для управления вычислительными ресурсами ОС ВО-СХД, позволяющие предельно быстро производить масштабирование ВО-СХД;

3. произведен анализ свойств разработанных алгоритмов: доказана оптимальность алгоритмов по основному критерию «время масштабирования»; доказана полиномиальность вычислительной сложности аппроксимационного алгоритма; экспериментально показана эффективность алгоритмов, по сравнению с существующими.

Практическая значимость. Предложенные модель и алгоритмы могут быть использованы при разработке ОС промышленных ВО-СХД, что позволит улучшить такой технико-экономический показатель ВО-СХД, как затраты на аренду вычислительных ресурсов.

Внедрение результатов. Предложенные алгоритмы миграции данных в ВО-СХД, а также разработанная программная система, реализующая данные алгоритмы, использовались при выполнении НИОКР ЦНИТ-9/Мк «Cache algorithms with user demand prediction for cloud storages» (Алгоритмы кэширования с прогнозированием пользовательских требований для облачных хранилищ), проводимом в СПбГЭТУ. Срок выполнения: с 01.2010 до 12.2010. Источник финансирования - внебюджет.

На защиту выносятся следующие результаты и положения:

1. математическая модель миграции данных в ВО-СХД;

2. аппроксимационный алгоритм миграции данных для управления вычислительными ресурсами в ВО-СХД;

3. доказательство оптимальности предложенного алгоритма по первому критерию задачи миграции данных в ВО-СХД, а также доказательство его полиномиальности.

Апробация работы. Предлагаемые решения и результаты диссертационной работы докладывались и обсуждались на международных и всероссийских научно-технических конференциях в 2008-2011 гг.

Публикации. Материалы диссертации опубликованы в 5 научных работах, из них 2 статьи в рецензируемых журналах, рекомендованных ВАК, 3 работы в научных трудах конференций, из которых одна работа на английском языке в трудах международной конференции сообщества IEEE.

Структура и объем диссертации. Диссертация состоит из введения, пяти глав, заключения, списка литературы, включающего 77 наименований. Основная часть работы изложена на 113 страницах машинописного текста. Работа содержит 23 рисунка, 5 таблиц и 4 приложения общим объемом 18 страниц.

Содержание работы

Во Введении обоснована актуальность диссертационной работы, сформулирована цель и аргументирована научная новизна исследований, показана практическая значимость полученных результатов, представлены выносимые на защиту научные положения.

В первой главе анализируются основные подходы к организации СХД. Описываются архитектуры СХД и развитие архитектур, связанное с ростом потребностей бизнеса в данных. Описана концепция интеллектуальных СХД и различные типы интеллектуальных систем: устройства внешней памяти (DAS, Direct-attached Storage), сетевые системы хранения (NAS, Network Attached Storage) и сети хранения данных (SAN, Storage Area Network). Проанализированы основные подсистемы входящие в состав сетевой ОС распределенных хранилищ (рис. 1): коммуникационная подсистема (front-end), кэш и серверная подсистема (back-end).

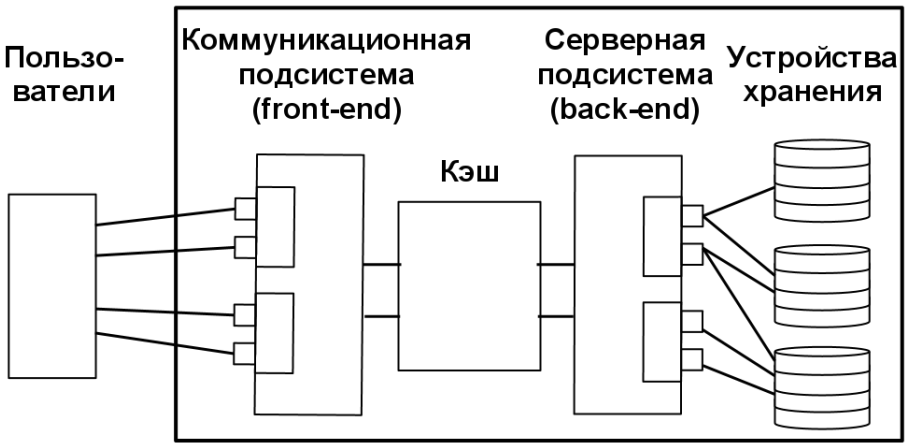

Рисунок 1. Структура СХД

Эффективность функционирования СХД зависит от отображения логической структуры на физическую. В главе **исследуется процедура переконфигурации СХД**, которая служит повышению эффективности отображения логической структуры на физическую организацию за счет перемещения элементов данных с высокозагруженных устройств хранения на менее

загруженные устройства. Эта процедура выполняется в компоненте переконфигурации серверной подсистемы СХД (рис. 2).

Рисунок 2. Серверная подсистема СХД

Выделяют три фазы процедуры переконфигурации в СХД: вычисление (псевдо-) оптимального плана распределения данных; вычисления (псевдо-) оптимального плана миграции данных, согласно текущему расположению данных и (псевдо) оптимальному плану распределения; выполнение процедуры миграции данных, согласно плану миграции.

Длительность процедуры переконфигурации определяется длительностью процедуры миграции данных, т.к. миграция связана с перемещением большого количества элементов данных между устройствами хранения, причем ее выполнение не должно приводить к снижению производительности СХД. Таким образом, **было выявлено**, что от эффективности вычисления плана миграции напрямую зависит время выполнения переконфигурации СХД, и, соответственно, эффективность работы СХД.

Во второй главе рассматривается концепция облачных вычислений, или «cloud computing», согласно которой вычислительные ресурсы не покупаются, а предоставляются в виде услуги, т.е. арендуются. **Описана ис-**

тория термина «облачные вычисления», связь концепции облачных вычислений с GRID системами. Описаны наиболее популярные сервисы, предоставляющие ресурсы, в соответствии с этой концепцией: Amazon Web Services[1], Microsoft Azure[2] и другие. Такие службы, являющиеся вычислительной инфраструктурой для других служб или сервисов, относятся к классу IaaS сервисов (Infrastructure As A Service).

С практической точки зрения, наиболее интересными устройствами, использующими арендованные ресурсы, являются ВО-СХД, которые способны масштабироваться, т.е. динамически менять состав устройств хранения во время функционирования (рис. 3). **Произведен анализ** существующих ВО-СХД и особенностей их реализации. Основной задачей ВО-СХД, кроме обеспечения должного качества обслуживания пользователей, является минимизация использования арендованых IaaS ресурсов.

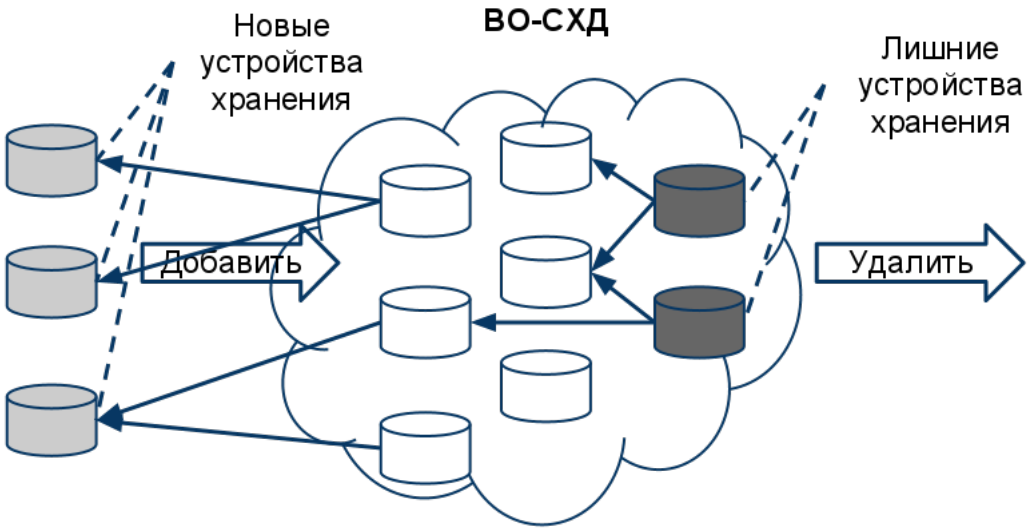

Рисунок 3. Добавляемые (светло-серые) и удаляемые (темно-серые) устройства хранения

Масштабирование ВО-СХД, т.е изменение состава устройств хранения, производится во время процедуры переконфигурации. Минимизация использования арендованных ресурсов требует предельно быстрого выполнения процедуры масштабирования, для чего требуется минимизировать время мигра-

[1] http://aws.amazon.com

[2] http://www.microsoft.com/windowsazure/

ции данных с учетом специфики ВО-СХД. В главе был **сделан вывод** о необходимости разработки новых алгоритмов миграции данных для ОС ВО-СХД, способных сократить время масштабирования до завершения полной процедуры миграции. Алгоритмы миграции данных в облачных хранилищах должны разделять процедуру миграции на две части: масштабирование СХД и остаточную миграцию (рис. 4). Основной целью алгоритмов должно являться сокращение времени масштабирования на некоторую величину Δ, при возможной потере эффективности общего времени переконфигурирования СХД.

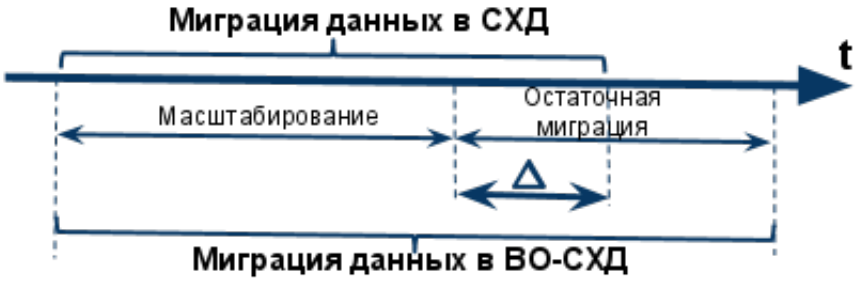

Рисунок 4. Сокращение времени масштабирования в процедуре миграции данных (Δ)

В третей главе анализируются существующие математические модели СХД и алгоритмы миграции данных. В главе показано, что входными данными задачи миграции данных являются выходные данные задачи распределения данных и текущая конфигурация СХД, т.е. новое(целевое) распределение данных по устройствам хранения и текущее распределение. Эти данные можно представить в виде ненаправленного мультиграфа G демонстрирующего требования по перемещению элементов данных между устройствами хранения. Мультиграф G является моделью задачи миграции данных (рис. 5). В результате чего, план миграции можно разбить на шаги.

О п р е д е л е н и е 0.1. *Модель задачи миграции данных - это граф* $G = \langle D, E, W \rangle$, *в котором выполняются условия:* $\forall v, w \in D, W(v, w) = 0$ *если* $(v, w) \notin E$ *и* $\forall v, w \in D \Rightarrow W(v, w) = W(w, v)$, *где* D *- устройства хранения,* $E \subseteq D \times D$ *- операции перемещения данных,* $W : E \to \mathbb{N}$ *- весовая функция мультиграфа.*

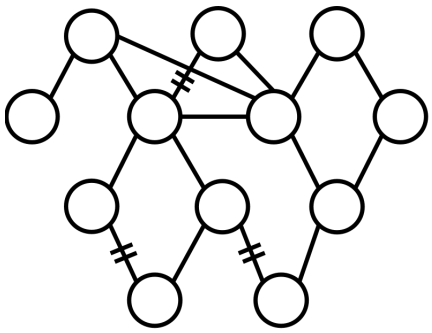

Рисунок 5. Модель миграции данных (двумя черточками выделены ребра кратные двум)

Задача миграции данных заключается в составлении плана миграции элементов данных СХД, при котором каждое устройство хранения в один момент времени участвует не более, чем в одной операции перемещения данных. Т.о. задача миграции данных сводится к задаче раскраски ребер мультиграфа, т.е. к задаче поиска хроматического индекса мультиграфа χ', которая принадлежит классу NP-полных задач. Задачу можно выразить более формально на основе модели миграции данных:

О п р е д е л е н и е 0.2. *Задача миграции данных в СХД заключается в поиске реберной раскраски мультиграфа модели миграции данных G (определение 0.1), обладающей минимальным количеством цветов.*

В главе **приводится описание ряда алгоритмов** из теории графов, которые в дальнейшем используются для разработки алгоритма миграции данных в ВО-СХД.

А л г о р и т м 0.1 (Алгоритм Шеннона) *Полиномиальный аппроксимационный алгоритм Шеннона для раскраски ребер мультиграфа имеет вычислительную сложность $O(|E|(\chi' + |D|))$, где E - множество ребер мультиграфа, D - множество вершин, χ' - хроматический индекс графа.*

Задача раскраски ребер двудольного мультиграфа является частным случаем задачи раскраски ребер мультиграфа, но она принадлежит классу

полиномиальных задач. Т.о. для ее решения существуют оптимальные алгоритмы с полиномиальной вычислительной сложностью.

А л г о р и т м 0.2 (Раскраска ребер двудольного мультиграфа) *Полиномиальный, оптимальный (не аппроксимационный) алгоритм раскраски ребер двудольного мультиграфа имеет вычислительную сложностью* $O(|E|\log|E|)$, *где* E *- множество ребер.*

В четвертой главе предложена модель миграции данных ВО-СХД (определение 0.3), которая построена на основе модели миграции СХД (определение 0.1) путем выделения подмножества масштабируемых устройств хранения D_S. D_S определяются как подмножество устройств хранения, которые необходимо вывести из строя или добавить к ВО-СХД в результате переконфигурации. Причем среди операций передачи данных не может быть операций передачи с одного масштабирующего устройства на другое.

О п р е д е л е н и е 0.3. *Моделью миграции данных ВО-СХД будем называть модель миграции данных с явно выделенным масштабирующим подмножеством* D_S*:* $G = \langle D, E, W, D_S\rangle$, *в котором выполняется условия:*

$$\begin{aligned}
&\forall v, w \in D, W(v,w) = 0 \text{ } \textit{если} \text{ } (v,w) \notin E \text{ - } \textit{из модели миграции;}\\
&\forall v, w \in D \Rightarrow W(v,w) = W(w,v) \text{ - } \textit{из модели миграции;}\\
&\forall v, w \in D_S \subseteq D \Rightarrow W(v,w) = 0 \text{ - } \textit{из определения } D_S.
\end{aligned} \tag{1}$$

Основываясь на предложенной модели 0.3, сформулирована задача миграции данных в ВО-СХД в терминах теории графов. Для ее точной формулировки введены понятия: масштабирующий подграф G_{scal} и остаточный подграф G_{res} модели миграции ВО-СХД. Для чего введено понятие подмножества смежных устройств хранения D_{adj} как подмножества устройств, смежных с устройствами из D_S. Первый подграф G_{scal} включает все элементы G

(устройства и операции передачи данных), имеющие отношение к масштабированию (т.е. все устройства D_S и D_{adj} и ребра их соединяющие), второй подграф G_{res} включает все остальные элементы графа G, не вошедшие в G_{scal}. На рисунке 6 изображена модель задачи миграции данных, разделенная на два подграфа G_{scal} и G_{res}. Черным цветом отмечены масштабируемые устройства D_S, серым - устройства D_{adj}. На основе предложенной модели **сформулирована задача миграции данных** в ВО-СХД в виде многокритериальной оптимизационной задачи.

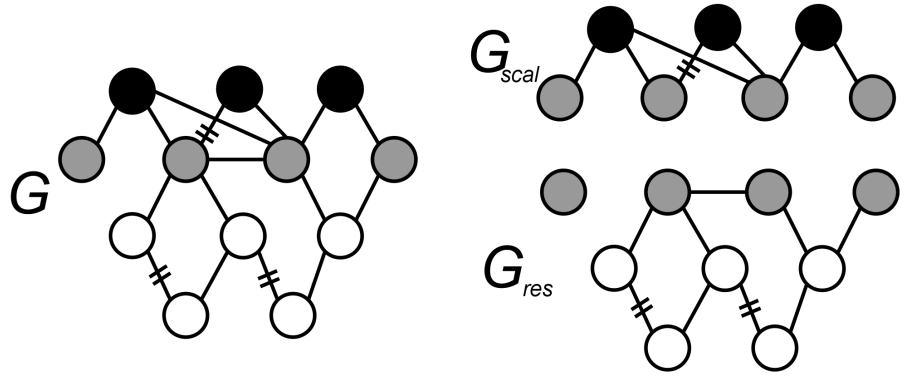

Рисунок 6. Разбиение модули миграции ВО-СХД G на два подграфа: G_{scal} и G_{res}

О п р е д е л е н и е 0.4. *Задача миграции данных в ВО-СХД G - это многокритериальная задача оптимизации плана миграции. Основным критерием задачи является план миграции в подграфе G_{scal}, а второй критерий - план миграции в подграфе G_{res}.*

Для решения сформулированной задачи **предложен переборный алгоритм** миграции данных в ВО-СХД, а также **предложен полиномиальный аппроксимационный алгоритм**.

А л г о р и т м 0.3 (Полиномиальная миграции данных)

1. *Выделить подграф G_{scal} из графа G на основе подмножества D_S.*

2. *Выделить подграф G_{res} из G на основе подграфа G_{scal}.*

3. *Использовать полиномиальный алгоритм раскраски двудольного мультиграфа (алгоритм 0.2) для вычисления плана миграции в G_{scal}.*

4. *Использовать аппроксимационный полиномиальный алгоритм Шеннона (алгоритм 0.1) для вычисления плана миграции мультиграфа G_{res}.*

5. *Получить общий план миграции ВО-СХД путем последовательного объединения планов миграции подграфа G_{scal} с планом подграфа G_{res}.*

А л г о р и т м 0.4 (Переборная миграция данных) *Переборный алгоритм миграции аналогичен полиномиальному алгоритму. Единственным отличием является использование оптимального переборного алгоритма раскраки ребер мультиграфа на шаге 4.*

Алгоритмы используют свойство двудольности подграфа G_{scal}, которое доказывается на основе его определения. С практической точки зрения, двудольность G_{scal} означает, что во время масштабирования устройства D_S не обмениваются данными между собой. Двудольность позволяет использовать существующие точные полиномиальные алгоритмы (а не аппроксимационные) для получения оптимального плана миграции в подграфе G_{scal}.

Доказывается оптимальность алгоритмов по первому критерию задачи миграции данных в ВО-СХД, а также **доказывается полиномиальность аппроксимационного алгоритма** (0.3).

Т е о р е м а 0.1 (Оптимальность алгоритмов 0.3 и 0.4) *Предложенные алгоритмы 0.3 и 0.4 оптимальны по первому критерию задачи миграции данных в ВО-СХД (определение 0.4).*

Т е о р е м а 0.2 (Полиномиальность алгоритма 0.3) *Предложенный алгоритм миграции данных в ВО-СХД 0.3 имеет полиномиальную вычислительную сложность* $\max\{O(|E|\log|E|), O(|E|(\chi' + |D|))\}$. *Где E - множе-*

ство ребер, D - множество вершин (устройств хранения), Δ - максимальная степень вершин.

В пятой главе приводится описание реализации разработанных алгоритмов миграции данных в ВО-СХД. **Разработана структура программы** и интерфейсы взаимодействия различных модулей. Подробно описаны модули программы, реализующие алгоритм разделения модели миграции ВО-СХД и алгоритмы раскраски ребер мультиграфа. **Программа реализована** на языке программирования Python.

При помощи реализованного программного средства экспериментально проверяется доказанное выше свойство оптимальности времени масштабирования ВО-СХД, а также определяется выигрыш во времени, который будет получен при использовании предложенных алгоритмов. Используемая методика экспериментальной оценки эффективности алгоритма призвана ответить на следующие вопросы: на сколько уменьшается время (количество шагов) масштабирования; на сколько увеличивается время полной процедуры миграции. Оценки времени производятся в процентах, относительно общего времени миграции с использованием традиционного алгоритма миграции данных. Используются следующие оценки (рис. 7):

- $P_{rel} = \dfrac{\chi'(G) - \chi'(G_{scal})}{\chi'(G)}$ - это относительный выигрыш от масштабирования, соответствующий сэкономленному времени по сравнению с традиционным алгоритмом;
- $L_{rel} = \dfrac{\chi'(G_{scal}) + \chi'(G_{res})}{\chi'(G)} - 1$ - это относительный проигрыш времени от разделения графа, соответствующий дополнительно затраченному времени на миграцию от общего времени миграции данных.

Эксперимент проводился на нескольких случайных графах различного размера (табл. 1). $|D|$ - количество вершин графа, $|E|$ - количество ребер,

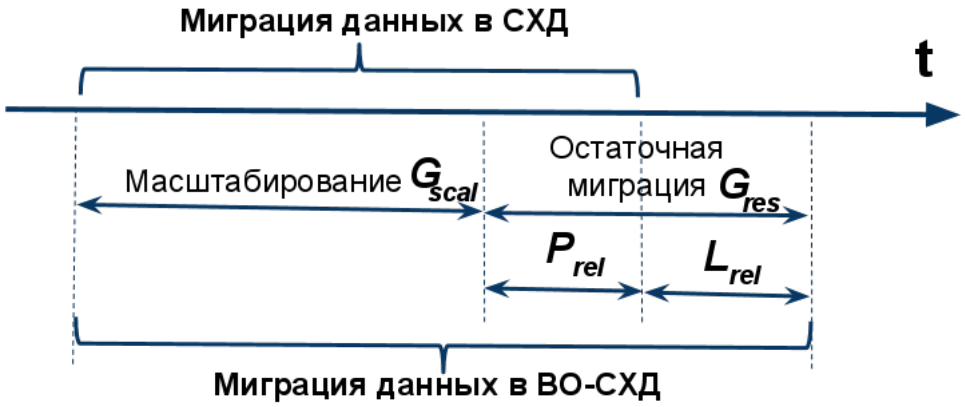

Рисунок 7. Оценка эффективности алгоритма миграции данных ВО-СХД

Таблица 1. Результаты экспериментов

№	$\|E\|$	$\|D\|$	$\|D_{scal}\|$	P_{rel}	L_{rel}	T
1	14	12	3	40%	20%	<0:01
2	30	20	5	50%	33%	<0:01
3	37	22	6	57%	29%	<0:01
4	39	25	6	33%	50%	<0:01
5	46	28	7	43%	43%	<0:01
6	103	37	16	50%	50%	<0:01
7	719	12	6	23%	27%	0:13
8	1067	12	3	21%	8%	0:48
9	1176	25	6	16%	21%	4:24
10	2784	28	7	18%	34%	21:02

$|D_{scal}|$ - количество масштабируемых вершин, T - время работы алгоритма.

Экспериментально показано, что предложенные алгоритмы способены обеспечивать значительную экономию времени масштабирования P_{rel} на 16-57%, по сравнению с существующими алгоритмами. При этом общее время миграции данных L_{rel} может увеличиться на 8-50%. Также в главе описаны возможные способы сокращения общего времени миграции.

В Заключении представлены основные результаты и выводы по диссертационной работе.

Основные результаты работы

В диссертационной работе, согласно поставленным целям, были разработаны: математическая модель, алгоритмы и программное обеспечение, предназначенные для организации эффективного масштабирования ВО-СХД, имеющие важное научное и практическое значение. Результаты:

1. решена техническая задача уменьшения времени масштабирования ВО-СХД во время выполнения процедуры переконфигурации;

2. разработана модель миграции данных ВО-СХД, которая учитывает возможность изменения состава устройств хранения ВО-СХД;

3. на основе предложенной модели, сформулирована задача миграции данных в ВО-СХД в виде многокритериальной оптимизационной задачи, основной критерий - «время масштабирования»;

4. разработаны переборный и аппроксимационный алгоритмы миграции данных для ОС ВО-СХД, способные производить масштабирование ВО-СХД за минимальное время;

5. математически доказана оптимальность предложенных алгоритмов по критерию минимума времени масштабирования ВО-СХД, также доказана полиномиальность аппроксимационного алгоритма;

6. разработано программное средство, реализующее предложенные алгоритмы;

7. экспериментально показано, что предложенные алгоритмы способны обеспечивать экономию времени масштабирования на 16-57% (по сравнению с существующими алгоритмами), при этом общее время миграции данных может увеличиться на 8-50%.

Публикации в изданиях, рекомендованных ВАК России:

1. Петров Д. Л. Оптимальный алгоритм миграции данных в масштабируемых облачных хранилищах // Управление большими системами: сборник трудов (электронный журнал). М. Институт проблем управления РАН. 2010. № 30. С. 180–197.

2. Петров Д. Л. Динамическая модель консолидированного облачного хранилища данных // Известия СПбГЭТУ «ЛЭТИ». СПб. 2010. № 4. С. 17–21.

Другие статьи и материалы конференций:

3. Petrov D. L., Tatarinov Y. Data migration in the scalable storage cloud (Миграция данных в масштабируемых облачных хранилищах) // IEEE International Conference on Ultra Modern Telecommunications. 2009. Pp. 1–4.

4. Петров Д. Л., Красюк В. Консолидация распределенных хранилищ данных: модели и алгоритмы // Материалы IX международной конференции-семинара. Высокопроизводительные параллельные вычисления на кластерных системах. 2009. С. 316–318.

5. Захаров А. С., Петров Д. Л. Реализация и экспериментальное исследование алгоритмов раскраски ребер мультиграфов // XVI международная открытая научная конференция. Современные проблемы информатизации в анализе и синтезе технологических и программно телекоммуникационных систем. 2011. С. 320–324.