РАЗРАБОТКА ТЕХНОЛОГИИ ХРАНЕНИЯ ДАННЫХ ДЛЯ ПАРАЛЛЕЛЬНЫХ СХД

К.В. Бородулин

Южно-Уральский госуниверситет, Челябинск

Описывается создание технологии параллельной системы хранения данных для высокопроизводительных вычислительных комплексов с пиковой производительностью до 0.5 Пфлопс.

Введение

На текущий момент большинство суперкомпьютеров имеют кластерную архитектуру, состоящую более чем из 1000 узлов [1]. Все узлы связаны высокопроизводительной сетью, позволяющей передавать до 40 Гбит/с, что обеспечивает высокую производительность на вычислительных задачах типа LINKPACK. Но при расчете инженерных задач производительность вычислений может сильно снижаться из-за того, что инженерные вычисления производят большое число обменов (сохранение состояния итераций, контрольных точек, чтение исходных данных) с системой хранения данных (СХД) [2].

Идея проекта состоит в увеличении скорости обмена данными между узлами суперкомпьютера и системой хранения данных путем организации параллельной системы хранения. Таким образом, цель – создание технологии параллельной системы хранения данных для высокопроизводительных вычислительных комплексов с пиковой производительностью до 0.5 Пфлопс. Для достижения данной цели необходимо решить следующие задачи:

- разработать алгоритмы обработки и хранения данных в параллельной СХД;
- разработать методы хранения и обработки метаданных в параллельной СУБД;
- реализовать прототип программной системы, использующий разработанные алгоритмы.

1. Обзор существующих технологий

Современные параллельные системы хранения данных (например, Panasas [2]) используют выделенный сервер для хранения метаданных и атрибутов файлов, и при повреждении данного сервера вся информация о файлах (например, на каких узлах сохранялся данный файл) будет утеряна. Также такие системы не позволяют установку сторонних компонентов, что значительно повышает стоимость ремонта и обслуживания системы.

Недостатки распределенных ФС, например lustre [4], заключаются в том, что при использовании высокопроизводительной сети скорость передачи данных с СХД ограничивается контроллером. К примеру, СХД суперкомпьютера «СКИФ-Аврора» использует контроллер, подключенный к сети суперкомпьютера по сети Infiniband и к полке хранения по интерфейсу SAS 6 Гбит/с. Таким образом, производительность сети Infiniband (40 Гбит/с) не используется полностью, т.к. контроллер не может записать такое количество данных на полку хранения за один момент времени и вынужден использовать очередь для данных.

Технология параллельной системы хранения данных позволяет увеличить скорость обмена данных между суперкомпьютером и СХД за счет использования большого числа параллельных операций чтения/записи.

2. Описание технологии

Параллельная система хранения данных состоит из следующих компонентов (рис. 1):

- FS – клиент, установленный на узле суперкомпьютера, служащий для соединения с СХД;
- Director – узел СХД, служащий для связи клиентов с СХД;
- Storage – узлы хранения, на которых хранятся блоки данных;
- СУБД – параллельная СУБД (MongoDB [6]), в которой хранятся метаданные и атрибуты файлов.

Принцип работы состоит в следующем: модуль *FS* соединяется с узлом *Director* и передает команды для чтения/записи и работы с атрибутами файлов. При записи блока *FS* посылает команду WriteBlock() на узел *Director*. Он выбирает незанятый узел *Storage* и передает *FS* номер выбранного узла, на который передается блок данных. Узел *Storage*, на который поступили данные, разбивает блок на сегменты и по высокопроизводительной сети передает их на узлы *Storage*, выбираемые по определенным критериям. После получения сегментов узлы *Storage* записывают метаинформацию и хеш сегмента о них в *СУБД*.

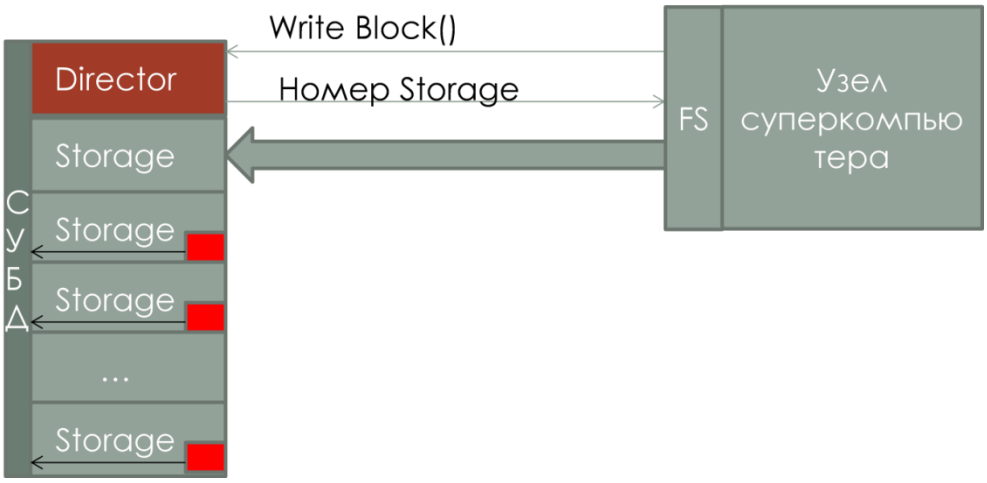

Рис. 1. Архитектура параллельной системы хранения данных

При операции чтения данных FS посылает команду ReadBlock() узлу Director. Он запрашивает информацию о сегментах блока в СУБД и посылает команды узлам Storage отправить данные сегменты в FS. Далее Director посылает контрольную сумму сегментов на FS для проверки целостности переданных данных.

Предложенная технология позволяет устранить точку отказа сервера метаданных за счет использования параллельной СУБД, что повышает надежность системы (каждый из узлов системы может автоматически выступить в роли Director, если основной узел выйдет из строя). Также повышается скорость работы с атрибутами, что важно при работе большого количества пользователей. Использование хешей позволяет обеспечить блочную дедупликацию данных за счет удаления сегментов, хеши которых совпадают. Также технология позволяет создавать моментальные снимки (снапшоты) за счет разрешения метаинформации о сегментах.

Литература

1. TOP500 List – November 2011 (1-100) – [http://top500.org/list/2011/11/100].
2. Panasas parallel storage: hpc benefit for computer aided engineering applications – [http://www.panasas.com/sites/default/files/uploads/docs/panasas_parallelstoragecae_sb_lr_1016.pdf].
3. Welch B., Unangst M. Clustered and Parallel Storage System Technologies // 7th USENIX Conference on File and Storage Technologies (FAST '09).
4. Сайт ФС lustre – [http://wiki.lustre.org/index.php/Main_Page].
5. Сайт проекта ОМЕГА – [http://omega.susu.ru/].
6. Сайт СУБД MongoDB – [http://www.mongodb.org/].
7. Российский рынок систем хранения данных за последний год увеличился наполовину – [http://expert.ru/expert/2012/20/kamera-hraneniya/].