

МОДЕЛЬ ПРОИЗВОДИТЕЛЬНОСТИ РЕПЛИЦИРОВАННОЙ БАЗЫ ДАННЫХ 2_{гс}

Афонов И.В., Ладыженский Ю.В.
Донецкий Национальный Технический Университет

В данном докладе рассматривается характеристика модели производительности распределённой и реплицированной базы данных (БД) 2_{гс}. Данная модель позволяет получить характеристики производительности системы – среднее время отклика и пропускную способность системы. Модель позволяет анализировать производительность системы в зависимости, от многих факторов включая качество выбора данных для репликации и выбора узлов для размещения копий.

Репликация данных – механизм синхронизации содержимого нескольких копий базы данных. Репликация может быть синхронной или асинхронной. В большинстве реальных систем применяется асинхронная репликация, когда данные после изменения на одном из узлов, распространяются на другие узлы в течение некоторого времени, а не сразу после изменения. Репликация также может быть полной либо частичной. При полной репликации – все данные размещены во всех базах данных. При частичной – часть данных размещена в части баз данных, а к остальным, не реплицированным данным, доступ производится удалённо.

Математическая модель 2_{гс} представляет собой комплекс моделей основных компонентов реплицированной базы данных. Критерием производительности данной модели является среднее время отклика и пропускная способность системы.

В таблице 1 представлены основные компоненты распределённой системы хранения данных (СХД) и методы их моделирования.

Таблица 1 – Компоненты модели 2_{гс} и способы моделирования

Компонент	Метод моделирования
Узлы	Сеть из m M/G/1
Сеть	M/G/1
Репликация	Частичная, 2 измерения
Качество репликации	Выбор информации для репликации и узлов для хранения копий

Таблица 1(Продолжение) – Компоненты модели 2гс и способы моделирования

Доступ к данным	Locality (<i>b</i> % запросов могут быть выполнены локально)
Обработка транзакций	Более 2 типов транзакций
Разрешение конфликтов	Игнорируется

В модели используется стратегия управления обновлениям «primary copy». Данная стратегия подразумевает выполнения обновлений только на одной копии данных, которая считается главной, в то время как запросы на выборку производятся с наиболее удобных узлов. Обновления распространяются на ведомые копии асинхронно. К преимуществам данной стратегии можно отнести низкую вероятность возникновения конфликтов и относительно небольшую нагрузку на сеть при частых обновлениях. Недостатком данной стратегии является то, что клиент должен быть всегда подключен к главному серверу.

Отличительной характеристикой данной модели является введение в модель показателей качества выбора данных для реплицирования и выбора узлов для размещения копий. Выбор данных для репликации и решение где разместить копии оказывает огромное влияние на производительность системы в целом.

Показатель качества выбора данных для репликации оказывает большое влияние на выполнение запросов на обновление данных, в то время как показатель качества размещения копий оказывает большое влияние на обработку запросов на выборку.

Выбор данных для репликации, которые должны часто обновляться вызывает повышенную загрузку сети во время распространения изменений на ведомые копии базы данных. Пример графика зависимости пропускной способности системы от показателя качества выбора данных для репликации показан на рисунке 1.

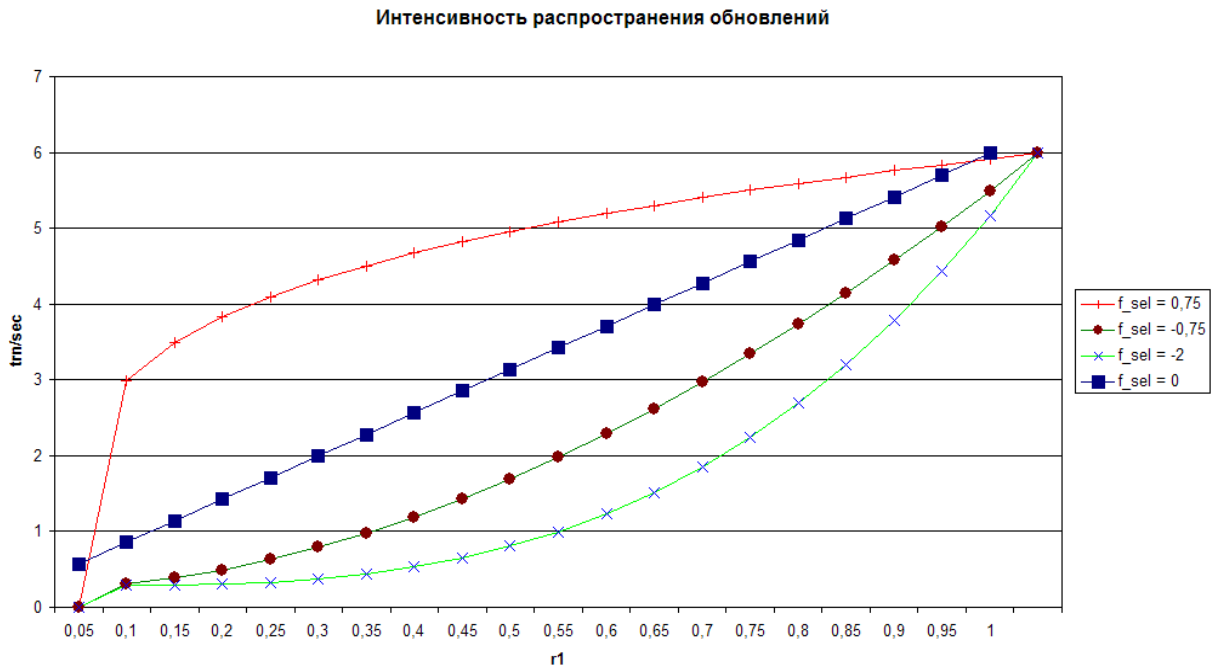


Рисунок 1 – График зависимости пропускной способности системы от показателя качества выбора репликации.

Неудачный выбор данных для репликации приводит к нелинейному и быстрому возрастанию нагрузки с увеличением доли реплицируемых данных. Удачный выбор данных не приводит к значительному увеличению нагрузки на систему, и функция возрастает медленнее линейной.

Кроме того, что выбор данных для репликации оказывает влияние на производительность обработки запросов на выборку, гораздо большее влияние на скорость их обработки оказывает размещение копий данных. Даже если все данные, к которым наиболее вероятен доступ, реплицированы, производительность всей системы не будет значительно увеличиваться, пока эти данные не будут доступны для запросов на локальных узлах.

Литература

- [1] Jim Gray, Pat Helland, Patrick O’Neil, Dennis Shasha. *The Dangers of Replication and a Solution*. Microsoft, 2002.
- [2] Matthias Nicola, Matthias Jarke. *Performance Modeling of Distributed and Replicated Databases*. Technical University of Aachen, 2000

Дата поступления в редакцию 16.05.2007.