УДК 004.738

ИССЛЕДОВАНИЕ МЕТОДОВ БАЛАНСИРОВКИ НАГРУЗКИ В ГЛОБАЛЬНЫХ СЕТЯХ

Петренко А.С., магистрант, Червинский В.В., к.т.н., доц.
(Донецкий национальный технический университет, г. Донецк, Украина)

Экстенсивный рост трафика в Интернет приводит к увеличению количества запросов к популярным сайтам. В связи с этим пользователи могут ощущать длительные задержки при доступе к информации. Изменение инфраструктуры сайта на локальный кластер не обеспечивает полного решения проблемы так как канал между кластером и глобальной сетью может стать узким местом данной инфраструктуры. Более эффективным решением является распределение серверов географически так, чтобы они располагались в отдельных сетях. Роль балансировки нагрузки в таких сетях возрастает, потому что распределяющий объект может осуществлять ее на основе как загрузки сети и серверов так и на основе расстояния между клиентом и серверами.

В данной статье осуществляется классификация подходов маршрутизации клиентских запросов на сервера сайта. Классификация архитектур распределенных Web-серверов в данной статье осуществляется преимущественно на основе объекта, который осуществляет распределение входящих запросов.

Существует 4 класса методов балансировки нагрузки [1]:

- клиентские;
- на базе DNS;
- диспетчерские;
- серверные.

В клиентских методах выбор сервера реализуется на стороне клиента либо случайным образом, либо с помощью механизмов интеллектуального выбора. В первом случае балансировка нагрузки и доступность серверов не может быть гарантирована, во втором же присутствует большая задержка, вызванная мониторингом состояния серверов клиентской программой. В связи с этим данный класс методов на практике практически не применяется.

В методах на базе DNS распределение запросов осуществляется авторизированным DNS-сервером домена. Прозрачность архитектуры для пользователя реализуется на прикладном уровне – сайт имеет один URL.

Однако, контроль распределения запросов со стороны DNS ограничен наличием между клиентом и DNS-сервером, осуществляющим балансировку, большого числа промежуточных DNS-серверов, которые могут кэшировать DNS-записи с целью уменьшения объема трафика. DNS-сервер, осуществляя разрешение имен, определяет период TTL, в течении которого кэш данной записи может храниться. В течении данного периода запросы, проходящие через сервера, хранящие кэш, не будут достигать авторизированного DNS-сервера кластера и балансировка нагрузки для них проводиться не будет, они будут направляться на сервер определённый в кэше.

Контроль со стороны управляющего балансировкой DNS-сервера слабый, поскольку если он задает TTL близким к нулю ряд промежуточных серверов проигнорирует это значение.

В рамках методов на базе DNS существуют алгоритмы с постоянным значением TTL, которые, в связи с описанной проблемой, осуществляют контроль над очень малой долей запросов. Наряду с ними существуют алгоритмы с динамическим TTL [2], осуществляющие кроме балансировки выбор значения TTL для каждой записи (так меньшее значение TTL может быть присвоено запросам от доменов с большим количеством пользователей и запросам направляемым на менее производительный сервер). Алгоритмы первого типа в глобальных сетях не применяются, в отличие от динамических, которые легко масштабируются так как требуют наличия только информации которая может быть динамически получена DNS-сервером (частота запросов соответствующая каждому домену и производительность каждого сервера).

Однако данные методы при выборе сервера не берут во внимание расстояние между клиентом и сервером. Политики динамического назначения TTL в совокупности с механизмом вычисления расстояния до клиента способны обеспечить лучшую производительность.

Диспетчерские методы реализуют прозрачность для пользователя на сетевом уровне – каждому кластеру ставится в соответствие виртуальный IP-адрес, который является адресом специализированного устройства осуществляющего маршрутизацию запросов – диспетчера.

Основываясь на механизме маршрутизации запросов в данной группе методов выделяют два подхода:

- перезапись адресов в пакете;
- перенаправление средствами HTTP – диспетчер посылает клиенту ответ, в котором указан выбранный сервер.

Первый подход подразумевает большой уровень накладных расходов на изменение адресов, в то время как второй увеличивает число открытых TCP-сессий на стороне клиента.

Серверные методы используют двухуровневый механизм балансировки запросов: изначально они распределяются по множеству серверов с помощью DNS-сервера аналогично методам на базе DNS; затем каждый сервер способен перенаправить полученный запрос любому другому серверу. Данный подход позволяет преодолеть большинство ограничений наложенных на подходы на основе DNS. Перенаправление серверами реализуется теми же механизмами маршрутизации что и при диспетчерском подходе.

В статье рассмотрена классификация методов балансировки нагрузки, перечислены недостатки каждого метода. Наилучшей масштабируемостью обладают методы на базе DNS и серверные. Для применения в глобальных сетях объект, осуществляющий балансировку, должен учитывать не только загрузку серверов, но также загрузку каналов связи и расстояние между клиентом и сервером. Таким образом должна применяться комплексная метрика при выборе конкретного сервера.

Перечень ссылок

1. Таненбаум Э., Ван Стеен М. – Распределённые системы. Принципы и парадигмы. – СПб.: Питер, 2003.

2. M. Colajanni, P.S. Yu, V. Cardellini, “Dynamic load balancing in geographically distributed heterogeneous Web-servers”, Amsterdam, The Netherlands, May 1998.