УДК 004.724.4(045)

Жуков И.А., д-р техн. наук
Кулаков Ю.А., д-р техн. наук
Шпак И.Ю.

АДАПТИВНАЯ МНОГОПУТЕВАЯ МАРШРУТИЗАЦИЯ

Институт компьютерных технологий
Национального авиационного университета

Предложен алгоритм адаптивной многопутевой маршрутизации (АМР) для динамического конструирования трафика в пределах автономной системы. Он отличается от известного алгоритма оптимизированной многопутевой маршрутизации отсутствием необходимости хранить и обновлять информацию о всей сети в каждом ее узле. Алгоритм АМР основан на механизме сообщений об обратном давлении, который позволяет осуществлять распределение нагрузки на локальном уровне, а также прогнозировать размеры служебного трафика

Введение

В течении последних лет многопутевая маршрутизация стала темой многих исследований, особенно в части динамического конструирования трафика.

В сети современного Интернет сервис-провайдера, стоимость каналов связи обычно определена фиксировано (рис. 1). Если маршрутизация осуществляется по принципу минимизации стоимости каналов связи, то трафик всегда передается от источника к получателю по одному и тому же пути. Такой подход оправдан в случае отсутствия перегрузки каналов связи. Если же перегрузки возникают, то информация все равно будет передаваться по загруженному каналу, даже если параллельно существует свободный канал. Среди способов решения данной проблемы, предложенных в последнее время, предпочтительнее является пересылка напрямую (*straightforward*) использующая стандартную *IP*-маршрутизацию [1] и делающая попытки глобально оптимизировать стоимость канала для заданной таблице маршрутизации. Основным преимуществом такого метода является отсутствие необходимости изменять стек протоколов. Однако, приблизительный расчет таблицы маршрутизации для действующей *IP*-сети является нетривиальной задачей как показано в [2]. Более того, учитывая типичные изменения таблицы маршрутизации в течении дня [3], необходимо регулярно повторять процедуру оптимизации, что в свою очередь усложняет обслуживание сети.

В отличии от традиционной архитектуры *IP*-сети, в которой путь между парой маршрутизаторов определяется всецело в соответствии с установками стоимости каналов связи, технология многопротокольного переключения по меткам (*MPLS*) [4] позволяет явно выбрать и установить переключаемые по меткам пути (*LSP*) между парой маршрутизаторов в автономной системе. Количество предложений сосредоточено на выполнении конструирования трафика интеллектуальной системой управления *LSP*. Однако, основанное на *MPLS* конструирование трафика наследует основные проблемы усложнения управления.

Более простым подходом к конструированию трафика, применимым к архитектуре обычной *IP*-сети, является адаптация маршрутных метрик к условиям текущей загрузки в сети, как это было в ранней сети *ARPAnet* [5]. В этом случае стоимость каналов связи динамически определялась в зависимости от задержки пакетов, которые использовались как индикаторы перегрузки. Но, в связи со скачкообразными изменениями трафика (так как многопутевая маршрутизация подвержена влиянию даже при изменении стоимости всего одного канала связи) та-

кая схема приводит к нестабильности сети при высокой степени нагрузки [6].

К протоколам, ориентированным на оптимальное распределение нагрузки, относится оптимизированный многопутевой протокол (*OMP*) [6] – это дополнение для конструирования трафика, для известных протоколов маршрутизации по состоянию каналов. Такие протоколы ориентированы на оптимальное распределение нагрузки, основанное на том, что каждый маршрутизатор X имеет глобальную информацию о загрузке всех каналов связи в сети. Располагая такой информацией, маршрутизатор X может переключать трафик с загруженного канала на более свободный, осуществляя распределение нагрузки. Для поддержания каждого узла в актуальном состоянии, он распространяет и регулярно обновляет информацию о загрузке подсоединенных к нему каналов связи, используя встроенный в протокол механизм распространения (*flooding*) информации о состоянии каналов связи. Этот механизм включается либо по истечению времени, прошедшего с момента последнего обновления, либо по количеству изменений загрузки в последнем измерении.

Тем не менее, более детальное исследование *OMP* раскрывает несколько важных недостатков этого протокола, таких как сложность (и поэтому потребление ресурсов) структур хранения данных необходимых для содержания всего множества путей между всеми возможными парами источников и получателей, а также сопутствующая непредсказуемая перегрузка сигналами, необходимыми для распространения обновлений состояния каналов.

Адаптивная многопутевая маршрутизация (AMP)

Адаптивная многопутевая маршрутизация (*AMP*) разрабатывалась для того чтобы избежать перечисленных недостатков. В контексте *OMP*, основной идеей алгоритма *AMP* является перевод масштаба взаимодействия каждого узла с глобального уровня на локальный. Таким образом, любой произвольный узел X, не знает о состоянии всех возможных каналов между ним и всеми узлами сети. Узлу X известно только состояние каналов, соединяющих его с ближайшими соседями. Распространение информации о загрузке каналов в сети сводится к так называемому механизму обратного давления. Для того чтобы понять эту концепцию, можно представить сеть Интернет как систему соединяющихся однонаправленных, пористых резиновых труб, по которым течет вязкая жидкость. Как только жидкость наталкивается на препятствие (ситуация, при которой канал загружен) тут же локально повышается давление и далее возможно два варианта: давление направляется в сторону обратную направлению изначального движения (обратное давление), которое приведет к перераспределению давления между ближайшими трубами (жидкость вязкая) и достижению нового равновесия; возможно также что при продолжающемся воздействии давления часть жидкости начнет просачиваться через поры труб, что равносильно потере пакетов в сети, которого мы постараемся избежать применяя первый вариант развития событий, т.е. распределение нагрузки.

Для того чтобы лучше объяснить взаимодействие между локальным и глобальным распространением информации о загрузке каналов, представим произвольный узел сети A и Ω_A – произвольный набор соседей узла A. Также определим AB как произвольный направленный канал от узла A к узлу B. Основной механизм данного алгоритма представлен на Рис. 1. В отличии от *OMP*, в котором увеличение использования канала Y_0X приводит к тому, что все узлы сети разгружают некоторые свои пути, содержащие отрезок Y_0X, в *AMP* только один узел реагирует – Y_0. Так как Y_0 является передающим узлом, он старается переключить трафик на альтернативные пути передачи. Также он периодически отправляет сообщение об обратном давлении своим ближайшим соседям $N_j \in \Omega_{Y0}$, информируя их

о своем вмешательстве в связи с перегрузкой канала Y_0X. Все узлы $N_j \in \Omega_{Y0}$ поочереди передают эту информацию своим ближайшим соседям, которые, в свою

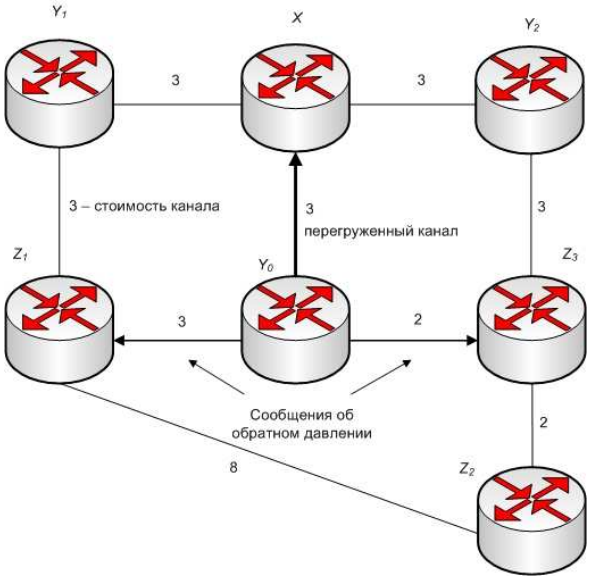

Рис. 1. Действие механизма обратного давления

очередь, передают такие сообщения дальше в соответствии с ситуациями перегрузки, с которыми они сталкиваются. Такой квази-рекурсивный механизм обеспечивает распространение информации о загрузке каналов связи по всей сети, создавая точно прогнозируемый служебный трафик, потому что сообщения об обратном давлении посылаются строго периодически.

Подсчет множества путей, распределение нагрузки, метрики загрузки канала

Многопутевая маршрутизация отличается от других протоколов внутренней маршрутизации (таких как *OSPF* [7]) тем, что задействует не единственный наилучший (минимальная стоимость) путь к получателю, а использует более одного пути для передачи. Стоит отметить, что так называемые многопутевые равноценные подходы (*ECPM*) используют пересылку напрямую, позволяя использовать несколько путей с одинаковой минимальной стоимостью передачи и разбиваю трафик равномерно по ним.

Для того чтобы увеличить количество возможных путей для передачи, но при этом не усложнить процесс до *MPLS*-подобного или маршрутизации от источника, [6] предлагает использовать «расслабленный» критерий выбора лучшего пути. Основная идея заключается в том, что любой соседний узел, который ближе (меньше по стоимости) к получателю чем текущий узел, является допустимым узлом для следующего перехода. Такое условие сразу делает невозможным образование маршрутных зацикливаний, так как стоимость узла получателя принудительно уменьшается каждым узлом на пути.

Возвращаясь к рис. 1, оценим рассмотренный выше критерий на примере сети из 7 узлов. Для узла X минимальная стоимость пути к узлу Z_2 достигается по пути $XY_0Z_3Z_2$ и равна 7. Так как стоимость к получателю Z_2 для Y_2 – соседу X равна $5<7$, путь $XY_2Z_3Z_2$ также становится доступным.

Использование критерия «расслабленного» выбора пути позволяет задействовать множество путей между отправителем и получателем, но в тоже время нуждается в распределения нагрузки между ними. Как показано в [8], распределение нагрузки в *AMP* динамично, но он скорее консервативно переносит нагрузку, в то время как в *OMP* распределение нагрузки осуществляется не в результате сложного процесса принятия решения, а по периодическим интервалам времени.

Если в *AMP* узел A имеет множество путей к узлу B, то он должен обладать механизмом распределения трафика для узла B по этим путям не допуская нарушения порядка следования пакетов. Для обеспечения этого требования *AMP* применяет хэш-функцию к адресам отправителя и получателя и разделяет хэш-пространство среди всех доступных путей с помощью отсечек. По аналогии с [6], такой механизм обеспечивает неравномерное разбиение трафика по путям и динамическое распределение нагрузки достигается нужными изменениями отсечек. Как показано в [9], *CRC*-16 (16-ти битная

циклическая избыточная проверка) хэш-функция обладает очень хорошей производительностью распределения нагрузки. Это достигается благодаря равномерному распределению адресных пар отправитель/получатель в пространстве решений при использовании реального трафика.

Как подходящая метрика для нагрузки канала в эластичном трафике, таком как *TCP*, наш алгоритм вводит эквивалентную нагрузку *EL*,как предложено в [6]:

$$EL = \max\{\rho, \rho \times K \times \sqrt{P}\}, \qquad (1)$$

где P – это вероятность потери пакетов и K – масштабный коэффициент, определяющий границы потери пакетов, при которых значение *EL* превышает загрузку ρ канала, определяемую как:

$$\rho = \frac{CarriedTrafficVolume}{LinkCapacity \times TimePeriod}. \qquad (2)$$

Уравнение (1) отображает динамическое поведение эластичного трафика, проходящего по каналу, так как *TCP*-поток замедляется практически обратно пропорционально корню квадратного вероятности потери пакетов в сужающемся канале [10]. В наших расчетах мы принимаем значение K равное 10, что соответствует превышению значения *EL* нагрузки канала ρ для вероятности потери пакетов более 1%.

Рекурсивное обратное давление как сигнальный механизм AMP

Как уже упоминалось, основной новизной *AMP* является сигнальный механизм. На Рис. 2 изображена типичная ситуация, в которой узел Y_0 должен принять решение по нагрузке для своих передающих каналов. Далее, как пример возьмем канал Y_0X для описания информации, необходимой для Y_0 для каждого из своих исходящих каналов. Кроме эквивалентной нагрузки на канал, которую можно измерить напрямую, Y_0 нуждается в информации о расстоянии на которое маршрутизируется трафик от Y_0 через X с учетом загрузки каналов XY_i, а также далее по каналам после узлов Y_i, где $i \geq 1$. Такая информация содержится в сообщениях об обратном давлении.

Для более детального объяснения механизма, рассмотрим узел X, который получает трафик от узла Y_0 по каналу Y_0X и пересылает это трафик узлам *Yi* по каналам XY_i где $i \geq 1$. Будем считать, что узел X (как и любой другой узел) имеет возможность измерять только эквивалентную нагрузку *EL* на своих исходящих каналах, т.е. XY_i, где $i \geq 1$, согласно (1). Пока Y_0 в некоторой степени отвечает за нагрузку канала XY_i, где $i \geq 1$, он не может напрямую определить наличие перегрузки на этом канале. Поэтому узел X, у которого есть прямая связь с каналом XY_i, может информировать узел Y_0 о ситуации с загрузкой канала, так как это принципиально важно для принятия решения о распределении нагрузки узлом Y_0. Это делается периодически с помощью отправки сообщений об обратном давлении. Это не обозначает что X информирует Y_0 только о (напрямую измеренной) нагрузке на канал XY_i, где $i \geq 1$, так как ситуация далее за узлами Y_i, где $i \geq 1$ также нужна. С другой стороны узел X не имеет прямого доступа к более дальним узлам, но получает сообщения об обратном давлении от узлов Y_i, которые информируют его о ситуации с загрузкой каналов за узлами Y_i, где $i \geq 1$, т.е. как раз о том, о чем дополнительно необходимо знать узлу Y_0. Итак, любое сообщение об обратном давлении, отправляемое из узла X к узлу Y_0 должно содержать явную информацию о полной загрузке канала XY_i, а также косвенную информацию о ситуации на дальнейших каналах, которую передает Y_i узлу X, $i \geq 1$. Далее приведем более формальное описание содержания сообщения об обратном давлении, как функцию f от $2n$ параметров, где n – количество исходящих каналов узла X. Пусть сообщение об обратном давлении – $BM(A, B)$, где A, B – значение параметров информации, передаваемой от узла A, узлу B. Тогда получим:

$$BM(X, Y_0) = f(EL_{XY_1}..., EL_{XY_n}, BM(Y_1, X),..., BM(Y_n, X)). \qquad (3)$$

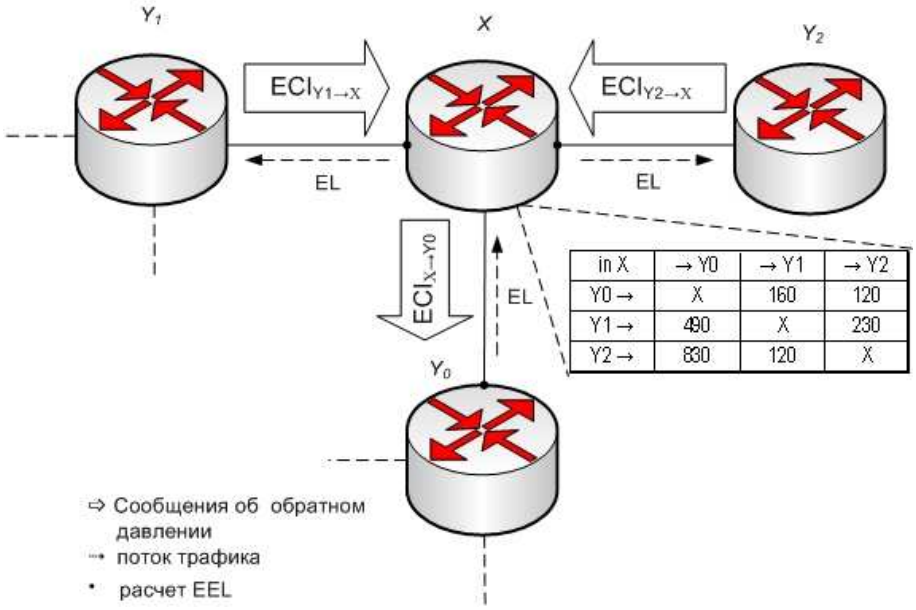

Рис. 2. Генерация сообщения о полной загрузке

Для того, чтобы сохранять *BM* маленьким, *f* должно отображать параметр *2n* в скалярном описании ситуации загрузки Y_0. Для определения *f* мы, во-первых, уменьшим количество входных параметров до одного на каждый исходящий канал, просуммировав ситуацию на каждом канале XY_i с помощью функции *g*:

$$g_i = g(EL_{XY_i}, BM(Yi, X)) \forall i = 1,..., n. \quad (4)$$

Так как ни исходящий канал, ни сеть далее не должны быть перегружены, g должна быть функцией максимума: при условии что $BM(Y_i,X)$ эквивалентно скалярному $ECI_{Yi \to X}$ (так называемый Полный индикатор загрузки), мы получаем определение эффективной эквивалентной нагрузки (*EEL*):

$$EEL_{XY_i} = \max\{EL_{XY_i}, ECI_{Y \to X}\}, \quad i = 1,2,..,n \quad (5)$$

Отметим что, $ECI_{Yi \to X}$ объединяет информацию о загрузке исходящего канала Y_i, такую же, как содержит *BM*-сообщение от Y_i к *X*. Точно также, $ECI_{X \to Y0}$ посылается как *BM*-сообщение от *X* к Y_0 и может быть рассмотрено как результат дальнейшего упрощения функции *f* в выражении (3), где n разных параметров g_i из (4) суммируются функцией *h*:

$$ECI_{X \to Y_0} = h(g_1,..., g_n). \quad (6)$$

Мы используем взвешенную сумму для подсчета *h*, где вес канала XY_i соответствует соотношению между трафиком в канале XY_i, который поступил от Y_0 через *X*, и полным трафиком в канале XY_i. Таким образом, взвешенная сумма предоставляет сжатую версию всей информации, которая доступна узлу *X* про вклад Y_0 в ситуацию по загрузке исходящих каналов, и поэтому называемой полный индикатор загрузки:

$$ECI_{X \to Y_0} = \sum_{Y_i \in \Omega_x / Y_0} \frac{\beta_{XY_i}(Y_0)}{\beta_{XY_i}} \bullet EEL_{XY_i}. \quad (7)$$

Напомним, что Ω_X это множество все соседних узлов узла *X*, XY_i – исходящий канал между *X* и Y_i, β_{XYi} (*Y0*) – это количество байт отправленных из узла Y_0 через *X* к Y_i и β_{XYi} – это общее количество байт переданных от любых узлов $\in \Omega_{X \setminus Yi}$ через *X* к Y_i.

Коротко подводя итог, $ECI_{X \to Y0}$ – это одноразмерный параметр, который описывает степень, в которой узел Y_0 влияет на ситуацию с загрузкой в сети, как это оценивает узел *X*. $ECI_{X \to Y0}$ описывается в соответствии с (3) и параметрами функции *f*, которая после (4) и (6) имеет вид:

$$BM(X,Y_0) = ECI_{X\to Y_0} = f(EL_{XY_1}...,EL_{XY_n}, BM(Y_1,X),..., BM(Y_n,X)) = h(g_1,...,g_n) = h(g(EL_{XY_1}, ECI_{Y_1\to X}),..., g(EL_{XY_n}, ECI_{Y_n\to X})) \quad (8)$$

Формулировка, данная в (8) представляет рекурсивную структуру механизма обратного давления, так как $ECI_{Yi\to X}$ – это аналог информации отправляемой узлом Y_i к и $BM(Y_i, X)$.

Для того, чтобы гарантировать быстрое распространение информации о нагрузке в сети, интервалы между последовательными *BM*-сообщениями в каждом отдельном канале нужно установить малыми, порядка одной секунды. Если мы примем во внимание маленький размер *BM*-сообщения (зависит от протокола, обычно порядка 50 байт) и пропускную способность современных каналов связи (обычно измеряется Мегабитами), то получим следующее: *AMP*-сигнализация обычно занимает менее чем $3*10^{-4}$ % емкости канала.

В заключении отметим, что расчет $\beta_{XY_i}(Y_0)$ в (7) требует от узла *X* точного измерения трафика на входящем канале Y_0X к исходящему каналу XY_i, i=1,2,…,n. Такое измерение содержится в так называемой Входящей/Исходящей Матрице, которая, в свою очередь, хранится в узле *X* (Рисунок 2). Эта матрица хранится для каждой пары узлов (P,Q), $P,Q \in \Omega_X$ $P \neq Q$, количество байт передаваемых между этими узлами через *X*.

Выводы

Представлен алгоритм адаптивной многопутевой маршрутизации как средство эффективного динамического конструирования трафика. В отличие от известных алгоритмов, которые используют глобальную систему оповещения о загрузке каналов, наш алгоритм использует систему локального оповещения, что делает поток служебной информации минимальным и четко прогнозируемым, а также позволяет сократить затраты ресурсов маршрутизаторов, так нет необходимости хранить информацию о всех возможных множествах путей. Результаты моделирования показывают преимущество предложенного алгоритма над существующими.

Список литературы

1. *B. Fortz, M. Thorup*. Internet Traffic Engineering by Optimizing OSPF Weights. Proc. IEEE Infocom, Tel Aviv, Israel, 2000. –P. 519–528.

2. *A. Medina, N. Taft, K.* Salamatian, S. Bhattacharyya, C. Diot: Traffic Matrix Estimation: Existing Techniques and New Directions. ACM SIGCOMM, Pittsburg, PA, 2002.

3. *S. Bhattacharyya, C. Diot, J. Jetcheva, N. Taf.* POP-Level and Access-Link-Level Traffic Dynamics in a Tier-1 POP. ACM SIGCOMM Internet Measurment Workshop, San Francisco, CA, 2001.

4. *E. Rosen, A. Viswanathan, R. Callon.* Multiprotocol Label Switching Architecture. IETF RFC 3031, 2001.

5. *A. Khanna, J. Zinky.* The Revised ARPANET Routing Metric. ACM SIGCOMM Symposium on Communications Architectures and Protocols, Austin, TX, 1989.

6. *C. Villamizar*. OSPF Optimized Multipath (OSPF-OMP). IETF Internet Draft, 1999.

7. *I. Gojmerac, T. Ziegler, P. Reichl.* Adaptive Multi-Path (AMP) – a Novel Routing Algorithm for Dynamic Traffic Engineering. Technical report FTW-TR-2008-007, Vienna, 2008.

8. *J. Moy*. OSPF version 2. IETF RFC 2328, 1998.

9. *Z. Cao, Z. Wang, E. Zegura.* Perfomance of Hashing-Based Schemes for Internet Load Balancing. IEEE Infocom, Tel Aviv, Israel, 2000.

10. *M. Mathis, J. Semke, J. Mahdavi, T. Ott.* The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm. ACM Computer Communications Review, 27(3), 1997.