КАЧЕСТВО ОБСЛУЖИВАНИЯ В СЕТЯХ IP

Г.Г. Яновский, заведующий кафедрой сетей связи
СП6ГУТ им. проф. М.А. Бонч-Бруевича, доктор технических наук

В данной статье дается обзор методов поддержки качества обслуживания в сетях, построенных на базе IP-ориентированных протоколов. Рассматриваются рекомендации Y.1540 и Y.1541, определяющие сетевые характеристики и нормы качества обслуживания в сетях IP. Представлен анализ механизмов поддержки качества обслуживания в пакетных сетях, специфицированных в Рекомендации МСЭ-Т Y.1291. Представлены основные подходы к решению задачи обеспечения гарантированного качества обслуживания в IP-сетях, базирующиеся на моделях интегрированных и дифференцированных услуг.

Качество обслуживания (Quality of Service, QoS) является предметом активных исследований и стандартизации на протяжении всей истории развития телекоммуникаций. Существенный вклад в развитие различных аспектов концепции QoS внес Международный союз электросвязи, включая, в том числе, разработку норм и требований к показателям качества обслуживания, стандартизацию сетевых механизмов, обеспечивающих необходимые показатели QoS, а также формулировку основополагающих определений.

Среди стандартов, посвященных качеству обслуживания в электросвязи, одно из центральных мест занимает Рекомендация МСЭ E.800. В ней качество обслуживания определяется как "суммарный эффект рабочих характеристик обслуживания, который определяет степень удовлетворенности пользователя данной службой". Расширяя концепцию качества обслуживания, отвечающую Рекомендации E.800, Рекомендация МСЭ G.1000 разделяет рабочие характеристики обслуживания на функциональные компоненты и связывает их с сетевыми характеристиками, определенными в ряде рекомендаций МСЭ – таких как I.350, Y.1540 и Y.1541.

В дополнение к Рекомендации МСЭ G.1000, определяющей структуру связей между рабочими характеристиками (производительностью, надежностью, потерями, задержкой и др.) и характеристиками сети, Рекомендация МСЭ G.1010 содержит спецификации требований со стороны приложений, ориентированных на конечного пользователя. Качество обслуживания нашло отражение в большом числе статей и книг, среди которых отметим монографии [1 – 3].

Исторически, первые системы оценок и механизмов поддержки качества обслуживания были разработаны для традиционных видов электросвязи – телеграфии и телефонии. Понятно, что сегодня при широком применении сетей передачи данных, быстром внедрении широкополосных технологий и замене телеграмм на сообщения электронной почты параметры качества обслуживания и механизмы их поддержки в телеграфных сетях становятся все менее актуальными. При построении и эксплуатации ТфОП задача обеспечения гарантированного качества обслуживания состоит в том, чтобы обслуживание телефонного вызова осуществлялось с соблюдением всех установленных норм, в том числе, и заданных показателей качества передачи речи. Совокупность этих норм и соответствующих численных значений базируется на документах МСЭ и ETSI.

Модель услуг в ТфОП была основана на принципе установления соединения и в дальнейшем (70-е – 80-е годы прошлого столетия) была распространена на такие

технологии передачи данных, как X.25, Frame Relay и широкополосные цифровые сети интегрального обслуживания (B-ISDN), основанные на модели виртуальных каналов. В сетях B-ISDN рекомендации МСЭ (в частности, I.356 и I.610) и Форума ATM определяют параметры качества обслуживания и способы их измерений для межконцевых соединений.

В отличие от упомянутых выше технологий в классических сетях IP применяется метод доставки, полностью исключающий любую форму организации соединений – как физических, так и виртуальных. Этот метод основан на рассылке пакетов-дейтаграмм. Качество доставки в традиционных сетях IP базируется на принципе так называемой "наилучшей попытки" (Best effort). Концепция "наилучшей попытки" предполагает, что пользователи справедливо разделяют доступные сетевые ресурсы, трафик передается со скоростью, максимально возможной в данных условиях загрузки ресурсов сети, но при этом не гарантируется обеспечение любого предварительно определенного уровня качества обслуживания. Очевидно, что такой подход к обслуживанию означает следующее: отсутствуют различия между разными видами трафика, нет гарантии в доставке пакетов в правильном порядке, и что он будет доставлен в требуемое время или вообще будет доставлен, и т. д.

Концепция "наилучшей попытки" была достаточно эффективной для приложений, где можно передавать данные не в реальном времени (электронная почта, передача файлов). Кроме того, с учетом переизбытка сетевых ресурсов в транспортных сетях, построенных на базе волоконно-оптических линий связи, принцип "наилучшей попытки" в определенной степени позволяет обеспечить сегодня требования телефонии (голос поверх IP) и других приложений реального времени.

Однако, как только возникает недостаток ресурсов, ведущий к увеличению вероятности потерь пакетов и росту их задержек, для приложений реального времени необходимые показатели качества обслуживания не могут быть обеспечены. Прежде всего, это объясняется основным принципом функционирования IP-сетей – передачей данных в дейтаграммном режиме, т. е. без установления соединений и без управления. С появлением новых приложений, особенно реального времени (интерактивная передача речи, видеотелефония и видеоконференции), вопрос о гарантированном качестве обслуживания в сетях IP становится одним из наиболее сложных. Это объясняет, почему качество обслуживания в сетях IP остается предметом постоянного внимания МСЭ, ETSI, IETF и других организаций стандартизации в электросвязи.

Сегодня общепризнанно, что сети с коммутацией каналов и пакетов постепенно эволюционируют в направлении создания общей инфраструктуры, базирующейся на протоколах семейства IP. Этот процесс получил название конвергенции. Инфраструктура, возникшая в результате конвергенции, должна будет обеспечивать транспортировку трафика телефонных сетей, сетей телевидения и трафика приложений, традиционно использующих сети Интернет. Подобный сценарий конвергенции предлагает как экономический выигрыш, получаемый благодаря объединению технологий, так и определяет развитие сектора телекоммуникаций через создание новых услуг.

Однако, процесс конвергенции до настоящего времени протекает достаточно медленно. И здесь мы вновь возвращаемся к проблеме обеспечения необходимого качества обслуживания, которая является одним из основных тормозящих факторов в процессе конвергенции сетей и услуг и построении единой сети на базе IP, рассматриваемой сегодня как сеть следующего поколения (Next Generation Network, NGN). Чтобы полностью реализовать преимущества конвергенции в будущих IP-ориентированных сетях, необходимо разработать новые принципы распределения ресурсов сетей и

управления трафиком, которые будут гарантировать различные уровни показателей качества обслуживания для большого и разнообразного числа приложений, реализуемых конечными пользователями.

При этом разделение ресурсов и процессы управления трафиком должны быть скоординированы в условиях наличия большого числа разнообразных приложений с существенно отличающимися требованиями к рабочим характеристикам сети (табл. 1). Детальное рассмотрение рабочих характеристик, определяющих качество обслуживания, и соответствующих норм будет проведено в следующих разделах.

Таблица 1.
Чувствительность различных приложений к сетевым характеристикам

Тип трафика	**Уровень чувствительности к сетевым характеристикам**			
	Полоса пропускания	**Потери**	**Задержка**	**Джиттер**
Голос	Очень низкий	Средний	Высокий	Высокий
Электронная коммерция	Низкий	Высокий	Высокий	Низкий
Транзакции	Низкий	Высокий	Высокий	Низкий
Электронная почта	Низкий	Высокий	Низкий	Низкий
Telnet	Низкий	Высокий	Средний	Низкий
Поиск в сети "от случая к случаю"	Низкий	Средний	Средний	Низкий
Постоянный поиск в сети	Средний	Высокий	Высокий	Низкий
Пересылка файлов	Высокий	Средний	Низкий	Низкий
Видеоконференция	Высокий	Средний	Высокий	Высокий
Мультикастинг	Высокий	Высокий	Высокий	Высокий

Работы МСЭ по стандартизации качества обслуживания в сетях IP

В рамках работ МСЭ по стандартизации качества обслуживания в сетях IP предполагаются следующие этапы решения задачи обеспечения QoS для сетей, построенных на базе IP-ориентированных протоколов:

- создание согласованного общего набора рабочих характеристик сетей IP и норм для него;
- внедрение сетевых механизмов, которые будут обеспечивать заданные показатели качества обслуживания в конфигурации "терминал-терминал";
- вложение нормированных значений показателей качества обслуживания в протоколы сигнализации;
- разработка архитектуры сетевых механизмов поддержки.

В 2002 г. ИК 13 МСЭ-Т опубликовала два международных стандарта, которые отвечают первому из перечисленных этапов. Рекомендация МСЭ Y.1540 описывает стандартные сетевые характеристики для передачи пакетов в сетях IP. Рекомендация МСЭ Y.1541 [5] определяет нормы для параметров, определенных в Y.1540, между двумя граничными сетевыми интерфейсам – точками подключения оконечных терминальных устройств. Кроме того, в этой рекомендации специфицированы шесть классов качества обслуживания в зависимости от приложений.

Эти рекомендации важны для всех участников телекоммуникационного сценария – операторов и провайдеров, производителей оборудования и конечных пользователей. Сетевые операторы и провайдеры будут использовать их при планировании,

развертывании и оценке сетей IP в соответствии с требованиями конечных пользователей к качеству обслуживания. Производители будут опираться на эти рекомендации при создании оборудования, которое должно отвечать спецификациям сетевых провайдеров. Наконец, конечные пользователи (в первую очередь, корпоративные) смогут применить рекомендации Y.1540 и Y.1541 при оценке характеристик реально функционирующих IP-сетей с позиций соответствия этих характеристик требованиям потребителей. Рассмотрим некоторые детали рекомендаций Y.1540 и Y.1541, касающиеся основных сетевых характеристик, связанных с обеспечением QoS в сетях IP.

Рекомендация МСЭ Y.1540

В Рекомендации Y.1540 рассматриваются следующие сетевые характеристики, как наиболее важные по степени их влияния на сквозное качество обслуживания (от источника до получателя), оцениваемое пользователем:

- производительность сети;
- надежность сети/сетевых элементов;
- задержка;
- вариация задержки (джиттер);
- потери пакетов.

Производительность сети (или скорость передачи данных) пользователя определяется как эффективная скорость передачи, измеряемая в битах в секунду. Следует отметить, что значение этого параметра не совпадает с максимальной пропускной способностью сети, ошибочно называемой (причем, довольно часто) полосой пропускания. Минимальное значение производительности обычно гарантируется провайдером услуг, который, в свою очередь, должен иметь соответствующие гарантии от сетевого провайдера.

В Рекомендации Y.1540 не приведены нормативные характеристики производительности сети, которые различаются для различных приложений. Вместе с тем, в Рекомендации Y.1541 отмечено, что параметры, связанные с эффективной скоростью передачи, могут быть определены через дескриптор трафика IP-сети, описанный в Рекомендации МСЭ Y.1221.

Надежность сети/сетевых элементов. Пользователи обычно ожидают высокий уровень надежности от систем связи. Надежность сети может быть определена через ряд параметров, из которых наиболее часто используется коэффициент готовности, вычисляемый как отношение времени простоя объекта к суммарному времени наблюдения объекта, включающему время простоя и время между отказами. В идеальном случае коэффициент готовности должен быть равен 1, что означает стопроцентную готовность сети. На практике коэффициент готовности оценивается числом "девяток". Например "три девятки" означают, что коэффициент готовности составляет 0,999, что соответствует 9 часам времени недоступности (простоя) сети в год. Готовность сети ТфОП оценивается величиной "пять девяток", что означает 5,5 мин. простоя в год. В табл. 2 приведены данные по времени простоя для различного количества "девяток".

Таблица 2.
Коэффициенты готовности и соответствующие значения времени простоя оборудования

Коэффициент готовности	Время простоя
0,99	3,7 дней в год

0,999	9 часов в год
0,9999	53 минуты в год
0,99999	5,5 минут в год
0,99999999	30 секунд в год

Необходимо отметить, что обеспечение коэффициента готовности "пять девяток" в сетях IP, построенных на традиционном оборудовании данных (серверы, маршрутизаторы), является достаточно серьезной проблемой. Причина этого состоит в том, что обработка информационных потоков в сетях IP в значительной части базируется на программном обеспечении (а не на аппаратном, как это имеет место в ТфОП). В то же время статистика отказов сетевого оборудования показывает, что надежность программного обеспечения примерно в два раза ниже надежности аппаратного обеспечения.

Параметры доставки пакетов IP. В общем случае сеанс связи состоит из трех фаз – установления соединения, передачи информации и разъединения соединения. В Рекомендации Y.1540 из трех фаз сеанса связи рассматривается только вторая – фаза доставки пакетов IP. Такой подход отражает природу сетей IP, не ориентированных на установление соединений. Спецификацию рабочих характеристик и параметров QoS для двух других фаз (установление и разъединение соединения) планируется провести в дальнейшем.

Рекомендация МСЭ-Т Y.1540 определяет следующие параметры, характеризующие доставку IP-пакетов.

Задержка доставки пакета IP (IP packet transfer delay, IPTD). Параметр IPTD определяется как время $(t_2 - t_1)$ между двумя событиями – вводом пакета во входную точку сети в момент t_1 и выводом пакета из выходной точки сети в момент t_2, где $(t_2 > t_1)$ и $(t_2 - t_1) <= T_{max}$.

В общем, параметр IPTD определяется как время доставки пакета между источником и получателем для всех пакетов – как успешно переданных, так и пораженных ошибками.

Средняя задержка доставки пакета IP – параметр, специфицированный в Рекомендации Y.1540, определяется как средняя арифметическая величина задержек пакетов в выбранном наборе переданных и принятых пакетов. Значение средней задержки зависит от передаваемого в сети трафика и доступных сетевых ресурсов, в частности, от пропускной способности. Рост нагрузки и уменьшение доступных сетевых ресурсов ведут к росту очередей в узлах сети и, как следствие, к увеличению средних задержек доставки пакетов.

Речевая информация и, отчасти, видеоинформация являются примерами трафика, чувствительного к задержкам, тогда как приложения данных в основном менее чувствительны к задержкам. Когда задержка доставки пакета превышает определенные значения T_{max}, такие пакеты отбрасываются. В приложениях реального времени (например, в IP-телефонии) это ведет к ухудшению качества речи. Ограничения, связанные со средней задержкой пакетов IP, играют ключевую роль для успешного внедрения технологии Voice over IP (VoIP), видео-конференций и других приложений реального времени. Этот параметр во многом будет определять готовность пользователей принять подобные приложения.

Вариация задержки пакета IP (IP packet delay variation, IPDV). Параметр V_k, характеризует вариацию задержки IPDV. Для IP-пакета с индексом k этот параметр определяется между входной и выходной точками сети в виде разности между

абсолютной величиной задержки X_k при доставке пакета с индексом k, и определенной эталонной (или опорной) величиной задержки доставки пакета IP, $d_{1,2}$, для тех же сетевых точек:

$$V_k = X_k - d_{1,2}.$$

Эталонная задержка доставки пакета IP, $d_{1,2}$, между источником и получателем определяется как абсолютное значение задержки доставки первого пакета IP между данными сетевыми точками. Вариация задержки пакета IP, или джиттер, проявляется в том, что последовательные пакеты прибывают к получателю в нерегулярные моменты времени. В системах IP-телефонии это, к примеру, ведет к искажениям звука и в результате к тому, что речь становится неразборчивой.

Коэффициент потери пакетов IP (IP packet loss ratio, IPLR). Коэффициент IPLR определяется как отношение суммарного числа потерянных пакетов к общему числу принятых в выбранном наборе переданных и принятых пакетов. Потери пакетов в сетях IP возникают в том случае, когда значение задержек при их передаче превышает нормированное значение, определенное выше как T_{max}. Если пакеты теряются, то при передаче данных возможна их повторная передача по запросу принимающей стороны. В системах VoIP пакеты, пришедшие к получателю с задержкой, превышающей T_{max}, отбрасываются, что ведет к провалам в принимаемой речи. Среди причин, вызывающих потери пакетов, необходимо отметить рост очередей в узлах сети, возникающих при перегрузках.

Коэффициент ошибок пакетов IP (IP packet error ratio, IPER). Коэффициент IPER определяется как суммарное число пакетов, принятых с ошибками, к сумме успешно принятых и пакетов, принятых с ошибками.

Рекомендация МСЭ Y.1541

Рекомендация Y.1540 определяет численные значения параметров, специфицированных в ней, которые должны выполняться в сетях IP на международных трактах, соединяющих терминалы пользователей. Нормы на параметры разделены по различным классам QoS, которые определены в зависимости от приложений и сетевых механизмов, применяемых для обеспечения гарантированного качества обслуживания. В табл. 3 [5] представлены нормы на определенные выше сетевые характеристики.

Значения параметров, приведенные в таблице, представляют собой, соответственно, верхние границы для средних задержек, джиттера, потерь и ошибок пакетов. В Рекомендации Y.1541 представлены спецификации набора параметров, которые связаны с измерением реальных значений сетевых характеристик – периода наблюдений, длины тестовых пакетов, числа пакетов и т. д. В частности, при оценке качества передачи пакетов речи в IP-телефонии минимальный интервал наблюдения должен быть порядка 1 – 20 с при типичной скорости передачи 50 пакетов/с. Рекомендуемый интервал измерений для задержки, джиттера и потерь должен составлять не менее 60 с.

Таблица 3
Нормы для характеристик сетей IP с распределением по классам качества обслуживания

Сетевые характеристики	Классы QoS					
	0	1	2	3	4	5
Задержка доставки пакета IP, IPTD	100 мс	400 мс	100 мс	400 мс	1 с	Н
Вариация задержки пакета IP, IPDV	50 мс	50 мс	Н	Н	Н	Н
Коэффициент потери пакетов IP, IPLR	$1x10^{-3}$	$1x10^{-3}$	$1x10^{-3}$	$1x10^{-3}$	$1x10^{-3}$	Н
Коэффициент ошибок пакетов IP, IPER	$1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$	Н

Примечание: Н – не нормировано

Рекомендация Y.1541 устанавливает соответствие между классами качества обслуживания и приложениями:

- Класс 0 – приложения реального времени, чувствительные к джиттеру, характеризуемые высоким уровнем интерактивности (VoIP, видеоконференции);
- Класс 1 – приложения реального времени, чувствительные к джиттеру, интерактивные (VoIP, видеоконференции);
- Класс 2 – транзакции данных, характеризуемые высоким уровнем интерактивности (например, сигнализация);
- Класс 3 – транзакции данных, интерактивные;
- Класс 4 – приложения, допускающие низкий уровень потерь (короткие транзакции, массивы данных, потоковое видео);
- Класс 5 – традиционные применения сетей IP.

Архитектура сетевых механизмов обеспечения качества обслуживания в сетях IP

Помимо определения сетевых параметров и спецификации норм для них ИК 13 МСЭ-Т проводит в настоящее время работы по идентификации и стандартизации сетевых механизмов, обеспечивающих QoS в IP-ориентированных сетях. В мае 2004 г. была принята Рекомендация МСЭ Y.1291 [6], описывающая архитектурную модель для поддержки качества обслуживания в сетях с пакетной передачей.

Сетевые механизмы должны использоваться в комбинации с характеристиками качества обслуживания, формируемыми в зависимости от приложений. При разработке архитектуры сетевых механизмов учитывалось, что различные услуги будут иметь разнообразные требования к характеристикам сети. Например, для телемедицины точность доставки играет более существенную роль, чем суммарная средняя задержка или джиттер, тогда как для IP-телефонии джиттер и задержка являются ключевыми характеристиками и должны быть минимизированы.

С учетом тенденции постоянного расширения числа приложений с различными требованиями к характеристикам качества обслуживания архитектура поддержки QoS должна включать в себя широкий набор общих сетевых механизмов, как существующих, так и перспективных, подлежащих разработке.

Архитектура поддержки QoS определяет набор сетевых механизмов, называемых конструктивными блоками. В настоящее время определен начальный набор

конструктивных блоков, отвечающих трем логическим плоскостям: плоскости контроля, плоскости данных (информационной плоскости) и плоскости административного управления (см. рисунок).

Плоскость контроля. Механизмы QoS контрольной плоскости оперируют с путями, по которым передается трафик пользователей, и включают в свой состав:

- управление допуском (Admission Control, AC);
- маршрутизацию для QoS (QoS routing);
- резервирование ресурсов (Resource reservation).

Плоскость данных. Эта группа механизмов оперирует непосредственно с пользовательским трафиком и включает в себя:

- управление буферами (Buffer management);
- предотвращение перегрузок (Congestion avoidance);
- маркировку пакетов (Packet marking);
- организацию и диспетчеризацию очередей (Queuing and scheduling);
- формирование трафика (Traffic shaping);
- правила обработки трафика (Traffic policing);
- классификацию трафика (Traffic classification).

Плоскость административного управления. Эта плоскость содержит механизмы QoS, имеющие отношение к эксплуатации, администрированию и управлению сетью применительно к доставке пользовательского трафика. В число механизмов QoS на этой плоскости входят:

- измерения (Metering);
- заданные правила доставки (Policy);
- восстановление трафика (Traffic restoration);
- соглашение об уровне обслуживания (Service Level Agreement).

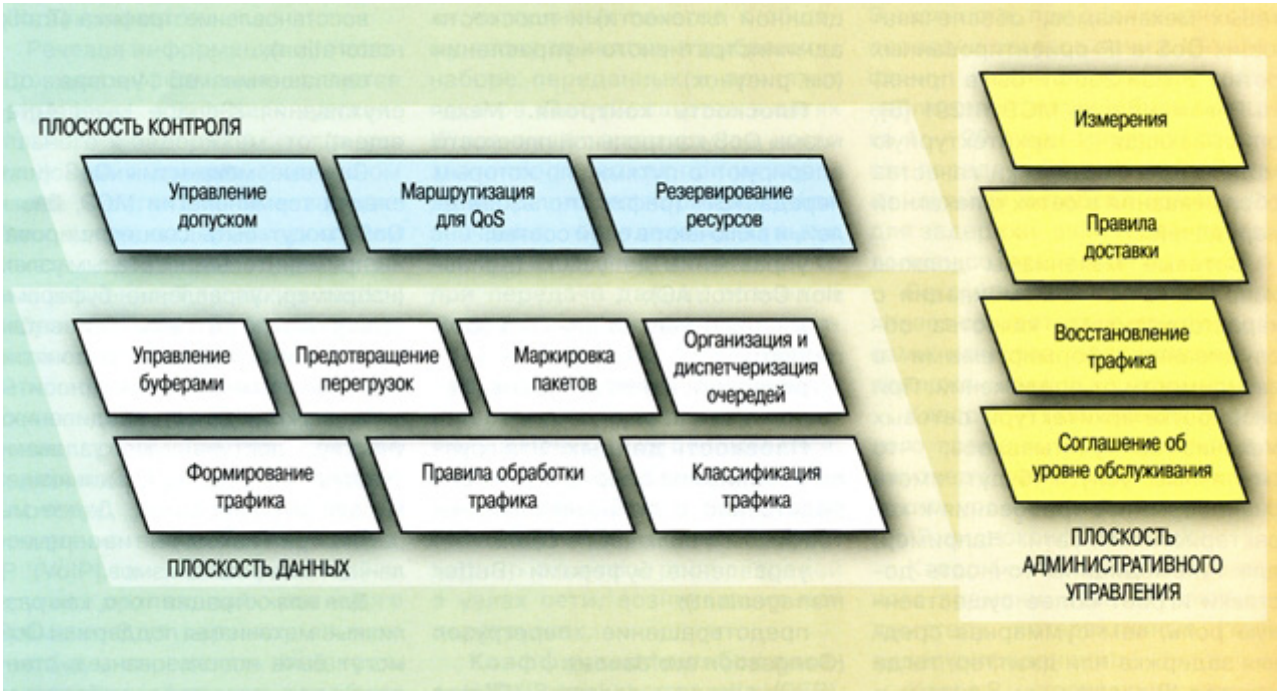
Сетевые механизмы QoS (или, следуя терминологии МСЭ, блоки QoS) могут быть специфицированы применительно к сетевым узлам (например, управление буферами узлов) или к сетевым сегментам (маршрутизация QoS), где понятие "сетевой сегмент" может относиться к межконцевому соединению, участку доступа, межузловому участку или участку, соединяющему две и более сетей. Далее мы рассмотрим некоторые из перечисленных выше механизмов.

Для иллюстрации того, как различные механизмы поддержки QoS могут быть использованы в стандартизованных методах обеспечения требуемых показателей качества обслуживания, мы рассмотрим два наиболее широко применяемых в настоящее время подхода при решении задачи обеспечения качества обслуживания: интегрированные (IntServ) и дифференцированные услуги (DiffServ).

Механизмы поддержки качества обслуживания в сетях IP

Как было отмечено выше, переход к сетям следующего поколения, построенным на базе стека протоколов IP, возможен только при условии, что для большого числа приложений будут обеспечены соответствующие показатели качества обслуживания. Для достижения этой цели был разработан ряд механизмов борьбы с задержками и потерями, которые в соответствии с разрабатываемой Рекомендацией МСЭ-TY.1291 разделены по трем

плоскостям – плоскости контроля, плоскости данных и плоскости административного управления.



Механизмы QoS в плоскости контроля

Управление допуском (Call Admission Control). Этот механизм контролирует новые заявки на пропуск трафика через сеть, определяя, может ли вновь поступающий трафик привести к перегрузке сети или к ухудшению уровня качества обслуживания для уже имеющегося в сети трафика. Обычно управление допуском построено на определенном наборе правил администрирования, контроля и управления сетевыми ресурсами.

Эти правила могут быть специфицированы в соответствии с потребностями сетевого провайдера или базироваться на соглашении между провайдером и пользователем и включать в свой состав различные параметры QoS. Для удовлетворения требований определенных служб (например, при чрезвычайных обстоятельствах), соответствующему трафику может быть присвоен высший приоритет при доступе в сеть.

Маршрутизация QoS (QoS routing) обеспечивает выбор пути, который удовлетворяет требованиям к качеству обслуживания для конкретного потока данных. Выбираемый путь может отличаться от кратчайшего пути. Процесс определения пути предполагает знание требований к качеству обслуживания со стороны потока данных и наличие информации о доступных сетевых ресурсах.

В настоящее время предложено большое число возможных методов определения наилучшего пути по критерию QoS. Как правило, в вычислениях наилучшего пути в маршрутизации QoS учитывается либо одна сетевая характеристика, либо две (производительность и задержка, стоимость и производительность, стоимость и задержка и т. д.), с тем, чтобы сделать процесс вычислений приемлемым для инженерных расчетов.

Резервирование ресурсов (Resource reservation). В целом, необходимым условием для обеспечения резервирования ресурсов является наличие ресурсов в сети. Резервирование ресурсов широко использовалось в сетях ATM при формировании постоянных

виртуальных соединений. В IP-ориентированных сетях наиболее типичным механизмом резервирования является механизм, базирующийся на протоколе RSVP.

Механизмы QoS в плоскости данных

Управление буферами (Buffer management). Управление буферами (или очередями) состоит в управлении пакетами, стоящими в узлах в очереди на передачу. Основные задачи управления очередями – минимизация средней длины очереди при одновременном обеспечении высокого использования канала, а также справедливое распределение буферного пространства между различными потоками данных. Схемы управления очередями различаются, в основном, критерием, по которому отбрасываются пакеты, и местом в очереди, откуда производится сброс пакетов (начало или конец очереди). Наиболее простым критерием для сброса пакетов является достижение очередью определенного порога, называемого максимальной длиной очереди.

Более распространены сегодня так называемые механизмы активного управления очередями. Типичным примером является алгоритм RED (Random Early Detection – раннее случайное обнаружение перегрузки). При использовании алгоритма RED поступающие в буфер пакеты сбрасываются на основании оценки средней длины очереди. Вероятность сброса пакетов растет с ростом средней длины очереди.

Предотвращение перегрузок (Congestion avoidance). Механизмы предотвращения перегрузок поддерживают уровень нагрузки в сети ниже ее пропускной способности. Обычный способ предотвращения перегрузок состоит в уменьшении трафика, поступающего в сеть. Как правило, команда уменьшить трафик влияет, в первую очередь, на низкоприоритетные источники. Одним из примеров механизмов предотвращения перегрузок является механизм окна в протоколе TCP.

Маркировка пакетов (Packet marking). Пакеты могут быть промаркированы в соответствии с определенным классом обслуживания. Маркировка обычно производится во входном пограничном узле, где в специальное поле заголовка (Type of Service в заголовке IP или DS-байт в заголовке DiffServ, см. ниже) вводится определенное значение. Кроме того, маркировка применяется для тех пакетов, которые могут быть удалены в случае перегрузки сети.

Организация и планирование очередей (Queuing and scheduling). Цель механизмов этой группы – выбор пакетов для передачи из буфера в канал. Большинство дисциплин обслуживания (или планировщиков) основаны на схеме "первый пришел – первый обслуживается". Для обеспечения более гибких процедур вывода пакетов из очереди был предложен ряд схем, основанных на формировании нескольких очередей. Среди них, в первую очередь, необходимо назвать схемы приоритетного обслуживания. Другой пример гибкой организации очереди – механизм взвешенной справедливой буферизации (Weighted Fair Queuing, WFQ), когда ограниченная пропускная способность на выходе узла распределяется между несколькими потоками (очередями) в зависимости от требований к пропускной способности со стороны каждого потока.

Еще одна схема организации очереди основана на классификации потоков по классу обслуживания (Class-Based Queuing, CBQ). Потоки классифицируются в соответствии с классами обслуживания и затем размещаются в буфере в различных очередях. Каждой очереди выделяется определенный процент выходной пропускной способности в зависимости от класса, и очереди обслуживаются по циклической схеме.

Формирование трафика (Traffic shaping). Формирование или управление характеристиками трафика предполагает контроль скорости передачи пакетов и объема потоков, поступающих на вход сети. В результате прохождения через специальные формирующие буферы уменьшается пачечность исходного трафика, и его характеристики становятся более предсказуемыми. Известны два механизма обработки трафика – Leaky Bucket ("дырявое ведро") и Token Bucket ("ведро с жетонами"). Алгоритм Leaky Bucket регулирует скорость пакетов, покидающих узел. Независимо от скорости входного потока, скорость на выходе узла является величиной постоянной. Когда ведро переполняется, лишние пакеты сбрасываются.

В противоположность этому, алгоритм Token Bucket не регулирует скорость на выходе узла и не сбрасывает пакеты. Скорость пакетов на выходе узла может быть такой же, как и на входе, если только в соответствующем накопителе ("ведре") есть жетоны. Жетоны генерируются с определенной скоростью и накапливаются в ведре. Алгоритм характеризуется двумя параметрами – скоростью генерации жетонов и размером памяти (размером "ведра") для них. Пакеты не могут покинуть узел, если в ведре нет жетонов. И наоборот, сразу пачка пакетов может покинуть узел, израсходовав соответствующее число жетонов.

Правила обработки трафика (Traffic policing). Этот блок принимает решение о том, соответствует ли поступающий от транзитного узла к транзитному узлу трафик заранее согласованным правилам обработки или контрактам. Обычно несоответствующие пакеты отбрасываются. Отправители могут быть уведомлены об отброшенных пакетах и обнаруженных причинах, а также о соблюдении соответствия в будущем, обусловленного соглашениями SLA.

Классификация трафика (Traffic classification). Классификация трафика может быть проведена на потоковом или пакетном уровне. На входе в сеть в узле доступа (пограничном маршрутизаторе) пакеты классифицируются для того, чтобы выделить пакеты одного потока, характеризуемого общими требованиями к качеству обслуживания. Затем трафик подвергается процедуре нормирования (механизм Traffic Conditioning). Нормирование трафика предполагает измерение его параметров и сравнение результатов с параметрами, оговоренным в контракте по трафику, известному как соглашение об уровне обслуживания (Service Level Agreement, SLA, см. ниже). Если условия SLA нарушаются, то часть пакетов может быть отброшена. Магистральные маршрутизаторы, составляющие ядро сети, обеспечивают пересылку пакетов в соответствии с требуемым уровнем QoS.

Механизмы QoS в плоскости административного управления

Измерения (Metering). Измерения обеспечивают контроль параметров трафика – например, скорость потока данных в сравнении с согласованной в SLA скоростью. По результатам измерений могут быть реализованы определенные процедуры – такие, как сброс пакетов и применение механизмов Leaky Bucket и Token Bucket.

Заданные правила доставки (Policy). Под правилами доставки здесь понимается набор правил, используемых для контроля и административного управления доступом к сетевым ресурсам. На основе таких правил поставщики услуг могут осуществлять реализацию механизмов в плоскости управления и плоскости данных. Возможными применениями правил доставки являются маршрутизация по заданным правилам, фильтрация пакетов на основе заданных правил (маркировка или отбрасывание пакетов), регистрация заданных потоков, правила обработки, связанные с безопасностью.

Восстановление трафика (Traffic restoration). Под восстановлением трафика в данной рекомендации понимается реакция сети, смягчающая последствия в условиях отказа. Восстановление трафика рассматривается на различных уровнях эталонной модели процессов. На физическом уровне при использовании SDH надежность обеспечивается автоматической защитной коммутацией. На канальном уровне транспортных сетей восстановление трафика обеспечивается специальными механизмами, развитыми для кольцевых и ячеистых структур. Соответствующие процедуры предусмотрены в технологии ATM. Восстановление на сетевом уровне (протокол IP) осуществляется с помощью технологии MPLS.

Соглашение об уровне обслуживания (Service Level Agreement). Одним из основных понятий в концепции обеспечения требуемого уровня качества обслуживания в современных сетях является соглашение об уровне обслуживания. Первые SLA-контракты были разработаны в середине 90-х годов при предоставлении услуг передачи данных с использованием технологий Frame Relay, ATM и IP. Необходимость подобных контрактов была вызвана возрастающими требованиями к операторам со стороны клиентов, чей бизнес все больше зависел от надежной и своевременной передачи информации. Контракт SLA предполагает повышенную ответственность поставщика услуг, дисциплинирует его. В какой-то степени это дисциплинирует и заказчика, поскольку заключению соглашения предшествует этап анализа требований к уровню сервиса.

Соглашение SLA, называемое в ряде источников контрактом по трафику, представляет собой контракт между пользователем и провайдером услуг/сетевым провайдером. В контракте определяются основные характеристики (профиль) трафика, формируемого в оборудовании пользователя, и параметры QoS, предоставляемые провайдером. Соглашение SLA может включать в себя также и ценовые характеристики. Техническая часть SLA специфицирует набор параметров и их значения, которые вместе определяют уровень обслуживания, обеспечиваемый трафику пользователя со стороны сетевого провайдера.

Контракт SLA может быть статическим (согласовывается на длительный период – месяц, год и т. п.) или динамическим (определяется для каждого сеанса). В последнем случае для запроса требуемого уровня QoS должен использоваться сигнальный протокол (например, RSVP). Соглашения SLA, прежде всего, предполагают четко регламентированные обязательства поставщика услуг по обеспечению их качества (время предоставления услуги, например, круглосуточно или только в рабочие дни; время реакции на инцидент; время выезда персонала к заказчику; время закрытия инцидента и т. д.), а также штрафные санкции за нарушение регламента. Из опыта зарубежных сетевых провайдеров известно, что стоимость SLA добавляется к стоимости гарантийного обслуживания и в ряде случаев может быть *в* несколько раз выше стоимости гарантийного обслуживания.

Модель предоставления интегрированных услуг (IntServ)

Процесс превращения сети Интернет в середине 90-х годов из академической в коммерческую инфраструктуру, рост числа узлов и количества пользователей, применение для разнообразных приложений с различными требованиями к качеству обслуживания – все эти факторы определили быстрое развитие механизмов поддержки QoS. В ответ на новые условия, возникшие в сетях IP, Комитет IETF предложил большой набор моделей и механизмов для обеспечения качества обслуживания в сетях Интернет, которые разделяются на две категории в соответствии с названиями рабочих групп

Комитета IETF, разрабатывающих эти модели и механизмы – интегрированных услуг и дифференцированных услуг.

Рабочая группа Integrated Services Working Group разрабатывала модель предоставления интегрированных услуг (или IntServ), основанную на принципе интегрированного резервирования ресурсов. Модель IntServ была разработана для поддержки приложений реального времени, чувствительных к задержкам. Механизмы, реализующие модель интегрированных услуг, должны обеспечивать взаимодействие всех сетевых устройств для поддержки любого уровня QoS вдоль пути передачи определенного потока пакетов.

Наиболее детально среди механизмов группы IntServ проработан протокол RSVP (Resource ReSerVation Protocol), спецификация которого (RFC 2205, [7]) была принята Комитетом IETF в 1997 г. Механизмы группы IntServ относятся к группе методов, гарантирующих "жесткое" или абсолютное качество обслуживания. Протокол RSVP является наиболее известным представителем группы механизмов интегрированного обслуживания. По существу, RSVP представляет собой протокол сигнализации, в соответствии с которым осуществляется резервирование и управление ресурсами с целью гарантии "жесткого" качества обслуживания. Резервирование производится для определенного потока IP-пакетов перед началом передачи этого потока. Идентификация потока (определение пакетов, принадлежащих одному потоку) осуществляется по специальной метке, размещаемой в основном заголовке каждого пакета IPv6. После резервирования пути начинается передача пакетов данного потока, обслуживаемых на всем межконцевом соединении с заданным качеством.

Протокол RSVP является только протоколом сигнализации. Для обеспечения требуемого качества обслуживания на фазе переноса пакетов трафика он должен быть дополнен одним из существующих протоколов маршрутизации, а также набором механизмов управления трафиком, включающих управление допустимостью соединений (CAC), классификацию трафика, управление и планирование очередей, а также другие механизмы, составляющие основу архитектуры механизмов поддержки QoS, рассмотренную выше.

Несмотря на возможности протоколов группы IntServ в плане обеспечения требуемых показателей QoS, реализация и развертывание методов интегрированного обслуживания связаны с определенными трудностями, особенно в территориально распределенных сетях. В частности, необходимо учитывать возможность перегрузки маршрутизаторов и переполнения накопителей в сетевых узлах при большом числе одновременно обслуживаемых потоков. Необходимо также признать, что протоколы группы IntServ не отвечают требованиям масштабируемости.

Достаточно высокими оказываются и требования к маршрутизаторам с точки зрения набора обязательных механизмов (RSVP, CAC и др.). Поэтому во второй половине 90-х годов (именно в этот период был отмечен взрывной рост сетей Интернет) начались работы по созданию моделей и механизмов предоставления дифференцированных услуг (Diff-Serv). Эти работы проводятся группой Differentiated Services Working Group Комитета IETF.

Модель предоставления дифференцированных услуг

Модель дифференцированных услуг (Differentiated Services, DiffServ) является логическим продолжением работ IETF над архитектурой IntServ. Недостатки, заложенные в самом принципе модели IntServ (жесткие гарантии качества обслуживания, низкий уровень

масштабирования) привели к необходимости создания более гибких механизмов обеспечения QoS. Общая характеристика принципов предоставления дифференцированных услуг (RFC-2475, [8]) была опубликована в декабре 1998 г., а более детальные спецификации появились в середине 1999 г. Методы DiffServ составляют группу механизмов, которые в отличие от методов IntServ обеспечивают относительное или "мягкое" качество обслуживания.

Основная идея механизмов DiffServ состоит в предоставлении дифференцированных услуг для набора классов трафика, отличающихся требованиями к показателям качества обслуживания. Как и в случае механизмов IntServ, для реализации дифференцированных услуг широко применяются механизмы, входящие в состав рассмотренной выше архитектуры поддержки QoS в сетях IP.

Одним из центральных понятий модели DiffServ является соглашение об уровне обслуживания, входящее в состав механизмов QoS на плоскости менеджмента. В модели DiffServ архитектура сети представляется в виде двух сегментов – пограничных участков и ядра. На входе в сеть в узле доступа (пограничном маршрутизаторе) пакеты классифицируются (механизм Traffic classification) для того, чтобы выделить пакеты одного потока, характеризуемого общими требованиями к качеству обслуживания. Затем трафик подвергается процедуре нормирования (механизм Traffic conditioning). Нормирование трафика предполагает измерение его параметров и сравнение результатов с параметрами, оговоренным в контракте SLA. Если условия SLA нарушаются, то часть пакетов может быть отброшена. При необходимости поток пакетов проходит через устройство профилирования (механизм Traffic shaping). Магистральные маршрутизаторы, составляющие ядро сети, обеспечивают пересылку пакетов в соответствии с требуемым уровнем QoS.

Требования к необходимому набору показателей качества обслуживания задаются в специальном однобайтовом поле каждого пакета – в октете Type of Service (ToS) протокола IPv4 или в октете Traffic Class (TC) протокола IPv6. Отметим, что в модели DiffServ это поле называется DS-байтом.

Содержание DS-байта определяет вид предоставляемых услуг. Первые два бита определяют приоритет пакета, следующие четыре – требуемый класс обслуживания пакета в узле, и два бита остаются неиспользуемыми. Класс обслуживания здесь означает механизм обработки и продвижения пакета из данного узла к следующему (Per-Hop Behavior, PHB) в соответствии с необходимым качеством обслуживания. Таким образом, с помощью поля DS можно определить до 32 различных уровней качества обслуживания.

В стандартах IETF RFC 2598 и RFC 2597 были определены два класса услуг для модели DiffServ. В спецификации RFC 2598 описан класс "срочной доставки" (Expedited Forwarding, EF), обеспечивающий наивысший из возможных уровней качества обслуживания (Premium Service) и применяемый для приложений, требующих доставки *с* минимальными значениями задержки и джиттера.

Второй класс обслуживания, получивший название "гарантированной доставки" (Assured Forwarding, AF), представлен в спецификации RFC 2597. Класс гарантированной доставки поддерживает уровень качества обслуживания более низкий, чем класс срочной доставки, но более высокий, чем обслуживание по принципу "наилучшей попытки" (Best effort). Внутри этого диапазона QoS класс AF определяет четыре типа трафика и три уровня отбрасывания пакетов. Таким образом, класс AF обеспечивает возможность обслуживания

до 12 разновидностей трафика в зависимости от набора требуемых показателей качества обслуживания.

Обработка пакетов в соответствии с определенными уровнем приоритета и типом трафика осуществляется специальными схемами обслуживания очередей, обеспечивающими контроль задержек и джиттера пакетов и исключение возможных потерь.

Среди основных механизмов управления очередями отметим приоритетное обслуживание (Priority Queuing), взвешенное справедливое обслуживание (Weighted Fair Queuing) и обслуживание в соответствии с механизмом PHB (Class-Based Queuing).

Относительная простота классификации трафика в модели DiffServ и отсутствие механизмов сквозного (end-to-end) резервирования ресурсов определяют широкие возможности применения дифференцированных услуг по сравнению с механизмами IntServ. Применение механизмов DiffServ в магистральном ядре сети позволяет использовать их для обработки агрегированного трафика, который может объединяться в пограничных сегментах сети. Такой подход может оказаться эффективным, например, в IP-телефонии, когда множество речевых потоков объединяются в один агрегированный, характеризуемый одинаковыми требованиями к показателям качества обслуживания.

По-видимому, механизмы DiffServ все же не могут гарантировать такой же уровень QoS, какой можно получить в цифровых телефонных сетях, базирующихся на коммутации каналов (например, в ISDN). Вместе с тем, можно ожидать, что в будущих сетях доля служб, требующих такой уровень качества, будет относительно небольшой, тогда как для приложений с менее критическими требованиями к QoS модели и механизмы дифференцированных услуг будут способны обеспечить необходимый уровень качества обслуживания.

Заключение

В заключение отметим следующее. Решение задачи обеспечения требуемого качества обслуживания в сетях IP, безусловно, может быть достигнуто прямым путем – на основе предоставления гарантированной полосы пропускания, повышения производительности сетевых устройств – маршрутизаторов и шлюзов, использовании магистралей с высокими пропускными способностями.

Однако, наиболее целесообразным представляется применение гибких методов, которые обеспечивают требуемые показатели качества обслуживания при эффективном использовании ресурсов сети для большого набора различных приложений, включая и наиболее критичные аудио- и видеоприложения реального времени.

Литература

1. McDysan. QoS and Traffic Management in IP and ATM Networks // McGraw-Hill. 2000.
2. Е.А. Кучерявый. Управление трафиком и качество обслуживания в сети Интернет//СПб, Наука и Техника. 2004.
3. Р. Кох, ГГ. Яновский. Эволюция и конвергенция в электросвязи//М., Радио и связь. 2001.
4. МСЭ-Т Recommendation Y.1540. IP Packet Transfer and Availability Performance Parameters//December 2002.
5. МСЭ-Т Recommendation Y.1541. Network Performance Objectives for IP-Based Services//May 2002.
6. МСЭ-Т Recommendation Y.1291. An Architectural Framework for Support of Quality of Service in Packet Networks //May 2004.
7. L. Zhang, R. Braden. Resource reservation Protocol//RFC-2205, September 1997.
8. S. Blake et al. Architecture for Differentiated Services//RFC-2475, December 1998.