Лекция 10
Качество сетевого обслуживания (служба QoS). Понятие QoS и основы построения службы QoS

Ещё в 1990-х годах появился достаточно широкий ряд типов сетевых приложений, нуждающихся в определенном уровне качества используемых сетевых соединений. К такого рода приложениям относятся, например, IP-телефония (VoIP-приложения), системы видеоконференцсвязи, интерактивной удалённой графики (сетевые игры, видео по запросу и пр.), приложения реального времени и ряд других типов приложений. В связи с этим возникла потребность в создании специальных средств обеспечения качества сетевого обслуживания (Quality of Service - QoS). Основными составляющими требований приложений к QoS являются высокая пропускная способность используемых сетевых соединений, низкие задержки в доставке пакетов данных и низкий уровень потерь пакетов данных (вплоть до недопустимости таких потерь).

В настоящей лекции рассматривается круг вопросов, связанных с организацией предоставления QoS сетевым приложениям.

10.1. Классификация потребностей сетевых приложений в QoS и формулирование требования к параметрам QoS

Сетевые приложения различных типов предъявляют разный уровень требований к указанным выше параметрам QoS. При этом приложения, более требовательные по одному из параметров, могут быть менее требовательными к другим. В настоящем параграфе проводится классификация приложений по их требованиям к различным параметрам QoS, а также рассматриваются вопросы формулирования требований к необходимым параметрам QoS.

10.1.1. Классификация сетевых приложений по их потребностям в предоставляемом QoS

Основными параметрами рассматриваемой ниже классификации являются:

1) скорость передачи данных, выполняемой приложением, и её относительная предсказуемость
2) чувствительность приложений к задержкам пакетов
3) чувствительность приложений к потерям и искажениям пакетов.

Классификация приложений по скорости выполняемой ими передачи данных

Скорость передачи данных сетевым приложением практически всегда не является постоянной величиной, а может варьироваться в очень широких пределах. Поэтому при классификации приложений по скорости передаваемого ими трафика важным показателем является не только максимальное или среднее значение этой скорости, но и определяющее относительную предсказуемость скорости трафика соотношение между этими параметрами, характеризуемое так называемым *коэффициентом пульсации трафика*. Этот коэффициент $K_п$ вычисляется по формуле $K_п = V_{max} : V_{ср}$ как отношение максимального значения мгновенной скорости к средней скорости.

По величине значения этого коэффициента сетевые приложения обычно разбивают на 2 класса: приложения с потоковым и с пульсирующим трафиком.

- ***Потоковый*** (*stream*) или равномерный трафик соответствует значениям $K_п$ < 10:1. Для приложений с этим типом трафика может быть организована передача данных через сеть со сравнительно *постоянной скоростью передачи*, CBR (Constant Bit Rate), и имеет легко вычислимую верхнюю верхнюю границу, равную $CBR*K_п$. Примерами приложений с потоковым трафиком являются приложения передачи аудио или видео информации. При этом для классической цифровой телефонии скорость передачи голосовой информации просто постоянна и, как известно читателю из гл.1 составляет 64 Кбит/сек. Отметим, что постоянная (а не средняя) скорость передачи информации (Constant Information Rate - CIR) может обеспечиваться как некоторыми технологиями передачи данных (например, ATM), так и некоторыми протоколами более высокого уровня, которые будут рассмотрены далее в настоящей главе.
- ***Пульсирующий*** (*burst* - взрывной) трафик характеризуется высокой степенью непредсказуемости текущего значения *переменной скорости передачи* (Variable Bit Rate - VBR). При этом возможны более или менее продолжительные периоды полного отсутствия трафика, чередующиеся с периодами пиковой загрузки, как правило, гораздо менее продолжительными. При этом в промежутки пиковой загрузки приложения с пульсирующим трафиком зачастую способны заполнить этим трафиком всю пропускную способность используемых каналов передачи данных (на наиболее “узкой” части маршрута передачи данных). Типичными примерами приложений с пульсирующим трафиком относятся службы FTP и HTTP, для которых периоды выбора очередного пересылаемого информационного объекта (файла/страницы) зачастую могут быть гораздо более продолжительными, чем промежутки времени, требуемые для пересылки этого объекта.

Формулирование требований “достаточной” скорости передачи данных для приложений с потоковым и пульсирующим трафиком может требовать задания различных параметров требований к QoS, рассматриваемым в следующем пункте настоящего параграфа.

Классификация приложений по их чувствительности к задержкам пакетов

Задержка (*latenc*y) – это время доставки пакета от источника к получателю. Величина задержки для различных пакетов одного и того же сетевого соединения может быть не постоянной, а варьироваться в зависимости, например, от времени нахождения в очередях маршрутизаторов по пути следования пакетов. Для некоторых сетевых приложений наличие задержек практически незаметно, другие приложения могут накладывать достаточно жёсткие ограничения на максимально допустимую величину задержки. Отметим также, что некоторые приложения, не очень чувствительные к относительно постоянным по величине задержкам, могут быть весьма чувствительны к высокой степени вариации задержек пакетов (в других словах - “разбросу”, дисперсии их значений), называемой также джиттером (jitter).

Рассмотрим типы сетевых приложений, упорядоченные по степени чувствительности к задержкам, начиная с наименее чувствительных.

- ***Асинхронные***. Практически не накладывают никаких ограничений на время задержек. Они также называются приложениями с «эластичным» трафиком. Ярким примером таких приложений является служба электронной почты.
- ***Синхронные***. Чувствительны к задержкам, но допускают их практически без ущерба для функциональности. Примерами являются службы FTP и HTTP - небольшая дополнительная задержка (к “времени реакции” службы) в начале пересылки файла или страницы практически неощутима пользователем.

- ***Интерактивные***. Задержки заметны пользователю, доставляют ему определённый дискомфорт, но не влекут отказов в функциональности приложений. Хорошим примером является текстовый редактор удаленных файлов, например, редактор системы Google Docs.
- **Изохронные**. Для таких приложений имеется порог чувствительности к величине задержки, при превышении которого резко ухудшается функциональность приложения. Например, во многих источниках указывают, что при передаче голосовой информации превышение задержками порога 100-150 мсек приводит к резкому снижению качества голоса, воспроизводимого принимающей стороной. Однако, читатели видели множество телемостов, проводимых с использованием спутниковых каналов, минимальная величина задержек на которых, между прочим, составляет около 300 мсек. При общении через такие телемосты заметны значительные паузы в начале ответов собеседников, но далее и голос, и изображение принимаются нормально. Но и аудио и видео приложения очень чувствительны к величине джиттера, превышение которыми порога 10 мсек в соответствии с данными Рекомендаций ITU-T G.712 приводит резкому ухудшению качества воспроизведения голосовой информации, не говоря о качестве воспроизведения видео. Однако спутниковые каналы включают лишь 2 основных хопа (канала между маршрутизаторами): Земля - спутник и спутник - Земля. Поэтому минимизируя число достаточно высокоскоростных хопов на Земле можно добиться очень малых значений величины джиттера и обеспечить очень качественное воспроизведение дошедших с заметной задержкой голоса и изображения.
- ***Сверхчувствительные***. Для таких приложений, относящихся к приложениям реального времени, превышение задержками некоторого порога (зачастую очень малого) влечёт полный отказ в функциональности приложения. К числу таких приложений относятся, в частности, приложения управления в реальном времени сложными техническими объектами и системами. Опоздание управляющего сигнала на те или иные исполнительные механизмы системы и вызванная этим запоздалая реакция может привести к аварийной ситуации.

Отметим, что зачастую вместо приведенной детальной классификации используют более грубую, разделяющую приложения по степени их чувствительности к задержкам всего на 2 класса: асинхронные и синхронные. При этом в класс асинхронных приложений грубой классификации относят приложения, допускающие задержки в несколько секунд. Таким свойством, очевидно, обладают лишь первых 2 класса детальной классификации. А все остальные классы приложений детальной классификации (начиная с интерактивных) попадают в класс синхронных приложений в грубой классификации.

Классификация приложений по их чувствительности к потерям и искажениям пакетов

По степени чувствительности приложений к потерям и искажениям пакетов их обычно делят на 2 класса: чувствительные и нечувствительные.

- ***Чувствительные к потерям пакетов приложения*** значительно, а иногда и полностью, утрачивают свою функциональность при потере или искажении хотя бы одного пакета. К такого рода приложениям относятся приложения, пересылающие алфавитно-цифровую информацию, включая символьные файлы, файлы архивов (со сжатыми файлами), зашифрованную информацию, бинарные исполнимые файлы и пр. При этом если потеря или искажение пакета с важной информацией символьного файла может лишь более или менее сильно уменьшить степень полезности этого файла, то искажение хотя бы одного бита

файлов остальных указанных типов может привести к невозможности открыть архив, расшифровать зашифрованную информацию или корректно выполнить программу. Отметим, что большинство сетевых приложений относятся именно к рассматриваемому классу.

- ***Нечувствительные к потерям пакетов приложения*** допускают ограниченные потери пакетов практически без ущерба их функциональности. Однако, если будет превышен некоторый порог доли потерянных пакетов, функциональность таких приложений может резко ухудшиться. К такому рода приложений относятся, например, приложения, передающие через сеть аудио или видео информацию. Отметим, что устойчивость таких приложений к потерям информации как правило обеспечивается не сама собой, а достигается за счёт специального кодирования передаваемой информации с её декодирования при приёме, выполняемых специальными программами - кодеками. Допустимый порог потерь пакетов при этом сильно зависит от наличия и “продвинутости” применяемых кодеков. Так известные Рекомендации ITU-T G.712 определяют пороговое значение допустимой доли потерянных пакетов равным 1%. Но, как показано в работе (Singh H.P., Singh S., Singh J. Real Time Analysis of VoIP System under Pervasive Environment through Spectral Parameters // International Journal of Computer Applications. 2011. Vol. 31. № 2), за счёт использования существенно более продвинутого по сравнению со стандартным кодеком G.711a кодека Speex порог чувствительности к потерям может быть повышен до 10% (в 10 раз!), а порог чувствительности к джиттеру - до 15 мсек (в 1,5 раза по сравнению со значениями, указанными в Рекомендациях ITU-T G.712). Дополнительно отметим, что к настоящему времени кодек Speex, как указано на сайте этого кодека, к настоящему времени превзойдён по всем параметрам новым свободно распространяемым кодеком Opus, стандартизованным как RFC 6716. Этот стандарт объединяет технологии таких известных кодеков, как Skype SILK и Xipn.Org CELT.

В завершение рассмотрения 3-х основных характеристик сетевых приложений с точки зрения QoS отметим, что все они не являются сколь либо зависимыми. Набор характеристик, присущий конкретному приложению, в принципе, может быть достаточно произвольным, но на практике не для любого сочетания характеристик существует обладающее ими приложение. В качестве примеров характеристик распространённых типов приложений можно указать приложения с потоковым трафиком, изохронные, нечувствительные к потерям (VoIP приложения, средства видеоконференцсвязи и пр.) и приложения с пульсирующим трафиком, синхронные (в смысле детальной классификации типа трафика), чувствительные к потерям FTP и HTTP. Отметим также, что основным 4-рём наборам рассматриваемых характеристик, которым может быть поставлено в соответствие какое-либо приложение ещё в технологии ATM (см. лекцию 4) присвоены названия классов A, B, C и D, а остальные приложения отнесены к классу F.

10.1.2 Формулирование требований к параметрам QoS

Рассмотренные три характеристики сетевых приложений, применяемых индивидуальными и/или корпоративными пользователем при доступе к интернету, могут быть взяты в качестве рамочной основы для формулирования трёх групп требований пользователя к качеству сетевых услуг, предоставляемых ему оператором доступа в интернет (провайдером). Эти требования обычно формулируются в соглашении об уровне (сетевого) сервиса (Service Level Agreement - SLA), включаемом в договор о предоставлении услуг доступа в интернет в виде раздела этого договора или дополнительного соглашения к нему.

Рассмотрим вкратце три упомянутые группы требований к к предоставляемому провайдером QoS.

- ***Требования к параметрам пропускной способности.*** Могут включать либо требования гарантированной общей пропускной способности CIR (Consnant Information Rate), либо требования гарантированной средней пропускной способности CBR (Constant Bit Rate) и допустимого объёма пульсаций (превышений объёма передаваемого трафика по сравнению с объёмом, преданным со скоростью CBR) за определённый промежуток времени.
- ***Требования к параметрам задержек***. В число этих параметров могут быть включены максимально допустимая и средние величины задержки пакетов, а также максимально допустимые и средние значения джиттера.
- ***Требования к параметрам надёжности передачи данных*** обычно включают лишь максимальный процент потерянных и/или искажённых пакетов.

Отметим, что в каждую из этих групп требований входят требования к средним значениям некоторых параметров или некоторых процентных соотношений. А при вычислении таких значений важнейшим обстоятельством является то, за какой временн'ой период выполняется усреднение или вычисляется процент. Чем больше будет величина этого периода, тем менее жёсткими будут требования QoS. Так требование потерь не более 1% пакетов в течение суток будет выполнено, если на протяжение 14 минут все пакеты будут потеряны, а оставшуюся часть суток потерь совсем не будет. Но, очевидно, 14-минутное "отсутствие сети" в самый неподходящий (в соответствии с известным законом бутерброда) момент времени может совершенно не устроить заказчика SLA. Поэтому для того, чтобы сформулировать требования QoS в по возможности жёсткой форме необходимо зафиксировать в SLA по возможности малое значения периода усреднения показателей и вычисления процентных соотношений.

10.2. Общая организация службы QoS

Для того, чтобы обеспечить выполнение требований Qos необходима специальная служба. Это связано с тем, что, сетевые соединения, используемые приложениями, предъявляющими те или иные требования QoS к этим соединениям, как правило являются достаточно протяжёнными и проходят через множество промежуточных каналов передачи данных, естественно, соединённых через множество промежуточных маршрутизаторов. В таких обстоятельствах скорость передачи данных приложением ограничена самым медленным и/или перегруженным каналом, а весьма существенная часть времени задержки пакетов может составлять время их задержки на одном из самых медленных и/или перегруженных маршрутизаторов. Сетевым же приложениям требуется запрашиваемое QoS на протяжении всего маршрута "из конца в конец" (end-to-end), включая перегруженные участки этого маршрута. Для того, чтобы обеспечить это свойство необходима специальная распределённая служба, обеспечивающая для определённых сетевых соединений требуемое качество передачи данных из конца в конец маршрута их пересылки. Такое качество может быть обеспечено, например, путём предварительного резервирования ресурсов пропускной способности каналов и маршрутизаторов по всему маршруту пересылки данных "особых" сетевых соединений или путём специальной приоритетной обработки пакетов этих соединений в узких местах сетевых маршрутов.

Перейдём к рассмотрению общей организации службы QoS.

10.2.1. Базовая архитектура службы QoS

Базовая архитектура службы QoS строится из элементов трёх типов: Средств QoS уровня узла сети, протоколов сигнализации службы QoS и централизованных средств политики, распределения и учёта ресурсов обеспечения QoS.

- ***Средства QoS уровня узла сети*** выполняют обработку пакетов трафика, поступающих во входные порты этих узлов, необходимую для обеспечения требуемого качества обслуживания. При этом в качестве узлов сети канального уровня используются коммуникационные устройства канального уровня, такие, как коммутаторы в сетях Ethernet. В качестве узлов IP-сетей, очевидно, используются маршрутизаторы. На каждом из этих уровней имеются и могут использоваться специфичные для этого уровня реализации средств QoS уровня узла. Но в подавляющем большинстве случаев сетевые соединения, требующие обеспечения QoS, проходят через несколько сегментов канального уровня, то есть, через построенную из этих сегментов IP-сеть. Поэтому нас в первую очередь будут интересовать средства QoS уровня узла сети, применяемые на маршрутизаторах.
- ***Протоколы сигнализации QoS*** обеспечивают согласованную работу всех узлов сети по всему маршруту передачи данных из конца в конец с целью организации предоставления требуемого QoS на всём этом маршруте. По отмеченным выше соображениям нас будут интересовать протоколы сигнализации QoS, применяемые при реализации службы QoS на уровне IP.
- ***Централизованные средства политики, распределения и учёта ресурсов обеспечения QoS*** необходимы потому, что коммуникационных ресурсов узлов сети (например, пропускной способности их портов) может с очень высокой вероятностью не хватить для одновременного предоставления сколь либо высокого качества обслуживания всех проходящих через эти узлы сетевых соединений. Поэтому необходим централизованный распорядитель ресурсами QoS сети, ведущий учёт доступных для распределения ресурсов QoS и предоставляющий возможность выделения этих ресурсов для определённых сетевых соединений в соответствии с некоторыми правилами соответствующей политики.

Перейдем к обзорному рассмотрению указанных элементов базовой архитектуры глобальной службы QoS. Более детальное рассмотрение некоторых из них приводится в следующей лекции.

10.2.2 Обзор средств QoS уровня узла сети

Средства QoS уровня узла сети выполняют основную работу, требуемую для обеспечения QoS для проходящих через эти узлы потоков трафика, поскольку именно они выполняют всю работу по продвижению пакетов этого трафика с входных портов узла в выходные порты, ведущие по направлению к получателю трафика. В процессе этой работы как раз и формируются те значения показателей QoS, которые для каждого узла сети должны удовлетворять требованиям, сформулированным для передачи трафика из конца в конец от его отправителя к его получателю.

Средства QoS узла сети включают средства двух типов: средства обслуживания очередей и средства кондиционирования трафика.

Средства обслуживания очередей являются неотъемлемой частью любого коммуникационного устройства (коммутатора, маршрутизатора), работающего по принципу коммутации пакетов. В состав этих средств могут входить компоненты,

реализующие различные алгоритмы от бесхитростной и безприоритетной очереди FIFO (обслуживание строго в порядке поступления) до более сложных алгоритмов приоритетного и пропорционального обслуживания, а также некоторых форм изощрённого комбинирования этих алгоритмов. Более подробно эти алгоритмы рассматриваются в следующей лекции.

Средства кондиционирования трафика могут включать в свой состав средства классификации трафика, средства профилирования трафика и средства формирования трафика.

- ***Классификация трафика*** выполняется на пограничных коммуникационных устройствах сети (граничных коммутаторах сегментов; пограничных маршрутизаторах AS и, возможно, их подсетей, пограничных маршрутизаторах сетей MPLS). При выполнении классификации из общего потока входного трафика порта выделяются выделяются потоки (потоки индивидуальных соединений или агрегированные потоки), имеющие общие требования к QoS. При выполнении классификации могут учитываться множество различных параметров каждого передаваемого пакета, включая IP-адреса источника и получателя пакета, номера портов источника и получателя пакетов и пр. Результат классификации передаётся следующим по маршруту пересылки пакетов узлам сети через специальные, упомянутые ниже, поля заголовка пакета, играющие роль протокола сигнализации.
- ***Профилирование трафика на основе правил политики*** (policing). Каждому входному потоку узла, выявленному на стадии классификации трафика, может быть поставлен в соответствие некоторый набор параметров QoS, который может включать, например, согласованную среднюю скорость потока и согласованный максимальный объём допустимой пульсации. Такой набор параметров называется профилем трафика. Суть профилирования состоит в том, что параметры поступающего трафик потока сопоставляются с параметрами профиля и, в случае превышения при обработке некоторого пакета фактических значений параметров над значениями соответствующих параметров профиля, этот пакет отбрасывается или специальным образом маркируется для того, чтобы при определённом количестве повторных нарушений быть безусловно отброшенным. Информация о маркировке пакета передаётся в специальных битах соответствующего поля (для IP-пакета - DS-байта) заголовка пакета. Это поле выполняет функции протокола сигнализации. Отметим, что входящая в состав дополнительных функций маршрутизаторов и применяемая на пограничных маршрутизаторах функция шейпинга (shaping), по сути реализует профилирование агрегированных потоков входящего извне трафика (с безусловным отбрасыванием нарушающих профиль пакетов) для профиля, единственным параметром которого является максимально допустимая мгновенная скорость потока.
- **Формирование трафика** (shaping, здесь это слово имеет иной смысл, чем в названии “shaping” дополнительной функции маршрутизатора). Суть функции формирования трафика состоит в придании обрабатываемому этой функцией трафику требуемой временн’ой “формы”. Основным назначением этой функции является сглаживание пульсаций трафика с целью обеспечения более равномерных интервалов между передачей пакетов. Это способствует уменьшению очередей на коммуникационных устройствах, следующих за данным по направлению потока трафика и, как следствие, способствует уменьшению джиттера. А это, в свою очередь, способствует повышению качества передачи голосовых и видео потоков.

10.2.3. Краткий обзор протоколов сигнализации

Протоколы сигнализации QoS можно разбить на 2 класса. К первому классу относятся протоколы, ориентированные на поддержку гарантированного или “жесткого” качества обслуживания (см. следующую лекцию), ко второму - протоколы, поддерживающие реализацию требований QoS с максимальными усилиями (best efforts), но без гарантий того, что эти требования будут реализованы в полной мере (“мягкое” качество обслуживания).

В первом случае применяется протокол резервирования RSVP или его расширения. Средствами этого протокола в начале установления сетевого соединения на всех узлах сети, расположенных по маршруту передачи трафика соединения резервируются сетевые ресурсы, необходимые для гарантированного обеспечения каждым из узлов предъявленных требований к качеству сетевого обслуживания.

Во втором случае в качестве протокола сигнализации используется группа полей заголовка пакета. Так в применяемом в технологии Ethernet протоколе IEEE 802.1p, предназначенном для приоритетного обслуживания пакетов в очередях к коммутаторам информация о приоритете кадра передаётся в 3-х битах метки кадра. В протоколе IPv4 роль протокола сигнализации играет DS байт заголовка IP-пакета, ранее называвшийся байтом TOS. В протоколе IPv6 роль протокола сигнализации службы QoS играют поля “приоритет” и “метка потока”.