2. Требования к компьютерным сетям

Главным требованием, предъявляемым к сетям, является выполнение сетью того набора услуг, для оказания которых она предназначена: например, предоставление доступа к файловым архивам или страницам публичных web-сайтов, обмен электронной почтой в пределах предприятия или в глобальных масштабах, интерактивный обмен речевыми сообщениями IP-телефонии и т. п.

Все остальные требования — *производительность, надежность, совместимость, управляемость, защищенность, расширяемость и масштабируемость* — связаны с качеством выполнения основной задачи. И хотя все эти требования весьма важны, часто понятие «качество обслуживания» (Quality of Service, QoS) компьютерной сети трактуется более узко — в него включаются только две самые важные характеристики сети — производительность и надежность.

2.1. Производительность

Потенциально высокая производительность — это одно из основных преимуществ распределенных систем, к которым относятся компьютерные сети. Это свойство обеспечивается принципиальной, но не всегда практически реализуемой возможностью распараллеливания работ между несколькими компьютерами сети.

Существует несколько основных характеристик производительности сети:

1. время реакции;
2. скорость передачи данных;
3. задержка передачи и вариация задержки передачи.

Время реакции сети является интегральной характеристикой производительности сети с точки зрения пользователя. В общем случае время реакции определяется как интервал времени между возникновением запроса пользователя к какой-либо сетевой службе и получением ответа на этот запрос.

Значение этого показателя зависит от типа этой службы, от того, какой пользователь и к какому серверу обращается, а также от текущего состояния элементов сети — загруженности сегментов, коммутаторов и маршрутизаторов, через которые проходит запрос, загруженности сервера и т. п.

Поэтому имеет смысл использовать также и средневзвешенную оценку времени реакции сети, усредняя этот показатель по пользователям, серверам и времени дня (от которого в значительной степени зависит загрузка сети).

Время реакции сети обычно складывается из нескольких составляющих. В общем случае в него входит время подготовки запросов на клиентском компьютере, время передачи

запросов между клиентом и сервером через сегменты сети и промежуточное коммуникационное оборудование, время обработки запросов на сервере, время передачи ответов от сервера клиенту и время обработки получаемых от сервера ответов на клиентском компьютере.

Ясно, что разложение времени реакции на составляющие не интересует пользователя — ему важен конечный результат, однако для сетевого специалиста очень важно выделить из общего времени реакции составляющие, соответствующие этапам собственно сетевой обработки данных, — передачу данных от клиента к серверу через сегменты сети и коммуникационное оборудование.

Знание сетевых составляющих времени реакции дает возможность оценить производительность отдельных элементов сети, выявить узкие места и в случае необходимости выполнить модернизацию сети для повышения ее общей производительности.

Скорость передачи данных отражает объем данных, переданных сетью или ее частью в единицу времени. Пропускная способность уже не является пользовательской характеристикой, так как она говорит о скорости выполнения внутренних операций сети — передаче пакетов данных между узлами сети через различные коммуникационные устройства. Зато она непосредственно характеризует качество выполнения основной функции сети — транспортировки сообщений — и поэтому чаще используется при анализе производительности сети, чем время реакции.

Скорость передачи данных измеряется либо в битах в секунду, либо в пакетах в секунду. Пропускная способность может быть мгновенной, максимальной и средней.

Средняя скорость вычисляется путем деления общего объема переданных данных на время их передачи, причем выбирается достаточно длительный промежуток времени — час, день или неделя.

Мгновенная скорость отличается от средней тем, что для усреднения выбирается очень маленький промежуток времени — например, 10 мс или 1 с.

Максимальная скорость — это наибольшая мгновенная пропускная способность, зафиксированная в течение периода наблюдения. Максимально достижимая скорость передачи данных называется пропускной способностью элемента сети.

Чаще всего при проектировании, настройке и оптимизации сети используются такие показатели, как средняя и максимальная пропускные способности. Средняя пропускная способность отдельного элемента или всей сети позволяет оценить работу сети на большом промежутке времени, в течение которого в силу закона больших чисел пики и спады интенсивности трафика компенсируют друг друга. Пропускная способность позволяет

оценить возможности сети справляться с пиковыми нагрузками, характерными для особых периодов работы сети, например утренних часов, когда сотрудники предприятия почти одновременно регистрируются в сети и обращаются к разделяемым файлам и базам данных.

Пропускную способность можно измерять между любыми двумя узлами или точками сети, например между клиентским компьютером и сервером, между входным и выходным портами маршрутизатора. Для анализа и настройки сети очень полезно знать данные о пропускной способности отдельных элементов сети.

Из-за последовательного характера передачи данных различными элементами сети общая пропускная способность любого составного пути в сети будет равна *минимальной* из пропускных способностей составляющих элементов маршрута. Для повышения пропускной способности составного пути необходимо в первую очередь обратить внимание на самые медленные элементы. Иногда полезно оперировать *общей пропускной способностью сети,* которая определяется как максимальное количество информации, переданной между всеми узлами сети в единицу времени. Этот показатель характеризует качество сети в целом, не дифференцируя его по отдельным сегментам или устройствам.

Обычно при определении пропускной способности сегмента или устройства в передаваемых данных не выделяется трафик какого-то определенного пользователя, приложения или компьютера — подсчитывается общий объем передаваемой информации. Тем не менее для более точной оценки качества обслуживания такая детализации желательна, и в последнее время системы управления сетями все чаще позволяют ее выполнять.

Задержка передачи определяется как задержка между моментом поступления данных на вход какого-либо сетевого устройства или части сети и моментом появления их на выходе этого устройства. Этот параметр производительности по смыслу близок ко времени реакции сети, но отличается тем, что всегда характеризует только сетевые этапы обработки данных, без задержек обработки конечными узлами сети. Обычно качество сети характеризуют величинами *максимальной задержки передачи* и *вариацией задержки.* Не все типы трафика чувствительны к задержкам передачи, во всяком случае, к тем величинам задержек, которые характерны для компьютерных сетей, — обычно задержки не превышают сотен миллисекунд, реже — нескольких секунд. Такого порядка задержки пакетов, порождаемых файловой службой, службой электронной почты или службой печати, мало влияют на качество этих служб с точки зрения пользователя сети. С другой стороны, такие же задержки пакетов, переносящих голосовые данные или видеоизображение, могут приводить к значительному снижению качества предоставляемой пользователю информации —

возникновению эффекта «эха», невозможности разобрать некоторые слова, дрожание изображения и т. п.

Пропускная способность и задержки передачи являются независимыми параметрами, так что сеть может обладать, например, высокой пропускной способностью, но вносить значительные задержки при передаче каждого пакета. Пример такой ситуации дает канал связи, образованный геостационарным спутником. Пропускная способность этого канала может быть весьма высокой, например 2 Мбит/с, в то время как задержка передачи всегда составляет не менее 0,24 с, что определяется скоростью распространения электрического сигнала (около 300 000 км/с) и длиной канала (72 000 км).

2.2. Надежность и безопасность

Одной из первоначальных целей создания распределенных систем, к которым относятся и вычислительные сети, являлось достижение большей надежности по сравнению с отдельными вычислительными машинами.

Важно различать несколько аспектов надежности. Для технических устройств используются такие показатели надежности, как среднее время наработки на отказ, вероятность отказа, интенсивность отказов. Однако эти показатели пригодны только для оценки надежности простых элементов и устройств, которые могут находиться только в двух состояниях — работоспособном или неработоспособном. Сложные системы, состоящие из многих элементов, кроме состояний работоспособности и неработоспособности, могут иметь и другие промежуточные состояния, которые эти характеристики не учитывают. В связи с этим для оценки надежности сложных систем применяется другой набор характеристик.

Готовность, или *коэффициент готовности (availability),* означает долю времени, в течение которого система может быть использована. Готовность может быть улучшена путем введения избыточности в структуру системы: ключевые элементы системы должны существовать в нескольких экземплярах, чтобы при отказе одного из них функционирование системы обеспечивали другие.

Чтобы компьютерную систему можно было отнести к высоконадежным, она должна как минимум обладать высокой готовностью, но этого недостаточно. Необходимо обеспечить *сохранность данных* и защиту их от искажений. Кроме этого, должна поддерживаться *согласованность* (непротиворечивость) данных, например, если для повышения надежности на нескольких файловых серверах хранится несколько копий данных, то нужно постоянно обеспечивать их идентичность.

Так как сеть работает на основе механизма передачи пакетов между конечными узлами, то одной из характерных характеристик надежности является *вероятность доставки пакета*

узлу назначения без искажений. Наряду с этой характеристикой могут использоваться и другие показатели: вероятность потери пакета, вероятность искажения отдельного бита передаваемых данных, отношение потерянных пакетов к доставленным.

Другим аспектом общей надежности является *безопасность (security),* то есть способность системы защитить данные от несанкционированного доступа. В распределенной системе это сделать гораздо сложнее, чем в централизованной. В сетях сообщения передаются по линиям связи, часто проходящим через общедоступные помещения, в которых могут быть установлены средства прослушивания линий. Другим уязвимым местом могут быть оставленные без присмотра персональные компьютеры. Кроме того, всегда имеется потенциальная угроза взлома защиты сети от неавторизованных пользователей, если сеть имеет выходы в глобальные сети общего пользования.

Еще одной характеристикой надежности является *отказоустойчивость (fault tolerance).* В сетях под отказоустойчивостью понимается способность системы скрывать от пользователя отказ отдельных ее элементов. Например, если копии таблицы базы данных хранятся одновременно на нескольких файловых серверах, то пользователи могут просто не заметить отказ одного из них. В отказоустойчивой системе отказ одного из ее элементов приводит к некоторому снижению качества ее работы (деградации), а не к полному останову. Так, при отказе одного из файловых серверов увеличивается только время доступа к базе данных из-за уменьшения степени распараллеливания запросов, но в целом система продолжает выполнять свои функции.

2.3. Расширяемость и масштабируемость

Термины *расширяемость* и *масштабируемость* иногда используют как синонимы, но это неверно — каждый из них имеет четко определенное самостоятельное значение.

Расширяемость (extensibility) означает возможность сравнительно легкого добавления отдельных элементов сети (пользователей, компьютеров, приложений, служб), наращивания длины сегментов сети и замены существующей аппаратуры более мощной. При этом принципиально важно, что легкость расширения системы иногда может обеспечиваться в некоторых весьма ограниченных пределах. Например, локальная сеть Ethernet, построенная на основе одного сегмента толстого коаксиального кабеля, обладает хорошей расширяемостью, в том смысле, что позволяет легко подключать новые станции. Однако такая сеть имеет ограничение на число станций — оно не должно превышать 30-40. Хотя сеть допускает физическое подключение к сегменту и большего числа станций (до 100), но при этом чаще всего резко снижается производительность сети. Наличие такого ограничения и является признаком плохой масштабируемости системы при хорошей расширяемости.

Масштабируемость (scalability) означает, что сеть позволяет наращивать количество узлов и протяженность связей в очень широких пределах, при этом производительность сети не ухудшается. Для обеспечения масштабируемости сети приходится применять дополнительное коммуникационное оборудование и специальным образом структурировать сеть. Например, хорошей масштабируемостью обладает многосегментная сеть, построенная с использованием коммутаторов и маршрутизаторов и имеющая иерархическую структуру связей. Такая сеть может включать несколько тысяч компьютеров и при этом обеспечивать каждому пользователю сети нужное качество обслуживания.

2.4. Прозрачность

Прозрачность (transparency) сети достигается в том случае, когда сеть представляется пользователям не как множество отдельных компьютеров, связанных между собой сложной системой кабелей, а как единая традиционная вычислительная машина с системой разделения времени.

Прозрачность может быть достигнута на двух различных уровнях — на уровне пользователя и на уровне программиста. На уровне пользователя прозрачность означает, что для работы с удаленными ресурсами он использует те же команды и привычные ему процедуры, что и для работы с локальными ресурсами. На программном уровне прозрачность заключается в том, что приложению для доступа к удаленным ресурсам требуются те же вызовы, что и для доступа к локальным ресурсам. Прозрачность на уровне пользователя достигается проще, так как все особенности процедур, связанные с распределенным характером системы, маскируются от пользователя программистом, который создает приложение. Прозрачность на уровне приложения требует сокрытия всех деталей распределенности средствами сетевой операционной системы.

Сеть должна скрывать все особенности операционных систем и различия в типах компьютеров. Пользователь компьютера Macintosh должен иметь возможность обращаться к ресурсам, поддерживаемым UNIX-системой, а пользователь UNIX должен иметь возможность разделять информацию с пользователями Windows.

Концепция прозрачности может быть применена к различным аспектам сети. Например, прозрачность расположения означает, что от пользователя не требуется знаний о месте расположения программных и аппаратных ресурсов, таких как процессоры, принтеры, файлы и базы данных. Аналогично, прозрачность перемещения означает, что ресурсы должны свободно перемещаться из одного компьютера в другой без изменения своих имен. Еще одним из возможных аспектов прозрачности является прозрачность параллелизма, заключающаяся в том, что процесс распараллеливания вычислений происходит

автоматически, без участия программиста, при этом система сама распределяет параллельные ветви приложения по процессорам и компьютерам сети. В настоящее время нельзя сказать, что свойство прозрачности в полной мере присуще многим вычислительным сетям, это скорее цель, к которой стремятся разработчики современных сетей.

2.5. Поддержка разных видов трафика

Компьютерные сети изначально предназначены для совместного доступа пользователя к ресурсам компьютеров: файлам, принтерам и т. п. Трафик, создаваемый этими традиционными службами компьютерных сетей, имеет свои особенности и существенно отличается от трафика сообщений в телефонных сетях или, например, в сетях кабельного телевидения. Однако 90-е годы стали годами проникновения в компьютерные сети трафика мультимедийных данных, представляющих в цифровой форме речь и видеоизображение. Естественно, что для динамической передачи мультимедийного трафика требуются иные алгоритмы и протоколы и, соответственно, другое оборудование.

Главной особенностью трафика, образующегося при динамической передаче голоса или изображения, является наличие жестких требований к синхронности передаваемых сообщений. Для качественного воспроизведения непрерывных процессов, которыми являются звуковые колебания или изменения интенсивности света в видеоизображении, необходимо получение измеренных и закодированных амплитуд сигналов с той же частотой, с которой они были измерены на передающей стороне. При запаздывании сообщений будут наблюдаться искажения.

В то же время трафик компьютерных данных характеризуется крайне неравномерной интенсивностью поступления сообщений в сеть при отсутствии жестких требований к синхронности доставки этих сообщений. Например, доступ пользователя, работающего с текстом на удаленном диске, порождает случайный поток сообщений между удаленным и локальным компьютерами, зависящий от действий пользователя, причем задержки при доставке мало влияют на качество обслуживания пользователя сети. Сегодня практически все новые протоколы в той или иной степени предоставляют поддержку мультимедийного трафика.

Особую сложность представляет *совмещение* в одной сети традиционного *компьютерного* и *мультимедийного трафика.* Обычно протоколы и оборудование компьютерных сетей относят мультимедийный трафик к факультативному, поэтому качество его обслуживания оставляет желать лучшего. Сегодня затрачиваются большие усилия по созданию сетей, не ущемляющих интересы каждого из типов трафика. Наиболее близки к

этой цели сети на основе технологии ATM, разработчики которой изначально учитывали случай сосуществования разных типов трафика в одной сети.

2.6. Управляемость

Управляемость сети подразумевает возможность централизованно контролировать состояние основных элементов сети, выявлять и разрешать проблемы, возникающие при работе сети, выполнять анализ производительности и планировать развитие сети. В идеале средства управления сетями представляют собой систему, осуществляющую наблюдение, контроль и управление каждым элементом сети — от простейших до самых сложных устройств, при этом такая система рассматривает сеть как единое целое, а не как разрозненный набор отдельных устройств.

Хорошая система управления наблюдает за сетью и, обнаружив проблему, активизирует определенное действие, исправляет ситуацию и уведомляет администратора о том, что произошло и какие шаги предприняты. Одновременно с этим система управления должна накапливать данные, на основании которых можно планировать развитие сети. Наконец, система управления должна быть независима от производителя и обладать удобным интерфейсом, позволяющим выполнять все действия с одной консоли.

Решая тактические задачи, администраторы и технический персонал сталкиваются с ежедневными проблемами обеспечения работоспособности сети. Эти задачи требуют быстрого решения, обслуживающий сеть персонал должен оперативно реагировать на сообщения о неисправностях, поступающих от пользователей или автоматических средств управления сетью. Постепенно становятся заметны более общие проблемы производительности, конфигурирования сети, обработки сбоев и безопасности данных, требующие стратегического подхода, то есть *планирования* сети. Планирование, кроме этого, включает прогноз изменений требований пользователей к сети, вопросы применения новых приложений, новых сетевых технологий и т. п.

Полезность системы управления особенно ярко проявляется в больших сетях: корпоративных или публичных глобальных.

В настоящее время в области систем управления сетями много нерешенных проблем. Большинство существующих средств вовсе не управляют сетью, а всего лишь осуществляют *наблюдение* за ее работой. Они следят за сетью, но не выполняют активных действий, если с сетью что-то произошло или может произойти. Мало масштабируемых систем, способных обслуживать как сети масштаба отдела, так и сети масштаба предприятия.

2.7. Совместимость

Совместимость, или *интегрируемость,* означает, что сеть способна включать в себя самое разнообразное программное и аппаратное обеспечение, то есть в ней могут сосуществовать различные операционные системы, поддерживающие разные стеки коммуникационных протоколов, а также аппаратные средства и приложения от разных производителей. Сеть, состоящая из разнотипных элементов, называется *неоднородной*, или *гетерогенной*, а если гетерогенная сеть работает без проблем, то она является интегрированной. Основной путь построения интегрированных сетей — использование модулей, выполненных в соответствии с открытыми стандартами и спецификациями.

2.8. Качество обслуживания

В общем случае *качество обслуживания* (Quality of Service, QoS) определяет вероятностные оценки выполнения тех или иных требований, предъявляемых к сети приложениями или пользователями. Например, при передаче голосового трафика через сеть под качеством обслуживания чаще всего понимают гарантии того, что голосовые пакеты будут доставляться сетью с задержкой не более *N* мс, при этом вариация задержки не превысит *M* мс, и эти характеристики станут выдерживаться сетью с вероятностью 0,95 на определенном временном интервале. Файловому сервису нужны гарантии средней полосы пропускания и расширения ее на небольших интервалах времени до некоторого максимального уровня для быстрой передачи пульсаций. В идеале сеть должна гарантировать особые параметры качества обслуживания, сформулированные для каждого отдельного приложения. Однако по понятным причинам разрабатываемые и уже существующие механизмы QoS ограничиваются решением более простой задачи — гарантированием неких усредненных требований, заданных для основных типов приложений.

Чаще всего параметры, фигурирующие в разнообразных определениях качества обслуживания, регламентируют следующие показатели работы сети:

- скорость передачи данных;
- задержки передачи пакетов;
- уровень потерь и искажений пакетов.

Качество обслуживания гарантируется для некоторого потока данных. Напомним, что поток данных — это последовательность пакетов, имеющих некоторые общие признаки, например адрес узла-источника, информация, идентифицирующая тип приложения (номер порта TCP/UDP) и т. п.

Механизмы поддержки качества обслуживания сами по себе не создают пропускной способности. Сеть не может дать больше того, что имеет. Так что фактическая пропускная

способность каналов связи и транзитного коммуникационного оборудования — это ресурсы сети, являющиеся отправной точкой для работы механизмов QoS. Механизмы QoS только управляют распределением имеющейся пропускной способности в соответствии с требованиями приложений и настройками сети. Самый очевидный способ перераспределения пропускной способности сети состоит в управлении очередями пакетов.

Поскольку данные, которыми обмениваются два конечных узла, проходят через некоторое количество промежуточных сетевых устройств, таких как концентраторы, коммутаторы и маршрутизаторы, то поддержка QoS требует взаимодействия всех сетевых элементов на пути трафика, то есть «из конца в конец» («end-to-end», e2e). Любые гарантии QoS настолько хороши, насколько их обеспечивает наиболее «слабый» элемент в цепочке между отправителем и получателем. Поэтому нужно хорошо понимать, что поддержка QoS только в одном сетевом устройстве, пусть даже и магистральном, может весьма незначительно улучшить качество обслуживания или же совсем не повлиять на параметры QoS.

Реализация в компьютерных сетях механизмов поддержки QoS является сравнительно новой тенденцией. Долгое время компьютерные сети существовали без таких механизмов, и это объясняется в основном двумя причинами. Во-первых, большинство приложений, выполняемых в сети, были нетребовательными. Для таких приложений задержки пакетов или отклонения средней пропускной способности в достаточно широком диапазоне не приводили к значительной потере функциональности. Примерами нетребовательных приложений являются наиболее распространенные в сетях 80-х годов приложения электронной почты или удаленного копирования файлов.

Во-вторых, сама пропускная способность 10-мегабитных сетей Ethernet во многих случаях не являлась дефицитом. Так, разделяемый сегмент Ethernet, к которому было подключено 10-20 компьютеров, изредка копирующих небольшие текстовые файлы, не превышающие несколько сотен килобайт, позволял трафику каждой пары взаимодействующих компьютеров пересекать сеть так быстро, как это требовалось породившим этот трафик приложениям.

В результате большинство сетей работало с тем качеством транспортного обслуживания, которое обеспечивало потребности приложений. Правда, никаких гарантий относительно нахождения задержек пакетов или пропускной способности, с которой пакеты передаются между узлами, в определенных пределах эти сети не давали. Более того, при временных перегрузках сети, когда значительная часть компьютеров одновременно начинала передавать данные с максимальной скоростью, задержки и пропускная способность

становились такими, что работа приложений давала сбой — шла слишком медленно, с прерываниями сеансов и т. п.

Транспортный сервис, который предоставляли такие сети, получил название *best effort,* то есть сервис «с максимальными усилиями». Сеть старается обработать поступающий трафик как можно быстрее, но при этом никаких гарантий относительно результата своих усилий не дает. Примерами являются большинство популярных технологий, разработанных в 80-е годы: Ethernet, Token Ring, IP, X.25. Сервис «с максимальными усилиями» основан на некотором справедливом алгоритме обработки очередей, возникающих при перегрузках сети, когда в течение некоторого времени скорость поступления пакетов в сеть превышает скорость продвижения этих пакетов. В простейшем случае алгоритм обработки очереди рассматривает пакеты всех потоков как равноправные и продвигает их в порядке поступления *(First Input First Output, FIFO).* В том случае, когда очередь становится слишком большой (не умещается в буфере), проблема решается простым отбрасыванием вновь поступающих пакетов.

Очевидно, что сервис «с максимальными усилиями» обеспечивает приемлемое качество обслуживания только в тех случаях, когда производительность сети намного превышает средние потребности, то есть является избыточной. В такой сети пропускная способность достаточна даже для поддержания трафика пиковых периодов нагрузки. Также очевидно, что такое решение не экономично — по крайней мере, по отношению к пропускным способностям сегодняшних технологий и инфраструктур, особенно для глобальных сетей. Так как пиковые нагрузки и области, где они возникают, трудно предсказать, то такой путь не дает долговременного решения.

Тем не менее построение сетей с избыточной пропускной способностью, будучи самым простым способом обеспечения нужного уровня качества обслуживания, иногда применяется на практике. Например, некоторые поставщики сетевых услуг TCP/IP предоставляют гарантию качественного обслуживания, постоянно поддерживая определенный уровень превышения пропускной способности своих магистралей над потребностями своих клиентов.

В условиях, когда многие механизмы поддержания качества обслуживания только разрабатываются, использование для этих целей избыточной пропускной способности часто оказывается единственно возможным, хотя и временным решением.