МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДЕРЖАВНИЙ УНІВЕРСИТЕТ ТЕЛЕКОМУНІКАЦІЙ
Навчально-науковий інститут Інформаційних технологій
(назва інституту (факультету))

Комп'ютерних наук
(назва кафедри)

ПЛАН КОНСПЕКТ ЛЕКЦІЙ

з дисципліни «Конвергентна мережна інфраструктура»

за спеціальністю 124 Системний аналіз
(шифр та повна назва напряму (спеціальності))

Спеціалізації

Укладач(і): к.т.н. Сєрих С.О.
(науковий ступінь, вчене звання, П.І.Б. викладача)

Конспект лекцій розглянутий та схвалений на засіданні кафедри Комп'ютерних наук
(повна назва кафедри)

Протокол № 8 від «11» лютого 2019 року
Завідувач кафедри Вишнівський В. В.

Модуль 1 Моделювання та проектування високошвидкісних мереж, впровадження мережевих рішень конвергентної мережевої інфраструктури
Тема 2. Моделювання та проектування високошвидкісних мереж

Лекція № 4

Тема лекції: Модель якості роботи мережі

1. Складові якості роботи мережі. Матрична модель якості мереж.
2. Первинні та вторинні параметри якості.
3. Еталонні з'єднання та можливі процеси подій у мережах.

Виконати самостійне завдання № 4.

1. Вивчити питання лекції.
2. Визначити чисельне значення первинних параметрів на кожному рівні ієрархії мереж доступу за завданням лабораторного заняття № 4.

Література:

1. Сосновский О.А Телекоммуникационные системы и компьютерные сети. – Минск: БГЭУ, 2007.-176с.
2. ВоробієнкоП.П. Телекомунікаційні та інформаційні мережі: підручник/П.П. Воробієнко, Л.А. Нікітюк, П.І. Резніченко. К.: САММІТ-Книга, 2010. – 708с.
3. Гніденко М.П., Вишнівський В.В., Сєрих С.О., Зінченко О.В., Прокопов С.В. Конвергентна мережна інфраструктура. – Навчальний посібник. – Київ: ДУТ, 2019. – 179 с.
4. Соколов В. Ю. Інформаційні системи і технології : Навч. посіб. К.: -ДУІКТ, 2010. - 138 с.
5. Олифер Виктор, Олифер Наталия. Компьютерные сети. Принципы, технологии, протоколы. (Учебник для вузов). — ISBN 978-5-496-01967-5. 5-е изд. — СПб.: Питер, 2016. — 992 с.

Вступ

Висока вартість протяжних каналів передачі даних і складність просто підвищення швидкості передачі даних за рахунок прокладки додаткових волоконно-оптичних жил обумовлює надзвичайно економне відношення до пропускної спроможності каналу в глобальних мережах. Для нормальної роботи додатків в таких умовах потрібне застосування методів забезпечення якості обслуговування (Quality of Service, QoS). Іншим могутнім стимулом втілювання механізмів QoS є прагнення до передачі по пакетним мережам всіх типів інформацій, в тому числі і мультимедійної - голосу, зображення, відеоданих.

Трафік таких застосувань відрізняється чутливістю до затриманням - одного з параметрів, який регулюється за допомогою механізмів QoS.

Тому в більшості технологій, спеціально розроблених для глобальних мереж передачі даних - Frame Relay, ATM, - механізми QoS являються вбудованими.

1. Складові якості роботи мережі та якості їх обслуговування

1.1 Основні поняття якості обслуговування

Основною рушійною силою розвитку мережі є додатки. У відповідь на постійні вимоги додатків, що ростуть, до пропускної спроможності мережі з'являються нові високошвидкісні технології. Перенесення в комп'ютерні мережі нового вигляду трафіку, наприклад IP-телефонії., аудіо і відео віщання, привів до появи нових вимог, пов'язаних із забезпеченням низького рівня затримок пакетів, підтримкою групової доставки пакетів і т.д.

Просте підвищення пропускної спроможності мережі вже більше не є гарантією того, що різноманітні застосування, що працюють в мережі, отримають те обслуговування, яке їм потрібне. Екстенсивний шлях розвитку, коли канали і комунікаційне устаткування мережі замінюється іншим, на порядок продуктивнішим, дає швидкий бажаний результат - додаткам зазвичай дістається необхідна для їх якісної роботи частка пропускної здібності, при цьому затримки із-за очікування пакетів в чергах (невід'ємна особливість мереж цього типу) стають невідчутними. Проте з перебігом часу кількість користувачів мережі росте, змінюються самі застосування – як правило, вони стають більш ресурсоємними, і мережа починає випробовувати перенавантаження.

Потрібні нові механізми якості обслуговування, що враховують все багато різноманіття вимог, що пред'являються додатками до мережі, і що надають кожному застосуванню той рівень якості обслуговування, який йому необхідний - чи то через об'єктивні потреби додатку, чи то відповідно зі згодою про якість обслуговування, заключними між клієнтом і оператором мережі. Такі механізми і працюють у складі служби QoS, що є свого роду «розпорядником» пропускної спроможності каналів і виробником комунікаційних пристроїв, контролюючи і регулюючи споживання ресурсів мережі трафіком окремих користувачів і їх груп.

1.2 Типи QoS

Поняття якості обслуговування в пакетних мережах є суто статистичним. В умовах, коли пакети передаються кінцевими вузлами в мережу у випадкові моменти часу, черги в комунікаційних пристроях теж представляють собою випадкові процеси, що приводить до того, що миттєва швидкість потоку даних і затримки пакетів також носять випадковий характер. Тому всі параметри, якими вимірюється якість обслуговування в пакетних мережах, є статистичними. Як правило, це середнє значення (математичне очікування) і варіації (дисперсії) швидкості інформаційного потоку і затримок пакетів. Немає сенсу говорити про затримку окремого пакету або про швидкість потоку на дуже маленькому проміжку часі, сумірному з часом передачі одного пакету. Які-небудь числові оцінки якості обслуговування можуть бути на практиці зміряні шляхом усереднювання відповідних величин на якому-небудь заздалегідь обумовленому проміжку часу.

Типи QOS розрізняються по ступеню «строгості», тобто по тому, наскільки твердо сервіс QOS може гарантувати забезпечення певних значень характеристик QoS- пропускної спроможності, затримок, варіацій затримок, рівня втрат пакетів і т.п.

Можна виділити три типи служб QoS.

1 Сервіс з максимальними зусиллями, який також можна назвати відсутністю

QoS, забезпечує взаємодію кінцевих вузлів без яких би не було б гарантій. Типові представники таких послуг - класичні мережі Ethernet або IP, які не роблять ніяких відмінностей між пакетами окремих користувачів і додатків і обслуговують ці пакети на основі принципу FIFO (першим прийшов - першим обслужений).

2 Сервіс з перевагою (званий також «м'яким» сервісом QoS) - деякі типи трафіку обслуговуються краще, ніж інші. Мається на увазі швидша обробка, в середньому більше пропускної спроможності і менше втрат даних. Це статистична перевага, а не чисельно виражені гарантії. Точні значення параметрів QoS, які отримають додаток в результаті роботи служби QoS цього типу, невідомі і залежать від характеристик пропонованого мережі трафіку. Наприклад, якщо високо пріоритетний трафік пропонує мережі в даний момент часу низьку інтенсивність своїх пакетів, то низько пріоритетний трафік може в цей час отримувати вельми якісне обслуговування - значну пропускну спроможність і низькі затримки. Проте при зміні ситуації, коли високо пріоритетний трафік починає посилати в мережу свої пакети з високою інтенсивністю, низько пріоритетний трафік може взагалі якийсь час не обслуговуватися.

3 Гарантований сервіс (званий також «жорстким», або «істинним», сервісом QoS) дає статистичні численні гарантії різним потокам трафіку. Зазвичай такий вид QoS заснований на попередньому резервуванні мережених ресурсів для кожного з потоків, що отримав гарантії обслуговування. Трафік, якому виділили ресурси, гарантовано має при проходженні через мережу ті параметри пропускної спроможності або затримок, які визначені для нього в числовому вигляді (звичайно, якщо джерела цього трафіку не порушують обумовлених для них умов і генерують не більше пакетів, чим передбачалося). Служби такого типу здатні, наприклад, гарантувати додатку обумовлену пропускну спроможність, не зменшується ні за яких обставин, якою б переобтяженою мережа не ставала. Слід зазначити, що гарантії носять статистичний характер, тобто можна гарантувати деяке числові значення якого-небудь параметра тільки з деякою вірогідністю, хай дуже високою, не рівною 1. Наприклад, з вірогідністю 0,999 можна стверджувати, що затримка пакету не перевищить 100 мс, отже, один пакет з 1000 може затриматися в мережі і на більший час. Другою складовою такого режиму роботи служби QoS є вхідний контроль потоків, яким дані гарантії. Дійсно, дотримувати дані гарантії можна тільки в тому випадку, якщо інтенсивності вхідних в мережу потоків не перевищують граничних нижче:

зазначених значень. Інакше потік споживатиме більше ресурсів, чим йому виділялося, а значить, іншим потокам дістанеться менше, ніж було заплановано, і гарантії для них дотримані не будуть.

Ці три підходи до роботи служб QoS не виключають, а доповнюють один одного.

В результаті їх комбінування дозволяє врахувати різноманітні вимоги додатків і різні умови роботи мережі.

Деяким застосуванням достатньо обслуговування з максимальними зусиллями, тоді як іншим обов'язкове жорстке дотримання гарантій за рахунок попереднього резервування.

Трафік додатків, для яких параметри якості обслуговування не абсолютно важливі (тобто додаток, а точніше, його користувач може змиритися з деяким погіршенням реактивності мережі), обробляється службою другого типу, яка диференційовано розподіляє ті, що залишилися після резервування ресурси між декількома

пріоритетними класами трафіку. Якщо одному з диференційованих класів потрібна на деякий час велика пропускна здатність, то його потреби задовольняються за рахунок трафіку диференційованих класів нижчого пріоритету або трафіку, обробленого службою обслуговування з максимальними зусиллями.

Трафік додатків, для яких обов'язковою умовою є деякий гарантований рівень пропускної спроможності і/або затримок (наприклад трафік відео конференції або трафік, що поступає від вимірювальної системи реального часу), обслуговується службою QoS третього типу, представляючи гарантований сервіс.

1.3 Згода про рівень обслуговування

Якість обслуговування може розглядатися з двох позицій. По-перше, з точки зору споживача транспортних послуг мережі, коли характеристики якостей; обслуговування виступають як деякі бажані умови, що забезпечують нормальну роботу додатків. В цьому випадку мова йде про потрібний якостей(обслуговування. По-друге, з погляду постачальника послуг" (або адміністратора корпоративної мережі), для якого якість обслуговування - це фактичні характеристики роботи мережі, спостережені в результаті її моніторингу або в результаті аналітичного дослідження або імітаційного моделювання мережі.

Природною основою нормальної співпраці постачальника і споживача є договір, який в даному випадку називається угодою про рівень обслуговування (Service Level Agreement, SLA). У цьому договорі постачальник послуг і його клієнт визначають наступні позиції.

1 Параметри якості обслуговування трафіку, які цікавлять споживача і які згоден підтримувати постачальник (за показаннями середня пропускна спроможність, максимальні затримки і варіації затримок, максимальна інтенсивність втрат даних, коефіцієнт готовності сервісу, максимальний час відновлення сервісу після відмови і т. д.).

2 Методи вимірювання параметрів якості обслуговування.

3 Визначення плати за обслуговування. Система оплати може бути достатньо складною, особливо якщо угода передбачає декілька вирівняний якості обслуговування, які оплачуються за різними тарифами.

4 Санкції за порушення зобов'язань постачальника послуг із-за забезпечення належної якості обслуговування, а також за відхилення параметрів трафіка користувача від обумовлених значень. Ці санкції можуть виражатися у вигляді штрафів або в іншій формі, наприклад у формі надання сервісу протягом деякого часу безкоштовно або за зниженим тарифом.

5 Як і любий договір, угода SLA по взаємної згоди постачальника і клієнта може включати велику кількість різних додаткових статей. Наприклад, статтю, що обумовлює умови переходу до більш якісному обслуговуванню або обслуговуванню з різним рівнем якості в залежності від дня тижня або часу доби.

6 Угоду може включати також правила кондиціонування трафіку користувача, тобто обробки трафіку, який виходить за обговорені межі, наприклад трафіку з більшою середньою інтенсивністю на значному проміжку часу. Ці правила можуть включати формальні способи розпізнавання різних потоків призначеного для користувача трафіку (що потрібно у тому випадку, коли угода обумовлює диференційоване обслуговування певних потоків). Також правила кондиціонування можуть визначати умови відкидання або маркіровки пакетів-порушників (помічені

пакети відкидатимуться мережею не завжди, а тільки у тому випадку, коли мережні пристрої випробовують перевантаження)! В деяких випадках мережне устаткування виконує згладжування пульсацій трафіку для поліпшення якості роботи додатку (наприклад, що працює з голосом), а також для зменшення затримок в транзитних вузлах мережі, оскільки більш рівномірно пакети, що поступають, зменшують перевантаження мережних пристроїв.

Багато постачальників послуг пропонують своїм клієнтам типові контракти SLA.

В них не тільки визначений перелік характеристик якості обслуговування, але навіть і їх конкретні числові значення, наприклад: «затримка пакетів, усереднена цінна за місяць, не перевищуватиме 100 мс при передачі між будь-якими двома Вузлами мережі». Типові контракти полегшують життя постачальникам послуг, оскільки для їх реалізації можна обійтися без засобів гарантованої підтримки якості обслуговування. Потрібно тільки підтримувати приблизно постійний рівень запасу пропускної спроможності і пропонувати в SLA ті значення параметрів QoS, які демонструє працююча мережа.

Згода SLA полягає не тільки між постачальником послугами публічних мереж і корпоративними клієнтами. Достатньо популярним останнім часом стало укладення подібних контрактів між інформаційним відділом підприємства - постачальником транспортних послуг, і споживачами – функціональними відділами підприємства.

1.4 Вимоги до якості обслуговування додатків різних типів.

Сучасна тенденція конвергенції мереж різних типів, привела до необхідності перенесення мережею всіх видів трафіку, а не тільки традиційного для комп'ютерних мереж трафіку додатків доступу до файлів і електронної пошти.

Характеристики QoS особливо важливі у тому випадку, коли мережа передає одночасно трафік різного типу, наприклад трафік веб-додатків і голосовий трафік. Це пов’язано з тим, що різні типи трафіку пред'являють різні вимоги характеристикам QoS. Добитися одночасного дотримання всіх характеристик QoS для всіх видів трафіку дуже складно. Тому зазвичай використовують наступний підхід: класифікують всі види трафіку, що існують в мережі, відносячи кожен з них до одного з поширених типових видів трафіку, а потім добиваються одночасного виконання певної підмножини з набору вимог для цих типів трафіку.

До теперішнього часу виконана велика робота по класифікації трафіку додатків.

Як основні критерії класифікації були прийняті три характеристики трафіку:

1 відносна передбаченість швидкості передачі даних;

2 чутливість трафіку до затримок пакетів;

3 чутливість трафіку до втрат і спотворень пакетів.

1.4.1 Передбаченість швидкості передачі даних.

Відносно передбаченості швидкості передачі даних трафік додатків ділиться на два великі класи:

1 потоковий трафік (stream); 2 пульсуючий трафік (burst).

Додатки з потоковим трафіком породжують рівномірний потік даних, який

поступає в мережу з постійною бітовою швидкістю (Constant Bit Rate, CBR). При використанні методу комутації пакетів трафік таких додатків є послідовність пакетів

однакового розміру (рівного В біт), наступних один за одним через один і той же інтервал часу Т (рис. 8.1).

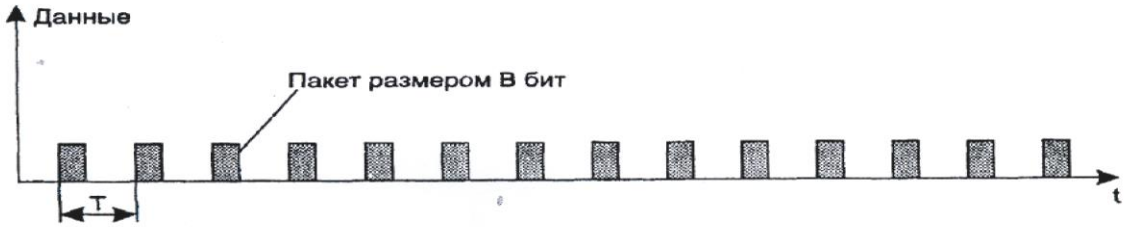


Рис. 8.1. Потоковый трафик

Постійна швидкість потокового трафіку (CBR) може бути обчислена шляхом
усереднювання на одному періоді: CBR - B/T бит/с.

У загальному випадку постійна швидкість потокового трафіку менше
номінальної максимальної бітової швидкості протоколу, за допомогою якого
передаються дані, оскільки між пакетами існують паузи. Так максимальна швидкість передачі даних за допомогою протоколу Ethernet – 10 Base-T складає 9,76 Мбит/с (випадок кадрів максимальної довжини), що менше за номінальну швидкість цього протоколу, рівну 10 Мбит/с

Додатки з пульсуючим трафіком відрізняються високим ступенем
непередбаченості, коли періоди мовчання змінюються пульсацією, протягом якої пакети «щільно» слідують один за одним. В результаті трафік характеризується змінною бітовою швидкістю (Variable Bit Rate, VBR), що ілюструє рис. 4.2.

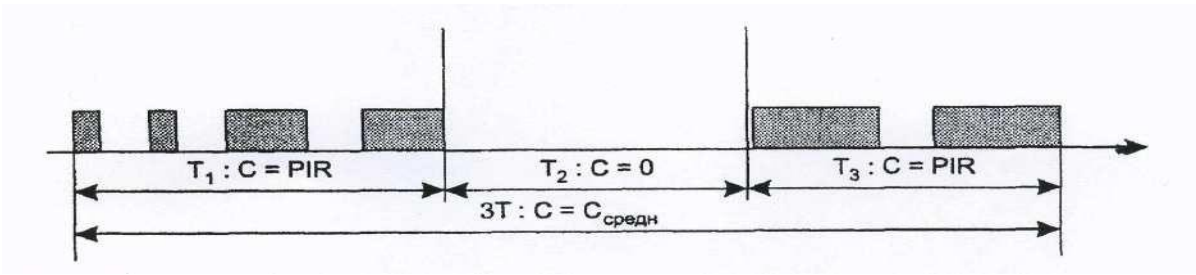


Рис. 4.2. Пульсуючий трафік

На малюнку показано три періоди вимірювань Т1, Т2 і Т3. Для спрощення розрахунків прийняте, що пікові швидкості на першому і третьому періодах рівні між собою і рівні PIR, а всі три періоди мають однакову тривалість Т. Враховуючи це, можна обчислити величину пульсації, яка рівна кількості бітів, переданих на періоді пульсації:

$$B = PIR \times T.$$

Таким чином, величина пульсації для періодів Т1і Т3 рівна В, а на періоді Т2 -
нулю.

Для приведеного прикладу можна підрахувати коефіцієнт пульсації. (Нагадаємо, що він рівний відношенню пікової швидкості на якому-небудь невеликому періоді

часу до середньої швидкості трафіку, зміряної на тривалому періоді часу.) Оскільки пікова швидкість на періоді T1 (або T3) рівна B/T, а середня швидкість на сумарному періоді T1 + T2 + T3 рівна 2B/3T, то коефіцієнт пульсації рівний 3/2.

Практично будь-який трафік, навіть трафік потокових застосувань, має не нульовий коефіцієнт пульсації. Просто значення коефіцієнтів пульсації у потокового і пульсуючого трафіків істотно розрізняються. У додатків з пульсуючим трафіком він зазвичай знаходиться в межах від 2:1 до 100:1, а у потокових застосувань близький до 1:1. У локальних мережах коефіцієнт пульсації зазвичай вище, ніж в глобальних, оскільки на магістралях глобальних мереж трафік є сумою трафіків багатьох джерел, що по закону великих чисел приводить до згладжування результуючого трафіку.

1.4.2 Чутливість трафіку до затримок пакетів.

Ще один критерій класифікації додатків за типом трафіку - чутливості до затримок пакетів і їх варіацій. Далі перераховані основні типи додатків в порядку підвищення чутливості до затримок пакетів.

1 Асинхронні застосування. Практично немає обмежень на якийсь час затримки
(еластичний трафік). Приклад такого застосування - електронна пошта.

2 Інтерактивні застосування. Затримки можуть бути відмічені користувачами, але вони не позначаються негативно на функціональності додатків. Приклад такого - текстовий редактор, що працює з видаленим файлом.

3 Ізохронні застосування. Є поріг чутливості до варіацій затримок, при перевищенні якого різко знижується функціональність застосування. Приклад - передача голосу, коли при перевищенні порогу варіації затримок в 100-150 мс різко знижується якість відтворного голосу, а при ≥450 втрачається контакт
Надчутливі до затримок застосування. Затримка доставки даних зводить функціональність додатку до нуля. Приклад - додатки, які управляють технічним об'єктом в реальному часі. При запізнюванні управляючого сигналу на об'єкті може відбутися аварія.

Взагалі кажучи, інтерактивність додатку завжди підвищує його чутливість до затримок. Наприклад, широкомовна розсилка аудіо інформації може витримувати значні затримки передачі пакетів (залишаючись чутливим до варіацій затримок), а інтерактивний телефонний або телевізійна розмова їх не терпить, що добре помітно при трансляції розмови через супутник. Тривалі паузи в розмові вводять співбесідників в оману, часто вони втрачають терпіння і починають чергову фразу одночасно.

Разом з приведеною вище класифікацією, тонко диференціюючою чутливістю додатків до затримок і їх варіацій, існує і більш грубіше ділення додатків за цією ж ознакою на два класи - асинхронні і синхронні. До асинхронних відносять ті застосування, які нечутливі до затримок передачі даних в дуже широкому діапазоні, аж до декількох секунд, а решта всіх застосувань, на функціональність яких затримки впливають істотно, відносять до синхронних застосувань.

Інтерактивні застосування можуть відноситися як до асинхронних (наприклад,
текстовий редактор), так і до синхронних (наприклад, відео конференція).

1.4.3 Чутливість трафіку до втрат і спотворень пакетів

Нарешті, останнім критерієм класифікації додатків є їх чутливість до втрат пакетів. Тут зазвичай ділять додатки на дві групи.

2 Додатки, чутливі до втрати даних. Практично всі додатки, що передають алфавітно-цифрові дані (до яких відносяться текстові документи, коди програм, числові масиви і т. п.), володіють високою чутливістю до втрати окремих, навіть невеликих, фрагментів даних. Такі втрати часто ведуть до повного знецінення інших, успішно прийнятої інформації. Наприклад, відсутність хоч би одного байта в коді програми робить її абсолютно непрацездатною. Всі традиційні мережеві застосування (файловий сервіс, сервіс баз даних, електронна пошта і т. д.) відносяться до цього типу додатків.

2 Додатки, стійкі до втрати даних. До цього типу відносяться багато застосувань, що передають трафік з інформацією про інерційні фізичні процеси. Стійкість до втрат пояснюється тим, що невелика кількість відсутніх даних можна визначити на основі прийнятих. Так, при втраті одного пакету, що несе декілька послідовних вимірів голосу, відсутні виміри при відтворенні голосу можуть бути замінені апроксимацією на основі сусідніх значень. До такого типу відноситься велика частина додатків, що працюють з мультимедійним трафіком (аудіо і відео додатки). Проте стійкість до втрат має свої межі, тому відсоток втрачених пакетів не може бути великим (наприклад, не більше 1 %), Можна відзначити також, що не будь- який мультимедійний трафік такий стійкий до втрат даних, наприклад, скомпресований голос і відео зображення дуже чутливі до втрат, тому відносяться до першого типу додатків.

2.4.1 Класи додатків

Взагалі кажучи, між значеннями трьох характеристик якості обслуговування (відносна Передбаченість швидкості передачі даних; чутливість трафіку до затримок пакетів; чутливість трафіку до втрат і спотворенням пакетів) немає строгого взаємозв'язку. Тобто додаток з рівномірним потоком може бути як асинхронним, так і синхронним, а, наприклад, синхронне застосування може бути як чутливим, так і нечутливим до втрат пакетів. Проте практика показує, що зі всього різноманіття можливих поєднань значень цих трьох характеристик є декілька таких, які охоплюють велику частину використовуваних сьогодні застосувань.

Наприклад, наступне поєднання характеристик додатку «породжуваний трафік - рівномірний потік, додаток ізохронний, стійкий до втрат» відповідає таким популярним застосуванням, як IP-телефонія, підтримка відео конференцій, аудіо віщання через Інтернет. Існують і такі поєднання характеристик, для яких важко привести приклад додатку, наприклад: «породжуваною трафік - рівномірний потік, додаток асинхронний, чуттєве до втрат».

Стійких поєднань характеристик, що описують певний клас додатків, існує не так вже багато. Так, при стандартизації технології ATM, яка спочатку розроблялася для підтримки різних типів трафіку, було визначено 4 класи додатків: A, B, C і D. Для кожного класу рекомендується використовувати власний набір характеристик QoS, Крім того, для всіх застосувань, не включених ні в один з цих класів, був визначений клас X, в якому поєднання характеристик додатку може бути довільним. Класифікація ATM є на сьогодні найбільш детальною і загальною,

вона не вимагає від нас знання технологій, використовуваних для передачі цих типів
трафіка, тому приведемо її тут (табл. 8.1).

Таблиця 8.1 . Класи трафіку

Клас трафіку	Характеристики
A	Постійна бітова швидкість, чутливість до затримок, передача зі встановленням з'єднання (наприклад, голосовий трафік, трафік телевізійного зображення). Параметри QoS: пікова швидкість передачі даних, затримка, джітер
B	Змінна бітова швидкість, чутливість до затримок, передача зі встановленням з'єднання (наприклад, компресований голос, ком пресування відео зображення). Параметри QOS: пікова швидкість передачі даних, пульсація, середня швидкість передачі даних, затримка, джітер
C	Змінна бітова швидкість, еластичність, передача зі встановленням з'єднання (наприклад, трафік комп'ютерних мереж, в яких кінцеві вузли працюють по протоколах зі встановленням з'єднань, - frame relay, X.25, TCP). Параметри QoS: пікова швидкість передачі даних, пульсація, середня швидкість передачі даних
D	Змінна бітова швидкість, еластичність, передача без встановлення з'єднання (наприклад, трафік комп'ютерних мереж, в яких кінцеві вузли працюють по протоколах без встановлення з'єднань, - IP/UDP, Ethernet). Параметри QoS не визначені
X	Тип трафіку і його параметри визначаються користувачем

Приведена класифікація додатків лежить в основі типових вимог до параметрів і механізмів забезпечення якості обслуговування в сучасних мережах.Первинні та вторинні параметри якості.

2.1 Параметри якості

Трьом критеріям класифікації додатків (передбаченість швидкості передачі даних, чутливість до затримок і чутливість до втрат і спотворення) відповідають три групи параметрів, використовуваних при визначенні і завданні необхідної якості обслуговування.

1 Параметри пропускної спроможності. До таких параметрів відносяться середня, максимальна (пікова) і мінімальна швидкості передачі даних.

2 Параметри затримок. Використовується середня і максимальна величини затримок, а також середнє і максимальне значення варіацій затримок, тобто відхилень між пакетних інтервалів в трафіку, що прибуває, в порівнянні з витікаючи.

3 Параметри надійності передачі. Використовується відсоток втрачених пакетів, а також відсоток спотворених пакетів.

При визначенні всіх цих параметрів важливо, на якому періоді і вимірюється даний параметр. Чим менше цей період, тим більше жорсткими є відповідно вимоги якості обслуговування і тим важче для мережі їх витримати. Тому постачальники послуг IP-мереж, які випробовують складнощі із забезпеченням QoS, віддають перевагу говорити в угодах SLA про середньомісячної характеристики., тоді як постачальники послуг мереж frame relay і ATM, що мають в своєму розпорядженні могутні засоби QoS, здатні гарантувати параметри, усереднені на періоді в декілька секунд.

Параметри якості обслуговування можуть бути віднесені до пропонованого трафіку, що породжується додатками користувача. Ці ж параметри можуть характеризувати можливості мережі по обслуговуванню цього трафіку. Хай, наприклад, у користувача є додаток, який породжує рівномірний потік з постійною швидкістю N. При укладенні угоди SLA з поставником послуг користувач бере на себе зобов'язання, що пропонований трафік 1риложения не перевищуватиме

максимальну швидкість N. Постачальник, в свою чергу, гарантує з боку мережі, що мінімальна величина пропускної спроможності, що надається цьому застосуванню, буде не менша N. Це необхідно для забезпечення прийнятної якості обслуговування трафіку даного додатку.

Для додатків з пульсуючим трафіком якість обслуговування краще всього характеризується середньою швидкістю і максимальною швидкістю, яка потрібна в період пульсації. Зазвичай при цьому обмовляється або максимальний час пульсації, протягом якого додаток передає дані з максимальною швидкістю, або максимальний об'єм даних, який можна передати у вигляді пульсації. Часто використовується також варіант із завданням максимальної і мінімальної меж швидкості. В цьому випадку додатку гарантується пропускна здатність на рівні мінімальної межі, достатня для його задовільного функціонування, а саме застосування зобов'язується не направлять в мережу трафік з швидкістю, що перевищує максимальну межу.

Параметри якості обслуговування зазвичай обмовляються в угоді про рівень обслуговування SLA між користувачем мережі і постачальником послуг. Після укладення угоди користувач і постачальник послуг повинні належним чином набудувати свої програмні і апаратні засоби, щоб вони відпрацьовували обумовлені параметри. Деякі технології дозволяють автоматизувати процес взаємного узгодження параметрів QOS між технікою користувача і устаткуванням постачальника. Наприклад, в технології ATM при встановленні з'єднання ці параметри узгоджуються за допомогою процедури, званою трафік-контрактом.

Висновки

1 Залежно від строгості дотримання гарантій забезпечення визначених параметрів обслуговування - пропускної спроможності, затримок, варіацій затримок, рівня втрат пакетів і т.п. - розрізняють наступні типи QoS: обслуговування в міру можливості (best effort), обслуговування з надаванням переваги одного потоку (м'який сервіс QoS) і гарантоване обслуговування (жорсткий сервіс QoS).

2 Безліч додатків може бути розділена на класи в залежності від вимог, що пред'являються ними до якості обслуговування. У основу класифікації покладені три основні характеристики трафіку, що породжується додатками: відносна передбаченість швидкості передачі даних чутливість трафіку до затримок пакетів, чутливість трафіку до втрат і спотворень пакетів.

3 В угоді про рівень обслуговування (Service Level Agreement, SLA) поставник послуг і його клієнт визначають: параметри якості обслуговування трафіка, які цікавлять споживача і які згоден підтримувати постачальник, методи вимірювання цих параметрів, фінансові зобов'язання обох сторін, правила «кондиціонування» призначеного для користувача трафіку і ін.

4 При визначенні і завданні необхідної якості обслуговування в SLA використовуються: параметри пропускної спроможності (середня, максимальна і мінімальна швидкості передачі даних), параметри затримок (середня і максимальна величини затримок, а також середнє і максимальне значення варіацій затримок), параметри надійності передачі (відсоток втрачених пакетів, а також відсоток спотворених пакетів).

2.2 Служба QoS Модель служби QoS

Мережа - це розподілене середовище, що складається з великої кількості

пристроїв для підтримки різних технологій і протоколів. Тому достатньо складно примусити її дотримувати єдині вимоги по якісному обслуговуванню різних видів трафіку на всьому протязі складеного шляху від одного конечного вузла до іншого, тобто «з кінця в кілець» (end-to-end). Навіть забезпечення простої «зв'язності» мережі «з кінця в кінець» для складеної неоднорідної мережі є непростим завданням, яка представляє собою основу функцію класичних засобів стеків протоколів. Завдання ж просування пакетів в такій мережі із заданими параметрами якості обслуговування істотно складніше ніж перша. Особливо якщо в мережі існують численні потоки даних з куп але сумісними характеристиками, наприклад пульсуючий файловий графік і синхронний голосовий.

Для рішення поставлених задач в мережі необхідна служба QoS. Ця служба має розподілений характер так як її елементи повинні бути присутніми у всіх мережних пристроях, що просувають пакети: комутаторах, маршрутизаторів, серверах доступу. З іншого боку, роботу окремих мережних пристроїв та забезпеченню підтримки QoS потрібно скоординувати, щоб якість обслуговування була однорідною уздовж всього шляху, по якому слідують пакети потоку. Тому служба QoS повинна включати також елементи централізованого управління, за допомогою яких адміністратор мережі може погоджено конфігурувати механізми QoS в окремих пристроях мережі.

Базова архітектура служби QoS включає елементи трьох основних типів, представлених на рис.8.3:

1 засоби QoS вузла, що виконують обробку трафіку, що поступає у вузол, відповідно до вимог якості обслуговування;

2 протоколи QoS-сигналізації для координації роботи мережних елементів по підтримці якості обслуговування «з кінця в кінець»;

3 централізовані функції політики, управління і обліку QoS, дозволяють адміністраторам мережі цілеспрямовано впливати на мережеві елементи для розділення ресурсів мережі між різними видами трафіку

2.3 Засоби QoS вузла

Засоби QoS вузла являються основним виконавчим механізмом служби QoS, оскільки саме вони безпосередньо впливають на процес просування пакетів між вхідними і вихідними інтерфейсами комутаторів і маршрутизаторів і, отже, визначають внесок даного пристрою в характеристики якості обслуговування мережі. Засоби QOS вузла можуть включати механізми двох типів:

1 механізми обслуговування черг;

2 механізми кондиціонування трафіку.

Механізми обслуговування черг є необхідним елементом будь-якого. пристрою, що працює за принципом комутації пакетів. Вони можуть підтримувати різні алгоритми обробки пакетів, які потрапили в чергу, від найпростіших типу FIFQ (першим прийшов - першим обслужений) дуже складних, підтримуючих обробку декількох класів потоків, наприклад алгоритмів пріоритетного, або зваженого, обслуговування. За умовчанням в мережних пристроях діє алгоритм черги FIFO, але він достатній тільки для реалізації обслуговування з максимальними зусиллями, а для підтримки «дійсних» сервізів QoS потрібні складніші механізми.

Механізми другого типу (кондиціонування трафіку) можуть реалізовуватися або не реалізовуватися в мережному вузлі, підтримуючому QoS. Річ у тому, що

забезпечення потрібної якості обслуговування завжди означає створення таких вимог, коли швидкість просування трафіку потоку узгоджується. із швидкістю надходження цього трафіку у вузол мережі. Черги виникають в ті періоди часу, коли швидкість надходження трафіку стає більше швидкості його пересування. Механізми обслуговування черг розраховані на роботу якраз в періоди перевантаження і потрібні для того, щоб потоки якомога менше чи страда7" від існування таких періодів. Затримки від очікування в чергах повинні укладатися в параметри потоку. Механізми кондиціонування трафіку вирішують задачу створення умов якісного обслуговування трафіку іншим способом - за рахунок зменшення швидкості надходження потоку :в даний вузол настільки, щоб вона завжди залишалася менше, ніж швидкість просування цього потоку.

Механізм кондиціонування трафіку зазвичай включає виконання декількох їх функцій.

1 Класифікація трафіку. Ця функція виділяє із загальної послідовності пакетів, що поступають в пристрій, пакети одного потоку, що має загальні вимоги до якості обслуговування. Класифікація може виконується на основі різних формальних ознак пакету - адрес джерела і призначення, ідентифікаторів додатків, значення пріоритету пакета, значення мітки потоку.

2 Профілізація трафіка на основі правил політики (policing). Для кожного вхідного потоку є відповідний йому набір параметрів QoS. Цей набір часто називають профілем трафіка Профілювання трафіку має на увазі перевірку відповідності кожного вхідного потоку параметрам його профілю. У разі порушення параметрів профілю (наприклад, перевищення тривалості пульсації або середньої швидкості) відбувається відкидання або маркіровка пакетів цього потоку. Відкидання деяких пакетів знижує інтенсивність потоку і приводить його параметри у відповідність з указаними в профілі. Маркіровка пакетів без відкидання потрібна для того, щоб пакети всі були обслужені даним вузлом (або подальшими по потоку), але з якістю обслуговування, відмінним від указаного в профілі (наприклад, зі збільшеним значенням затримки). Для перевірки відповідності вхідного трафіку заданому профілю механізм кондиціонування виконує вимірювання параметрів потоку. Для цього зазвичай використовується один з відомих алгоритмів, наприклад алгоритм «дірявого відра» (leaky bucket).

3Формування трафіку (shaping). Ця функція призначена для додання минулому профілізацію трафіку потрібної тимчасової «форми». В основному за допомогою даної функції прагнуть згладити пульсації трафіку, щоб вихід пакетів з пристрою був більш рівномірним, ніж вхід. Згладжування пульсацій зменшує черги в мережних пристроях, які оброблятимуть трафік далі по потоку. Його також доцільно використовувати для відновлення тимчасових співвідношень трафіку додатків, працюючих з рівномірними потоками, наприклад голосових застосувань.

Механізми кондиціонування трафіку можуть підтримуватися кожним вузлом мережі або реалізовуватися тільки в прикордонних пристроях. Останній варіант часто використовують постачальники послуг, кондиціонуючи трафік своїх клієнтів.

Мережі з комутацією пакетів були спочатку розроблені для передачі асинхронного трафіку, так що із затримками можна було миритися. Проте сьогодні, коли мережі передачі даних почали переносити різні типи трафіку, у тому числі і чутливого до затримок, питання забезпечення показників QoS вийшли на перше місце.

Для розуміння механізмів підтримки QoS необхідно досліджувати процес утворення черг в мережних пристроях і зрозуміти найбільш істотні чинники, що впливають на довжину черги.

Існує гілка прикладної математики, предметом якої є процеси утворення черг. Ця дисципліна так і називається - теорія черг. Ми не заглиблюватимемося в математичні основи цієї теорії, приведемо тільки деякі її виводи, істотні для проблеми QoS, що розглядається нами.

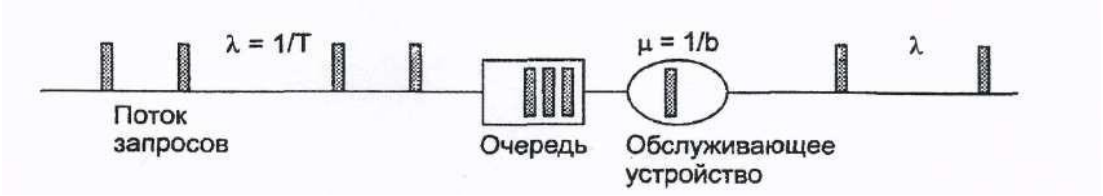


Рис. 8.4. Модель M/M/1

На рис. 8.4 показана найбільш проста модель теорії черг, відома під назвою M/M/1.

Основними елементами моделі є:

1 вхідний потік абстрактних заявок на обслуговування;

2 буфер;

3 обслуговуючий пристрій;

4 вихідний потік обслужених заявок.

Заявки поступають на вхід буфера у випадкові моменти часу. Якщо у момент надходження заявки буфер порожній і обслуговуючий пристрій вільний, то заявка відразу ж передається в цей пристрій для обслуговування. Обслуговування також триває випадковий час.

Якщо у момент надходження заявки буфер пустий, але обслуговуючий пристрій зайнятий обслуговуванням заявки, що раніше поступила, то заявка чекає його завершення в буфері. Як тільки обслуговуючий пристрій завершує обслуговування чергової заявки, вона передається на вихід, а прилад вибирає з буфера наступну заявку (якщо буфер не порожній). Що виходять з обслуговуючого пристрою заявки утворюють вихідний потік. Буфер вважається нескінченним, тобто заявки ніколи не втрачаються через те, що вичерпана ємкість буфера.

Якщо прибула заявка застає буфер не порожнім, то вона стає в чергу і чекає обслуговування. Заявки вибираються з черг порядку поступлення, тобто дотримується дисципліна обслуговування першим прийшов – першим обслужений (First-In, First-Out, FIFO).

Теорія черг дозволяє оцінити для цієї моделі середню довжину черги і середній час очікування заявки в черзі залежно від характеристик вхідного потоку і часу обслуговування.

Тут 1 означає, що моделюється один обслуговуючий пристрій, перша буква M позначає тип розподілу інтервалів надходження заявок (марковське розподілення друга - тип розподілу значень часу обслуговування (теж марковське).

Вважатимемо, що нам відоме, що середній час між проступанням заявок рівний T. Це означає, що інтенсивність надходження заявок, яка традиційно позначається в теорії черг символом X, рівна

$$\lambda = 1/T \text{ заявок в секунду.}$$

Випадковий процес надходження заявок описується в цій моделі функцією розподілу інтервалів між надходженнями заявок. Для спрощення отримання

компактних аналітичних результатів зазвичай вважають, що ці інтервали описуються так званим марковським розподілом (інша назва - пуассонівське), щільність якого показана на рис. 8.5 З малюнка видно, що вхідний потік є істотно пульсуючим, оскільки є не нульова вірогідність того, що інтервал між заявками буде дуже невеликим, близьким до нуля, а також того, що він буде дуже великим. Середнє відхилення інтервалів також рівне Т, так що стандартне відхилення рівне Т/Т = 1.

Також вважатимемо, що середній час обслуговування заявки рівний Ь. Це означає, що обслуговуючий прилад здатний просувати заявки на вихід з інтенсивністю 1/Ь = μ Знову ж таки для отримання аналітичного результату вважають, що час обслуговування - це випадкова величина з пуассонівською щільністю розподілу.

Ухвалення таких припущень дає простій результат для середнього часу очікування заявки в черзі, яке ми позначимо w:

w=p/ 1-p

Тут через p позначено відношення λ/μ

Параметр p називають коефіцієнтом використання (utilization) обслуговуючого приладу. Для будь-якого періоду часу цей показник рівний відношенню часу зайнятості обслуговуючого приладу до величини цього періоду.

Залежність середнього часу очікування заявки w від p ілюструє рис. 8.6. Як видно з поведінки кривої, параметр p грає ключову роль в утворенні черги. Якщо значення p близько до нуля, то і середній час очікування дуже близько до нуля. А це означає, що заявки майже ніколи не чекають обслуговування буфері (у момент їх приходу він виявляється порожнім), а відразу потрапляють в обслуговуючий пристрій. І навпаки, якщо p наближається до 1, то час очікування росте дуже швидко і нелінійно (і в межі рівно нескінченності). Така поведінка черги інтуїтивна зрозуміло, адже p - це відношення середньої інтенсивності вхідного потоку до середньої інтенсивності його обслуговування. Чим ближче середні значення інтервалів між пакетами до середнього часу обслуговування тим складніше обслуговуючому пристрою справлятися з навантаженням.

За допомогою моделі M/M/1 можна моделювати мережу з комутацією пакетів рис..8.7.

Хай вхідний потік заявок грає роль потоку пакетів, що поступають на вхід інтерфейсу комутатора, буфер моделі відповідає буферу вхідного інтерфейсу комутатора, а абстрактний обслуговуючий пристрій моделює процесор, оброблювальний пакети і що направляє їх на вихідний інтерфейс. Отже, середній час обслуговування заявки відповідає середньому часу просування пакету процесором комутатора з вхідного буфера у вихідний канал.

Зрозуміло, що приведена модель дуже спрощено описує процеси, які трапляються в мережі. Вона не враховує багатьох особливостей обробки пакетів, на¬пример кінцевого розміру буфера комутатора, ненульового часу проступання пакету в буфер і інших. Проте вона дуже корисна для розуміння основних чинників, що впливають на величину черги.

Мережні інженери добре знайомі з графіком, представленим на рис. 8.6. Вони інтерпретують цей графік як залежність затримок в мережі від її загрузки. Параметр p моделі відповідає коефіцієнту використання мереженого ресурсу, який бере участь в передачі трафіку, тобто інтерфейсу комутатора, процесора комутатора, каналу або середовища, що розділяється.

У приведеному графіку є і щось несподіване. Важко уявити, що обслуговуючий пристрій (мережений ресурс) практично перестає справлятися з своїми обов'язками, коли його коефіцієнт використання наближається до 1. Адже в цьому випадку навантаження не перевищує його можливостей, а тільки наближується до цієї межі. Інтуїтивно не дуже зрозуміла також причина існування черг при значеннях р в околицях 0,5. Інтенсивність обробки трафіку удвічі перевищує інтенсивність навантаження, а черги існують!

Такі парадоксальні на перший погляд результати характерні для систем, в яких протікають випадкові процеси. Оскільки А, і ц - це середні значення інтенсивносте й потоків на великих проміжках часу, то на невеликих проміжках часу вони можуть істотно відхилятися від цих значень. Черга створюється на тих проміжках, на яких інтенсивність надходження пакетів набагато перевершує інтенсивність обслуговування.

Перевантаження ресурсів може привести до повної деградації мережі, коли, не дивлячись на те що мережа передає пакети, корисна швидкість передачі даних виявляється рівною нулю. Це відбувається в тому випадку, якщо затримки доставки всіх пакетів f перевершують деякий поріг, і пакети по тайм-ауту відкидаються вузлом призначення, як застарілі. Якщо ж протоколи, що працюють в мережі, використовують надійні процедури передачі даних на основі квітування і повторної передачі загублених пакетів, то процес перевантаження наростатиме лавино подібно.

Існує ще один важливий параметр; надаючи безпосередній вплив в мережах з комутацією пакетів. Цим параметром є варіація інтервалів вхідного потоку пакетів, тобто пульсація вхідного трафіку. Ми аналізували поведінку моделі теорії черг , що вхідний потік описується пуассоновським розподілом, котре має досить велике стандартне відхилення варіації (нагадаємо, що середня варіація його рівна Т при середньому значенні інтервалу Т, а коефіцієнт варіації рівний 1). А що буде, якщо варіація інтервалів вхідного потоку буде менша? Або вхідний потік буде над пульсуючим?

На жаль, моделі теорії черг не дають для цих випадків простих аналітичних залежностей, подібних до формули (1). Тому для отримання результатів доводиться застосовувати методи імітую чого моделювання мереж або проводити вимірювання в реальній мережі.

На рис. 8.8 показано сімейство залежностей w від р, отриманих для різних значень коефіцієнта варіації CV вхідного потоку. Імітаційна модель враховує фіксовану затримку в мережі. Одна з кривих, у якої CV - 1, відповідає пуассоновському вхідному потоку. З малюнка видно, що чим менше пульсує вхідний потік (CV наближається до нуля), тим менше проявляється ефект лавиноподібного утворення черги при наближенні коефіцієнта завантаження ресурсу до 1. І навпаки, чим більше CV, тим раніше (при менших значеннях р) починає цей ефект виявлятися. З поведінки графіків на малюнку можна зробити два висновки.

Для оцінки значень затримок в чергах на комутаторах мережі недостатньо інформації про коефіцієнт завантаження р, необхідно також знати параметри пульсації трафіку.

Для зниження рівня затримок потрібно знижувати значення р і згладжувати трафік.

Механізми забезпечення QoS можуть бути:.

1. Робота в недовантаженому режимі

Отже, через мережу одночасно протікає велика кількість потоків. Кожен з них вимагає обслуговування відповідно до певних вимог QoS. Кожен потік проходить на шляху проходження від одного кінцевого вузла до іншого через декілька комутаторів, і в кожному з комутаторів він проходить через дві черги - до процесора комутатора і до вихідного каналу комутатора. Ми вже з'ясували, що головним чинником, що впливає на величину затримок, а значить, і на якість обслуговування, є коефіцієнт використання ресурсу. Тому для забезпечення певної якості обслуговування важливо щоб коефіцієнт використання кожного ресурсу (тобто процесорів і комутаторів), який обслуговує потік на шляху його проходження, не перевищував певної величини.

Найпростішим способом забезпечення вимог QOS для всіх потоків являється робота мережі в недовантаженому режимі, коли всі комутатори і канали працюють на 20-30 % від своєї максимальної продуктивності.

Проте це зводить «нанівець» основну гідність мережі з комутацією пакетів, а саме її високу продуктивність при передачі пульсуючого трафіку.

2. Введення різних класів обслуговування

Єдино прийнятним для практики є забезпечення якості обслуговування в навантаженій мережі. Для спрощення розуміння поки ділитимемо всі потоки на два класи - чутливий до затримок (трафік реального часу, наприклад голосовий) і еластичний, такий, що допускає великі затримки, але чутливих до втрат даних.

Ми точно не знаємо залежність затримок від коефіцієнта використання ресурсу, але знаємо загальний вид цієї залежності. Якщо ми забезпечимо для чутливого до затримок трафіку коефіцієнт завантаження кожного ресурсу не більше 0,2, то, очевидно, що затримки в кожній черзі будуть невеликими і швидше за все прийнятними для багатьох типів додатків цього класу. Для еластичного трафіку, слабо чутливого до затримок, можна допустити більш високий коефіцієнт завантаження, але не більше 0,9. Для того, щоб пакети цього класу не втрачалися, потрібно передбачити для них буферну пам'ять, достатню для зберігання всіх пакетів періоду пульсації. Ефект від такого розподілу загрузки ресурсу ілюструє рис. 8.9

Затримки чутливого до затримок трафіку рівні w_s, а затримки еластичного трафіку - w_c

Щоб добитися різних коефіцієнтів використання ресурсів для різних класів трафіку, потрібно в кожному комутаторі для кожного ресурсу підтримувати дві різні черги. Алгоритм вибірки пакетів з черг повинен віддавати перевагу черзі чутливих до затримок пакетів. Якби всі пакети першої черги обслуговувалися пріоритетно, а пакети другої черги - тільки тоді, коли перша черга порожня, то для трафіку першої черги трафік другої черги фактично перестав би існувати. Тому якщо відношення середньої інтенсивності пріоритетного трафіку A,j до продуктивності ресурсу μ рівне 0,2, то і коефіцієнт завантаження для нього рівний 0,2. А ось для еластичного трафіку, пакети якого завжди чекають обслуговування пріоритетних пакетів, коефіцієнт завантаження підраховується по-іншому. Якщо середня інтенсивність еластичного трафіку рівна λ2, то для нього ресурс буде завантажений на $(\lambda 1 + \lambda 2)/\mu$. Отже якщо ми хочемо, щоб для еластичного трафіку коефіцієнт завантаження складав 0,9, то його інтенсивність повинна знаходитися із співвідношення $\lambda 2/\mu = 0,7$.

Основна ідея, лежача в основі всіх методів підтримки характеристик QoS,

полягає в наступному: загальна продуктивність кожного ресурсу повинна бути розділена між різними класами трафіку нерівномірно

Можна ввести більш ніж два класи обслуговування і старатися, щоб кожен клас працював на своїй частині кривої затримок. Якщо така задача вирішена, то можна забезпечити поліпшення характеристик QOS за рахунок інших методів, наприклад знижуючи пульсацію трафіку. Залишилося з'ясувати, яким чином можна забезпечить такі умови для різних класів трафіку в кожному вузлі мережі.

Це завдання вирішується протягом всього часу існування пакетних мереж, тобто вже більше тридцяти років. Довгий час пакетні мережі передавали тільки еластичний трафік, тому основними вимогами QoS були мінімізація втрат пакетів і утримання коефіцієнта навантаження кожного ресурсу мережі не вище 0,9. Методи, вирішальні це завдання, носять назву методів контролю перевантаження.

З появою на початку 90-х чутливого до затримок трафіку ситуація ускладнилась і почалися пошуки нових методів. Власне, термін «якість обслуговування» з'явився саме в цей час, відображаючи детальніші і диференційовані вимоги різних типів трафіку до мережі.

3. Еталонні з’єднання та можливі процеси подій у мережах.

Основу засобів підтримки QoS в мережних елементах складають черги і алгоритми обробки цих черг. Ці механізми використовуються в будь-якому мережному пристрої, яке працює на основі механізму комутації пакетів, - в маршрутизаторі, в комутаторі локальної або глобальної мережі, в кінцевому вузлі виключення складають тільки повторювачі, які пакетів не розрізняють, працюють на рівні потоків бітів).

Черга потрібна для обробки періодів тимчасових перевантажень, коли мережний пристрій не може передавати пакети на вихідний інтерфейс в тому темпі у якому вони поступають для виконання такого просування. Якщо причиною перевантаження є процесорний блок мережного пристрою, то для тимчасового зберігання необроблених пакетів використовується вхідна черга, тобто черга, пов'язана з вхідним інтерфейсом. Би тому ж випадку, коли причина перевантаження полягає в обмеженій швидкості вихідного інтерфейсу (а вона завжди обмежена швидкістю підтримуваного протоколу), то пакети тимчасово зберігаються у вихідній черзі.

Головним по ступеню впливу на виникнення черг чинником є коефіцієнт навантаження пристрою (utilization) - відношення середньої інтенсивності вхідного трафіку пристрою до середньої інтенсивності просування пакетів на вихідний інтерфейс.

Якщо коефіцієнт навантаження більше одиниці, значить, інтенсивність вхідного трафіку постійно вище, ніж інтенсивність просування пакетів на вихідний інтерфейс. Тому черга в пристрої існує завжди, її швидкість намагалась йти до нескінченності, коли б не кінцевий розмір буфера, відведеного під зберігання пакетів, що стоять в черзі. Але і у тому випадку, коли коефіцієнт навантаження менше одиниці, черга теж може існувати, більш того, мати достатньо значну середню довжину.

Це відбувається тоді, коли є деяка варіація інтервалів поступлення пакетів в пристрій - чим більше ця варіація, тим більше середня довжина черги. Варіація

інтервалів надходження пакетів є другим важливим чинником, що впливає на поведінку черг, після коефіцієнта навантаження. При пульсуючому характері багатьох типів трафіку комп'ютерних мереж, коли коефіцієнт пульсацій рівний 100:1 або більш, черги можуть бути значними. Якщо ж ця варіація відсутня, тобто пакети прибувають строго через певні проміжки часу, як це відбувається у трафіку типу рівномірного потоку, то черга при коефіцієнті завантаження, меншому 1, не виникає. Вплив пульсацій трафіку на появу затримок обслуговування добре відомий користувачам сегментів Ethernet, що розділяються. Навіть при значеннях коефіцієнта навантаження сегменту 0,5 затримки доступу до мережі бувають значними, що примушує використовувати ці мережі з коефіцієнтом навантаження сегменту не більше 0,3.

Наслідком виникнення черг являється погіршення якості обслуговування трафіку. Утворюються затримки передачі пакетів, що носять до того ж не постійний характер, а це означає, що ростуть варіації затримок. Крім того, при тривалих пульсаціях черги можуть зростати настільки, що пакети не поміщаються в буферну пам'ять мережних пристроїв і втрачаються.

Оцінка можливої довжини черг в мережених пристроях була б дуже корисною, якби допомагала оцінити параметри якості обслуговування при відомих характеристиках трафіку. Проте поведінку черг є імовірнісний процес, на який впливає багато чинників, особливо при складних алгоритмах обробки черг, що використовують пріоритети або зважене обслуговування різних потоків. Хоча для аналізу черг і розроблена спеціальна область прикладної математики - теорія масовому обслуговування (queuing theory), - вона може дати кількісні оцінки тільки дуже простих ситуацій, не відповідних реальним умовам роботи мережних пристроїв. Тому служба QOS використовує для підтримки гарантованого рівня QoS достатньо складну модель, вирішальну завдання комплексно. Це робиться за допомогою наступних методів:

1 за рахунок попереднього резервування смуги пропускання для трафіку з відомими параметрами (наприклад, значеннями середньої інтенсивності і величини пульсації);

2 Про примусової профілізації вхідного трафіку, що дозволяє підтримують коефіцієнт навантаження пристрою на потрібному рівні;

3 використання складних алгоритмів управління чергами.

Найчастіше в маршрутизаторах і комутаторах застосовуються наступні алгоритми обробки черг:

4 традиційний алгоритм FIFO;

5 пріоритетне обслуговування (Priority Queing), яке також називають «подавляю чим»;

6 зважене обслуговування (Weighted Queing, WQ).

Кожен алгоритм розроблявся для вирішення певних завдань і специфічним чином впливає на якість обслуговування різних типів трафіка в мережі. Можливо і комбіноване застосування цих алгоритмів.

3.1 Традиційний алгоритм FIFO

Принцип традиційного алгоритму FIFO полягає в тому, що у разі перевантаження пакети поміщаються в чергу, а при припиненні перевантаження передаються на вихід в тому порядку, в якому поступили, тобто «першим прийшов - першим пішов»

(First In - First Out, FIFO). У всіх пристроях з комутацією пакетів - це алгоритм обробки черг за умовчанням. Гідністю його являється простота реалізації і відсутність потреби в конфігурації. Та все ж він має і корінний недолік - неможливість диференційованої обробки пакетів різних потоків. Всі пакети стоять в загальній черзі на рівних підставах - і пакети чутливого до затримок голосового трафіка, і пакети нечутливого до затримок, але дуже інтенсивного трафіку резервного копіювання, тривалі пульсації якого можуть надовго затримати голосовий пакет.

Черги FIFO необхідні для нормальної роботи мережних пристроїв, але вони недостатні для підтримки диференційованої якості обслуговування.

3.2 Пріоритетне обслуговування.

Алгоритми пріоритетного обслуговування дуже популярні в багатьох областях обчислювальної техніки, зокрема в операційних системах, коли одним додаткам потрібно віддати перевагу перед іншими при обробці їх в мультипрограмному змішанні. Застосовуються ці алгоритми і для переважної по порівнянню з іншими обробки одного класу трафіку

Механізм пріоритетного обслуговування заснований на розділенні всього мережного З трафіку на невелику кількість класів і подальшого призначення кожному класу деякої числової ознаки - пріоритету.

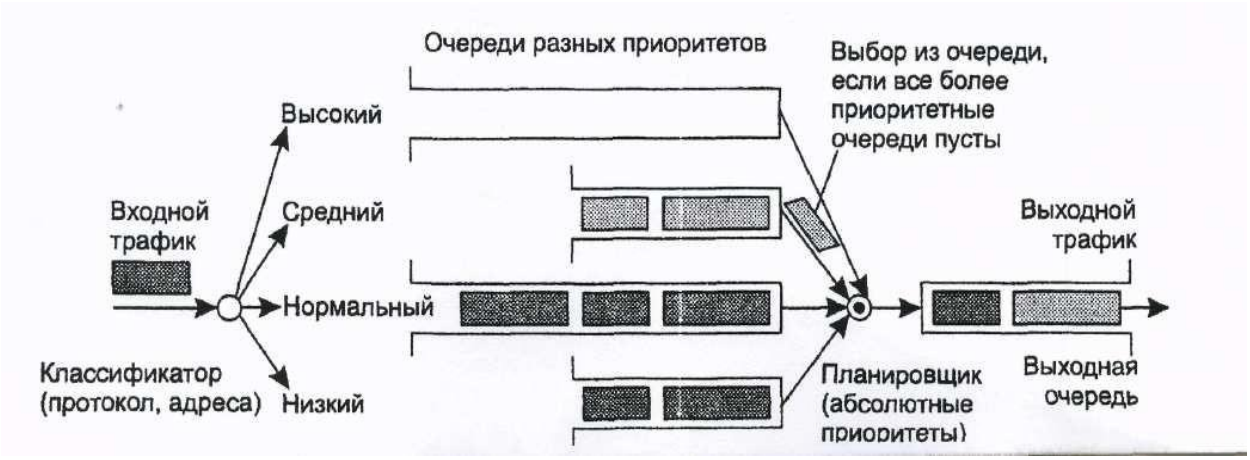
Класифікація трафіку є окремим завданням. Пакети можуть розбиватися на пріоритетні класи на підставі різних ознак: адреси призначення, адреси джерела, ідентифікатора додатку, що генерує цей трафік, будь-яких інших комбінацій ознак, які містяться в заголовках пакетів. Правила класифікації пакетів є частиною політики адміністрування мережі.

Точка класифікації трафіку може розміщуватися в кожному комунікаційному пристрої. Рішення, що більш масштабується рішенням - розміщення функцій класифікації трафіку в одному або декількох пристроях, розташованих на межі мережі (наприклад, в комутаторах корпоративної мережі, до яких підключаються комп'ютери користувачів, або у вхідних маршрутизаторах мережі постачальника послуг). В цьому випадку необхідне спеціальне поле в пакеті, якому можна запам'ятати призначене значення пріоритету, щоб їм могла скористатися решта мережних пристроїв, оброблювальних трафік після класифікую чого пристрою. Таке поле є в заголовку багатьох протоколів. У тих же випадках, коли спеціального поля пріоритету в заголовку немає, розробляється додатковий протокол, який вводить новий заголовок з таким полем (так відбулося, наприклад, з протоколом Ethernet).

Пріоритети можуть призначатися_не тільки комутатором або маршрутизатором, але і додатком на вузлі-відправнику. Необхідно також враховувати, що якщо в мережі відсутня централізована політика призначення пріоритетів, кожен мережний пристрій може не погодитися з пріоритетом, призначеним даному пакету в іншій точці мережі. В цьому випадку воно перепише значення пріоритету у відповідності з локальною політикою, прийнятою безпосередньо на даному пристрої.

У мережному пристрої, що підтримує пріоритетне обслуговування, є декілька черг (буферів), по одній для кожного пріоритетного класу. Пакет, що поступив в період перевантажень, поміщається в чергу, відповідну його пріоритетному классу1. На рис. 8.10 приведений приклад використання чотирьох пріоритетних черг з високим, середнім, нормальним і низьким пріоритетом. До тих пір, поки з

пріоритетнішої черги не будуть вибрані все наявні в ній пакети, пристрій не переходить до обробки наступною, менше пріоритетної черги. Тому пакети з низьким пріоритетом обробляються тільки тоді, коли порожні всі вищестоящі черги: з високим, середнім і нормальним пріоритетами.



Розмір буфера мережного пристрою визначає максимальну довжину черги чекаючи обслуговування пакетів, якщо пакет поступає при заповненому буфері, то він просто відкидається. Зазвичай за умовчанням всім пріоритетним чергам відводяться однакові буфери, але багато пристроїв вирішують адміністратору призначати кожній черзі буфер індивідуального розміру. Розмір буфера визначається в ідеальному випадку так, щоб його вистачало з деяким запасом для зберігання черги середньостатистичної довжини. Проте встановити це значення достатнє складно, оскільки воно змінюється залежно від навантаження мережі, тому потрібне постійне і тривале спостереження за роботою мережі. У загальному випадку, чим вище значущість трафіку для підприємства, чим більше інтенсивність і пульсації, тим більший розмір буфера потрібен цьому трафіку. На прикладі, приведеному на рис. 8.10 для трафіку вищого і нормального пріоритету вибрані великі розміри буферів, а для решти двох класів - менші. Мотиви ухваленого рішення для вищого пріоритету очевидні, а трафік нормального пріоритету має, очевидно, високу інтенсивність і значний коефіцієнт пульсацій.

Пріоритетне обслуговування черг забезпечує висока якість обслуговування для пакетів з найпріоритетнішої черги. Якщо середня інтенсивність їх надходження в пристрій не перевершує пропускної спроможності вихідного інтерфейсу (і продуктивності внутрішніх просуваючи блоків самого пристрою), то пакети вищого пріоритету завжди отримують ту пропускну спроможність, яка їм потрібна. Рівень затримок високо пріоритетних пакетів також мінімальний. Проте він не нульовий і залежить в основному від характеристик потоку цих пакетів - чим вище за пульсацію потоку і його інтенсивність, тим імовірніше виникнення черги, утвореної пакетами даного високо пріоритетного потоку. Трафік решти всіх пріоритетних класів майже прозорий для пакетів вищого пріоритету. Слово «майже» відноситься до ситуації, коли високо пріоритетний пакет вимушений чекати завершення обслуговування низько пріоритетного пакету, якщо його прихід співпадає за часом з початком просування низько пріоритетного пакету на вихідний інтерфейс.

Що ж до решти пріоритетних класів, та якість їх обслуговування буде нижча, ніж у пакетов~самого1шсокого пріоритету, причому рівень зниження може бути дуже істотним. Якщо коефіцієнт навантаження вихідного інтерфейсу, визначуваний

тільки трафіком вищого пріоритетного класу, наближається в якийсь період часу до одиниці, то трафік решти класів_просто на цей час замерзають. Тому пріоритетне обслуговування зазвичай застосовується для класу трафіку, чутливого до затримок, маючого невелику інтенсивність. За таких умов обслуговування цього класу не дуже ущемляє обслуговування решти трафіку. Наприклад, голосовий трафік чутливий до затримок, але його інтенсивність зазвичай не перевищує 8-16 Кбит/с, так що при призначенні йому вищого пріоритету збиток решті класів трафіку буде не дуже значним.

Проте в мережі можуть спостерігатися і інші ситуації. Наприклад, відео трафік також вимагає пріоритетного обслуговування, але має набагато більше високу інтенсивність. Для таких випадків розраховані алгоритми обслуговування черг, що дають низько пріоритетному трафіку деякі гарантії навіть в періоди підвищення інтенсивності високо пріоритетного трафіку.

Уважний читач, очевидно, вже звернув увагу на те, що при описанні роботи пріоритетних черг фігурували не окремі потоки, а класи трафіку. Це важлива особливість, яка відноситься не тільки до пріоритетних алгоритмів, але і до багатьом іншим механізмам підтримки якості обслуговування.

Мережа може обслуговувати трафік з різним ступенем гранулярності. Окремий потік є мінімальною одиницею обслуговування, яку приймають до уваги механізми забезпечення заданих параметрів QoS. Якщо ми забезпечуємо кожному потоку власні параметри QoS, то це підтримання якості обслуговування на рівні потоків. Якщо ми об'єднуємо декілька потоків в єдиний потік і перестаємо розрізняти окремі потоки забезпеченні параметрів QoS, то це підтримка якості обслуговування на рівні класів трафіку. Такі класи також називають агрегатами трафіку.

3.3 Зважені черги

Алгоритм зважених черг розроблений для того, щоб можна було пре¬доставить всім класам трафіку певний мінімум пропускної спроможності або гарантувати деякі вимоги до затримок. Під вагою даного класу розуміється відсоток такою, що надається класу трафіку пропускної спроможності від повної пропускної спроможності вихідного інтерфейсу.

При зваженому обслуговуванні так само, як при пріоритетному, трафік ділиться на декілька класів, і для кожного класу ведеться окрема черга пакетів. Але з кожною чергою зв'язується не пріоритет, а відсоток пропускної спроможності ресурсу, що гарантується даному класу трафіку при перевантаженнях цього ресурсу. Для вхідного потоку таким ресурсом є процесор, а для вихідного потоку (після виконання комутації) - вихідний інтерфейс.

Приклад

Показане на рис. 8.12 пристрій для 5 класів трафіку підтримує 5 черг до вихідного інтерфейсу комутатора. Цим чергам при перевантаженнях виділяється відповідно 10 %, 10%, 30 %, 20 % і 30 % пропускній спроможності вихідного інтерфейсу.

Досягається поставлена мета тим, що черги обслуговуються послідовно і циклічно, і в кожному циклі обслуговування з кожної черги вибирається таке число байтів, яке відповідає вазі даної черги. Наприклад, якщо цикл перегляду черг в даному прикладі рівний одній секунді, а швидкість вихідного інтерфейсу складає 100 Мбит/с, то при перевантаженнях в кожному циклі першої черги приділяється 10

% часу, тобто 100 мс і вибирається 10 Мбіт даних, з другої - теж 10 Мбіт, з третьої - 30 Мбіт, з четвертої - 20 Мбіт, з п'ятої - 30 Мбіт.

В результаті кожному класу трафіку дістається гарантований мінімум пропускної здатності, що у багатьох випадках є бажанішим результатом, ніж придушення низько пріоритетних класів високо пріоритетним.

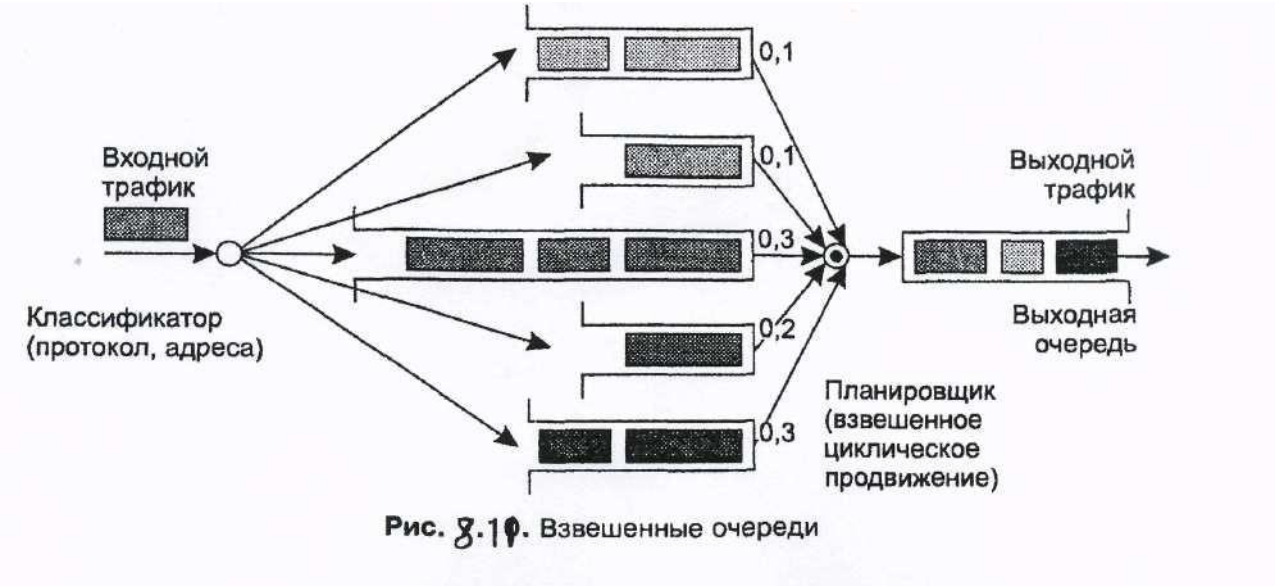


Рис. 8.10. Взвешенные очереди

Так як трафік тередається пакетами, а не бітами, то реальне розподілення пропускної спроможності між класами трафіку завжди трохи відрізняється від планованого. Наприклад, замість 10 % перший клас трафіку може отримати при перевантаженнях 9 або 12 %. Чим більше час циклу, тим точніше дотримуються необхідні пропорції між класами трафіку, оскільки з кожної черги вибирається велике число пакетів і вплив розміру кожного пакету усереднюється.

З іншого боку, тривалий цикл приводить до великих затримок передачі пакетів. Наприклад, при вибраному нами для прикладу циклі в одну секунду затримка може скласти одну секунду і більше - адже арбітр повертається до кожної черги не частіше, ніж раз в секунду, крім того, в черзі може знаходитись більш за один пакет. Тому при виборі часу циклу потрібно забезпечити баланс між точністю дотримання пропорцій пропускної спроможності і прагненням до зменшення затримки.

Для нашого прикладу час циклу в 1000 мкс є прикладом такого балансу. З одного боку, воно забезпечує обслуговування черги кожного класу кожні 1000 мкс, а значить - нижчий рівень затримок. З іншого боку, цього час досить, щоб вибрати з кожної черги в середньому по декілька пакетів (першій черзі в нашому прикладі відводитиметься 100 мкс, що достатньо для передачі у вихідний канал одного пакету Fast Ethernet або десяти пакетів Gigabit Ethernet).

На рівень затримок і варіації затримок пакетів для деякого класу трафіку при зваженому обслуговуванні в значній мірі впливає коефіцієнт використання. В цьому випадку коефіцієнт підраховується як відношення інтенсивності вхідного трафіку класу до пропускної спроможності, виділеної цьому класу відповідно до його ваги. Наприклад, якщо ми виділили першій черзі 10 % від загальної пропускної спроможності вихідного інтерфейсу, тобто 10 Мбит/с, а середня інтенсивність потоку, який потрапляє в цю чергу, рівна 3 Мбит/с, то коефіцієнт використання для цього потоку складе 3/10 = 0,3. Залежність на рис. 8.6 показує, що затримки при такому значення коефіцієнта використання будуть незначними. Якби інтенсивність вхідного потоку цієї черги була 9 Мбит/с, то черги були б значними, а при

перевищенні межі 10 Мбит/с частина пакетів потоку постійно б відкидалася із-за переповнювання черги.

Якісна поведінка черги і, відповідно, затримок тут виглядає приблизно так само, як і у разі черги FIFO - чим менше коефіцієнт загрузки, тим менше середня довжина черги і тим менше затримки.

Як і для пріоритетного обслуговування, при зваженому обслуговуванні адміністратор може призначати різним класам черг буфери різних розмірів. Зменшення розмірів буферів для черг веде до зростання числа втрат пакетів при перевантаженнях, та зате знижує час очікування для тих пакетів, які не були відкинуті і потрапили в чергу.

Існує також такий вид зваженого обслуговування, як зважене справедливе обслуговування (Weighted Fair Queuing, WFQ). У разі подібного обслуговування пропускна спроможність ресурсу ділиться між всіма потоками порівну , тобто «справедливо».

Увага

Зважене обслуговування забезпечує необхідні співвідношення між інтенсивностями трафіку різних черг тільки в періоди перевантажень, коли кожна черга постійно заповнена. Якщо ж яка-небудь з черг порожня (тобто для трафіку даного класу поточний період не є періодом перевантаження), то при перегляді черг вона пропускається, і її час обслуговування розподіляється між рештою черг в відповідності з їх вагою. Тому в окремі періоди трафік певного класу може володіти більшою інтенсивністю чим відповідний процент від пропускної.

Комбіновані алгоритми обслуговування черг

Кожен з описаних підходів має свої достоїнства і недоліки. Пріоритетне обслуговування, забезпечуючи мінімальний рівень затримок для черги найвищого пріоритету, не дає ніяких гарантій відносно середньої пропускної здатності для трафіку черг нижчих пріоритетів. Зважене обслуговування забезпечує заданий розподіл середньої пропускної спроможності, але не враховує вимог до затримок.

Існують комбіновані алгоритми обслуговування черг. У найбільш популярному алгоритмі подібного роду підтримується одна пріоритетна черга при обслуговуванні решти черг відповідно до зваженого алгоритмом. Зазвичай пріоритетна черга використовується для чутливого до затримкам трафіку, а решта - для еластичного трафіку декількох класів. Кожен клас еластичного трафіку отримує деякий мінімум пропускної спроможності при перевантаженнях. Цей мінімум обчислюється як відсоток від пропускної здатності, що залишилася від пріоритетного трафіку. Очевидно, що потрібно якось обмежити пріоритетний трафік, щоб він не поглинав всю пропускну здатність ресурсу. Зазвичай це робиться засобами профілізації трафіка, які розглядаються далі.

Алгоритми управління чергами не запобігають перевантаженням, а лише деяким «справедливим» чином в умовах дефіциту перерозподіляють ресурси між різними потоками або класами трафіку. Алгоритми управління чергами відносяться до механізмів контролю перевантажень, які починають працювати, коли мережа вже переобтяжена.

Існує інший клас засобів, які носять назву механізмів запобігання перенавантаження. Очевидно, що запобігти перевантаженню мережі можна в тому випадку, коли сумарна інтенсивність всіх потоків, що передаються кожним інтерфейсом кожного комутатора мережі, менше пропускної спроможності цього

інтерфейсу. Добитися цього можна двома способами - збільшуючи пропускну здатність інтерфейсу або зменшуючи інтенсивності потоків.

Перший варіант відноситься до засобів проектування і планування мережі і тому тут не розглядається.

Другий варіант - зменшення інтенсивності потоків - можна реалізувати також двома принципово різними способами. Перший спосіб заснований на використанні механізму зворотного зв'язку, за допомогою якого переобтяжений вузол мережі, реагуючи на перевантаження, просить попередні вузли, розташовані уздовж маршруту проходження потоку (або потоків, належних одному класу), тимчасово знизити швидкість трафіку. Після того, як перенавантаження в даному вузлі зникне, він посилає інше повідомлення, що дозволяє підвищити швидкість передачі даних. Інший спосіб базується на передчасному резервуванні пропускної здатності для потоків, що протікають через мережу. Для цього йому необхідна попередня інформація про інтенсивності потоків. Принципи резервування ресурсів ми розглянемо пізніше, а зараз зупинимося на механізмах зворотного зв'язку.

Учасники зворотного зв'язку

Існує декілька механізмів зворотного зв'язку. Вони відрізняються інформацією, яка передається по зворотному зв'язку, а також тим, який тип вузла генерує цю інформацію і хто реагує на цю інформацію - кінцевий вузол (комп'ютер) або проміжний (комутатор або маршрутизатор).

На рис. 8.12 показані різні варіанти організації зворотного зв'язку.

Зворотний зв'язок 1 організована між двома кінцевими вузлами мережі. Це найбільш радикальний спосіб зниження навантаження на мережу, так як тільки кінцевий вузол може понизити швидкість надходження інформації в мережу. Проте цей вид зворотного зв'язку не відносять до методів контролю перевантажень, оскільки його призначення - боротьба з перевантаженнями вузла призначенням, а не з перевантаженнями мережних пристроїв. Принципово ця та ж сама проблема, так як вона є наслідком тимчасового перевищення швидкості надходження пакетів в ресурс над швидкістю обробки цих пакетів. Тільки ресурсом в даному випадку виступає не комутатор мережі, а кінцевий вузол. Але традиційно за цим видом зворотного зв'язку закріпилася власна назва - контроль потоку. Пристрої мережі не беруть участь в роботі цього виду механізму зворотного зв'язку, вони тільки передають відповідні повідомлення між кінцевими вузлами. Не дивлячись на різні назви, в методах контролю перевантаження і контролю потоку використовуються загальні механізми.

При організації зворотного зв'язку важливо враховувати час передачі інформації по мережі. У високошвидкісних глобальних мережах за час передачі повідомлення про перевантаження вузла призначення вузол-джерело може встигнути передати через мережу тисячі пакетів, так що перевантаження не буде ліквідовано вчасно. З теорії автоматичного управління відомо, що затримки в контурі зворотного зв'язку можуть приводити до багатьом небажаним ефектам, первинним цілям. Наприклад, в системі можуть початися коливальні процеси, і вона ніколи не зможе прийти в рівноважний стан. Подібні явища спостерігалися на ранній стадії розвитку Інтернету, коли із-за не удосконалення алгоритмів зворотного зв'язку і алгоритмів маршрутизації в ньому виникли ділянки перевантажень, які періодично переміщалися по мережі. Причина такої проблеми інтуїтивно зрозуміла - затримка в

контурі зворотного зв'язку приводить до того, що регулюючий елемент отримує застарілу інформацію про стан регульованого елементу. В даному випадку вузол-джерело отримує інформацію про стан черги вузла призначення із затримкою. Тому можливі ситуації, коли вузол-джерело починає знижувати швидкість передачі інформації, коли насправді черга у вузлі призначення вже не існує, і, навпаки, підвищувати швидкість передачі інформації в той момент, коли вузол призначення почав випробовувати перевантаження. Для боротьби з такими явищами в контур зворотного зв'язку зазвичай вводиться інтегруючий елемент, який на кожному кроці обробляє не тільки поточне повідомлення зворотного зв'язку, але і декілька попередніх повідомлень, що дозволяє врахувати динаміку зміни ситуації і реагувати адекватно.

Зворотний зв'язок 2 організована між двома сусідніми комутаторами. Комутатор повідомляє сусіда, що знаходиться вище за течією потоку, що він відчуває перевантаження і його буфер заповнився до критичної величини. Отримавши таке повідомлення, сусід, розташований вище за течією, повинен понизити на деякий час швидкість передачі даних у напрямі переобтяженого ком¬мутатора і тим самим вирішити проблему перевантаження. Це менш ефективне для мережі в цілому рішення, оскільки потік продовжуватиме текти від вузла-джерела з тією ж швидкістю, що і раніше. Але для комутатора, який випробовує перевантаження, це є хорошим виходом, оскільки він отримує час для того, щоб розвантажити чергу, що переповнилася. Правда, проблема переноситься в комутатор, розташований вище за течією, в якому тепер може виникнути перевантаження, оскільки він починає передавати дані з свого буфера з меншою швидкістю. Гідністю такого методу є низька затримка зворотного зв'язку, оскільки вузли є сусідами (якщо вони, звичайно, не сполучені супутниковим каналом).

Зворотний зв'язок 3 організована між проміжним комутатором вузлом-джерелом. Повідомлення зворотного зв'язку хоч і передаються декількома комутаторами мережі у напрямі вузла-джерела, але вони на нього не реагують.

Зворотний зв'язок 4. Тут, як і у разі зворотного зв'язку 1, повідомлення про перенавантаження породжується вузлом призначення і передається вузлу-джерелу. Проте є і важлива відмінність: в даному випадку кожен проміжний комутатор реагує на це повідомлення. По-перше, він знижує швидкість передачі даних в направленні вузла призначення, по-друге, він може змінити зміст повідомлення. Наприклад, якщо вузол призначення просить понизити швидкість 30 Мбит/с, то проміжний комутатор може понизити цю величину 20 Мбит/с, оцінивши стан свого буфера. Крім того, породити повідомлення зворотного зв'язку може будь-який комутатор мережі, а не тільки вузол призначення.

При описі різних варіантів організації зворотного зв'язку ми маємо на увазі, що повідомлення про перевантаження йде в напрямі, зворотному напрямі передачі інформації користувача (власне, тому цей механізм так і називається). Проте деякі комунікаційні протоколи не передбачають можливості генерації подібних повідомлень проміжними вузлами. У таких умовах часто використовується штучний прийом - передача повідомлення про перевантаження вузлу призначення, який перетворить його в повідомлення зворотного зв'язку і відправляє в потрібному напрямі, тобто у напрямі джерела. Цей варіант показаний на малюнку як зворотний зв'язок 5.

Інформація зворотного зв'язку

У вживаних сьогодні методах зворотного зв'язку використовуються наступні основні типи повідомлень:

1 ознака перевантаження;

2 максимальна швидкість передачі;

3 максимальний об'єм даних (кредит); 4 непрямі ознаки.

Ознака перевантаження не говорить про ступінь перевантаженості мережі або вузла, він тільки фіксує факт наявності перевантаження. Реакція вузла, що отримав таке повідомлення, може "побут ь різною. У деяких протоколах вузол зобов'язаний припинити передачу інформації в певному напрямі до тих пір, поки не буде отримано інше повідомлення зворотного зв'язку, що вирішує продовження передачі. У інших протоколах вузол поводиться адаптивний, він знижує швидкість на деяку величину і чекає реакції мережі. Якщо повідомлення з ознакою перевантаження продовжують поступати, то він продовжує зниження швидкості.

У другому типі повідомлень указується максимальна швидкість передачі, тобто поріг швидкості, який повинен дотримувати джерело або проміжний вузол, розташований вище за течією потоку. В цьому випадку обов'язково потрібно враховувати час передачі повідомлення по мережі, щоб виключити коливальні процеси в мережі і забезпечити потрібну швидкість реакції на перевантаження. Тому в територіальних мережах такий спосіб зазвичай реалізується силами всіх комутаторів мережі (зворотний зв'язок 4 в нашому прикладі).

Повідомлення «максимальний об'єм даних» широко використовується в пакетних мережах алгоритмі ковзаю чого вікна (див. розділ 6). Цей алгоритм дозволяє не тільки забезпечувати надійну передачу даних, але і реалізувати зворотний зв'язок для контролю потоку між кінцевими вузлами. Параметром, несучим інформацію зворотного зв'язку, є «вікно» - число, тісно пов'язане з поточним розміром вільного простору в буфері приймаючого вузла. Вікно також називають кредитом, який той, що приймає дає вузлу, що передає

Вузол, що передає, може з будь-якою швидкістю передати об'єм інформації (або окрема кількість пакетів, якщо вікно вимірюється в пакетах), відповідний кредиту. Але якщо кредит вичерпаний, то вузол, що передає, не має права передавати інформацію, поки не отримає наступний кредит. При перевантаженнях приймаючий вузол зменшує розмір вікна, тим самим знижуючи навантаження. Якщо ефект перевантаження зникає, то приймаючий вузол збільшує розмір вікна. Недоліком цього алгоритму є те, що він працює тільки в протоколах зі встановленням з'єднання.

І, нарешті, що в деяких випадках передає вузол визначає, що приймаючий вузол (або вузли) випробовує перевантаження, по деяких непрямим признакам без отримання повідомлення зворотного зв'язку. Такими непрямими ознаками можуть бути факти втрати пакетів. Для того, щоб протокол міг виявляти факти втрат пакетів, це повинен бути протокол зі встановленням з'єднання. Тоді закінчення тайм-ауту або прихід дубліката позитивної квитанції побічно свідчать про те, що пакет втрачений. Проте втрата пакету не завжди свідчить про перевантаження мережі. Перевантаження мережі - це тільки одна з можливих причин втрати пакету, іншою причиною може бути ненадійна робота комунікаційних пристроїв (відмови устаткування, спотворення даних із-за перешкод). Але оскільки реакція на перевантаження і ненадійну роботу мережі повинна бути однаковою і полягати в зниженні швидкості передачі, то неоднозначність причини втрати пакету не є

проблемою.

Прикладом протоколу, що використовує неявну інформацію про перевантаження, являється протокол TCP. Цей протокол за допомогою явної інформації зворотного зв'язку (про розмір вікна) здійснює контроль потоку, а за допомогою неявної (втрати пакетів, дублікати квитанцій) - контроль перевантаження.

Резервування ресурсів і комутація пакетів

Як вже було сказано вище, ще одним механізмом запобігання перенавантаження в мережі, разом із зворотним зв'язком, є резервування ресурсів. Головна ідея резервування полягає в тому, щоб обмежити рівень перевантажень деякою прийнятною величиною. Ця величина повинна бути такій, щоб алгоритми контролю перевантаження, що працюють в комутаторах мережі, справлялися з короткочасними перевантаженнями і без зворотного зв'язку забезпечували необхідні значення характеристик QOS.

Резервування ресурсів в мережах з комутацією пакетів принципово відрізняється від подібної процедури в мережах з комутацією каналів. У мережах з комутацією каналів для кожного каналу резервується (виділяється) фіксована частка пропускної спроможності лінії зв'язку (фізичного каналу). Потік передається через мережу з постійною швидкістю, рівній зарезервованій для нього пропускній спроможності. При цьому пропускна спроможність з'єднання завжди закріплена за цим потоком, вона не може динамічно перерозподілятися серед інших потоків. Попереднє резервування є невід'ємною властивістю мережі з комутацією каналів.

У мережах з комутацією пакетів резервування не є обов'язковим. Інколи у визначенні методу комутації пакетів відсутність резервування фігурує як основна властивість цього типу мереж. Але і в тих випадках, коли резервування в мережах з комутацією пакетів виконується, воно відрізняється від резервування ресурсів в мережах з комутацією каналів тим, що тут враховується пульсуючий характер трафіку і можливість динамічного перерозподілу пропускної спроможності мережі між потоками (агрегатами).

Резервування полягає в тому, що всі мережні пристрої уздовж проходження потоку повинні виділити цьому потоку (агрегату) деяку частину пропускної спроможності своїх інтерфейсів і продуктивності процесорів, рівну середній необхідній швидкості передачі даних потоку. Пояснимо це на прикладі.

Приклад

Припустимо, що в початковому стані ресурси мережі, показаної на рис. 8.13. не були зарезервовані. Потім було вирішено виділити деякі ресурси мережі потоку 1. Для цього необхідно знати, принаймні, такий параметр потоку, як середню потрібну швидкість передачі даних. Припустимо, що ця швидкість для потоку 1 рівна 15 Мбит/с, а пропускні спроможності всіх каналів зв'язку (а значить, і інтерфейсів комутаторів) рівні 100 Мбит/с. Для спрощення вважатимемо, що кожен вхідний інтерфейс оснащений власним процесором, продуктивність якого перевищує продуктивність даного інтерфейсу, так що процесор не може бути вузьким місцем, і ми не прийматимемо його в розрахунок при ухваленні рішення про виділення ресурсів.

Потік 1 може бути прийнятий на обслуговування, тому що всі інтерфейси на його шляху володіють достатньою для його обслуговування продуктивністю ($15 < 100$). Тому резервування виконується, і кожен інтерфейс уздовж шляху потоку

запам'ятовує, що він вже виділив 15 Мбит/с своєї продуктивності потоку 1.

Допустимо, що після цього виникла потреба в резервуванні ресурсів для потоку 2, який володіє середньою швидкістю передачі даних 70 Мбит/с. Таке резервування також може бути зроблене, оскільки у всіх інтерфейсів уподовж маршруту потоку 2 вільна (не зарезервована) пропускна спроможність інтерфейсів перевищує 70 Мбит/с. У тих інтерфейсів, через які проходять як потік 1, так і потік 2 (інтерфейси 13 і ц комутаторів S2 і S3 відповідно), залишається 85 Мбит/с вільної пропускної спроможності, а у решти інтерфейсів - 100 Мбит/с. Після резервування у інтерфейсів і3 і I1 залишається по 15 Мбит/с вільної пропускної спроможності.

Також виявляється успішною спроба резервування пропускної спроможності для потоку 3, середня швидкість якого рівна 10 Мбит/с. Проте резервування для потоку 4, середня швидкість якого 20 Мбит/с, виявляється неможливим, оскільки у інтерфейсів і3 і I1 залишилося тільки по 5 Мбит/с вільної пропускної спроможності.

Цей приклад показує, що мережа відмовляється прийняти на обслуговування потік, якщо вона не може гарантувати йому необхідний рівень якості обслуговування. Ми, звичайно, спростили схему резервування ресурсів. В дійсності, мережа може гарантувати потоку не тільки дотримання його середньою швидкістю, про яку ми говорили в прикладі, але і забезпечити інші характеристики QOS, такі як максимальна затримка, максимальна варіація затримки і допустимий рівень втрат даних. Проте для цього мережа повинна знати деякі додаткові параметри потоку, наприклад його максимальний рівень пульсації, щоб зарезервувати необхідний простір в буфері.

Вільна пропускна спроможність для чутливого до затримок трафіку і для еластичного трафіку повинна при резервуванні враховуватися окремо. Щоб забезпечити для пріоритетного трафіку прийнятний рівень затримок і їх варіацій, максимальна сумарна резервована пропускна спроможність не повинна перевищувати 30-50 % від загальної пропускної спроможності кожного ресурсу. Для ілюстрації цього скористаємося попереднім прикладом. Хай ми вирішили відвести чутливому до затримок трафіку 30 % пропускної спроможності ресурсів. Тоді, якщо чутливими до затримок є потоки 1 і 3, то резервування для них можливо. Якщо ж такими потоками являються потоки 1 і 2, то немає, скільки сумарна середня швидкість цих потоків рівна 85 Мбит/с, що ільш ніж 30 Мбит/с (30 % від100 Мбит/с).

Якщо ми маємо на увазі, що чутливий до затримок графік буде обслуговуватись в пріоритетній черзі, то при резервуванні пропускної спроможності для еластичного трафіку потрібно враховувати, що йому може бути виділена тільки та частина пропускної спроможності, яка залишилася від чутливого до затримок трафіку. Наприклад, якщо потоки 1 і 3 є чутливими до затримок і ми виділили їм необхідну середню пропускну спроможність 30 Мбит/с, то для еластичних потоків залишається тільки 70 Мбит/с свободою пропускної здатності.

Що ж міняється в мережі після того, як в ній виконано резервування? Нічого принципово нового. Просто мережа виявляється завантаженою раціональним чином. У ній немає ресурсів, які працюють з перевантаженням. Механізми організації черг як і раніше забезпечують тимчасову буферізацію пакетів в періоди пульсацій. Оскільки ми планували завантаження ресурсів з розрахунку середніх швидкостей передачі даних, то на періодах пульсацій протягом деякого обмеженого часу швидкості потоків можуть перевищувати середні швидкості, так що механізми боротьби з перевантаженнями як і раніше потрібні. Для забезпечення необхідних

середніх швидкостей потоків на періодах перевантажень відповідні потоки можуть обслуговуватися за допомогою зважених черг. Головна перевага методу комутації пакетів також зберігається: якщо деякий потік не витрачає відведеної йому пропускній спроможності, то вона може бути використана для обслуговування іншого потоку. Нормальною практикою є резервування пропускної спроможності тільки для частини потоків, тоді як інші потоки обслуговуються без резервування, отримуючи обслуговування по можливості (з максимальними зусиллями). Тимчасово вільна пропускна спроможність може для таких потоків виділятися динамічно, без порушення узятих зобов'язань по обслуговуванню потоків, для яких ресурси зарезервовані.

Мережа з комутацією каналів подібного перерозподілу ресурсів виконати не може, оскільки у неї в розпорядженні немає одиниць інформації, що незалежно адресуються, - пакетів!

Приклад.

Проілюструємо принципову відмінність резервування ресурсів в мережах з комутацією пакетів і в мережах з комутацією каналів на прикладі автомобільного трафіку. Хай в деякому місті вирішили забезпечити деякі привілеї для руху машин швидкої допомоги, В ході обговорення цього проекту виникли дві конкуруючі ідеї його реалізації. Перший варіант передбачав виділення для автомобілів швидкої допомоги окремої смуги на всіх дорогах міста, недоступною для другого транспорту ні за яких умов, навіть якщо в якийсь період часу машин швидкої допомоги на дорозі немає. У другому випадку для машин швидкої допомоги також виділялась окрема смуга, але у відсутності привілейованих машин по ній дозволялось рухатися і іншому транспорту. У разі ж появи машини швидкої допомоги автомобілі, що займають виділену смугу, зобов'язані були її звільнити. Неважко відмітити, що перший варіант відповідає принципу резервування в мережах з комутацією каналів - пропускна спроможність виділеної смуги монопольно використовується автомобілями швидкої допомоги і не може бути перерозподілена навіть тоді, коли вона їм не потрібна. Другий варіант є аналогією резервування в мережах з комутацією пакетів. Пропускна спроможність дорогі тут використовується ефективніше, але для потоку автомобілів швидкої допомоги такий варіант менш сприятливий, оскільки при необхідності звільнення смуги виникають перешкоди, що створюються непривілейованими машинами.

Повертаючись від автомобільного трафіку до мереж з комутацією пакетів, слід відзначити: для того, щоб дотримати гарантії обслуговування кожного потоку, описаної схеми резервування недостатньо. Ми припустили, що точно знаємо середню пропускну спроможність і параметри пульсацій потоків.

На практиці такі відомості не завжди бувають достовірними. А що трапиться, якщо потік поступатиме в мережу з швидкістю, що перевищує ту, яку ми враховували при резервуванні? І ще одне важливе питання залишається відкритим - як забезпечити автоматичне резервування пропускної спроможності уздовж маршруту проходження потоку? Для вирішення поставлених завдань в мережі необхідна система забезпечення якості обслуговування, в яку крім механізмів управління чергами входять деякі додаткові механізми.

Системи забезпечення якості обслуговування, засновані на резервуванні

Система забезпечення якості обслуговування має розподілений характер, оскільки її елементи повинні бути присутніми на всіх мережних пристроях,

просуваючи пакети: комутаторах, маршрутизаторах, серверах. З іншого боку, роботу окремих мережних пристроїв по підтримці характеристик QoS потрібно координувати, щоб якість обслуговування була однорідною уздовж всього шляху, по якому слідують пакети потоку. Тому служба QoS повинна включати також елементи централізованого управління, за допомогою яких адміністратор мережі міг би погоджено конфігурувати механізми підтримання характеристик QoS в окремих пристроях мережі.

1Система забезпечення якості обслуговування, що базується на резервуванні ресурсів, складається з підсистем декількох типів (рис. 8.14)

2 механізмів обслуговування черг; 3 протоколу резервування ресурсів;

4 механізмів кондиціонування трафіку.

Механізми обслуговування черг використовуються в періоди тимчасових перенавантаження. При цьому зазвичай застосовуються комбінації пріоритетної черги з чергами із зваженим обслуговуванням.

Протокол резервування ресурсів потрібний для автоматизації процедури резервування на всьому шляху проходження деякого потоку, тобто «з кінця в кінець». Протокол резервування є аналогом протоколів встановлення з'єднання в мережах з комутацією каналів, тому він іноді називається сигнальним протоколом, відповідно до термінології, прийнятої для цього типу мереж.

Повідомлення протоколу резервування ресурсів роблять два «проходи» по мережі. Спочатку джерело генерує повідомлення, яке проходить мережу в прямому напрямку до приймача інформації. У цьому повідомленні протоколу резервування міститься так званий профіль трафіку, тобто такі його характеристики, як середня швидкість, параметри пульсації, а також вимоги до якості обслуговування, наприклад до рівня затримок. На підставі цього профілю і вимог QoS кожен комутатор на шляху проходження потоку приймає рішення про можливість або неможливість виконати резервування для цього потоку. Якщо він «погоджується» виконати резервування, то повідомлення передається далі, а комутатор робить відмітки про параметри проведеного резервування. Якщо всі комутатори уздовж маршруту згодні із запрошеними параметрами резервування, то останній комутатор передає нове повідомлення протоколу резервування у зворотному напрямі. При проходженні цього повідомлення кожен комутатор фіксує параметри резервування для даного потоку.

Ініціювати роботу сигнального протоколу може не тільки кінцевий вузол, але і проміжний пристрій. В цьому випадку гарантоване обслуговування потоку виконуватиметься не на всьому шляху проходження трафіку, а тільки в межах визначеного у частка мережі, що знижує якість обслуговування.

Протокол резервування ресурсів дозволяє виконувати резервування як для окремих потоків, так і для класів трафіку. Принципи його роботи в обох випадках залишаються однаковими. Проте ініціатором резервування ресурсів для класу трафіку є не кінчений вузол, якого цікавить особистий потік, а один з комутаторів мережі. Таким комутатором частіше всього стає прикордонний комутатор мережі постачальника послуг, який приймає потоки різних користувачів.

У мережах з віртуальними каналами функції протоколу резервування ресурсів зазвичай виконує протокол встановлення віртуального каналу - це є його додатковим завданням. У датаграмних мережах протокол резервування є самостійним протоколом. Прикладом такого протоколу є протокол резервування ресурсів

(ReSource reservation Protocol, RSVP), який працює в IP-мережах. Резервування може виконуватися і вручну адміністратором мережі, який повинен конфігурувати параметри резервування для кожного потоку в кожному комутаторі мережі.

Механізми кондиціонування трафіку стежать за тим, щоб поточні параметри потоків відповідали заявленим при резервуванні. Це свого роду контрольно-пропускні пункти, які перевіряють трафік на вході в комутатор. Без таких механізмів неможливе виконання гарантій обслуговування трафіку. Якщо середні швидкості потоків або пульсації перевищать той рівень, який був вказаний при резервуванні, то затримки і втрати пакетів вище допустимих. Таке перевищення може відбутися з різних причин. Скажімо, тому, що достатньо важко точно оцінити параметри трафіку. Передчасне вимірювання середньої швидкості і пульсації можуть не дати правильного результату, тому що ці характеристики можуть мінятися з часом, і через тиждень вони вже не відповідатимуть дійсності. Крім того, не можна виключати умисного спотворення характеристик трафіку, що можливо при використанні комерційних послуг

Механізм кондиціонування трафіку зазвичай виконує декілька функцій.

1 Класифікація трафіку. Ця функція виділяє із загальної послідовності пакетів, що поступають в пристрій, пакети одного потоку, загальні вимоги, що має, до якості обслуговування. У мережах з віртуальними каналами ознакою потоку є мітка віртуального шляху, тому додаткової класифікації не вимагається. У датаграмних мережах такої ознаки, як правило, немає, тому класифікація виконується на основі декількох формальних ознак пакету - адрес джерела і призначення, ідентифікаторів застосувань і т.п. Без класифікації пакетів в датаграмних мережах неможливо забезпечити необхідну якість обслуговування.

2 Профілізація трафіку. Для кожного вхідного потоку в кожному комутаторі є відповідний йому набір параметрів QoS, тобто профіль трафіку. Профілізація трафіку має на увазі перевірку відповідності кожного вхідного потоку параметрам його профілю. Існують алгоритми, які дозволяють виконати таку перевірку автоматично в темпі надходження пакетів на вхідний інтерфейс комутатора. Прикладами алгоритмів профілізації є алгоритми «дірявого відра» і «відра маркерів». Ці алгоритми будуть розглянуті при вивченні окремих технологій, таких як IP, Frame Relay і ATM.

У разі порушення параметрів профілю (наприклад, перевищення продовжування пульсації або середньої швидкості) відбувається відкидання або маркіровка пакетів цього потоку. Відкидання деяких пакетів знижує інтенсивність потоку і приводить його параметри у відповідність з вказаними в профілі. Маркіровка пакетів без відкидання потрібна для того, щоб пакети все ж таки були обслужені даним вузлом (або подальшими по потоку)» але з якістю обслуговування, відмінною від вказаного в профілі

Формування трафіку (shaping). Ця функція призначена для придання минулому профілізацію трафіку потрібної тимчасової «форми». В основному за допомогою даної функції прагнуть згладити пульсації трафіку, щоб пакети на виході пристрою з'являлися більш рівномірно, ніж на вході. Згладжування пульсацій скоротить черги в мережних пристроях, які оброблятимуть трафік далі по потоку. Його також потрібно використовувати для відновлення тимчасових співвідношень трафіку додатків, що працюють з рівномірними потоками, наприклад голосових додатків.

Механізми кондиціонування трафіку можуть підтримуватися кожним вузлом мережі або реалізовуватися тільки в прикордонних пристроях. Останні варіант часто використовують постачальники послуг, кондиціонуючи трафік своїх клієнтів.

Існує принципова відмінність поведінки описаної системи для забезпечення середньої швидкості потоку, з одного боку, і для забезпечення необхідних порогів затримок і варіацій затримок, з іншого боку.

Необхідне значення середньої швидкості обслуговування забезпечується за рахунок конфігурування відсотка пропускної спроможності, що виділяється, при зваженому обслуговуванні. Тому мережа може виконати запит на будь-яке значення середньої швидкості для потоку, якщо воно не перевищує вільної пропускної спроможності ресурсів мережі уздовж цього потоку.

Проте мережа не може конфігурувати алгоритм пріоритетного обслуговування так, щоб він строго забезпечив який-небудь заздалегідь заданий поріг затримок і їх варіації. Напрям пакетів в пріоритетну чергу тільки дозволяє гарантувати, що затримки будуть достатньо низькими - істотно менше, ніж у пакетів, які обробляються по лгоритму заженого обслуговування. Але аналітично оцінити кількісні значення затримок дуже важко. Яким же чином постачальник послуг може виконати свої зобов'язання перед клієнтами?

Як правило, вирішення даної проблеми ґрунтується на вимірюванні трафіку в мережі. Постачальник послуг повинен організувати пріоритетне обслуговування трафіка з однією або декількома пріоритетними чергами, вимірюючи затримки реального трафіку і обробляючи результати статистичними методами. Це означає, що він повинен будувати гістограми розподілу затримок для різноманітних шляхів проходження потоків і визначати середні затримки, середні варіації, максимальні затримки і максимальні варіації для кожного класу трафіку, чутливого до затримок. На підставі цих характеристик постачальник вибирає деякі граничні значення характеристик QoS, які він може гарантувати своїм клієнтам. Зазвичай такі граничні значення вибираються з деяким запасом, щоб при появі деякої кількості нових клієнтів мережа могла дотримувати заявлені гарантії.

Інжиніринг трафіка

При розгляді системи забезпечення якості обслуговування, заснованої на резервуванні, ми не стали піднімати питання маршрутів проходження потоків через мережу. Точніше, ми вважали, що вони якимсь чином вибрані, причому цей вибір робиться без урахування вимог QoS. І в умовах заданості маршрутів ми прагнули забезпечити проходження по цих маршрутах такого набору потоків, для якого можна гарантувати дотримання вимог QoS. Очевидно, що задачу підтримки вимог QoS можна вирішити більш ефективно, якщо вважати, що маршрути проходження трафіку не фіксовані, а також підлягають вибору. Це дозволило б мережі обслуговувати більше потоків з гарантіями QoS при тих же характеристиках самої мережі, тобто пропускній спроможності каналів і продуктивності комутаторів і маршрутизаторів.

Завдання вибору маршрутів для потоків (або класів) трафіку з урахуванням дотриманням вимог QoS вирішують методи інжинірингу трафіку (Traffic Engineering, TI). За допомогою цих методів прагнуть добитися ще однієї мети - по можливості максимально і збалансовано завантажити всі ресурси мережі, щоб мережа при заданому рівні якості обслуговування володіла як можна більш високою

сумарною продуктивністю.

Методи ТІ, як і інші розглянуті раніше методи, засновані на резервуванням ресурсів. Тобто вони не тільки дозволяють знайти раціональний маршрут для потоку, але і резервують для нього пропускну спроможність ресурсів мережі, що знаходяться уздовж цього маршруту.

Методи інжинірингу трафіку є порівняно новими для мереж з комутацією пакетів. Це пояснюється багато в чому тим, що передача еластичного трафіку не пред'являла строгих вимог до параметрів QoS. Крім того, Інтернет довгий час не був комерційною мережею, тому максимальне використання ресурсів не вважалося першочерговим завданням для IP-технологій, лежачих в основі Інтернету.

Сьогодні ситуація змінилася. Мережі з комутацією пакетів повинні передавати різні види трафіку із заданою якістю обслуговування, максимально використовують можливості своїх ресурсів. Проте для цього їм потрібно змінити деякі, що стали вже традиційними, підходи до вибору маршрутів.

Недоліки традиційних методів маршрутизації

Основним принципом роботи протоколів маршрутизації в мережах з комутацією пакетів ось вже довгий час є вибір маршруту на основі топології мережі без урахування інформації про її поточне завантаження.

Для кожної пари «адреса джерела - адреса призначення» такі протоколи вибирають єдиний маршрут, не зважаючи на інформаційні потоки, що протікають через мережу. В результаті всі потоки між парами кінцевих вузлів мережі йдуть по найкоротшому (відповідно до деякої метрики) маршруту. Вибраний маршрут може бути раціональнішим, наприклад, якщо в розрахунок приймається номінальна пропускна спроможність каналів зв'язку або затримки, що вносяться ними, або менш раціональним, якщо враховується тільки кількість проміжних маршрутизаторів між початковим і кінцевим вузлами.

Традиційні методи маршрутизації розглядають якнайкращий вибраний маршрут як єдиного можливого, навіть якщо існують інші, хоч і декілька гірші маршрути.

Класичним прикладом неефективності такого підходу є «риба» - мережа з топологією, приведеною на рис. 8.15. Не дивлячись на те що між комутаторами A і E існує два шляхи (верхній - через комутатор B і ніжній - через комутатори C і D), весь трафік від комутатора A до комутатора E відповідно до традиційних принципів маршрутизації прямує по верхньому шляху. Тільки тому, що нижній шлях трохи (на одну ділянку ретрансляції) довший, ніж верхній, він ігнорується, хоча міг би працювати «паралельно» з верхнім шляхом.

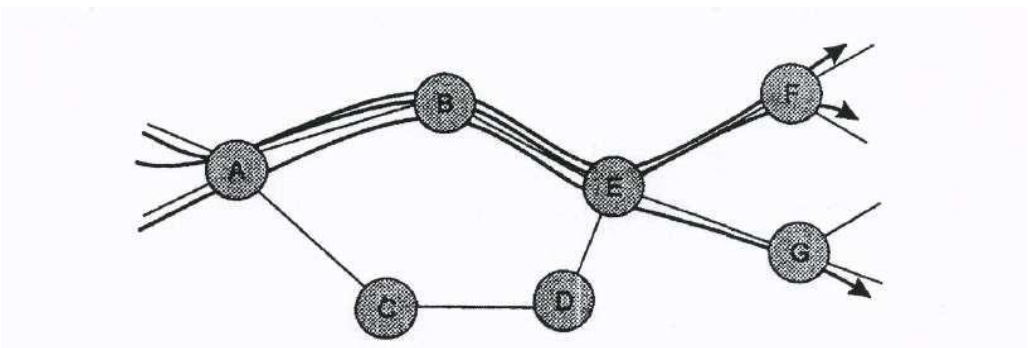


Рис. 8.15 Неефективність найкоротших шляхів

Такий підхід приводить до того, що навіть якщо найкоротший шлях

переобтяжений, пакети все одно посилаються по цьому шляху. Так, в мережі, представлений на рис. 8.15 верхній шлях продовжуватиме використовуватися навіть тоді, коли його ресурсів перестане хапати для обслуговування трафіку від комутатора А до комутатору Е, а нижній шлях простоюватиме, хоча, можливо, ресурсів комутаторів В і З вистачило б для якісної передачі цього трафіку.

У наявності явна збитковість методів розподілу ресурсів мережі - одні ресурси працюють з перевантаженням, а інші не використовуються зовсім. Традиційні методи боротьби з перевантаженнями цю проблему вирішити не можуть, потрібні якісно інші механізми.

Методи інжинірингу трафіку

Початковими даними для методів інжинірингу трафіку є:

1 характеристики мережі, що передає, - її топологія, а також продуктивність складових її комутаторів і ліній зв'язку (рис. 8.16);

2 зведення про запропоноване навантаження мережі, тобто про потоки трафіку, які мережа повинна передати між своїми прикордонними комутаторами (рис. 8.17)

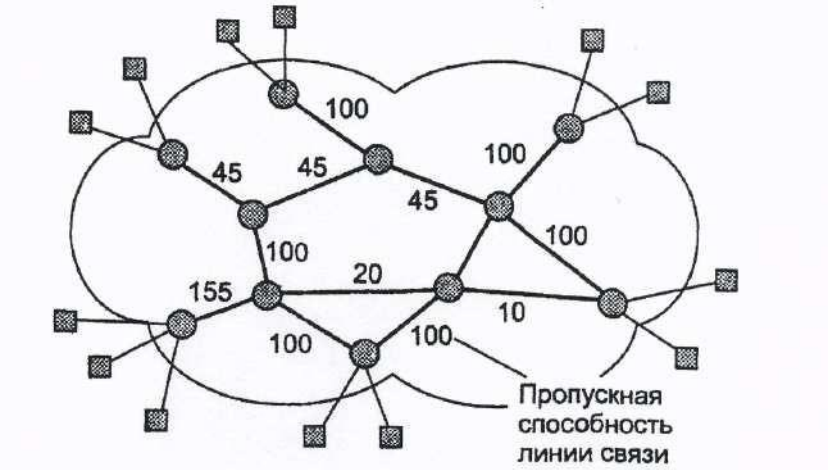


Рис. 8.16 Топологія мережі і продуктивність її ресурсів

Хай продуктивність процесора кожного комутатора достатня для обслуговування трафіку всіх його вхідних інтерфейсів, навіть якщо трафік потрапляє на інтерфейс з максимально можливою швидкістю, рівній пропускній спроможності інтерфейсу. Тому при резервуванні ресурсів вважатимемо ресурсами пропускну спроможність ліній зв'язку між комутаторами, який визначає також пропускну спроможність двох інтерфейсів, зв'язаних цією лінією.

Кожен потік характеризується точкою входу в мережу, точкою виходу з мережі і профілем трафіку. Для отримання оптимальних рішень можна використовувати детальнийопис кожного потоку, наприклад враховувати величину можливою пульсації трафіку або вимоги QoS. Проте оскільки кількісно оцінити їх вплив на роботу мережі досить складно, а вплив цих параметрів на характеристики QoS менш значущо, то для знаходження субоптимального розподілення шляхів проходження потоків через мережу, як правило, враховуються тільки їх середні швидкості передачі даних, що і показане на рис. 8.17.

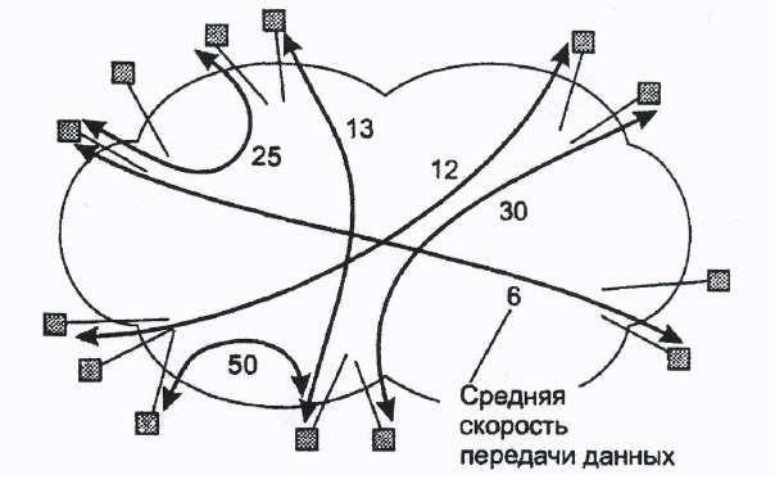


Рис. 8.17. Запропоноване навантаження

Методи ТІ частіше працюють не з окремими потоками, а з агрегованими потоками, які є об'єднанням декількох потоків. Оскільки ми шукаємо загальний маршрут для декількох потоків, то агрегувати можна тільки потоки, що мають загальні точки входу в мережу і виходу з мережі. Агреговане завдання потоків дозволяє спростити завдання вибору шляхів, оскільки при індивідуальному розгляді кожного призначеного для користувача потоку проміжні комутатори повинні зберігати дуже великі об'єми інформації, оскільки індивідуальних потоків може бути дуже багато. Необхідно, проте, підкреслити, що агрегація окремих потоків в один можливо тільки у тому випадку, коли всі потоки, що становлять, пред'являють одні і ті ж вимоги до якості обслуговування. Далі в цьому розділі ми скорочено користуватимемося терміном «потік» як для індивідуального потоку, так і для агрегованого, оскільки принципи ТІ від цього не міняються.

Завдання ТІ полягає у визначенні маршрутів проходження потоків трафіку через мережу, тобто для кожного потоку потрібно знайти точну послідовність проміжних комутаторів і їх інтерфейсів. При цьому маршрути повинні бути такими, щоб всі ресурси мережі були навантажені до максимально можливого рівня, а кожен потік отримував необхідну якість обслуговування.

Максимальний рівень використання ресурсів вибирається так, щоб механізми контролю перевантаження могли забезпечити необхідну якість обслуговування. Це означає, що для еластичного трафіку максимальне значення вибирається не більше, ніж 0,9, а для чутливого до затримок трафіку - не більше, чим 0,5. Оскільки звичайне резервування проводиться не для всіх потоків, то потрібно залишити частину пропускної спроможності для вільного використання. Тому приведені максимальні значення зазвичай зменшують до 0,75 і 0,25 відповідно. Для спрощення міркувань ми вважатимемо далі, що в мережі передається один вид трафіку, а потім покажемо, як узагальнити методи ТІ для випадку трафіку декількох типів.

Існують різні формальні математичні визначення завдання ТІ. Ми тут обмежимося найбільш простим визначенням, тим більше що сьогодні воно найчастіше використовується на практиці.

Вважатимемо, що рішенням задачі ТІ є такий набір маршрутів для заданого безлічі потоків трафіку, для якого всі значення коефіцієнтів використання ресурсів уздовж маршруту проходження кожного потоку не перевищують деякого заданого порогу К максимальне.

На рис. 8.19 показане одне з можливих рішень задачі, ілюструють яку рис. 8.16 і 8.17. Знайдені маршрути гарантують, що максимальний коефіцієнт використання будь-якого ресурсу для будь-якого потоку не перевищує 0,6.

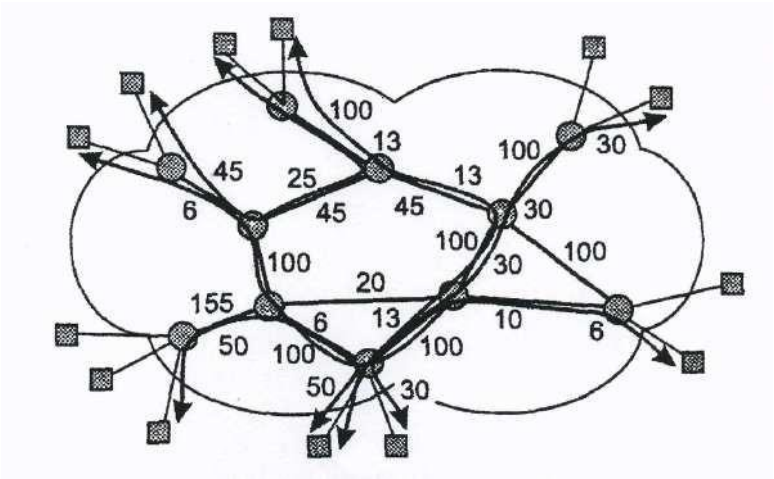


Рис. 8.19- Розподіл навантаження по мережі - вибір шляхів
^ передачі трафіку

Рішення задачі ТІ можна шукати по-різному. По-перше, можна шукати його передчасно, у фоновому режимі. Для цього потрібно знати початкові дані: топологію і продуктивність мережі, а також вхідні і вихідні точки потоків трафіку і середню швидкість передачі даних в них. Після цього завдання раціонального розподілу шляхів проходження трафіку при фіксованих точках входу і виходу, а також заданому рівні максимального значення коефіцієнта використання ресурсу можна передати деякій програмі, яка, наприклад, за допомогою направленого перебору варіантів знайде рішення. Результатом роботи програми будуть точні маршрути для кожного потоку з вказівкою всіх проміжних комутаторів.

По-друге, можна вирішувати задачу ТІ в оперативному режимі, доручивши її самим комутаторам мережі. Для цього використовуються модифікації стандартних протоколів маршрутизації. Модифікація протоколів маршрутизації полягає в тому, що вони повідомляють один одному не тільки топологічну інформацію, але і текуче значення вільної пропускної спроможності у кожного ресурсу.

Після того, як рішення знайдене, потрібно його реалізувати, тобто утілити в таблицях маршрутизації. На цьому етапі може виникнути проблема - в тому випадку, якщо ми хочемо прокласти ці маршрути в датаграмной мережі. Річ у тому, що таблиці маршрутизації цих мереж враховують тільки адреси призначення пакетів. Комутатори і маршрутизатори таких мереж (наприклад, IP-мереж) не працюють з потоками, для них потік в явному вигляді не існує, кожен пакет при його просуванні є незалежною одиницею комутації. Можна сказати, що таблиці просування цих мереж відображають тільки топологію мережі (напрями просування до певних адрес призначення).

Тому привнесенні методів резервування в датаграмних мережах трапляється з великими труднощами. У протоколах резервування, подібних згаданому раніше протоколу RSVP, використовується деякий додатковий набір ознак крім адреси призначення, щоб визначити потік для датаграмного маршрутизатора. При цьому поняття потоку потрібне тільки на етапі резервування, а при просуванні пакетів як і раніше працює традиційна для цього типу мереж схема, що враховує тільки адресу

призначення.

Тепер представимо ситуацію, коли у нас є декілька потоків між двома х кінцевими вузлами, і ми хочемо направити їх по різних маршрутах. Ми прийняли таке рішення, виходячи з балансу завантаження мережі, коли вирішували задачу інжинірингу трафіку. Датаграмний комутатор або маршрутизатор не має можливості реалізувати наше рішення, тому що для всіх цих потоків у нього в таблиці просування є тільки один запис, відповідний загальній адресі призначення пакетів цих потоків. Змінювати логіку роботи комутаторів і маршрутизаторів датаграмних мереж досить недоцільно, оскільки це дуже принципова модернізація.

Тому методи інжинірингу трафіку сьогодні використовуються тільки в мережах з віртуальними каналами, для яких не складає труднощів реалізувати найдене рішення для групи потоків. Кожному потоку (або групі потоків з однаковими маршрутами) виділяється віртуальний канал, який прокладається відповідно до вибраного маршруту. Методи інжинірингу трафіку успішно застосовуються в мережах ATM і Frame Relay, що працюють на основі техніки віртуальних каналів. IP-мережі також спираються на методи TI, коли ті використовуються в мережах ATM або Frame Relay, що працюють в складеній мережі, побудованій на ос¬нове протоколу IP. Існує також нова технологія MPLS, яка розроблена спеціально як засіб того, що привнесло техніки віртуальних каналів в IP-мережі. На основі технології MPLS в IP-мережах можна також вирішувати задачі TI.

Інжиніринг трафіку різних класів

При рішенні задачі інжинірингу трафіку ми вважали, що всі потоки трафіку пред'являли однакові вимоги до якості обслуговування. Тобто користувачів мережі задовольняло, що всі потоки обслуговуються із заданою середньою швидкістю (вона, природно, у кожного потоку своя, що відрізняється від інших).

Реальнішою є ситуація, коли у кожного користувача мережі є декілька класів трафіку, і ці класи відрізняються різними вимогами до якості обслуговування.

Методи TI, що враховують наявність в мережі трафіку з різними вимогами QoS, вирішують проблему точно так, як і методи резервування ресурсів окремих вузлів. Якщо у нас є, наприклад, два класи трафіку, то ми задамся двома рівнями максимального використання ресурсів.

Для досягнення такого результату з кожним ресурсом повинні бути зв'язані два лічильники вільної пропускної спроможності - один для пріоритетного трафіку, другий для еластичного трафіку. При визначенні можливості проходження маршруту через конкретний ресурс для пріоритетного трафіку середня інтенсивність нового потоку повинна порівнюватися з вільною пропускний спроможністю для пріоритетного трафіку.

Якщо вільну пропускну спроможність досить і новий потік буде проходити через даний інтерфейс, то значення середньої швидкості передачі даних для нового потоку необхідно відняти як з лічильника завантаження пріоритетного трафіку, так і з лічильника завантаження еластичного трафіку, оскільки пріоритетний трафік завжди обслуговуватиметься перед еластичним і створить додаткове навантаження для еластичного трафіку. Якщо ж завдання TI вирішується для еластичного трафіку, то його середня швидкість передачі даних порівнюється з вільною пропускною спроможністю лічильника еластичного трафіку і у разі позитивного рішення значення цієї швидкості віднімається тільки з лічильника еластичного трафіку,

оскільки для пріоритетного трафіку еластичний трафік прозорий.

Модифіковані протоколи маршрутизації повинні поширювати по мережі інформацію про два параметри вільної пропускної спроможності - для кожного класу трафіку окремо. Якщо ж завдання узагальнюється для випадку передачі через мережу трафіку декількох класів, то, відповідно, з кожним ресурсом повинне бути зв'язане стільки лічильників, скільки класів трафіку існує у мережі, а протоколи маршрутизації повинні поширювати вектор вільних пропускних спроможностей відповідної розмірності.

Висновки

Якість обслуговування в його вузькому сенсі фокусує увагу на характеристиках і методах передачі трафіку через черги комунікаційних пристроїв. Методи забезпечення якості обслуговування займають сьогодні одне з найважливіших місць в арсеналі технологій мереж з комутацією пакетів, оскільки без їх застосування неможлива робота сучасних мультимедійних застосувань, таких як IP-телефонія, відео і радіомовлення, інтерактивне дистанційне навчання і т, п.

Характеристики QoS відображають негативні наслідки перебування пакетів в чергах, які виявляються в зниженні швидкості передачі, затримках пакетів і їх втратах.

Пріоритетні і зважені черги, а також резервування і зворотний зв'язок дозволяють гарантувати якість обслуговування для чутливого до затримок і еластичного трафіка.

Алгоритм ковзаю чого вікна забезпечує не тільки надійну передачу пакетів, але і являється ефективним засобом зворотного зв'язку.

Архітектура заснованої на резервуванні системи підтримки якості обслуговування включає:

1 механізми черг;

2 протоколи резервування, що дозволяють автоматично виділяти необхідні ресурси для «крізного» потоку;

3 засобу кондиціонування трафіку, що виконують класифікацію, профілізацію і формування трафіку.

Методи інжинірингу трафіку полягають у виборі раціональних маршрутів проходження потоків через мережу. Вибір маршрутів забезпечує максимізацію завантаження ресурсів мережі при одночасному дотриманні необхідних гарантій відносно параметрів якості обслуговування трафіку.

Виконати самостійне завдання.

1. Вивчити питання лекції.
2. Визначити чисельне значення первинних параметрів на кожному рівні ієрархії мереж доступу за завданням лабораторного заняття № 4.

Література:

1. Стеклов В.К., Беркман Л.Н. Телекомунікаційні мережі. Київ, “Техніка”, 2001.- с.321-340.
2. Стеклов В.К., Беркман Л.Н. Проектування телекомунікаційних мереж. Київ, “Техніка”, 2003. – с.192-198.
3. Олифер В.Г., Олифер Н.А. Компьютерные сети. Принципы, технологии, протоколы: учебник для вузов. 4изд.-СПб.: 2010.- с.184-224.