УДК 621.396.662.072.078 : 004.738 (043.3)

Моденов С. Ю., аспірант (Тел. +380 (93) 600 85 39. E-mail: modenovs@mail.ru)
(Национальный авиационный университет, г. Киев)

АНАЛІЗ СПЕЦІАЛІЗОВАНИХ КОМП'ЮТЕРНИХ МЕРЕЖ МЕТОДАМИ ТЕОРІЇ МАСОВОГО ОБСЛУГОВУВАННЯ

Моденов С. Ю. Аналіз спеціалізованих комп'ютерних мереж методами теорії масового обслуговування. У роботі розглянуті особливості застосування методів теорії масового обслуговування до спеціалізованих мереж реального часу з різнорідним трафіком. Зроблено розрахунки зростання черг за наявності перенавантаження мережних комутаційних вузлів при різних степенях самоподібності трафіку. Показано, що для дотримання високого рівня коефіцієнта використовування мережі потрібно застосовувати методи регулювання або вирівнювання (policing and shaping) інтенсивності самоподібного трафіку, придушувати активність джерел, що перенавантажують мережу, та обирати розміри буферів більшими, ніж це витікає з результатів класичного аналізу черг.

Ключові слова: комп'ютерна мережа, система масового обслуговування, різнорідний трафік, самоподібність, параметр Херста, аналіз черг

Моденов С. Ю. Анализ специализированных компьютерных сетей методами теории массового обслуживания. В работе рассмотренны особенности применения методов теории массового обслуживания к специализированным сетям реального времени с разнородным трафиком. Сделаны расчеты роста очередей при наличии перегрузки сетевых коммутационных узлов при разных степенях самоподобия трафика. Показано, что для соблюдения высокого уровня коэффициента использования сети, нужно применять методы регулирования или выравнивания (policing and shaping) интенсивности самоподобного трафика, подавлять активность источников, которые перегружают сеть, и выбирать размеры буферов больше, чем это следует из результатов классического анализа очередей.

Ключевые слова: компьютерная сеть, система массового обслуживания, разнородный трафик, самоподобие, параметр Херста, анализ очередей

1. Вступ

Мета цієї роботи – описати практичні методи застосування теорії масового обслуговування (ТМО) до комп'ютерних мереж спеціального призначення та критичного застосування, зокрема, до мереж для систем управління виробничими процесами, які за визначенням є системами реального часу. Виникає багато випадків, коли важливо прогнозувати вплив деякої зміни в конструкції та/або топології мережі: або очікується зріст навантаження на мережу, або планується модифікація чи розширення мережі.

Продуктивність мережі – системна характеристика. В інтерактивній системі або у системі реального часу параметр продуктивності – час реакції. У багатьох випадках забезпечення потрібної продуктивності – головна проблема. Для її вирішення можливі такі підходи.

1. Аналіз, заснований на фактичних значеннях. Він робиться після практичної реалізації проекту.
2. Аналіз на основі принципу подібності. Робиться екстраполяція результатів функціонування існуючої мережі на майбутню мережу більшого масштабу.
3. Розробка аналітичної моделі, заснованої на теорії масового обслуговування.
4. Розробка і дослідження імітаційної моделі.

Перший метод – практично метод проб та помилок. Це досить дорогий та неефективний метод.

Другий метод є більш перспективним. Однак недоліком цього методу є те, що поведінка більшості систем при поточних змінах навантаження не співпадає з інтуїтивно очікуваними результатами.

Типовий приклад, взятий з роботи [1], наведений на Рис. 1. Верхня лінія показує, як змінюється затримка відповіді $t_{â}$ системи з розподілом ресурсів при збільшенні коефіцієнту використання $k_{â}$ мережі (нагадаємо, що коефіцієнт використання мережі є відношення навантаження на мережу до пропускної спроможності мережі).

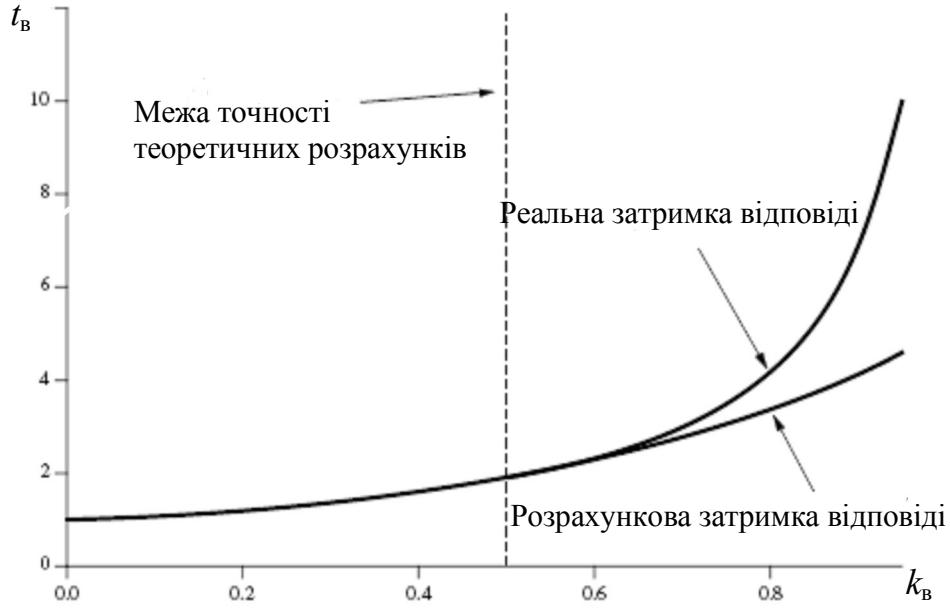

Рис. 1.

Можна зробити висновок, що при простій екстраполяції реальна якість роботи системи не відповідає розрахунковій вже при $k_в \geq 0,6\ldots0,7$, а при $k_в \geq 0,8\ldots0,9$ мережа взагалі починає працювати "на себе", передаючи втрачені пакети знову і знову.

Метод використання аналітичних моделей у вигляді набору рівнянь, у результаті рішення яких можуть бути отримані бажані параметри (затримка відповіді, продуктивність, і т.п.) – більш точний інструмент прогнозу. Для комп'ютерів, операційних систем, мережних технологій, інших практичних задач аналітичні моделі, засновані на теорії масового обслуговування, забезпечують прийнятну збіжність теорії та практики. Трудність використання ТМО полягає в тому, що для розв'язання рівнянь і отримання рішень у замкнутій формі треба робити цілий ряд спрощуючих припущень.

Останній метод – імітаційна модель. Тут, використовуючи спеціалізовані мови програмування для створення імітаційних моделей, можна з достатньою гнучкістю та детальністю моделювати реальні процеси та об'єкти і уникати введення багатьох припущень, потрібних при використанні теорії масового обслуговування напряму. Проте, в більшості випадків, імітаційна модель не може служити у якості першого етапу аналізу. Точність результатів імітаційного моделювання в усіх випадках обмежена точністю вхідних даних.

За результатами порівняльного аналізу методів оцінювання продуктивності мережі можна зробити висновок, що найбільш придатною є комбінація теоретичних методів, зокрема, теорії масового обслуговування, з методами імітаційного моделювання. Розглянемо моделі ТМО, які доцільно застосовувати до аналізу комп'ютерних мереж.

2. Моделі теорії масового обслуговування

2.1. Модель одноканальної системи. Найпростішу систему масового обслуговування (СМО) зображено на Рис. 2. Центральний елемент системи – сервер, який обслуговує деякі заявки. Ці заявки поступають в систему обслуговування. Якщо сервер вільний, заявка обслуговується негайно. Інакше заявка, що прибуває, стає в чергу. Коли сервер завершив обслуговування заявки, вона відбуває. Якщо є заявки, що чекають в черзі, одна з них негайно поступає на обслуговування до сервера. Сервер в цій моделі може виконувати функцію обслуговування заявок.

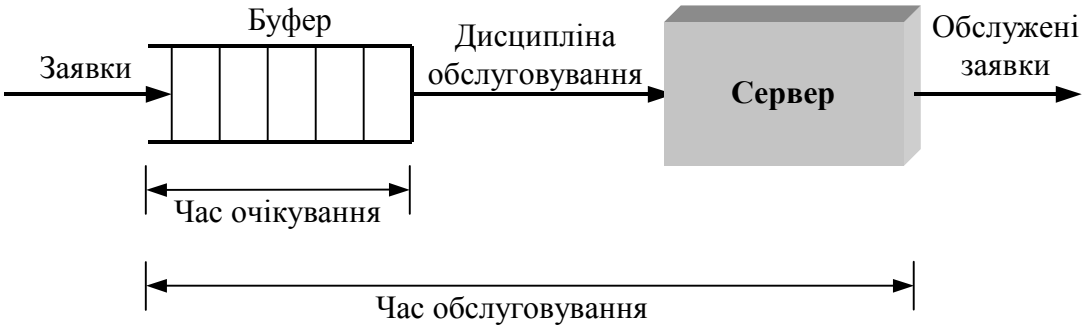

Рис. 2. Одноканальна система масового обслуговування

2.2. Параметри потоку заявок. Незважаючи на зовнішню простоту схеми, зображеної на Рис. 2, тут ілюструються деякі важливі параметри, пов'язані із моделями масового обслуговування. Заявки прибувають в буфер з деякою середньою інтенсивністю λ (число заявок у секунду).

У будь-який даний час, в черзі буде знаходитись певна кількість заявок (нуль або більше); об означимо середнє число заявок у черзі через w, середнє число заявок, що обслуговуються – через ρ, а середній час очікування T_w. Цей час усереднюється по всім заявкам, що поступають на вхід, з урахуванням тих, які не чекають взагалі. Сервер обслуговує заявки, що поступають, з середнім часом обслуговування T_s. Це часовий інтервал між посилкою заявки до сервера і виходу обслуженої заявки з сервера. Інтенсивність обслуговування μ – це число обслужених заявок за одиницю часу. Загальне середнє число заявок, що знаходяться в системі, в тому числі заявка, що обслуговується (якщо вона є) і заявки, що очікують обслуговування (якщо вони є), означимо r і середній час, впродовж якого заявка знаходиться в системі (чекає своєї черги і обслуговується,) - T_r; цей час розглядаємо як середній час загального знаходження заявки в системі (очікування плюс обслуговування).

Якщо ми припускаємо, що місткість черги нескінченна, то ніякі заявки ніколи в системі не втрачаються; вони тільки затримуються впродовж часу очікування та обслуговування. При цих обставинах, середнє число відправлених заявок дорівнює середньому числу прибуваючих заявок у одиницю часу.

При збільшенні інтенсивності прибуття заявок на вхід системи час знаходження заявок в системі також збільшується, що призводить до заторів. Черга стає довшою, час очікування збільшується. При $\rho = 1$, тобто $\lambda = \mu$, сервер насичується, працюючи 100% часу. Тому теоретична максимальна інтенсивність вхідного потоку пов'язана з середнім часом обслуговування T_s як $\lambda_{\max} = 1/T_s$.

Проте при насиченні системи, коли $\rho \to 1$, черга зростає практично до нескінченності. На практиці при обмеженому розмірі буферної пам'яті та наявності обмежень на затримку відповіді зазвичай обмежують інтенсивність вхідного потоку в одноканальній системі від 70% до 90% відносно теоретичного максимуму.

2.3. Модель багатоканальної системи. На Рис. 3 зображено узагальнену модель багатоканальної системи обслуговування з загальним буфером. За винятком інтенсивності обслуговування, всі параметри, використані при аналізі одноканальної системи, мають той же сенс. Якщо ми маємо N ідентичних серверів з однаковою інтенсивністю обслуговування кожним сервером, що дорівнює ρ, то можна вважати, що середня інтенсивність обслуговування системи в цілому дорівнює $N\rho$; цей останній термін часто співвідносять з інтенсивністю трафіку u, що чисельно дорівнює інтенсивності вхідного потоку заявок λ. Теоретичний максимум відносної інтенсивності обслуговування дорівнює $N \times 100\%$, а теоретичний максимум інтенсивності вхідного потоку є $\lambda_{\max} = N/T_s$.

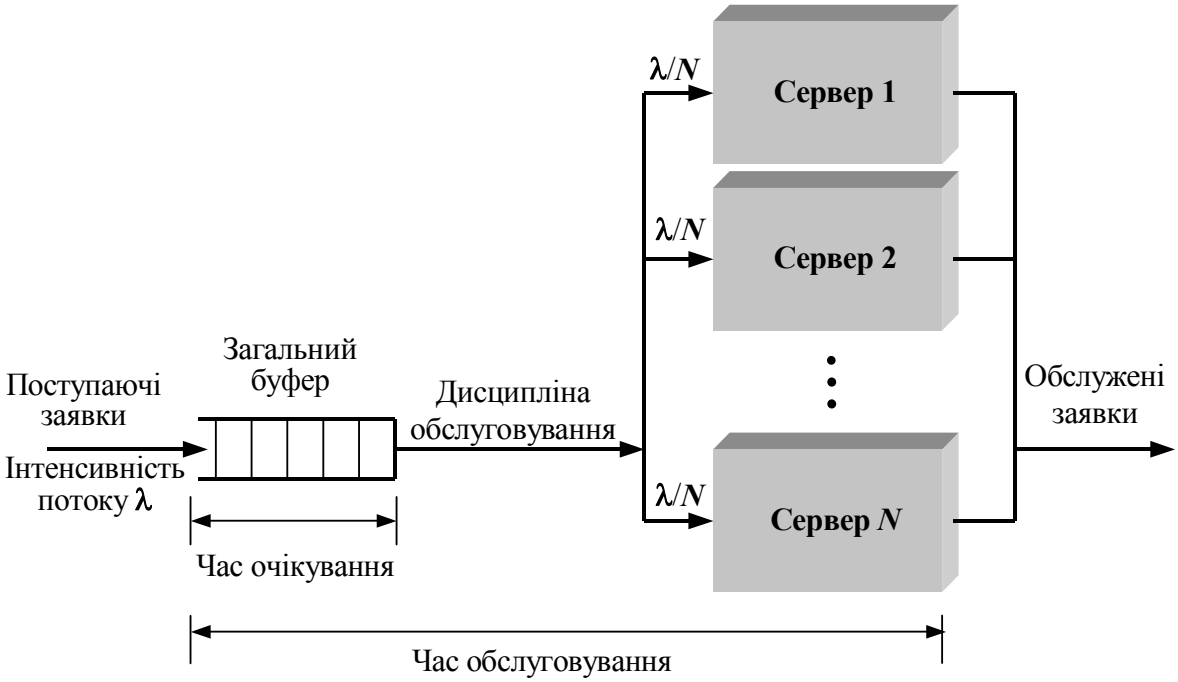

Рис. 3. Багатоканальна система обслуговування з загальною буферною пам'яттю (загальна черга з заданою дисципліною обслуговування)

На Рис. 4 зображено багатоканальну систему з розділеною буферною пам'яттю, що можна трактувати як паралельну структуру з одно канальних систем обслуговування. Хоча зміни в структурі не є принциповими, робочі характеристики зображеної системи можуть істотно відрізнятися від тої, що розглянуто раніше.

Ключові характеристики для черги з декількома обслуговуючими пристроями аналогічні характеристикам для одноканальної системи. Припускається нескінченний об'єм буферної пам'яті і нескінченний розмір черги, з розподілом черги між всіма обслуговуючими пристроями (серверами). Звичайно вважають, що реалізується дисципліна обслуговування в порядку надходження (*FIFO*). Для випадку багатоканальної системи обслуговування, якщо всі сервери передбачаються ідентичним, вибір специфічного сервера для чергової заявки не впливає на час обслуговування.

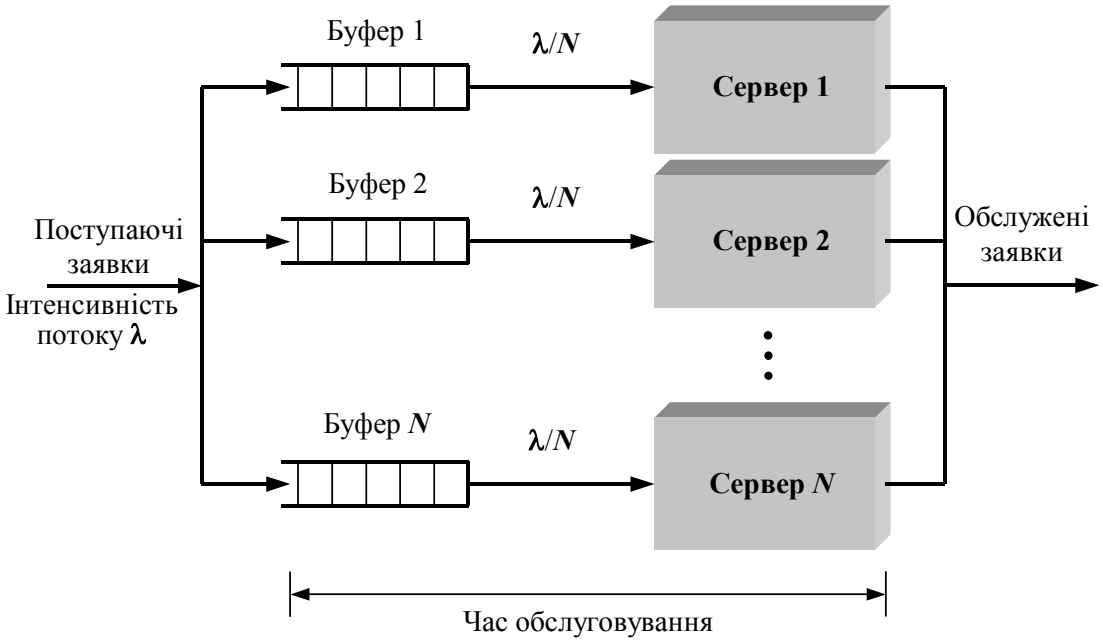

Рис. 4. Багатоканальна система обслуговування з роздільною буферною пам'яттю (індивідуальні черги з заданими дисциплінами обслуговування)

3. Основні співвідношення теорії масового обслуговування для самоподібного трафіку

Для подовження аналізу зробимо деякі спрощуючі припущення. Ці припущення, звичайно, можуть погіршити адекватність моделі реальним ситуаціям, але в більшості випадків, результати будуть достатньо точні для вирішення задач планування і розробки конкретних проектів.

Наведемо деякі співвідношення, справедливі, наприклад, для стаціонарних та ергодичних процесів приходу заявок на обслуговування. Ці співвідношення можуть бути корисні як асимптотичні наближення реальних процесів.

Для оцінювання середнього розміру черзі r за умов стаціонарності та ергодичності процесу приходу заявок використовуються так звані формули Літтла [2]:

- для одноканальної системи обслуговування $r = \lambda T_r$, $r = w + \rho$;
- для N-канальної системи обслуговування $\rho = \lambda T_r / N$, $u = \lambda T_s = \rho N$, $r = w + N\rho$,

де $T_r = T_w + T_s$.

Відповідно можна через формули Літтла зв'язати число ρ з інтенсивністю приходу заявок λ та часом знаходження заявки в системі T_s. Воно дорівнює $\rho = \lambda T_s$.

Таким чином, для аналізу СМО необхідно мати таку апріорну інформацію:

- інтенсивність вхідного потоку заявок;
- середній час обслуговування;
- число каналів обслуговування.

На основі даної інформації можна отримати асимптотичні оцінки середнього числа заявок у черзі, середній час очікування та загальний час знаходження заявки в системі.

Показано, що трафік даних, що циркулює в цифрових мережах, і, зокрема, у комп'ютерних мережах з комутацією пакетів, має так звані самоподібні, або фрактальні, властивості [3]. Самоподібність являє собою властивість процесу зберігати своє поводження

і зовнішні ознаки при розгляді в різному масштабі. Для часових послідовностей масштабуємою величиною є час. Виходячи з визначення самоподібності, можна стверджувати, що часові і спектральні характеристики випадкового процесу (у нашому випадку – трафіку) при зміні масштабу усереднення будуть описуватися тими самими рівняннями, функціями, але з відповідними масштабними коефіцієнтами. Іншими словами, самоподібність якого-небудь процесу (явища) можна трактувати як інваріантність до змін чи масштабу розміру.

Потоки пакетів самоподібного трафіку розподілені не по закону Пуассона, а по іншим імовірнісним законам з так званими "важкими хвостами". Це розподіли Парето, Вейбулла, логарифмічно-нормальний розподіл, гамма-розподіл, бета-розподіл та деякі інші, менш популярні.

Наприклад, вираз для щільності імовірності розподілу Парето має наступний вид:

$$f(x) = \frac{\alpha}{k}\left(\frac{k}{x}\right)^{\alpha+1},$$

де k і a $(a,k) < 0$ – параметри розподілу.

Відповідно функція імовірності та середнє значення:

$$F(x) = 1 - \left(\frac{k}{x}\right)^{\alpha} \quad (x > k; \alpha > 0); \qquad E[X] = \frac{\alpha}{\alpha - 1}k \quad (\alpha > 1).$$

Реальні випадкові процеси, звичайно, зберігають властивість самоподібності тільки до певної межі. Ця міра статистичної усталеності процесу при багаторазовому масштабуванні визначається так званим параметром Херста чи параметром самоподібності. Випадковий процес $x(t)$ є статистично самоподібним з параметром Херста H $(0{,}5 \le H \le 1)$, якщо для будь-якого $a > 0$ процес $x(at)/a^H$ має ті ж статистичні характеристики, що і сам процес $x(t)$:

– математичне очікування $\mathrm{M}[x(t)] = \dfrac{\mathrm{M}[x(at)]}{a^H}$;

– дисперсія $\mathrm{D}[x(t)] = \dfrac{\mathrm{D}[x(at)]}{a^{2H}}$;

– кореляційна функція $R(t,\tau) = \dfrac{R(at, a\tau)}{a^{2H}}$.

Чим більше H, тим довше зберігається властивість самоподібності при багаторазовому масштабуванні. При $H = 0{,}5$ ця властивість практично відсутня.

Кореляційні функції самоподібних процесів з великим параметром Херста загасають повільніше, ніж у звичайних випадкових процесів, причому спадання має, як правило, коливальний характер. Установлено, що спадання постійної складової кореляційної функції відбувається за законом $c_1 t^{-c_2 a}$, де c_1, c_2 – константи, a – параметр масштабу. Відповідно і спектральна щільність процесу теоретично прагне до нескінченності при частоті, що наближується до нуля.

Такі специфічні характеристики властиві різнорідному трафіку типу *Triple Play* (мова – відео – дані) та *Quadro Play* (мова – відео – дані плюс мобільні абоненти). Фізично вони обумовлені високим ступенем групування пакетів на клієнтських ділянках, у маршрутизаторах і вузлах комутації. Навіть якщо джерело породжує регулярний потік пакетів, дані до споживача доставляються серіями, що перемежовуються інтервалами простою [3]. Причинами цього є обмежена швидкість роботи мережних пристроїв, недостатній обсяг буферів і ін. [4].

Крім того, самоподібній трафік має особливу структуру, що зберігається при багаторазовому масштабуванні - у реалізації, як правило, є присутнім деяка кількість викидів при відносно невеликому середньому рівні трафіку. Через такі сплески навантаження характеристики мережі також погіршуються: збільшуються втрати, затримки, джиттер пакетів при проходженні через вузли мережі.

Методи розрахунку вимог до мереж нових поколінь (пропускної здатності каналів, ємності буферів і ін.) засновані на марковських моделях і формулах Ерланга чи Літтла, що з успіхом використовувалися при проектуванні телефонних мереж, можуть давати невиправдано оптимістичні рішення і приводити до недооцінки навантаження.

При самоподібній природі трафіку залежність середньої тривалості черги (відповідно, необхідного розміру буфера) q від середнього коефіцієнта використання має наступний вид:

$$q = \frac{\rho^{1/2(1-H)}}{(1-\rho)^{H/(1-H)}} .$$

При $H=0,5$ ця формула спрощується: $q = \rho/(1-\rho)$, що являє собою класичний результат СМО з найпростішим вхідним потоком і показово розподіленим часом обслуговування ($M/M/1$). Для системи з детермінованим часом обслуговування ($M/D/1$) класичний результат виглядає в такий спосіб:

$$q = \frac{\rho}{1-\rho} - \frac{\rho^2}{2(1-\rho)} .$$

Наведемо ***приклад розрахунку*** залежності середнього числа заявок в одноканальній системі від коефіцієнту використання системи $\rho = \lambda/\mu$. Розрахунки зроблено по наведених формулах як для найпростішого (Пуассонівського) потоку заявок, так і для самоподібних потоків, проведені розрахунки, результати яких зображені на Рис. 6.

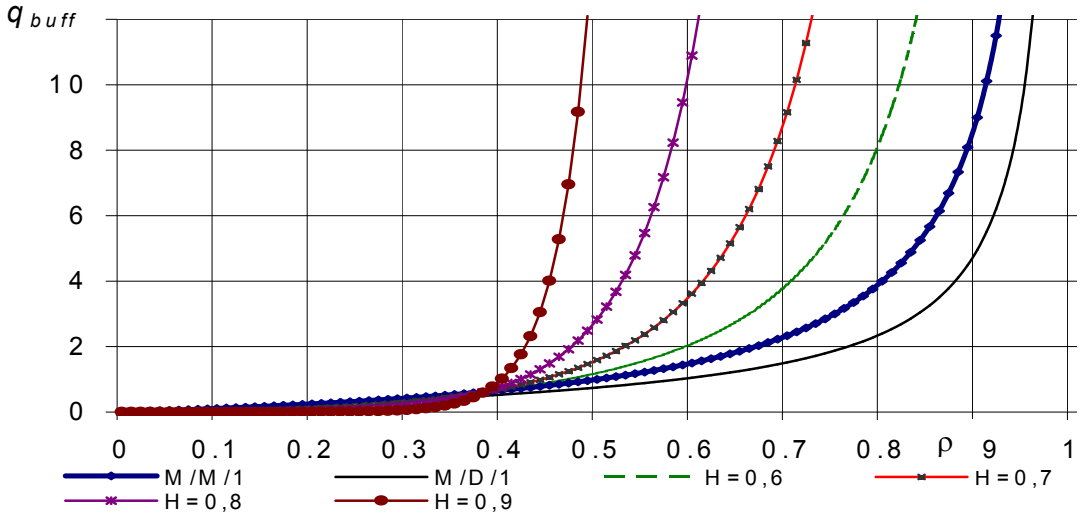

Рис. 6. Залежності довжини черги заявок (потрібної пам'яті буфера q_{buff}) від коефіцієнта використання ρ для різних моделей вхідного трафіку.

На графіках добре видно, що для самоподібного трафіку вже при $\rho \approx 0,4$ потрібно збільшувати ресурси пропускної спроможності мережі та пам'яті буферних пристроїв, у порівняні з класичною моделлю M/M/1, що вважається найменш сприятливої в порівнянні з іншими (наприклад, з постійним чи гауссівським розподіленим часом обслуговування). Швидкість росту необхідного обсягу пам'яті росте при збільшенні параметру Херста, що

обумовлено, в основному, ступенем групування однорідних пакетів і сплесками навантаження на мережу [5].

Можна також зробити висновок, що просте нарощування буферної пам'яті (апаратним чи програмним способом) є малоефективним. При очікуваному збільшенні частки трафіку даних у загальному обсязі ступінь самоподібності буде збільшуватися, і залежність $\rho\left(q_{buff}\right)$ буде зростати все більш різко. Крім того, просте збільшення об'ємів буферної пам'яті може призвести до зворотного ефекту – збільшення затримок доставки даних із-за довгого перебування пакетів у черзі. Більш ефективними методами усунення впливу самоподібності трафіку є формування та згладжування трафіку (policing and shaping) – так звані "діряве відро" або "маркерне відро" [6].

4. Висновки

Судячи з результататів проведеного аналізу, для великих значень H (при високому ступені самоподібності трафіку) потреби в буферній пам'яті починають швидко зростати вже при незначному коефіцієнті використовування мережі. За умови, що число елементів буферної пам'яті повинне бути не менше середнього числа заявок, розмір буфера повинен вибиратися залежно від бажаного коефіцієнта використовування мережі. Якщо бажано мати високий рівень коефіцієнта використовування мережі, потрібно застосовувати методи регулювання або вирівнювання (policing and shaping) інтенсивності самоподібного трафіку, придушувати активність джерел, що перевантажують мережу, та обирати розміри буферів більшими, ніж це витікає з результатів класичного аналізу черг.

Література

1. Столлингс В. Современные компьютерные сети / В. Столлингс. – 2-е изд. – Санкт-Петербург : Питер, 2003. – 783 с.

2. Гнеденко Б. В. Введение в теорию массового обслуживания / Б. В. Гнеденко, И. Н. Коваленко. – Москва : Наука, 1987. – 336 с.

3. Виноградов Н. А. Анализ потенциальных характеристик устройств коммутации и управления сетями новых поколений / Н. А. Виноградов // Зв'язок. – 2004. – №4. – С. 10-17.

4. Tanenbaum A. S. Computer Networks / Andrew S. Tanenbaum, David J. Wetherall. – 5th Ed. – Prentice Hall, Cloth, 2011. – 960 pp.

5. Fillatre L. Sequential Non-Bayesian Network Traffic Flows Anomaly Detection and Isolation // [Електронний ресурс] / L. Fillatre, I. Nikiforov, S. Vaton, P. Casas // – Режим доступу: http://iie.fing.edu.uy/investigacion/grupos/artes/publicaciones/casas_iwap2008_FillatreFinal.pdf

6. Chang Shu, Nick A. Vinogradov. The Method of Adaptive Shaping of the Traffic Flows of Calculating Networks / Chang Shu, Nick A. Vinogradov // Proceedings the Fourth Congress "Aviation in the XXI Century", (Safety in Aviation and Space Technologies), V.1, Kiev, National aviation university, 2010, Sept. 21 – 23. – PP. 18.13-18.16.

Дата надходження в редакцію: 21.04.2015 р. Рецензент: д.т.н., проф. Ю. В. Кравченко