**О. Кузьмін, Л.Журавчак**
Національний університет"Львівська політехніка"

**УДК 681.3.06(075)**

# ОЦІНКА ЧАСУ ВІДГУКУ HTTP-СЛУЖБИ WEB-СЕРВЕРА



***Проаналізовано продуктивність HTTP-служби WEB-сервера для різних конфігурацій при зростанні кількості одночасно працюючих користувачів. Продуктивність визначається часом відгуку на запит до WEB-сервера.***

***In the paper the analysis of productivity HTTP-protocol of WEB-server for different of the configurations will be carried out at increase of quantity of the simultaneously operating users. The productivity is estimated by response time on inquiry to WEB-server.***

## 1. Вступ

*WEB* є системою, яка постійно удосконалюється та розширюється за рахунок появи нових компонентів та їх розвитку. Зростає її популярність, що призводить до збільшення її користувачів. У цих умовах велике значення відіграє час відгуку на запит до *WEB*-сервера, оскільки він значною мірою впливає як на популярність *WEB*-сайта, так і на прибуток, який може отримувати його власник. Тому важливо знати, як може змінюватися час відгуку при зміні робочого навантаження і що є стримуючим фактором для його зменшення. У роботі розглядається методика оцінки продуктивності *HTTP*-служби *WEB*-сервера з використанням методу імітаційного моделювання.

## 2. Методика оцінки

*WEB*-сервер можна навести множиною апаратно-програмних ресурсів, які забезпечують якість обслуговування запитів до відповідних служб. Кожний запит ініціює обчислювальний процес, який виконується *WEB*-сервером. Тому в основу методології імітаційного моделювання покладено поняття обчислювального процесу [1,2], який обслуговується взаємопов'язаними системами масового обслуговування (СМО), якими є апаратно-програмні ресурси *WEB*-сервера. Час відгуку на запит *WEB*-служби буде складатися з часу очікування обслуговування та часу обслуговування окремими СМО. Для оцінки цього часу треба визначити складність обчислювального процесу, яка характеризується кількістю необхідного процесорного часу, адресним простором процесу (кількістю необхідної оперативної пам'яті), кількістю операцій введення – виведення, кількістю даних, які передаються за одну операцію введення – виведення.

Розглянемо процес роботи *WEB*-сервера при отриманні *HTTP*-запиту від користувача:

1. Встановлення з'єднання з користувачем.
2. Пошук інформації за запитом користувача засобами файлової системи.
3. Синтаксичний розбір *HTML* сторінки та передавання її користувачу.

4. Здійснення запитів до баз даних чи виконання скриптів для правильного відображення змісту сторінки.
5. При необхідності завантаження рисунків та інших об'єктів сторінки.
6. Розірвання з'єднання.

Кожен з етапів даного процесу характеризується певними затратами часу, які пов'язані насамперед з використанням обчислювальних ресурсів центрального процесора, пам'яті та файлової системи.

Затрати часу на отримання даних з носіїв інформації файлової системи (дискової підсистеми) в основному пов'язані з їхньою швидкодією і можуть бути отримані з технічної документації. Ці затрати містять час доступу до даних ( пов'язаний з необхідністю пошуку даних ) та саме час передавання, який враховує швидкодію як самого пристрою, так і каналів передачі даних.

Для виміру часу роботи процесора в режимі обслуговування *HTTP*-запитів необхідно протестувати систему на певних визначених *HTML*-документах заданої складності з одночасним веденням протоколу використання процесора протягом етапів обслуговування. З отриманих даних можна знайти середнє значення та відхилення часу обслуговування процесором тестових завдань. Оцінивши складність тестового та спостережуваних завдань, можна пропорційно знайти часові затрати на виконання останніх.

Оцінити складність *HTML*-документа можна, виходячи з того, який обсяг даних необхідно опрацювати. Наприклад, він може містити роботу із синтаксичного розбору та перетворення в пекети *HTML* та витрати на передавання зображень, мультимедіа тощо.

Кількість операцій введення – виведення можна вважати такою, що дорівнює кількості звертань до дискової підсистеми, що містить зчитування опису сторінки, зображень та іншої інформації.

Кількість даних, яка передається за одну операцію введення – виведення, можна знайти, поділивши відповідну кількість даних, яка передається, на кількість звертань.

Важливою характеристикою є також інтенсивність, з якою користувачі звертаються за новою інформацією до сервера.

Дана множина параметрів визначає модель робочого навантаження такого інформаційного ресурсу, як сервер. Кожний з параметрів має свій випадковий характер, який визначається відповідним йому законом розподілу.

### 3. Результати досліджень

Досліджувався вплив на час відгуку *HTTP*-служби зростання кількості користувачів *WEB*-сервера. Розглядались декілька різних конфігурацій *WEB*-сервера: однопроцесорна та двопроцесорна з різними підсистемами введення – виведення. До складу однопроцесорної конфігурації входять два дискові накопичувачі зі швидкістю обертання дисків 5400 об/хв з інтерфейсом під'єднання ATA100 EIDE та 256 Мб оперативної пам'яті.

Сервер використовується для опрацювання запитів користувачів на отримання інформації у вигляді *HTML* сторінок, середній обсяг яких становить 20 Кбайт *HTML*-тексту. Причому середня кількість рисунків, які розміщуються на одній сторінці, до-

рівнює 2, а їх середній обсяг становить 50 Кбайт з середньоквадратичним відхиленням 10% від середнього значення. Кожна сторінка здійснює запит до бази даних, що вимагає в середньому передачу з дискової підсистеми 10 Мбайт службової інформації для отримання результату.

Експериментальним шляхом було встановлено, що час, який необхідний процесору для опрацювання запиту на завантаження однієї *WEB*-сторінки розміром 20 Кбайт з 5-ма рисунками по 20 Кбайт, становить в середньому 50 мс із середньоквадратичним відхиленням 15% від середнього значення [3]. Розподіл інформації за жорсткими дисками є рівномірним. Час, який потрібний на опрацювання запиту до бази даних, становить близько 100 мс.

Користувачі посилають запити на отримання інформації з періодичністю 9 – 11 с та використовують паралельно завантаження декількох сторінок, середня кількість яких – три водночас.

Згідно з попередніми даними параметри імітаційної моделі [4] будуть мати такі значення (дисципліна обслуговування для всіх пристроїв – *FIFO*):

Для джерела заявок :

- Режим роботи оперативний з розподілом часу між надходженням заявок за експоненціальним законом. Для цього знайдемо інтенсивність $\lambda$ потоку заявок, що дорівнюватиме $10/60 = 0.166667$ $c^{-1}$ для 10 заявок за 1 хвилину.
- Пріоритет заявок однаковий, вважаючи, що користувачі мають одинакові права доступу до даного ресурсу.
- Кількість завдань в пакеті має дорівнювати трьом, враховуючи кількість одночасно завантажуваних сторінок.
- Закон розподілу потрібного обсягу оперативної пам'яті приймемо за нормальний, виходячи з обсягу сторінки з зображеннями та додатковими витратами в розмірі 20% від загальної величини для збереження додаткової службової інформації.
- Кількість операцій введення – виведення визначимо як кількість звертань до жорсткого диску з метою завантаження *HTML*-тексту та відповідних зображень. У результаті задамо Пуассонівський закон розподілу з матсподіванням, яке дорівнює 4-м, виходячи із початкових даних.
- Процесорний час опрацювання завдання приймемо за нормальним розподілом з параметрами: матсподіванням $T = 100 + 50(20+2\cdot20)\ /(20+5\cdot20) = 130$ мс середньоквадратичним відхиленням $\sigma = 015\cdot T = 19.5$мс.
- Обсяг даних за одну операцію введення – виведення приймемо таким, що дорівнює загальному обсягу сторінки, поділеному на кількість звертань до диску. Отже, $V = (120 + 10 \cdot 1024)/4 = 2590$ Кбайт. Для нормального закону розподілу задамо це значення як математичне сподівання, а відхилення з умови усереднення вимірювань буде становити 259Кбайт.
- Кількість пристроїв дорівнюватиме двом з однаковими ймовірностями їх вибору, що дорівнюють 0.5.
- Для пам'яті задамо режим функціонування з динамічним розподілом та загальним обсягом 262144 Кбайт.
- Режим роботи процесора без квантування часу.

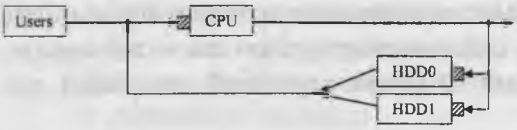


Рис.1. Схема організації однопроцесорного WEB-сервера

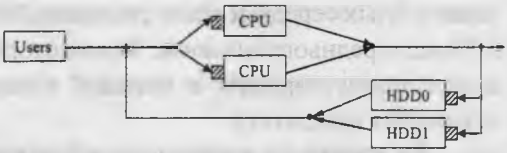


Рис.2. Схема організації двопроцесорного WEB-сервера

Для пристроїв введення – виведення інформації задамо нормальний розподіл часу їх роботи з врахуванням часу доступу до них, що становить 21 мс для заданого класу, виходячи з експериментальних даних тестування, та середньоквадратичним відхиленням 10%, що становить 2.1 мс.

Швидкість передачі каналами введення – виведення аналогічно задаємо нормально розподіленою з параметрами 28000 Кбайт/с та 280 Кбайт/с відповідно.

Топологічні схеми однопроцесорної та двопроцесорної конфігурації *WEB*-сервера наведені на рис.1 і рис.2 відповідно.

Результати проведених досліджень відображені в табл. 1 і 2 та діаграмах 1 і 2 для однопроцесорної та двопроцесорної конфігурації відповідно при збільшенні кількості запитів від тридцяти до трьохсот, що відповідає одночасній роботі від 10 до 100 користувачів з *WEB*-сервером.


Діаграма 1


Діаграма 2

Як видно з отриманих даних, збільшення часу відгуку має експоненціальний характер і швидкість його наростання збільшується з появою нових під'єднань користувачів, що при значній їх кількості робить використання системи неприйнятним для заданої якості обслуговування. Отже, важливо отримати граничне значення кількості запитів, при яких сервер функціонує в нормальному режимі. Вважається, що якість обслуговування є прийнятною, якщо час відгуку не перевищує 4 с. Наближено при наявності в системі 90 користувачів (що відповідає 270 запитів за хвилину ) ми виходимо за граничне значення для системи з дисками, що обертаються із швидкістю 5400 обертів на хвилину.

Перехід на двопроцесорну конфігурацію в незначному ступені зменшує час відгуку на запит *WEB*-сервера при однакових параметрах дискової підсистеми.

*Таблиця 1*

Результати моделювання WEB-сервера

| Кількість користувачів | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|
| Час відгуку для IDE 5400 | 0,833 | 0,906 | 0,98 | 1,1 | 1,247 | 1,428 | 1,781 | 2,279 | 3,731 | 6,621 |
| Час відгуку для IDE 7200 | 0,741 | 0,794 | 0,847 | 0,932 | 1,026 | 1,145 | 1,354 | 1,59 | 2,082 | 3,017 |
| Час відгуку для SCSI 10000 | 0,628 | 0,659 | 0,695 | 0,756 | 0,809 | 0,878 | 0,985 | 1,102 | 1,31 | 1,59 |

*Таблиця 2*

Результати моделювання WEB-сервера на двопроцесорній платформі

| Кількість користувачів | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|
| Час відгуку для IDE 5400 | 0,813 | 0,887 | 0,951 | 1,067 | 1,221 | 1,39 | 1,742 | 2,22 | 3,65 | 6,52 |
| Час відгуку для IDE 7200 | 0,712 | 0,762 | 0,809 | 0,887 | 0,977 | 1,096 | 1,287 | 1,51 | 1,958 | 2,878 |
| Час відгуку для SCSI 10000 | 0,6 | 0,626 | 0,654 | 0,705 | 0,75 | 0,809 | 0,901 | 0,999 | 1,145 | 1,387 |

Проаналізувавши ці результати, можна зробити висновок, що вплив продуктивності процесорного блоку на час відгуку є значно меншим, ніж вплив продуктивності дискової підсистеми.

Тому, щоб підняти продуктивність системи загалом, слід перш за все збільшувати швидкодію пристроїв введення – виведення, що є більш економічно вигідно.

## 4. Висновки

Запропонована методика дослідження параметрів серверних систем, а саме часу відгуку, може бути застосована для аналізу реально функціонуючих рішень з метою отримання відповідей на запитання про граничну продуктивність системи, її запас, максимальну кількість одночасно під'єднаних користувачів.

Отримавши ці дані, адміністратор сервера може прийняти організаційні рішення про оновлення потужностей серверного обладнання відповідно до цілей організації. Крім цього, можна визначити "вузькі" місця існуючого обладнання, вивчити їх взаємовплив та оцінити їх.

Цю методику легко перенести на інші застосування, знайшовши параметри роботи *FTP*-сервера, сервера баз даних, *e-mail* тощо, а також їх суміщення.

Цінність застосування імітації як методу дослідження полягає в тому, що на даний час немає універсальних аналітичних методів досліджень проблем такого типу. Наприклад, теорія СМО може розглядати ці системи тільки частково при експоненційних розподілах характеристик елементів, опираючись на розроблений математичний апарат, а для імітаційних моделей такого обмеження не існує.

1. Кузьмин А.В., Лукащук Л.А. Пакет прикладных программ для имитационного моделирования вычислительных управляющих комплексов (СИМВУК) // Тезисы докладов конференции "Информационно-измерительные системы и точность в приборостроении". – Москва. – 10-11 ноября 1982 г.

2. Кузьмин А.В., Лукащук Л.А. Универсальная имитационная модель вычислительного управляющего комплекса: Сб. "Теория и практика построения информационно-вычислительных систем". – Изд-во СГУ, 1982. – С. 33 – 35.

3. Производительность *WEB*-служб. Анализ, оценка и планирование: Пер. с англ./ Д.Менаске, В.Альмейда. – Спб: ООО "ДиаСофтЮП", 2003. – 480 с.

4. The methodology and software of simulation modeling of computing structures. A.Kuzmin, L.Juravchak. Proceedings of the VIIth International Conference CADSM 2003,"The Experience of Designing and Application of CAD Systems in Microelectronics",IEEE:03 EX618, 18-22 February 2003,Ukraine, pp.421-423.