A Measurement Based Memory Performance Evaluation of High Throughput Servers

Garba Isa Yau and Abdul Waheed
E-mail: {dygarba, awaheed}@ccse.kfupm.edu.sa
Computer Engineering Department
King Fahd University of Petroleum and Minerals
Dhahran 31261, Saudi Arabia

abstract

Performance of high throughput servers largely depends on memory issues, like cache, main and virtual memory, and disk. According to Moore's law, processor speed have been roughly doubling every eighteen months while memory and disk only get faster at about 10% per year. Bottleneck in server performance has shifted from processors to memory and disk. Memory hierarchy performance is a limiting factor in the performance of high throughput servers. In this work, we conduct measurement-based performance study of key high throughput servers namely: streaming media servers, web servers and web proxy servers, and a routing server based on Linux kernel. Our aim is to identify where on-chip and memory becomes bottleneck in the performance of this high throughput servers. We conducted some initial experiments to substantiate on our claim that memory hierarchy is a limiting factor for these high throughput servers. Our initial results are highly motivating

Keywords: streaming media, encoding rate, cache miss, page fault, video on demand, IP forwarding.

1 Introduction

Internet is becoming more pervasive and critical for our day-to-day transaction. More bandwidth is becoming available on the public internet while for organizational intranets, bandwidth is now hardly an issue. While performance bottleneck is gradually shifting away from network bandwidth and the CPU speed is ramming up, the potential bottleneck of most internet servers is likely to be memory and IO. Memory performance is obviously not catching up with CPU [**?**]. While streaming servers and Web servers operate at the network edge, the network core is dominated by switches and routers. We are particularly interested in accessing the memory performance of these network devices: streaming media servers, Web servers and IP forwarding on a general purpose computing platform.

Streaming servers are characterized by high-bandwidth demand and stringent timing-requirement. Memory becomes an issue especially because of the poor temporal locality and large data working set of continuous data. These characteristics of continuous data does not favor the memory hierarchy. In streaming media servers, data is periodically and at constant rate fetched from the disk, buffered in the memory and copied into the network buffer where it is delivered to the client. Strict timing must be maintained to ensure acceptable quality presentation at the client end. In the case of Web servers, strict timing is not an issue, however, high latency will discourage visitors. Apart from timing constrains, streaming and Web servers exhibit similarity in their working mechanism.

Low cost and good performance of general purpose computing platform attracts the implementation of routing and IP forwarding on these platforms [**?**, **?**, **?**, **?**, **?**]. Another advantage of implementation of IP forwarding on general platform is that it is highly flexible and configurable. New features can be added easily and tuning for performance is somewhat easier compared with hardware routers. Hardware routers normally have the TCP/IP protocol stack optimized for the purpose of routing and IP forwarding. However, general purpose software running on these platforms do not have such optimization and most of the issues related to this remain open research questions.

Our work focuses on studying the memory performance of these applications deployed on general purpose computing platforms. We conducted measurement based performance study of representative streaming media servers, Web servers and IP forwarding on Linux platform.

2 Multimedia Streaming Application

3 Web Applications

4 Routing and IP Forwarding

5 Conclusion and Future Work

Acknowledgement The authors would like to thank King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia for supporting this research effort.

References

[1] J. L. Hennessy and D. A. Patterson, *Computer Architecture: A Quantitative Approach*. Morgan Kaufmann Publishers, Inc., 1996.

[2] R. Morris, E. Kohler, J. Jannotti, and M. F. Kaashoek, "The click modular router," *17th ACM Symposium on Operating Systems Principles (SOS '99) - Operating Systems Review*, vol. 34, pp. 217–231, December 1999.

[3] E. Kohler, *The Click Modular Router*. PhD thesis, Massachusetts Institute of Technology, February 2001.

[4] O.-I. Lepe-Aldama and J. Garcia-Vidal, "A performance model of a pc based ip software router," *IEEE ICC*, 2002.

[5] X. Qie, A. Bavier, L. Peterson, Scott, and Karlin, "Scheduling computations on a software-based router," *In Proc. IEEE Joint International Conference on Measurement Modeling of Computer Systems (SIGMETRICS)*, June 2001.

[6] J. R. Hess, D. C. Lee, S. J. Harper, M. T. Jones, and P. M. Athanas, "Implementation and evaluation of a prototype reconfigurable router," *Seventh Annual IEEE Symposium on Field-Programmable Custom Computing Machines*, April 1999.