AASCIT Communications
Volume 3, Issue 4
ISSN: 2375-3803

American Association for Science and Technology

The Performance Trends in Computer System Architecture

E. Kesavulu Reddy Department of Computer Science, Sri Venkateswara University, Tirupati, India

Received: June 19, 2016; **Accepted**: July 20, 2016; **Published**: August 12, 2016

Keywords

Microprocessor, Cloud-Based Services, Networking and Computing, Computer Architecture

Computer architectures have entered a watershed as the quantity of network data generated by user applications exceed the data-processing capacity of any individual computer end-system. Performance evaluation is at the foundation of computer architecture research and development. Contemporary microprocessors are so complex that architects cannot design systems based on intuition and simple models only. Adequate performance evaluation methods are absolutely crucial to steer the research and development process in the right direction. Performance evaluation is non-trivial to multiple aspects to such as picking workloads, selecting an appropriate modeling or simulation approach, running the model and interpreting the results using meaningful metrics. Performance data may drive research and development in a wrong direction. Design needs to reduce costs, save power and increase performance in a multi-scale approach that has potential application from nano scale to data-centre-scale computers.

Introduction

High-performance computers (HPCs) are forecasting to the rise of the data centre, providing our favourite search engine, storage for our family photos and pivot for much corporate, government and academic computing. The modern application presumes network connectivity and increasingly the common application presumes high-performance network connectivity to ensure timely working of each network operation. Cloud usage has driven network demand, yet the cloud itself is not well defined; a cloud may likely be provisioned by a company as a subscription service itself provided as a *loss leader*. Cloud-based services may be presented as one or more of a provision of storage, computing resource or specific services.

In contrast with *Cloud* as offering of services both physical and virtual for the purpose of discussion in this paper, data centre is the physical manifestation of equipment whose purpose is to provide cloud services. The data centre can be provided with modest physical equipment provisioned within a small department machine-room to hyper-data centre's such as that of Google.

The effects of high-performance network applications are reflected in the increasing need for network bandwidth. Cisco's Global Cloud Index [1] for 2013–2018, new technologies, such as the Internet of Things, will further increase the demand on networking, computing and storage resources and thus, growing the associated data to 3.6-fold by 2018 [1]. In addition, global consumer storage requirement is expected to increase to 19.3. Extra byte (annually) by 2018 which suggests current computing architectures do not provide ultimate answers to forthcoming challenges.

While the move to cloud-based computing may improve performance, and reduce operational costs such as power consumption, it does not solve the following challenges: Non- suitability for academic researchers in providing services and applications to users [2]. Requires more servers [3] and cost-effectiveness.

Trends in Computer Architecture

Major Trends Affecting Microprocessor Performance and Design

In a competitive processor, some of the major trends affecting microprocessor performance are:

- Increasing number of Cores
- Clock Speed
- Number of Transistors

Increasing Number of Cores

Multi-core processors are referred to as a single computing component with two or more independent central processing unit called “cores”. The multi-core processor enables users to have boosted performance, improved power consumption and parallel processing that allows multiple tasks to be performed simultaneously. The development of microprocessors for desktops and laptops today is expanding from core i3, core i5, and core i7 presently. This results in using several chips in the CPUs. In the year 2017, it is estimated that embedded processors shall sport 4,096 cores, servers shall have 512 cores and desktop chips shall be using 128 cores.

Clock Speed

Clock speed is defined as the frequency at which a processor executes instructions and/or data is processed. The clock speed is measured in megahertz (MHz) or gigahertz (GHz). It is a quartz crystal which vibrates and sends beats or pulse to each component that is synchronized with it. (PC computer notes, 2003). The speed of microprocessors measured in megahertz (MHz) processes one million instructions per second. Besides that, the microprocessor that runs in gigahertz (GHz), is able to process a billion instructions per second.

In modern technology, most CPU runs in gigahertz range. For instance, a 3GHz Microprocessor and a 3.6GHz is faster than a 500MHz microprocessor as it six times slower. The speed of the computer is fast when the frequency of the microprocessor is higher.

Number of Transistors

The number of transistors available on the microprocessor has a massive effect on the performance of CPU. For instance, in microprocessor 8088, it takes about 15 clock cycles to execute instructions, with this we can assume that on one 16-bit multiplication of the 8088 processors, it takes about 80 cycles.

According to Moore’s Law, the number of transistors on a chip roughly doubles every two years. As a result, the scale gets smaller and transistor counts increases at regular pace to provide improvements in integrated circuit functionalities and performance while decreasing costs.

By increasing the number of transistors, it allows a technology known as pipelining. The execution of instruction overlaps in the pipelined architecture. For instance, it might take five clock cycles to execute each of the instructions; the five instructions may be in different stages for executions and we can deduce that one instruction is completed at every clock cycle. Most modern processors have multiple instruction decoders with its very own pipeline that allows multiple instruction streams, where one instruction is completed at each clock cycle with a lot of transistors used in the microprocessors.

Microprocessor Design Goals for Laptops, Servers, and Desktops and Embedded System

Microprocessor in laptops, servers, desktops varies as they have unique forms varying from each other. Laptop is small and portable; a version that functions as a computer for use anytime and anywhere. The Microprocessor design goal is with emphasize on power consumption. Laptop uses battery power; it would be inconvenient for laptop users to carry the battery adapter wherever they go and thus, the microprocessor in a laptop ensures that it consumes lesser power compared to a desktop computer. Besides that, the processors also help in cooling the laptops as they produce a lot of heat when they are in use which might damage the internal hardware of the laptop. To ensure that the laptops have the required cooling requirement, the processors allow the laptop to lower the clock speed and bus speed. Cooling requirements is also achieved, when the processors makes the laptop to run in a lower operating voltage which also helps in less power consumption.

A server is a computer or device on a network that works together with the network resources. Generally, serves runs24*7 hours to function efficiently in a network and avoid disruption in the server operations may be disastrous than the failure of a desktop computer. The microprocessor design for a server ensures that the server's uptime is stable, always available and reliable to use by having larger cache memory. The cache memory in the server is higher than the desktops and embedded systems. The microprocessor design implemented for servers helps in controlling the heat released; i.e. the microprocessor relative size for a server is 2U (3.5-in thick) or 1U (1.75-in thick) in size and permits the servers to implement large cooling system as it runs 24*7. Whereas, a desktop computer is also personal computer that is used regularly at a single location and it is not portable. The microprocessor design goal for a desktop also ensures that it supports job scheduling and multitasks an operation which helps it performs more than one job at a time. The microprocessor design goal for an embedded system focuses on power consumption. The power consumption of an embedded microprocessor is based on the relative size of the microprocessor; i.e. embedded system uses a very small amount of power which reduces the power consumption of the system. The microprocessor design goal of an embedded system would be memory management through code density; which is the amount of space engaged by executable programs in an embedded system. The microprocessor is aimed to lower the code density.

Optimizing Performance on POWER8 Processor-Based Systems

The optimization performance guidance is organized into three broad categories:

Lightweight Tuning and Optimization Guidelines

Lightweight tuning covers simple prescriptive steps for tuning application performance onPOWER8 processor-based systems. These simple steps can be carried out without detailed knowledge of the internals of the application that is being optimized and usually without modifying the application source code. Simple system utilization and performance tools are used for understanding and improving your application performance.

Deployment Guidelines

Deployment guidelines cover tuning considerations that are related to the: Configuration of a POWER8 processor-based system to deliver the best. This section presents some guidelines and preferred practices. Understanding logical partitions (LPARs), energy management, I/O configurations, and using multi-threaded cores are examples of typical system considerations that can impact application performance.

Deep Performance Optimization Guidelines

Deep performance analysis covers performance tools and general strategies for identifying and fixing application bottlenecks. This type of analysis requires more familiarity with performance tools and analysis techniques, sometimes requiring a deeper understanding of the application internals, and often requiring a more dedicated and lengthy effort.

Another approach that exploits the economies of scale by using commodity components is represented by Rack Scale [3] and the Open Compute project. The Rack Scale architecture is usually referred to by three key concepts [4]: the disaggregation of the compute, memory and storage resources; the use of silicon photonics as a low-latency, high-speed fabric; and, finally, software that combines disaggregated hardware capacity over the fabric to create 'pooled systems'. Rack scale is not well defined, it can refer to a large unit, filling part of a rack; it may also refer to a single rack [5]and it sometimes also refers to a small number of racks [6]. Several commercial products have tried to address the growing computing needs. These machines range in size from one to10 rack units and sometimes contain more than 1000 cores, divided between many small server units.

Limitations of Current-Day Architectures

The computing industry historically relied on increased microprocessor performance as transistor density doubled [7], while power density limits [8] led to multi-processing [9]. Common servers today consist of multiple processors, each consisting of multiple cores, and increasingly a single machine runs a hypervisor to support multiple virtual machines (VMs). A hypervisor provides to each VM an emulation of the resources of a physical computer. Upon each VM, a more typical operating system and application software may operate. The hypervisor allocates each VM memory and processor time. While a hypervisor gives access to other resources, e.g. network and storage, limited guarantees (or constraints) are made on their usage or availability. While VMs are popular, permitting consolidation and increasing the mean utilization of machines, the hypervisor has limited ability to isolate competing resource use or mitigate the impact of usage between VMs

Resource isolation is not the only challenge for scaling computing architectures. General purpose central processing units (CPUs) are not designed to handle the high packet rates of new networks. Doing useful work on a 100 Gbps data stream exceeds the limits of today's processors. This is despite the modern CPU intra-core/cache ring-bus achieving a peak interface rate of3Tbps [9], and a peak aggregate throughput that grows proportionally with the number of cores. A data stream of 100 Gbps, with 64 byte packets, is a packet rate of 148.8M packets per second; thus a 3GHz CPU has only 20 cycles per packet: significantly less than required even just to send or receive. The inefficiency of packet processing by the CPU remains a great challenge, with a current tendency to offload to an accelerator on the network interface itself [10]. On February 11, 2016

In-memory processing and the use of remote direct memory access as the underlying communications system is a growing trend in large-scale computing. Architectures such as scale out non-uniform memory access (NUMA) [12] for rack-scale computers are very sensitive to latency and thus have latency-reducing designs [13]. However, they have limited scalability due to intrinsic physical limitations of the propagation delay among different elements of the system. A fiber used for inter-server connection has a propagation delay of 5 ns/m; thus, within a standard height rack, the propagation delay between the top and bottom rack units is approximately 9 ns, and the round-trip time to fetch remote data is 18 ns.

While for current generation architectures this order of latency is reasonable [12], it indicates scale-out NUMA machines at data-centre scale(with each round-trip taking at least 1μs) are not plausible, as the round-trip latency alone is many magnitudes the time-scale for memory retrieval off local random access memory or the latency contribution of any other element in the system.

Photonics has advanced hand in hand with network-capacity growth. However, photonics has its own limitations [10]: the minimum size for photonic devices is determined by the wavelength of light, e.g. optical waveguides must be larger than one-half of the wavelength of the light in use.

Limitations are faced at several levels in the system hierarchy: from the practical limitations of physics to the increasing *impedance mismatch* between processor clock speed and network data rates.

The Gap Between Networking and Computing

The silicon vendors for both computing and networking devices operate in the same technological ecosystem. CPU manufacturers often had access to the newest fabrication processes and the leading edge of shrinking gate size. Furthermore, in the past 20 years, the interconnect rate of networking devices doubled every 18 months, whereas computing system I/O throughput doubled approximately every 24 months. At the interface between network and processor PCI-Express, the dominant processor-I/O inter connect, the third generation of which was released in 2010, achieves 128 Gbps over 16 serial links. The fourth generation expected in 2016 aims to double this bandwidth. The limitation of existing computinginterconnects vexes major CPU vendors [14].

General purposes processors are extremely complex devices whose traits cannot be limited to specifications such as data path bandwidth or I/O inter connect. Subsequently, we evaluate the performance of CPUs using the Standard Performance Evaluation Corporation (SPEC) CPU2006benchmark [15] and contrast this with the improvement in network-switching devices and computing interconnect.

Conclusion

Many microprocessors are created for different purpose based on system design needs and thus newer systems would get to perform additional tasks which helps ease the user's job. For instance, future embedded systems could increase their capability such as emphasizing on the cache design and bus architecture as well as with the desktops by increasing the performance of the system through development of new embedded microprocessor design to achieve set goals. An architecture that places networking at the centre of the machine require a fusion of knowledge from computer systems, network design, operating systems and applications; yet it will provide for a revolution, solving issues forced on servers by current approaches and architectures.

An important part of the success of novel computing architecture has always been a combination of hardware and software synergetic integrations. System application innovators believe that, irrespective of the absolute correctness of our architecture for every purpose, the innovation opportunities enabled will also have long-lasting consequences and benefits.■

Acknowledgments

I am most thankful to Sri Venkateswara University, Tirupati, Andhra Pradesh, India for providing facilities and support in the research work. And I'm also thankful to my Research Supervisors Prof. M. Padmavathamma for M. Phil and Prof. P. Govindarajulu for PhD, allowing me to do research work independently

Dr. E. Kesavulu Reddy

I am Dr. E. Kesavulu Reddy and work as an Assistant Professor in Dept. of. Computer Science, Sri Venkateswara University College of Commerce Management and Computer Science, Tirupati (AP)-India. My research areas of interest in the field of Computer Science are Elliptic Curve Cryptography-Network Security, Data Mining, and Neural Networks.
E-mail:ekreddysvu2008@gmail.com

References

[1] Cisco. 2014 Cisco global cloud index: forecast and methodology, 2013–2018.

[2] Chen Y, Sion R. 2011 To cloud or not to cloud? musings on costs and viability. In Proc. ACMSymp. on Cloud Computing, Cascais, Portugal, 26–28 October 2011, pp. 29:1–29:7. New York.

[3] Costa P, Ballani H, Narayanan D. 2014 Rethinking the network stack for rack-scale computers. In Proc. HotCloud, Philadelphia, PA, 17–18 June 2014. Berkeley, CA

[4] Intel. 2014 Intel rack scale architecture: faster service delivery and lower TCO. May 2015.

[5] Kyathsandra J, Dahlen E. 2013 Intel Rack Scale architecture overview. In Proc. INTEROP, LasVegas, NV, 2–6 May 2013. San Francisco, CA.

[6] Falsafi B. 2015 Heterogeneous memory and its impact on rack-scale computing. In Proc. 2nd Int. Workshop on Rack-scale Computing, Bordeaux, France, 21 April 2015.

[7] Moore GE. 1965 Cramming more components onto integrated circuits. Electronics 38, 114–117.

[8] Patterson DA, Hennessy JL. 2013 Computer organization and design: the hardware/softwareinterface. London, UK.

[9] Park C, Badeau R, Biro L, Chang J, Singh T, Vash J, Wang B, Wang T. Tbpsonchipring interconnect for 45 nm 8-core enterprise xeon processor. In Proc. IEEE Int. Solid-StateCircuits Conf., San Francisco, CA, 7–11 February 2010, pp. 180–181. Piscataway.

[10] Gallenmüller S, Emmerich P, Wohlfart F, Raumer D, Carle G. Comparison of frameworks for high-performance packet IO. In Proc. ACM/IEEE Symp. on Architectures for networking and Communications Systems, Santa Clara, CA, 17–18 March 2016, pp. 29–38. New York.

[11] Novakovi´c S, Daglis A, Bugnion E, Falsafi B, Grot B. In Proc. Int. Conf. on Architectural Support for Programming Languages and Operating Systems, Salt Lake City, UT,1–5 March 2014, pp. 3–18. New York.

[12] Daglis A, Novakovic S, Bugnion E, Falsafi B, Grot B. 2015 Manycore network interfaces for in-memory rack-scale computing. In Proc. Int. Symp. in Computer Architecture, Portland, OR,13–17 June 2015. New York.

[13] Zilberman N, Watts PM, Rotsos C, Moore AW. 2015 Reconfigurable network systems and software-defined networking. Proc. IEEE 103, 1102–1124.

[14] Intel. 2014 Intel re-architects the fundamental building block for high-performance computing. Intelr%e-architects-the-fundamental-building-block-for-high-performance-computing (accessed January 2015).

[15] Henning JL. 2006 SPEC CPU2006 benhmark descriptions. ACM SIGARCH Comput. Arch. News 34, 1–17.