Memory Systems and Interconnects for Scale-Out Servers

THÈSE N^{O} 6682 (2015)

PRÉSENTÉE LE 18 SEPTEMBRE 2015

À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS
LABORATOIRE D'ARCHITECTURE DE SYSTÈMES PARALLÈLES
PROGRAMME DOCTORAL EN INFORMATIQUE ET COMMUNICATIONS

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Stavros VOLOS

acceptée sur proposition du jury:

Prof. W. Zwaenepoel, président du jury
Prof. B. Falsafi, directeur de thèse
Prof. R. Balasubramonian, rapporteur
Prof. Y. Xie, rapporteur
Prof. P. Ienne, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2015

Abstract

The information revolution of the last decade has been fueled by the digitization of almost all human activities through a wide range of Internet services. The backbone of this information age are scale-out datacenters that need to collect, store, and process massive amounts of data. These datacenters distribute vast datasets across a large number of servers, typically into memory-resident shards so as to maintain strict quality-of-service guarantees.

While data is driving the skyrocketing demands for scale-out servers, processor and memory manufacturers have reached fundamental efficiency limits, no longer able to increase server energy efficiency at a sufficient pace. As a result, energy has emerged as the main obstacle to the scalability of information technology (IT) with huge economic implications.

Delivering sustainable IT calls for a paradigm shift in computer system design. As memory has taken a central role in IT infrastructure, memory-centric architectures are required to fully utilize the IT's costly memory investment. In response, processor architects are resorting to manycore architectures to leverage the abundant request-level parallelism found in data-centric applications. Manycore processors fully utilize available memory resources, thereby increasing IT efficiency by almost an order of magnitude.

Because manycore server chips execute a large number of concurrent requests, they exhibit high incidence of accesses to the last-level-cache for fetching instructions (due to large instruction footprints), and off-chip memory (due to lack of temporal reuse in on-chip caches) for accessing dataset objects. As a result, on-chip interconnects and the memory system are emerging as major performance and energy-efficiency bottlenecks in servers.

Abstract

This thesis seeks to architect on-chip interconnects and memory systems that are tuned for the requirements of memory-centric scale-out servers. By studying a wide range of data-centric applications, we uncover application phenomena common in data-centric applications, and examine their implications on on-chip network and off-chip memory traffic. Finally, we propose specialized on-chip interconnects and memory systems that leverage common traffic characteristics, thereby improving server throughput and energy efficiency.

Key words: cloud, scale-out, datacenters, interconnects, memory systems, DRAM

Contents

1 Introduction

Over the past five decades, information technology (IT) has gone through multiple phases. Driven by the continuous digitization of human activities, IT has transitioned from being compute-centric, to being network-centric, to being data-centric. IT was born in the form of *mainframes* to fulfill the need for number manipulation. The daily interaction between individuals and computers gave rise to *personal computers*, which were instantly made available in enterprises and homes, and digitized new flavors of activities, such as text writing. The need for fast interaction and communication among individuals in enterprises caused a shift in IT toward *networked computers*. A massive paradigm shift in IT followed when computer networks led to the invention of *Internet* due to the desire of individuals to connect and interact with each other across the world. Today, almost all our daily activities have been digitized through a wide range of Internet services, such as online banking, social networking, and video streaming. The information revolution of the last decade has been fueled by the digitization of all kinds of data, granting to individuals ubiquitous access to data and capturing information of value to business and societies.

Technology innovations and advancements in semiconductor fabrication industry have been powering IT with scalable computing platforms for decades. Two powerful paradigms have enabled computing scalability: *Moore's Law* and *Dennard Scaling*. Moore's Law postulates that fabrication advancements enable reduction in transistor size, doubling transistor density

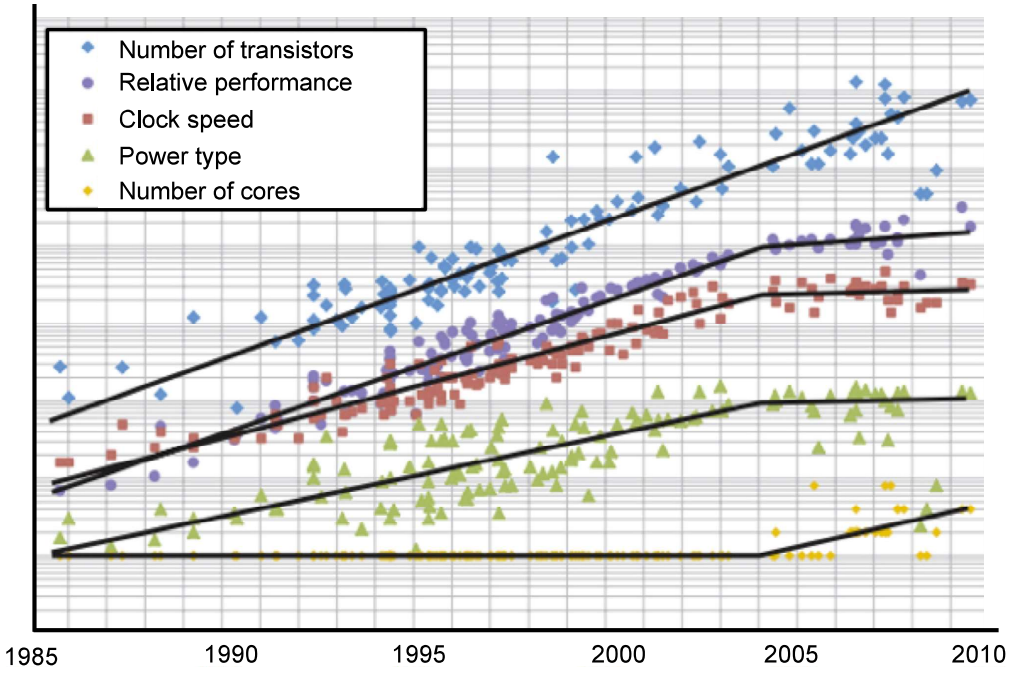


Figure 1.1 – Scaling trends for the transistor count, clock frequency, number of cores, and single-thread performance of processor chips. Source: Graph created by C. Batten based on data from M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten.

for processor and memory chips approximately every two years, while Dennard Scaling states that as transistors get smaller, their power density stays constant due to reduction of chip voltages. As shown in Figure 1.1, smaller and faster transistors have allowed processor designers to scale performance exponentially through higher core clock frequencies and micro-architectural advancements until 2004, and higher core counts for the last decade (2004-today) once frequency scaling became prohibitive due to chip-level cooling, thermal, and power constraints. Moore's Law coupled with Dennard Scaling have enabled an exponential increase in computing energy efficiency, and have been fueling IT with scalable computing platforms for over four decades.

1.1 Data-Centric Information Technology Meets Energy Wall

Data has taken a central stage in our world, driving the skyrocketing demands of the IT sector for computational resources. The backbone of today's IT is large-scale datacenters, which host

a myriad of IT services and consequently collect, store, and process massive amounts of data. State-of-the-art datacenters deployed by technology giants, such as Google and Microsoft, host tens of thousands of servers, and have huge acquisition costs ($100+ M), occupy enormous space (same as a football-sized pitch), and have vast power footprints (5-20 MW). Datacenter energy footprint has been estimated to be at 1.3% of global energy usage [66], and to grow at 20% per year due to a rapid pace of deployment of new datacenters.

The information revolution of the last decade has been accompanied by three IT colliding trends. First, the central role of data in our world has resulted in a rapid increase in data that needs to be collected, stored, and processed. IDC estimates a 300-fold-increase in the size of the digital universe over the span of 15 years, totaling over 40 zetabytes (i.e., over 40 billion terabytes) by 2020 [47]. Second, memory density and bandwidth cannot scale up at a sufficient pace, no longer satisfying the massive memory requirements of data-centric IT in an energy-efficient way [129]. Third, the semiconductor manufacturing industry has reached its fundamental efficiency limits, entering a post-Dennard Scaling Era, where on-chip voltages cannot be scaled down at a sufficient pace [27, 41, 45], no longer being able to increase energy efficiency of computing platforms exponentially.

With the growth of the digital universe outpacing technology scaling, energy is becoming the main scalability bottleneck to IT with huge economic and ecological implications. Based on projections, a ten-fold-increase in datacenter energy efficiency is required in the next decade to make IT sustainable. Achieving this goal, however, will require rethinking datacenter design, calling for innovation across all layers of the data-centric computing stack.

1.2 Toward Specialized Memory-Centric Servers

Datacenter operators rely on *scale-out* architectures to deliver a scalable data-centric computing platform. In essence, scale-out datacenters distribute the vast datasets of IT services across a large number of servers, and typically exhibit a low degree of inter-server communication as servers mostly handle independent requests that do not share any state. IT services rely on

in-memory processing to boost throughput and lower response latency [12, 19, 94]. As a result, DRAM accounts for a significant share of both acquisition and operating costs of datacenters [9, 66, 83]. Maximizing datacenter efficiency calls for architectures that exhibit high memory resource utilization and minimize the overhead to access memory.

Although there has been a shift toward data-centric IT, servers – the heart of datacenters – still employ processor-centric architectures that were proposed in the early stages of compute-centric IT. In these systems, memory along with storage and networking are built around the processor. The mismatch between the requirements of data-centric IT and traditional computer system architectures [29] leads to severe under-utilization of datacenters' memory resources [28, 67, 80], and consequently poor datacenter efficiency [33].

Delivering sustainable IT infrastructure calls for a paradigm shift in computer system design toward *specialized memory-centric* architectures. Specialized memory-centric architectures seek to ensure efficient usage of memory resources by designing the entire computing stack – including processors [37, 78], networking [93], and software – around memory, and to fit the unique characteristics of data-centric applications [28, 29].

1.2.1 What do data-centric workloads need?

Data-centric workloads, or scale-out workloads, exhibit abundant request-level (e.g., online services) and/or data-level parallelism (e.g., analytics) [21, 28]. The existing parallelism is leveraged through multi-threaded software stacks, where incoming requests in online services are assigned to individual worker threads, and datasets in analytics processing are partitioned and processed by multiple processes or threads [28, 65]. While threads are running on individual cores, they need to access instructions and data.

Instructions. These workloads deploy complex and deep software stacks and heavily use third-party libraries, resulting in multi-MB-sized application instruction working sets [2, 28, 29, 30]. Furthermore, the workloads spent significant fraction of their execution time in the operating system, mainly for network activity [28, 29, 73], resulting in even larger instruction

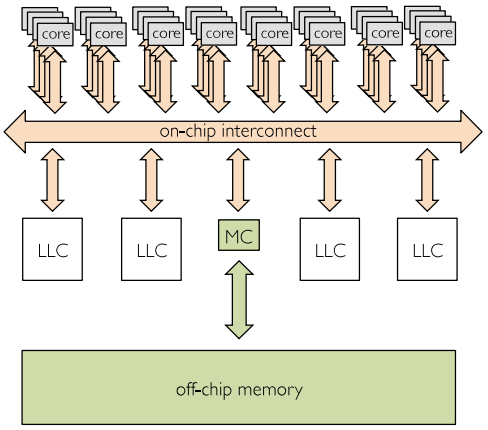


Figure 1.2 – Architecture of a specialized memory-centric server.

working sets. Because of large instruction footprints and their read-only nature, the shared instruction working set is accommodated in the last-level cache (i.e., the last level of the on-chip cache hierarchy). As a result, all cores frequently access the last-level cache (LLC) to fetch instructions.

Data. These workloads need to access vast memory-resident datasets for retrieving request- or task-dependent objects. Due to the disparity between dataset sizes (several tens of GBs) and on-chip cache capacity (few tens of MBs), there is negligible temporal dataset reuse in on-chip caches, resulting in a high incidence of off-chip memory accesses to fetch dataset objects. To allow for constant-time (or sub-linear-time) dataset object retrievals, the datasets are typically organized as pointer-intensive indexing data structures (e.g., a hash table or a tree). Accesses to pointer-intensive data structures, however, result in a limited degree of ILP and MLP within each thread due to high data dependency.

1.2.2 Processor architecture

Specialized processors [37, 78] employ a large number of cores with customized complexity to strike for a balance between available instruction-, memory-, and thread-level parallelism (Figure 1.2). Based on the observation that last-level caches in data-centric workloads exploit mostly instruction-level temporal reuse, specialized processors reduce last-level cache ca-

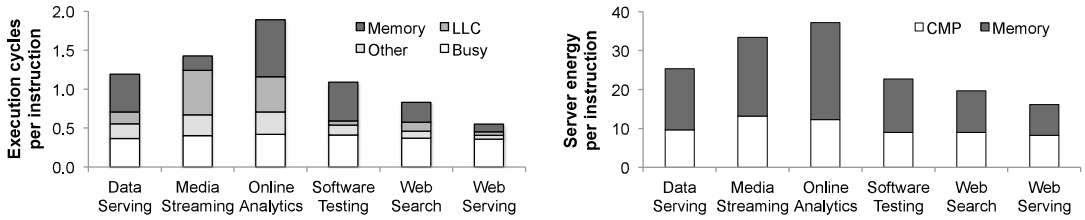
pacity (a few MBs) to free area and power resources in favor of more cores, and to provide fast access to LLC-resident instructions. This specialized processor architecture delivers an order of magnitude higher server throughput over conventional processors while minimizing processor energy consumption in the face of long-latency memory stalls, thereby fully utilizing available memory resources and reducing the energy overhead to access memory.

1.2.3 Efficiency bottlenecks

With specialized processors improving server and datacenter efficiency by almost an order of magnitude [33], the server efficiency bottlenecks are shifting to the memory system, including on-chip interconnects and the on-chip and off-chip memory subsystems. Manycore processors exhibit high incidence of accesses to the last-level-cache (LLC) for fetching instructions, and off-chip memory for fetching dataset objects. Maximizing efficiency calls for on-chip interconnects and memory systems that provide efficient access to LLC-resident instruction footprints and memory-resident datasets.

On-chip interconnects. They serve as the means of communication between cores and the last-level cache. They play a pivotal role in ensuring the performance and power scalability of server manycore chips as they provide the path to performance-critical LLC-resident instructions [28, 29], and communication power is emerging as a significant fraction of the total chip power [35, 118]. Designing an efficient on-chip interconnect is challenging as achieving both low latency and high bandwidth objectives comes at the cost of area/power overheads, prohibitive under fixed area/power budgets.

Multicore processors, featuring 2-16 cores, have relied on conventional crossbar interconnects [112] to achieve uniform and low network latency as well as high bandwidth by connecting each core to all last-level-cache banks. However, as the number of cores (and consequently number of ports) grow, crossbar interconnects face scalability limitations as their area and power footprints scale quadratically with the crossbar radix (i.e., number of crossbar ports) [110], and hence require prohibitive amount of on-chip resources.



(a) Execution time (b) Server energy

Figure 1.3 – Efficiency bottlenecks in memory-centric servers.

Tiled interconnects, referred to as Networks-on-Chip (NoCs), are emerging as the architecture of choice for providing a scalable interconnect for manycore processors [126] by employing packet-switched architectures and regular topologies, and decoupling the number of cores from the router radix. Their power footprint, however, is emerging as a significant obstacle in the quest for efficient manycore chips [35], accounting for as high as 40% of chip power [39, 118], calling for architectures that maximize NoC performance for a given power budget.

Memory System. The memory system plays a key role in IT efficiency as it hosts and provides access to the vast datasets of data-centric applications. With memory capacity driving memory system design, DRAM manufacturing industry has focused on improving DRAM density rather than DRAM efficiency. Over the past decade, DRAM density improved by 16x (256 Mb in 2004 to 4 Gb in 2014) as opposed to other DRAM parameters, such as latency and frequency. For instance, DRAM latency improved only by 50% (60 ns in 2004 to 40 ns in 2014). Due to the ever-increasing reliance of servers on DRAM, memory system is emerging as the major efficiency bottleneck as it has to serve frequent accesses from many cores to DRAM:

- **Latency.** Over the past decades, memory access latency has steadily increased relatively to the computation time, and hence a significant fraction of the server execution time is spent on waiting for memory accesses to be served by main memory (Figure 1.3a).
- **Bandwidth.** The growth rate in core count outpaces bandwidth scaling of conventional memory interfaces, driving designs into a memory bandwidth wall [49, 105, 129]. Conventional interfaces employ parallel buses to connect the processor to a set of dual-inline memory modules (DIMMs). Unfortunately, parallel interfaces exhibit poor signal

integrity which limits bus frequency [129]. Furthermore, the low pin-count scalability limits the number of memory channels integrated on a commodity processor [49]. Thus, high memory capacity requires that multiple DIMMs are deployed per memory channel, degrading signal integrity, and consequently lowering the bus frequency further.

- **Energy.** Memory energy is emerging as a major energy-efficiency bottleneck in servers, accounting for 48-62% of total server energy (Figure 1.3b) primarily due to architectural choices in memory interface design and DRAM organization.
  - High-speed memory interfaces require energy-intensive DIMM-side clock recovery circuits which are kept active [81] regardless of the bus utilization, resulting in high static power consumption.
  - DRAM memory uses a page-based organization, whereby the first access to a page must activate (or open) the page, requiring significant energy. Once a page is open, subsequent accesses to that page are served from the row buffer, avoiding the high energy and latency cost of a page activation. However, inter- and intra-thread contention on row buffer resources in manycore server processors prevent memory systems from fully exploiting row buffer locality. As a result, page activations are a major contributor to memory energy.

1.3 Thesis Goals

This thesis proposes novel on-chip interconnects and memory systems tuned for the requirements of memory-centric scale-out servers.

Thesis Statement

Architecting high-throughput and energy-efficient memory-centric scale-out servers requires tuning their on-chip interconnect and memory system to fit the common traffic access characteristics of data-centric applications.

1.4 Efficient On-Chip Communication

While today's on-chip interconnects are designed to provide low-latency and high-bandwidth core-to-core and core-to-LLC communication, our study of server workloads demonstrates that their on-chip network activity is dominated by core-to-LLC communication consisting of short requests (for instructions and clean data) and associated long responses.

We propose Cache-Coherence Network-on-Chip (CCNoC), a specialized on-chip interconnect to fit the bimodal network traffic characteristics of servers via a pair of asymmetric request and response networks. The networks are tuned for the type of traffic traversing them and differ in their datapath width and router micro-architecture. CCNoC improves on-chip interconnect area/power efficiency and boosts server throughput under fixed area/power budgets.

1.5 Efficient LLC-Memory Communication

Improving memory system efficiency requires amortizing the costly DRAM page activations over multiple row buffer accesses. Although temporal locality at the LLC in scale-out workloads (due to vast datasets and large reuse distances) is scarce, spatial locality is abundant. Our study of scale-out applications shows that these applications commonly operate on coarse-grained objects (e.g., database rows, memory-mapped files) that are accessed through a pointer-intensive indexing data structure (e.g., a hash table, a tree). However, due to absence of information within the memory hierarchy about memory access patterns, last-level caches and memory controllers fail to exploit the coarse-grained memory accesses of scale-out applications.

We propose Bulk Memory Prediction and Streaming (BuMP) to identify accesses to coarse-grained objects, and trigger bulk transfers of coarse-grained objects between processor and off-chip memory. In doing so, BuMP exploits the spatial locality of scale-out applications and leverages the coarse granularity at which off-chip memory is organized, thereby improving server throughput and memory energy efficiency.

1.6 Scalable Off-Chip Memory Systems

Emerging SerDes-connected memory (SCM) can break the pin bandwidth constraints of modern DDR interfaces and provide the required bandwidth, but at a significant power cost and high latency overhead due to large point-to-point memory networks required to host the vast datasets of scale-out workloads.

Our study of scale-out workloads shows that their memory access distributions are skewed, and hence a small portion of memory accounts for bulk of memory activity. This phenomenon primarily originates from the skewed dataset access distribution found in scale-out applications. For instance, a small fraction of popular users and their pictures in image sharing services account for the majority of user activity. However, in real-world setups with memory sizes of 100s of GBs, the hot dataset exceeds the capacity of on-chip and die-stacked caches.

We introduce MeSSOS, a Memory System for Scale-Out Servers. MeSSOS employs SCM modules as a high-bandwidth cache (HBC) in front of conventional DRAM. As the HBC is effective in filtering most of memory accesses, DCM modules can be clocked at low frequency, thus enabling high memory capacity at relatively low static power overhead. Overall, MeSSOS satisfies the required memory bandwidth and capacity requirements of a scale-out server while minimizing the power consumption of underlying memory technologies and interfaces.

1.7 Thesis Contributions

Through a combination of analytic modeling models, trace-driven analysis, and cycle-accurate full-system simulation of manycore servers, we demonstrate:

- **Bimodal on-chip network traffic.** On-chip network traffic in servers consists of short requests for instructions and clean data, and their associated long responses. In particular, 95% of all network messages fall into one of the two categories.

- **Inefficiency of existing on-chip interconnects.** Existing single- and multi-network on-

chip interconnects are sub-optimal as they do not exploit the bimodal network traffic characteristics of servers, leading to incorrect use of available network resources.

- **Efficient on-chip communication.** On-chip interconnect efficiency can be maximized through a pair of asymmetric request and response networks. Each network is optimized for the dominant traffic type, and hence the networks have different datapath width, different buffer architecture, and different pipeline length. Specialization allows for reducing communication delay and area/energy footprints of crossbars and buffers.

- **Bimodal off-chip memory traffic.** Memory accesses in servers occur at fine and coarse granularities. In particular, 59-79% of all memory accesses fall into memory pages with high access density while the majority of remaining access go to low-density pages.

- **Inefficiency of existing memory systems.** Existing memory systems do not exploit the coarse granularity at which memory is accessed and organized due to inter- and intra-core contention on memory resources. State-of-the-art prefetching [113] and scheduled writeback [116] mechanisms can exploit limited row buffer locality as they target only a subset of types of memory accesses. Because row buffer locality is poorly exploited, memory page activations are a major contributor to memory energy.

- **Granularity prediction.** The first access within a page is highly accurate in predicting the granularity at which the page will be accessed for both memory reads (corroborating prior work on spatial footprint prediction [68, 113]) and memory writes.

- **Efficient off-chip communication.** Memory system performance and energy efficiency can be improved by exploiting the coarse-grained memory access patterns. A simple predictor can identify high-density pages and enforce bulk transfers between processor and off-chip memory upon the first access to a page, with minimal on-chip power (50mW) and low storage (14KB) overhead.

- **Dataset popularity.** Dataset popularity in servers is highly skewed, so that a hot portion of memory (5-10%) serves the bulk of memory accesses (65-95%). In real-world setups

with memory sizes of hundreds of GBs, the hot portion of memory considerably exceeds the capacity of on-chip and die-stacked caches.

- **Scalable off-chip memory systems.** Conventional DRAM and emerging SerDes-connected memory show complementary characteristics with regards to capacity, bandwidth, and power consumption. Skewed dataset access distributions can be leveraged for architecting a heterogeneous memory system by employing SerDes-connected memory as a cache in front of conventional DRAM.

- **Die-stacked DRAM caches are obsolete.** There is a disparity between die-stacked cache capacity and the hot dataset size. Due to their inability in exploiting temporal reuse, die-stacked caches provide marginal system efficiency gains when integrated to a system that utilizes emerging SerDes-connected memory modules.

The rest of this thesis is organized as follows. In Chapter 2, we introduce CCNoC, a specialized on-chip interconnect for efficient core-LLC communication. In Chapter 3, we present BuMP, a low-cost enhancement to the on-chip memory system for efficient LLC-memory communication. In Chapter 4, we present MeSSOS, a scalable off-chip memory system organization for satisfying the required memory bandwidth and capacity of a scale-out server. Finally, we discuss related work in Chapter 5, and conclude in Chapter 6.