TITLE:

Parallel Memory System Architectures for Packet Processing in Network Virtualization(Digest_要約)

AUTHOR(S):

Korikawa, Tomohiro

CITATION:

Korikawa, Tomohiro. Parallel Memory System Architectures for Packet Processing in Network Virtualization. 京都大学, 2021, 博士(情報学)

ISSUE DATE:

2021-03-23

URL:

https://doi.org/10.14989/doctor.k23326

RIGHT:

学位規則第9条第2項により要約公開

京都大学学術情報リポジトリ
KURENAI 紅
Kyoto University Research Information Repository

学位論文内容の要約

●氏名：郡川　智洋

●論文題目：Parallel Memory System Architectures for Packet Processing in Network Virtualization
（ネットワーク仮想化におけるパケット処理のための並列メモリシステムアーキテクチャ）

Network virtualization aims to reduce the capital expenditure and the operating expenditure of network infrastructure by leveraging commercial off-the-shelf hardware such as general-purpose computers and virtual network functions instead of dedicated network equipment. Network functions comprise various packet processing tasks such as classification, searching, modification, and queueing, each of which issues memory accesses and requires high memory performance. Conventional network equipment comprises purpose-built, dedicated components such as processors, memory devices, and bus architecture to satisfy the specifications and requirements of network services. Large-scale service providers such as telecom operators have depended on conventional equipment to satisfy the service level agreement and to accommodate various and a large amount of traffic from subscribers, which prevents the service providers from benefiting from network virtualization. This thesis studies four problems about parallel memory system architectures for packet processing in network virtualization. Each problem corresponds to the main memory parallelism, integration of on-chip cache memories of the CPU with the parallel main memory, capacity and parallelism of the on-chip cache memories in the presence of parallel main memory, and accumulated latency of data transfers between processors and memories when there are multiple packet processing tasks with memory accesses, respectively.

Firstly, this thesis proposes a memory system architecture that uses a three-dimensional (3D)-stacked memory to increase the main memory parallelism. In current general-purpose computers such as servers based on x86 central processing unit (CPU), the main memory parallelism is much less than the number of CPU cores, which limits packet processing performance in network virtualization. This work augments main memory parallelism by leveraging both channel-level parallelism and bank interleaving of a 3D-stacked dynamic random access memory (DRAM). In the 3D-stacked DRAM, a database for packet processing is split into partial databases, each of which is allocated to each set of memory channel and bank. A hash-function-based distributor distributes incoming memory requests to an appropriate memory channel-bank set that has the corresponding partial database for the requests. This work introduces an analytical model of the proposed architecture for two traffic patterns, one with random memory request arrivals and one with bursty arrivals. The numerical results observe that the proposed memory system architecture increases packet processing performance up to around 80 Gbps for the smallest-sized Internet Protocol packets involving random and bursty memory request arrivals.

Secondly, this thesis proposes a memory system architecture that integrates on-chip private cache memories with the off-chip 3D-stacked memory to reduce memory access latency in the existence of main memory parallelism. In general-purpose computers, CPUs usually have several levels of on-chip cache memories to obscure the main memory latency. The on-chip cache memories comprise private cache memories that belong to each CPU core and the last-level-cache (LLC) that is shared among all the CPU cores. The proposed architecture integrates the private

cache memories of each CPU core with the 3D-stacked DRAM-based main memory. This work explores the integration of the cache memories with the 3D-stacked DRAM, considering two reference architectures, one with private cache memories and shared LLC and one without any cache memories. The results observe that the proposed architecture reduces latency by 58 % and 1.8 % and increases throughput by 104 % and 1 % with reducing the blocking probability by 91 % and 18 %, compared to the reference architectures with private cache memories and shared LLC and that without any cache memories, respectively.

Thirdly, this thesis proposes a memory system architecture that uses the 3D-stacked memory, the on-chip private cache memories, and on-chip LLC slices to increase capacity and parallelism of the on-chip cache memories in the integration with the off-chip parallel main memory. The on-chip shared LLC in the latest CPU comprises multiple LLC slices, each of which belongs to one of the CPU cores and can be accessed from each CPU core via a mesh or ring bus. The proposed architecture integrates the LLC slices with the on-chip private cache memories and the off-chip 3D-stacked DRAM. The cached data is distributed to each LLC slices according to a memory address-based hash function so that CPU cores can access the LLC slices in parallel. This work analyzes the memory performance dependency on the number of assigned LLC slices and CPU cores. The results observe that the proposed architecture reduces latency by 62 % and 12 % and increases throughput by 108 % and 2 % with reducing the blocking probabilities by 96 % and 50 %, compared to the reference architectures with private cache memories and shared LLC and that with private cache memories and without shared LLC, respectively.

Fourthly, this thesis proposes a memory system architecture that uses a network of 3D-stacked memories to increase throughput and reduce accumulated latency of data transfers between processors and memories when there are multiple packet processing tasks with memory accesses. Packets that enter the memory network receive packet processing without data transfers between the memories and the processors until the packet processing is completed. Packet processing task is allocated in the user-defined programmable logic in the logic layer of each 3D-stacked DRAM. The results observe that the proposed architecture increases throughput and reduces the accumulated latency when there are multiple packet processing tasks, compared to the architecture with 3D-stacked DRAM-based parallel main memory, where every memory access requires data transfers between the processors and memories.

This thesis is organized as follows. Chapter 1 introduces the background of packet processing, computer architectures of dedicated equipment and general-purpose computers, and major hardware devices in computers. Chapter 2 describes related works. Chapter 3 presents the parallel memory system architecture using the interleaved 3D-stacked memory. Chapter 4 presents the parallel memory system architecture using the interleaved 3D-stacked memory and the on-chip private cache memories. Chapter 5 presents the parallel memory system architecture using the interleaved 3D-stacked memory, the on-chip private cache memories, and the on-chip LLC slices. Chapter 6 presents the memory system architecture using the network of interleaved 3D-stacked memories. Finally, chapter 7 concludes this thesis.