Title: Deadline-Aware Shared Memory Management in Heterogeneous Accelerator Systems

Author: Ayushi Agarwal (2018ANZ8503)
Amar Nath and Shashi Khosla School of Information Technology, IIT Delhi

Abstract:

Integrating domain-specific hardware accelerators (HWA) on modern heterogeneous systems-on-chip (SoCs) has enabled the deployment of complex applications, such as vision, natural language processing, autonomous driving, and augmented reality, in small form factors. The high memory bandwidth requirements and strict deadlines imposed by hardware accelerators pose challenges in their integration into the system's memory hierarchy. The system-level shared cache, or the last-level cache (LLC), is a critical resource shared among multi-core processors, GPUs, and hardware accelerators in heterogeneous systems. It significantly reduces the bottleneck at the off-chip memory and delivers high performance. With the integration of accelerators on the shared cache gaining momentum, the strict Quality of Service (QoS) requirements of accelerators, or deadlines, can lead to severe performance degradation of processor cores. Thus, managing the shared cache efficiently between cores and accelerators becomes crucial. Off-chip memory management between processor cores and accelerators with strict QoS requirements, or deadlines in CPU-HWA systems, has been explored by prior works. Given the architectural differences between DRAM and cache systems, the off-chip memory management strategies cannot be extended to the shared cache. While prior works have addressed shared cache space and bandwidth management for multiprocessor cores, the management of shared cache between cores and accelerators remains largely unexplored.

We characterize Internet Protocol Security (IPsec), a high-throughput application, by collecting memory traces of this application running on the accelerators and CPU cores of the NXP LX2160A SoC. We utilize this characterization to design a simulation infrastructure for simulating IPsec on potential domain-specific architectural extensions and conduct a design-space exploration across various general-purpose memory management policies. We propose and evaluate APPAMM, an application-specific predictive packet-aware memory management policy using the knowledge of IPsec to improve performance for next-generation SoCs.

To overcome the limitations of state-of-the-art heterogeneous full-system simulators, we integrate cycle-accurate memory access traces from detailed accelerator simulators into gem5, enabling the simulation of various accelerators, such as Cryptographic, Ethernet I/O, and Machine Learning (AI/ML), alongside processor cores. We

identify a new problem of dynamically partitioning the shared last-level cache bandwidth between processor cores and accelerators. State-of-the-art shared cache controllers do not support optimizing the system performance with the accelerator's deadline constraints. We propose and evaluate a novel cache request arbitration and scheduling policy, FLASH, which efficiently exploits deadline-awareness and the dynamic progress of the accelerator to vary the shared cache bandwidth allocation between processor cores and accelerators to meet the deadline given for the accelerator while minimizing the impact on the performance of cores. The policy takes into consideration the shared cache access characteristics of applications to conditionally schedule or defer requests from cores at the shared cache.

State-of-the-art cache management techniques employ reuse-aware bypassing of accesses from cores, utilizing reuse predictors to enhance performance. In our subsequent work, we demonstrate that the reuse analysis performed on processor applications does not necessarily hold for accelerators, and show that the state-of-the-art reuse predictors perform sub-optimally for accelerators. We propose a novel clustering-based methodology, LERN, for learning and predicting the reuse behavior of hardware accelerators at the shared cache. We then propose and evaluate a deadline and reuse-aware cache management strategy, HyDRA, which explores a novel tradeoff between reuse and deadline awareness for performance efficiency. It utilizes LERN to dynamically predict the reuse behavior of accelerator accesses and make bypass decisions, thereby improving system throughput while meeting accelerator deadlines. By controlling the bypass aggressiveness according to the deadline, HyDRA achieves a balance between deadline and reuse awareness.

While QoS-aware and performance-oriented cache management improves system performance and meets accelerator deadlines, it inadvertently increases off-chip data movements, thereby impacting off-chip memory energy consumption. Since off-chip memory energy contributes disproportionately to the total system energy, efficient shared cache management is crucial for minimizing off-chip memory energy without sacrificing performance and deadlines, but remains unaddressed by state-of-the-art policies. We introduce an unexplored tradeoff between reuse and energy awareness in the presence of deadlines. A deadline- and reuse-aware policy might aggressively bypass accelerator accesses, potentially worsening the dynamic off-chip memory energy. However, an energy-aware policy attempts to reduce bypass aggressiveness, sensing high off-chip memory contention. We propose and evaluate a novel deadline and energy-aware approach, HADES, to minimize off-chip memory energy by utilizing application characteristics and variations in the off-chip memory access volume to balance dynamic scheduling and bypass aggressiveness of accelerators at the shared cache.

Viva Details:

Date: Wednesday, 03 June 2026
Time: 11:30 – 12:30 PM
Location: ANSK School of IT, SIT 001 (Ground Floor)