MOCHA: Modelling and Optimising Complex Heterogeneous Architectures

Shuai Zhao, Xiaotian Dai, Wanli Chang and Iain Bate
University of York, UK

January 29, 2020

Abstract

Emerging computing applications impose higher performance demand for hardware architectures. As both the high performance and low latency have to be satisfied, traditional Commercial Off-The-Shelf (COTS) architectures are often insufficient due to the inefficiency and inflexibility to apply them in implementation of performance critical systems.

Recently, high-performance heterogeneous architectures, such as Multi Processor Systems-on-Chip (MPSoC), have attracted significant research attention due to its computation capability and its widespread adoption by industry. However, the use of MPSoC inevitability increases the complexity of software design and poses challenges in modelling, scheduling and analysis of the system. In addition, timing-critical systems that adopt MPSoCs impose a higher demand for real-time systems, making its applications and the underlying resources (e.g., processors, networks and memory) ever more complex to control and analyse. In a typical heterogeneous system, challenges include meeting deadlines and providing timing accuracy. Timing accuracy is defined here as the degree of variability with which an action is completed. The complexity is introduced by shared resources at the architectural level.

For example, one major challenge is to model and analyse caches that are shared by multi-cores. Considering the execution of transaction process that could be scheduled and executed on a number of cores, which creates data and task dependencies. In this case, there are clear trade-offs between each transaction running on a single core which might reduce data cache misses but increase the instruction cache misses, or alternatively running each function of the transaction that is statically assigned to an individual core, in which case increases the cache misses but potentially reduces the instruction cache misses.

The MOCHA (Modelling and Optimising of Complex and Heterogeneous Architectures) project focuses at complex heterogeneous systems (e.g., MP-

SoCs) and aims to address the challenges brought by the complexity of such architectures in terms of system scheduling, task allocation, memory management, etc., and delivers novel scheduling and simulation methods that are deployed on a digital twin with timing guarantee considered.

Our vision to the solution are multi-fold. First, we propose a multi DAG dynamic scheduling for generic heterogeneous systems, to improve system schedulability, compared to traditional real-time scheduling strategies. The proposed scheduling algorithm is based on feedback-based scheduling to provide flexibility and favour critical tasks during execution. In addition, proposed scheduling is cache-aware and can reduce cache miss rates (for both instruction and data) by precisely allocating processing tasks to a set of distributed and heterogeneous processors. Timing accuracy is also considered favouring certain critical tasks, in the context of the mixed-criticality system. Then, a novel memory management model is presented to provide safe and cost efficient memory access, along with techniques such as cache locking and scratchpad to further improve system performance. At last, we present a configurable simulator of generic heterogeneous architectures as the digital twinning of complex systems. The simulator can facilitate the development of novel scheduling algorithms and resource management strategies, and can also be used as a test-bed to evaluate the performance of various task allocation methods, scheduling algorithms and memory management models.