Adaptive Heterogeneous Throttling for On-Chip Networks


Kevin Chang


Chris Fallin


Rachata Ausavarungnirun


Onur Mutlu

The network-on-chip (NoC) is a primary shared resource in a chip multiprocessor (CMP) system. As core counts continue to increase and applications become increasingly data-intensive, the network load will also increase, leading to more congestion in the network. This network congestion can degrade system performance if the network load is not appropriately controlled. Prior works have proposed source-throttling congestion control, which limits the rate at which new network traffic (packets) enters the NoC in order to reduce congestion and improve performance. These prior congestion control mechanisms have shortcomings that significantly limit their performance: either 1) they are not application-aware, but rather throttle all applications equally regardless of applications' sensitivity to latency, or 2) they are not network-load- aware, throttling according to application characteristics but sometimes under- or over-throttling the cores.

In this work, we propose Adaptive Heterogeneous Throttling, or AHT, a new source-throttling congestion control mechanism based on two key principles: application-aware throttling and network-load-aware throttling rate adjustment. First, we observe that only network-bandwidth-intensive applications (those which use the network most heavily) should be throttled, allowing the other latency-sensitive applications to make faster progress without as much interference. Second, we observe that the throttling rate which yields the best performance varies between workloads; a single, static, throttling rate under- throttles some workloads while over-throttling others. Hence, the throttling mechanism should observe network load dynamically and adjust its throttling rate accordingly. While some past works have also used a closed-loop control approach, none have been application-aware. AHT is the first mechanism to combine application-awareness and network-load-aware throttling rate adjustment to address congestion in a NoC.

We evaluate AHT using a wide variety of multiprogrammed workloads on several NoC-based CMP systems with 16-, 64-, 144-cores and compare its performance to two state-of-the-art congestion-control mechanisms. Our evaluations show that AHT consistently provides higher system performance and fairness than prior congestion-control mechanisms.

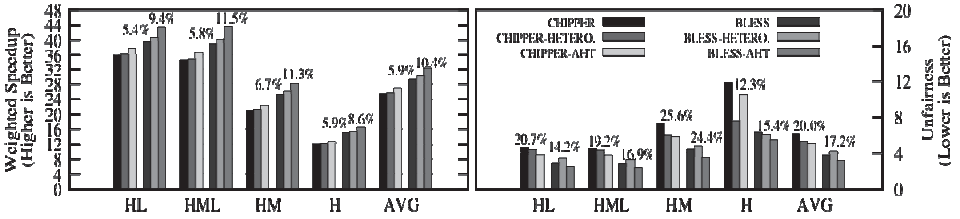

Fig. 1. Performance and fairness of AHT vs. common baselines.