Reducing Response Time with Preheated Caches

Dr. Frank Bellosa
Karlsruhe Institute of Technology, Germany

11/7/2016 at 11:00 am
CoRE 305

Abstract

CPU performance is increasingly limited by thermal dissipation, and soon aggressive power management will be beneficial for performance. Especially, temporarily idle parts of the chip (including the caches) should be power-gated in order to reduce leakage power. Current CPUs already lose their cache state whenever the CPU is idle for extended periods of time, which causes a performance loss when execution is resumed, due to the high number of cache misses when the working set is fetched from external memory. In a server system, the first network request during this period suffers from increased response time. We present a technique to reduce this overhead by preheating the caches in advance before the network request arrives at the server: Our design predicts the working set of the server application by analyzing the cache contents after similar requests have been processed. As soon as an estimate of the working set is available, a predictable network architecture starts to announce future incoming network packets to the server, which then loads the predicted working set into the cache. Our experiments show that, if this preheating step is complete when the network packet arrives, the response time overhead is reduced by an average of 80%.

Faculty Host: Uli Kremer