Scalable Computing Software Laboratory Technical Report
Department of Computer Science
Illinois Institute of Technology

LPM: Layered Performance Matching in Memory

Hierarchy

Yu-Hang Liu
Department of Computer Science
Illinois Institute of Technology
ylin242@iit.edu

Xian-He Sun
Department of Computer Science
Illinois Institute of Technology
sun@iit.edu

August 01, 2014

Technical Report No. IIT/CS-SCS2014-08

http://www.cs.iit.edu
10 West 31st Street, Chicago, IL 60616

LIMITED DISTRIBUTION NOTICE: This report has been submitted for publication outside of IIT-SCS and will probably be copyrighted if accepted for publication. It has been issued as a Technical Report for early dissemination of its contents. In view of the transfer of copyright to the outside publisher, its distribution outside of IIT-SCS prior to publication should be limited to peer communications and specific requests. After outside publication, requests should be filed only by reprints or legally obtained copies of the article (e.g. payment of royalties).
ABSTRACT
Data access has become the preeminent performance bottleneck of computing. In this study, we propose a Layered Performance Matching (LPM) approach to match the memory performance with the computing performance. The basic idea of LPM is that the performance of each layer of a memory hierarchy should be optimized and can be optimized to match the request of its adjacent higher layer as closely as possible. The LPM model simultaneously considers both memory concurrency and memory locality. It reveals the fact that increasing the effective overlapping between hits and misses of the higher layer will ease the performance impact of the lower layer, where the terms “pure miss” and “pure miss penalty” are introduced to measure the “effectiveness” of such hit-miss overlapping. By discriminating common miss and pure miss, LPM has reshaped the conventional computer hardware and software design principle of “locality is always good” and calls for a joint consideration of locality and concurrency. We first introduce the motivation and concept of the LPM model. Next, the formulation and measurement of layered matching are derived. Then, layered performance optimization algorithms are presented to carry LPM automatically. Finally, case studies are provided to illustrate and confirm the correctness and practical value of the newly proposed LPM approach. Our evaluation shows that memory system performances can be improved up to 230 times with optimized layered performance matching. Without changing the hardware configurations, by simply scheduling data allocation following the matching principle in heterogeneous multicore systems, we also have achieved more than 20% performance improvement in our case studies. Analysis and experimental results show LPM is feasible and provides a novel and efficient way to cope with the ever severe memory wall problem, the ever complex memory system, and to optimize the ever vital memory performance.

Categories and Subject Descriptors
D.3.3 [Computer Systems Organization]: Performance of Systems

General Terms
Performance, Algorithms, Theory, Measurement, Verification

Keywords
Memory wall, memory stall time, layered performance matching (LPM), memory concurrency, concurrent average memory access time (C-AMAT)