Decoupled Compressed Cache: Exploiting Spatial Locality for Energy-Optimized Compressed Caching

Somayeh Sardashti and Professor David A. Wood
University of Wisconsin-Madison

Optimizing Memory Hierarchy for Energy

Maximize LLC effective capacity to reduce system energy!

Access to main memory vs. LLC:
6X Longer Latency
60X Higher Energy Cost

Why not double the LLC?
15%-30% of on-chip area
2X LLC Area

Exploiting Spatial Locality

Most workloads exhibit spatial locality.

Limited number of tags
Internal fragmentation
Energy-expensive compactions

Higher effective cache size
Low area overhead
Higher system performance
Lower system energy

Decoupled Compressed Cache (DCC)

DCC

Decoupled
Super-Blocks

Non-contiguous
Sub-Blocks

(Đ-Co)DCC

• DCC exploits spatial locality to improve compression effectiveness:
  • Uses decoupled super-blocking to track more blocks with low area overhead.
  • Compresses and allocates a block into non-contiguous data sub-blocks.
• (Đ-Co)(Compressed DCC):
  • Co-compacting blocks of a super-block to reduce internal fragmentation.

DCC Implementation

☑️ We integrate (Đ-Co)DCC with AMD Bulldozer LLC.
  • No need for an alignment network

☑️ We implement the tag match and the sub-selection logic in Verilog.
  • No need for an alignment network

Experimental Methodology

• We model a multicore system with GEMS.
• We use workloads from Commercial workloads, SPEC OMP, PARSEC, and SPEC CPU2006.
• We use Cacti to measure (Đ-Co)DCC power and area.

Evaluation

(Co)-DCC:
• Performs better than a conventional LLC of twice the capacity.
• Boosts system performance by 14% on average (up to 38%).
• Saves system energy by 12% on average (up to 39%).

Results

Normalized LLC Area

Normalized Effective LLC Capacity

Normalized System Energy

Normalized Runtime

Normalized LAT

Normalized Power

Normalized Energy

SPEC OMP

PARSEC

SPEC CPU2006

apache

jbb

oltp

zeus

ammp

applu

equake

mgrid

wupwise

black

canneal

freqmine

m1

m2

m3

m4

m5

m6

m7

m8

GEOMEAN

Intel Nehalem

Cores
Eight OOO cores, 3.2 GHz

L1S/L1D$
Private, 32-KB, 8-way

L2$
Private, 256-KB, 8-way

L3$
Shared, 8-MB, 16-way, 8 banks

Main Memory
4GB, 16 Banks, 800 MHz bus frequency DDR3

Access to main memory vs. LLC:
6X Longer Latency
60X Higher Energy Cost

Potential 3.9X larger LLC

VSC-2X achieves only 1.6X LLC effective capacity

Compaction LLC dynamic energy

Why not double the LLC?
15%-30% of on-chip area
2X LLC Area