Approaches on Improving Energy Efficiency in the Processor/Memory Hierarchy

Aditya Deshpande, Jeffrey Draper {adeshpan, draper} @ ISI.EDU
Information Sciences Institute :: Ming Hsieh Department of Electrical Engineering :: University of Southern California, Los Angeles, CA

Introduction

- Energy-efficiency is becoming a key metric in the design and development of systems
- On-chip caches form the back-bone of memory hierarchy and are impacted by the interplay between many factors — application, operating system, hardware configuration, etc.
- Data movement in the memory hierarchy identified as a major source of energy dissipation
- On-chip caches also occupy large chip real-estate and significant share of total leakage energy losses

Problems in Current Memory Hierarchy

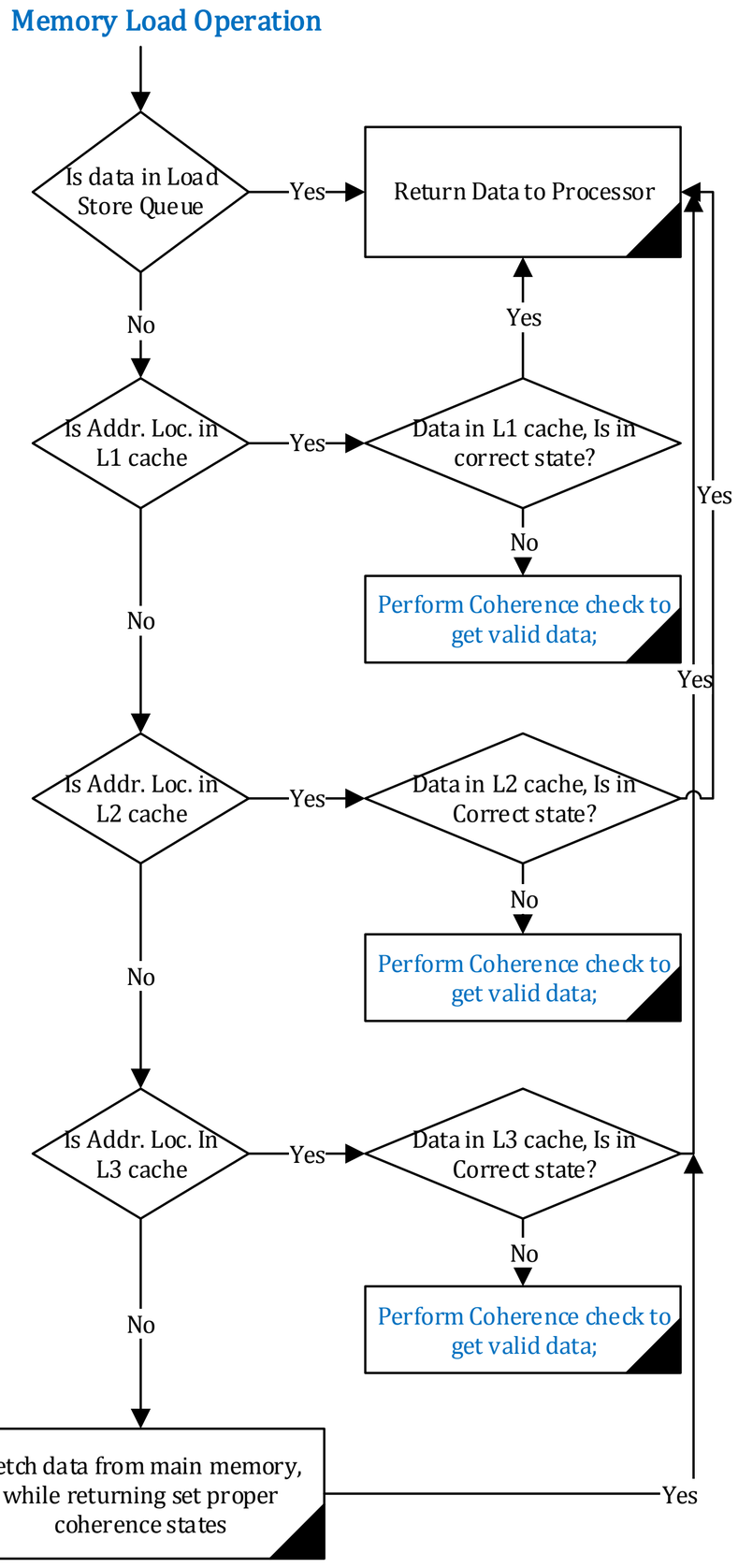
- Majority of advancements in the memory hierarchy have emphasized improving performance while neglecting energy efficiency in the system
- Current memory hierarchy is rigid, requiring applications to adapt to its configurations
- Memory hierarchy behavior is dictated by some obvious and some non-obvious parameters —

  Obvious: capacity, associativity, block-size, access-time, latency, cache-replacement policy, energy per access

  Non–Obvious: number of MSHRs per cache level, bits used for set -way associativity, data-transfer bus width, data-buffers

With energy-efficient computing gaining attention, the memory subsystem, which includes large on-chip caches and consumes a significant amount of overall energy, should be able to more effectively adapt to improve energy efficiency and in-turn performance during application execution

Data Movement in Memory Hierarchy



Some Ideas to Improve Energy Efficiency in Caches

Minimizing Data Movement in Memory Hierarchy

- In multicore processors overheads from managing coherence for the shared data generates additional interconnection network traffic in the hierarchy. With increasing caches sizes and poor scaling of wires, this "overhead" data movement consumes large energy.
- **Challenges:** Effectively managing coherence in an energy-constrained manner would require better models to measure traffic and movement characteristics.

Power Management of Memory Hierarchy

- Like DVFS in processors, the cache hierarchy could be designed to implement voltage and frequency scaling providing similar trade-offs between energy and cache-performance
- **Challenges:** Designing a meaningful granularity of voltage and frequency scaling without compromising storage cell stability and cache footprint

Reconfigurable Memory Hierarchy

- Memory hierarchy should adapt to application needs rather than forcing applications to tune to available resources. Hardware needs to be proactive in managing resources to software/application demands.
- **Challenges:** Ability to turn On/Off or reconfigure caches would require additional resources, and its management can potentially be a substantial overhead depending on the granularity of re-configurability

Some Observations

- With memory hierarchy consisting of multiple levels of caches, each level has strong impact on other levels in the hierarchy
- With multicore computing becoming the norm, maintaining correct data state across the hierarchy results in large amount of control messages passing through the interconnection fabric consuming energy
- The problem of data movement in the memory hierarchy is expected to exacerbate with increase in #cores, #levels and better parallel applications

Summary

- Memory hierarchy has been designed with performance in mind; it is time to rethink the structure of memory hierarchy with energy efficiency and data movement in mind
- Better utilization of available on-chip resources and flexibility/ adaptability towards application behavior would lead to higher energy efficiency and potentially higher performance