Optimizing Replication, Communication, and Capacity Allocation in CMPs
Zeshan Chishti, Michael D. Powell, and T. N. Vijaykumar

Presented by: Siddhesh Mhambrey
Wednesday 11/5/08

Summary by: Srivatsan Chellappa

Brief outline of the paper:
CMPs have the advantage of reduced latency over SMPs. However they face the problem of having limited on-chip memories. The overall performance of the CMPs can be improved if the available on-chip memory is utilized effectively. This paper explores techniques for the effective utilization of the memory available on CMPs such as
1) Placing read-only copies of the data close to the requesters for faster access.
2) Controlled replication to reduce capacity pressure by eliminating read only copies.
3) In-situ communication is introduced to avoid coherence miss of read-write-shared data.
4) Capacity stealing, i.e placing the private data that exceeds core capacity in neighbouring caches with lesser memory demand.
For this the authors have simulated a hybrid cache that implements these techniques, called NuRAPID (Non-Uniform Access with Replacement and Placement using Distance associativity).

Summary of the strengths of the paper as illustrated by the presentation & critiques:
1) All the techniques discussed do definitely help to improve efficiency of the CMPs by better management of the available on chip memory. This has been proven by the extensive simulations conducted by the authors over different benchmarks which show that these techniques out perform uniform shared caches by 13%.
2) The proposed policies are explained in detailed with methods to implement them like the modifications needed for the cache coherence protocol and the creation of C state which allows multiple processors to share the dirty block.
3) The new architecture proposed tries to implement all the suggested policies. Capacity stealing greatly exploits the low latency provided by CMPs to improve efficiency.

Summary of the Weaknesses of the paper as illustrated by the presentation & critiques:
1) A major drawback of the paper is that the NuRAPID architecture make extensive use of distance associatively, which works well for the simulated 4 core CMP but does not scale with the number of cores. The techniques may not be useful for higher cores or might produce diminishing returns. This has not been discussed at all in the paper.
2) The major challenge for current microprocessors and CMPs is power and heating. The paper makes no effort to study the thermal effects of their techniques.
3) The on-chip bus traffic is bound to increase because of these policies and this might cause bus arbitration stalls. The paper does not look at this major factor of latency.
4) The system designer now has a new problem of forming the policy for distance replacement along with the conventional data replacement policy. The implementation of promotion and demotion policies might actually take more time than the actual access from the remote cache.

Conclusion:
The Authors have put in laudable efforts to propose three ideas for the effective utilization of the CMP memory capacity which is shown to produce a 13% performance improvement for 4 core caches. A few issues like power, scalability and complexity could have been better addressed to make it a truly novel paper.