Computer Architecture Reading Group Notes

Date: 4/15/04
Discussion Leader: Samir Chopra
Notes: Njuguna Njoroge
Topic: Transactions

Paper
Hammond et. al. Transactional Memory and Coherence to appear in ISCA 2004.

Administrative Tasks

- Think about topics that you would like to be discussed at future group meetings. E-mail your suggestions to Kelly.
- Next Thursday's topic is Synchronization

Summary

The paper proposes a new shared memory model, which it calls Transactional Memory Coherence and Consistency (TCC). TCC's model is based on transactions, which is a group of instructions to be atomically executed. The idea behind the model is to eliminate the need for synchronization and its related complexity. The model is dependant on the increasing interprocessor bandwidth on parallel systems today.

Discussion

- The discussion began with Christos providing a clarification of the terms memory coherency and consistency.
    - Coherency: One address in shared-memory architecture should appear to have the same value to all processors, regardless of the actual value in each processor's cache or main memory. There needs to be a mechanism to enforce this.
    - Consistency: Writes and reads to memory, regardless of their origin, should appear in order—i.e. RAW, WAW, etc. hazards should be avoided.
- Programming Model
    - Ernesto points out that programming model may not be as easy it as the paper suggests.
    - Rebecca expresses concerns about how the programmer would know how to choose the appropriate size of the transactions.
    - Christos claims that merging transaction to form a bigger transaction is okay. Splitting them is a lot more difficult and dangerous.
- Hardware
    - Austin sees scalability being an issue.
    - Christos says that coherence is easier to implement because of coarse grain nature of transaction.
    - Christos also says that TCC is also more latency intolerant because:
        - It doesn't have to wait for a response.
        - Committing is a 1-way communication
    - On the downside:

        - Latency of arbitration can be issue since only 1-processor can commit at a given moment.
        - Ordered transactions can also introduce latencies
    - Bandwidth:
        - According to Christos, BW figures cited in paper seem reasonable—Authors consulted Bill Dally and he said it was okay☺
        - However, the paper assumes infinite bandwidth for the interprocessor buses and as Samir noted, miss-free caches are also being assumed. It would be interesting to see how these figures look on the real chip, which is taping out this summer…