Título: A Shared-Variable-Based Synchronization Approach to Efficient Cache Coherence Simulation for Multi-Core Systems
Citação: Cheng-Yang Fu, Meng-Huan Wu, Ren-Song Tsay, "A shared-variable-based synchronization approach to efficient cache coherence simulation for multi-core systems", Design, Automation & Test in Europe Conference & Exhibition (DATE), 2011, pp.1-6, 14-18 March 2011
Autor: Andrei Braga **RA:** 079713

Para manter a consistência de memória em uma arquitetura multi-core, é necessário empregar um sistema adequado para a manutenção de coerência das memórias cache. Parâmetros das memórias cache como tamanho e política de substituição são fundamentais para o desenvolvimento de um hardware em uma arquitetura multi-core. Os efeitos da manutenção de coerência das memórias cache é igualmente importante para o desempenho de um software de execução paralela. Assim, a simulação de um sistema de manutenção de coerência das memórias cache é crucial para ambas as frentes, hardware e software, em uma arquitetura multi-core.

Uma simulação de manutenção de coerência das memórias cache envolve vários simuladores, um para cada core. Para manter o tempo de simulação consistente para cada core, é necessária uma sincronização de tempo. As abordagens convencionais e intuitivas, de sincronizar a cada ciclo ou acesso de memória, propriciam uma simulação de baixo desempenho. À luz do conhecimento dos autores deste artigo, as abordagens existentes geram simulações com apenas uma de duas boas características: rapidez ou precisão. A principal contribuição deste artigo é, então, uma abordagem eficiente para a sincronização de tempo.

Para avaliar o desempenho da abordagem proposta, os autores realizam experimentos que a comparam com as quatro outras abordagens seguintes. A primeira (CB, do inglês cycle-based) é a abordagem que sincroniza cada simulador de um core em cada tempo de um ciclo. A simulação gerada por esta abordagem é garatidamente precisa. No entanto, a quantidade de sincronizações realizadas prejudicam a rapidez da simulação.

A segunda e a terceira abordagens são abordagens dirigidas a eventos. Abordagens dirigidas a eventos sincronizam os simuladores de cada core somente quando ocorrem determinados eventos. A segunda abordagem (MC, do inglês memory accesses with coherence actions) sincroniza, em vez de em cada tempo de um ciclo, no momento da execução do próximo evento de passagem de mensagem. A terceira abordagem (MA, do inglês memory accesses) faz a sincronização em cada ponto de acesso de memória. Na prática, estas abordagens podem gerar muitos eventos e as simulações obtidas também podem ter um desempenho ruim.

A quarta abordagem (SMC, do inglês shared variable access with coherence actions) faz uso da observação a seguir. Programas de execução paralela utilizam variáveis compartilhadas para se comunicar ou integarir um com o outro. Somente variáveis compartilhadas podem residir em mais de uma memória cache (em mais de um core) e apenas acessos a variáveis compartilhadas são importantes para a manutenção de coerência das memórias cache. Com esta observação, é possível sincronizar os simuladores de cada core apenas no acesso de variáveis compartilhadas e, assim, obter uma simulação com bom desempenho e precisão.

A abordagem proposta neste artigo (SMA, do inglês shared variable access) ainda se utiliza da seguinte análise. Para melhorar a eficiência da simulação produzida, os autores observam que o tratamento das *ações de coerência* em um simulador de um core pode ser adiado até se encontrar um evento de acesso de uma memória compartilhada. No entanto, é preciso processar as ações de coerência na ordem cronológica correta.

Os autores, então, realizam experimentos que permitem comparar as cinco abordagens citadas quanto a três quesitos: rapidez (milhões de instruções processadas por segundo), ônus causado pelas sincronizações e tempo total de simulação. Nos três quesitos, a abordagem SMA supera as outras. Quanto à rapidez, em especial, a abordagem SMA é de 18 a 44 vezes mais rápida que a abordagem CB, de 8 a 17 vezes mais rápida que abordagem MC e de 6 a 8 vezes mais rápida que a abordagem MA.

Em suma, os autores apresentam, comprovando com experimentos, uma abordagem eficiente para a sincronização de tempo em simulações de coerência das memórias cache em uma arquitetura multi-core. A abordagem faz uso das propriedades operacionais de manutenção de coerência das memórias cache e mantém eficientemente a sequência correta de simulação. Como um trabalho futuro, os autores propõem estender o esquema de sincronização proposto para a simulação de sistemas multitarefas.