Reducing Peak Power with a Table-Driven Adaptive Processor Core

Vasileios Kontorinis – Amirali Shayan – Rakesh Kumar – Dean M. Tullsen

MICRO – 2009

Autor do Resumo: Davi de Andrade Lima Castro – **RA**: 107072

A maioria das técnicas de redução de consumo atualmente empregadas nos processadores de alto desempenho buscam apenas reduzir o **consumo médio** de potência. Embora esta redução seja desejável e traga benefícios como o barateamento das soluções de resfriamento (além é claro da própria redução do consumo de energia), o **consumo de pico** impacta diretamente em vários aspectos, tais como a própria implementação física do circuito integrado, requerimentos de *packaging* e custos da fonte de alimentação.

A razão disto está no fato de que muitas decisões devem ser tomadas para cobrir um pior caso de consumo, embora este raramente ocorra em operações normais. Existe então uma parcela considerável de *over-design* que acarreta em maiores custos.

O artigo em questão propõe uma arquitetura altamente adaptativa, com componentes configuráveis centralizadamente controlados. É explorada uma observação experimental importante que mostra ser possível configurar minimamente alguns componentes sem prejudicar tanto o desempenho de uma aplicação, desde que os componentes mais importantes (*bottleneck*) para a aplicação estejam maximamente configurados.

Como o controle é **centralizado**, é possível então **garantir** um valor limite de **consumo de pico** escolhendo apenas certas combinações de configurações de forma a nunca ultrapassar este valor limite – apenas um certo conjunto dos componentes estará maximamente configurado. É esta garantia que possibilita tomar decisões baseadas no pico máximo escolhido (70%, 75% ou 80% do pico normal) durante o *design*.

A arquitetura é estruturada em três elementos: **componentes configuráveis, memória com as configurações possíveis** e **controle adaptativo**.

Exemplos de componentes configuráveis são: *cache*, *re-order buffer*, unidades de execução e outros. Eles são responsáveis por em torno de 50% do consumo de pico, porém como não é possível desligar completamente certos componentes, o valor mínimo possível encontrado para o valor limite de consumo de pico foi 70%.

O controle adaptativo é necessário porque cada aplicação possui sua própria configuração-ótima. As tarefas deste controle são três: *decidir quando mudar de configuração, qual configuração usar* (dentre as que se encontram na memória) e *avaliar a configuração escolhida*. O artigo propõe e avalia alguns algoritmos e os de melhor resultado utilizam *feedback* da performance atual para encontrar a configuração-ótima.

Os resultados obtidos foram muito positivos. Para a máxima redução de pico possível, 30%, a perda em desempenho é de 10%. Para uma redução de 25% do pico, temos uma penalidade de 5% no desempenho e para este caso estima-se uma redução de 5.3% da área do *die* (para a mesma variação na tensão de alimentação) ou uma redução de 26% na variação da tensão (para a mesma área). A redução de área é conseqüência direta da redução do número (ou capacitância) dos capacitores de desacoplamento, e estes estão relacionados com o pico do consumo de corrente.

Além disto, como os componentes configuráveis utilizam *power gating* (desligamento parcial de energia), a arquitetura também reduz o consumo médio de potência.

O *overhead* introduzido pelas técnicas é pequeno, visto que a memória de configuração é acessada em uma pequena parcela do tempo e possui tamanho em torno de apenas 400 bytes, e a lógica do controle adaptativo não faz parte do caminho crítico da arquitetura.

Como desvantagens têm-se o aumento da complexidade do teste e verificação do processador.