MuTARe: A Multi-Target, Adaptive Reconfigurable Architecture

Marcelo Brandalero [1], Antonio Carlos Beck F.[1]

[1]Universidade Federal do Rio Grande do Sul

Abstract. Power consumption, earlier a design constraint only in embedded systems, has become the major driver for architectural optimizations in all domains, from the cloud to the edge. Application-specific accelerators provide a low-power processing solution by efficiently matching the hardware to the application; however, since in many domains the hardware must execute efficiently a broad range of fast-evolving applications, unpredictable at design time and each with distinct resource requirements, alternatives approaches are required. Besides that, the same hardware must also adapt the computational power at run time to the system status and workload sizes. To address these issues, this thesis presents a general-purpose reconfigurable accelerator that can be coupled to a heterogeneous set of cores and supports Dynamic Voltage and Frequency Scaling (DVFS), synergistically combining the techniques for a better match between different applications and hardware when compared to current designs. The resulting architecture, MuTARe, provides a coarse-grained regular and reconfigurable structure which is suitable for automatic acceleration of deployed code through dynamic binary translation. In extension to that, the structure of MuTARe is further leveraged to apply two emerging computing paradigms that can boost the power-efficiency: Near-Threshold Voltage (NTV) computing (while still supporting transparent acceleration) and Approximate Computing (AxC). Compared to a traditional heterogeneous system with DVFS support, the base MuTARe architecture can automatically improve the execution time by up to 1:3x, or adapt to the same task deadline with 1:6x smaller energy consumption, or adapt to the same low energy budget with 2:3x better performance. In NTV mode, MuTARe can transparently save further 30energy in memory-intensive workloads by operating the combinatorial datapath at half the memory frequency. In AxC mode, MuTARe can further improve power savings by up to 50% by leveraging approximate functional units for arithmetic computations.

Resumo. Consumo de potência, antigamente um limitante no projeto apenas de sistemas embarca-dos, hoje é um dos principais objetivos de otimização em todos os domínios de dispositivos, desde a computação na núvem até a computação na borda. Aceleradores de propósito específico são capazes de fornecer uma solução para o processamento de baixa potência ao adequar o hardware à aplicação; porém, visto que, em diversos domínios, o hardware necessita executar uma ampla gama de aplicações, cada uma com diferentes requisitos computacionais, abordagens alternativas se fazem necessárias. Além disso, o mesmo hardware precisa se adequar, em tempo de execução, ao estado do sistema e tamanho da carga de trabalho, aumentando o poder computacional ao executar uma tarefa exigente e reduzindo-o quando

inativo. De forma a resolver estes problemas, esta tese apresenta um acelerador de proposito geral que pode ser acoplado a um conjunto heterogeneo decores e suporta DVFS, sinergisticamente combinando técnicas para uma melhor combi-nação entre diferentes aplicações e hardware quando comparado aos designs existentes hoje. A arquitetura resultante, MuTARe, provê uma estrutura regular e reconfigurável que é adequada para aceleração automática de código já existente através de tradução binária.Além disso, MuTARe também provê uma estrutura adequada para aplicar dois emergentes paradigmas de computação que podem aumentar a eficiencia de potência: computação no nível da tensão de threshold(mantendo a capacidade de aceleração transparente) e computação aproximativa. Comparado a um sistema heterogeneo tradicional com suporte a DVFS, a arquitetura MuTARe base pode automaticamente melhorar o tempo de execução em 1.3x, ou adaptar-se para o mesmo baixo tempo de execução com uma redução de 1.6x no consumo energético, ou adaptar-se para o mesmo baixo nível de energia com 2.3x melhor performance. No modo near-threshold, MuTARe pode melhorar o consumo de potência de forma transparente em mais 30% em tarefas que exigem bastante memória operando o circuito combinacional à metade da frequencia da memória. No modo computação aproximativa, MuTARe consegue melhorar o consumo de potência em até mais 50% usando unidades funcionais aproximativas para as computações.