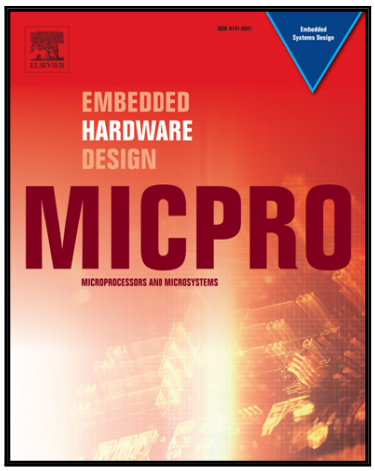

Optimizing Power Efficiency for 3D Stacked GPU-In-Memory Architecture

Wen Wen, Jun Yang, Youtao Zhang

PII: S0141-9331(17)30050-9
DOI: 10.1016/j.micpro.2017.01.005
Reference: MICPRO 2502

To appear in: *Microprocessors and Microsystems*

Received date: 1 February 2016
Revised date: 22 September 2016
Accepted date: 17 January 2017

Please cite this article as: Wen Wen, Jun Yang, Youtao Zhang, Optimizing Power Efficiency for 3D Stacked GPU-In-Memory Architecture, *Microprocessors and Microsystems* (2017), doi: 10.1016/j.micpro.2017.01.005

Optimizing Power Efficiency for 3D Stacked GPU-In-Memory Architecture

Wen Wen, Jun Yang

Department of Electrical and Computer Engineering
University of Pittsburgh
Pittsburgh, USA
Email: {wew55, juy9}@pitt.edu

Youtao Zhang

Department of Computer Science
University of Pittsburgh
Pittsburgh, USA
Email: zhangyt@cs.pitt.edu

Abstract

With the prevalence of data-centric computing, the key to achieving energy efficiency is to reduce the latency and energy cost of data movement. Near data processing (NDP) is a such technique which, instead of moving data around, moves computing closer to where data is stored. The emerging 3D stacked memory brings such opportunities for achieving both high power-efficiency as well as less data movement overheads. In this paper, we exploit power efficient NDP architectures using the 3D stacked memory. We integrate the programmable GPU streaming multiprocessors into the NDP architectures, in order to fully exploit the bandwidth provided by 3D stacked memory. In addition, we study the tradeoffs between area, performance and power of the NDP components, especially the NoC designs. Our experimental results show that, compared to traditional architectures, the proposed GPU based NDP architectures can achieve up to 43.8% reduction in EDP and 41.9% improvement in power efficiency in terms of performance-per-Watt.

Keywords: GPU; Stacked Memory; NoC; Power Efficiency