Uso de Memórias HMC no contexto de Sistemas Embarcados

Carlos Michel Betemps (16104417)

TEC-II - Hierarquias Avançadas de Memória - Proposta de Trabalho

PPGC - UFPel - 2017/1

Objetivos:

- Realizar um estudo de revisão sobre memórias HMC (Jeddeloh & Keeth, 2012, Zou et al. 2015, Beica, 2015, Pawlowski, 2011).
- Avaliar a utilização de Memórias HMC, em substituição às memórias DDR, como memória principal em sistemas embarcados (a partir de metodologia apresentada na próxima seção).
- Analisar resultados obtidos, apontando tendências, resultados interessantes e lições aprendidas.

Metodologia:

- Realizar levantamento de artigos que abordem o estado-da-arte sobre memórias HMC, focando em sua arquitetura de implementação, suas aplicações, vantagens de uso e problemas associados. Deverão ser utilizados artigos publicados em eventos e/ou periódicos. As pesquisas serão realizadas utilizando a máquina de busca do Google Scholar.
- Utilizar o simulador *gem5* para simular a arquitetura ARM (ARM Ltd.) - considerando a grande utilização de processadores ARM no contexto de sistemas Embarcados
    - Preparar o *setup* para a simulação a ser realizada no *gem5* (Binkert et al., 2011, gem5)
    - ***Verificar**: É necessário utilizar simulação com mais de um elemento de processamento (core)? Visando uma maior "pressão" sobre o sistema de memória.*
- Utilizar programas do benchmark *MiBench* (Guthaus et al. 2001) para execução no simulador *gem5*.
    - Realizar a geração de código com o compilador *gcc-arm-gnueabihf* (*cross-compiling*)
    - Programas para execução (pelo menos um programa em cada categoria do benchmark).
- Utilizar a ferramenta CACTI-3DD (Chen et al. 2012) para levantar dados de potência, área e tempo considerando as memórias HMC e DDR.
    - ***Risco***: dificuldade, até o momento, na obtenção da ferramenta;
    - ***Alternativas***: zsim-nvmain, 3D-Memory-Simulator,
- Utilizar o simulador CasHMC (Jeon & Chung, 2017) para levantar dados de latência e largura de banda. O simulador recebe como entrada traços (*traces*) de uso de memória como entrada. Como saída devolve informações como latência e largura de banda.
    - Caso necessário, como ***alternativa***, pode ser analisado o uso do simulador HMC-Sim (Leidel & Chen, 2014)
- Realizar as simulações considerando as seguintes configurações para a hierarquia de memória:
    - **L1i&d**: tamanho de **32 KB**, associativa **8-way**, tamanho de linha **64 B** (proposta inicial)
    - **Memória Principal**: 512MB (confirmar limitação do *gem5*).
        - DDR
        - HMC
    - Executar simulações com as seguintes hierarquias de memória:
        - L1 + DDR (ddr) - base
        - L1 + HMC (hmc)
        - L1 + L2 + DDR (l2+ddr)
        - L1 + L2 + HMC (l2+hmc)
    -
- Geração de estatísticas, na execução de cada programa em cada diferente configuração, para posterior levantamento de estimativas de tempo de execução, consumo energético, latência, largura de banda e área; com base nas estatísticas geradas pelo *gem5*, CACTI-3DD e CasHMC.
- Análise dos dados levantados, geração de resultados e gráficos de interesse, discussão sobre os resultados, apresentação de conclusões e encaminhamento de possíveis trabalhos futuros.

Motivação

Vários estudos abordaram o uso de memórias HMC e correlatas. Focando em um escopo mais amplo, especificamente em tecnologias 3D, o trabalho de Zou et al., (2015) aponta a integração de memórias 3D em

arquiteturas heterogêneas, possibilitando a integração de tecnologias diferentes no mesmo chip. Já o trabalho de Beica (2015) apresenta uma revisão das tecnologias 3D com integração via TSV (*Through-Silicon Via*), focando em aplicações e tendências de mercado. Suresh et al. (2014) apresenta uma avaliação da aplicação das tecnologias emergentes de memória no contexto de HPC e aplicações intensiva em dados, experimentando arquiteturas híbridas de memórias voláteis e não-voláteis. Santos et al. (2016) apresenta experimentos com uso de HMC (com reduzida latência) em aplicações *streaming* e aponta situações em que o uso de caches L3 é desnecessário. Outro estudo lida com questões de desempenho e energia de uma memória HMC Gen2 na execução de aplicações centradas em dados. Emulação e Execução em uma placa FPGA são combinadas (Gokhale et al., 2014). Alves et al. (2016) apresenta extensões na memória HMC para possibilitar o processamento-em-memória (PIM) de operações vetoriais, visando evitar a contenção nos canais de comunicação e poluição na memória cache. *Active Memory Cube* (AMC) é a arquitetura de processamento-em-memória apresentada no trabalho de Nair et al., (2015), a mesma apresenta um conjunto de unidades de processamento implementadas na camada de base da memória (HMC).

Basicamente os trabalhos apontam as vantagens do uso de memórias HMC, ressaltando a melhora de latência, largura de banda, potência e densidade (Jeddeloh & Keeth, 2012). Dado que sistemas embarcados normalmente possuem requisitos restritos quanto a área e consumo energético, mas ao mesmo tempo requisitos exigentes quanto ao tempo de execução e capacidade de processamento, vislumbra-se a possibilidade da aplicação de memórias HMC em sistemas embarcados, visando o aumento de desempenho (latência e largura de banda) com um eficiente consumo energético e de área. Assim, o trabalho visa realizar uma revisão do estado da arte sobre memórias HMC e experimentos (simulados) que visam avaliar a possibilidade de aplicar memórias HMC como memória principal em arquiteturas de sistemas embarcados. Memórias HMC estão em pleno desenvolvimento e estudo. O *Hybrid Memory Cube Consortium* (HMC Consortium) reúne uma série de parceiros dedicados ao desenvolvimento desta tecnologia de memória.

Referências

- JEDDELOH, Joe; KEETH, Brent. Hybrid memory cube new DRAM architecture increases density and performance. In: VLSI Technology (VLSIT), 2012 Symposium on. IEEE, 2012. p. 87-88.
    - *Presents the HMC, a DRAM architecture that improves latency, bandwidth, power and density. Through-silicon vias (TSVs), 3D packaging and advanced CMOS performance enable a new approach to memory system architecture.*
- ZOU, Qiaosha et al. Heterogeneous architecture design with emerging 3D and non-volatile memory technologies. In: Design Automation Conference (ASP-DAC), 2015 20th Asia and South Pacific. IEEE, 2015. p. 785-790.
    - *In the paper is demonstrated the advantages of leveraging three-dimensional (3D) integration on heterogeneous architectures. With 3D die stacking, disparate technologies can be integrated on the same chip, such as the CMOS logic and emerging non-volatile memory, enabling a new paradigm of architecture design.*
- BEICA, Rozalia. 3D integration: Applications and market trends. In: 3D Systems Integration Conference (3DIC), 2015 International. IEEE, 2015. p. TS5. 1.1-TS5. 1.7.
    - *This paper will provide an overview of the different applications of 3D integration using TSV technology, including product announcements, reverse engineering and worldwide patent activities, highlighting the most active players and their activities. The HMC memory is a package incorporating a high-speed logic layer with a stack of memories. The performance benefits (increased performance, bandwidth, lower power consumption, lower latency, etc.) that TSV can bring have already been proven and various memory products have been adopted for servers, high performance computers and products. Cost continue to remain a limiting factor especially for the more complex devices within the consumer market.*
- JEON, Dong-Ik; CHUNG, Ki-Seok. CasHMC: A Cycle-accurate Simulator for Hybrid Memory Cube. IEEE Computer Architecture Letters, v. 16, n. 1, p. 10-13, January-June 2017.
    - *The paper presents a cycle-accurate simulator for hybrid memory cube called CasHMC. It provides a cycle-by-cycle simulation of every module in an HMC and generates analysis results including a bandwidth graph and statistical data. Furthermore, CasHMC is implemented in C++ as a single wrapped object that includes an HMC controller, communication links, and HMC memory.*

The simulation results includes number of requests, reads, writes, flow packets, bandwidth, and average latency.

- SURESH, Amoghavarsha; CICOTTI, Pietro; CARRINGTON, Laura. Evaluation of emerging memory technologies for HPC, data intensive applications. In: Cluster Computing (CLUSTER), 2014 IEEE International Conference on. IEEE, 2014. p. 239-247.
    - *In this paper, we evaluate the impact of emerging technologies on HPC and data-intensive workloads modeling a 5-level hybrid memory hierarchy design. The HMC memory is used as a LLC (last level cache) with/without a NVM (Non-volatile) memory as a main memory.*
- SANTOS, Paulo C; ALVES, M. A.; DIENER, M.; CARRO, L.; NAVAUX, P. O. Exploring cache size and core count tradeoffs in systems with reduced memory access latency. In: Parallel, Distributed, and Network-Based Processing (PDP), 2016 24th Euromicro International Conference on. IEEE, 2016. p. 388-392.
    - *In this paper is presented an evaluation of the L3 cache importance on a high performance processor using HMC also exploring chip area tradeoffs between the cache size and number of processor cores. The high bandwidth provided by HMC memories can eliminate the need for L3 caches, removing hardware and making room for more processing power. The evaluations show that performance increased 37% and the EDP improved 12% while maintaining the same original chip area in a wide range of parallel applications, when compared to DDR3 memories.*
- PAWLOWSKI, J. Thomas. Hybrid memory cube (HMC). In: Hot Chips 23 Symposium (HCS), 2011 IEEE. IEEE, 2011. p. 1-24. (presentation slides)
    - *The slides present some HMC memories issues, like its goals, problems, architecture and performance. Also present some HMC (Technology: HMC, 4 DRAM w/ Logic) performance values:*
    - *VDD IDD BW GB/ s Power (W) mW/ GB/ s pj/ bit real pJ/ bit*
    - *1.2 9.23 128.00 11.08 86.53 10.82 13.7*

- LEIDEL, John D.; CHEN, Yong. HMC-sim: A simulation framework for hybrid memory cube devices. Parallel Processing Letters, v. 24, n. 04, p. 1442002, 2014.
    - This work introduces a new simulation framework developed specifically for the Hybrid Memory Cube specification. We present a set of novel techniques implemented on an associated development framework that provide an infrastructure to flexibly simulate one or more Hybrid Memory Cube stacked die memory devices attached to an arbitrary core processor.
- GOKHALE, Maya; LLOYD, Scott; MACARAEG, Chris. Hybrid memory cube performance characterization on data-centric workloads. In: Proceedings of the 5th Workshop on Irregular Applications: Architectures and Algorithms. ACM, 2015. p. 7.
    - The study deals with the performance and energy of a Gen2 HMC on data-centric workloads through a combination of emulation and execution on an HMC FPGA board. An in-house FPGA emulator has been used to obtain memory traces for a small collection of data-centric benchmarks. (HBM → GPU. WideIO → mobile. HMC → server/HPC applications).
- ALVES, Marco A. Z.; DIENER, M.; SANTOS, P. C.; CARRO, L.. Large vector extensions inside the HMC. In: Design, Automation & Test in Europe Conference & Exhibition (DATE), 2016. IEEE, 2016. p. 1249-1254.
    - The paper introduces the HIVE (HMC Instruction Vector Extensions) architecture, which allows performing common vector operations directly inside the HMC, avoiding contention on the interconnections as well as cache pollution.
- NAIR, Ravi et al. Active memory cube: A processing-in-memory architecture for exascale systems. IBM Journal of Research and Development, v. 59, n. 2/3, p. 17: 1-17: 14, 2015.
    - This paper described a new processing-in-memory architecture called the Active Memory Cube, which implements a set of processing units in the base layer under a stack of DRAM dies (HMC - Hybrid memory Cube).
- CHEN, Ke; LI, S.; MURALIMANOHAR, N.; AHN, J. H.; BROCKMAN, J. B.; JOUPPI, N. P.. CACTI-3DD: Architecture-level modeling for 3D die-stacked DRAM main memory. In: Design, Automation & Test in Europe Conference & Exhibition (DATE), 2012. IEEE, 2012. p. 33-38.
    - The papers introduces CACTI-3DD, a major extension to CACTI, designed to support architecture-level modeling of modern and future DRAM main memories. It simultaneously models power, area, and timing of commodity 2D and 3D DRAM designs, a requirement for accurate architecture-level tradeoffs.
- Hybrid Memory Cube Consortium. URL: http://www.hybridmemorycube.org/

- Nathan Binkert, Bradford Beckmann, Gabriel Black, Steven K. Reinhardt, Ali Saidi, Arkaprava Basu, Joel Hestness, Derek R. Hower, Tushar Krishna, Somayeh Sardashti, Rathijit Sen, Korey Sewell, Muhammad Shoaib, Nilay Vaish, Mark D. Hill, and David A. Wood. The gem5 simulator. SIGARCH Comput. Archit. News, 39(2):1–7, August 2011.
- gem5. The gem5 Simulator - A modular platform for computer-system architecture research. http://gem5.org/Main_Page, 2017.
- M. R. Guthaus, J. S. Ringenberg, D. Ernst, T. M. Austin, T. Mudge, and R. B. Brown. Mibench: A free, commercially representative embedded benchmark suite. In Proceedings of the Workload Characterization, 2001. WWC-4. 2001 IEEE International Workshop, WWC '01, pages 3–14, Washington, DC, USA, 2001. IEEE Computer Society.
- ARM Ltd. ARM. http://www.arm.com/, 2017.

Mapa Mental da Proposta de Trabalho: