УДК 621.3.049.77:000.93

EMERGING ARCHITECTURES FOR PROCESSOR-IN-MEMORY CHIPS: TAXONOMY AND IMPLEMENTATION

Valery A. Lapshinsky

Peoples' Friendship University of Russia
Miklukho-Maklaya str., 6, Moscow, Russia, 117198
National Research Nuclear University MEPhI
Kashirskoe Shosse, 31, Moscow, Russia, 115409

The emergence of PIM (processing-in-memory) die and Date-Centric systems (DCS) and near-data processing approach (NDP) has given rise to the need of developing architectural taxonomy for multi-core PNM (processing near memory) hardware with multi-level memory structure. PIM die (in Russian technical literature usually used terms chips or crystals) considered as an effective alternative to conventional SRAM/DRAM/Flash-memory on Cache-CPU/Main Memory/Storage Class Memory and Storage levels. In the past decade, a few different methods to classify and to implement PIM die and DCS/NDP systems proposed. These approaches are either software interfacing with computing, hierarchical and massively parallel SIMD processing approaches etc. In this paper, presented summarized prolegomena for PIM die architecture and implementation. In particular, in form of basic PIM chips and nanostores.

Key words: processing-in-memory; processing near memory; near-data processing; Date-Centric systems; PIM memory taxonomy; basic PIM chips implementation; emerging memories chips and nanostores

1. Introduction

The huge and growing degree of integration of standard memory chips represents both an opportunities and challenges. On one hand, the ability to store large volumes of data can encourage the improvement of the characteristics of computer systems for various applications — evolution of exascale computing, Big Data and high performance analytics, data mining [2; 13; 14], network technologies including sensor networks and systems [12], improvement of the characteristics of mobile hardware (in the form of smartphones and tablets) and so on. On the other hand, this die integration is also creating a host of new problems. In particular, practical limits on die power dissipation restrict the increase in clock speed and scalability [2]. The current trend is transition from sequential to parallel data in-memory processing, again with die power dissipation limits and problems of synchronization techniques [7; 11].

Processing capabilities can be embedded into conventional SRAM/DRAM/Flash NOR/NAND memory [2; 5; 11; 16]. Now these capabilities becoming software-visible. They needs special software interfaces, threading packages or generally to modify system and application software [9]. Finally, there is the problem of massive industrial production of cheap PIM die with efficient architecture suitable for use on the different levels of the memory hierarchy.

At the same time, technology opportunities such as:

— the increased adoption of emerging nonvolatile memories (NVM)

— memristors or phase-change memory [2; 13; 14] and functional devices with nanometer dimensions and exotic non-linear current-voltage characteristics [15],

— optical communications on chips,

— multi-cores, and heterogeneous computing,

all provide a unique opportunity for an end-to-end redesign of data-centric solutions across both hardware and software [2; 3; 6; 10; 17].

2. Taxonomy prolegomena

2.1. Taxonomy based on software interface with computing

The proposed [9] taxonomy approach based on software interface with computing divide logic in memory die architecture into two classes: software-transparent and software-visible.

Software-transparent PIM die functions are non-computing and associated service functions (memory controllers, built-in self-testing etc.) [8; 9]. But software-visible class of PIM architecture with fixed-function or pre-defined operations (bounded or compound operands) and fully programmable (general or specific purpose — GP/SP) logic in memory requires software interface with computing.

Examples of software interface, advantages and disadvantages of the different variants of this architecture discussed in [9]. There is significant middle ground worth further exploration, which proposed taxonomy could serve as framework for classifying and better understanding the strength and weaknesses of different classes of PIM die architectures.

2.2. Taxonomy based on memory hierarchy

While implementing a PIM die and PNM memory systems changes may be made to one level of the memory hierarchy or on all levels (figure 1). Data movement impacts performance, power efficiency and reliability, three fundamental attributes of systems. NDP seeks to minimize data movement by computing at the most appropriate location in the hierarchy, considering the location of data and the information that needs to be extracted from the data [3]. In this case, the computation near memory require the creation of effective PIM die architectures of the following classes. PIM chips with processors for: 1) non-volatile persistent (NVP) storage, 2) storage class memory (SCM), 3) main memory or PIM and 4) cache memory.

Past proposals mostly focused, firstly, on the requirements and specific details for each local level of the hierarchy: processing complexity, data type, access pattern, response time etc. [13]. Secondly, focused on the PIM die architecture based on conventional SRAM/DRAM/Flash NOR/NAND chips. However, NVM cost trends suggest that in near future all memory hierarchy levels along with the standard memory and PIM chips will contain chips with NVM PIM architecture. Consequently, with evolution of nanotechnology next step will be creation effective nanostores, which collocate processors and emerging NVM on the same chip [13; 14]. Connected to one another they formed a NDP cluster. There is a wide range of possible NVM PIM die architectures and nanostore systems.

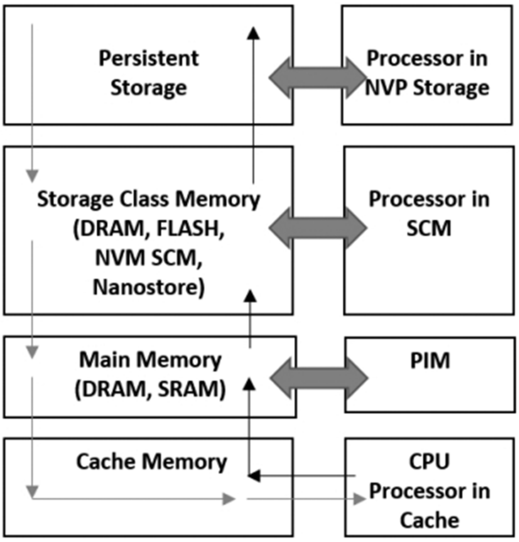

Figure 1. Taxonomy based on memory hierarchy
[Классификация на основе иерархии систем памяти]

2.3. Taxonomy based on SIMD massively parallel processing

SIMD (single instruction stream, multiple-data stream) PIM die architecture is old popular concept of placing computation capabilities to conventional SRAM/DRAM chips [5; 16]. In [11] presented a detailed and comparative taxonomy for PIM in SIMD fields of massively parallel processing (figure 2).

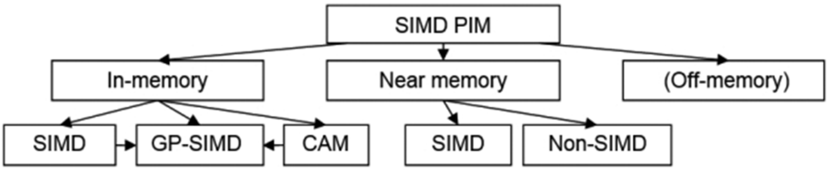

Figure 2. Taxonomy based on SIMD massively parallel processing
[Классификация на основе массовых параллельных вычислений типа SIMD]

According to the combination of number, computing power and location of the processing units/elements on the chip architecture of SIMD PIM crystals divided into the following classes: in-memory, near memory and off-memory.

In-memory architecture means implementation of a very large number of small (typically single bit) SIMD processing elements/units or units with associative functions on memory periphery, matching the number of memory rows and columns. Near memory architecture characterized by many processing units (usually clustered and including non-SIMD units, for example, sequential) which integrated with memory sections or blocks of memory. Architecture off-memory is a limiting case, when several units or accelerators (typically with SP engines) operated by main processors.

The proposed approach allowed to carry out synthesis and evaluations of GP-SIMD PIM die architecture, effectively combining massively parallel SIMD and sequential processing and associative functions and the limitations on power consumption and synchronization.

3. Implementation

The problem of mass production of PIM and PNM based on the traditional SRAM/DRAM/Flash chips can be solved by using a property either hardware or software reconfigurability of basic memory chips [1]. Transparent software architecture underlying crystal primarily intended for setup and implementation of service functions. This implementation considered as the architecture of the 1st generation basic crystals.

Today the integration level already reached a level where the basic PIM chip can contain not only the service functions and a set of processing elements and units for the processing directly in the memory, but also a number of specialized processing units and cores near memory. These cores/units may have a different architecture and purpose. Then the basic memory crystals with multi-core computing functionality that may appear in the future should be attributed to the architecture of the 2nd generation. This architecture can be called as multi-architecture [4]. Respective base crystals may have near universal architecture. They can be combined as parallel computing in-memory and sequential/parallel calculations near memory. 3rd generation — basic NVM GP-PIM die for nanostores.

It can be assumed that the emergence of universal basic PIM and PNM crystals is not far off, because the industry designed and planned for production such chips in the near future. Delay could face defeat in the market competition.

4. Conclusions

Based on the few various approaches for taxonomy and implementing PIM die can be concluded that there is a wide range of options for the PIM chips and PNM systems architecture. Some of these options implemented in practice as test samples, others now offered to consumers as commercial products. In this case, we are talking about the Hybrid Memory Cubes, produced by the 3-D technology using a TSV (Through-Silicon-Vias) interconnections [17].

Researches in the field of architecture of PIM die and, PNM and DSC systems, in fact, is only at the initial stage of its development. The idea of creating basic memory GP-PIM chips with multi-architectural and multi-core fillings suitable for mass production should become a driving force for long-term research and development.

Simultaneously with the searching effective architectural options for NVM PIM die and nanostores, using of innovative NVM memory elements, it is necessary to solve the problems of appropriate CAD-systems and test equipment development. You also need to standardize architecture of the PIM die. For example, on the basis of existing (used in the construction of systems-on-chip) and new (specially optimized for PIM) Intellectual Property (IP) elements and units.

REFERENCES

[1] *Aleksenko A.G., Lapshinsky V.A.* Evaluation of the effectiveness of the basic crystals of memory. Electronic engineering. Ser. Microelectronics. 1980. Vol. 4. P. 3—8 (In Russian).

[2] *Balasubramonian R.* et al. Near-Data Processing: Insights from a MICRO-46 Workshop. IEEE Micro. 2014, July-August. P. 36—42.

[3] *Burr G.W.* et al. Overview of candidate device technologies for storage-class memory. IBM J. Res.&Dev. 2008, JULY/SEPTEMBER. Vol. 52. No. 4/5. P. 12—21.

[4] *DeBenedictis E.P.* Will Moor's Law be Sufficient? IEEE 0-7695-2153-3/04. 2004. P. 1—12.

[5] *Elliott D.G., Snelgrove W.M., Cojocaru C. and McKenzie R* Computational RAM: Implementing Processors in Memory. *J. IEEE Design and Test of Computers*. 1999. № 16(1) (January-March). P. 32—41.

[6] *Freitas R.F. and Wilcke W.W.* Storage-class memory: The next storage system technology. *IBM J. Res. @ Dev*. 2008. Vol. 52. № (4/5). P. 439—447.

[7] *Kulkarni S.* et al. Taxonomy for Transactional Memory Systems. Int. J. of Computer Science and Information Technologies (IJCSIT). 2011. № 2(2). P. 766—775.

[8] *Lapshinsky V.A.* Basic crystals and CPU-oriented organization "smart" memory. Sensors & Systems. 2015. № 1. P. 81—86 (In Russian).

[9] *Loh G.H.* et al. A processing-in-Memory Taxonomy and a Case for Studding Fixed-function PIM. *In WoNDP: 1st Workshop on Near-Data Processing*. 2013. № 1. P. 1—6.

[10] *Meena J.S., Sze S.M., Chand U and Tseng T.Y.* Overview of emerging nonvolatile memory technologies. Nanoscale Research Letters. 2014. P. 1—33.

[11] *Morad A., Yavits L., Ginosar R.* GP-SIMD Processing-in-Memory. ACM Transactions on Architecture and Code Optimization. 2015, January. Vol. 11. Issue 4. P. 1—26.

[12] *Pershenkov V.S., Podlepecky B. I., Shagurin I.I., Bocharov Y.I.* Microelectronics applied to instrument-making. Sensors & Systems. 2015. № 1. P. 3—22. (In Russian)

[13] *Ranganathan P. Chang J. 2011.* IEEE Computer Society **44(1)**. P. 39—48.

[14] *Ranganathan P. and Chang J.* (Re)Designing Data-Centric Data Centers. IEEE Micro. 2012, JANUARY/FEBRUARY. P. 66—70.

[15] *Shibata T.* 2009. Solid-State Electronics. № 53. P. 1227—1241.

[16] *Stone H.S.* A Logic-in-Memory Computer. IEEE Trans. on Computer. 1970. V. C-19. N 1. P. 73—78.

[17] *Zhang D.P.* et al. TOP-PIM: Throughput-Oriented Programmable Processing in Memory. Proc. of the 23rd international symposium on High-performance parallel and distributed computing (HPDC'14), June 23–27, Vancouver, BC, Canada.

НОВЫЕ АРХИТЕКТУРЫ ДЛЯ ЧИПОВ «ПРОЦЕССОРЫ-В-ПАМЯТИ»: КЛАССИФИКАЦИЯ И РЕАЛИЗАЦИЯ

В.А. Лапшинский

Российский университет дружбы народов
ул. Миклухо-Маклая, д. 6, Москва, Россия, 117198
Национальный исследовательский ядерный университет «МИФИ»
Каширское шоссе, 31, Москва, Россия, 115409

Появление чипов типа «процессоры-в-памяти» (ПИМ) систем, ориентированных на данные (Date-Centric systems — DCS), и систем с вычислениями рядом с данными (near-data processing — NDP) настоятельно требует развития методов классификационного анализа архитектуры многоядерных чипов для вычислений рядом с многоуровневой структурой памяти. Чипы (в России в технической литературе обычно используются термины «кристаллы» или «интегральные схемы», ИС) ПИМ рассматриваются как эффективная альтернатива стандартным ИС SRAM/DRAM/Flash-памяти для различных уровней иерархии ЗУ: кеш, оперативной, промежуточной и внешней памяти. В последнее десятилетие были предложены различные способы классификации и методы реализации ПИМ чипов для создания систем типа DSC/NDP. Эти методы включают классификацию на основе программного интерфейса с вычислениями в памяти, иерархический подход и классификацию параллельных вычислений типа SIMD и др. В статье представлен предварительный обзор различных вариантов классификации архитектур ПИМ чипов и их реализации в виде ИС, в частности в виде базовых кристаллов общего применения и нанохранилищ.

Ключевые слова: вычисления в памяти; вычисления рядом с памятью; вычисления рядом с данными; системы, ориентированные на данные; классификация памяти ПИМ; базовые кристаллы ПИМ памяти общего применения; кристаллы новых типов и нанохранилища