G-6-01 (Invited)

HPP: A Novel Architecture for High Performance Processing

Donglin Wang[1], Zhiwei Zhang[1], Zijun Liu[1], Xueliang Du[1], Shaolin Xie[1], Hong Ma[1],
Guangxin Ding[1], Weili Ren[1], Fabiao Zhou[1], Wenqin Sun[1], and Huijuan Wang[1]

[1] Institute of Automation, Chinese Academy of Sciences
95 Zhongguancun East Road, Bejing, 100190, China
Phone: +86-136-8320-7729 E-mail: zhiwei.zhang@ia.ac.cn

Abstract

The HPP is an innovative architecture which targets on high performance computing with great power efficiency and outstanding computing performance. It is suitable for data intensive applications like supercomputing, machine learning and wireless communication. An example chip with four MaPU which is the first generation of HPP cores has been taped out successfully at TSMC 40nm low power process. The innovative architecture shows great energy efficiency over traditional CPU and GPGPU. Compared to MaPU, HPP has made more advancement on architecture. The chip with 32 HPP cores is being developed at TSMC 16nm FFC process and will be used commercially. The peak performance of the chip is 4.3TFLOPS and the power efficiency reaches up to 89.5FLOPS/W.

1. Introduction

Traditionally, various scientific domains have huge demand on high performance computing, such as climate modeling, earth subsurface modeling and sky simulation. Meanwhile, as the emerging field of machine learning and 5G wireless communication, sophisticated processing of large amounts of data poses significant challenge to the high performance processor design.

With sustained power supply, power consumption for indoor system with single processor is not an obvious problem. However, for supercomputers built with thousands of processors, the aggregate power consumption, increasing cost of deployment and maintenance would limit the scale of the system. For example, the most powerful supercomputer TaihuLight consumes 13.571MW and occupies 605 square meters of space [1]. To build such kind of system, dedicated and expensive cooling system and power station should be constructed.

The gap between peak performance and sustained performance [2] is the other challenge of the processor design today. It is reported that the mean utilization for various GPU benchmarks is only 45%. The main reason of this underutilization is memory stalls, which are caused by memory access latency and irregular access patterns [3].

In this paper, HPP design principles and techniques are presented for solving the mentioned problems above for the first time. The architecture of the first version HPP (MaPU) and the example chip with four MaPU cores implemented at TSMC 40LP process are described. HPP improvements of the chip which includes 32 HPP cores are described briefly.

2. General Instructions

The following principles and techniques adopted in HPP are proposed to increase the power efficiency while still providing high sustained processing performance.

(1) HPP is a domain-specific processor which can take advantage of the computing parallelism and data independence existing in the state-of-the-art digital signal processing algorithms. HPP is not designed as the general purpose application processor which should ensure program compatibility like Intel. The widely used techniques in traditional processor design like register rename, dynamic scheduling, and instruction issue window are not adopted. Simple control logic and massive function units are designed in HPP to save power and improve performance.

(2) Flexibility and efficiency when accessing internal SRAM memory is specially designed and optimized in HPP. The HPP use 512 bit width and the word length can be configured to 8 bits, 16bits, 32bit and 64bit. The memory system can provide equal efficiency when accessing the row and column of a matrix with different word length. However, current processers can only provide high efficiency in accessing the row but much lower efficiency when accessing the column of matrix, or vice vise.

(3) Unnecessary data movements in current processor design have resulted in great power waste and low computing efficiency. Several techniques are proposed to avoid the data movement problem in HPP. First, configurable forwarding logic is placed among the computing units. Output of the computing unit can be directly forwarded to the input of another computing unit in this way. This will significantly reduce the data movement between the computing units and register file. Second, compared with traditional processor design, a large capacity register file design is adopted for data reuse and data movement reduction between the register file and internal SRAM memory. Third, under the premise of meeting the area and power constraints, the capacity of internal SRAM memories in HPP core and SoC chip is designed to be as large as possible to reduce data movement between internal and external memories.

(4) For different digital signal processing applications, HPP can support enough flexibility and is programming friendly. For example, the protocol of the wireless communication applications is evolving rapidly. If programming flexibility is not supported, the life cycle of HPP processor will be very short. With the architecture similar to VLIW, function unit design in HPP is controlled by the microcode. For a microcode line, 14 microcodes are assigned to different slots and issued in the same clock cycle. Parameterized macro instructions like FFT, FIR and matrix multiplication

can be provided, which are hand optimized microcode.

(5) With the principle to provide advantage of programmability like common processors and high efficiency like ASIC, configurable forwarding logic is adopted in the design. According to the data flow characteristics of the aimed algorithms, the function units can be organized to form different data paths with configurable logic circuit.

As shown in Fig. 1, MaPU (the first generation of HPP) architecture is made up of three main components: microcode pipeline, multi-granularity parallel memory and scalar pipeline.

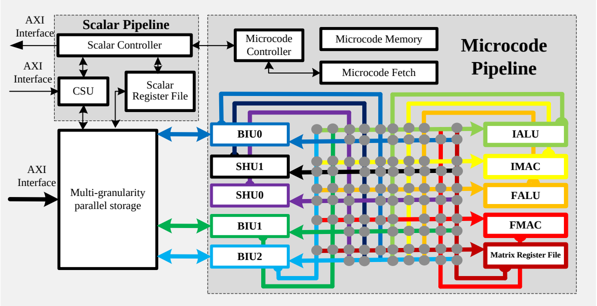

Fig. 1 The first generation of HPP architecture

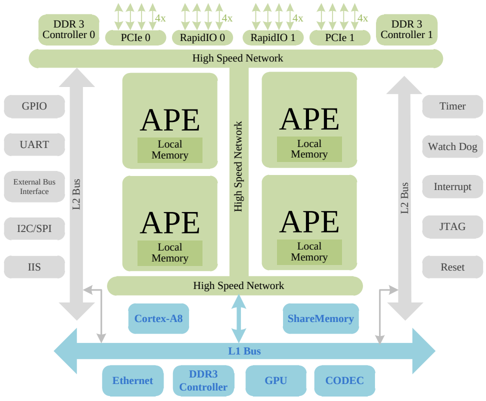

Fig. 2 Simplified SoC structure

The simplified diagram of the example chip design is shown in Fig. 2. The MaPU core of this chip is called APE, which stands for Algorithm Processing Engine. Other components like Cortex-A8 core, IPs, high speed IOs like DDR3, PCIe, RapidIO, and low speed interfaces are also integrated on the chip. Fig. 3 shows the final layout which was fabricated in TSMC 40LP process. The total area is 363.468mm^2. The APE can runs at 1GHz and the power consumption of the chip is 17W.

For single point FFT algorithm, the power efficiency of APE is 48.05 GFLOPS/W. It is shown that, APE has made significant energy efficiency improvement compared with the data presented in [4]. The actual energy efficiency of APE is almost 40x over Core i7 960(CPU) and 2x over Tegra K1(GPU) respectively.

Based on the first version, HPP has made several important improvements. (1) The capacity of microcode instruction memory is increased to support more complex algorithms. (2) The number of memory physical banks is reduced to mitigate the difficulty in physical placement design. (3) Block data accessing is supported to facilitate convolution operation in deep neural network. (4) Matrix transformation is supported in DMA design. (5) Microcode instruction compression is adopted to improve the utilization of instruction memory. (6) Double precision float point FMAC can support four SP FMAC operations. (7) Interrupt circuit is added to improve the interactive ability with the other cores. (8) The efficiency of computing unit is optimized to be more configurable and resources sharing. (9) A novel clock tree design tool is developed for the design and greatly reduced the clock network power. (10) New microcode type such as the ones supporting trigonometric functions and bit operations are added. (11) The internal SRAM can work at the same frequency as the function units which achieves up to 1.4GHz.

The commercial chip version of HPP which has integrated 32 HPP cores is being developed. It will be fabricated in TSMC 16nm FFC process. The frequency is 1.4GHz. The peak computing performance of the chip is 4.3TFLOPS and the power efficiency reaches up to 89.5FLOPS/W.

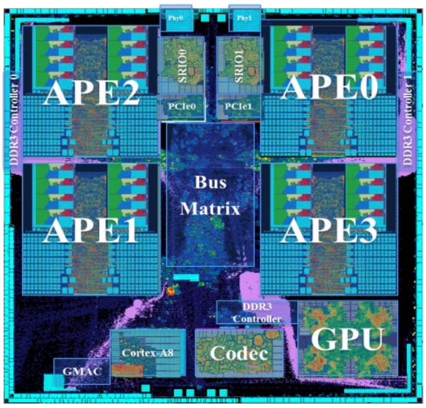

Fig. 3 Layout of the example chip

3. Conclusions

A novel architecture called HPP for high performance computing is presented. The first version of HPP fabricated in 40nm has been taped out successfully and shows significant advantage in computing/power efficiency. The more advanced commercial version of HPP fabricated at TSMC 16nm FFC process with much higher performance and energy efficiency is being developed.

Acknowledgements

I would like to express sincere thanks to all the contributors to the HPP project. This research was funded by the Strategic Pilot Projects of CAS.

References

[1] H. Fu et al., Sci. China Inf. Sci. 59(2016) 072001:2
[2] D. Parello et al., IEEE/ACM SC Conference (2002) 31
[3] A. Sethia et al., IEEE Trans on Computers 61 (2012) 1711
[4] M. H. Ionica et al., IEEE Micro 35 (2015) 6