Chapter 21
The Ultimate Data Flow for Ultimate Super Computers-on-a-Chip

Veljko Milutinović
Indiana University, USA

Miloš Kotlar
School of Electrical Engineering, University of Belgrade, Serbia

Ivan Ratković
https://orcid.org/0000-0002-0524-7227
Esperanto Technologies, Serbia

Nenad Korolija
Independent Researcher, Serbia

Miljan Djordjevic
University of Belgrade, Serbia

Kristy Yoshimoto
Indiana University, USA

Mateo Valero
BSC, Spain

ABSTRACT

This chapter starts from the assumption that near future 100BTransistor SuperComputers-on-a-Chip will include N big multi-core processors, 1000N small many-core processors, a TPU-like fixed-structure systolic array accelerator for the most frequently used machine learning algorithms needed in bandwidth-bound applications, and a flexible-structure reprogrammable accelerator for less frequently used machine learning algorithms needed in latency-critical applications. The future SuperComputers-on-a-Chip should include effective interfaces to specific external accelerators based on quantum, optical, molecular, and biological paradigms, but these issues are outside the scope of this chapter.

INTRODUCTION

Appropriate interfaces to memory and standard I/O, as well as to the Internet and external accelerators, are absolutely necessary, as depicted in the attached figure. Also, the number of processors in Figure 1, could be additionally increased if appropriate techniques are used, like cache injection and cache splitting

DOI: 10.4018/978-1-7998-7156-9.ch021

(Milutinovic, 1996). Finally, a higher speed could be achieved if some more advanced technology is used, like GaAs (Fortes, 1986; Milutinovic et al., 1986). Figure 1 is further explained with data in Table 1.

Figure 1. Generic structure of a future SuperComputer-on-a-Chip with 100 Billion Transistors.

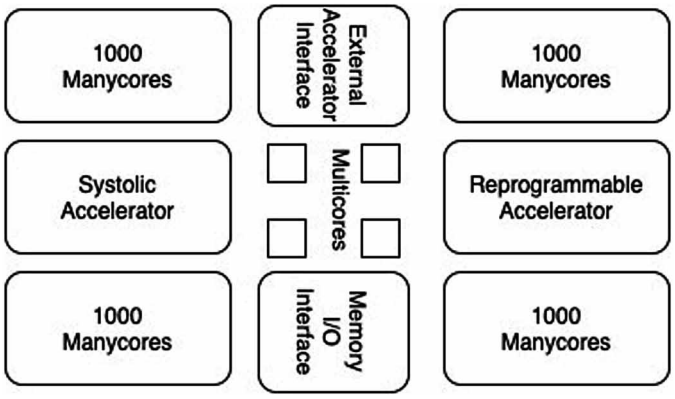

Table 1. Basically, current efforts include about 30 billion transistors on a chip, and this article advocates that, for future 100 billion transistor chips, the most effective resources to include are those based on the dataflow principle. For some important applications, such resources bring significant speedups, that would fully justify the incorporation of additional 70 billion transistors. The speedups could be, in reality, from about 10x to about 100x, and the explanations follow in the rest of this article.

Chip Hardware Type	Estimated Transistor Count
One Manycore with Memory	3.29 million
4000 Manycores with Memory	11 800 million (Techpowerup, 2020)
One Multicore with Memory	1 billion (Williams, 2019)
4 Multicore with Memory	4 billion
One Systolic Array	<1 billion (Fuchs et al., 1981)
One Reprogrammable Ultimate Dataflow	<69 billion (Xilinx, 2003)
Interface to I/O with external Memory	<100 million
Interface to External Accelerators	<100 million
TOTAL	<100 billion

Since the first three structures (multi-cores, many-cores, and TPU) are well elaborated in the open literature, this article focuses only on the fourth type of architecture, and elaborates on an idea referred to as the Ultimate DataFlow, that offers specific advantages, but requires a more advanced technology, other than today's FPGAs.

In addition, some of the most effective power reduction techniques are not applicable to FPGAs, which is another reason that creates motivation for research leading to new approaches for mapping of