Prototype Design of 30 GHz Superconducting Single-Flux-Quantum Microprocessor Towards Cryogenic General Purpose Computing

Koki Ishida[1], Masamitsu Tanaka[2], Takatsugu Ono[1], and Koji Inoue[1] †

[1] Kyushu University, [2] Nagoya University

† Academic advisor

ABSTRACT

Moore's Law, the observation in which the number of transistors in a chip doubles every 18 months, has so far been contributed to the evolution of computer systems, e.g., introducing manycore accelerations and large on-chip caches. Unfortunately, further transistor shrinking cannot be expected anymore, i.e., Moore's Law is coming to an end. Although manufacturing technologies have been progressing, there are some predictions that the end of Moore's Law will come by around 2025 to 2030, due to physical or economic reasons.

Cryogenic computing has a potential to break such critical problem, and a state-of-the-art implementation is available in the market as quantum computers. Although it can effectively be applied to specific purposes such as quantum annealing, there is a large gap regarding functionality between classical digital computing and the application-specific quantum acceleration. To make the cryogenic computing applicable to a wide range of emerging applications is a critical issue. Superconductor single-flux-quantum (SFQ) logic is a promising and practical VLSI technology for the general-purpose, cryogenic computing. Because of its ultra-high-speed and ultra-low-power natures, some researchers have so far greatly been contributed to developing SFQ devices and design technologies and SFQ microprocessors have been successfully demonstrated.

Although the SFQ microprocessors operate with outstanding clock frequency, e.g., several dozen GHz or even more than 100 GHz, unfortunately, their effective performance regarding "program execution speed" is comparable or worse than that of state-of-the-art CMOS microprocessors. The fundamental problem existing at behind the SFQ microprocessors is the lack of optimization from the viewpoint of microarchitecture, i.e., the structure of the SFQ microprocessors does not fully exploit the potential of SFQ logic and

the features of SFQ circuits. Specifically, these SFQ microprocessors employ bit-serial processing to reduce hardware and purse the ease of successful demonstrations. However, such bit-by-bit fine-grained operations make the execution time much longer, resulting in poor microprocessor performance. Moreover, the pipeline structure of these SFQ microprocessors is similar to that of conventional CMOS microprocessors in spite of the difference of device characteristics.

Therefore, we have studied the SFQ microprocessors architecture by standing on a device/circuit/architecture level co-designs. In our previous work, we have concerned SFQ circuits features and proposed bit-parallel processing and gate-level-pipelining with fine-grained, single instruction multiple threads (SIMT) execution for achieving extremely high-performance SFQ computing. Gate-level pipelining is the most fine-grained pipeline structure, in which one pipeline stage consists of only one logic gate. In general, this extremely deep pipeline structure cannot be applied in CMOS designs because the area overhead of pipeline registers becomes greater. However, this disadvantage does not appear in SFQ designs because of its device features. SFQ logic gates inherently have a kind of latch function, so that pipeline registers are not needed. In order to realize high-performance SFQ microprocessors by using gate-level pipelining, most pipeline stalls must be concealed. Current CMOS microprocessors adopt out-of-order execution to conceal pipeline stalls. However, SFQ logic uses a combination of weak voltage pulses for logical operations, the timing adjustment of pulses is one of the critical issues and that makes hard to implement such complex circuits. In addition, SFQ circuits cannot implement large-scale memory efficiently because of its low driving capability. Therefore, we use fine-grained SIMT execution that prepares as many threads as the number of pipeline stages and switches the thread to be executed every clock cycles. The fine-grained SIMT execution can conceal all pipeline stalls that are caused by data hazard while keeping the hardware simple. It can also save memory capacity because all threads execute the same instructions. Actually, we have designed 8 bit-parallel, gate-level-pipelined ALU as a part of microprocessor and successfully demonstrated over 56 GHz with 1.6 mW. However, the whole microprocessors' speed which considering the effect of wire delay has not been clarified. Moreover, the effect of cryocooler which needs to keep the microprocessor 4 K has also been indistinctness.

To solve these problems, we have designed a 4-bit SFQ microprocessor as a prototype and tested the operation by post-layout simulation. Specifically, the microprocessor employs RISC-based original instruction set architecture and implements

total 12 basic instructions, such as arithmetic, data transfer, and control instructions. As a result, the microprocessor has successfully operated at 30 GHz in all instructions and a test program consisting of 20 instructions including loop processing, which calculates matrix-vector product by iterative additions. The power consumption of the microprocessor is 6.6 mW. Based on these results, we have estimated the energy efficiency of 64-bit SFQ microprocessor including the cost of cryocooler. By applying state-of-the-art energy efficient SFQ circuit technology and process technology, the estimated energy efficiency with and without the cost of cryocooler are at most 47 GOPs/W and 31 TOPs/W, respectively.