High-Performance Computation Using a Single-Flux Quantum Accelerator

Mehdipour, Farhad
Graduate School of Information Science and Electrical Engineering, Department of Informatics, Kyushu University

Honda, Hiroaki
Institute of Systems, Information Technologies and Nanotechnologies

Kataoka, Hiroshi
Graduate School of Information Science and Electrical Engineering, Department of Informatics, Kyushu University

Inoue, Koji
Graduate School of Information Science and Electrical Engineering, Department of Informatics, Kyushu University

他

出版情報：International Technical Conference On Circuits/Systems, Computers and Communications. 24, pp.145-148, 2009-07
バージョン：
権利関係：

High-Performance Computation Using a Single-Flux Quantum Accelerator

Farhad Mehdipour†, Hiroaki Honda††, Hiroshi Kataoka†, Koji Inoue† and Kazuaki Murakami†

††Graduate School of Information Science and Electrical. Eng., Kyushu University, Fukuoka, Japan
†Institute of Systems, Information Technologies and Nanotechnologies, Fukuoka, Japan
E-mail: {farhad, kataoka}@c.csce.kyushu-u.ac.jp, dahon@isit.or.jp, {inoue,murakami}@i.kyushu-u.ac.jp

Abstract

A large-scale reconfigurable data-path (LSRDP) processor based on single-flux quantum circuits has been proposed to overcome the barriers originating from the CMOS technology. LSRDP is integrated to a general purpose processor in a high-performance computing system to accelerate the execution of data flow graphs extracted from scientific applications. The LSRDP micro-architecture and its specifications will be presented in this paper. A preliminary performance evaluation of the reconfigurable processor will be given as well.

Keywords: *Reconfigurable accelerators, Single-flux quantum Circuits.*

1. Introduction

Nowadays, providing high computational power to individual researchers in various scientific areas such as quantum chemistry, materials science, environmental issues and etc. is crucial for progress of the research and development. Although, continuing advances in manufacturing processes have made it possible for processor vendors to build increasingly faster, there is still a high demand to meet the required performance for specific applications. Generally as the most of computing systems are implemented by CMOS technology, there are some barriers in realizing powerful computing systems using this technology including high heat radiation, long interconnection delays and memory-wall problem [6].

As a solution, a desk-side tera-flop scale computer is introduced [5] which consists of a CMOS general purpose processor, a memory and a single-flux quantum (SFQ)-based Reconfigurable Large-Scale Data-Path processor (SFQ-LSRDP) as an accelerator (Fig. 1) [5]. Generally, a large memory bandwidth is demanded in conventional accelerators to perform calculations efficiently. Therefore, an on-chip memory is utilized for reduction of the required memory bandwidth. The proposed architecture is expected to be a 10TFLOPS desk-side computer with low electric power consumption and it is suitable for execution of the scientific applications demanding massive computations.

A SFQ circuit is based on the superconductor technology which includes low-power consumption and high-speed compared to the CMOS circuits [5]. A SFQ circuit has a smaller switching energy and high switching speed in comparison with CMOS circuit. In addition, since the SFQ pulse propagates in the speed of light, its transmission speed is not limited to the latency time of the electrical charge and discharge of CMOS gate capacitances. It is also suitable for pipeline processing and there is no additional cost for latch implementation.

In the proposed hybrid architecture, one main component is a large-scale reconfigurable data-path (LSRDP) based on SFQ to overcome the issues concerning to high electric power consumption, high heat radiation, difficulties in high-density packing as well as memory wall problem which limits the processing speed. LSRDP utilizes a data-path comprising reconfigurable interconnections to connect several floating point processing units together. According to Fig. 1, the LSRDP is augmented to GPP as an accelerator. Executing the most frequently executed portions of applications (represented as data flow graphs) or in other word the part of applications demanding massive and time-consuming computations is a main responsibility of the LSRDP. Critical segments are pulled-out from applications and their corresponding configuration bit-streams are generated. During execution of the application on the base processor, configurations associated to critical segments are loaded onto the LSRDP and executed to achieve higher performance and lower power consumption.

In this paper, detailed specifications of the LSRDP architecture and results of a primary performance evaluation of the overall architecture are presented.

2. LSRDP General Architecture and Specifications

Fig. 1 displays an overall architecture of a high performance computer consisting of a GPP, LSRDP as an accelerator and memory elements. Generally, LSRDP is a pipelined architecture comprising a two-dimensional array of processing elements (PEs) such that one PE can be connected through Operand Routing Networks (ORNs) to one or more PEs in the next row.

SFQ technology provides the LSRDP with a straightforward pipelined structure implementation. Each PE can be fed through input ports and the resultant data can be

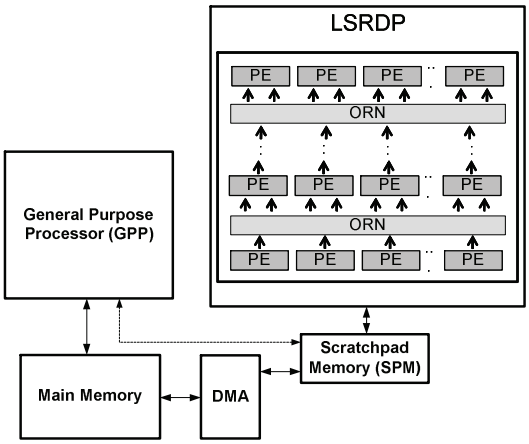

Fig. 1. Overall architecture of the SFQ-LSRDP

transferred to one or more PEs in the next row via ORN switches. Reverse data flow connections are not supported, which means that the flow of data in the array is only in one direction. The LSRDP should be an adaptable accelerator, since it is aimed to target various scientific applications. In order to satisfy this requirement, the architecture is featured with dynamically reconfigurable PEs and ORNs. Originally, an ORN consists of programmable switches. Through configuring control signals provided with PEs and ORN switches, the function of LSRDP can be determined at run time. Such flexibility makes it possible to implement various DFGs on the array.

A data flow graph (DFG) extracted from a target application program is mapped onto the LSRDP array. Since the cascaded PEs can generate a final result without temporally memorizing intermediate data, the number of memory load/store operations corresponding to spill codes can be reduced. Therefore, memory bandwidth required to achieve a high performance might decrease as well. Furthermore, since a loop-body mapped into the PE array is executed in a pipeline fashion, LSRDP can provide a high computing throughput.

Some assumptions and definitions on LSRDP architecture are presented here.

PE types: Each PE includes an FU for implementing desired operation and a TU (transfer unit) as a routing resource for transferring data to the next row. In LSRDP architecture, ORNs provide routing resources between consequent rows. It means, to connect two PEs locating on inconsequent rows one or more transfer units should be utilized. Since a unique implementation of PEs is preferred in SFQ technology, it is supposed that each PE has a general architecture including a functional unit (FU) and a Transfer unit (TU) and it is possible to use an FU for implementing a transfer unit as well. In addition, each PE has three inputs (two inputs for FU and one for TU) and two outputs (one from FU and another from TU).

Type and granularity of functional units: It is assumed that FUs can implement basic 64-bit double-precision floating point operations like e.g. ADD, SUB and MUL. Control instructions (branches) and direct memory accesses via PEs are not supported.

Layout: Layout of the LSRDP represents the type of FUs and their distribution. It is supposed that each FU can implement ADD/SUB and MUL operations.

Internal memory: 64-bit immediate registers are located in each PE in order to handle immediate values. The immediate values are transferred to the registers within configuration phase through a serial bit-stream.

Input/Output ports: A proper number of input/outputs ports are assigned to LSRDP with respect to the available memory bandwidth, LSRDP operation frequency, width of data bus and the number of memory read/write channels.

LSRDP dimensions: Fig. 1 shows that LSRDP is a matrix of PEs in which the height and width of LSRDP are the number of rows and columns, respectively. The LSRDP height and width are two important parameters which are determined during the design procedure and directly affect the number of resources and area of LSRDP.

Operand routing network (ORN): PEs of each row are connected to the PEs in the next row through ORNs as routing resources. The maximum connection length (MCL) is defined as the maximum horizontal distance of two PEs located in two consequent rows (Fig. 2). ORN size is determined base on the MCL value. The number of ORNs and their size affect the LSRDP area, energy consumption and its implementation cost as well.

ORNs' functionality is similar to a multiplexer however; ORNs are implemented as cross-bar switches. An ORN implementation has been introduced in [1], and here is a brief description on it. In the porposed architecture, the requirements of an ORN are as follows:

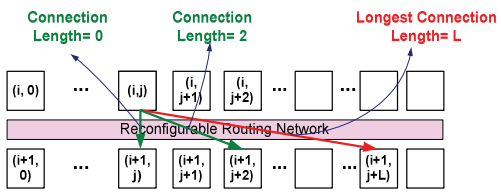

Fig. 2. Definition of the connection length and the maximum connection length (MCL)

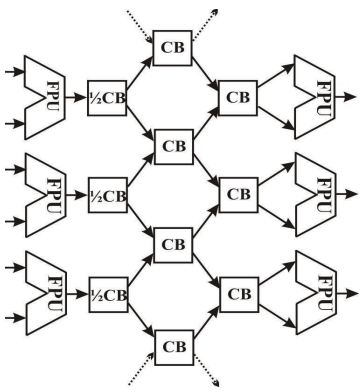

Fig. 3. Block diagram of a crossbar-based ORN

- each PE can be connected to one or more PEs in the next row;
- connections are among the PEs in the immediate vicinity of each other and the maximum number of the connections, *N*, is odd
- an FU output can be connected to either or both inputs of a PE in the next row.

To implement an ORN, crossbar switches are used as shown in Fig. 3. Due to the checker type arrangement of the crossbars, this type of architecture is naturally suitable for the ORN with an odd number of connections per FU. In order to support a multicasting network, the crossbar switches must be capable of performing two more functions in addition to 'cross' and 'bar': multicasting of either of the inputs. ½ CB is a crossbar switch with only one input, from which data can be sent to either or both outputs. The crossbar-based ORN has a regular pipelined structure that does not limit the performance of the LSRDP and can be reconfigured on the fly. It can also be easily redesigned for any given complexity by adding a necessary number of extra rows of crossbars.

Reconfiguration mechanism: LSRDP is a reconfigurable hardware that can be programmed within run-time. Upon reaching to a critical segment during application execution, a reconfiguration phase starts and the LSRDP configurable components including ORNs, immediate registers and PEs are reconfigured. Then, a DFG corresponding to the critical part of application is executed on LSRDP. To eliminate or to reduce the reconfiguration overhead time a pre-configuration can be performed, therefore after finishing the configuration stage and before the LSRDP operation, a wait state is required. In this manner, the reconfiguration phase would be overlapped with the GPP execution.

Fig. 4 shows the architecture of a PE and how it can be reconfigured during the configuration phase. Apart initializing immediate registers, the multiplexers, PEs and ORN micro-routing network should also be programmed using the configuration bits. A serial chain is used for configuring immediate registers, PEs and ORNs. In order to configure each component, the configuration bit-stream is serially transferred to the configuration registers. It might take a hundreds of cycles with respect to the configuration bit-stream size which directly depends on the LSRDP specifications.

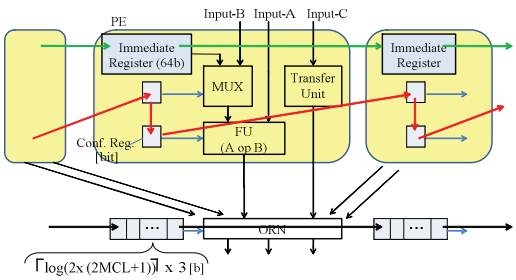

Fig. 4. Detailed architecture of a reconfigurable PE

External Memory: It is assumed that sixteen memory modules of 1800Mbps/pin are used in the memory array [2]. Each memory module uses one channel for input and the other channel for output. Data bus width for transferring data is 64bit and double-precision data (8 bytes) is subjected to the computations. Therefore, the data transfer rate is almost 24GB/s.

3. Preliminary Evaluation

To evaluate performance of the proposed LSRDP-based computer, a preliminary evaluation was accomplished based on simple analytical analysis and simulation. A base processor with the characteristic displayed in Table 1 was used. Also, Table 2 shows configuration of target reconfigurable processor comprising GPP and LSRDP.

Four applications are attempted as benchmark scientific applications including: one-dimensional heat (referred as Heat) and vibration equations (Vibration), two-dimensional Poisson equation (Poisson) [4], and recursion calculation part of electron repulsion integral (ERI [3]) as a quantum chemistry application. All calculations consist of ADD, SUB, and MUL operations.

In this paper, the experiments have been conducted for two applications Heat and Poisson. For each application, various sizes of DFGs (small-S, medium-M and large-L) are generated through expanding a basic DFG of each application. The larger DFG, more precise computations are possible. Fig. 5 and Fig. 6 show the performance on the GPP+LSRDP for abovementioned applications, respectively. Vertical axis denotes the normalized execution time (ratio of execution time on GPP+LSRDP to the execution time on GPP). In the figures, a breakdown of the execution time can be seen. For Poisson, in the 'Basic' bar, the largest portion of execution time (referred as 'Rearrange') is the time required for data rearrangement. Input/output data should be arranged in a proper order for the next execution step on the LSRDP. The second major fraction ('Stall') is the time elapsed for transferring data from scratchpad to main memory and vice versa. Fig. 5 depicts that the total execution time mainly consists of GPP time ('GPP'), LSRDP calculation time ('LSRDP') and communication time between the LSRDP and scratchpad memory ('Comm.').

To enhance the overall performance, a data reusing technique is employed to avoid the need for data rearrangement as well as the necessity of frequently reloading data from scratch pad memory. By using this technique the achievable speedup is improved considerably (Fig. 5). As a matter of fact, in the Poisson, by increasing DFG size, smaller speedup is achieved. That is because of particular property of this application. A large or a small DFG of Poisson can be chosen on the LSRDP; however by using smaller DFG, it is required to execute fewer instructions on the GPP rather than LSRDP.

Fig. 6 depicts the performance evaluation for the Heat application while data reusing is exploited. By using a larger DFG from the Heat application, total execution time decreases, hence performance rises. Total execution time is mainly includes the GPP time, the LSRDP computation time

and the communication time between LSRDP and scratchpad memory.

According to the graphs, only a small fraction of the overall execution time belongs to processing time on LSRDP and the main fraction concerns to the various overhead times and execution time on GPP. Consequently, reducing above overhead times will strongly improve the achievable speedup.

Table 1. Configuration of the base processor

Processor type	Out-of-order	
GPP operating frequency	3.2GHz	
Inst. issue width	4 instruction/cc	
Inst. decode width	4 instruction/cc	
Cache configuration	L1 data	64KB(128B Entry, 2way, 2cc)
	L1 instruction	64KB(64B Entry, 1way, 1cc)
	L2 unified	4MB(128B Entry, 4way, 16cc)
Latency of main memory	300cc	
L2 to main memory	Bus width	64 Bytes
	Freq	800 MHz

Table 2. Configuration of the reconfigurable processor GPP+LSRDP)

LSRDP operating frequency	80 GHz
Reconfiguration Latency	1cc
Latency SPM ←→LSRDP latency	1cc
Latency Main Memory ←→SPM	7500cc
Bandwidth SPM←→LSRDP	Max. 64 * 8 Bytes/cc
Bandwidth Main Memory←→ SPM	102.4GB/sec

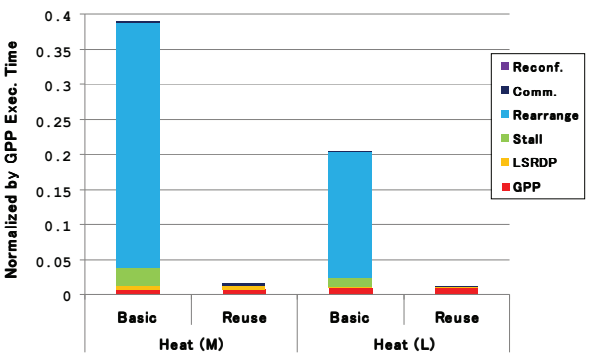

Fig. 5. Performance evaluation for Heat application

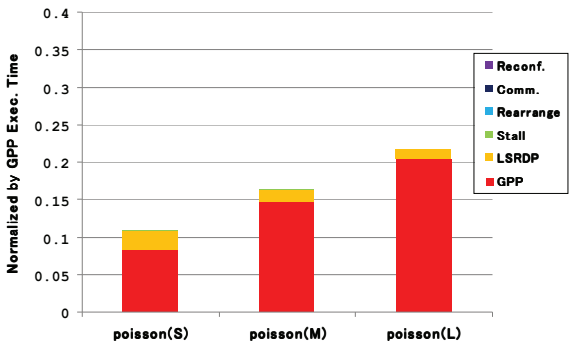

Fig. 6. Performance evaluation for Poisson application

4. Conclusion

A high-performance computer comprising an accelerator implemented by superconducting circuits was introduced. It is suitable for executing massive computational-intensive scientific applications. The LSRDP is the accelerator and it is implemented by means of SFQ circuits. Evidences from experiments demonstrate that the high-performance computer equipped with SFQ-LSRDP is promising for resolving issues originated from CMOS technology as well as achieving noticeable performances.

Acknowledgement

This research was supported in part by Core Research for Evolutional Science and Technology (CREST) of Japan Science and Technology Corporation (JST).

References

[1] A. Fujimaki, S. Iwasaki, K. Takagi, R. Kasagi, I. Kataeva, H. Akaike, M. Tanaka, N. Takagi, N. Yoshikawa, K. Murakami, "Demonstration of an SFQ-Based Accelerator Prototype for a High-Performance Computer," 2008 Applied Superconductivity Conference (ASC 2008), 2EZ01, Chicago, Aug 2008.

[2] Memory Roadmap, http://tw.renesas.co

[3] S. Obara and A. Saika, Efficient recursive computation of molecular integrals over Cartesian Gaussian Functions, J. Chem. Phys., Vol.84, pp.3963, 1986.

[4] W.H. Press, B.P. Flannery, S.A. Teukolsky, and T.W. Vetterling, Numerical Recipes in C, Cambridge University Press, 1988.

[5] N. Takagi, K. Murakami, A. Fujimaki, N. Yoshikawa, K. Inoue and H. Honda "Proposal of a desk-Side Supercomputer with Reconfigurable Data-Paths Using Rapid Single Flux Quantum Circuits," IEICE Trans. on Elec., E91-C(3):350-355, 2008.

[6]W. Wulf and S. McKee, Hitting the Memory Wall: Implications of the Obvious, ACM SIGArch Computer Architecture News, 23 (1):20-24, March 1995.