Mapping Scientific Applications on a Large-Scale Data-Path Accelerator Implemented by Single-Flux Quantum (SFQ) Circuits

Farhad Mehdipour¹, Hiroaki Honda², Hiroshi Kataoka¹, Koji Inoue¹, Irina Kataeva³, Kazuaki Murakami¹, Hiroyuki Akaike³, Akira Fujimaki³

¹ School of Information Science and Electrical Engineering, Department of Informatics, Kyushu University, Fukuoka, Japan
² Institute of Systems, Information Technologies and Nanotechnologies, Fukuoka, Japan
³ Department of Quantum Engineering, Nagoya University, Nagoya, Japan

Email: farhad@c.csce.kyushu-u.ac.jp
CREST-JST SFQ-RDP Project (2006~): A Low-power, high-performance reconfigurable processor based on single-flux quantum circuits

**Goals:**
- Discovering appropriate computation-intensive scientific applications
- Developing compiler tools
- Developing performance analyzing tools
- **Designing and Implementing the SFQ-LSRDP architecture**

Yokohama National Univ. SFQ-FPU chip, cell library

Nagoya Univ. SFQ-RDP chip, cell library, and wiring

Nagoya Univ. CAD for logic design and arithmetic circuits

Kyushu Univ. Architecture, Compiler and Applications

Superconducting Research Lab. (SRL) SFQ process

March 08-12, DATE 2010
Single-flux quantum (SFQ) vs. CMOS

- CMOS main issues in implementing a large accelerator:
  - High electric power consumption
  - High heat radiation
  - Difficulties in high-density packing

SFQ Features:
- High-speed switching and signal transmission
- Low power consumption
- Compact implementation (smaller area)
- Suitable for pipeline processing of data stream
Outline of the large-scale reconfigurable data-path (LSRDP) processor

Reconfigurable data-path components:
- A matrix of large number of floating-point Functional Units (FUs)
- Reconfigurable Operand Routing Network (ORN)
- Dynamic reconfiguration facilities
- Streaming Buffer (SB) for I/O ports

Features:
- Handling data flow graphs (DFGs) extracted from scientific applications
- Pipeline execution
- Burst transfer of input/output rearranged data from/to memory
- Reduced no. of memory accesses (to alleviate the memory wall problem)

Memory band width per MCM: 256GB/s
(=16GB/s \times 16 channels)
LSRDP architecture

- Processing Elements (PEs)
  - FU (Functional Unit) → 64-bit double-precision floating point operations including: ADD/SUB and MUL
  - TU (transfer unit) → routing resource for transferring data b/w inconsecutive rows

Diagram:
- Input ports
- Output ports
- Four various functionalities

March 08-12, DATE 2010
An overall perspective

---

(a) LRQP overall organization
(b) Layout types
M: MUL, A: ADD, BU: B

---

(d) PE's detailed structure
Immediate
Register (4 bits)
Immediate
Register
Conf. Reg.
MUX
Transf.
Reg
FU
(A + B)

(e) Definition of Maximum Connection Length (MUL)

(f) PE's various functionalities
• DFGs are manually generated

• DFG mapping results are employed throughout a quantitative design approach
Benchmark applications

- Finite differential method calculation of 2nd order partial differential equations
  - 1dim-Heat equation → Heat
  - 1dim-Vibration equation → Vibration
  - 2dim-Poisson equation → Poisson
- Quantum chemistry application
  - Electron Repulsion Integral calculation → ERI

**Ex.**: One-dim. heat equation for $T(x,t)$

$$\frac{\partial T(x,t)}{\partial t} = A \frac{\partial^2 T(x,t)}{\partial x^2}$$

Calculation by Finite Difference Method (FDM)

$$T(x_i,t_{j+1}) = D^* T(x_i,t_j) + B^* [T(x_{i-1},t_j) + T(x_{i+1},t_j)]$$
DFG mapping flow

- Placing DFG nodes on LSRDP
- Placing IO nodes
- Routing connections
- Routing Inp/Out connections
- Re-placing DFG nodes on LSRDP (considering IO nodes positions)
- Re-palcing output nodes

A mapping example

MCL = 3

March 08-12, DATE 2010
• In the above piece of map:
  • I1 and I9 are located far from each other
  • They are placed without noticing to the location of node 16 as their common child in the third row
  • Considering the location of common descendants will make them closer to each other
Proximity-factor based placement

- Proximity factor represents the amount of forces that two nodes exert to each other.
- \( S_{i,j} \): a set of common descendants for input nodes \( i \) and \( j \).
- \( D_{k;i}(=D_{k;j}) \): distance of common descendant node \( k \) to the input nodes \( i \) and \( j \).

\[
P = \begin{bmatrix}
\infty & p_{1,2} & \cdots & p_{1,n} \\
p_{2,1} & \infty & \cdots & p_{2,n} \\
\vdots & \vdots & \ddots & \vdots \\
p_{n,1} & p_{n,2} & \cdots & \infty
\end{bmatrix}
\]

\[
p_{i,j} = p_{j,i} = \begin{cases}
\infty & \text{if } i = j \\
\sum_{k \in S_{i,j}} \frac{1}{D_{k,i}} & \text{if } i \neq j
\end{cases}
\]

**Ex:**
\[
S_{1,2} = \{4, 6, 7\} \\
P(I_1, I_2) = p_{1,2} = \frac{1}{2} + \frac{1}{3} + \frac{1}{4} = \frac{13}{12}
\]
\[
S_{1,3} = \{7\} \\
P(I_1, I_3) = p_{1,3} = \frac{1}{4}
\]
\[
S_{2,3} = \{7\} \\
P(I_2, I_3) = p_{2,3} = \frac{1}{4}
\]

Inputs nodes \( I1 \) and \( I2 \) should be located in a closer distance rather than \( I3 \).
Input nodes placement alg.: Example

\[ C(l) = \sum_{i=l+1}^{r-1} p_{ij} \times \frac{1}{|i-l|} \]

\[ C(r) = \sum_{i=l+1}^{r-1} p_{ij} \times \frac{1}{|r-i|} \]

if \( C(l) > C(r) \) 

\( l = l+1, L[l] = j \)

else

\( r = r+1, L[r] = j \)

Placing the 1\(^{\text{st}}\) input node with the highest proximity factor

\[ \ldots \quad N/2-3 \quad N/2-2 \quad N/2-1 \quad 1 \quad N/2+1 \quad N/2+2 \quad N/2+3 \quad \ldots \]

Placing the 2\(^{\text{nd}}\) input node with the highest proximity factor

\[ \ldots \quad N/2-3 \quad N/2-2 \quad 2 \quad 1 \quad N/2+1 \quad N/2+2 \quad N/2+3 \quad \ldots \]

Placing \( i \)-th input node

\[ \ldots \quad N/2-K \quad \ldots \quad 2 \quad 1 \quad 3 \quad \ldots \quad N/2+M \quad \ldots \]

If \( C(l) > C(r) \):

\[ \ldots \quad i \quad \ldots \quad 2 \quad 1 \quad 3 \quad \ldots \quad N/2+M \quad \ldots \]

If \( C(r) > C(l) \):

\[ \ldots \quad N/2-K \quad \ldots \quad 2 \quad 1 \quad 3 \quad \ldots \quad i \quad \ldots \]

DFG nodes are placed using a slightly modified algorithm
Estimating the ORN area

(a)

\[ \text{Area(ORN)} = 1.5 \times W \times (4 \times MCL) \times \text{Area(CB)} \]

Number of rows = 1.5\( \times \)W

Number of columns = 4\( \times \)MCL

MCL = 1

Layout I: Area(PE) = 2.1 \times \text{Area(ADD)}
Layout II: Area(PE) = 1.1 \times \text{Area(ADD)}

(b)

\[ \text{Area(ORN)} = 2 \times W \times (6 \times MCL + 2) \times \text{Area(CB)} \]

Number of rows = 2\( \times \)W

Number of columns = 6\( \times \)MCL + 2

MCL = 1

Layout I: Area(PE) = 2.2 \times \text{Area(FU)}
Layout II: Area(PE) = 1.2 \times \text{Area(FU)}

(c)

\[ \text{Area(ORN)} = 1.5 \times M \times (4 \times MCL + 1) \times \text{Area(CB)} \]

Number of rows = 1.5\( \times \)M

Number of columns = 4\( \times \)MCL + 1

MCL = 2

Layout I: Area(PE) = 2.2 \times \text{Area(ADD)}
Layout II: Area(PE) = 1.2 \times \text{Area(ADD)}

Area(ORN-PEII) > Area(ORN-PEIII) > Area(ORN-PEI)
## Specifications of the benchmark DFGs

<table>
<thead>
<tr>
<th>DFG</th>
<th># of nodes</th>
<th># of inputs</th>
<th># of outputs</th>
<th># of pure ops</th>
<th>max. inp. nodes fan-out</th>
<th>Max. fan-out</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heat-8x1</td>
<td>34</td>
<td>6</td>
<td>4</td>
<td>16</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Heat-8x2</td>
<td>60</td>
<td>8</td>
<td>4</td>
<td>32</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Heat-16x2</td>
<td>172</td>
<td>16</td>
<td>12</td>
<td>96</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Poisson-3x3</td>
<td>62</td>
<td>18</td>
<td>1</td>
<td>33</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Vibration-4x2</td>
<td>48</td>
<td>8</td>
<td>4</td>
<td>24</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Vibration-8x2</td>
<td>136</td>
<td>16</td>
<td>12</td>
<td>72</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>ERI-1</td>
<td>20</td>
<td>8</td>
<td>3</td>
<td>9</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>ERI-2</td>
<td>76</td>
<td>16</td>
<td>9</td>
<td>51</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>ERI-3</td>
<td>89</td>
<td>14</td>
<td>9</td>
<td>66</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>ERI-4</td>
<td>67</td>
<td>19</td>
<td>1</td>
<td>47</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Max</td>
<td>172</td>
<td>19</td>
<td>12</td>
<td>96</td>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>
Evaluation results for various architectures

Connection length measurement:
- initial \( l_h = d_h \)
- modified \( l_{hv} = \frac{d_h}{d_v} \)

<table>
<thead>
<tr>
<th>nodes placement</th>
<th>Connection length measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>fan-out based</td>
</tr>
<tr>
<td>S2</td>
<td>proximity-factor based</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Layout-I</th>
<th>Layout-II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S1</td>
<td>S2</td>
</tr>
<tr>
<td>MCL</td>
<td>PE I</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>PE II</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>PE III</td>
<td>10</td>
</tr>
<tr>
<td>ORN size (overall)</td>
<td>PE I</td>
<td>25116</td>
</tr>
<tr>
<td>x CB</td>
<td>PE II</td>
<td>30600</td>
</tr>
<tr>
<td></td>
<td>PE III</td>
<td>18696</td>
</tr>
<tr>
<td>No. of PEs (overall)</td>
<td>PE I</td>
<td>580</td>
</tr>
<tr>
<td>x PE</td>
<td>PE II</td>
<td>634</td>
</tr>
<tr>
<td></td>
<td>PE III</td>
<td>627</td>
</tr>
<tr>
<td>Overall LSRDP Area x (KJJ)</td>
<td>PE I</td>
<td>36923</td>
</tr>
<tr>
<td></td>
<td>PE II</td>
<td>48083</td>
</tr>
<tr>
<td></td>
<td>PE III</td>
<td>35307</td>
</tr>
</tbody>
</table>

LAYOUT II and PE III result in smallest overall area
Conclusions & Future Work

- SFQ-LSRDP is a basic core of a high-performance low-power computer which is implemented by SFQ circuits

- LSRDP architecture is designed by using a quantitative approach on the results of the mapping process of the applications’ DFGs.

Future Work:

- Exploring design space to achieve minimum possible area
- Evaluating performance for various applications
- Discovering various sorts of appropriate applications

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