Textbook: VLSI ARRAY PROCESSORS
S.Y. Kung
Prentice-Hall, Inc.

INSTRUCTOR: CHING-LONG SU
E-mail: kevinsu@twins.ee.nctu.edu.tw
Chapter 4

Systolic Array Processors
Outline of Chapter 4

4.1 Introduction
4.2 Systolic Array Processors
4.3 Mapping DGs and SFGs to Systolic Arrays
4.4 Performance Analysis and Design Optimization
4.5 Systolic Arrays for the Transitive Closure and Dynamic Programming Problems
4.6 Systolic Design for Artificial Neural Network
4.7 Conclusion Remarks
4.8 Problems
4.1 Introduction

4.2 Systolic Array Processors

4.3 Mapping DGs and SFGs to Systolic Arrays

4.4 Performance Analysis and Design Optimization

4.5 Systolic Arrays for the Transitive Closure and Dynamic Programming Problems

4.6 Systolic Design for Artificial Neural Network

4.7 Conclusion Remarks

4.8 Problems
4.1 Introduction

1. Review the algorithm mapping onto SFG methodology
2. Discuss the cut-set systolization (retiming) method for systolic array design
4.1 Introduction
4.2 Systolic Array Processors
4.3 Mapping DGs and SFGs to Systolic Arrays
4.4 Performance Analysis and Design Optimization
4.5 Systolic Arrays for the Transitive Closure and Dynamic Programming Problems
4.6 Systolic Design for Artificial Neural Network
4.7 Conclusion Remarks
4.8 Problems
1. “A systolic system is a network of processors which rhythmically compute and pass data through the system.”

2. Systolic Array Processor avoids the classic memory access bottleneck problem.

3. Systolic Array Processor can solve the compute-bound and I/O-bound computations.
4.2 Systolic Array Processors

- Basic Configuration of Systolic Array

Diagram showing the basic configuration of a systolic array with Processing Elements (PE) and Memory.
4.2 Systolic Array Processors

Definition of Systolic Arrays

1. Synchrony: The data are rhythmically computed (timed by a global clock) and passed through the network.
2. Modularity and Regularity
3. Spatial Locality and Temporal Locality
4. Pipelinability
4.2 Systolic Array Processors

Example 1: Systolic Array for Convolution

\[ a_{out} = a_{in} \]
\[ b_{out} = b_{in} + a_{in} * W_i \]
4.2 Systolic Array Processors

Example 2: Hexagonal Systolic Array for Band Matrix Multiplication
4.2 Systolic Array Processors

Properties of Systolic Array

1. Simple and Regular Design
2. Concurrency and Communication
3. Balancing Computation with I/O
Clock Distribution Scheme for Synchronization of the Systolic Array System: H-tree Layout for the Balance of the Clock Circuit Delay

Linear Array  Square Array  Hexagonal Array
4.2 Systolic Array Processors

Systolic vs. SIMD vs. SFG Arrays

- Control Unit (Central Control)
  - Control Bus
  - Interconnection Network (Local)
    - Data Bus
  - Processing Uuit
    - Global Communication
  - SIMD Array
4.2 Systolic Array Processors

- Systolic vs. SIMD vs. SFG Arrays

Systolic Array

- Interconnection Network (Local)
- Control Unit
  - Processing Unit

Diagram shows the components of a systolic array processor with interconnections between the control units and processing units.
4.2 Systolic Array Processors

- Systolic vs. SIMD vs. SFG Arrays

SFG Array
4.1 Introduction
4.2 Systolic Array Processors
4.3 Mapping DGs and SFGs to Systolic Arrays
4.4 Performance Analysis and Design Optimization
4.5 Systolic Arrays for the Transitive Closure and Dynamic Programming Problems
4.6 Systolic Design for Artificial Neural Network
4.7 Conclusion Remarks
4.8 Problems
Three Stage of Canonical Mapping Algorithm for Systolic Array Design

1. Derive a (local) DG from the Algorithm
2. Map the DG to an SFG Array
3. Transform the SFG to a Systolic Array (i.e. retiming)
The major systolic array design gap is that most SFGs are not given in temporally localized form.

Systolic Array = SFG Array + Pipeline Retiming
4.3 Mapping DGs and SFGs to Systolic Arrays

- **Cut-Set Retiming Procedure**

1. **Timing Scale:** All delays $\mathbf{D}$ may be scaled, i.e. $\mathbf{D} \rightarrow \alpha \mathbf{D}$ (I/O Down Sample)

2. **Delay Transfer:** Given a cut-set of the SFG, which partitions the graph into two components, we can group the edges of the cut-set into two inbound edges and outbound edges.
4.3 Mapping DGs and SFGs to Systolic Arrays

- Data Transfer Rule

Inbound

Cut

Outbound

+kD

-kD
4.3 Mapping DGs and SFGs to Systolic Arrays

- **Systolization Procedure**

1. Selection of Basic Operation Modules
2. Applying Retiming Rules
3. Combination of Delay and Operation Modules
4.3 Mapping DGs and SFGs to Systolic Arrays

- **Systolization Procedure: Example of Lattice Filters**

![Diagram of Systolic Array]

- **Critical Path = 6 MAC**

- **Step 1. Time-Rescaled SFG for AR Lattice Filter**
4.3 Mapping DGs and SFGs to Systolic Arrays

- **Systolization Procedure: Example of Lattice Filters**

**Step 2. Retiming SFG for AR Lattice Filter**

- **Step 3. Systolic Array for AR Lattice Filter**
4.3 Mapping DGs and SFGs to Systolic Arrays

- Systolization Procedure:
  Example of Matrix Multiplication
All systolic arrays obtained from linear projections of the DG can be derived by the following two steps.

1. Mapping the DG onto SFGs by the SFG Projection Procedure
2. Mapping the SFG onto a Systolic Array by the Cut-Set Retiming
4.3 Mapping DGs and SFGs to Systolic Arrays

Retiming in the Sorting Systolic Arrays: Insertion Sorter
4.3 Mapping DGs and SFGs to Systolic Arrays

- Retiming in the Sorting Systolic Arrays: Selection Sorter

\[
\hat{d} = \hat{s} = [0,1]
\]
Retiming in the Sorting Systolic Arrays: Bubble-Sorter

\[ d = s = [1,1] \]
4.3 Mapping DGs and SFGs to Systolic Arrays

- Rotation of Schedule Vector for Insertion Sorter

\[ d = s = [1, 0] \]
4.3 Mapping DGs and SFGs to Systolic Arrays

Bit Level Systolic Arrays:
Example of Inner Product of Two Vectors

The inner product vector $c$ of two vectors $a$ and $b$ is computed as:

$$c = \sum_{k=1}^{n} a_k b_k$$

Assume that elements of $a$ and $b$ are $m$-bit integer.
4.3 Mapping DGs and SFGs to Systolic Arrays

**Bit Level Systolic Arrays:**

Example of Inner Product of Two Vectors

\[ c = \sum_{k=1}^{n} a_k \times b_k \]

\[ = a_1 \times b_1 + a_2 \times b_2 + \ldots + a_n \times b_n \]

\[ = 2^j \sum_{j=0}^{m-1} a_{1,j} \times 2^j \sum_{j=0}^{m-1} b_{1,j} + \ldots + 2^j \sum_{j=0}^{m-1} a_{n,j} \times 2^j \sum_{j=0}^{m-1} b_{n,j} \]
4.3 Mapping DGs and SFGs to Systolic Arrays

- **Bit Level Systolic Arrays:**

Example of Inner Product of Two Vectors

\[
c = \sum_{k=1}^{n} a_k \times b_k
\]

\[
= a_1 \times b_1 + a_2 \times b_2 + \ldots + a_n \times b_n
\]

\[
= 2^j \sum_{j=0}^{m-1} a_{1,j} \times 2^j \sum_{j=0}^{m-1} b_{1,j} + \ldots + 2^j \sum_{j=0}^{m-1} a_{n,j} \times 2^j \sum_{j=0}^{m-1} b_{n,j}
\]
4.3 Mapping DGs and SFGs to Systolic Arrays

- Bit Level Systolic Arrays: Example of Inner Product of Two Vectors