

## Optimizing the architecture of SFQ-RDP (Single Flux Quantum- Reconfigurable Datapath)

Mehdipour, Farhad

Graduate School of Information Science and Electrical Engineering, Kyushu University

Honda, Hiroaki

Institute of Systems, Information Technologies and Nanotechnologies (ISIT)

Kataoka, Hiroshi

Graduate School of Information Science and Electrical Engineering, Kyushu University

Inoue, Koji

Graduate School of Information Science and Electrical Engineering, Kyushu University

他

<https://doi.org/10.15017/14877>

---

出版情報 : SLRC プレゼンテーション, pp.1-, 2009-06-15. 九州大学システムLSI研究センター  
バージョン :  
権利関係 :



# Optimizing the Architecture of SFQ-RDP (Single Flux Quantum-Reconfigurable Datapath)

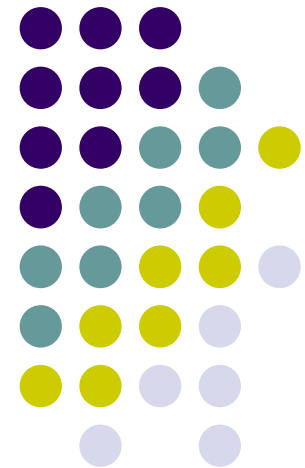
---

**F. Mehdipour\***, Hiroaki Honda\*\*, H. Kataoka\*, K. Inoue\* and K. Murakami\*

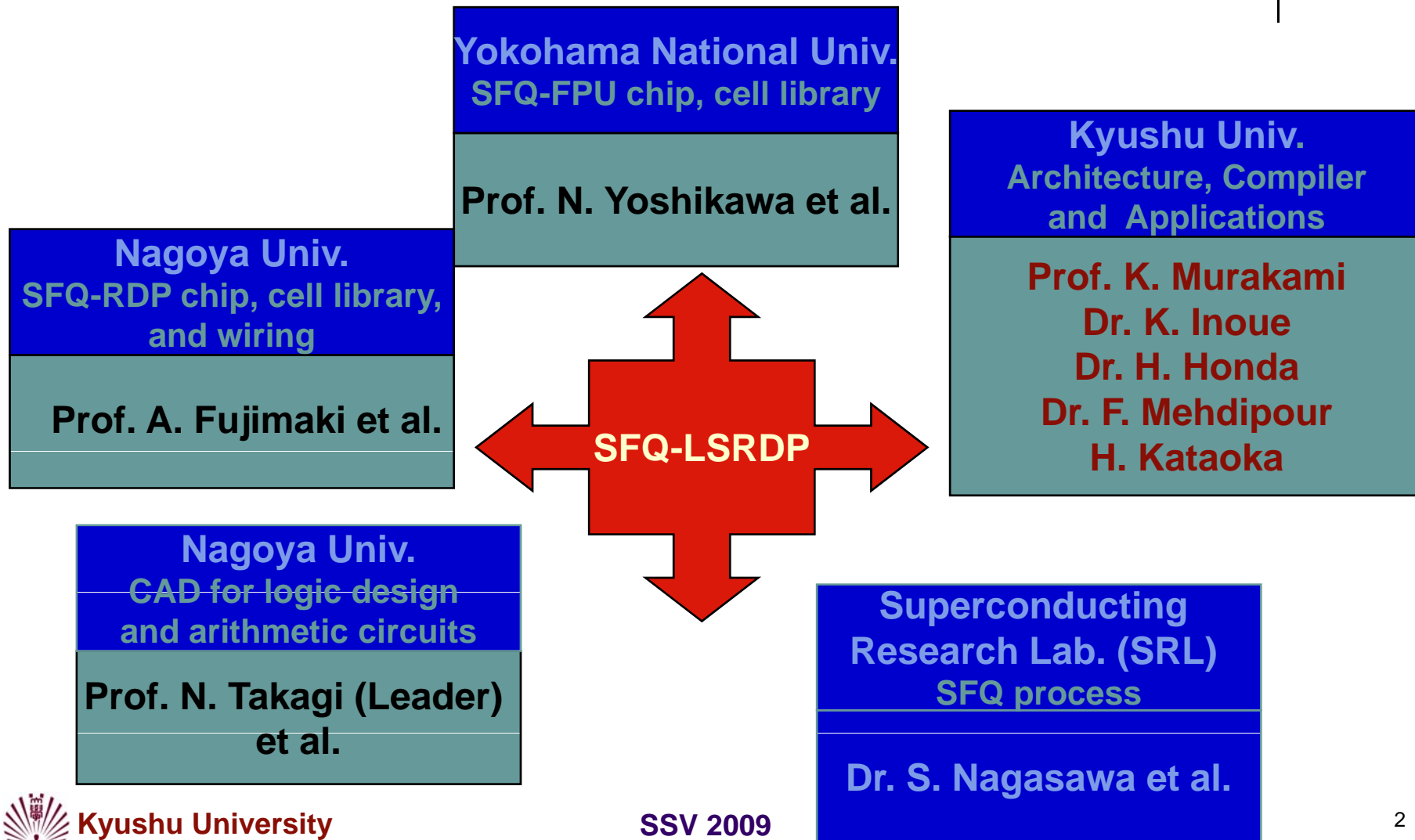
\*Graduate School of Information Science and Electrical Engineering, Kyushu University, Japan

\*\*Institute of Systems, Information Technologies and Nanotechnologies (ISIT), Fukuoka, Japan

E-mail: [farhad@c.csce.kyushu-ua.c.jp](mailto:farhad@c.csce.kyushu-ua.c.jp)



# CREST-JST (2006~): Low-power, high-performance, reconfigurable processor using single-flux quantum circuits

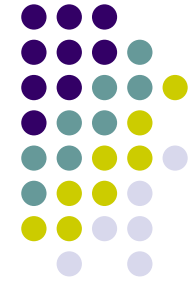


# Agenda



- Introduction
- Large-Scale Reconfigurable Data-Path (LSRDP)  
General Architecture and Specifications
- Design Procedure and Tool Chain
- Preliminary Results
- Conclusions and Future Work

# Introduction



- For performance improvement various accelerators are used with GPPs
  - PowerXcell, GPU, GRAPE-DR, ClearSpeed, etc.
  - Small size and low power consumption comparing to processors with similar performance

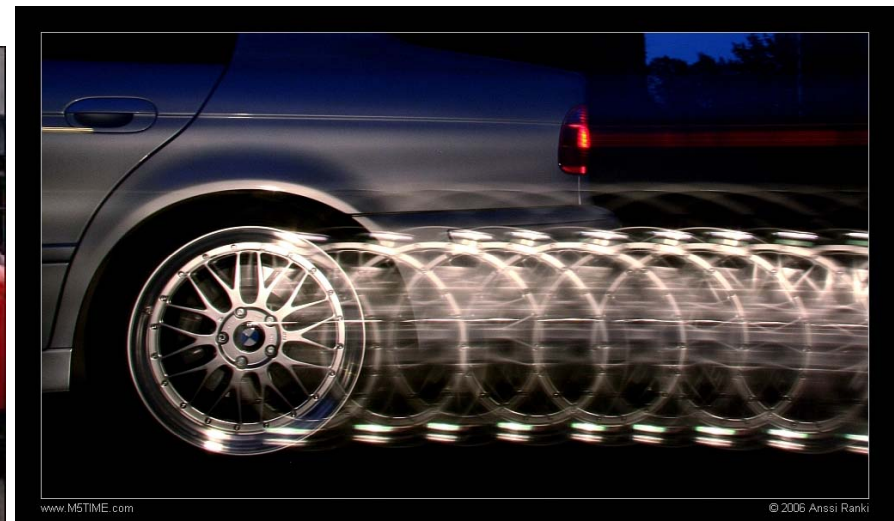


NVIDIA Tesla S1070  
<http://www.nvidia.com>

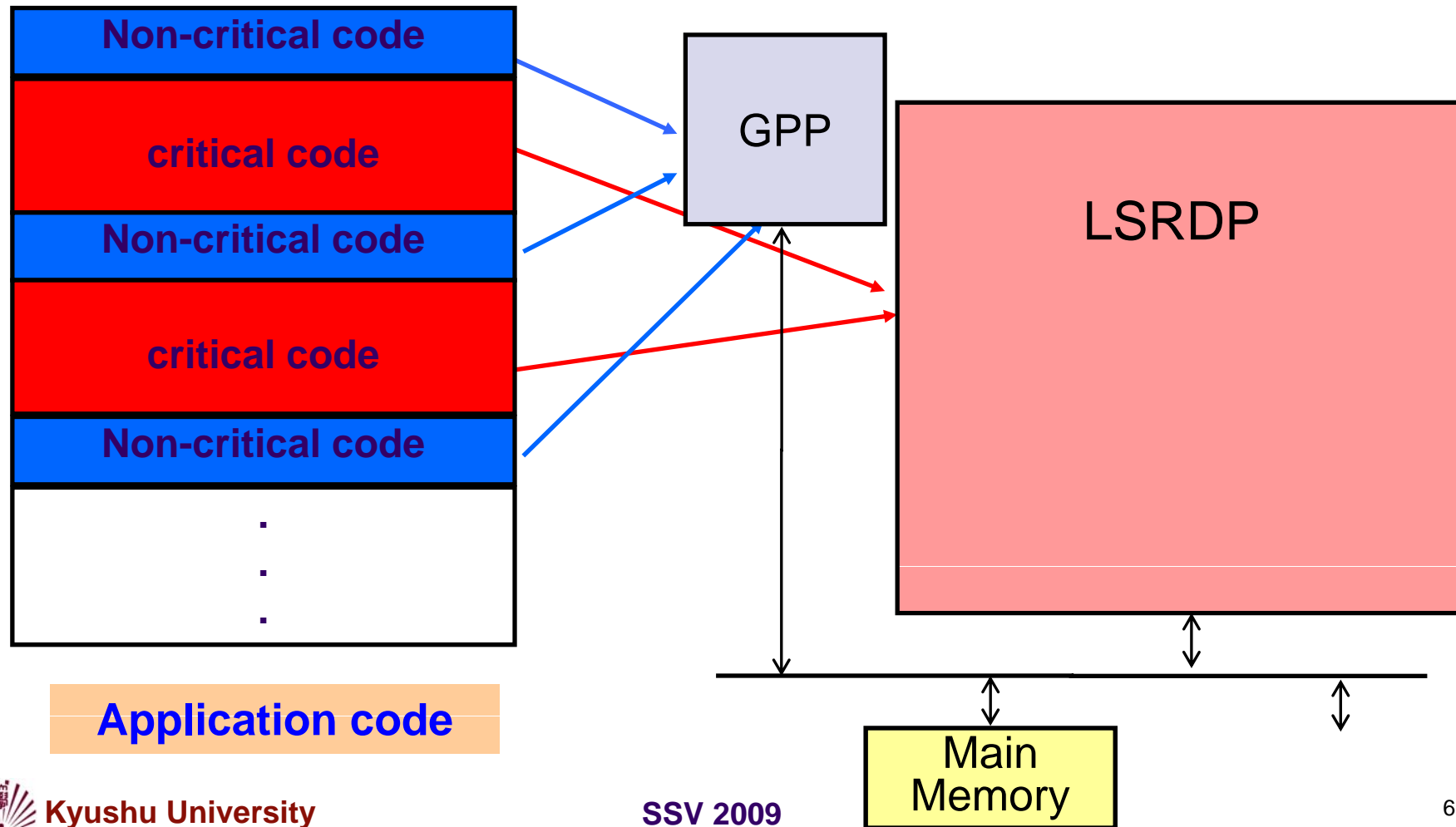
# Acceleration Through a Data-Path Processor



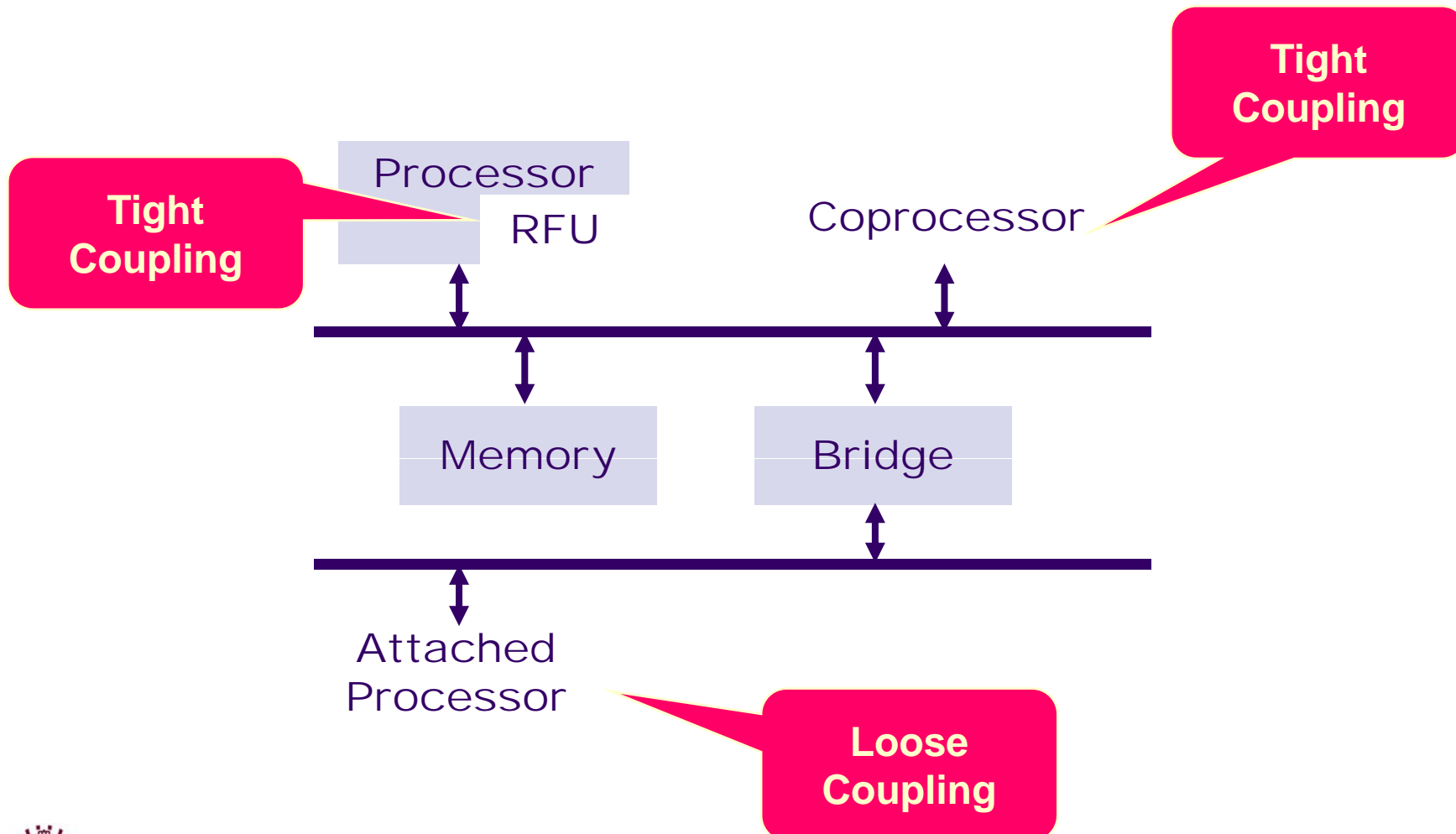
- Mechanism
  - Acceleration by using a data-path accelerator
  - Augmenting the accelerator to the base processor
  - Executes hot portions of applications on the accelerator



# How a Reconfigurable Processor Works



# Coupling an Accelerator to a Processor

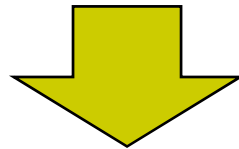


# Motivation



## Conventional accelerators:

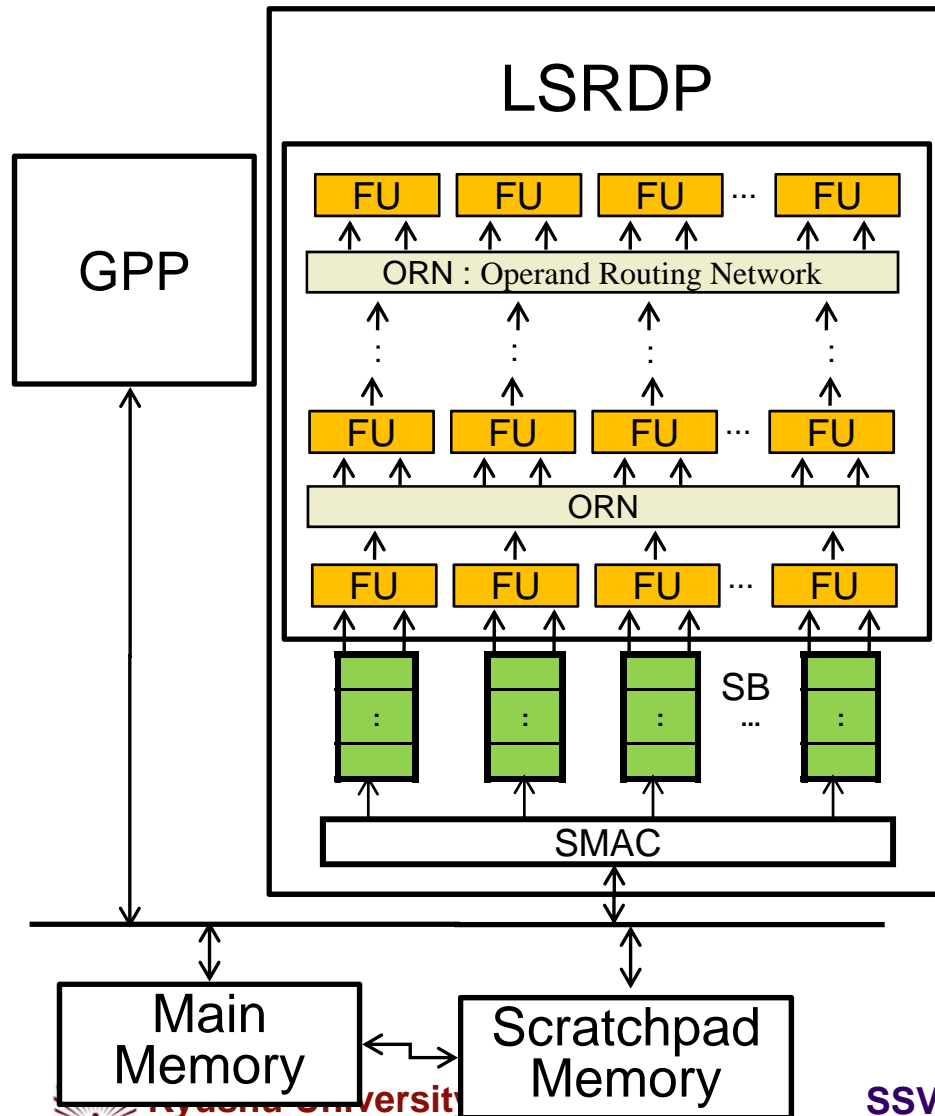
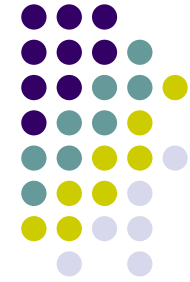
- A large memory bandwidth is demanded in conventional accelerators for high-performance computation
- On chip memories are often used to hide memory access latency



## Large-Scale Reconfigurable Data-Path (LSRDP):

- is introduced as an alternative accelerator
- reduces the no. of memory accesses by utilizing data-path

# Outline of Large-Scale Reconfigurable Data-Path (LSRDP) processor



- Reconfigurable data-path includes:

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

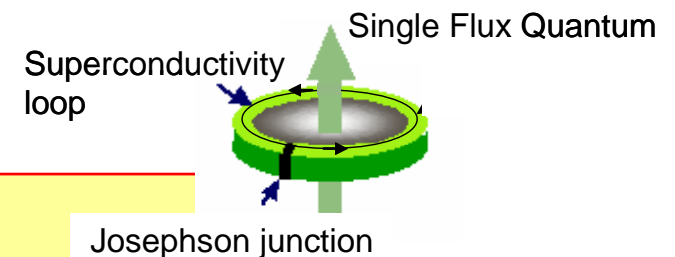
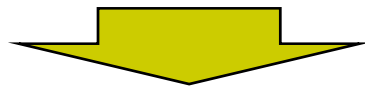
- Features:

- Data Flow Graphs (DFGs) extracted from critical calculation parts are directly mapped
- Pipeline execution
- Burst transfer is used for input /output rearranged data from/to memory

# Single-Flux Quantum (SFQ) against CMOS



- CMOS issues: (*if LSRDP has 32x32 FUs*)
  - high electric power consumption
  - high heat radiation and difficulties in high-density packing



- SFQ Features:
  - **High-speed switching** and signal transmission
  - **Low power consumption**
  - Compact implementation of a system (small area)
  - No cost for latch
  - Suitable for pipeline processing of data stream
  - **Serial bit-level processing**

# Goals of the Project

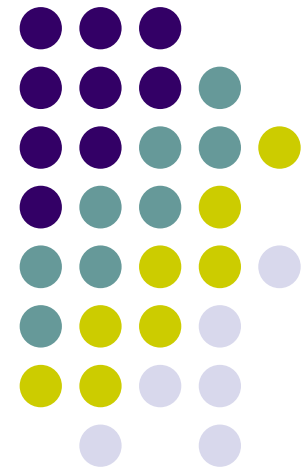


- Discovering appropriate scientific applications
- Developing compiler tools
- Developing performance analyzing tools

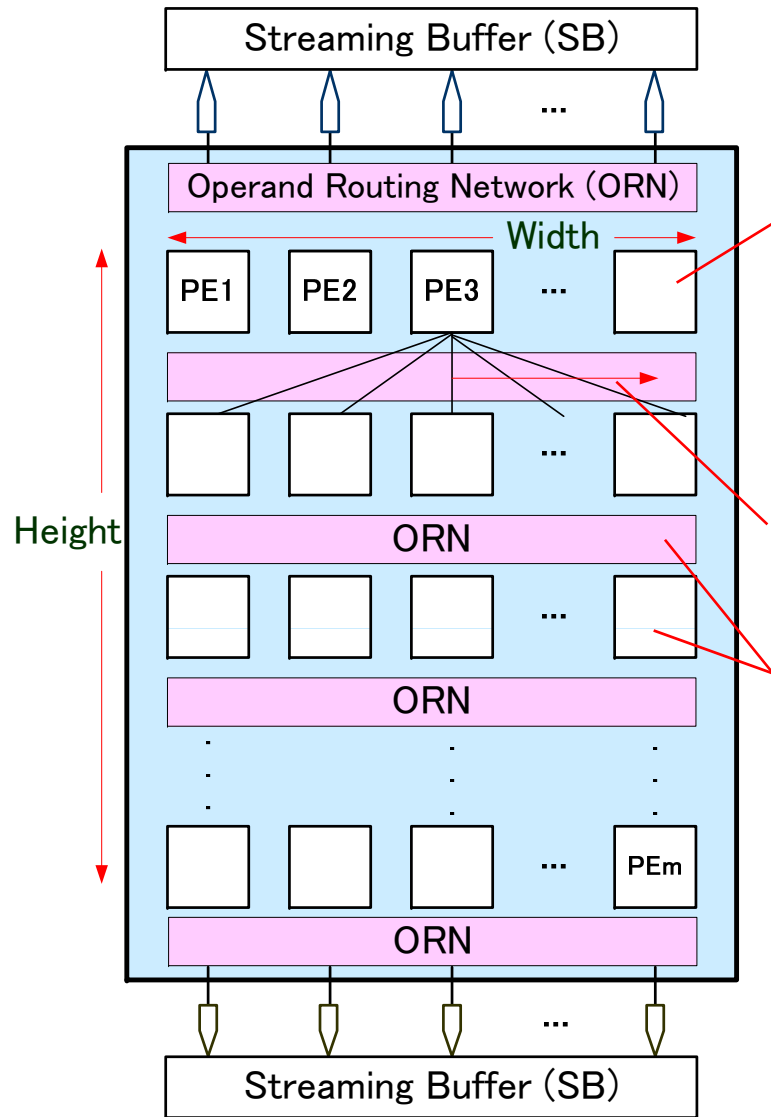
**Designing and Implementing SFQ-LSRDP architecture considering the features and limitations of SFQ circuits**

# LSRDP General Architecture and Specifications

---



# Parameters Should Be Decided Within the LSRDP Design Procedure



- Core structure: a rectangular matrix of PEs
- PE: combination of a Functional Unit (FU) and a data Transfer Unit (TU)

Width and Height ?

Maximum Connection Length (MCL) between consecutive rows?  
(impossible to implement full cross bar)

Layout: FU types  
(ADD/SUB and MUL)?

Reconfiguration mechanism?  
(PE, ORN, Immediate data)

- On-chip memory configuration?

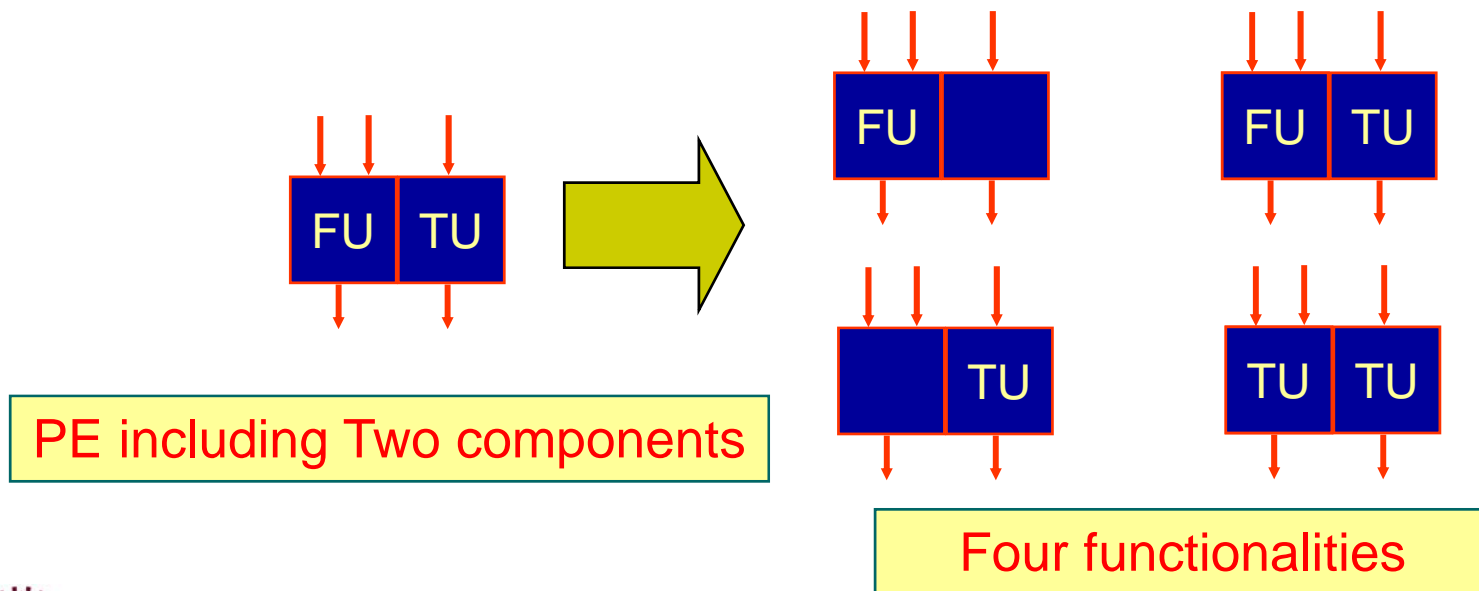


# LSRDP Architecture

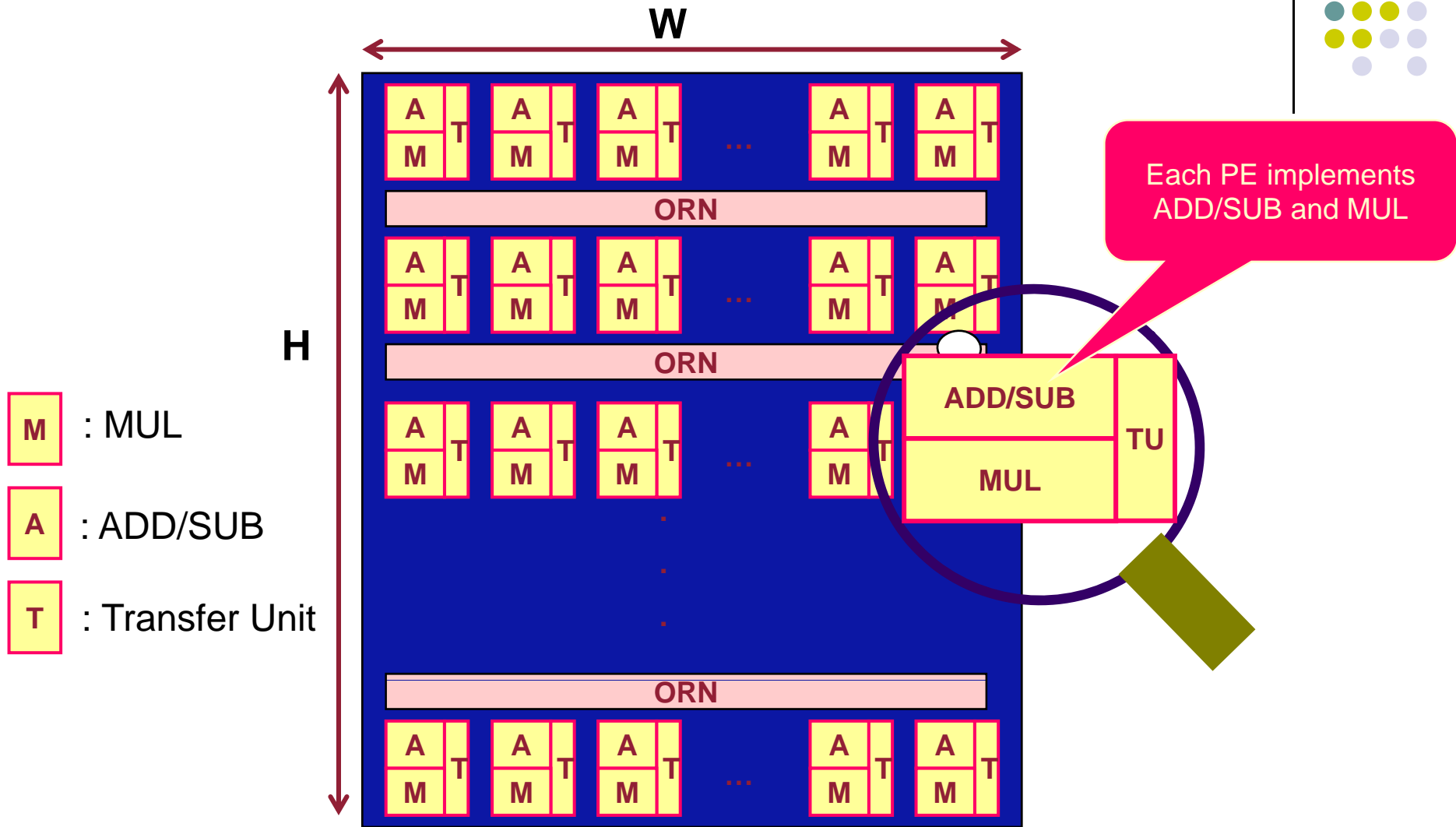


- **Processing Elements**

- FU
  - implements basic 64-bit double-precision floating point operations including: **ADD, SUB and MUL**
- TU (transfer unit) as a routing resource for transferring data from a row to an inconsecutive row

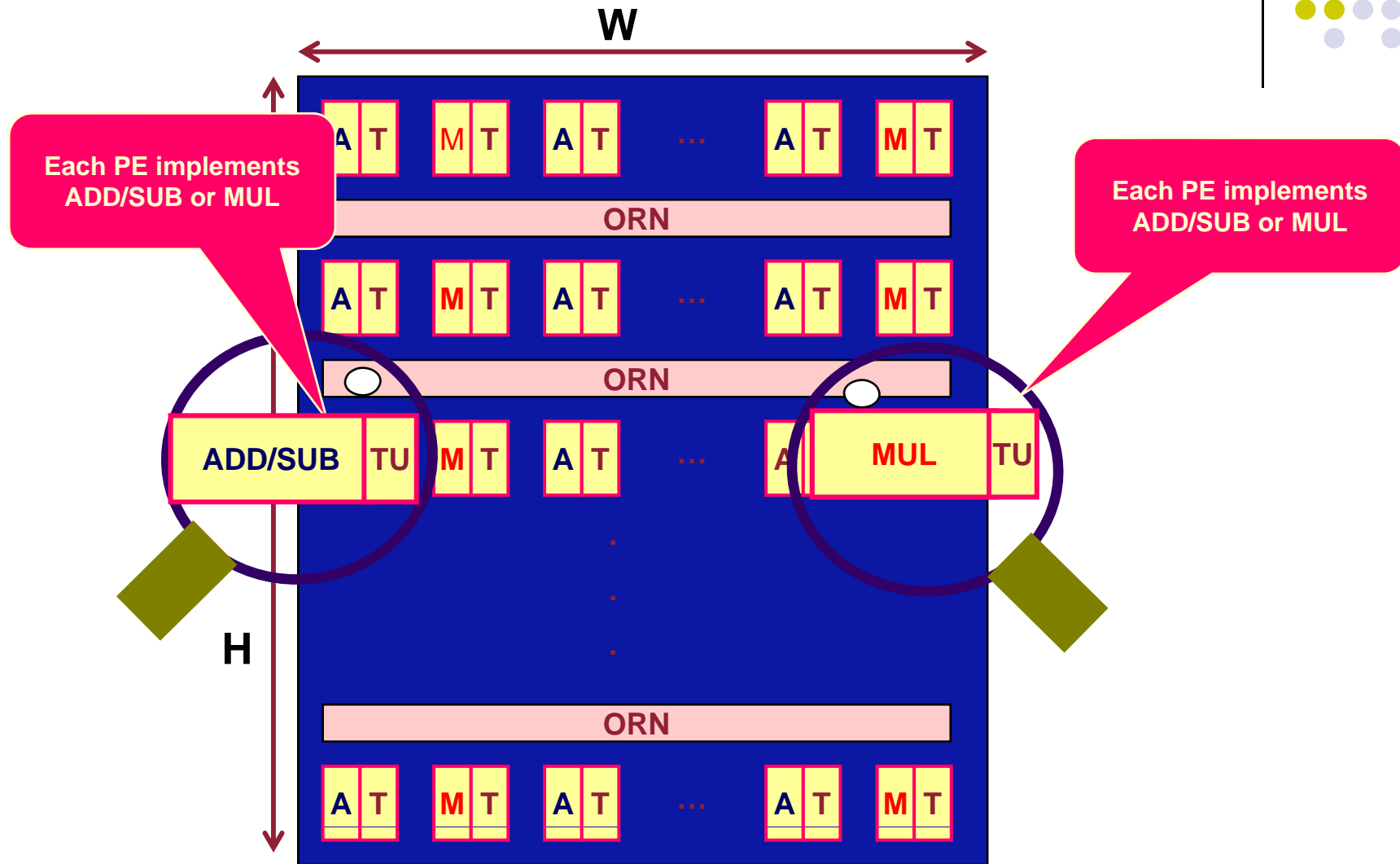


# Layout Types- Type I

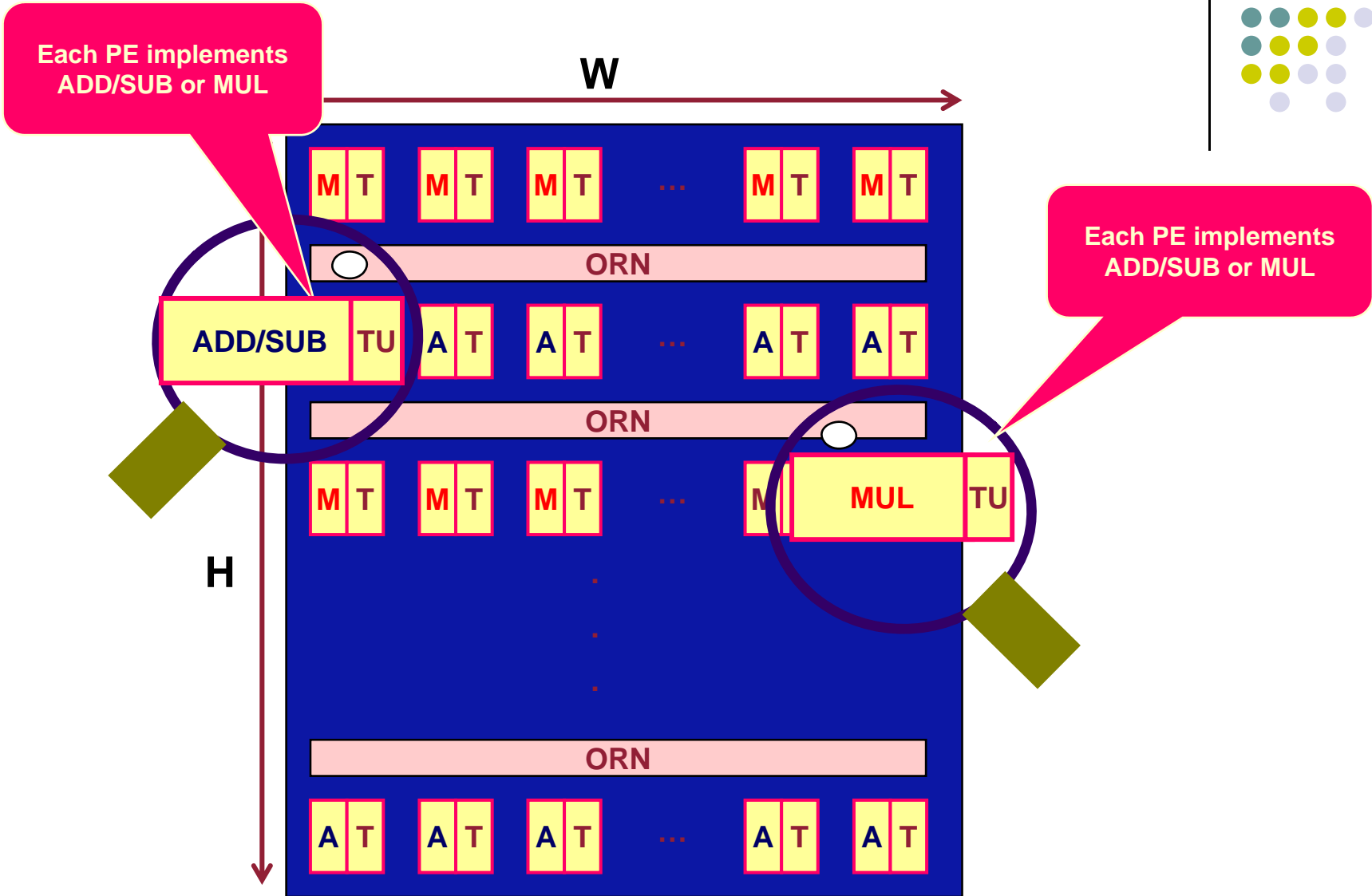


Flexible but consume a lot of resources

# Layout Types- Type II (Checked)

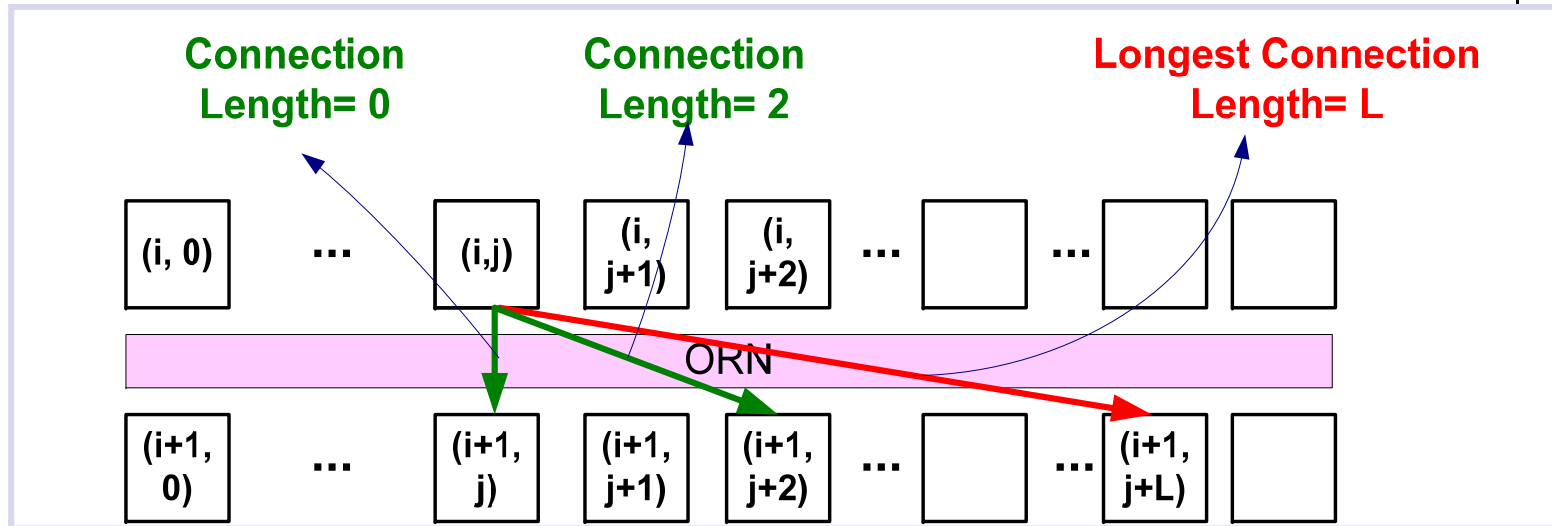


# Layout Types- Type III (Striped)



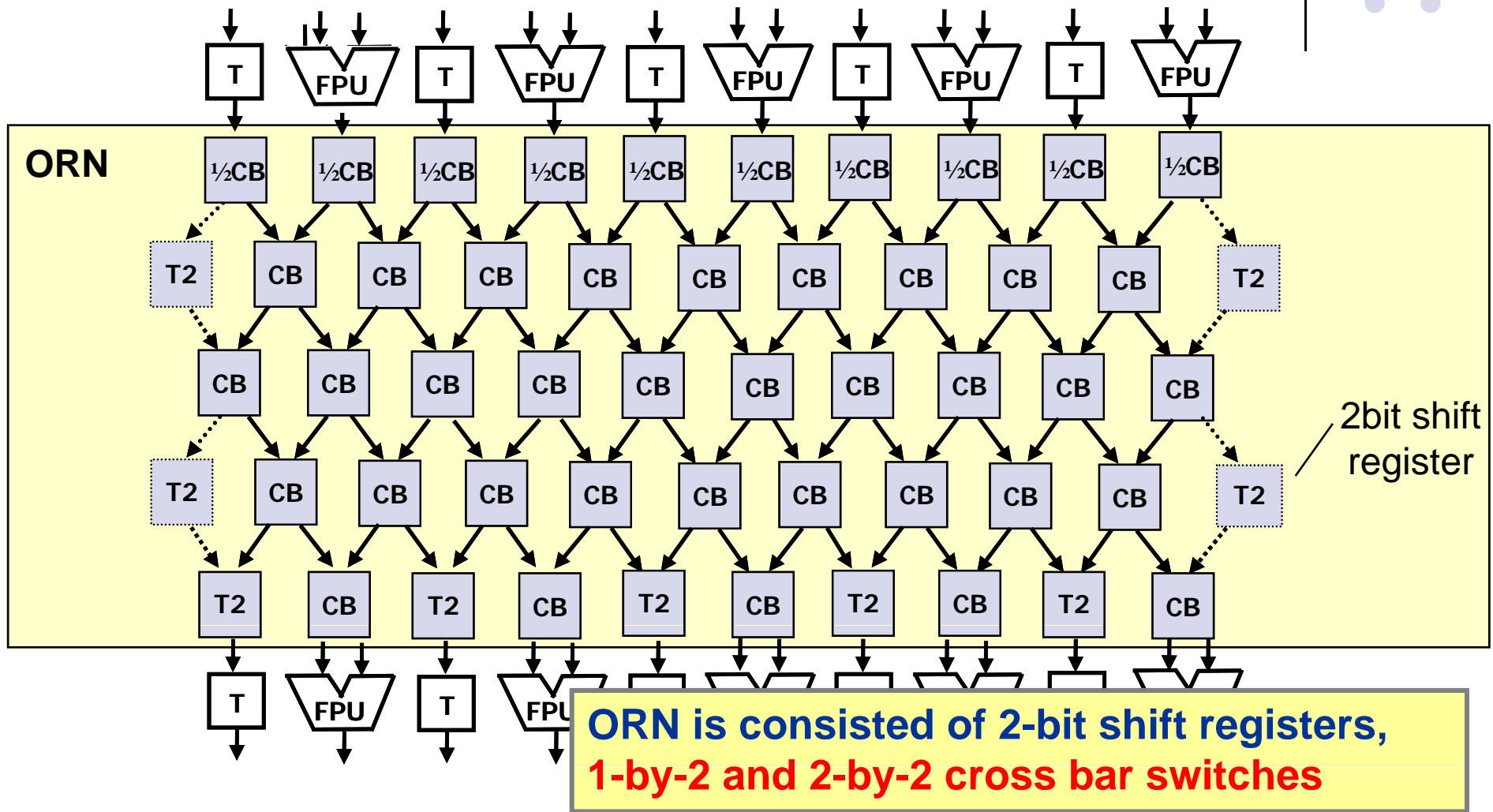
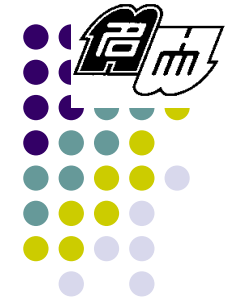
**Type II or III, which one is more efficient?**

# Maximum Connection Length (MCL)

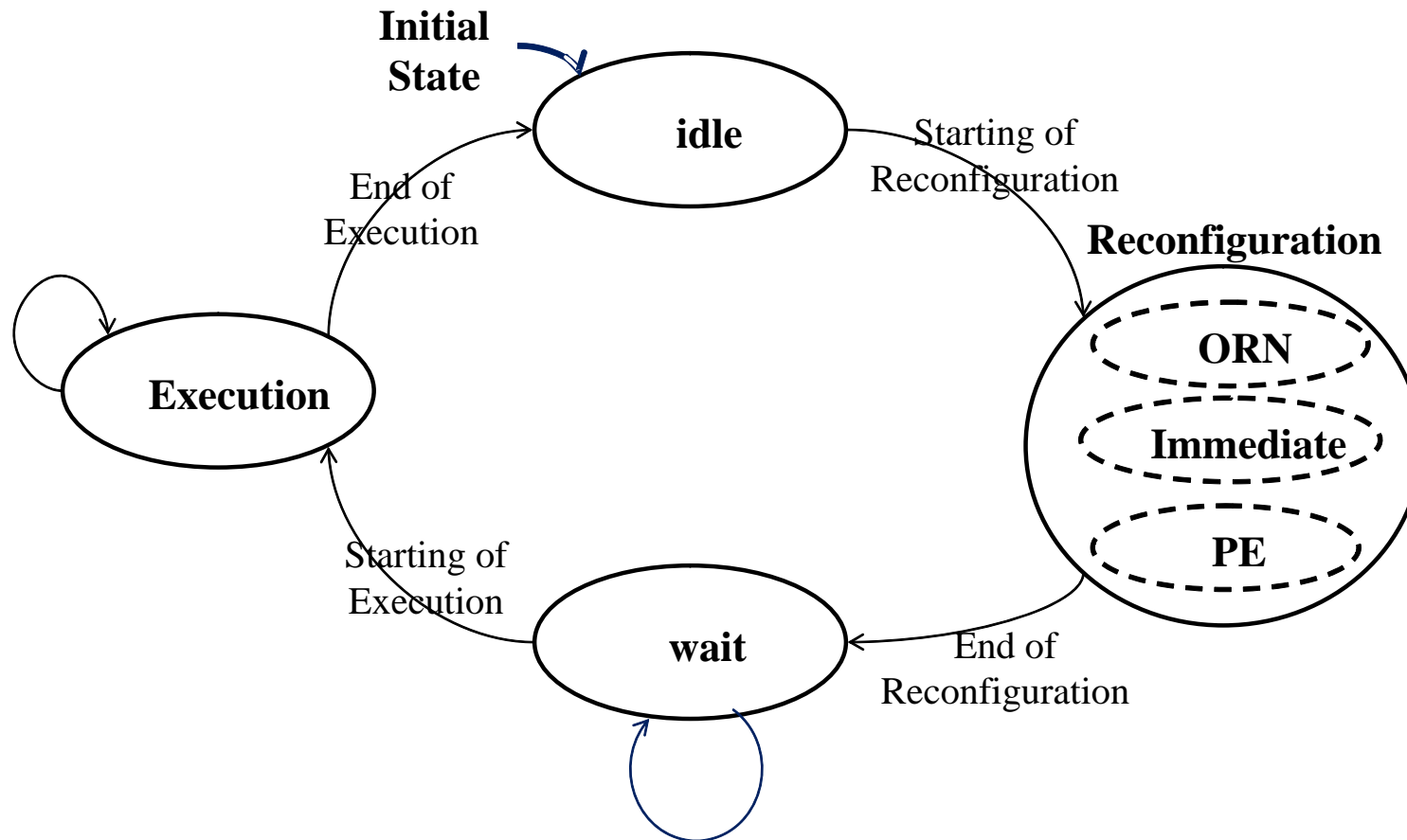


**MCL:**  
maximum horizontal distance  
between two PEs located in two subsequent rows

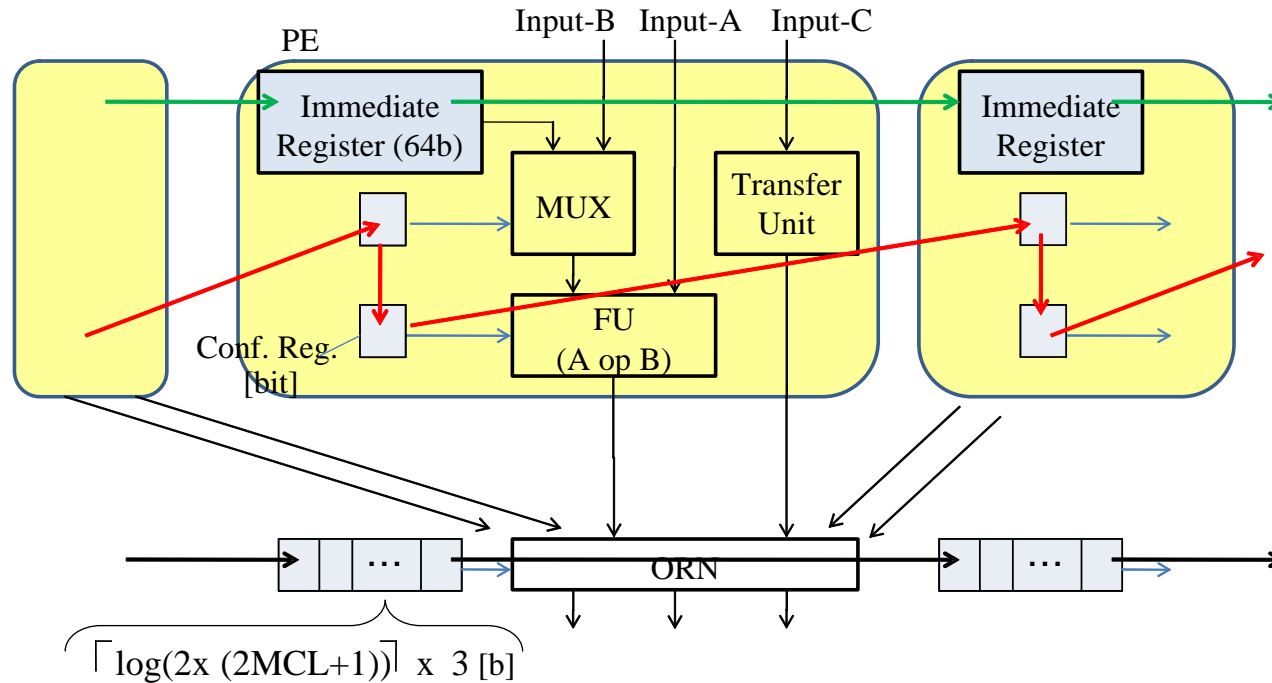
# An ORN Structure



# Dynamic Reconfiguration Mechanism



# Dynamic Reconfiguration Architecture

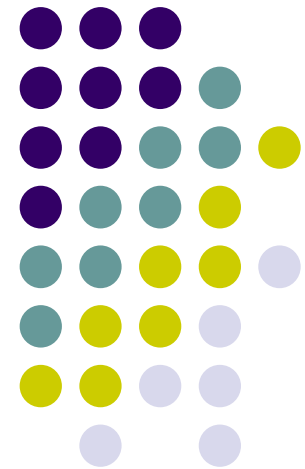


**Three bit-stream lines for dynamic reconfiguration of:**

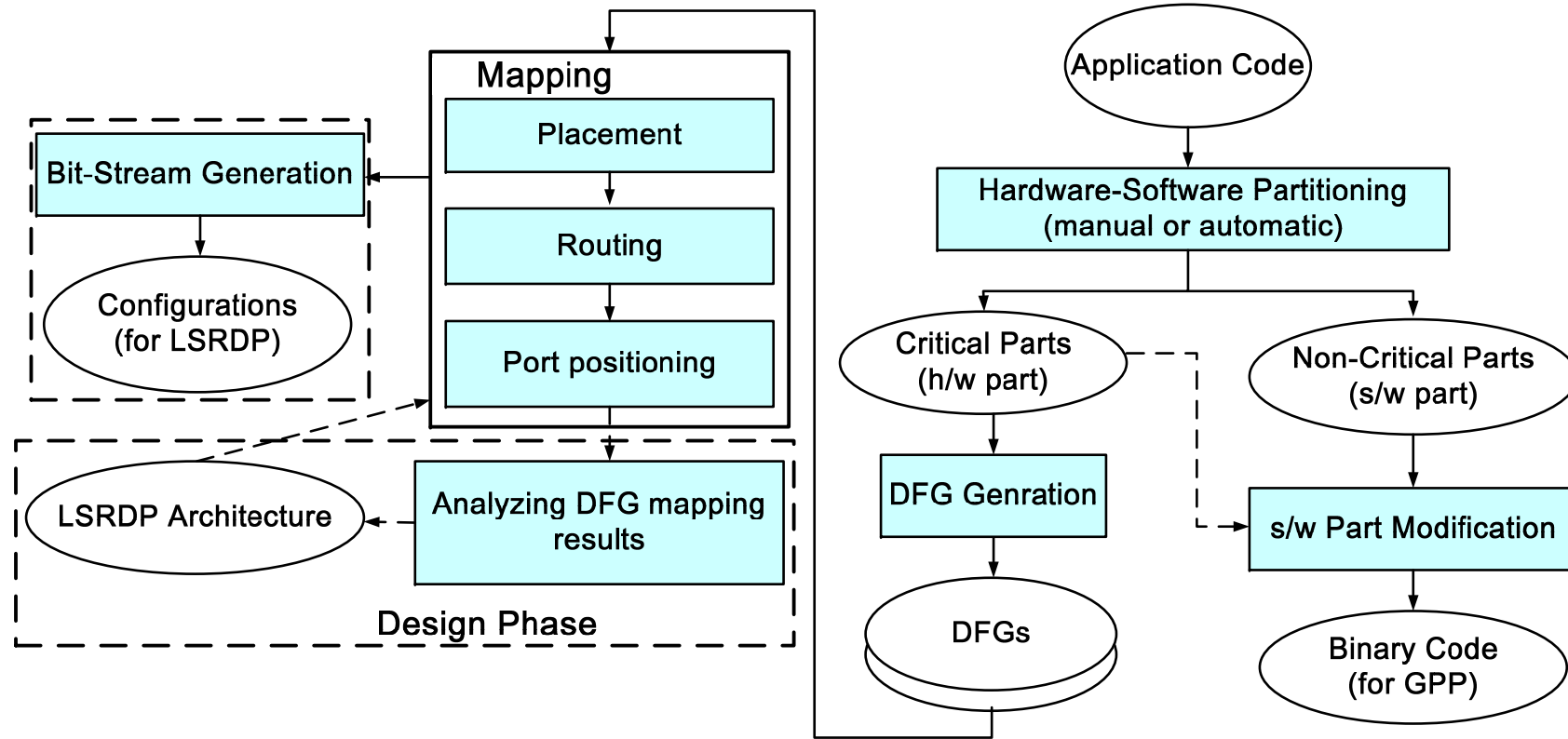
- Immediate registers (64bit) in each PE
- Selector bits for muxes selecting the input data of FUs
- Cross-bar switches in ORNs

# Design Procedure and Tool Chain

---



# Compiler and Design Flow



- DFGs are manually generated from critical parts of applications
- DFG mapping results are used for
  - Analyzing LSRDP architecture statistics
  - Generating LSRDP configuration bit-streams



# Benchmark Applications for Design Procedures



- Finite differential method calculation of 2<sup>nd</sup> order partial differential equations
  - 1dim-Heat equation (Heat)
  - 1dim-Vibration equation (Vibration)
  - 2dim-Poisson equation (Poisson)
- Quantum chemistry application
  - Recursive parts of Electron Repulsion Integral calculation (ERI-Rec)

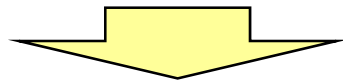
**Only ADD/SUB and MUL operations are used  
in the critical calculations of all above applications**

# DFG Extraction- Heat Equation



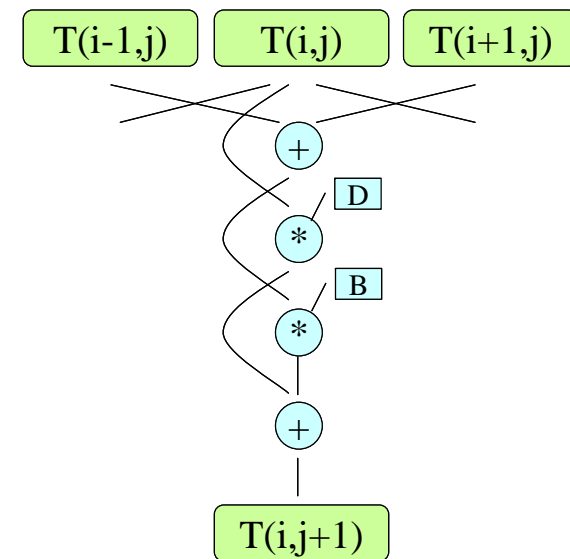
- 1-dim. heat equation for  $T(x,t)$

$$\frac{\partial T(x,t)}{\partial t} = A \frac{\partial^2 T(x,t)}{\partial x^2} \quad (A \text{ is const.})$$



- Calculation by Finite Difference Method (FDM)

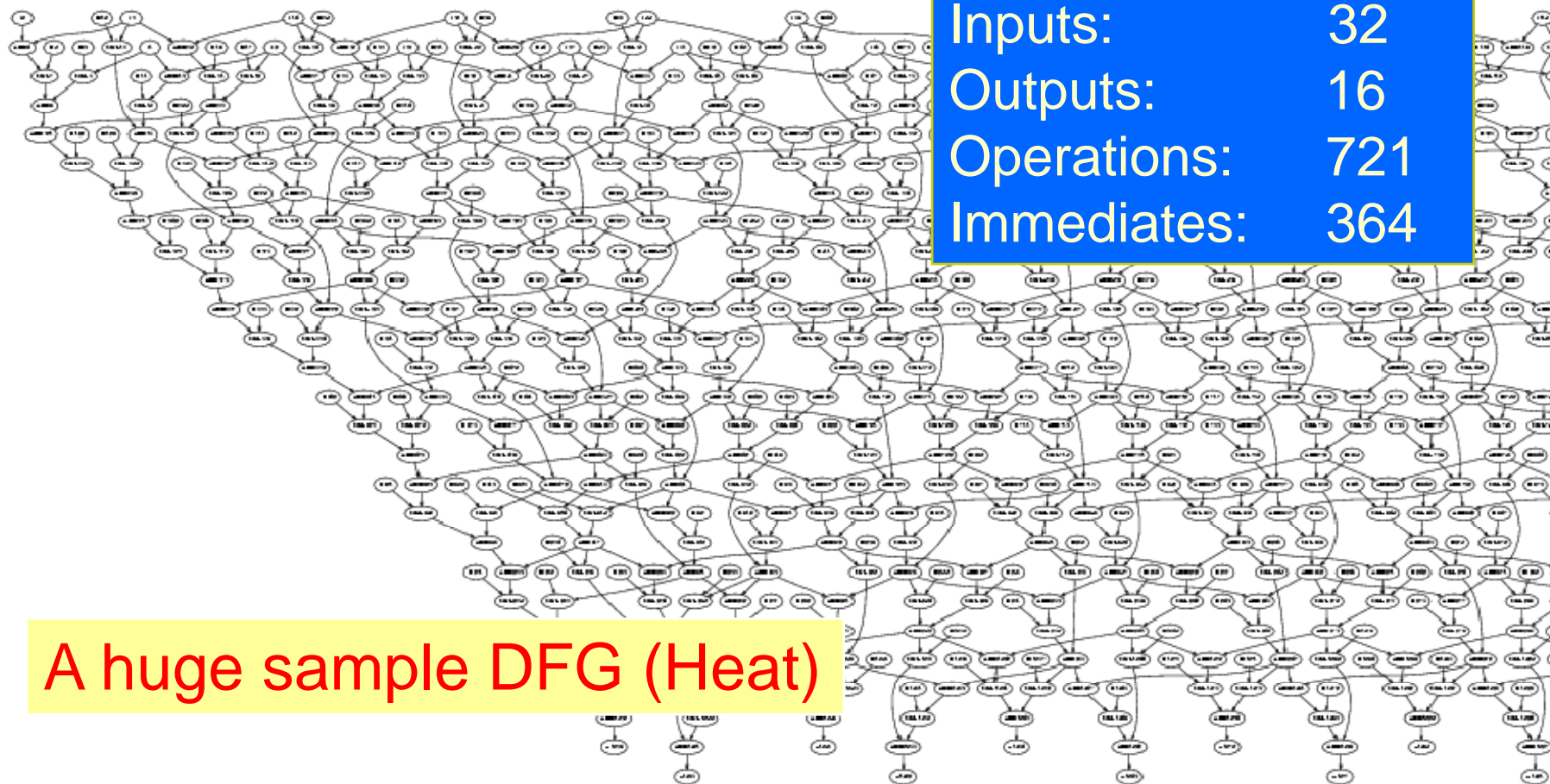
$$\begin{aligned} &T(x_i, t_{j+1}) \\ &= D * T(x_i, t_j) + B * [T(x_{i-1}, t_j) + T(x_{i+1}, t_j)] \end{aligned}$$



Basic DFG can be extended to horizontal and vertical directions to make a larger DFG

Basic DFG corresponding to minimum FDM calculation

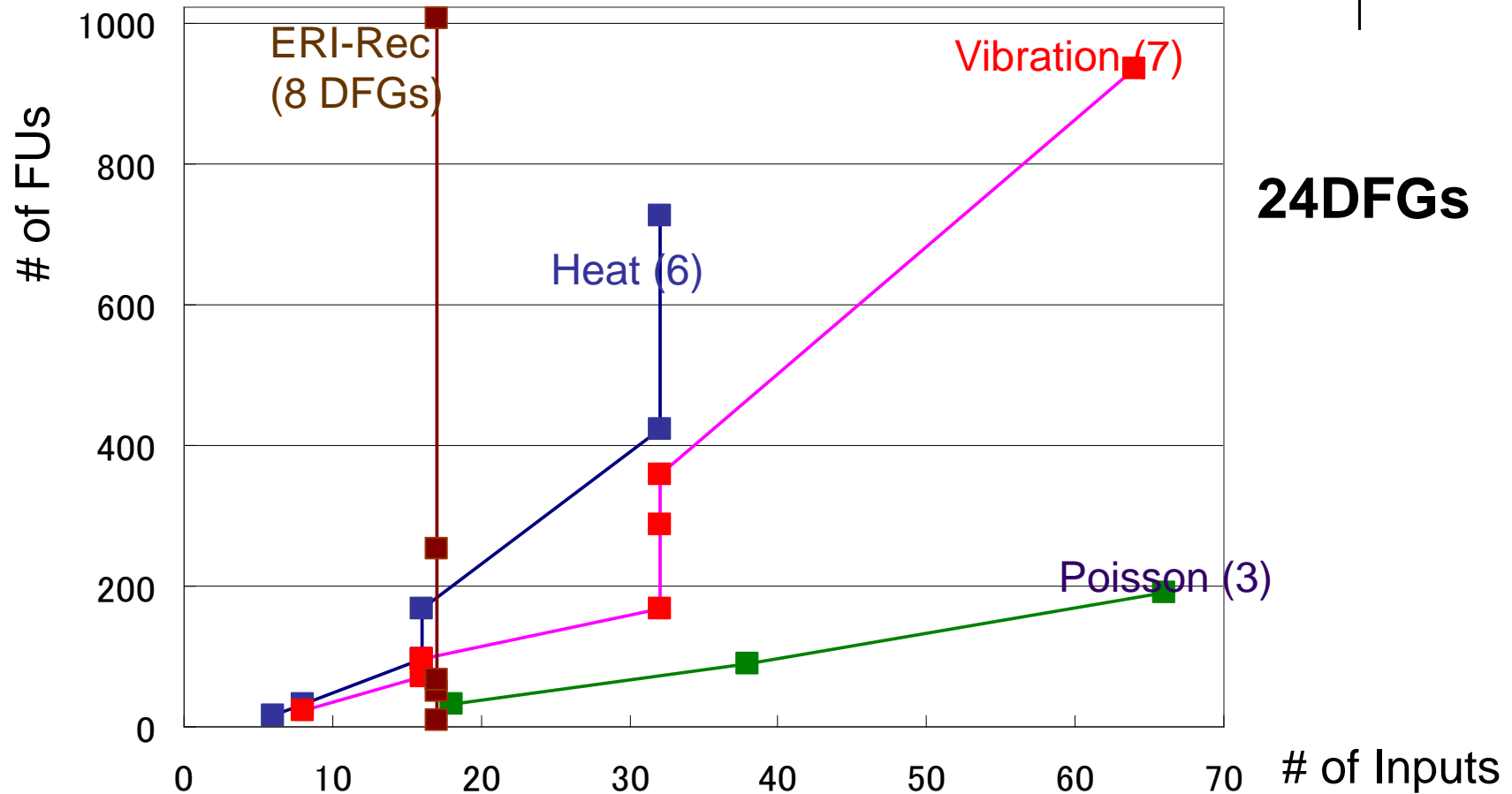
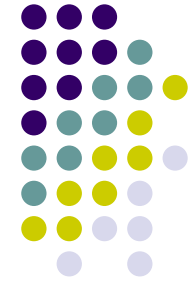
# Example of extracted DFGs- Heat



Inputs: 32  
Outputs: 16  
Operations: 721  
Immediates: 364

A huge sample DFG (Heat)

# DFG Distribution for each application



24DFGs

DFGs have different qualities in terms of the # of FUs, # of Inputs and Outputs

# DFG Classification

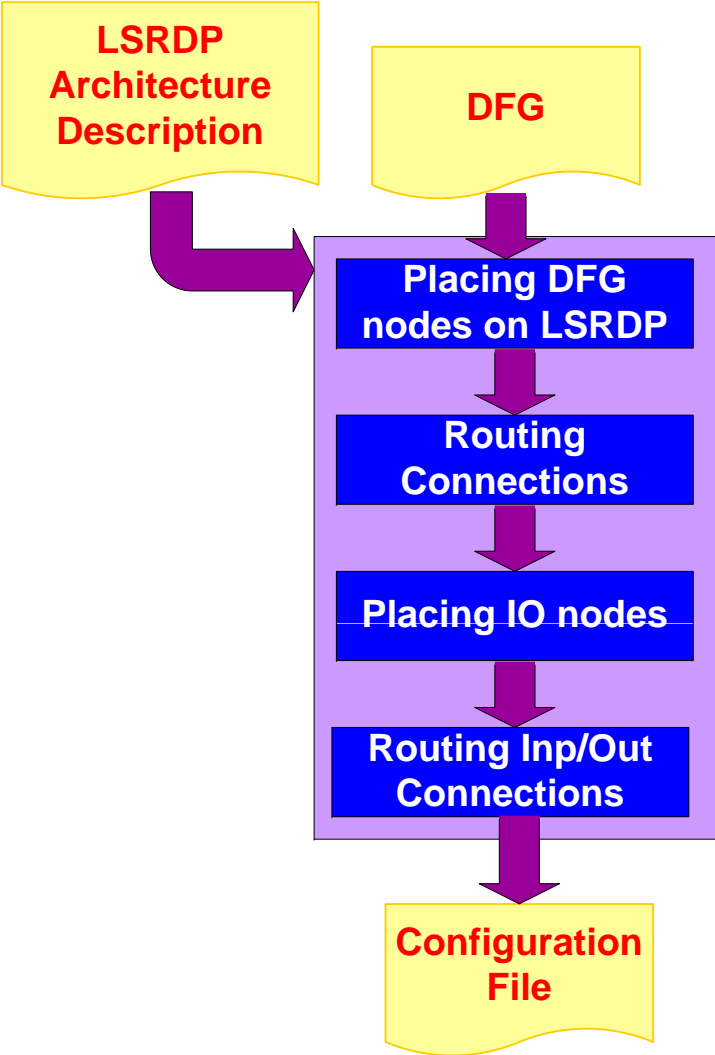


Class	# of FUs	# of Inputs	# of Outputs	# of DFGs
<b>RDP-S</b>	128	19	12	Heat (3) Poi (1) Vib (2) Eri (4)
<b>RDP-M</b>	512	19	12	Heat (1) Poi (1) Vib (1) Eri (4)
<b>RDP-L</b>	1024	38	24	Heat (2) Poi (1) Vib (2) Eri (5)
<b>RDP-XL</b>	> 1024	64	52	Heat (1) Poi (1) Vib (2) Eri (5)

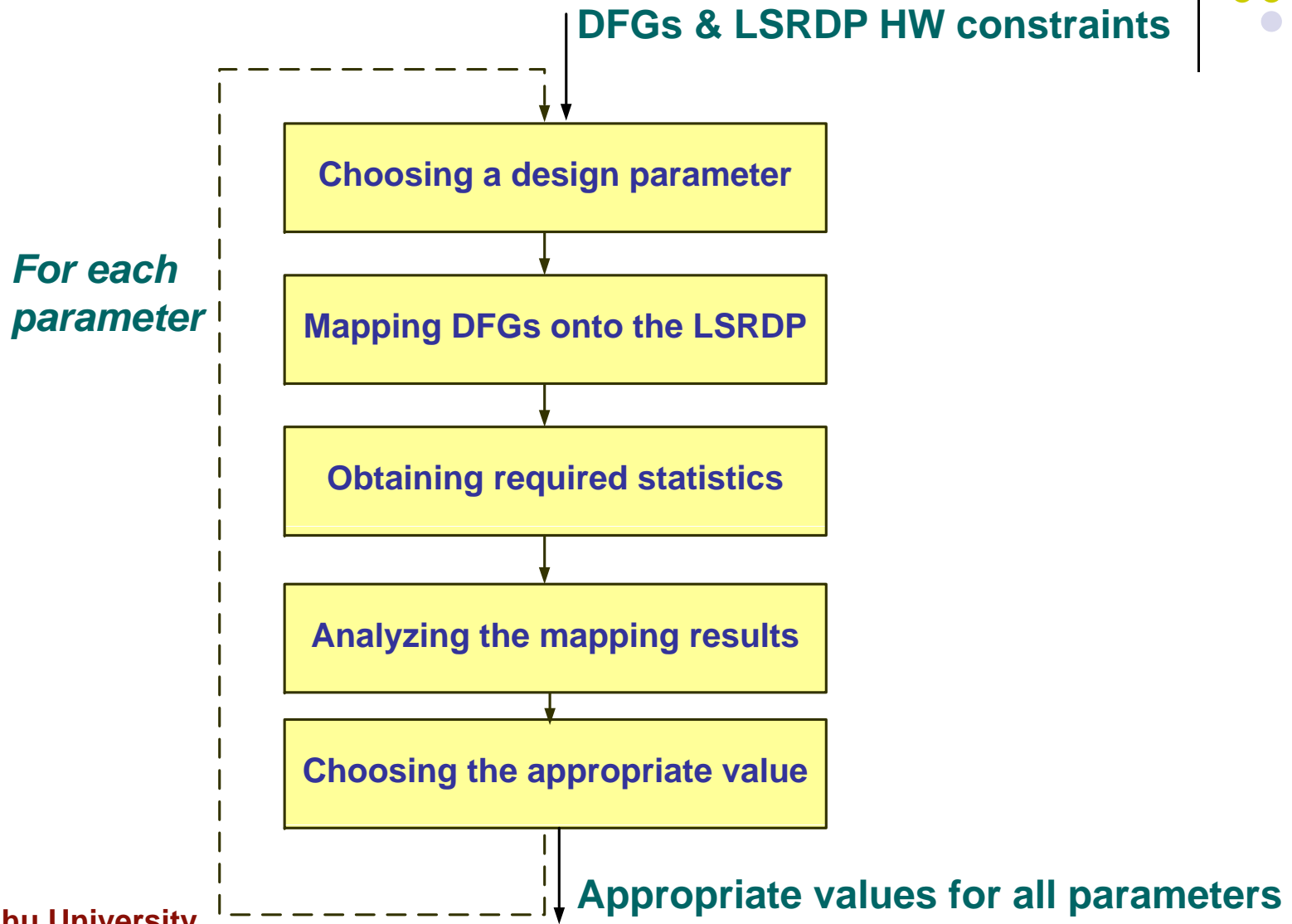
Totally,  
24 DFGs are prepared  
for benchmark Apps.

Due to broad range of DFG sizes  
DFGs are classified as S, M, L, XL with respect to their size  
and the number of Input/Output nodes  
=> LSRDP designing processes for S, M, L, XL, respectively

# Mapping DFGs onto LSRDP

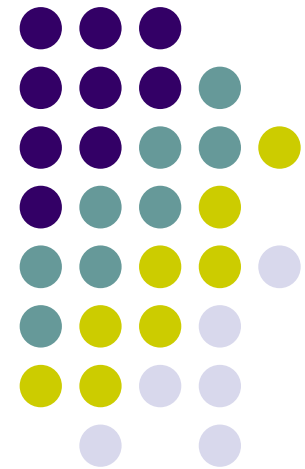


# LSRDP Design Procedure



# Preliminary Results

---



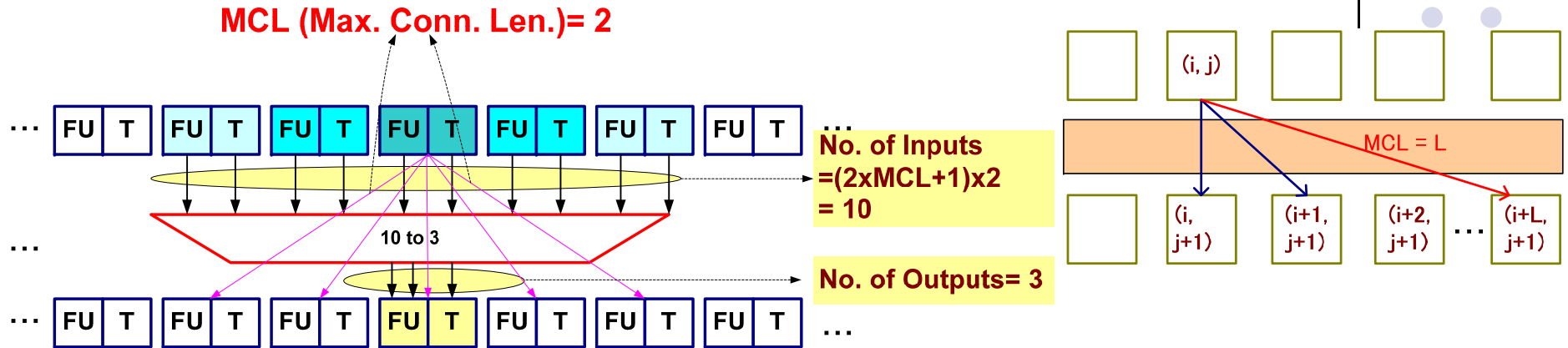
# LSRDP Specifications: Width & Height



	# of Input ports	# of Output ports	Width	Height
<b>LSRDP-S</b>	<b>19</b>	<b>12</b>	<b>16</b>	<b>16</b>
<b>LSRDP-M</b>	<b>19</b>	<b>12</b>	<b>32</b>	<b>16</b>
<b>LSRDP-L</b>	<b>38</b>	<b>24</b>	<b>64</b>	<b>32</b>

**LSRDP Dimensions and the number of input/output ports**

# LSRDP Specifications: MCL



LSRDP	MCL (avg/max)	ORN Size- No of Inps (avg/max), Outs
LSRDP-S	4/8	18/34, 3
LSRDP-M	5/9	22/38, 3
LSRDP-L	5/9	22/34, 3



# Analyzing Various LSRDP Layouts



	Layout	Size
Heat	I	8x3
	II	8x3
	III	8x4
Viriation	I	10x8
	II	10x8
	III	10x11
Poisson	I	10x10
	II	10x12
	III	15x18
ERI1	I	6x2
	II	9x3
	III	6x2
ERI2	I	10x10
	II	10x10
	III	15x8

**Layout I  $\simeq$  Layout II**

(Except ERI1 DFG which gives better size for Layout III)

**Layout II can be used instead of Layout I to obtain a smaller LSRDP**



# LSRDP at One Glance (1/2)

<b>Functional units</b>		ADD/SUB, MUL		
<b>Layout</b>		Type II (checker pattern)		
<b>Operations</b>		64-bit floating point		
<b>Processing structure</b>		Pipelined		
<b>PE structure</b>		FU, T, FU+T, T+T		
<b>LSRDP Size</b>		<b>Small</b>	<b>Medium</b>	<b>Large</b>
<b>No. of inp/out ports</b>		19/12	19/12	38/24
<b>Width/Height</b>		16/16	32/16	64/32
<b>Conf. bit-stream size</b>	<b>Imm. Regs</b>	16*16*64	32*16*64	64*32*64
	<b>ORNs</b>	16*BSS(ORN)	32* BSS(ORN)	64*BSS(ORN)
	<b>PEs</b>	16*16* 2	32*16*2	64*32* 2
<b>ORN</b>	<b>inputs, outputs</b>	22 , 3	26 , 3	26 , 3
	<b>Structure</b>	Cross-bar switch		
	<b>Conn. Type</b>	One-directional		

# LSRDP at One Glance (2/2)



<b>Internal memory</b>	<b>Type</b>	<b>Immediate registers</b>
	<b>Size and count</b>	<b>64-bit registers, One reg. for each PE</b>
	<b>Communication mechanism</b>	<b>Serial</b>
<b>External memory</b>	<b>No. of memory modules</b>	<b>16</b>
	<b>Date trans. rate</b>	<b>1800Mbps/pin</b>
	<b>Overall data trans. rate</b>	<b>24 GB/s</b>
	<b>Mem. to LSRDP bus width</b>	<b>64 bit</b>
	<b>Channels per module</b>	<b>Two</b>
<b>Reconf. mechanism</b>	<b>Bit serial configuration through a serial chain</b>	

# Preliminary Performance Evaluation



## Base processor configuration

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

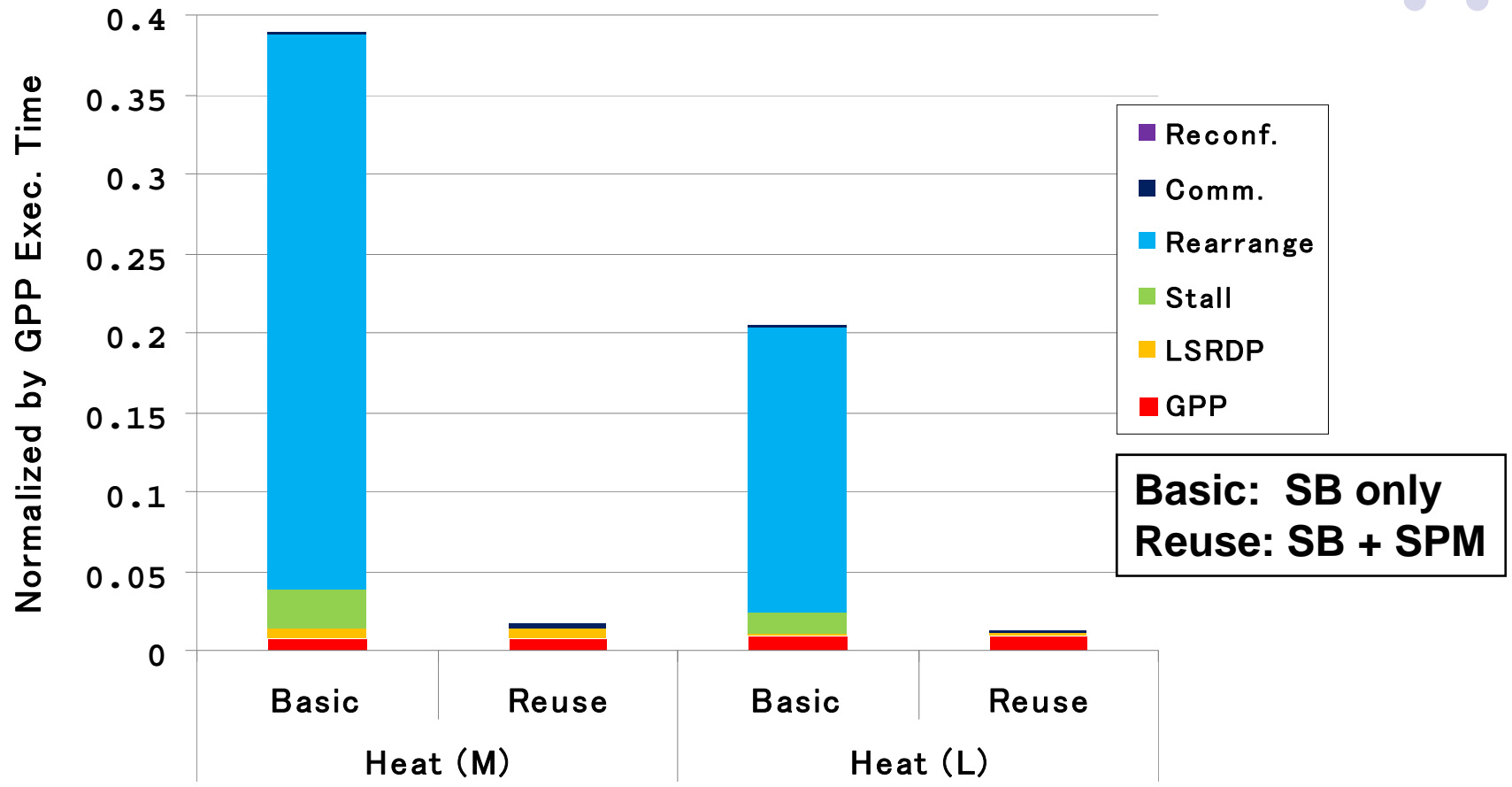
## GPP+LSRDP configuration

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

**GPP:** Exec. time measurement by means of a processor simulator  
**LSRDP:** Estimation by performance modeling

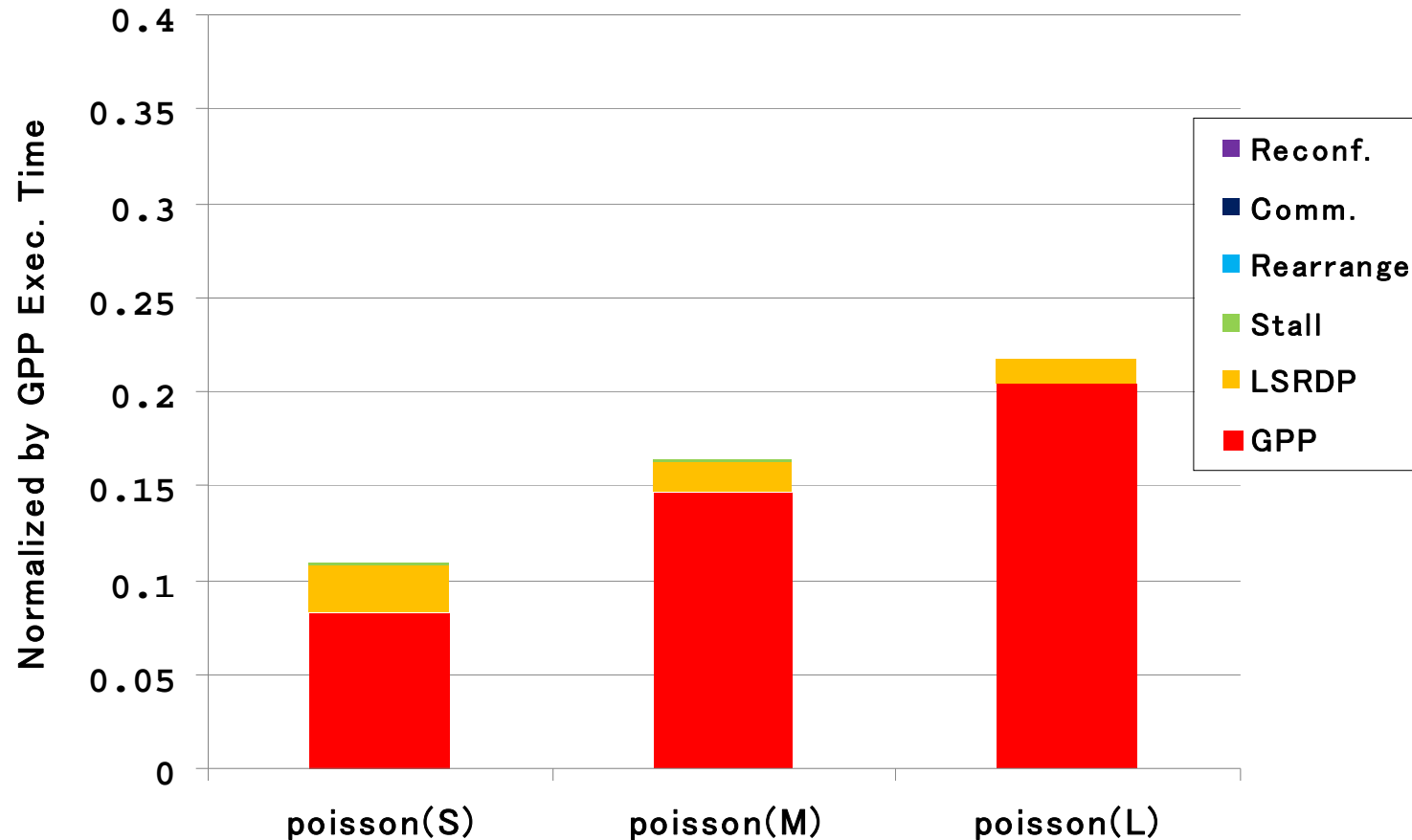
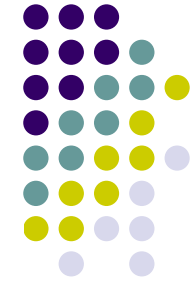


# Preliminary Performance Evaluation (Heat)



**Data reusing is employed to avoid the need for data rearrangement as well as frequently data retrieval from the scratchpad memory.**

# Preliminary Performance Evaluation (Poisson)



A small fraction is related to processing time on LSRDP and the main fraction concerns to various overhead times as well as the execution time on GPP

# Conclusions & Future Work



- A high-performance computer comprising an accelerator (LSRDP) implemented by superconducting circuits was introduced.
- 24 benchmark Data Flow Graphs (DFGs) were manually generated.
- LSRDP micro-architecture is designed based on characteristics of scientific applications via a quantitative approach.
- LSRDP is promising for resolving issues originated from CMOS technology as well as achieving considerable performances.

## Future Work:

- To achieve higher performance it is required to *reduce various overhead costs mainly related to data management part.*
- To reduce the implementation cost of LSRDP, we will focus on *reducing maximum connection length and ORN size.*

# Acknowledgement

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

