Reconfigurable Architectures FPGA as an accelerator

AMANO, Hideharu

hunga@am. ics. keio. ac. jp

PLD（Programmable Logic Device）

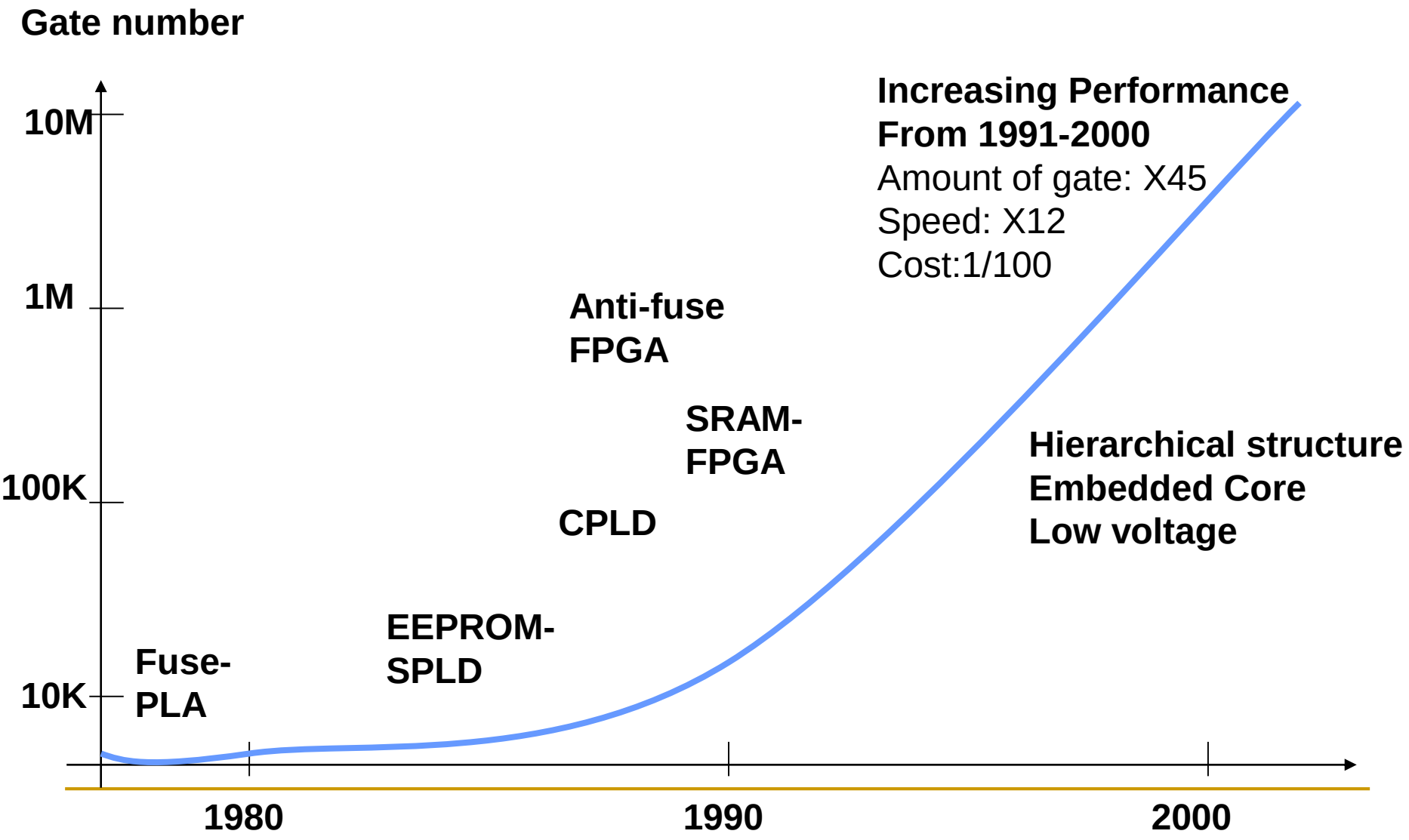
- Integrated Circuit whose logic function can be defined by users.

⟺ Standard IC, ASIC(Application Specific IC)

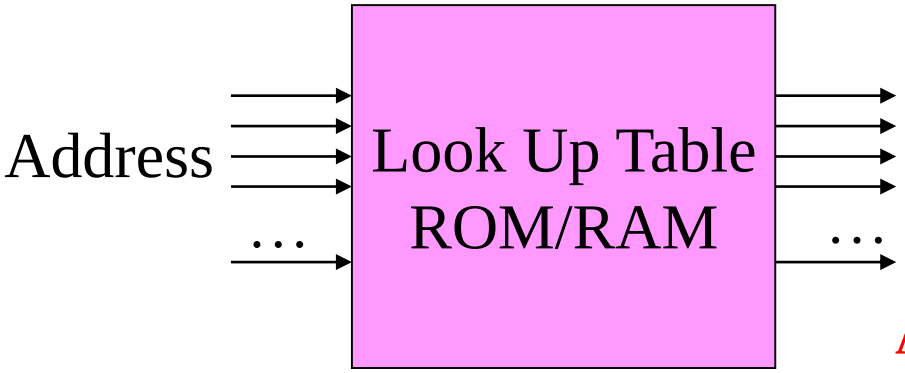
- SPLD（Simple PLD) / PLA（Programmable Logic Array)
  - Small scale IC with AND-OR array
- CPLD(Complex PLD)
  - Middle scale IC with AND-OR array
- FPGA（Field Progarmmable Gate Array）
  - Large scale IC with LUT

Caution! Terms are not well defined!

Rapidly development of PLD

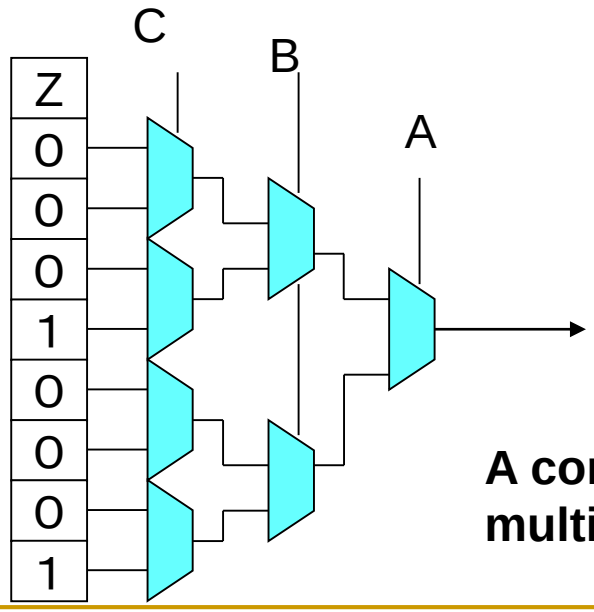


LUT : Look Up Table



A simple ROM/RAM can used as a random logic.

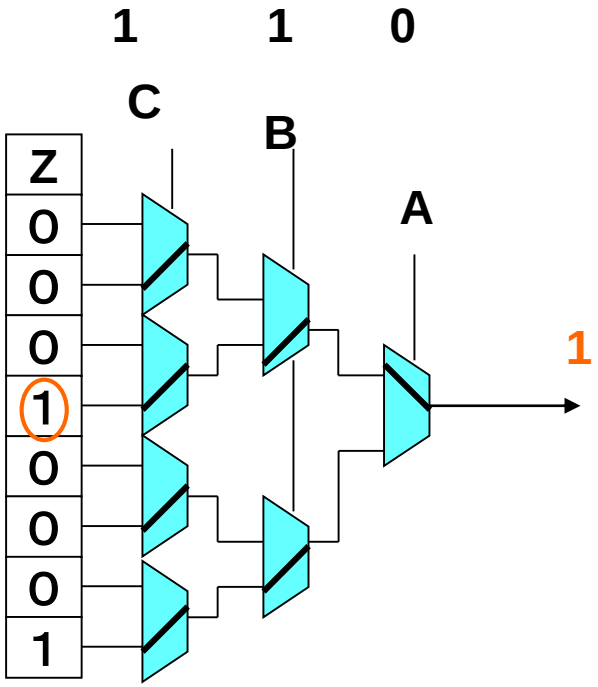
ABC	Z
000	0
001	0
010	0
011	1
100	0
101	0
110	0
111	1



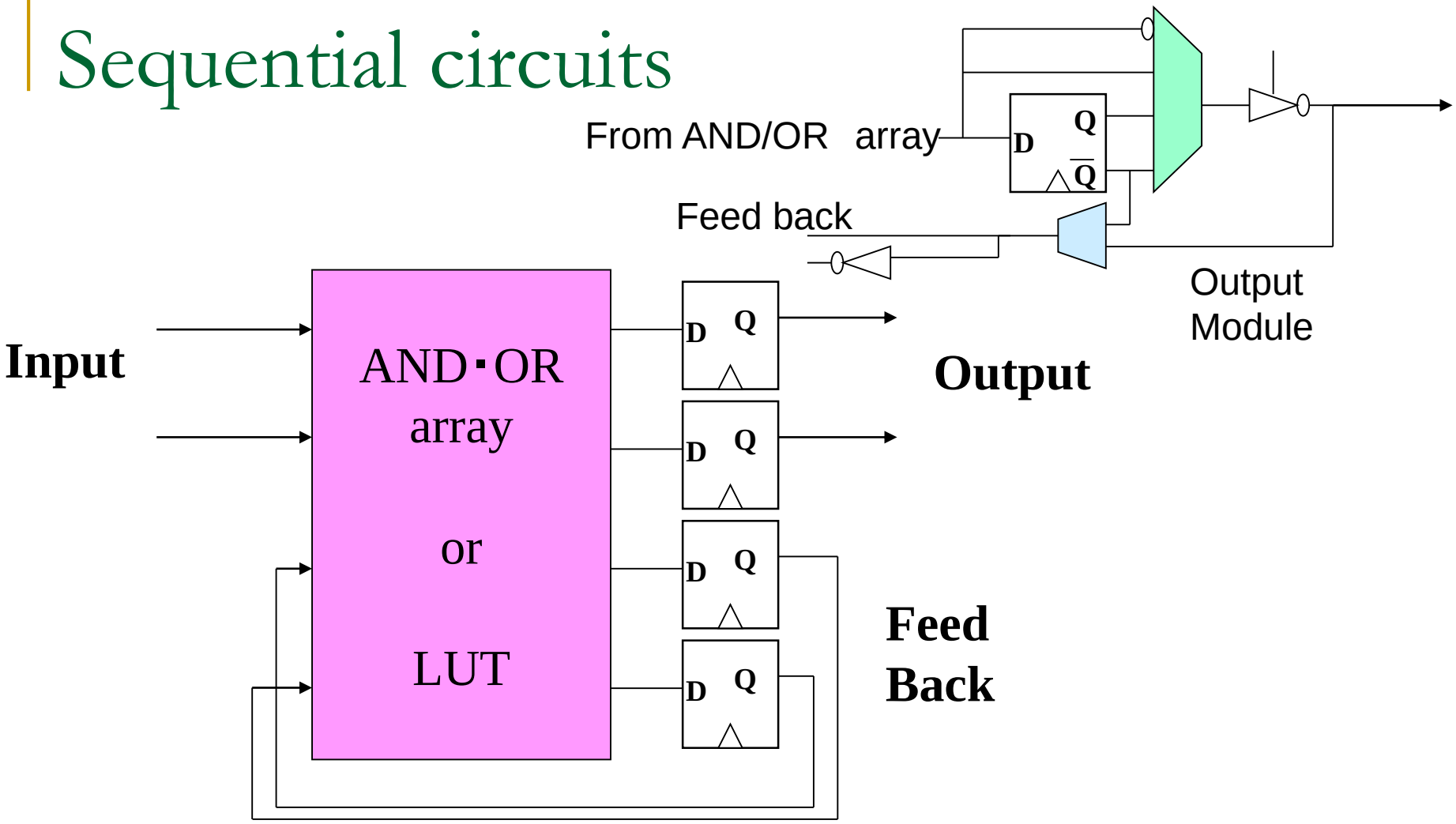
A combination of memory and multiplexers are commonly used.

An example using LUT : Look Up Table

ABC	Z
000	0
001	0
010	0
011	1
100	0
101	0
110	0
111	1

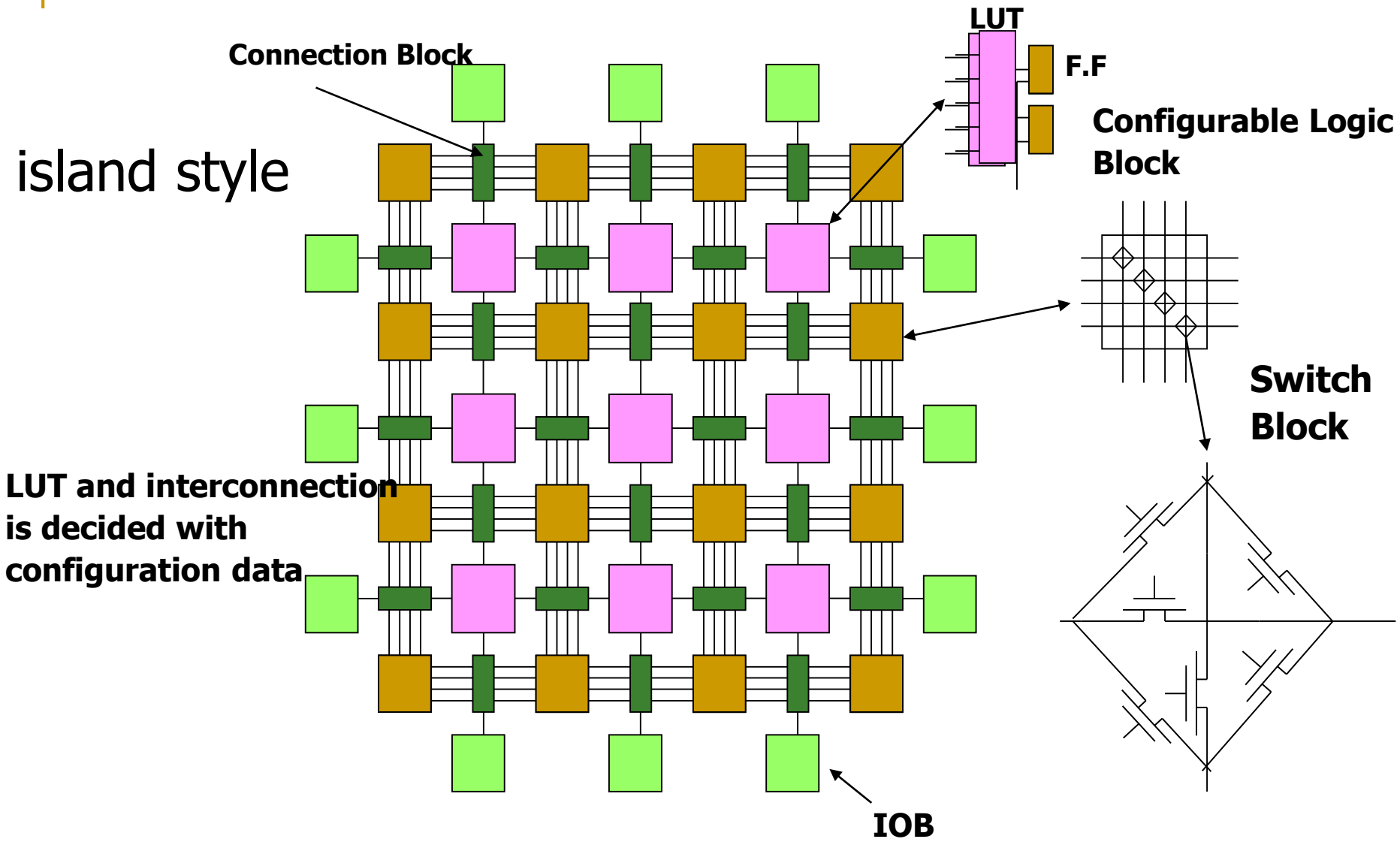


Sequential circuits

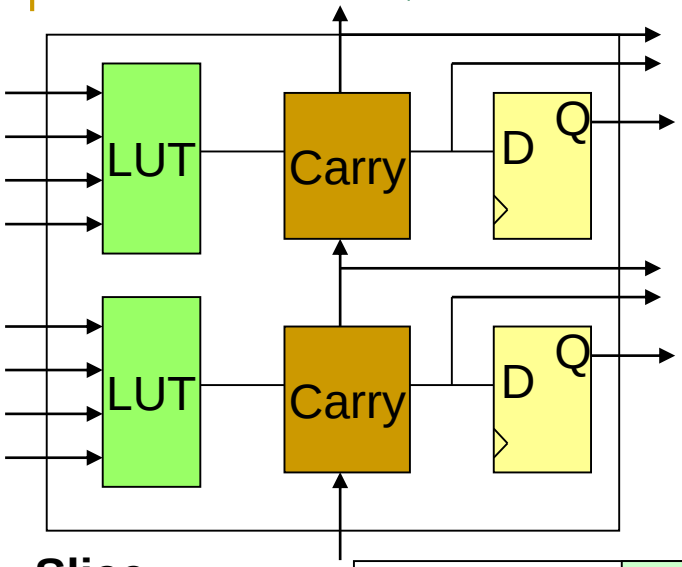


Sequential circuit (state machine) can be built by attaching Flip-flops and feed back loops.

FPGA(Field Programmable Gate Array)



Xilinx Virtex II



Slice

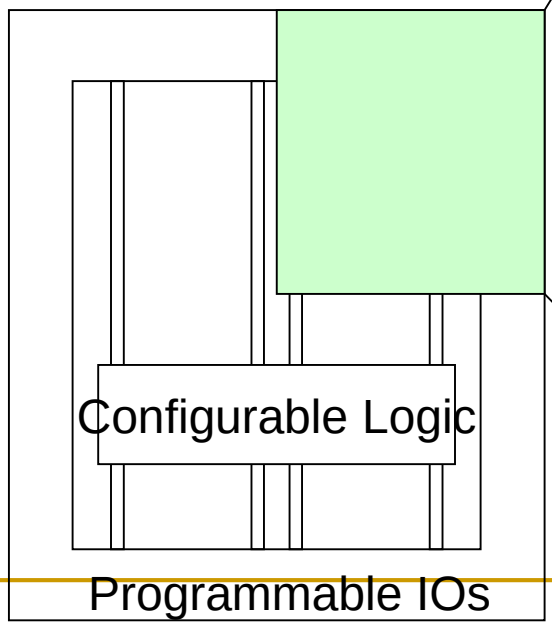
Slice X 2 → CLB (Configurable Logic Block)

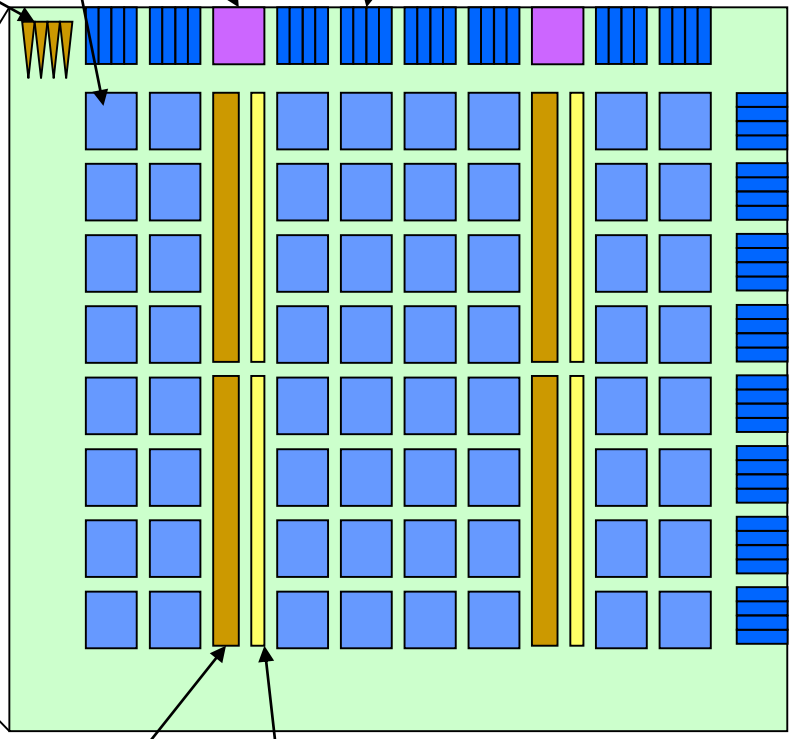
Global Clock MUX

DCM

IOB

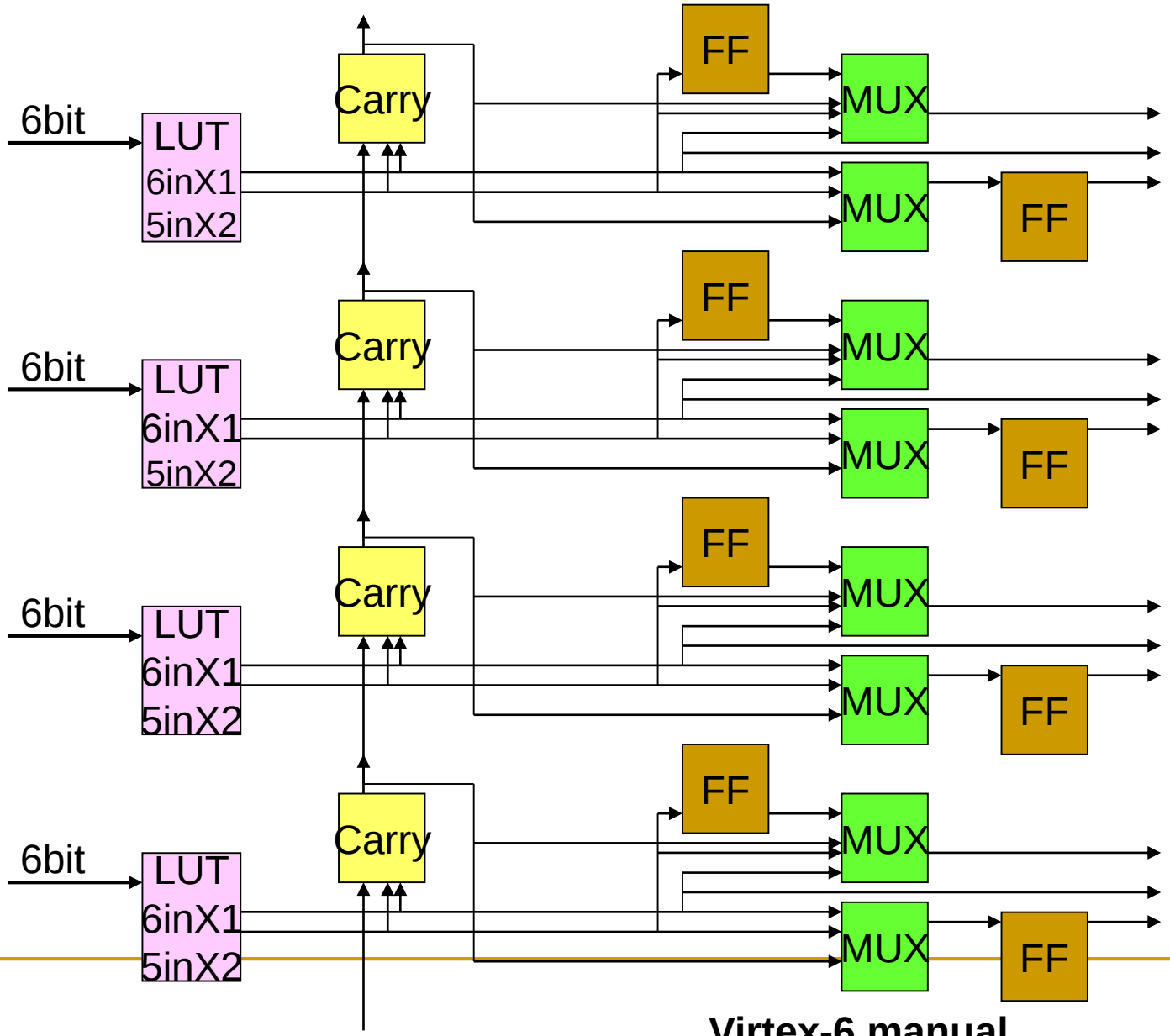
100000 CLBs
3Mbit





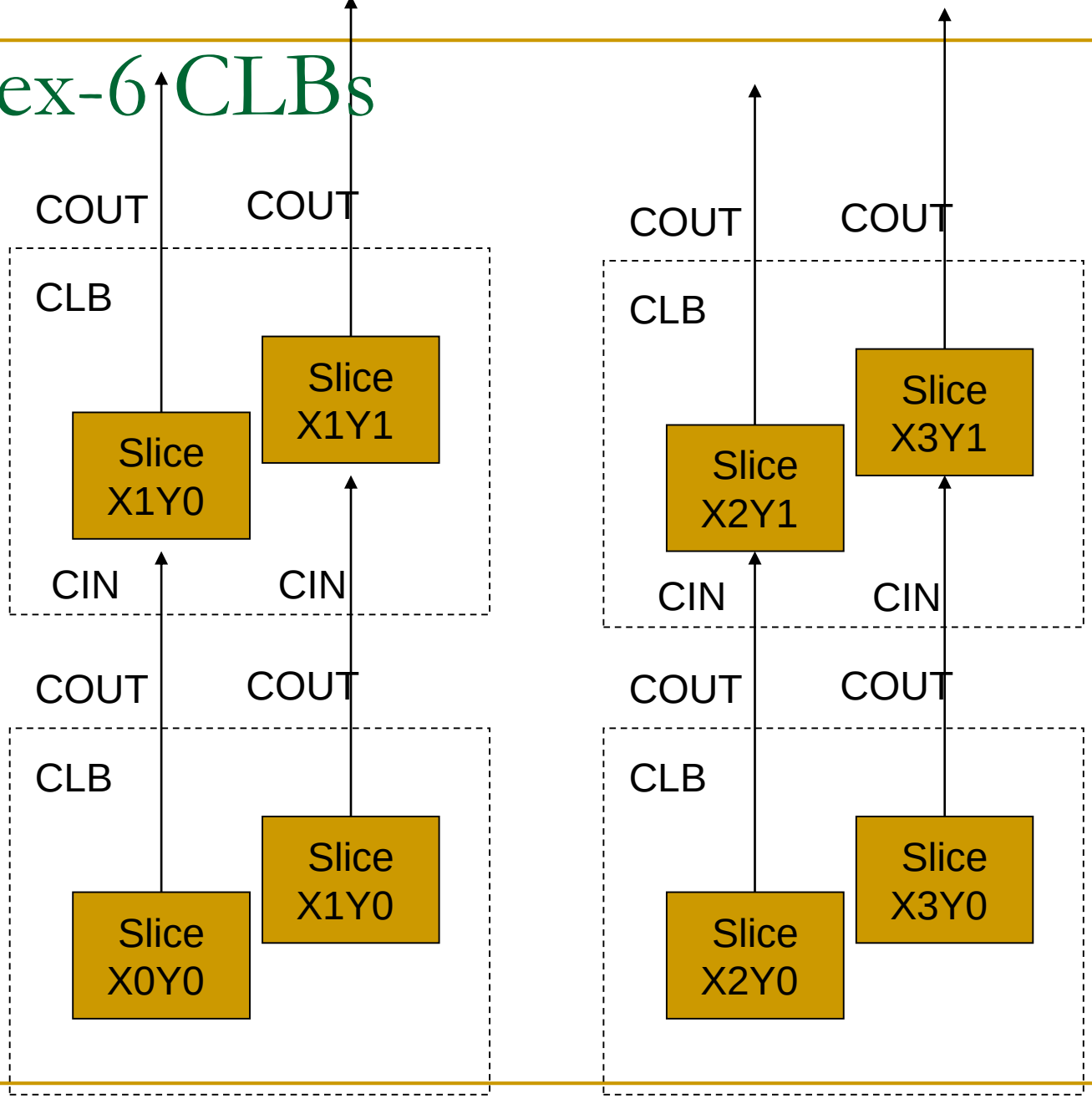
RAM **Multiplier**

Slice structure of Virtex-6



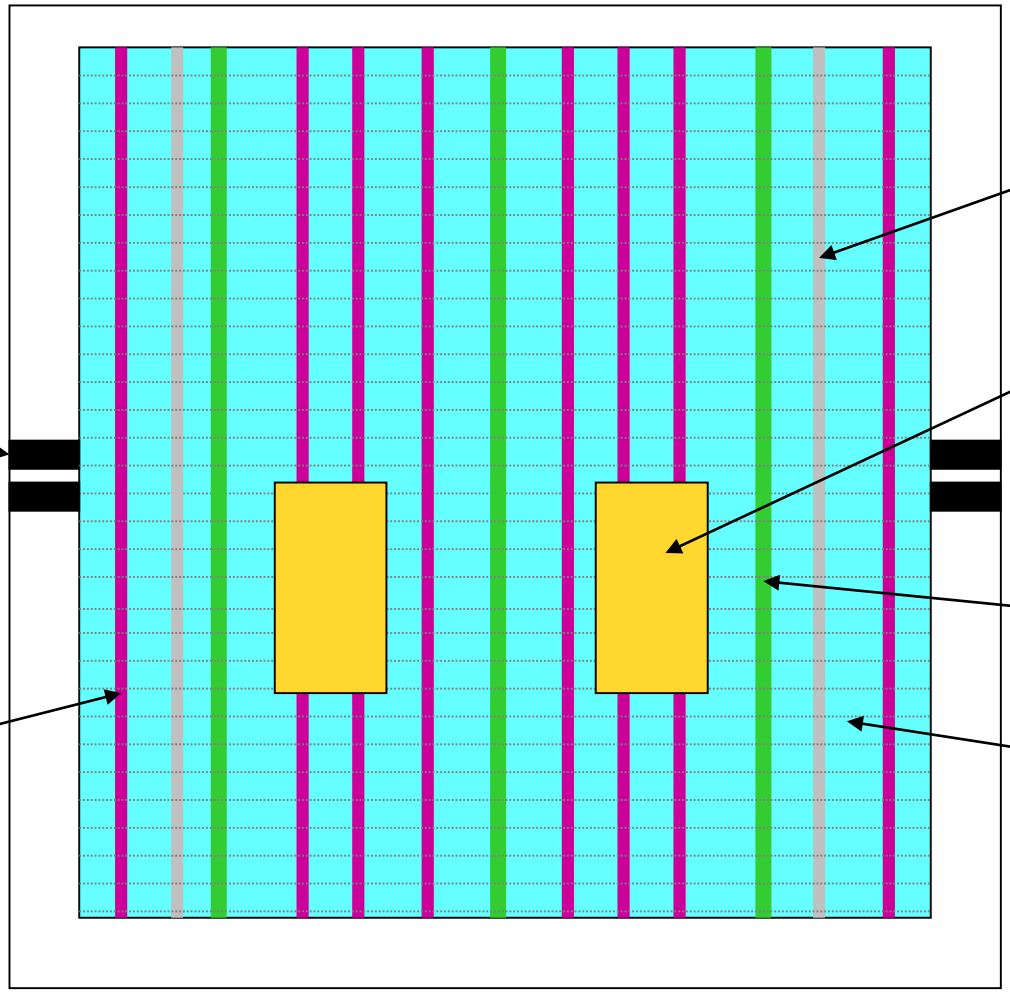
Virtex-6 manual

Virtex-6 CLBs



Virtex-6 manual

Altera Stratix II



DSP Blocks

Mega RAM Blocks

PLL

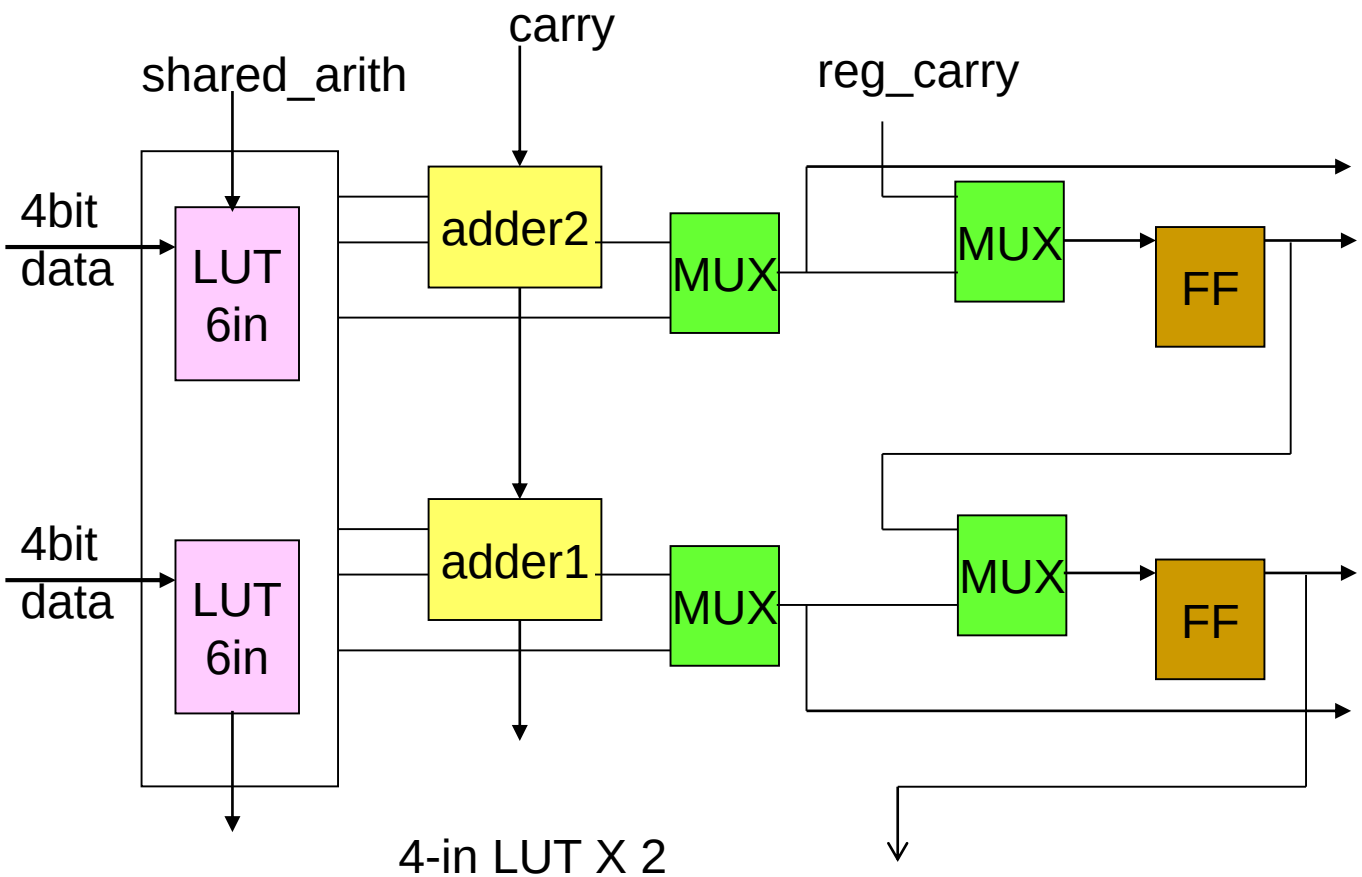
M4K RAM Blocks

M512 RAM Blocks

LAB : Logic Array Block consisting of 10 LE (4-input LUT and F.F.)

Hierarchical Interconnect

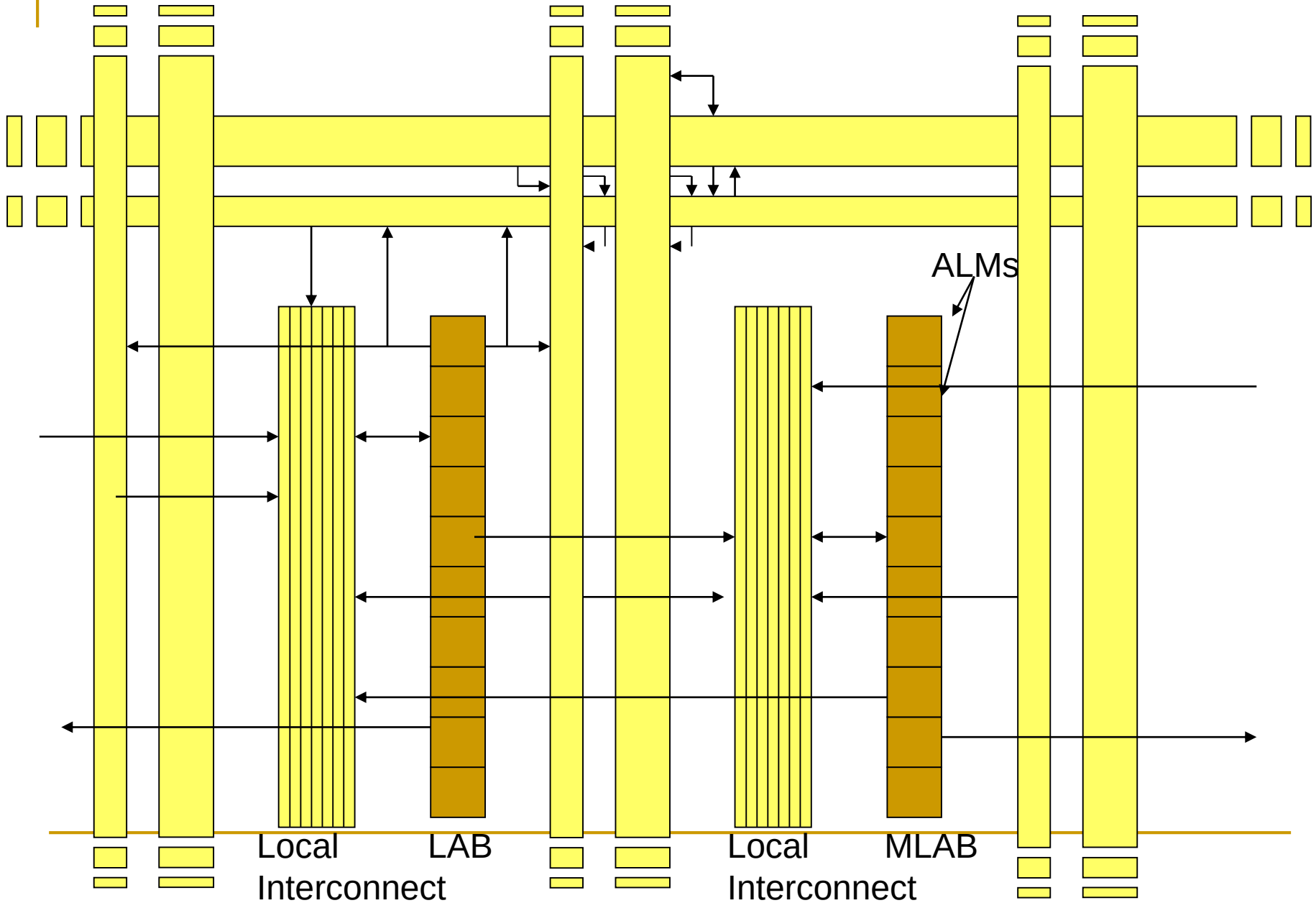
Stratix-IV ALM Structure



4-in LUT X 2
5-in LUT + 3-in LUT
5-in LUT + 4-in LUT 1-input shared
5-in LUT + 5-in LUT 2-input shared
6-in LUT
6-in LUT + 6-in LUT 4-input shared

Stratix-IV manual

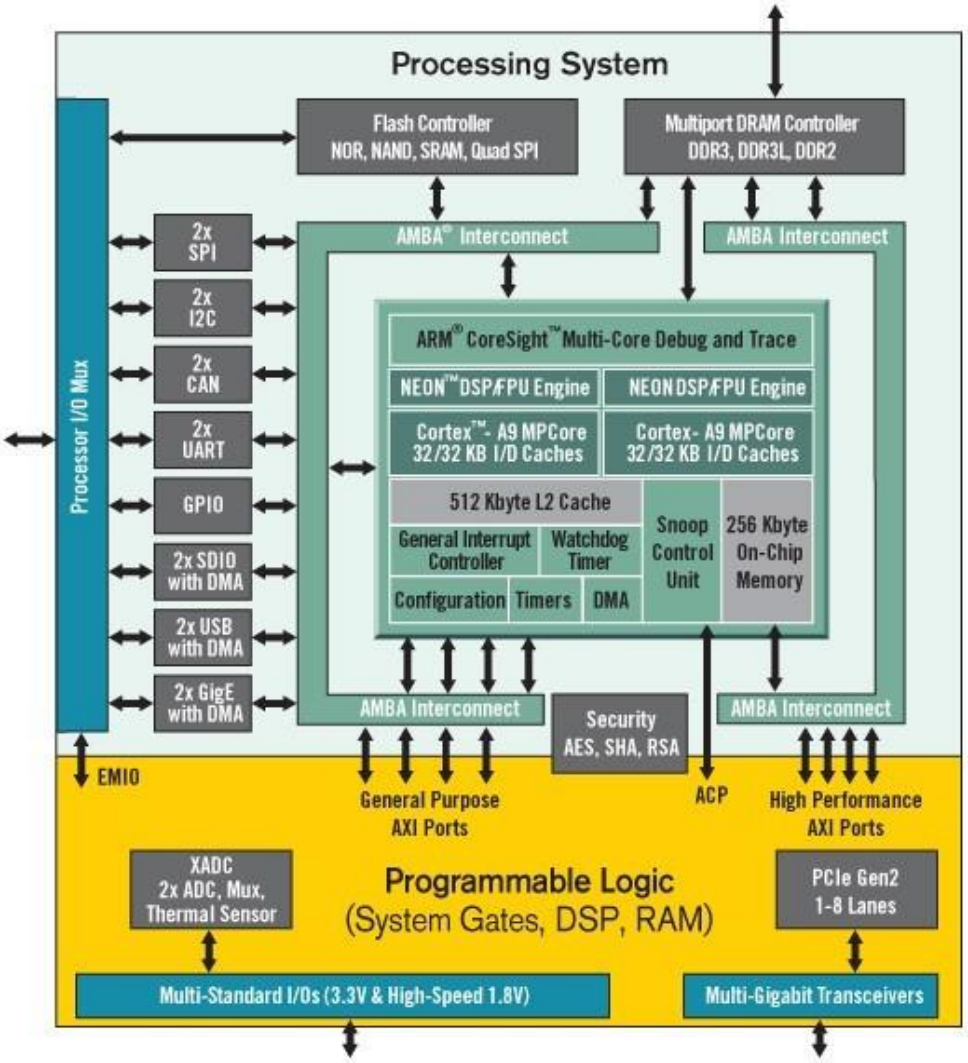
Stratix-IV LAB structure



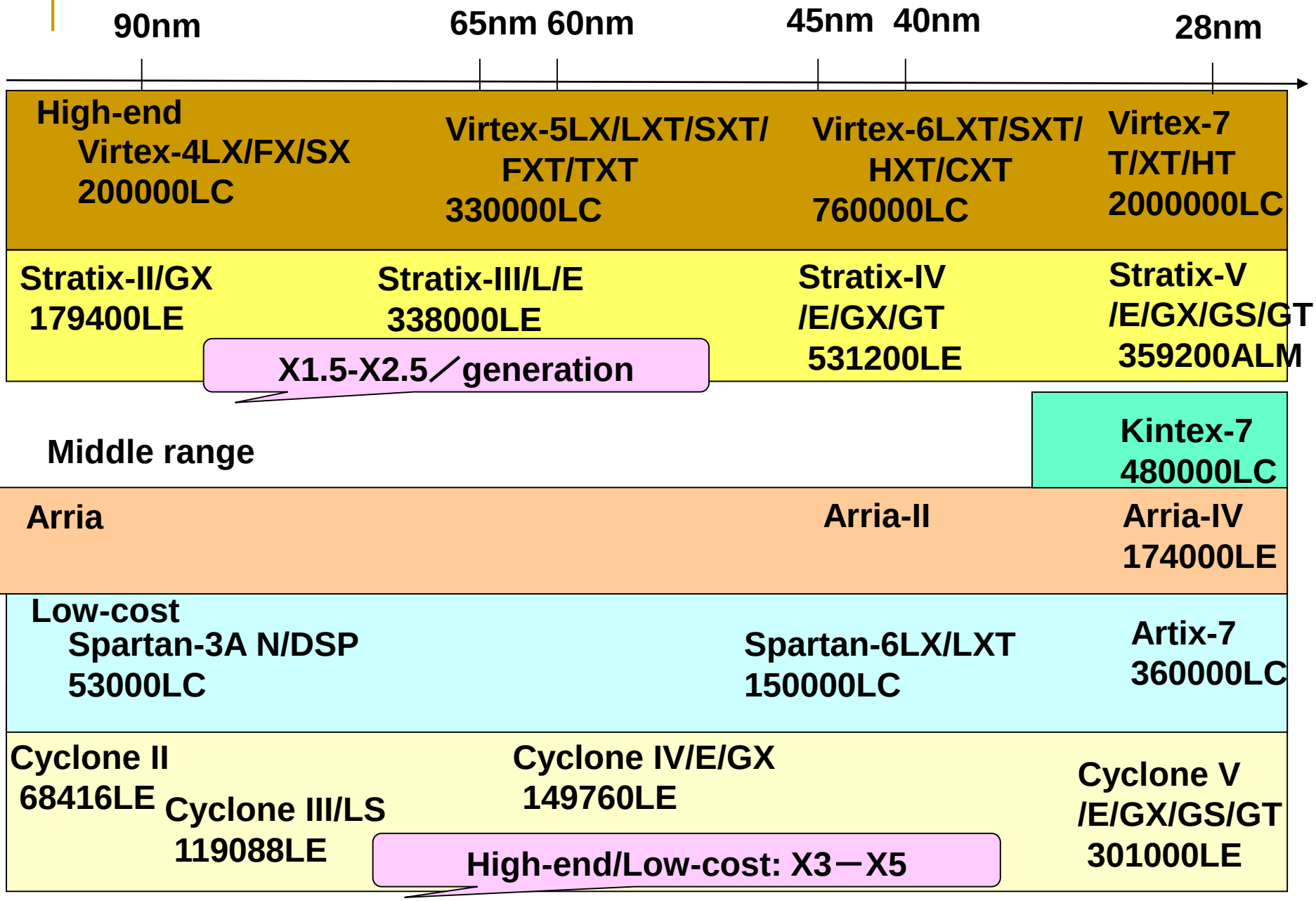
Stratix-IVマニュアルより

Zynq All programmable SoC

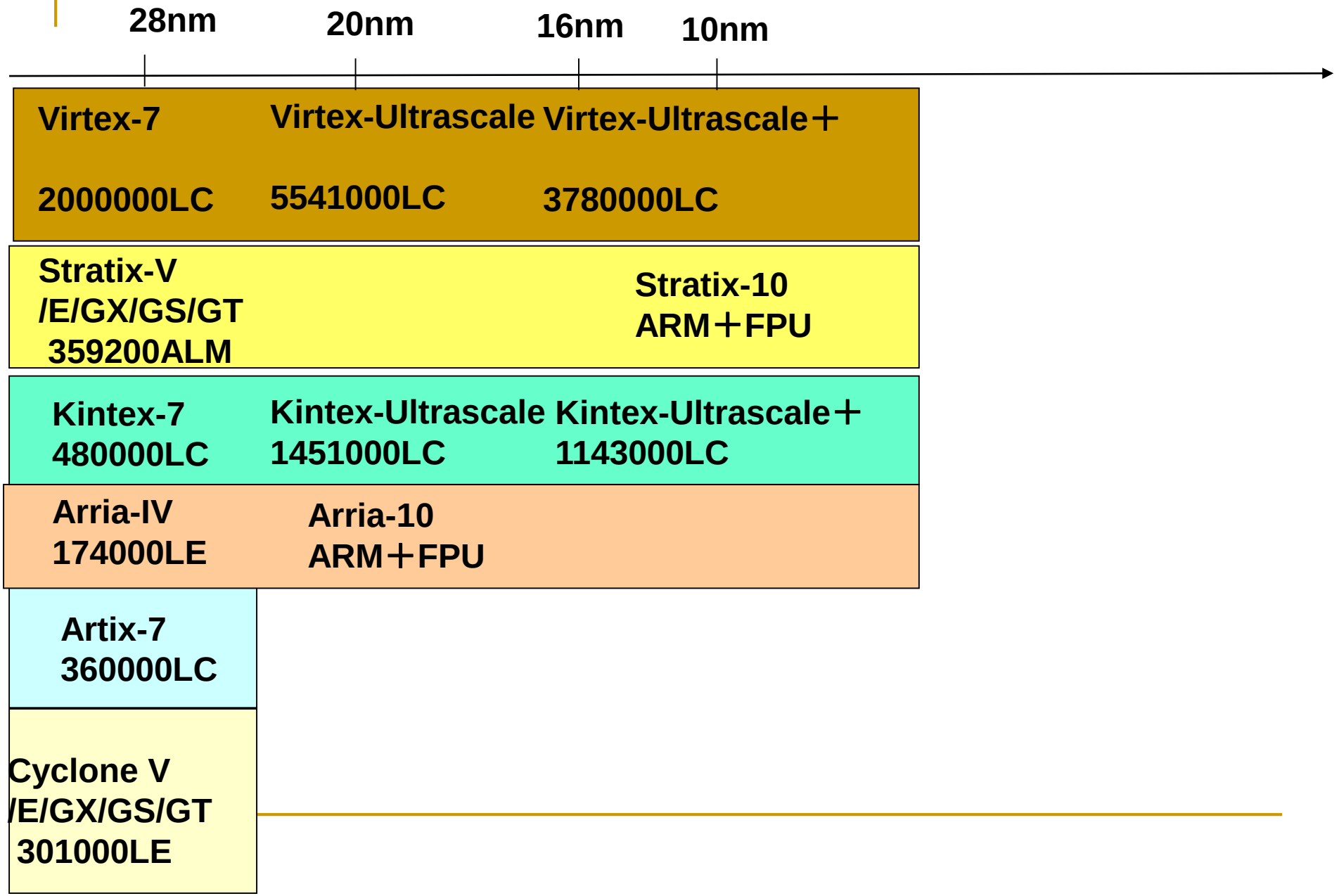
- ARM Cotex-A9（PS part) Dual Core CPU +
- 28nm Artix-7/Kintex-7 based FPGA（PL part)



Technologies vs. Product



Technology vs. Products (Cont.)

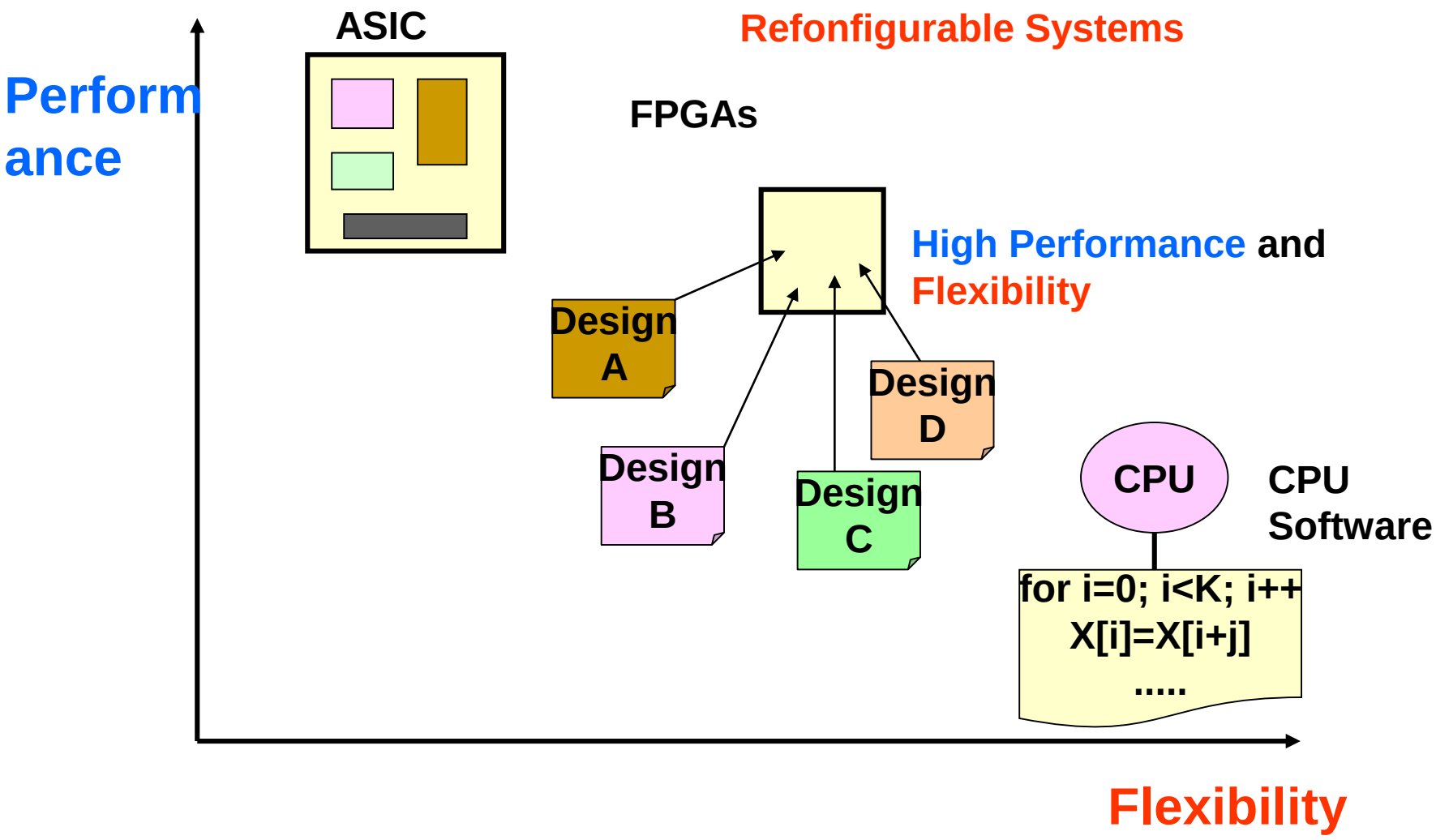


Design of PLDs

- Mostly designed with common HDL（Verilog-HDL, VHDL)
  - C level entry is used recently: Impulse-C, Vibado-HLS, SD-Accel(Xilinx), Open-CL, Intel-HLS(Intel)
- Synthesis, optimization, place and route is automatically done by vendors' tools.
  - Integration and combination of tools from various venders are used recently.
  - For large circuit, a long time is required especially for place and route.
  - Using IPs, clock/DLL adjustment is manually done.
  - Optimization techniques are different from vendors/products.

Reconfigurable System (Custom Computing Machine)

- A target algorithm is executed directly with a hardware on SRAM-style FPGA/PLDs.
  - High performance of special purpose machines.
  - High degree of flexibility of general purpose machines.
- A completely different execution mechanism from a stored program computers.


Refonfigurable Systems
Performance
ASIC
FPGAs
High Performance and Flexibility
Design A
Design B
Design C
Design D
CPU
CPU Software
for i=0; i<K; i++
X[i]=X[i+j]
.....
Flexibility

How enhance the performance ?

- Performance enhancement by hardware execution itself
  - The overhead of software execution (Instruction fetch, data load to registers, and etc.)
  - The overhead of using fixed size data.
  - The overhead of using only two way branches.

However, these benefits are not so large, for embedded CPU and DSP are highly optimized.

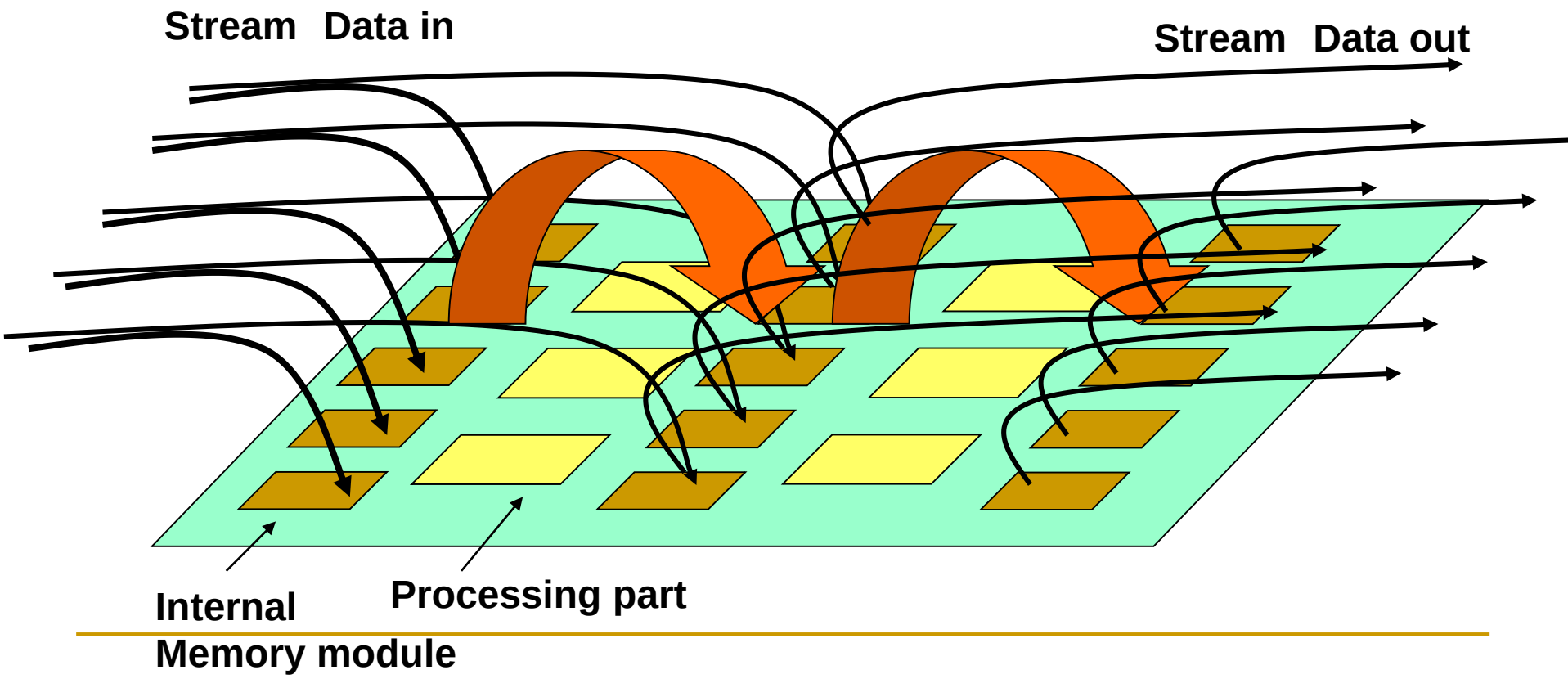
The key of performance improvement is parallel processing

Parallel processing in reconfigurable systems

- Various techniques can be used
  - SIMD execution
  - Pipelined structure
  - Systolic algorithm
  - Data driven control
- Parallel execution other than calculation
  - Parallel data access using internal memory units
  - Parallel data transfer including I/O accesses

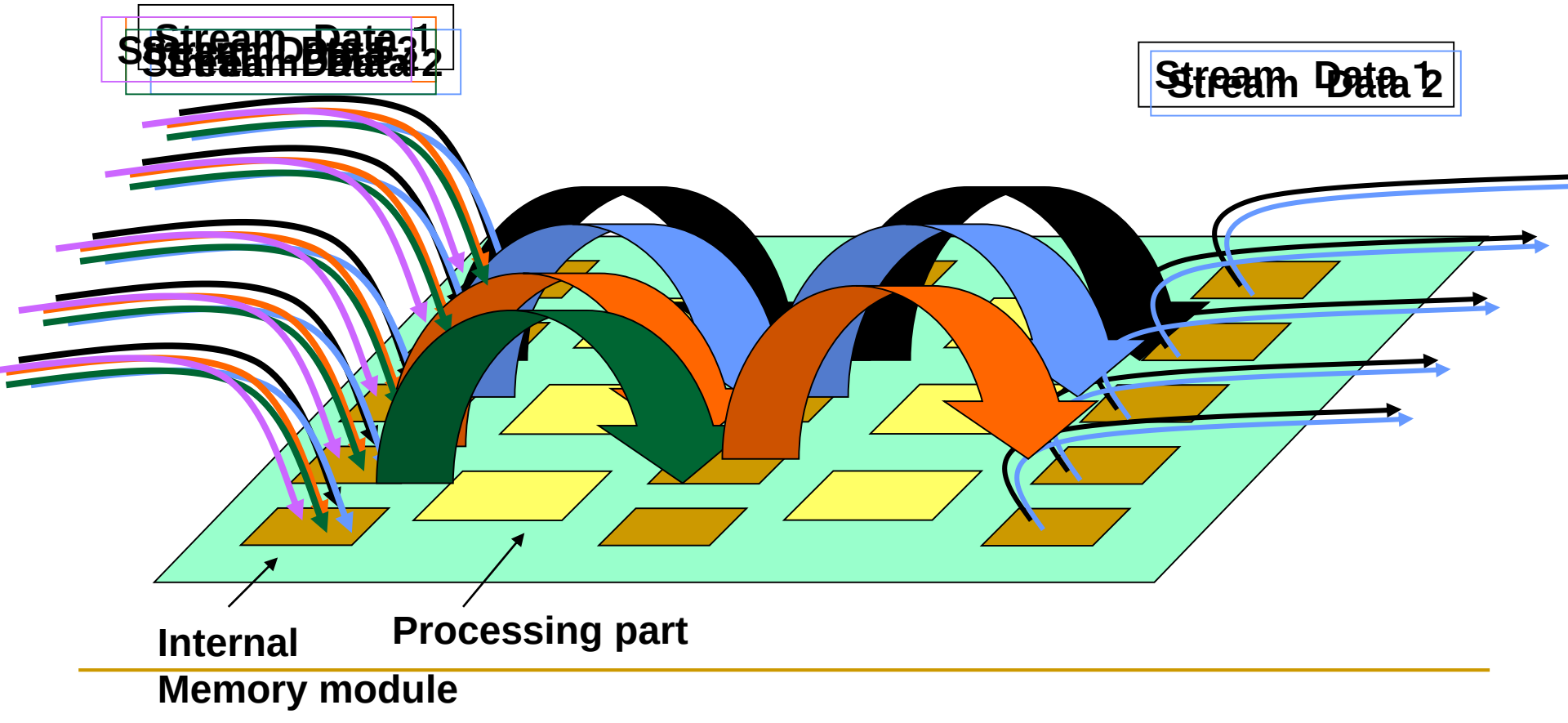
SIMD (Single Instruction-stream/ Multiple Data-stream)-like calculation

The same instruction is applied to different data stream
In Reconfigurable Systems, the operation is not required to be same
（SIMD-like calculation）

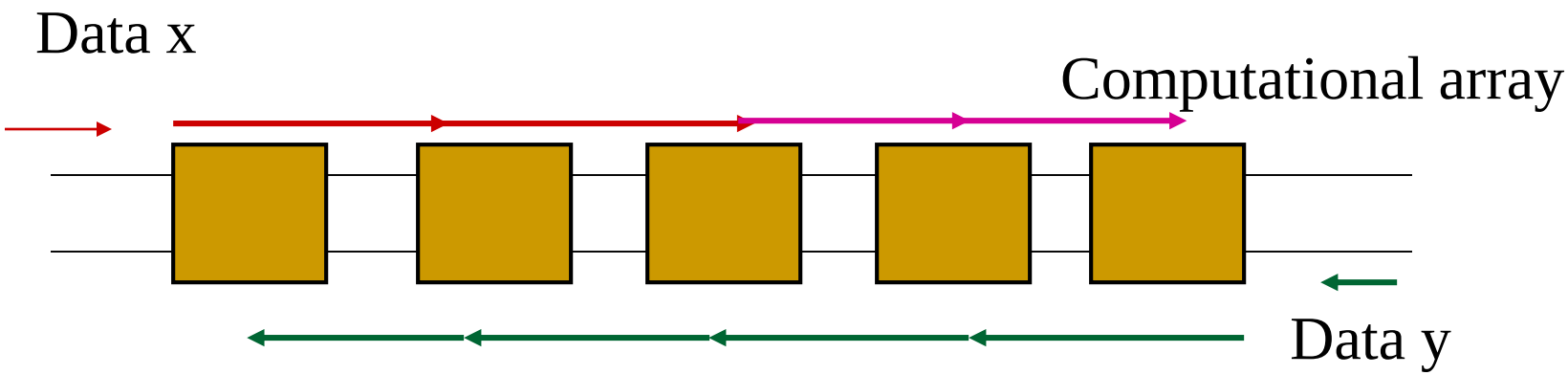


Pipelined structure

The stream is divided and inserted periodically.



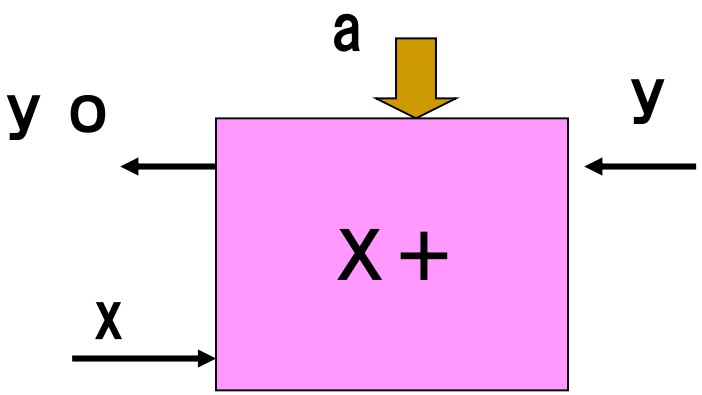
Systolic Algorithm



Data stream x, y are inserted with a certain interval.
When two stream meet each other, a calculation is executed.
→ Systolic: The beat of heart

Band matrix multiply y=Ax

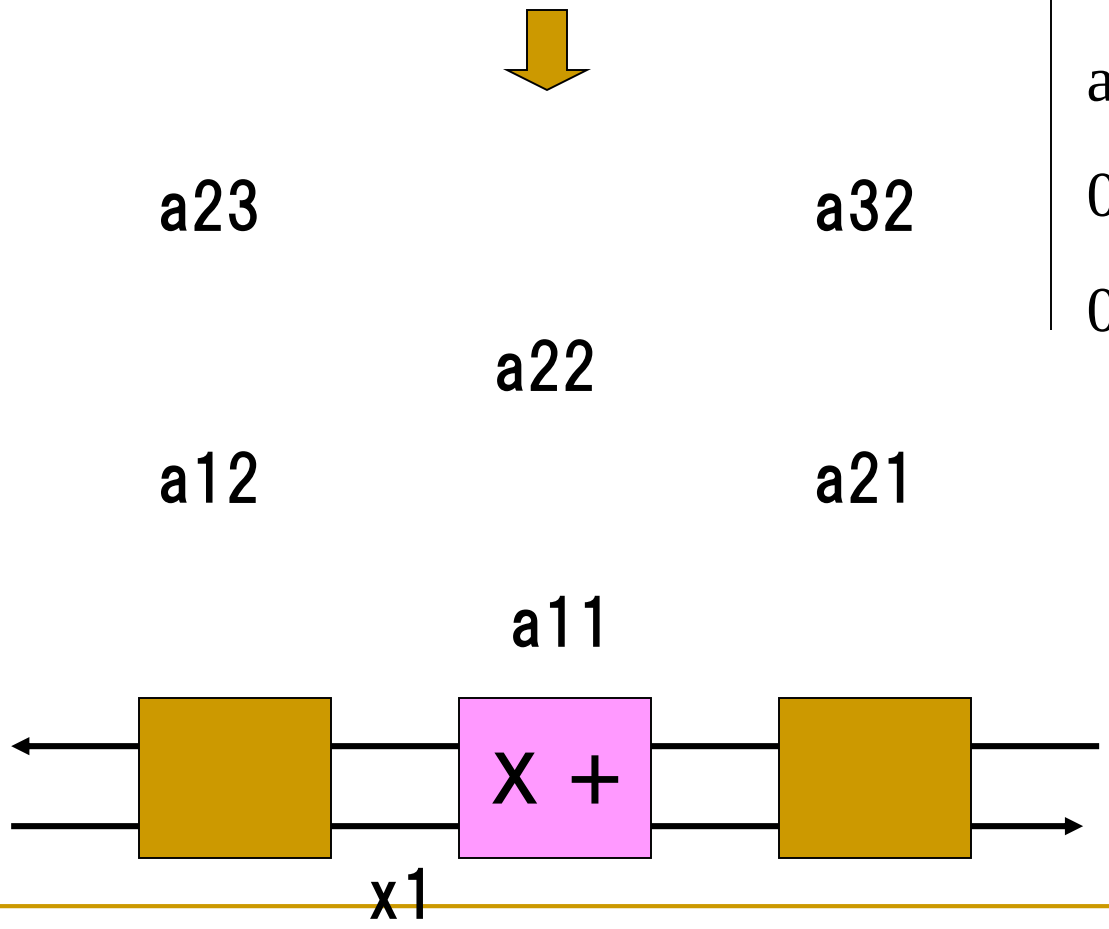
$$\begin{vmatrix} y0 \\ y1 \\ y2 \\ y3 \end{vmatrix} = \begin{vmatrix} a11 & a12 & 0 & 0 \\ a21 & a22 & a23 & 0 \\ 0 & a32 & a33 & a34 \\ 0 & 0 & a43 & a44 \end{vmatrix} \begin{vmatrix} x0 \\ x1 \\ x2 \\ x3 \end{vmatrix}$$



y o = a x + y i

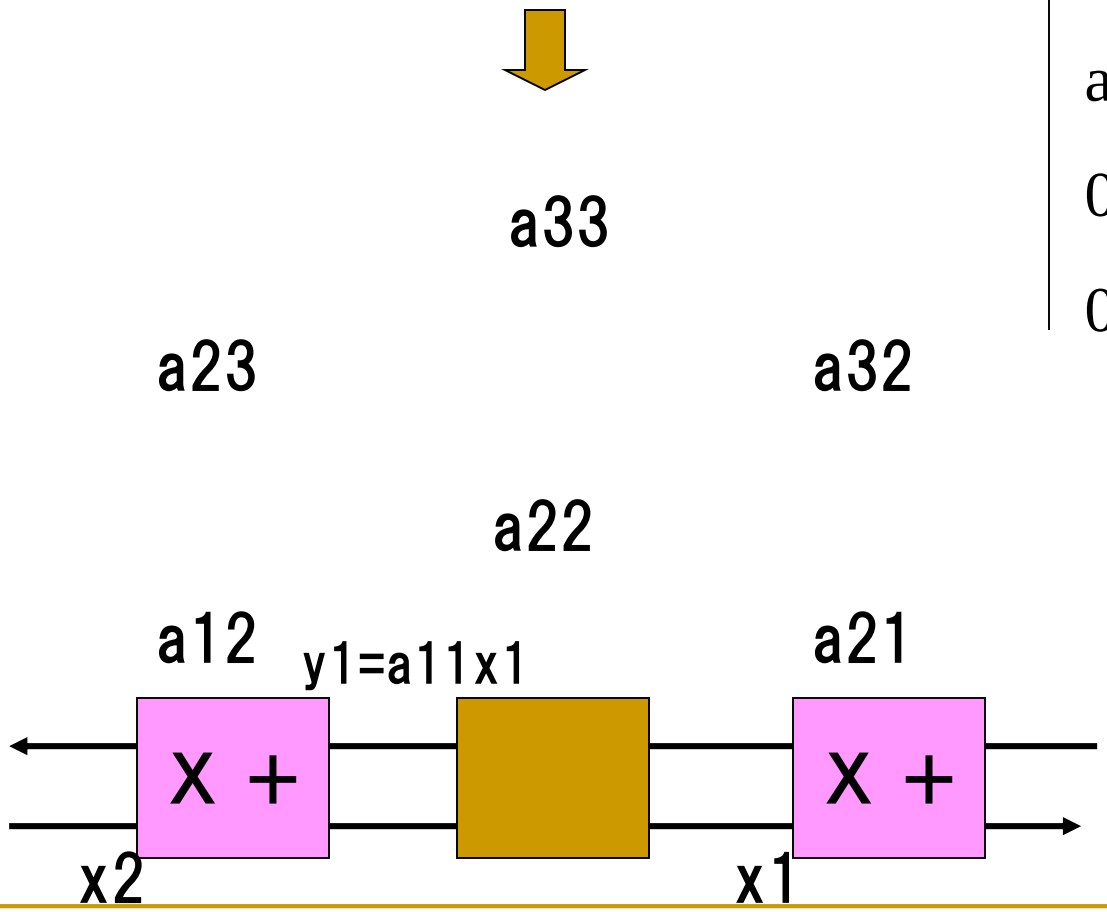
Band matrix multiply y=Ax

$$\begin{vmatrix} a11 & a12 & 0 & 0 \\ a21 & a22 & a23 & 0 \\ 0 & a32 & a33 & a34 \\ 0 & 0 & a43 & a44 \end{vmatrix}$$



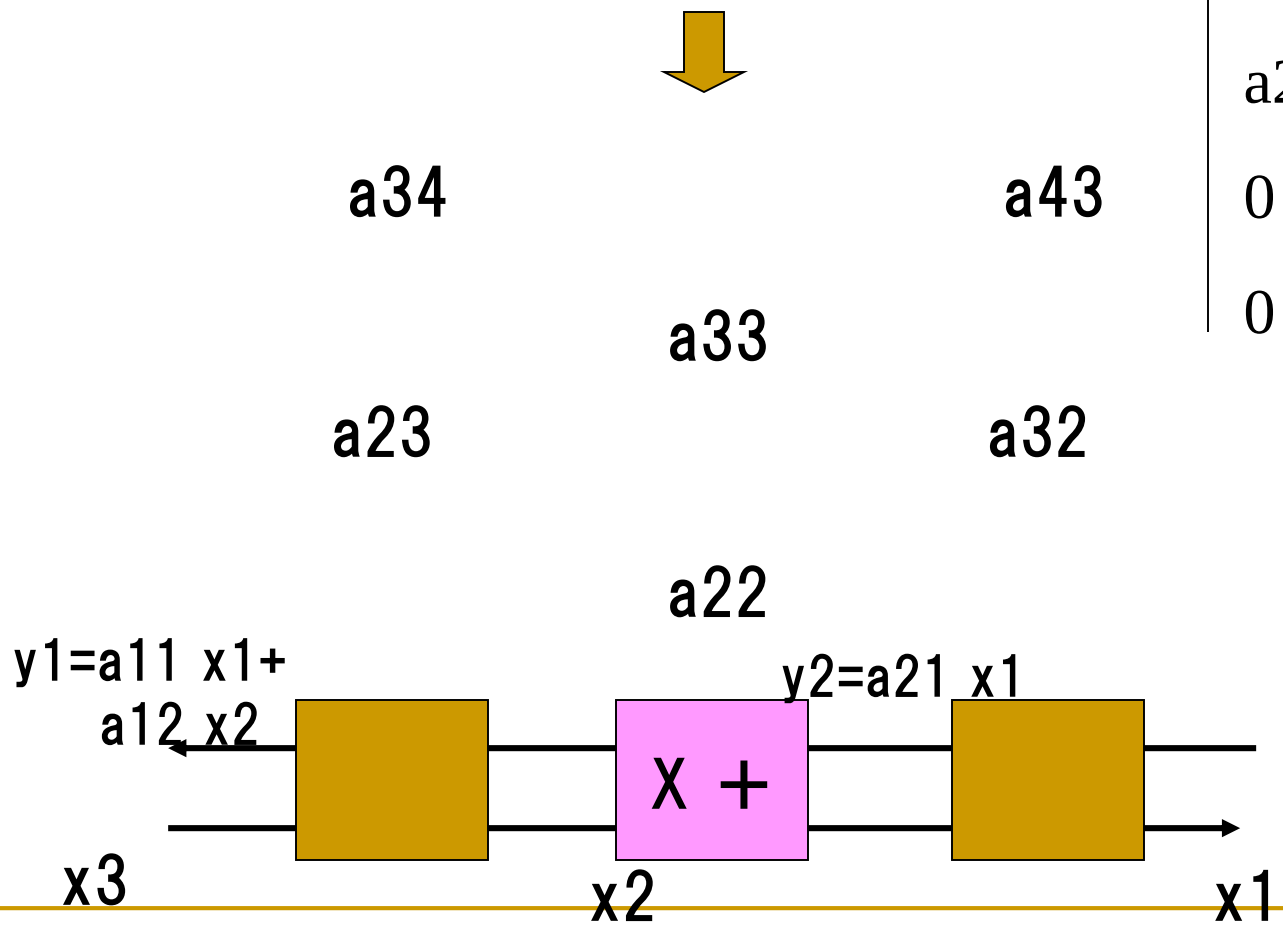
Band matrix multiply y=Ax

$$\begin{vmatrix} a11 & a12 & 0 & 0 \\ a21 & a22 & a23 & 0 \\ 0 & a32 & a33 & a34 \\ 0 & 0 & a43 & a44 \end{vmatrix}$$



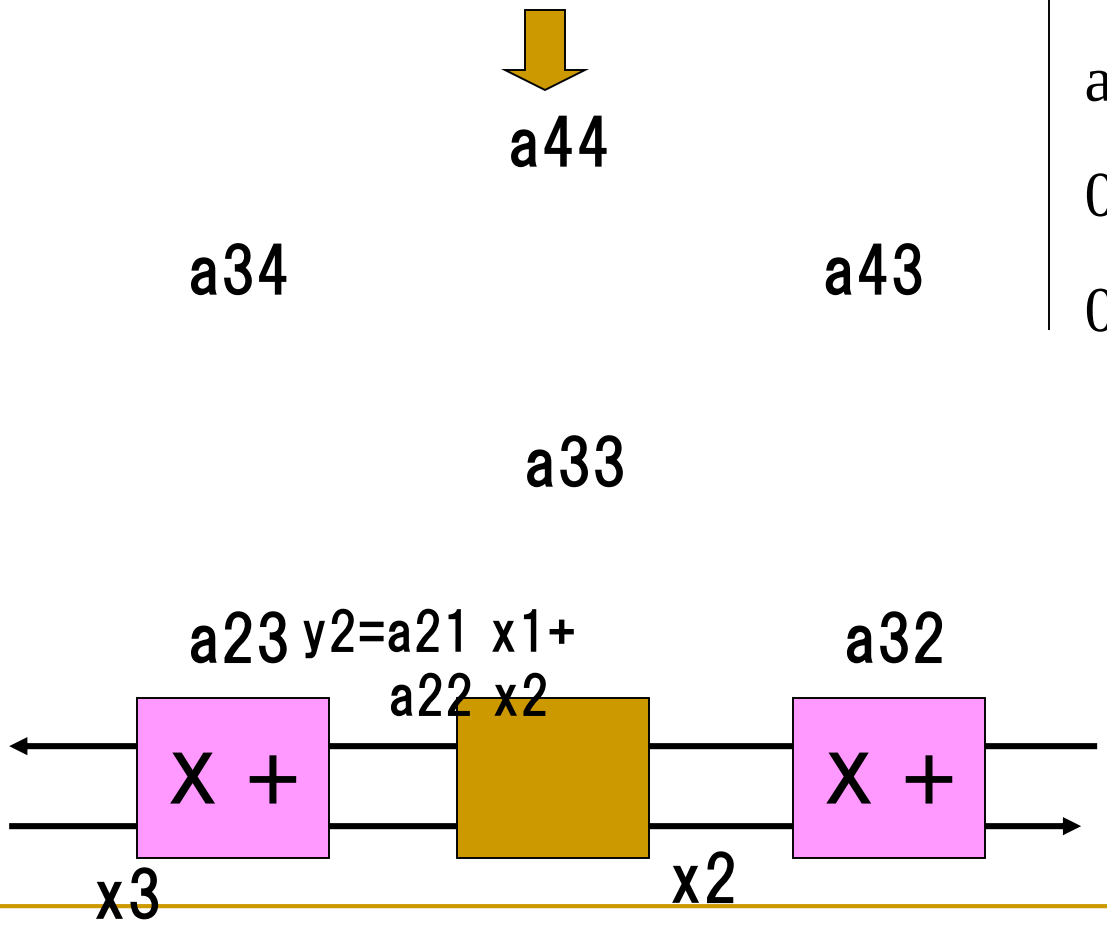
Band matrix multiply y=Ax

$$\begin{vmatrix} a11 & a12 & 0 & 0 \\ a21 & a22 & a23 & 0 \\ 0 & a32 & a33 & a34 \\ 0 & 0 & a43 & a44 \end{vmatrix}$$



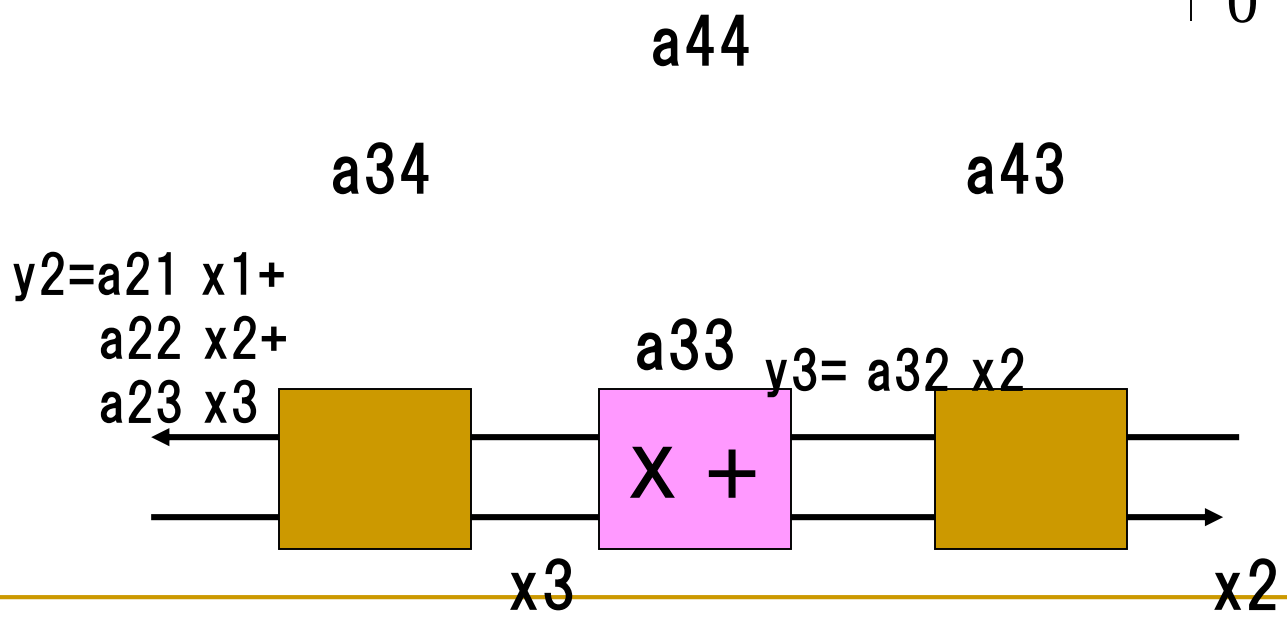
Band matrix multiply y=Ax

$$\begin{vmatrix} a11 & a12 & 0 & 0 \\ a21 & a22 & a23 & 0 \\ 0 & a32 & a33 & a34 \\ 0 & 0 & a43 & a44 \end{vmatrix}$$

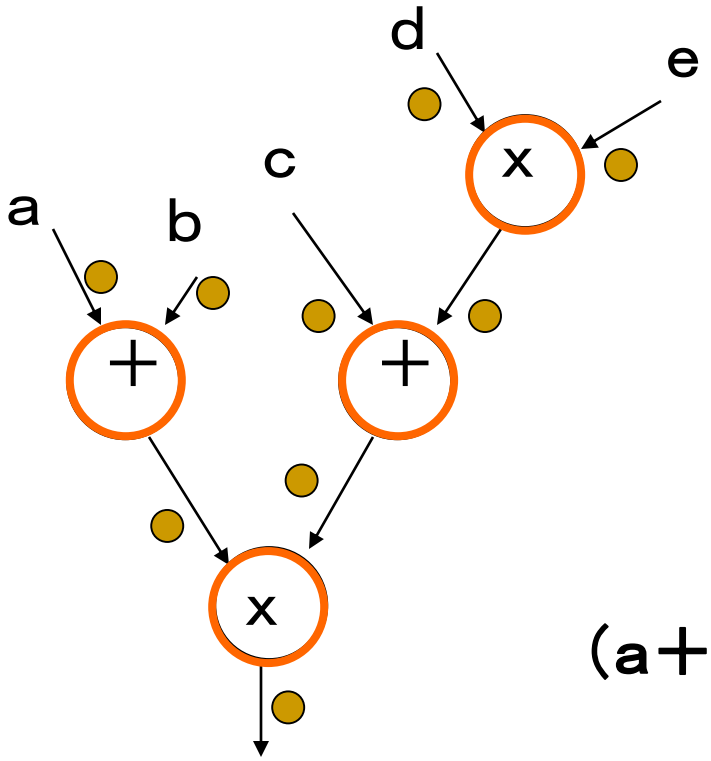


Band matrix multiply y=Ax

$$\begin{vmatrix} a11 & a12 & 0 & 0 \\ a21 & a22 & a23 & 0 \\ 0 & a32 & a33 & a34 \\ 0 & 0 & a43 & a44 \end{vmatrix}$$



Data flow algorithm



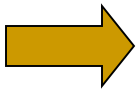
The process is activated with the available of tokens (data)

(a+b)x(c+(dxe))

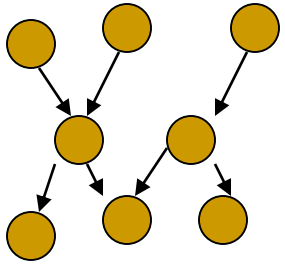
The overhead of synchronization is large.

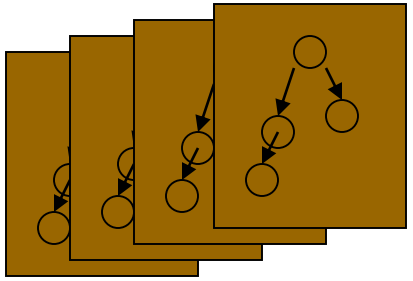
Data flow analysis and hardware generation

Data Flow Language



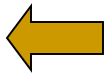
Data Flow Graph





Graph Decomposition

HDL Description

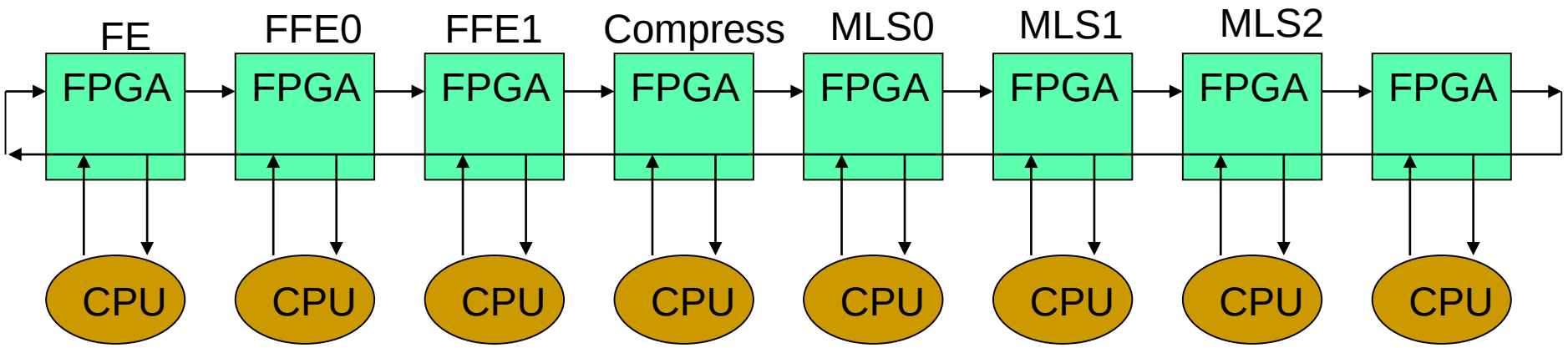


Configuration Data

Suitable for automatic generation of hardware

Microsoft's Catapult

Rank computation for Web search on Bing.
Task Level Macro-Pipelining (MISD)
FE: Feature Extraction
FFE: Free Form Expression: Synthesis of feature values
MLS: Machine Learning Scoring



FPGA: Altera's Stratix V

2-Dimensional Mesh is formed (8x6) for 1 cluster.

Historical flow of computer systems