

# ***Une introduction à la synthèse de haut-niveau***

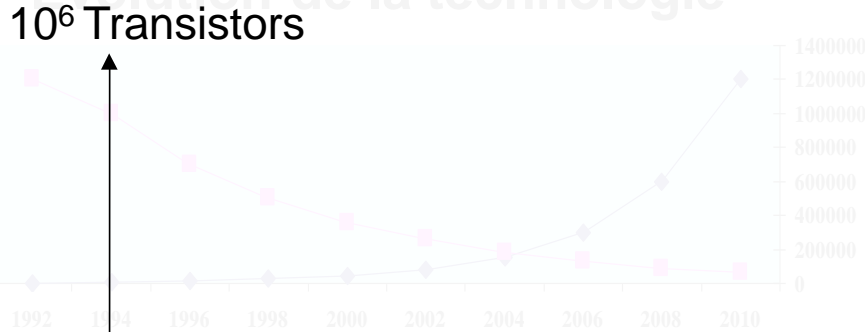
***(ou comment générer des architectures  
matérielles à partir du langage C)***

**Université de Bretagne-Sud  
Lab-STICC**

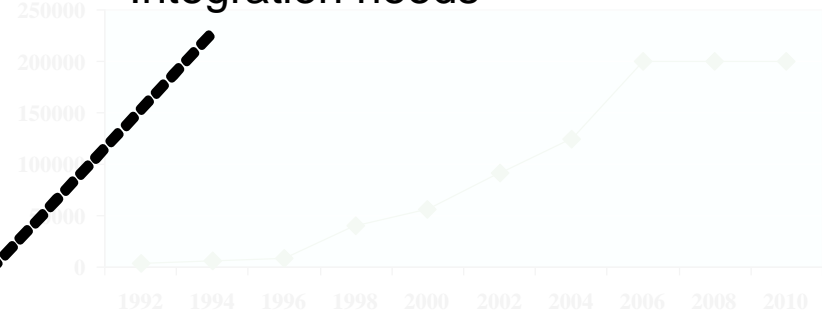
**Philippe COUSSY**  
**[philippe.coussy@univ-ubs.fr](mailto:philippe.coussy@univ-ubs.fr)**

# Productivity gap

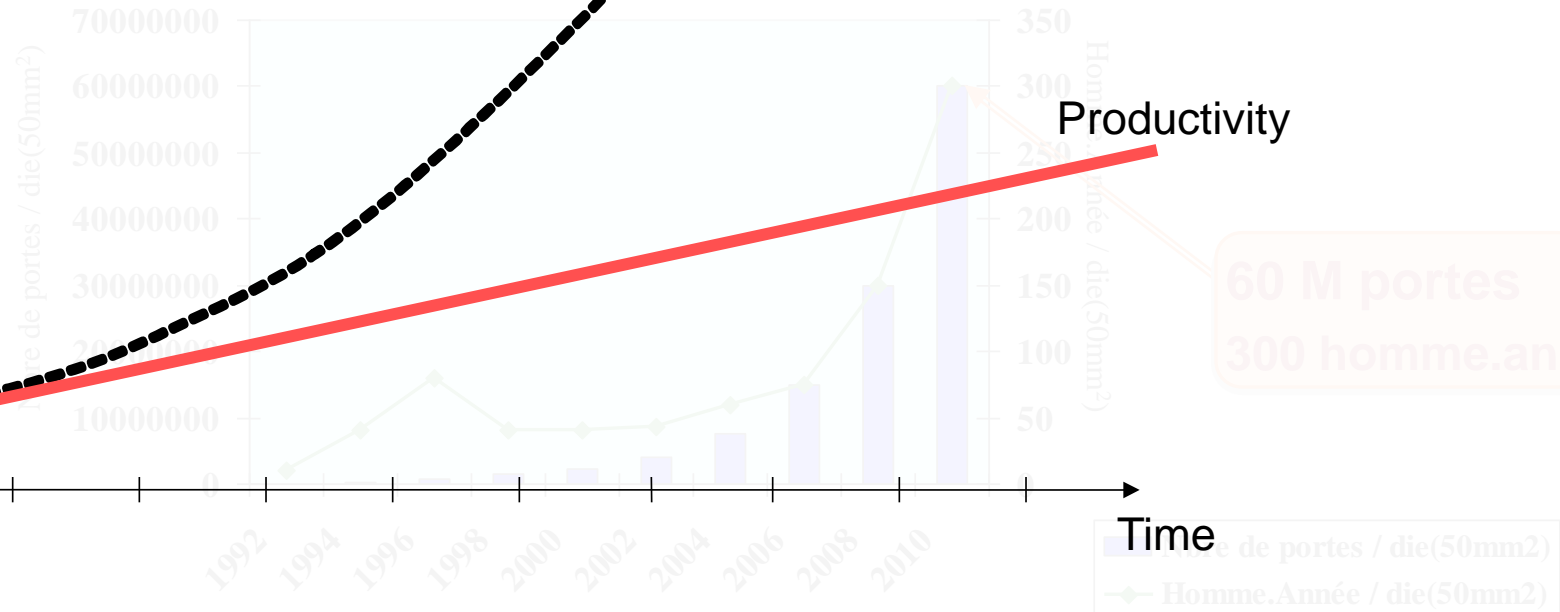
Evolution de la technologie



Evolution de la productivité



Evolution de la complexité des circuits

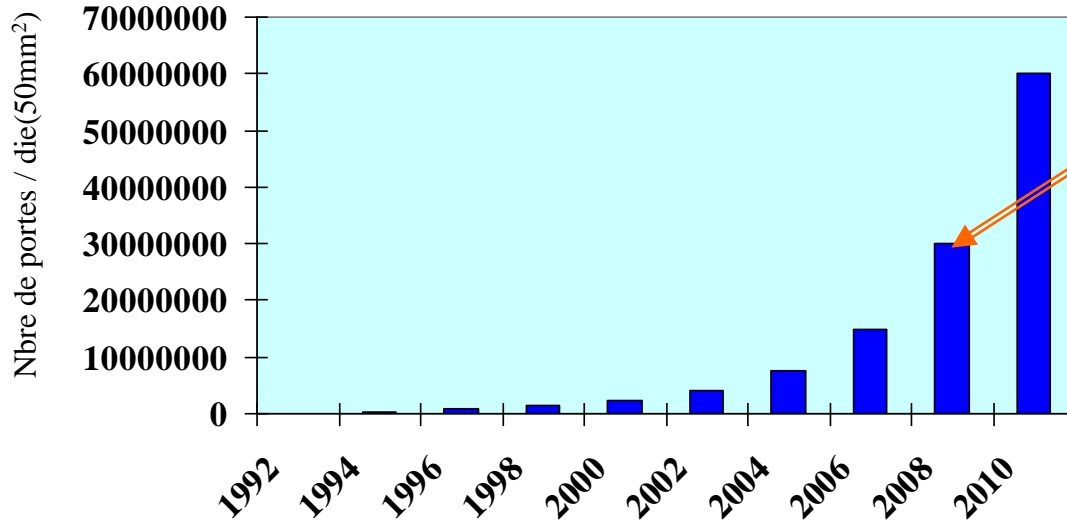
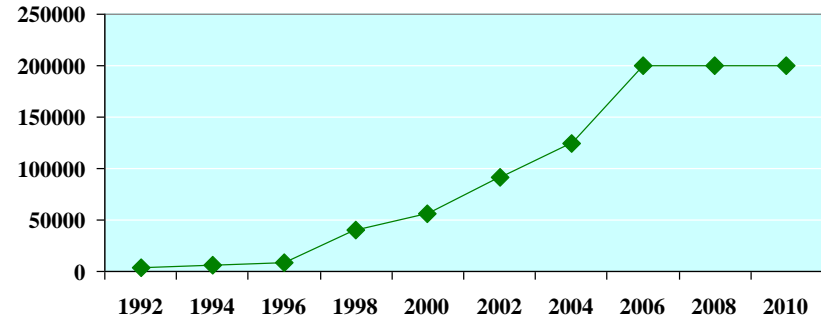
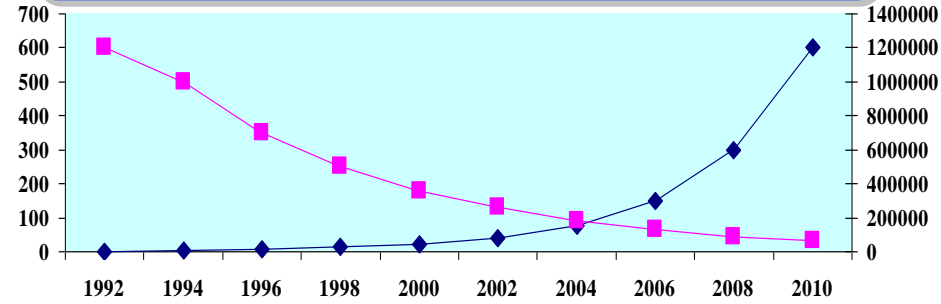


60 M portes  
300 homme.an

# Context

- Decreasing technology scale
- Increasing transistor density

- Designer productivity



- 30 M gates

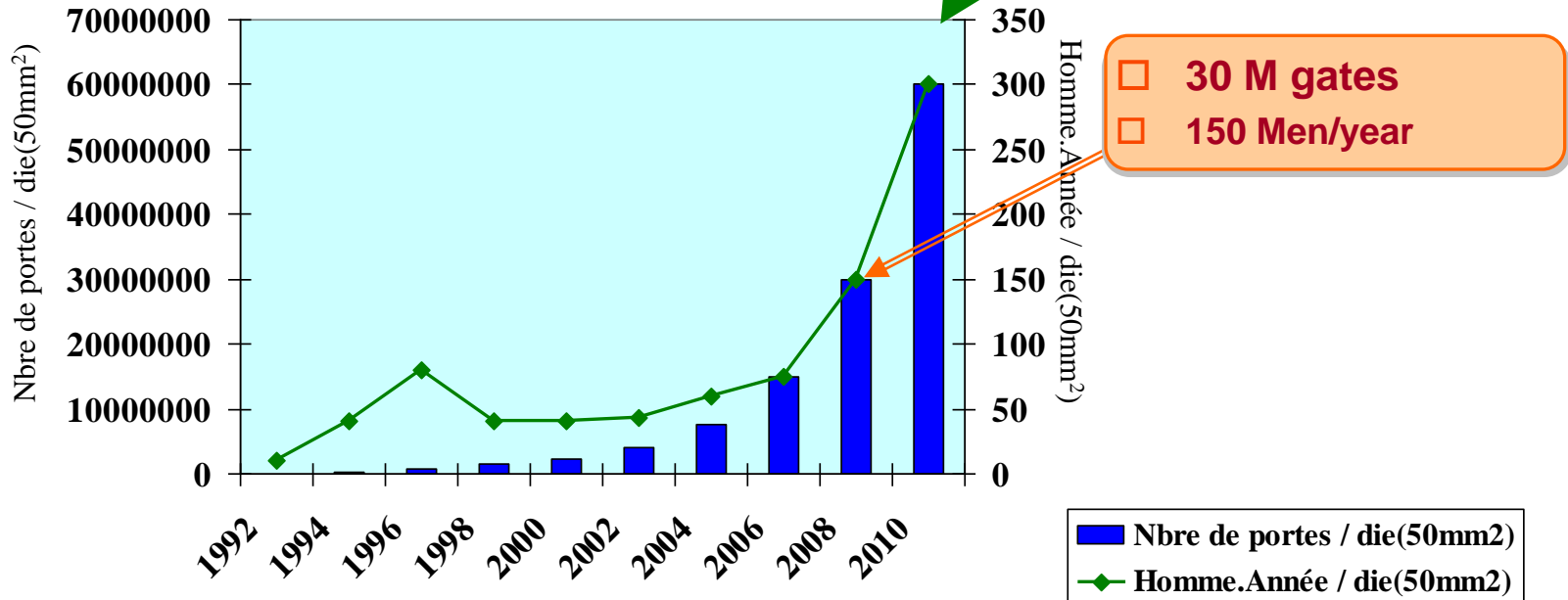
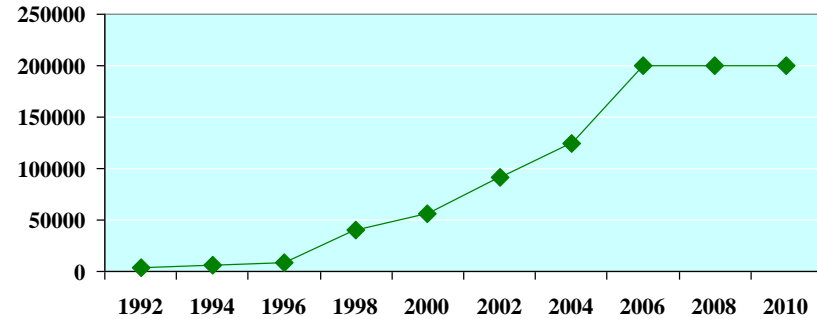
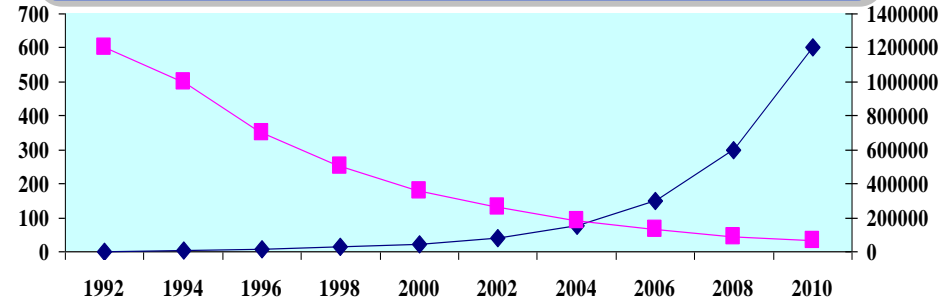
- Nbre de portes / die(50mm²)



# Context

- Decreasing technology scale
- Increasing transistor density

- Designer productivity



# Design methodologies

---

- **Synthesis and verification automation has always been key factors in the evolution of the design process**
  - Allow to explore the design space efficiently and rapidly
  - Correct by construction design

# Design methodologies

## □ Software domain

- Machine code (binary sequence)
- 1950s: concept of assembly language (and assembler)
  - *based on mnemonics*
  - *Maurice V. Wilkes de l'université de Cambridge*
- Later: High-level languages and compilers
  - *1951: First compiler*
    - (A-0 system) par Grace Hopper
  - *Fortran 1954-1957: First high-level language*
    - FORMula TRANslator
  - *Cobol 1959, Basic 1964, C 1972, C++ 1983...*

## □ High-level language

- Platform independent
- Follow the rules of human language
  - *with a grammar, a syntax and a semantic*
- Provide flexibility and portability
  - *by hiding details of the computer architecture*

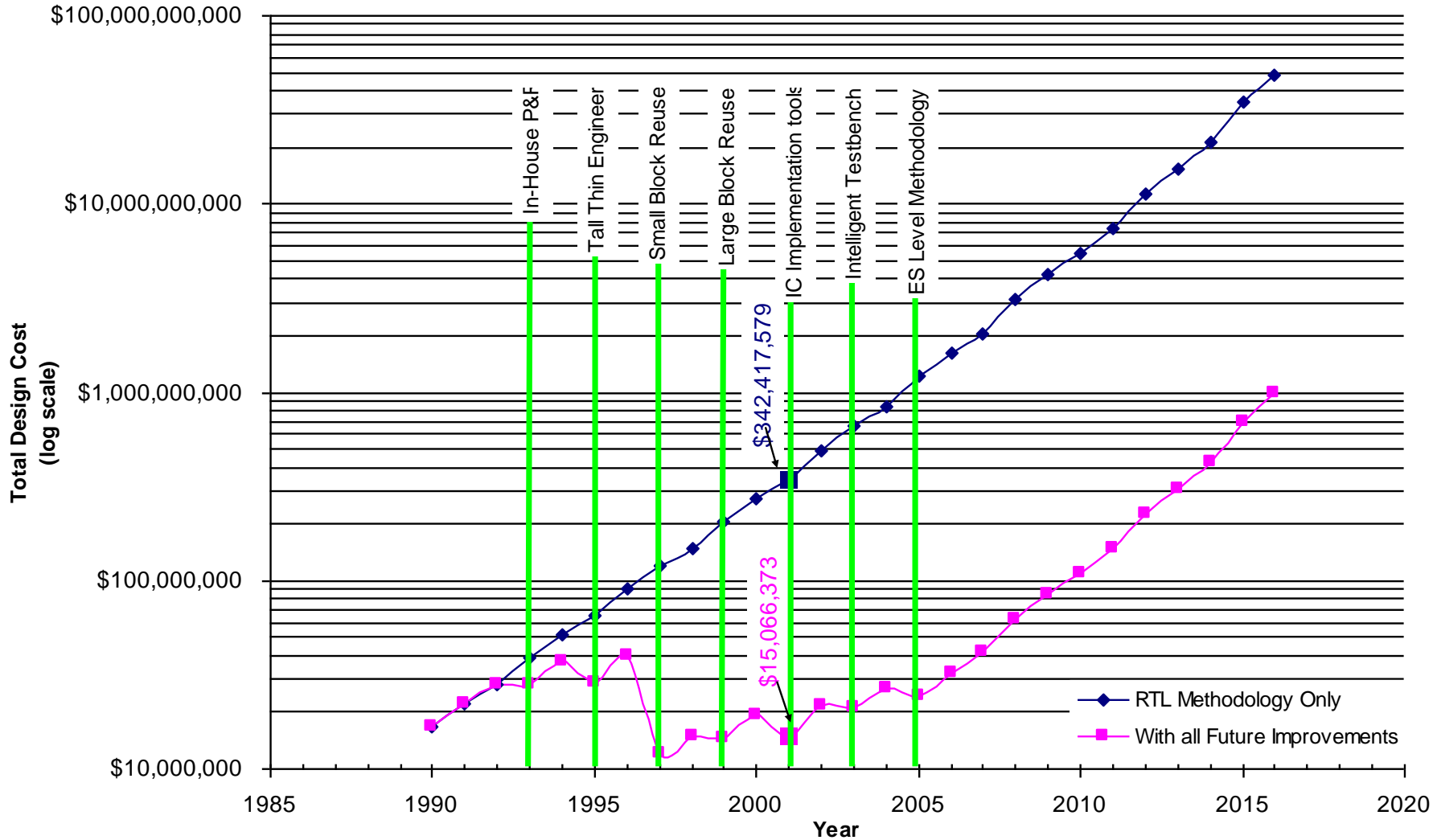
# Design methodologies

## □ Hardware domain

- 1960: IC were done by hand
  - *designed, optimized and laid out*
- 1970: Gate-level simulation
- end of 70: Cycle-based simulation
- 1980: Wide automation
  - *place & route, schematic circuit capture, formal verification and static timing analysis*
- Mid 1980: Hardware description language
  - *1986 Verilog, 1987 VHDL*
- 1990: logic synthesis
  - *VHDL and Verilog synthesizable subsets*
- Mid 1990:
  - *High-level synthesis (First gen),*
  - *Co-design, IP-core reuse...*
- 2000 : Electronic System Level ESL
  - *System level language*
    - SystemC, SystemVerilog...,
    - Virtual prototyping, Transaction Level Modelling TLM ...

# Design gap

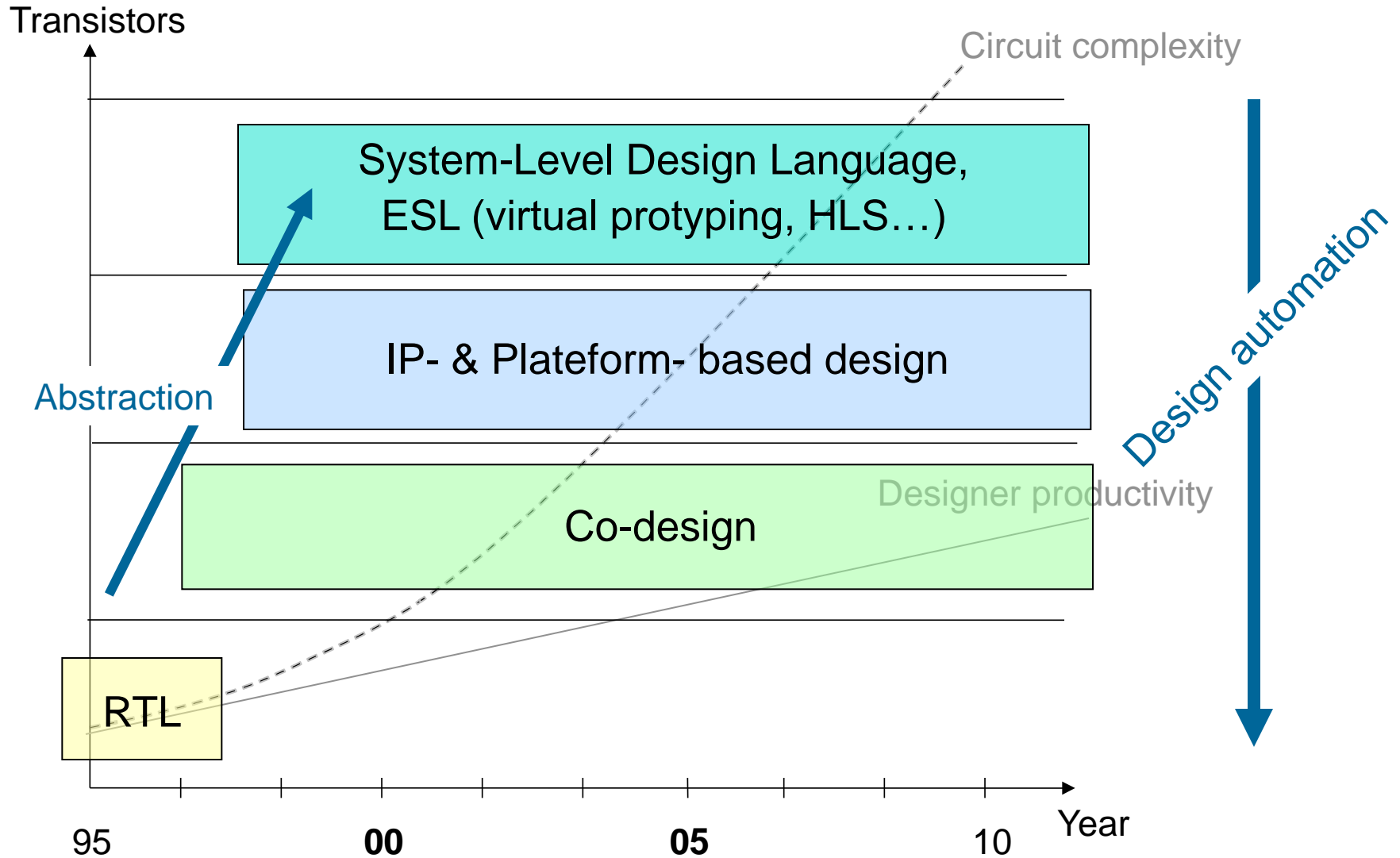
## SOC Design Cost Model



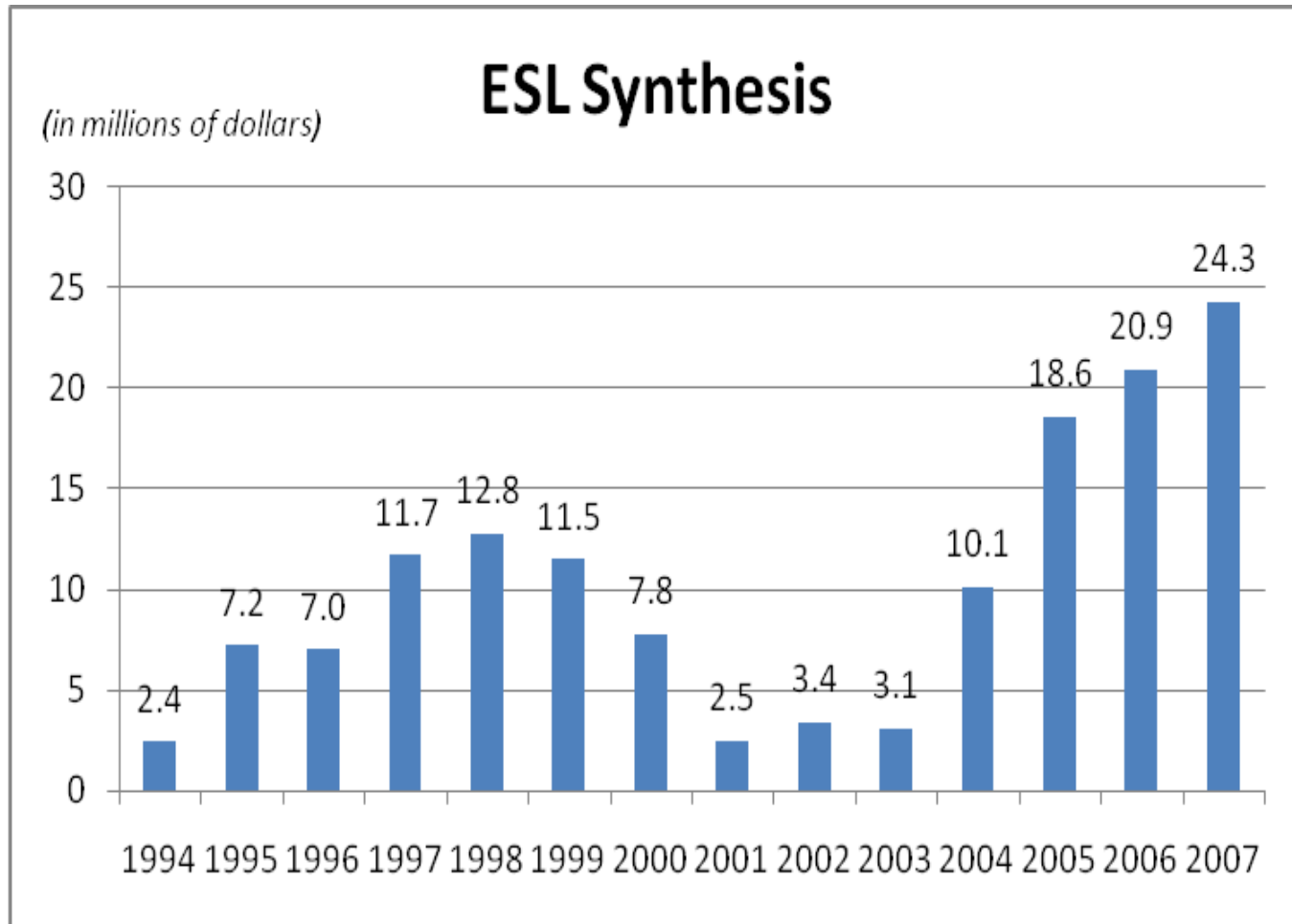
(A. B. Kahng, G. Smith, "A New Design Cost Model", For the 2001 ITRS)



# Electronic System Level Design (ESLD)



# ESL Market



# Outline

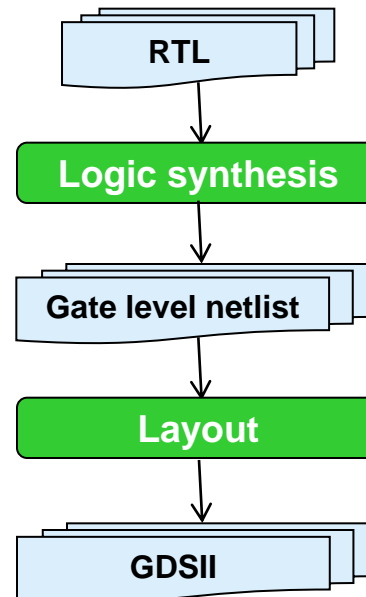
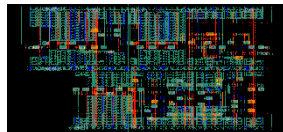
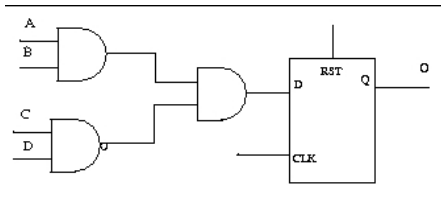
---

- Lab-STICC
- General context
- **High-Level Synthesis**
  - Brief introduction
  - “In details”
- **GAUT**
  - Overview
  - Results
- **Conclusion**
- **References**

# Typical HW design flow

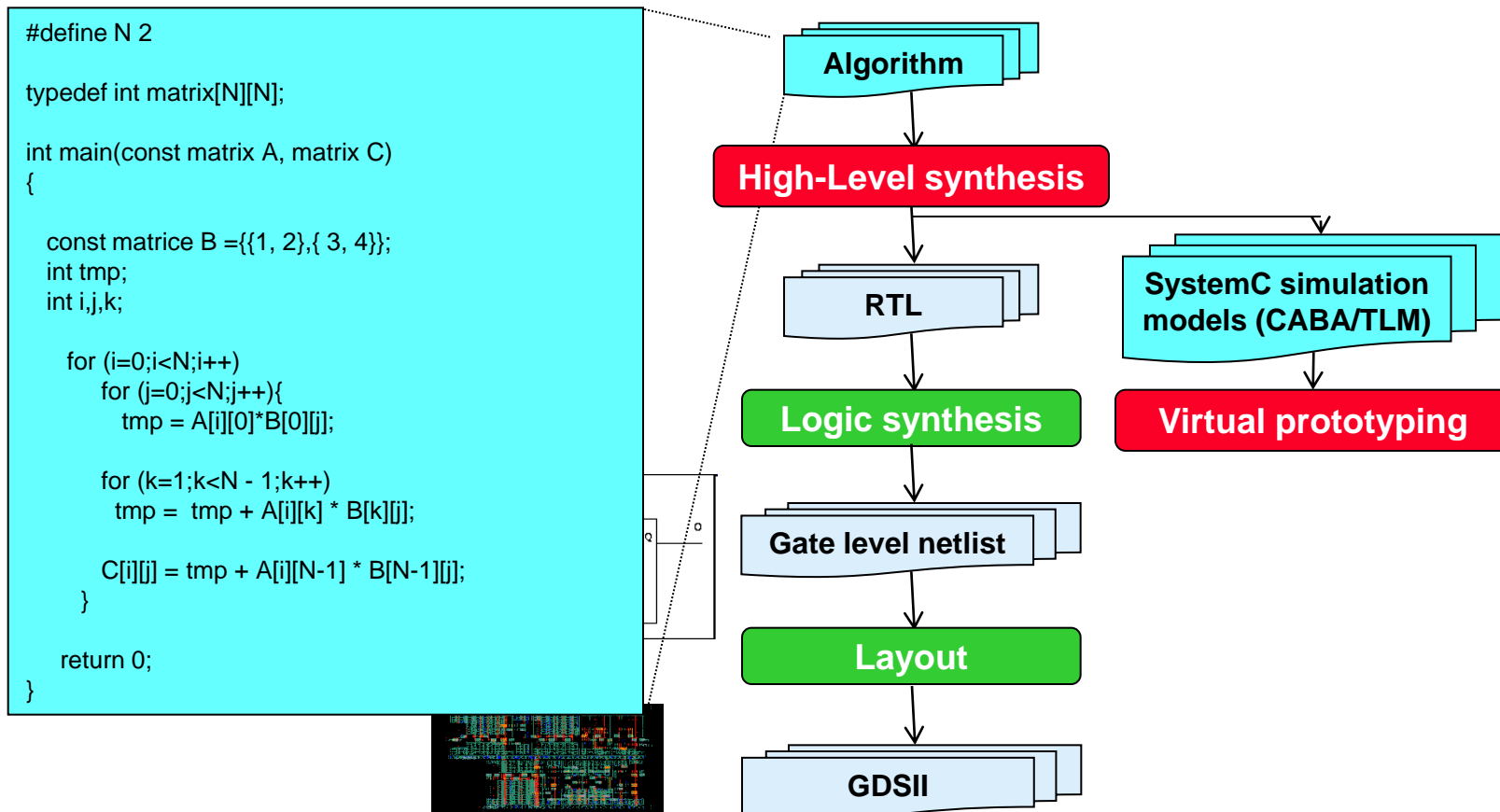
- Starting from a Register Transfer Level description, generate an IC layout

```
process( CLK, RST)
  if( RST = '1' ) then
    Q <= '0';
  else if rising_edge( CLK) then
    Q <= A and B and C nand D ;
```



# Typical HW design flow

- Starting from a functional description, automatically generate an RTL architecture



# High-level synthesis

---

- **Starting from a functional description, automatically generate an RTL architecture**
  
- **Constraints**
  - Timing constraints: latency and/or throughput
  - Resource constraints: #Operators and/or #Registers and/or #Memory, #Slices...
  
- **Objectives**
  - Minimization: area i.e. resources, latency, power consumption...
  - Maximization: throughput

# Synthesis steps

---

## **Compilation**

- Generates a formal modeling of the specification

## **Selection**

- Chooses the architecture of the operators

## **Allocation**

- Defines the number of operators for each selected type

## **Scheduling**

- Defines the execution date of each operation

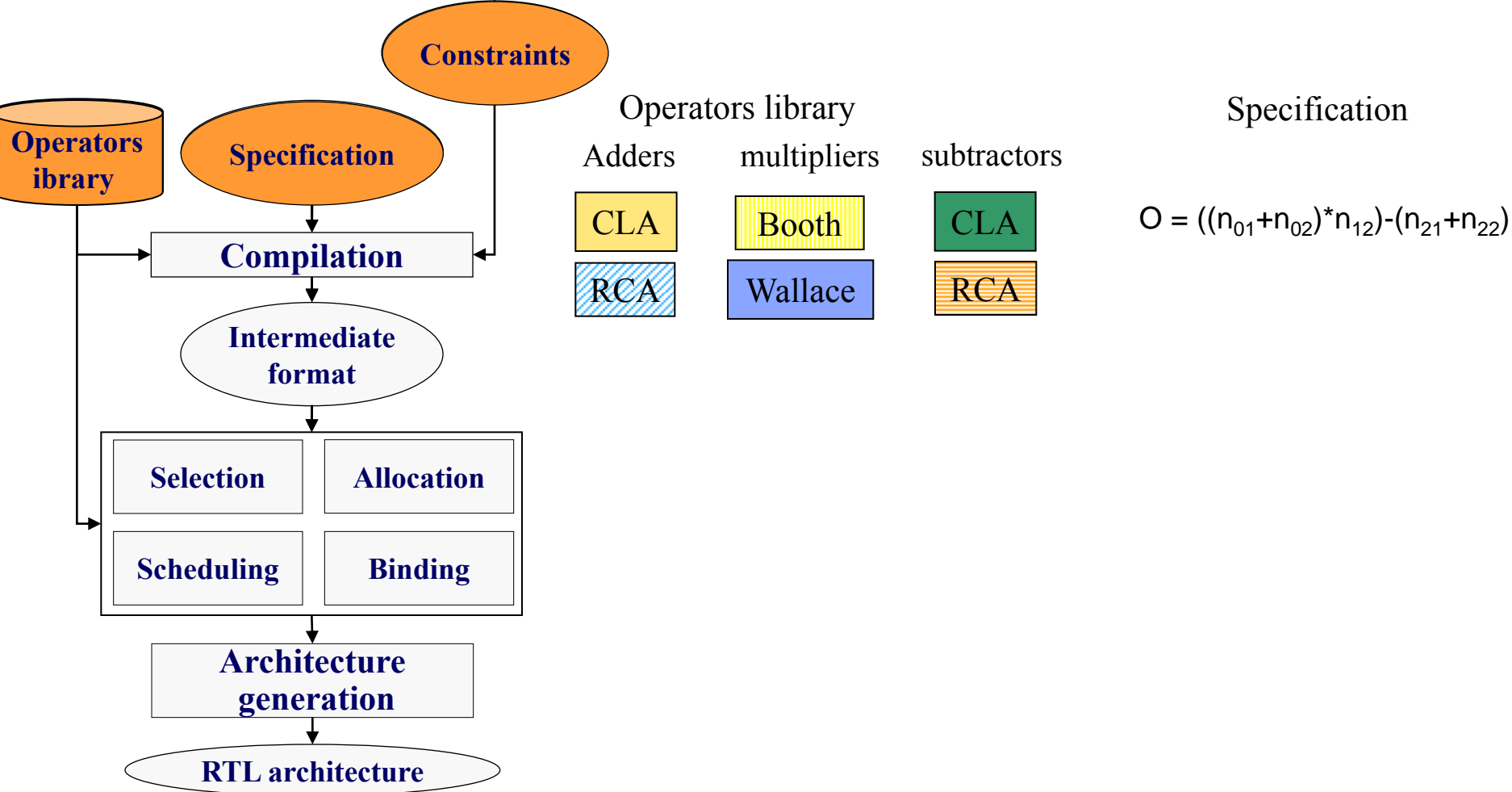
## **Binding (or Assignment)**

- Defines which operator will execute a given operation
- Defines which memory element will store a data

## **Architecture generation**

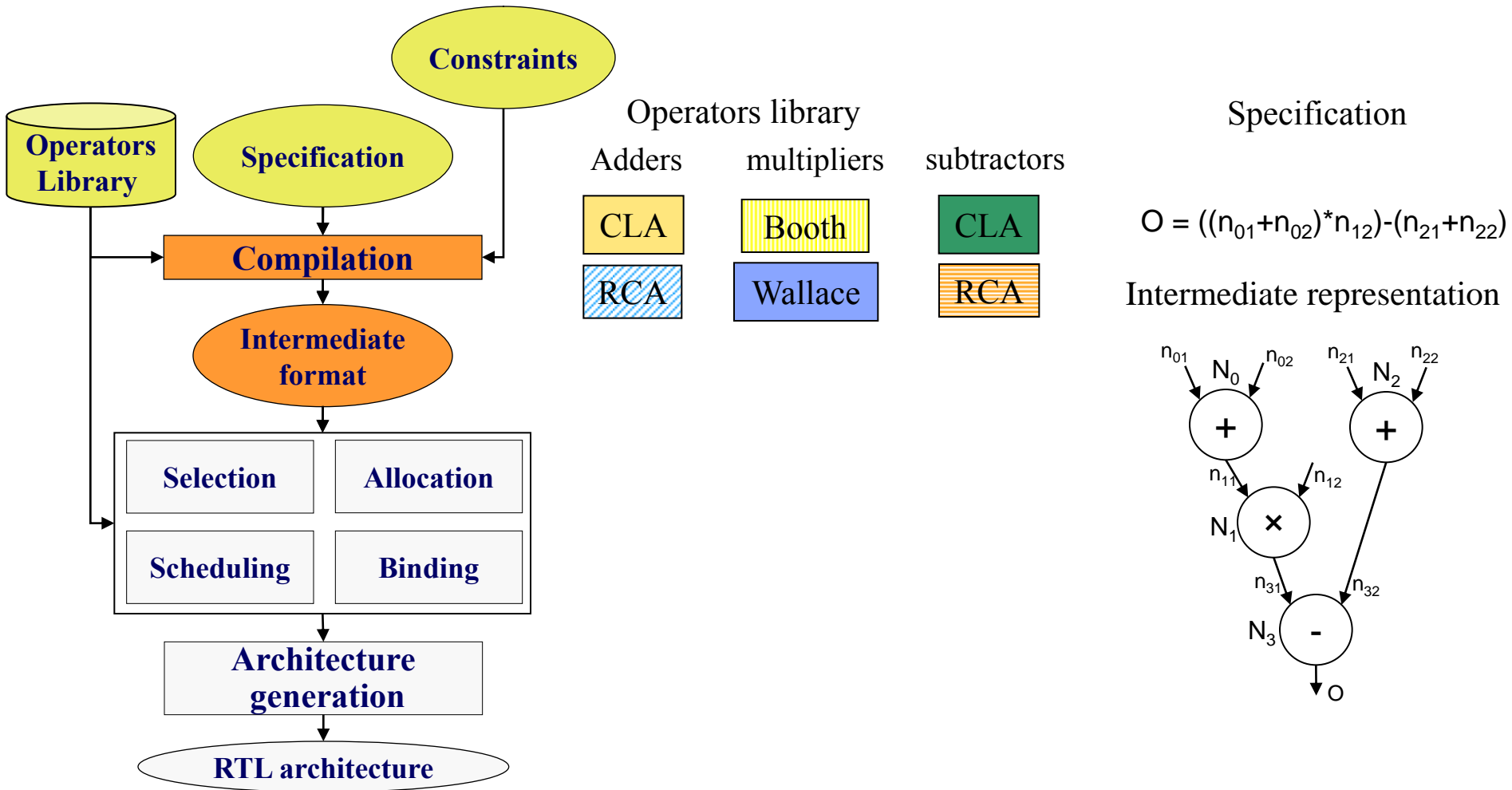
- Writes out the RTL source code in the target language e.g. VHDL

# HLS steps: inputs





# HLS steps: Compilation



# Synthesis steps

---

## Compilation

- Generates a formal modeling of the specification

## Selection

- Chooses the architecture of the operators

## Allocation

- Defines the number of operators for each selected type

## Scheduling

- Defines the execution date of each operation

## Binding (or Assignment)

- Defines which operator will execute a given operation
- Defines which memory element will store a data

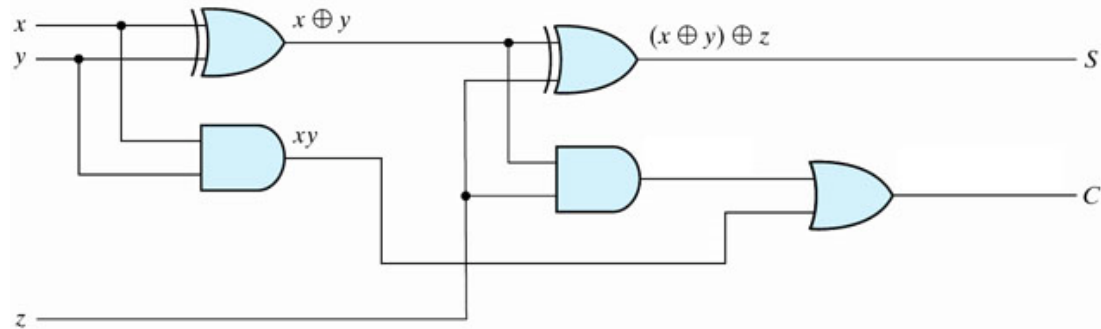
## Architecture generation

- Writes out the RTL source code in the target language e.g. VHDL

# Operator architecture

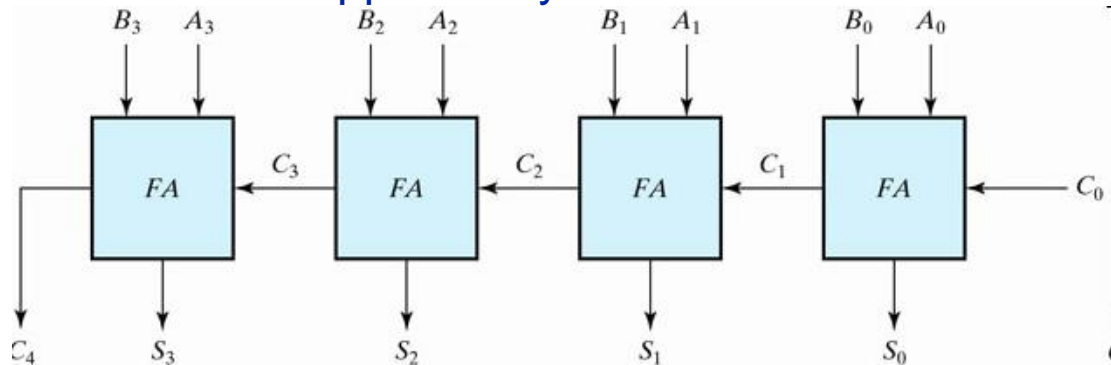
## □ Full 1-bit adder : $X + Y + Z$

- X, Y are the operands
- Z is the input carry



## □ Ripple Carry Adder

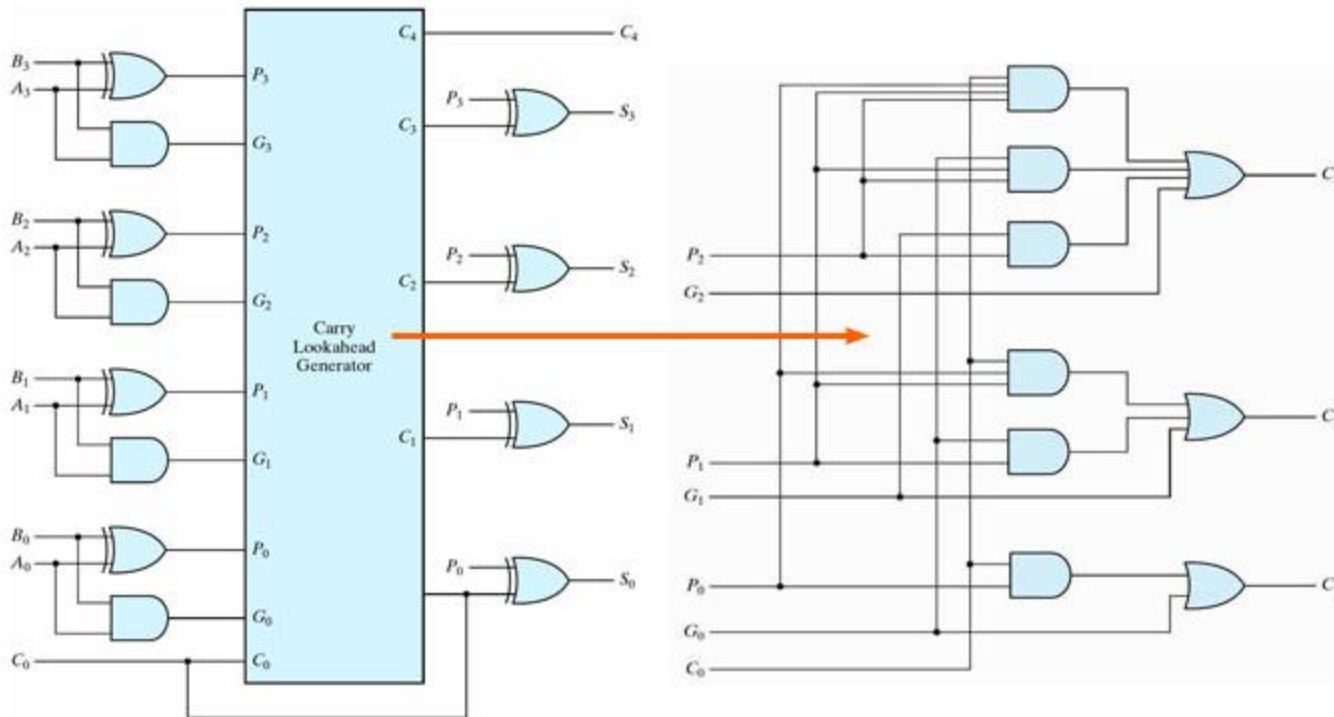
- Add two integers A and B
- Cascade of 1-bit adders  $\Rightarrow$  Ripple-Carry Adder



# Operator architecture

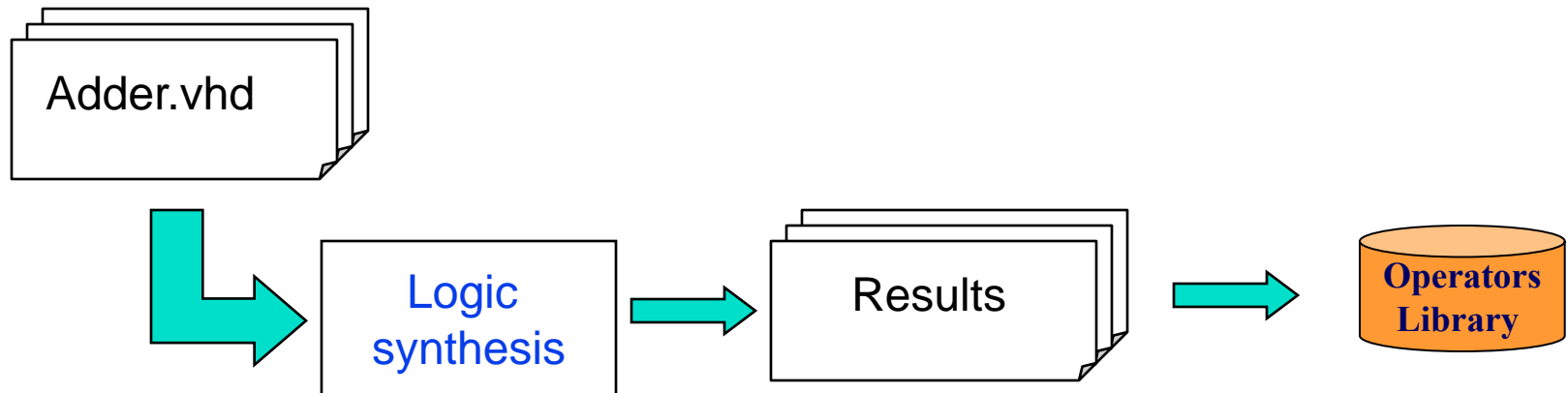
## □ Carry Look-ahead adder CLA

- Uses a carry generator to compute all the carries concurrently
  - *faster but also larger than the RCA*

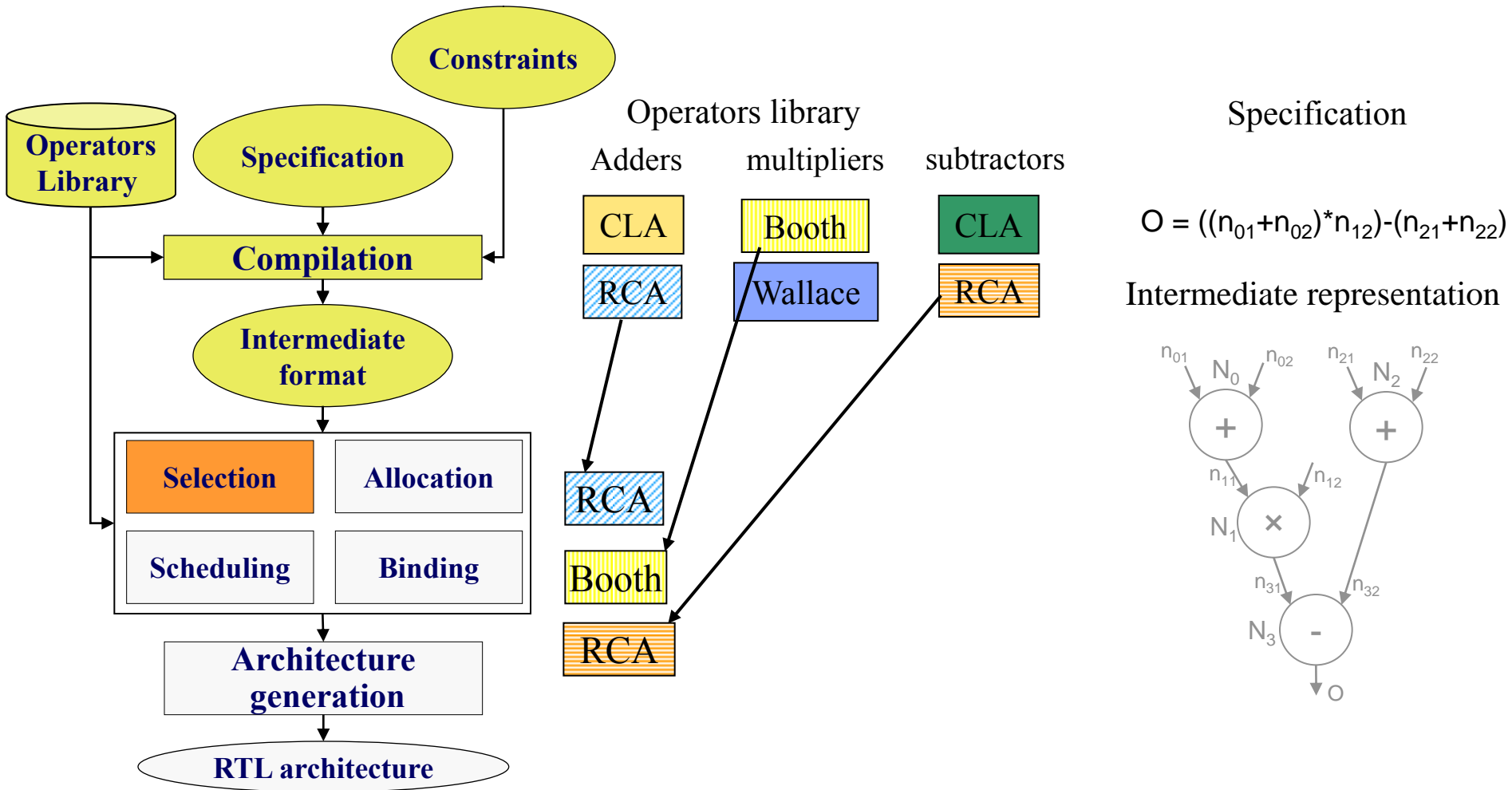


# Library characterization

- RTL architecture produced by HLS depends on the capabilities and characteristics of the operators
- Library processing reads the available libraries and determines the functional, timing, and area characteristics of the available parts.



# HLS steps: Selection



# Synthesis steps

---

## Compilation

- Generates a formal modeling of the specification

## Selection

- Chooses the architecture of the operators

## Allocation

- Defines the number of operators for each selected type

## Scheduling

- Defines the execution date of each operation

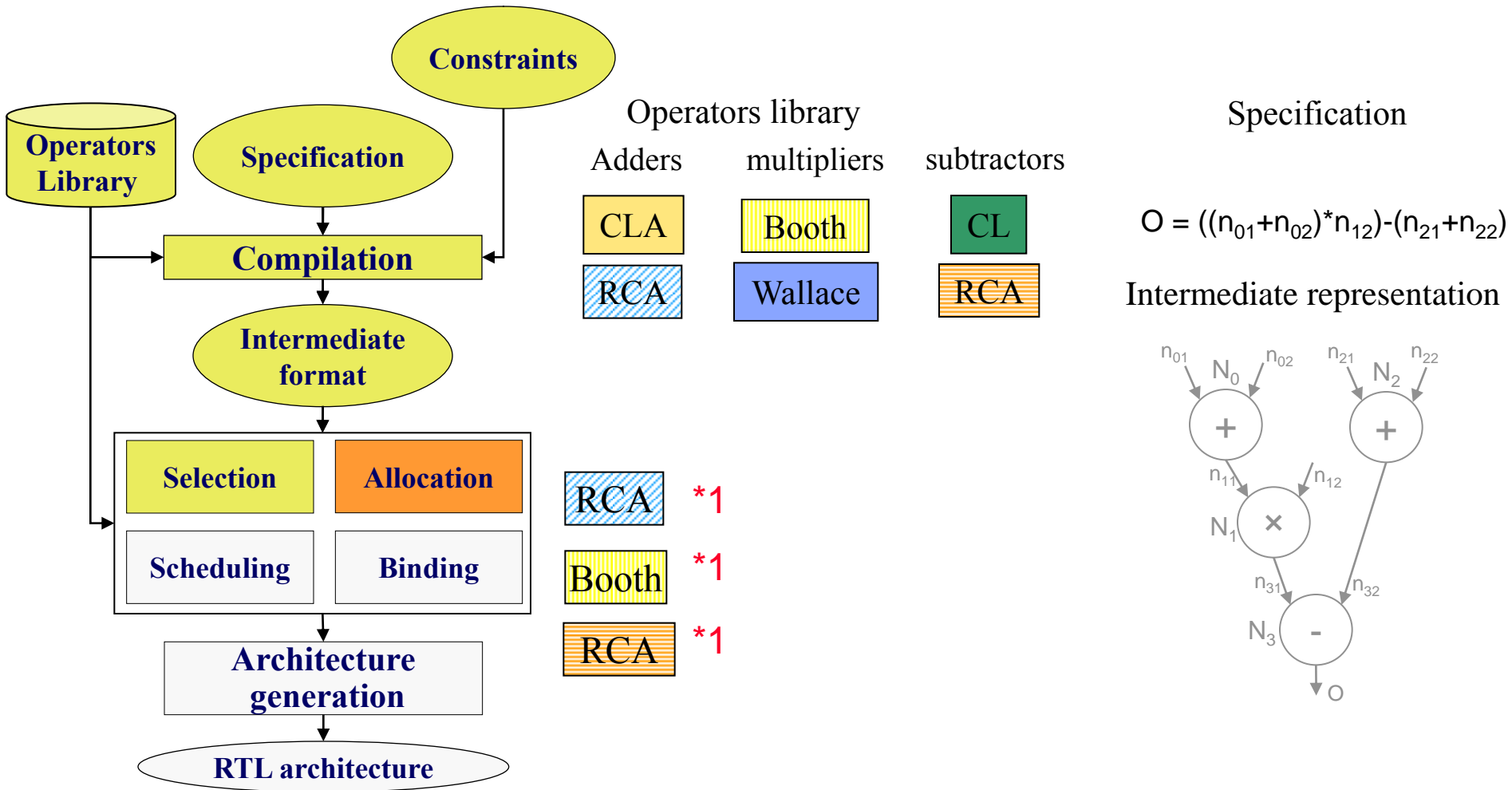
## Binding (or Assignment)

- Defines which operator will execute a given operation
- Defines which memory element will store a data

## Architecture generation

- Writes out the RTL source code in the target language e.g. VHDL

# HLS steps: allocation





# Synthesis steps

---

## Compilation

- Generates a formal modeling of the specification

## Selection

- Chooses the architecture of the operators

## Allocation

- Defines the number of operators for each selected type

## Scheduling

- Defines the execution date of each operation

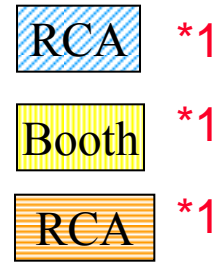
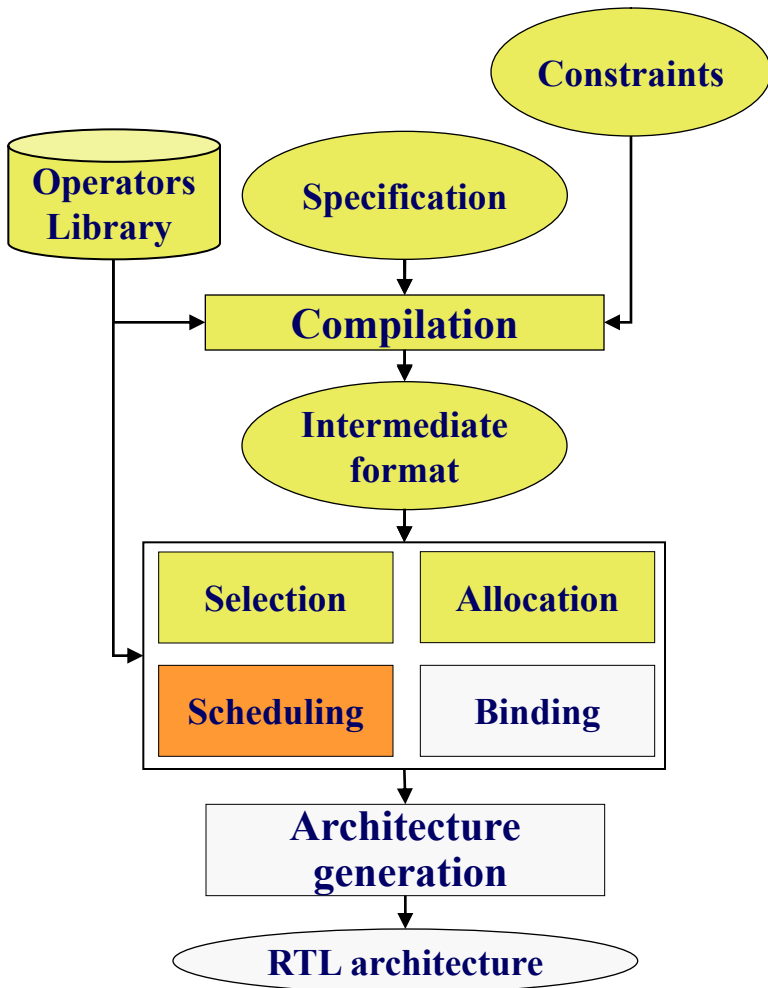
## Binding (or Assignment)

- Defines which operator will execute a given operation
- Defines which memory element will store a data

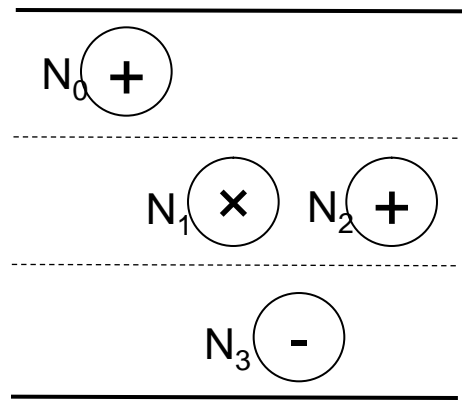
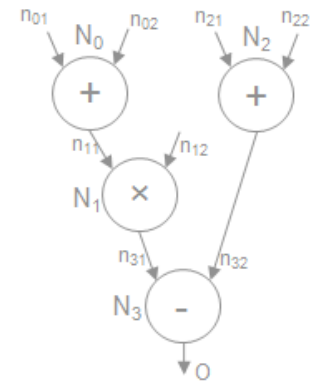
## Architecture generation

- Writes out the RTL source code in the target language e.g. VHDL

# HLS steps: scheduling



Intermediate representation



# Synthesis steps

---

## Compilation

- Generates a formal modeling of the specification

## Selection

- Chooses the architecture of the operators

## Allocation

- Defines the number of operators for each selected type

## Scheduling

- Defines the execution date of each operation

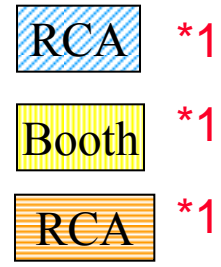
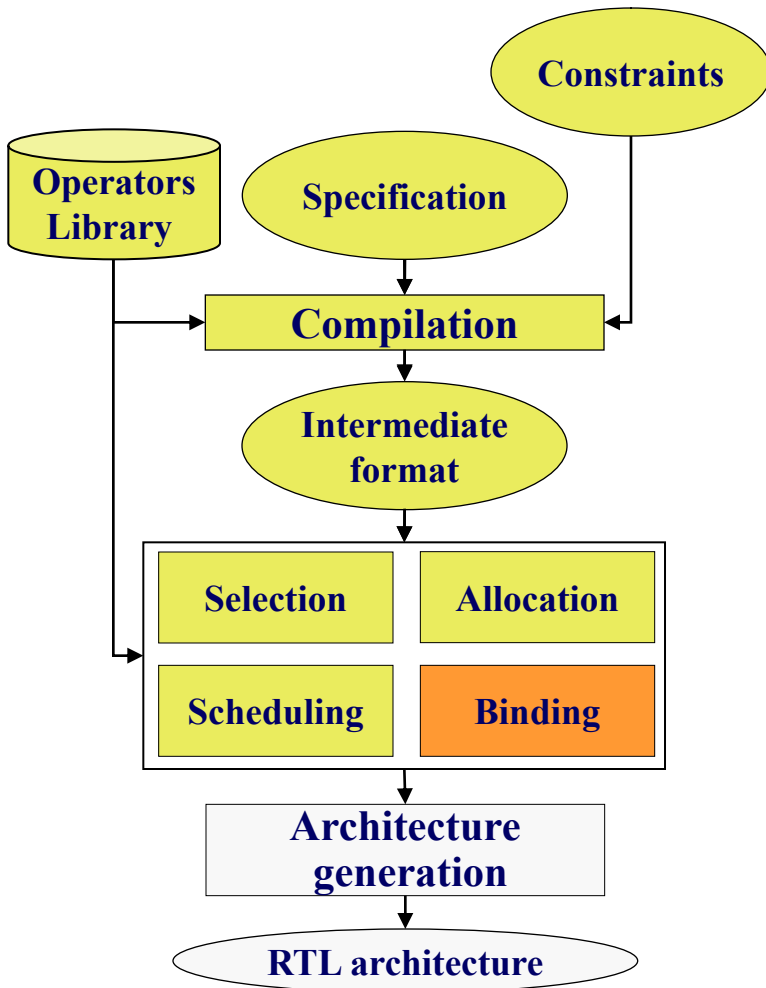
## **Binding (or Assignment)**

- Defines which operator will execute a given operation
- Defines which memory element will store a data

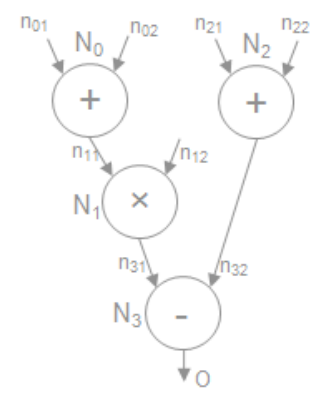
## Architecture generation

- Writes out the RTL source code in the target language e.g. VHDL

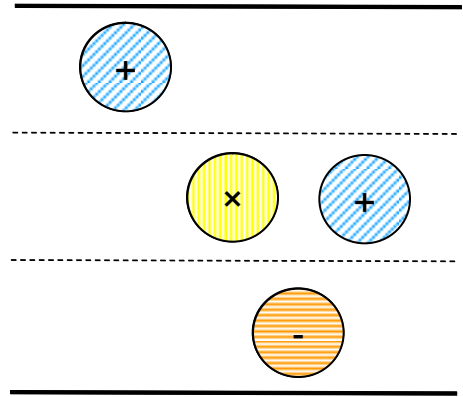
# HLS steps: binding



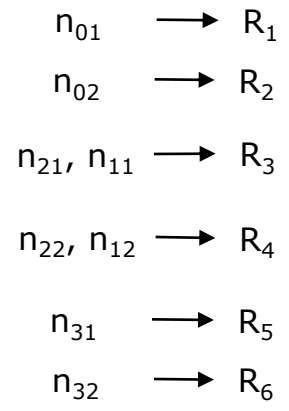
Intermediate representation



Operation binding



Data Binding



# Synthesis steps

---

## Compilation

- Generates a formal modeling of the specification

## Selection

- Chooses the architecture of the operators

## Allocation

- Defines the number of operators for each selected type

## Scheduling

- Defines the execution date of each operation

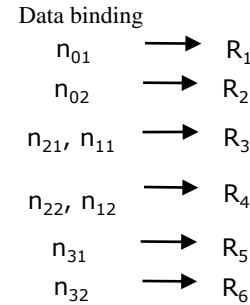
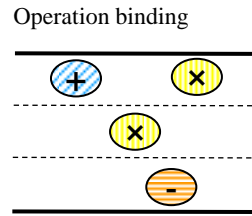
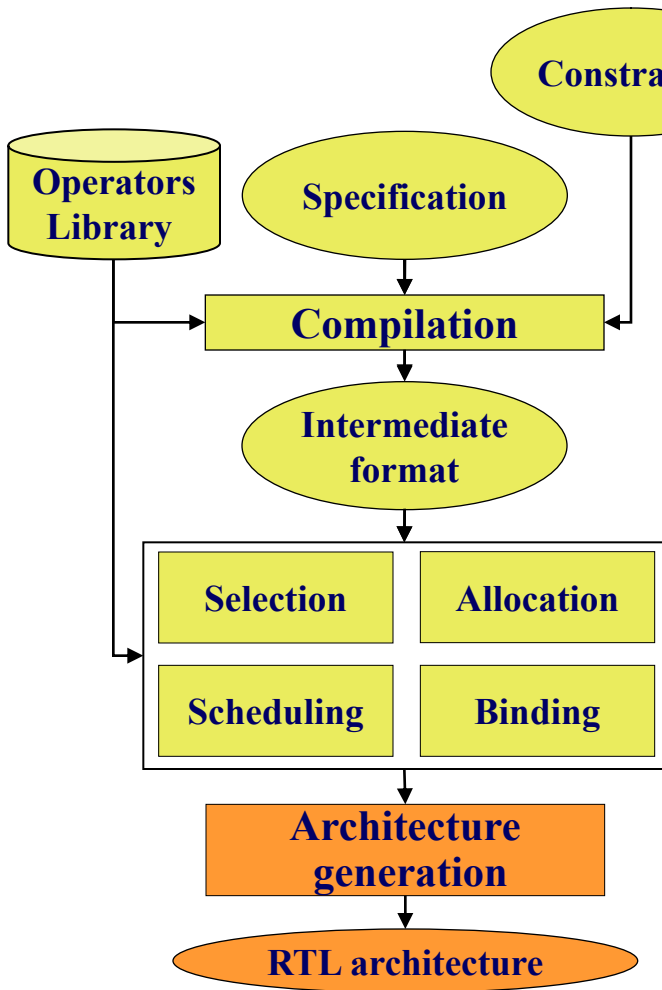
## Binding (or Assignment)

- Defines which operator will execute a given operation
- Defines which memory element will store a data

## **Architecture generation**

- Writes out the RTL source code in the target language e.g. VHDL

# HLS steps: output

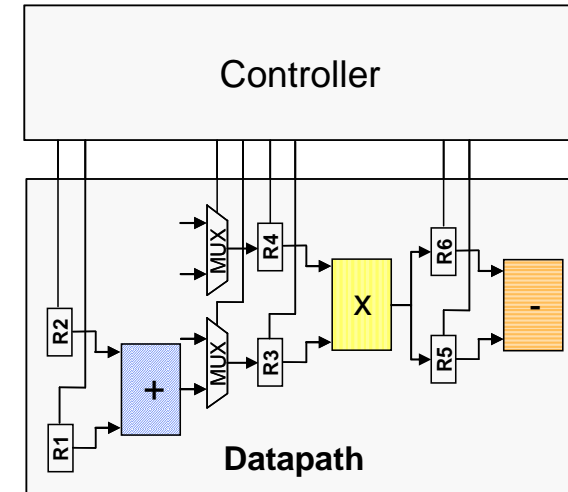


## Controller

- FSM controller
- Programmable controller

## Datapath components

- Storage components
- Functional units
- Connection components



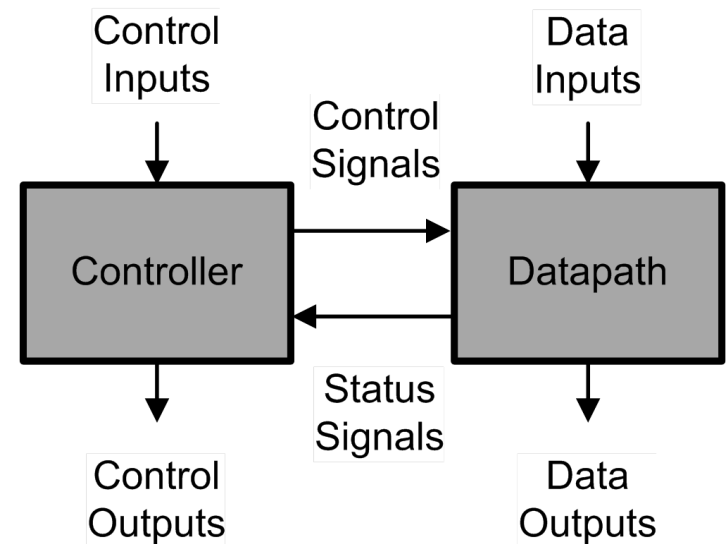
# RTL Architecture

## □ Controller

- FSM controller
- Programmable controller

## □ Datapath components

- Storage components
- Functional units
- Connection components

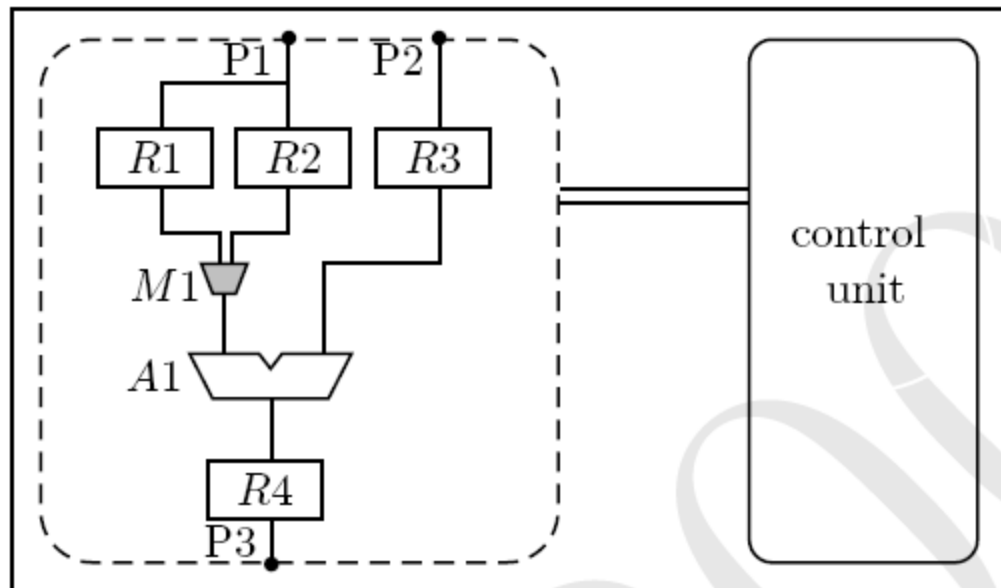


Source :  
Embedded System Design, © 2009, Gajski, Abdi, Gerstlauer, Schirner

# Example

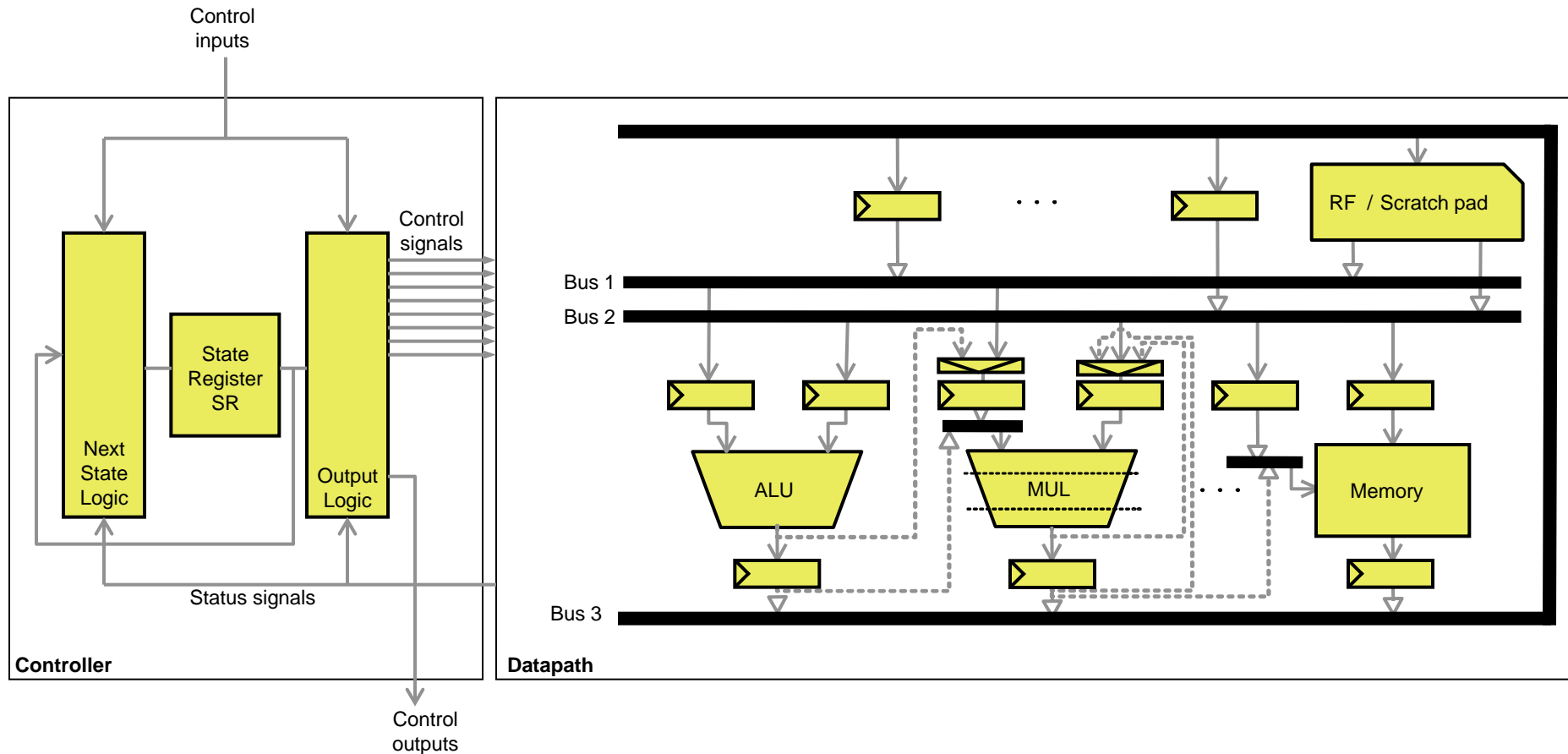
## □ This architecture performs the following operations:

- store two variables coming from the port P1 in R1 and R2
- store one variable coming from the port P2 in R3
- add the variables stored in R1 and R3 and put the result in R4
- add the variables stored in R2 and R3 and put the result in R4
- connect either R1 or R2 to A1
  - *the control unit manages this connection through M1*





# RTL architecture



Source :  
Embedded System Design, © 2009, Gajski, Abdi, Gerstlauer, Schirner

---

# **Problem examples and design flow**

# Resource constrained HLS

---

## □ Limited number of resources

- e.g.: 2 multipliers, 3 adders
- Pseudo architecture

## □ Schedule operations according to the available operators in the current control step

## □ Objectives

- Minimize the latency or maximize the throughput
  - *based on operations mobility i.e. operations urgency*

# Resource constrained HLS

---

## □ Limited number of resources

- e.g.: 2 multipliers, 3 adders
- Pseudo architecture

## □ Schedule operations according to the available operators in the current control step

## □ Objectives

- Minimize the latency or maximize the throughput
  - *based on operations mobility i.e. operations urgency*



**Allocation and then Scheduling**

# Time constrained HLS

---

## Latency constraint

- e.g. 5 clock cycles to process all the data

## Throughput constraint

- Cadency, initiation interval...
- e.g. process each 5 cycles a new set of input data

## Schedule operations by using operators as much as needed

## Objective

- Minimize the circuit area

# Time constrained HLS

---

## Latency constraint

- e.g. 5 clock cycles to process all the data

## Throughput constraint

- Cadency, initiation interval...
- e.g. process each 5 cycles a new set of input data

## Schedule operations by using operators as much as needed

## Objective

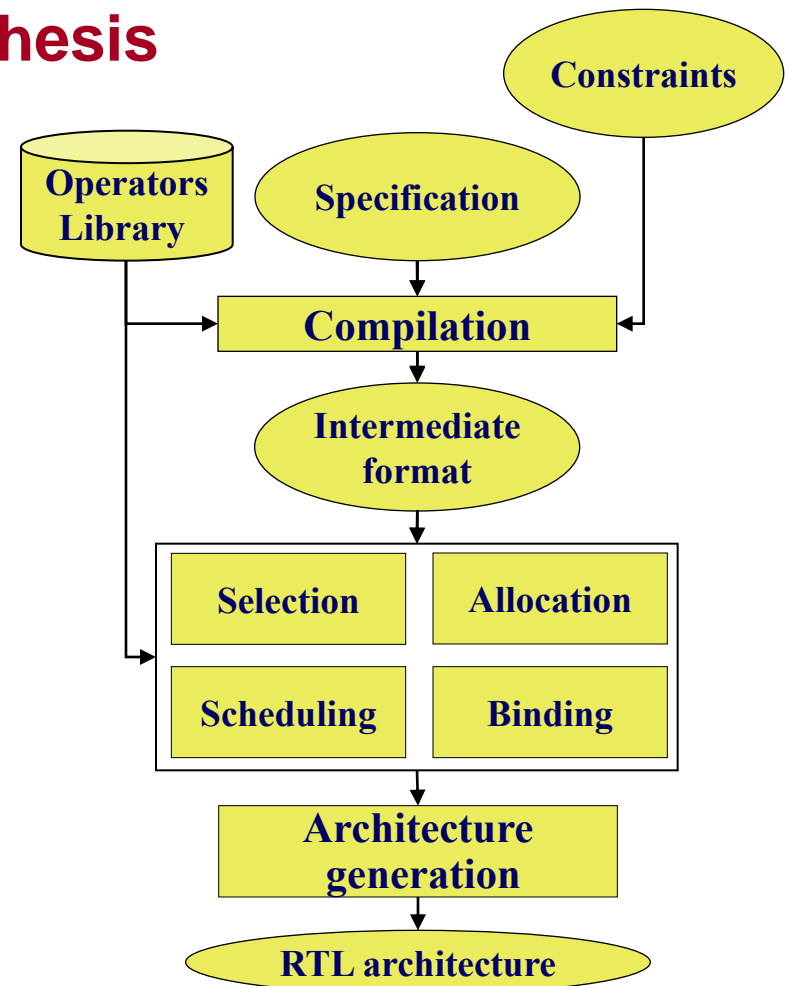
- Minimize the circuit area

⇒ **Scheduling and then Allocation**

# Design flows

## □ No unique design flow i.e. synthesis steps order

- Allocation → Scheduling
  - *resource constrained*
- Scheduling → Allocation
  - *Time constrained*
- Scheduling → Binding
- Binding → Scheduling
- Scheduling & Binding
- ...



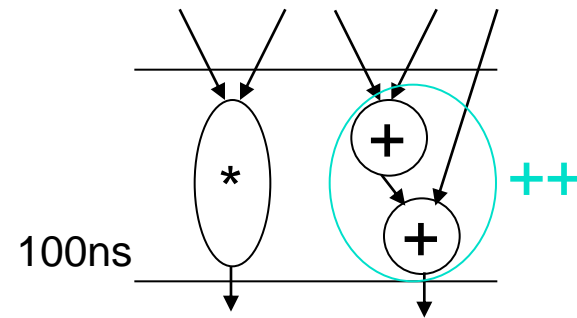
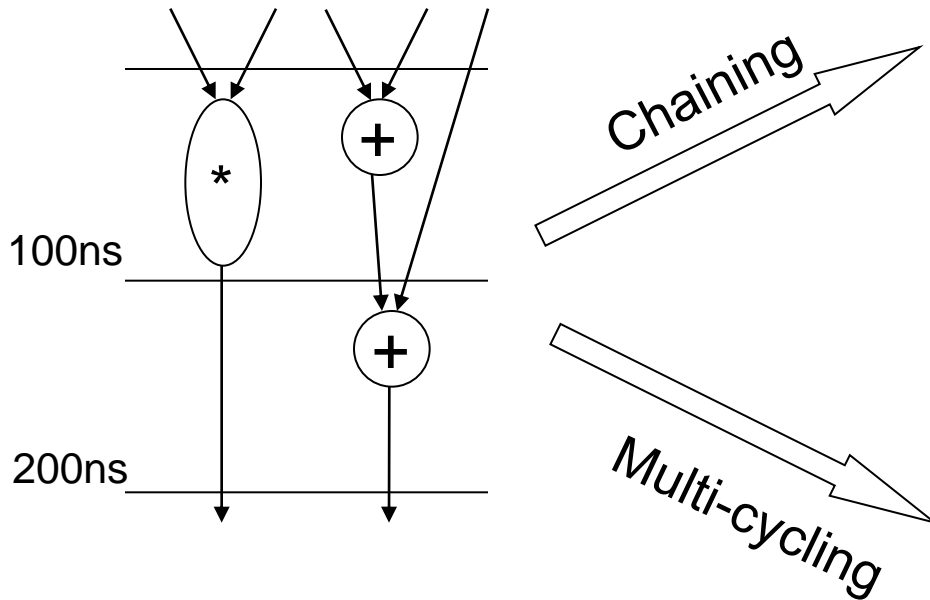
# And a lot of other problems...

---

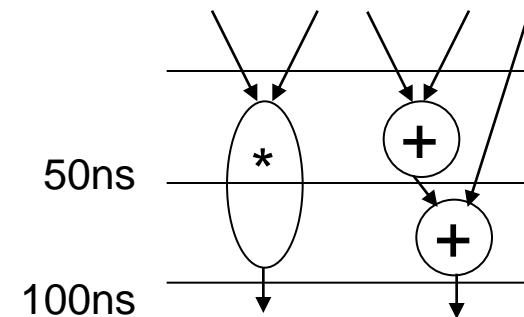
- Variable merging Storage Sharing
- Operation merging Operator sharing
- **Connection merging**
  - Bus sharing
- **Register merging**
  - Register file...
- **Chaining**
  - Several sequential operations in a cycle
- **Multi-cycling**
  - One operation takes more than one clock cycle to execute
- **Pipelining**
  - Pipelined Datapath, pipelined operator, pipelined controller
- ...



# Chaining, multi-cycling



Several sequential operations in a cycle



One operation takes more than one clock cycle to execute

# Outline

---

- Lab-STICC
- General context
- **High-Level Synthesis**
  - Brief introduction
  - “In details”
- **GAUT**
  - Overview
  - Results
- **Conclusion**
- **References**

# Synthesis steps

---

## **Compilation**

- Generates a formal modeling of the specification

## **Selection**

- Chooses the architecture of the operators

## **Allocation**

- Defines the number of operators for each selected type

## **Scheduling**

- Defines the execution date of each operation

## **Binding (or Assignment)**

- Defines which operator will execute a given operation
- Defines which memory element will store a data

## **Architecture generation**

- Writes out the RTL source code in the target language e.g. VHDL

# High-level synthesis goal

---

- **Starting from a functional description, automatically generate an RTL architecture**
  - Mathematic formula
  - Matlab/Simulink
  - C/C++/SystemC
  - ...

# Synthesizable models

---

## □ C for the synthesis:

- No pointer
  - *Statically unresolved*
  - *Arrays are allowed!*
  
- No standard function call
  - *printf, scanf, fopen, malloc...*
  
- Function calls are allowed
  - *Can be in-lined or not*
  
- Finite precision
  - *Bit accurate integers, fixed point, signed, unsigned...*
  - *Based on SystemC or Mentor Graphics data types*
    - *sc\_int, sc\_fixed*
    - *ac\_int, ac\_fixed*

# Synthesizable models

## □ C for the synthesis:

### ■ Finite precision

□ *bit accurate integer, fixed point, signed, unsigned...*

S/W C: Overflow checks everywhere.

```
unsigned int x, y, z, cy;

z = x + y;
if (0xFF..FF - x >= y)
    cy = 0; // bit 32
else
    cy = 1; // bit 32
```

H/W C: Check unnecessary.

```
sc_uint<32> x, y;
sc_uint<33> z;

z = x + y;
```

# Purely functional Example #1: a simple C code

```
#define N 16

int main(int data_in, int *data_out)
{ static const int Coeffs [N] = {98,-39,-327,439,950,-2097,-1674,9883,9883,-1674,-2097,950,439,-327,-39,98};

  int Values[N];
  int temp;
  int sample,i,j;

  sample = data_in;
  temp = sample * Coeffs[N-1];

  for(i = 1; i<=(N-1); i++){
    temp += Values[i] * Coeffs[N-i-1];
  }

  for(j=(N-1); j>=2; j-=1 ){
    Values[j] = Values[j-1];
  }

  Values[1] = sample;
  *data_out=temp;

  return 0;
}
```

# Purely functional example #2: bit accurate C++ code

```
#include "ac_fixed.h" // From Mentor Graphics
#define PORT_SIZE ac_fixed<16, 12, true, AC_RND, AC_SAT>
// 16 bits, 12 bits after the point, quantization = rounding, overflow = saturation
#define N 16
int main(PORT_SIZE data_in, PORT_SIZE &data_out)
{
    static const PORT_SIZE Coeffs [N]={1.1, 1.5, 1.0, 1.0, 1.7, 1.8, 1.2, 1.0, 1.6, 1.0, 1.5, 1.1, 1.9, 1.3, 1.4, 1.7};
    PORT_SIZE Values[N];
    PORT_SIZE temp;
    PORT_SIZE sample;

    sample= data_in;
    temp = sample * Coeffs[N-1];
    for(int i = 1; i<=(N-1); i++){
        temp = Values [i] * Coeffs[N-i-1] + temp;
    }

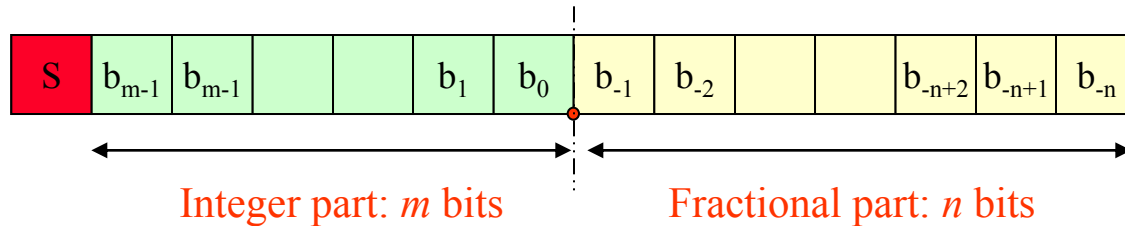
    for(int j=(N-1); j>=2; j-=1 ){
        Values[j] = Values [j-1];
    }
    Values[1] = sample;

    data_out=temp;
    return 0;
}
```



# Fixed-point

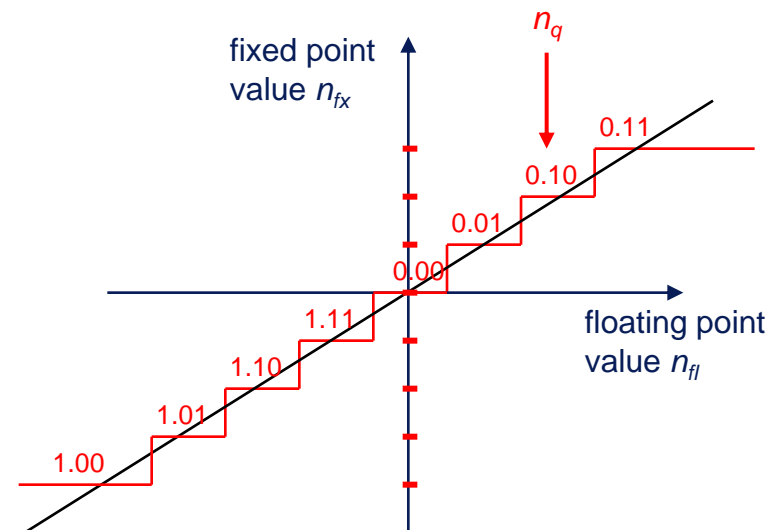
## □ Fixed point:



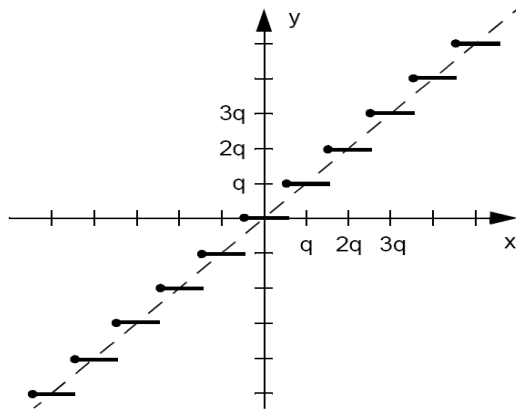
$b = m + n + 1$  bits  
format:  $(b, m, n)$

## □ Ac\_fixed<W,I,S,Q,O>

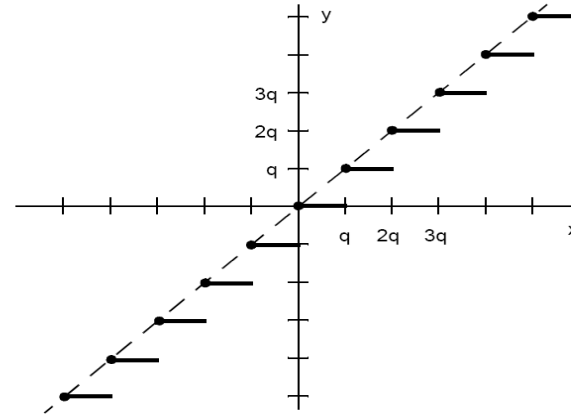
- W : word size ( $m+n+1$ )
- I : integer part size ( $m+1$ )
- S : signed or unsigned
- Q : rounding mode
- O : overflow mode
- Equivalent to SystemC data type `sc_fixed`



# Fixed point: rounding mode

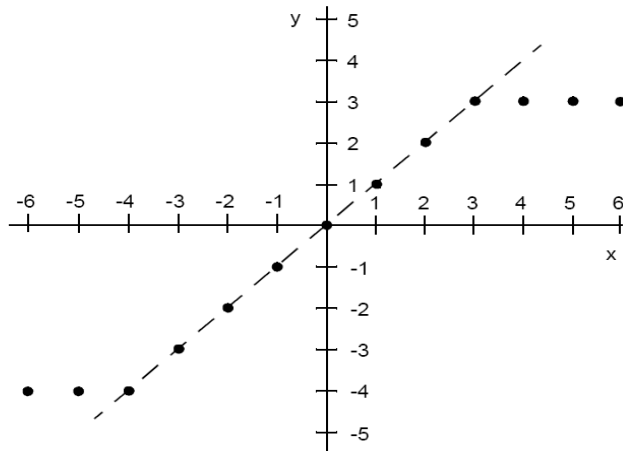


SC\_RND

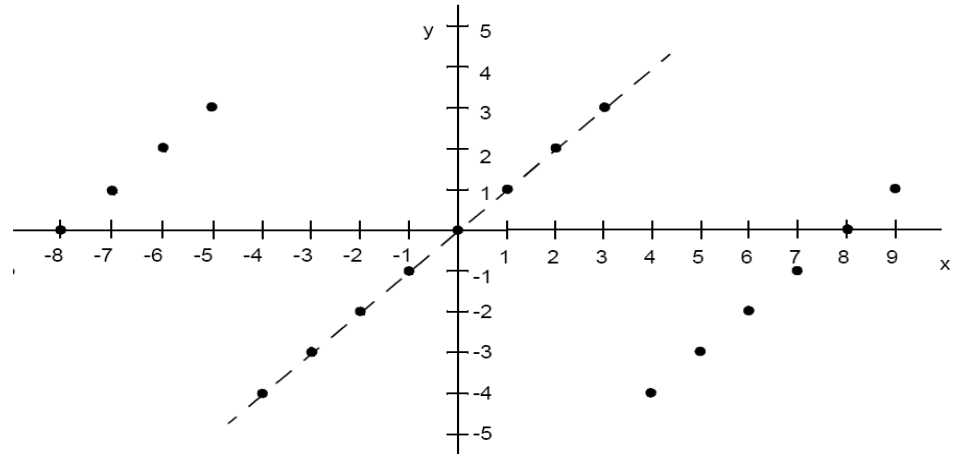


SC\_TRN

# Fixed point: overflow mode



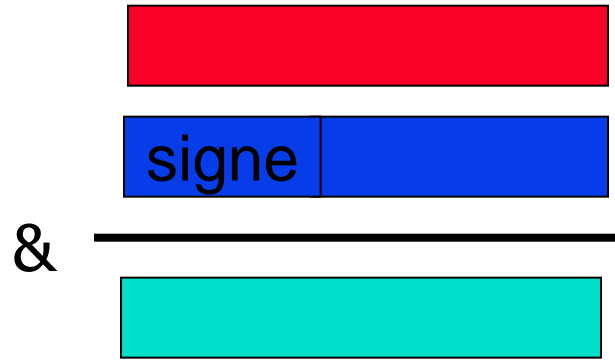
SC\_SAT



SC\_WRAP

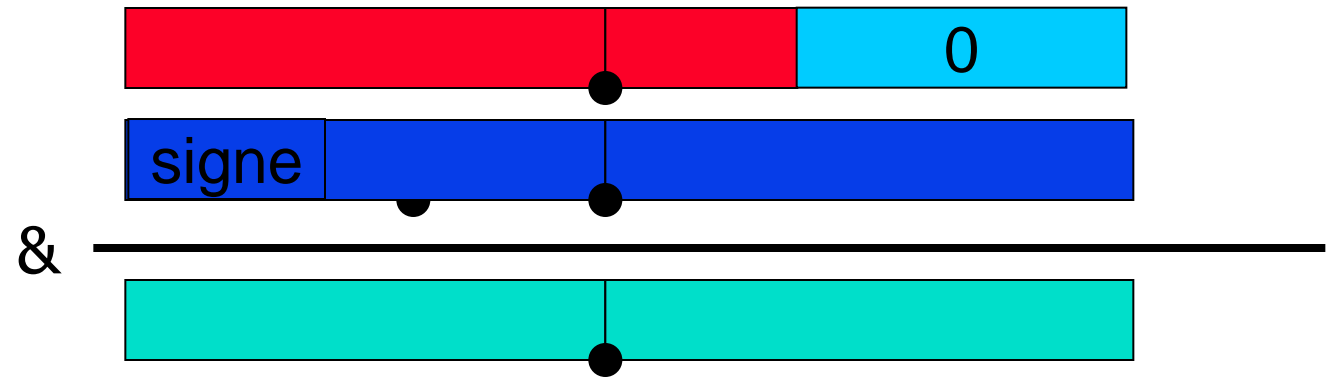
# Bit accurate operation

- Sign extension before the computation



## Fixed point operation

- Alignment before the computation



# High-level synthesis goal

---

- **Starting from a functional description, automatically generate an RTL architecture**
  - Algorithmic description
    - *no timing notion in the source code*
  - Behavioral description
    - *Notion of step / local timing constraints in the source code*
      - by using the wait statements of SystemC for example
  - The description can be
    - *“RTL oriented”*
    - *“Function oriented”*

# High-level synthesis

---

## □ Starting from a functional description, automatically generate an RTL architecture

### ■ Algorithmic description

- *No timing notion in the source code*
- *Mainly oriented toward data dominated application*
  - Highly processing algorithm like filters...
- *Initial description can be*
  - “RTL oriented”
  - “Function oriented”

### ■ Behavioral description

- *Notion of step / local timing constraints in the source code*
  - by using the wait statements of SystemC for example
- *Can be used for both data and control dominated application*
  - Interface controller, DMA...
  - Filters...

# High-level synthesis

---

## □ Starting from a functional description, automatically generate an RTL architecture

### ■ Algorithmic description

- *No timing notion in the source code*
- *Mainly oriented toward data dominated application*
  - Highly processing algorithm like filters...
- *Initial description can be*
  - “RTL oriented”
  - “Function oriented”

### ■ Behavioral description

- *Notion of step / local timing constraints in the source code*
  - by using the wait statements of SystemC for example
- *Can be used for both data and control dominated application*
  - Interface controller, DMA...
  - Filters...

# Behavioral description

## ■ Behavioral description

□ Notion of step / local timing constraints in the source code

■ by using the wait statements of SystemC for example

```
...
void addmul() {
    {
        sc_signal<sc_uint<32> > tmp1;
        tmp1 = 0;
        result = 0;
        wait();
    }
    while (1) {
        {
            tmp1 = b * c;
            wait();
        }
        {
            result = a + tmp1;
            wait();
        }
    }
}
...
```

Reset state

First state

Second state

Cycle-by-cycle FSM with reset state

# Function v.s. RTL description

```
01:  int OnesCounter(int Data) {
02:    int Ocount = 0;
03:    int Temp, Mask = 1;
04:    while (Data > 0) {
05:      Temp = Data & Mask;
06:      Ocount = Data + Temp;
07:      Data >>= 1;
08:    }
09:    return Ocount;
10: }
```

Function-based C code

```
01:  while(1) {
02:    while (Start == 0);
03:    Done = 0;
04:    Data = Input;
05:    Ocount = 0;
06:    Mask = 1;
07:    while (Data>0) {
08:      Temp = Data & Mask;
09:      Ocount = Ocount + Temp;
10:      Data >>= 1;
11:    }
12:    Output = Ocount;
13:    Done = 1;
14: }
```

RTL-based C code

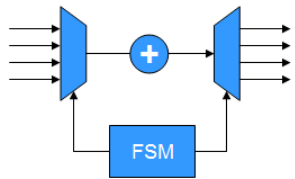


# High-level transformations

## □ Loops

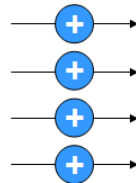
- loop unrolling
  - *None, partially, completely*
- Loop merging
- Loop tiling
- ...

```
for (i = 0; i<4; i++)  
{  
  r[i] = a[i] + b[i];  
}
```



No Unrolling  
1 Adder shared for 4 additions  
Latency = 4 cycles

```
r[0] = a[0] + b[0];  
r[1] = a[1] + b[1];  
r[2] = a[2] + b[2];  
r[3] = a[3] + b[3];
```



Unrolling = 4 (Full)  
4 Adders in parallel  
Latency = 1 cycle

```
for (i = 0; i<32; i++)  
{  
  a[i] = b[i] * c[i];  
}  
for (i = 0; i<16; i++)  
{  
  z[i] = a[i] + x[i];  
}
```

No Merging  
Loops execute sequentially  
Latency = 48 cycles

```
for (i = 0; i<32; i++)  
{  
  atmp = b[i] * c[i];  
  
  if (i<16)  
    z[i] = atmp + x[i];  
}
```

Merging Enabled  
Loops execute in parallel  
Latency = 32 cycles

# High-level transformations

---

## □ **Loops**

- Loop pipelining,
- loop unrolling
  - *None, partially, completely*
- Loop merging
- Loop tiling
- ...

## □ **Arrays**

- Arrays can be mapped on memory banks
- Arrays can be synthesized as registers
- Constant arrays can be synthesized as logic
- ...

## □ **Functions**

- Function calls can be in-lined
- Function is synthesized as an operator
  - *Sequential, pipelined, functional unit...*
- Single function instantiation
- ...

# Compilation

---

## □ Optimization

- Constant folding
- Dead code elimination
- Common sub-expression elimination
  - *Eliminate redundant operations*
- ...

## □ Formal model

- Inputs, outputs, and operations of the algorithm are identified
- Data and/or control dependencies are determined
- Intermediate representation is generated

# Control-Flow Graph CFG

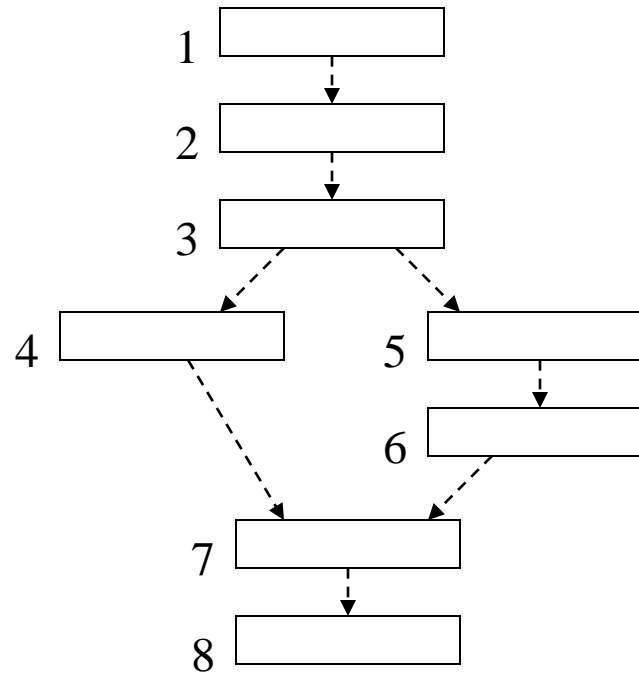
---

- **Exhibits operation sequences**
  - Through control dependencies
  
- **The sequence of operations comes directly from the source code**
  - The sequence is kept unchanged
    - *This limits the parallelism which should be limited if this representation is used to model control-oriented application*

# Example

```
1: t = a+b;  
2: u = a'-b';  
3: if (a<b)  
4:   v = t+c;  
   else  
   {  
5:   w = u+c';  
6:   v = w-d;  
   }  
7: x = v+e;  
8: y = v-e;
```

Source code



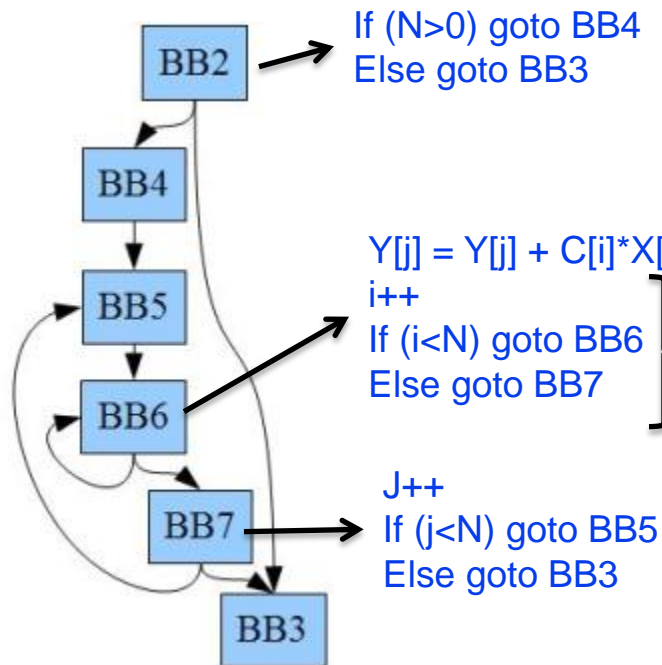
Graphical  
representation

# Example (2)

Filter.c



Compiler



If (N>0) goto BB4  
Else goto BB3

Y[j] = Y[j] + C[i]\*X[N-1-i]  
i++  
If (i<N) goto BB6  
Else goto BB7

J++  
If (j<N) goto BB5  
Else goto BB3

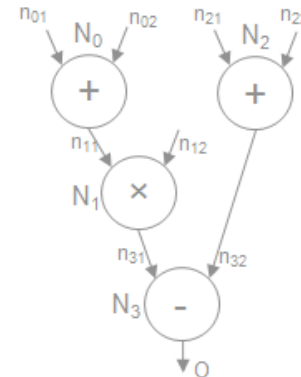
```
Void Filtre (int N, int C[N], int X[N], int Y[N]){  
    int i,j;  
    for (j =0; j<N; j++){  
        Y[j] = 0;  
        for (i=0; i<N, i++){  
            Y[j]= Y[j] + C[i]*X[N-1-i];  
        }  
    }  
}
```

# Data Flow Graph DFG

- Exhibits the parallelism between operations
  - Through data dependencies
    - *Variable node, operation node*

$$O = ((n_{01} + n_{02}) * n_{12}) - (n_{21} + n_{22})$$

Intermediate representation




# CDFG => DFG

---

## □ Exhibits the parallelism between operations

- Through data dependencies
  - *Variable node, operation node*

## □ Loops are completely unrolled

for  $i : 0 \rightarrow 2$   
 $c[i] = a[i] + b[i]$    $c[0] = a[0] + b[0]$   
 $c[1] = a[1] + b[1]$   
 $c[2] = a[2] + b[2]$

## □ Conditional assignments are transformed

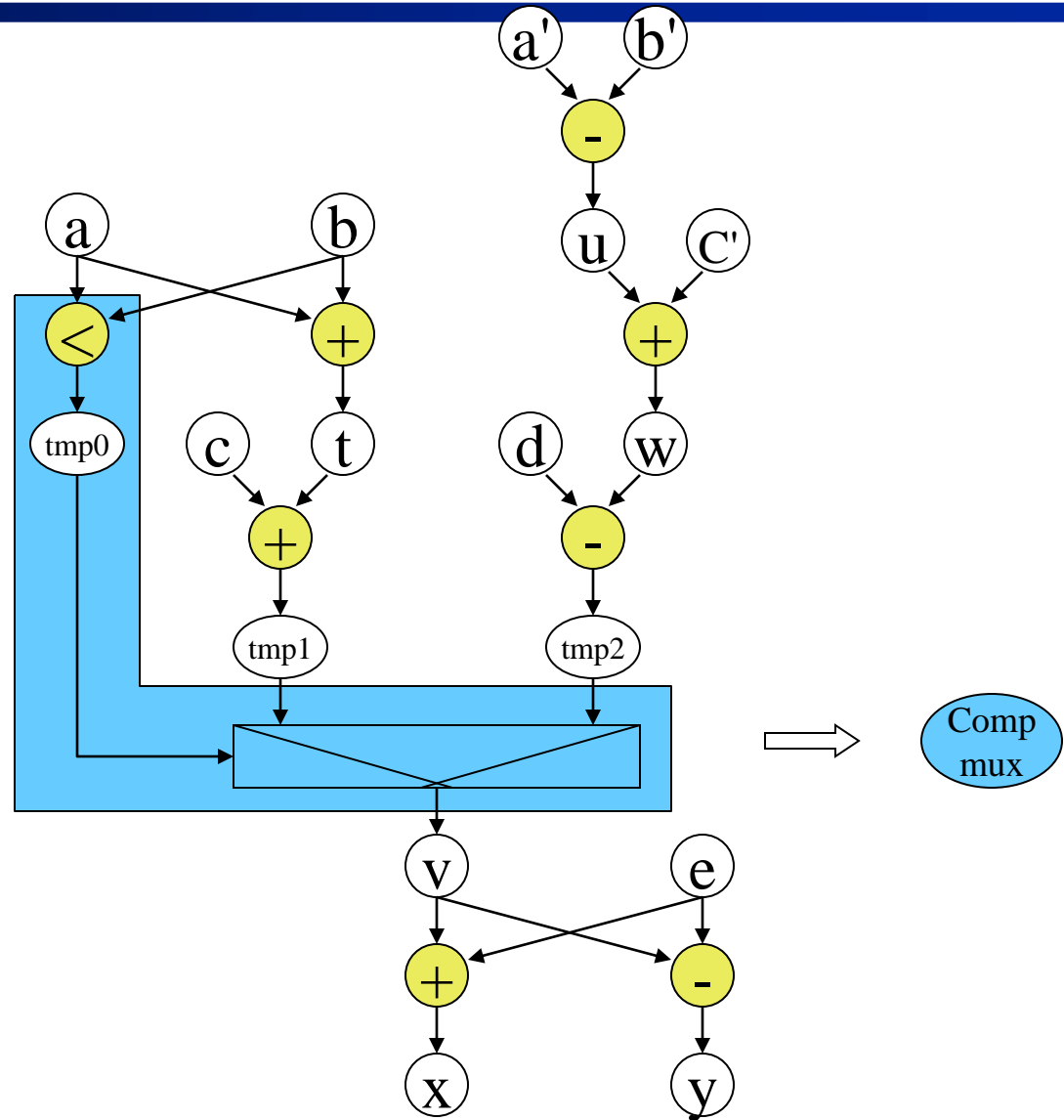
- i.e. *if/switch* constructs, are resolved by creating multiplexed values



# Example

```
1: t = a+b;
2: u = a'-b';
3: if (a<b)
4:   v = t+c;
  else
  {
5:   w = u+c';
6:   v = w-d;
  }
7: x = v+e;
8: y = v-e;
```

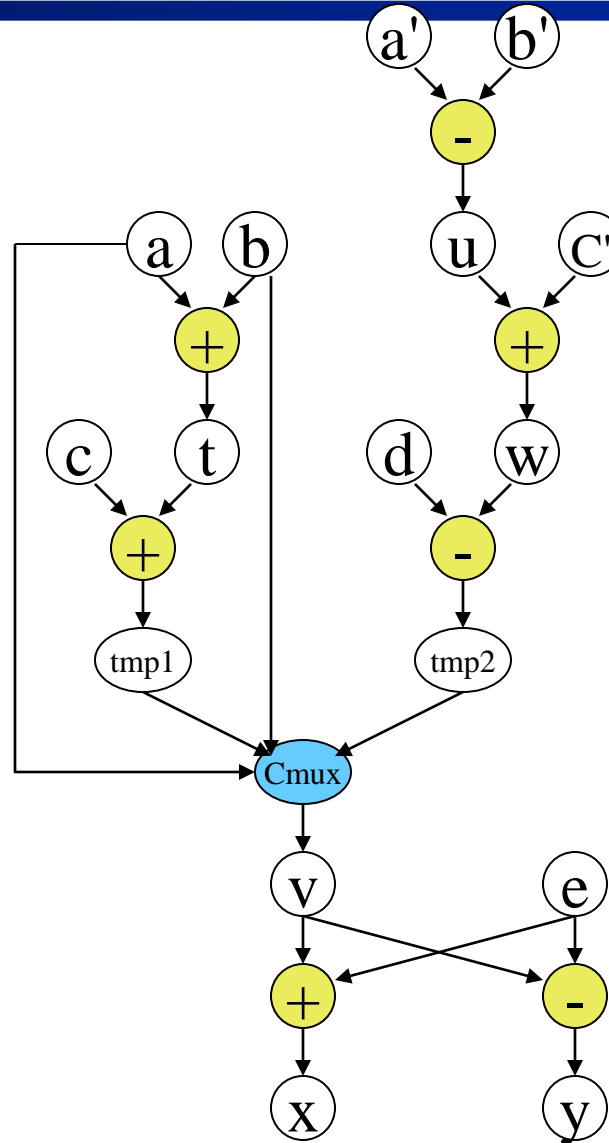
Source code



# Example

```
1: t = a+b;
2: u = a'-b';
3: if (a<b)
4:   v = t+c;
   else
   {
5:   w = u+c';
6:   v = w+d;
   }
7: x = v+e;
8: y = v-e;
```

Source code



# Data Flow Graph DFG

---

## □ Scheduling

- Resource constrained
  - *Latency minimization*
    - *List-Scheduling...*
  - *Throughput maximization*
    - *Modulo scheduling (IMS, SMS...)*
- Time constrained
  - *Resource minimization*
    - *Force-directed scheduling, ILP...*

## □ Linear FSM controller

- Worst execution time for the conditional assignments

# Synthesis steps

---

## Compilation

- Generates a formal modeling of the specification

## Selection

- Chooses the architecture of the operators

## Allocation

- Defines the number of operators for each selected type

## Scheduling

- Defines the execution date of each operation

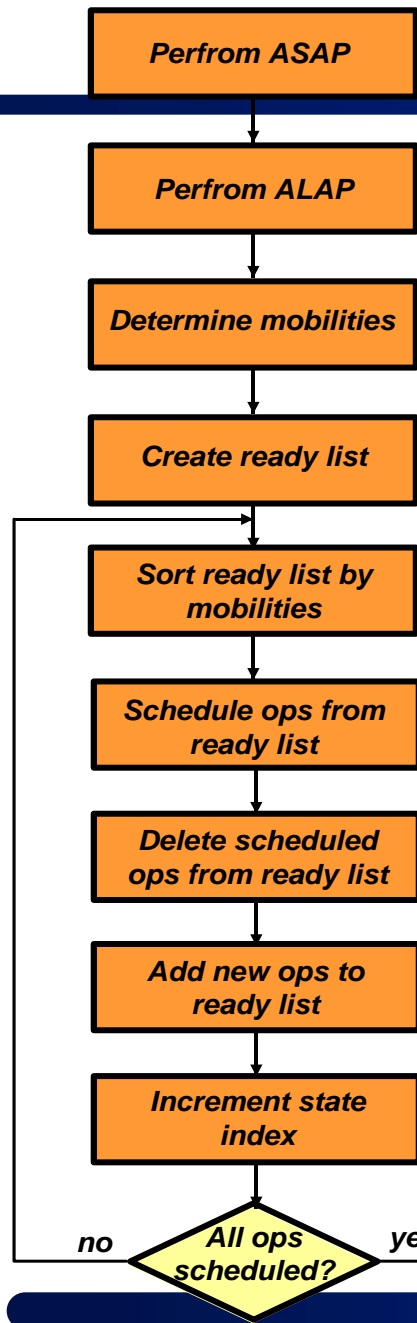
## Binding (or Assignment)

- Defines which operator will execute a given operation
- Defines which memory element will store a data

## Architecture generation

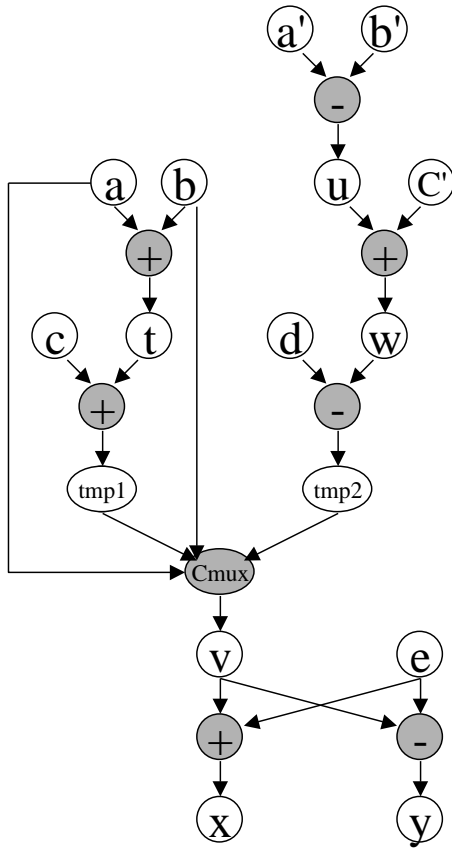
- Writes out the RTL source code in the target language e.g. VHDL

# Resource constrained scheduling

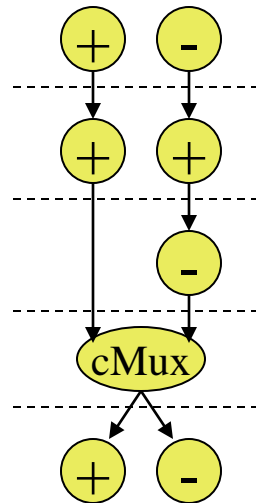


RC algorithm

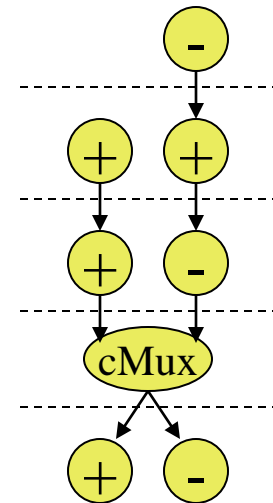
# List-scheduling



ASAP



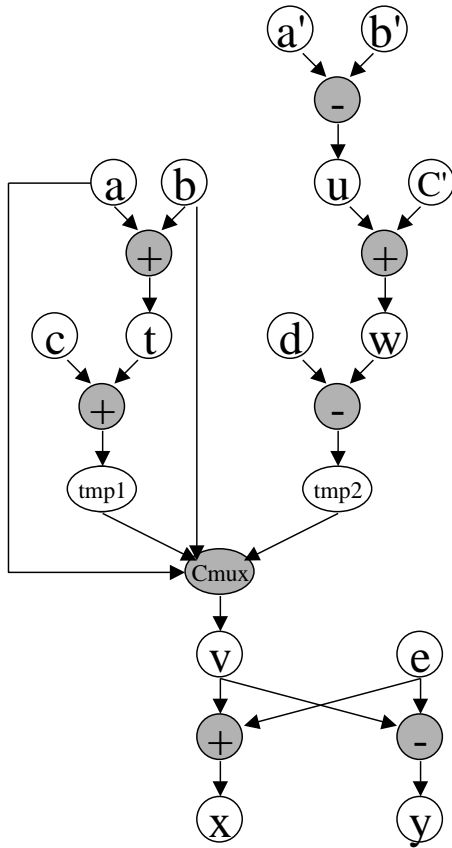
ALAP



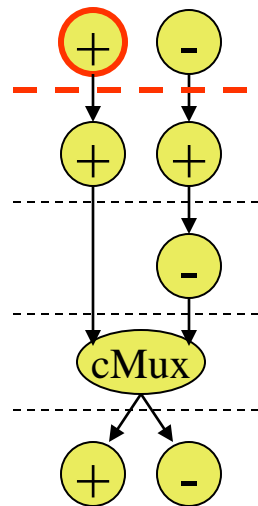
## Constraints

- 1 adder (1 cycle)
- 1 subtractor (1 cycle)
- 1 comparing component (1 cycle)
- No chaining

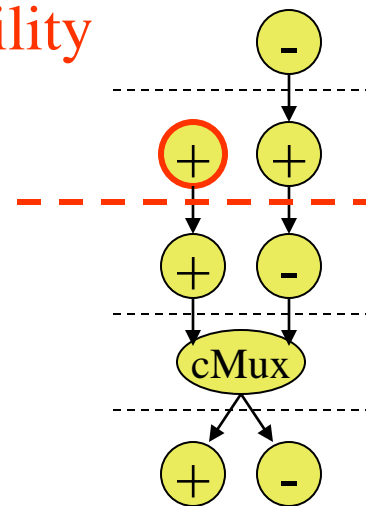
# List-scheduling



ASAP



ALAP

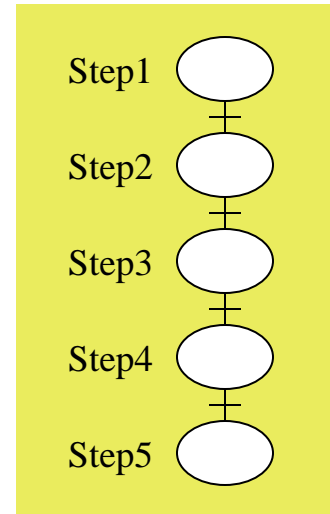
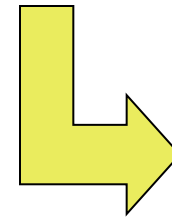
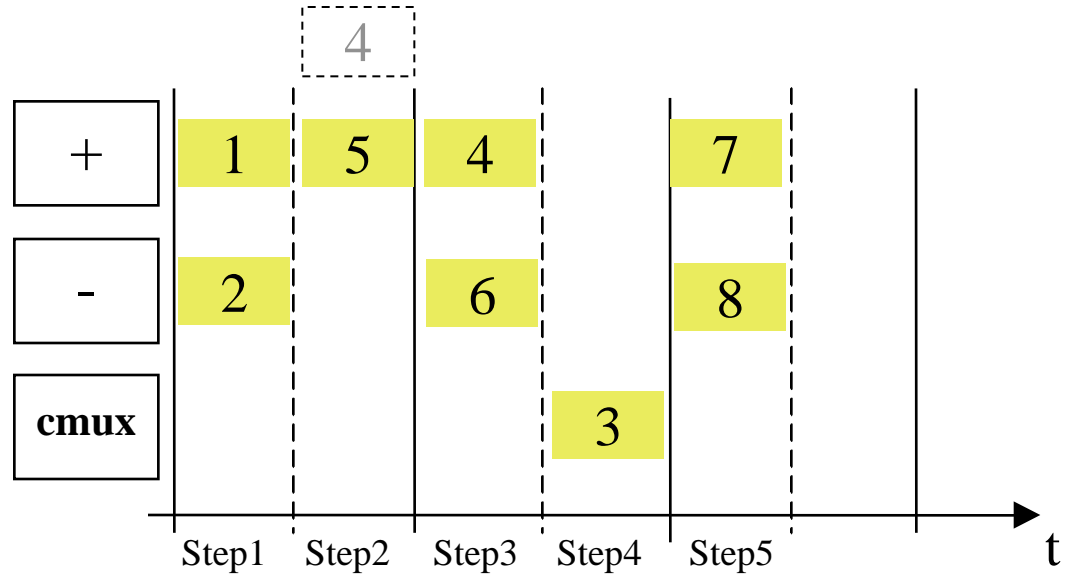
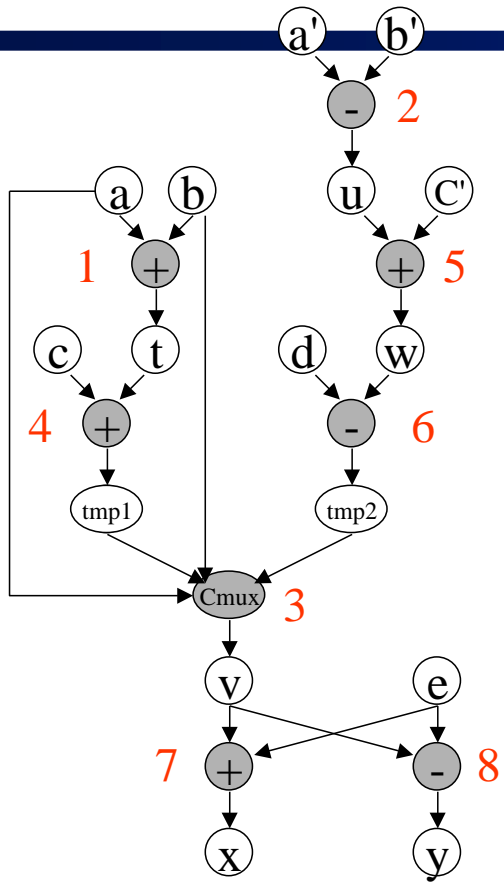


Priority = 1/Mobility

## Constraints

- 1 adder (1 cycle)
- 1 subtractor (1 cycle)
- 1 comparing component (1 cycle)
- No chaining

# List-scheduling



## Constraints

- 1 adder (1 cycle)
- 1 subtractor (1 cycle)
- 1 comparing component (1 cycle)
- No chaining



# List-based scheduling

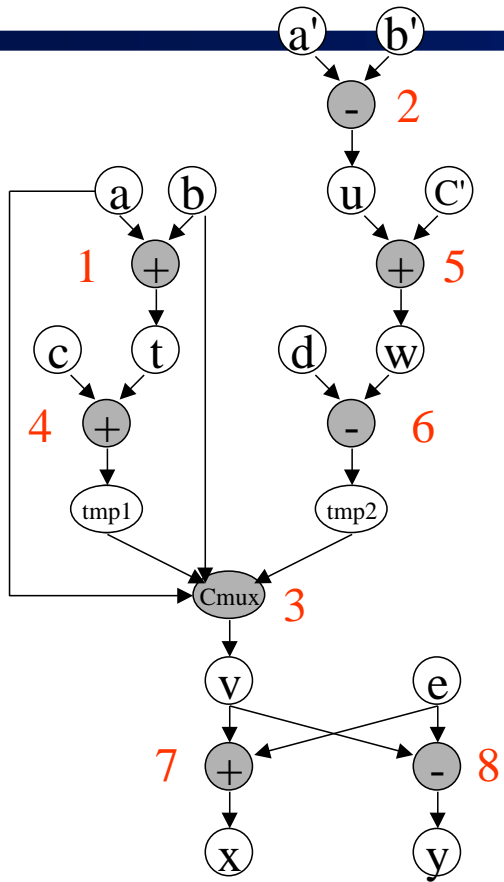
## □ Scheduling under throughput constraint (cadency)

- First operator allocation that a priori support the required parallelism
  - *In many HLS approach, an initial resource allocation is performed and subsequently modified during scheduling and/or binding => it is a lower bound*
- The average parallelism is calculated separately for each *type* of operation of the DGF

$$avr\_opr(type) = \left[ \frac{nb\_ops(type)}{\left[ \frac{II}{T(opr)} \right]} \right]$$

With  $II$  the Initiation Interval (cadency)  
 $nb\_ops(type)$  the number of operators of type  $type$   
 $T(opr)$  the propagation time of the operator (in cycles)

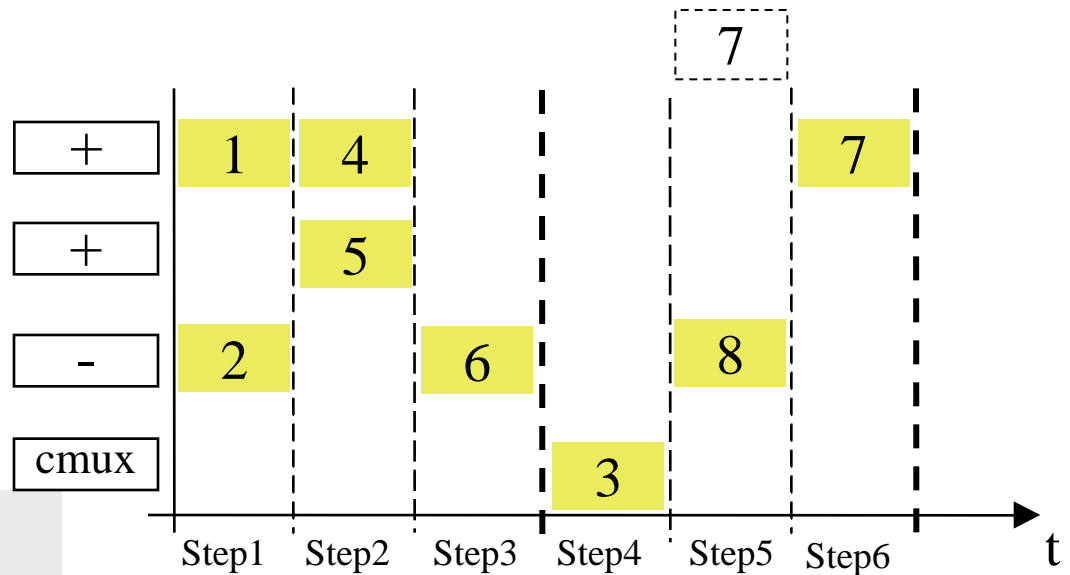
# List-based scheduling



$$\#adder = \lceil 4 * 1/3 \rceil = 2$$

$$\#sub = \lceil 3 * 1/3 \rceil = 1$$

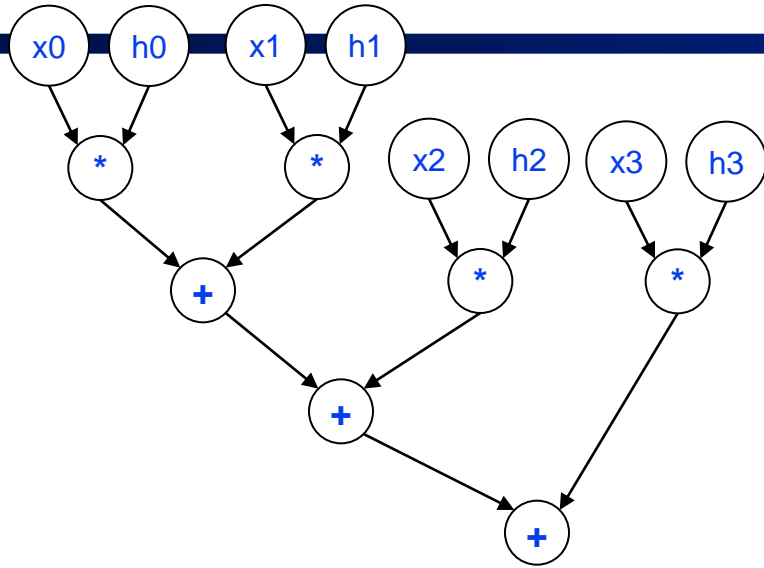
$$\#cmux = \lceil 1 * 1/3 \rceil = 1$$



## Constraint

- Throughput : one iteration each 3 cycles

# Impact on the memory



Input : X[N]

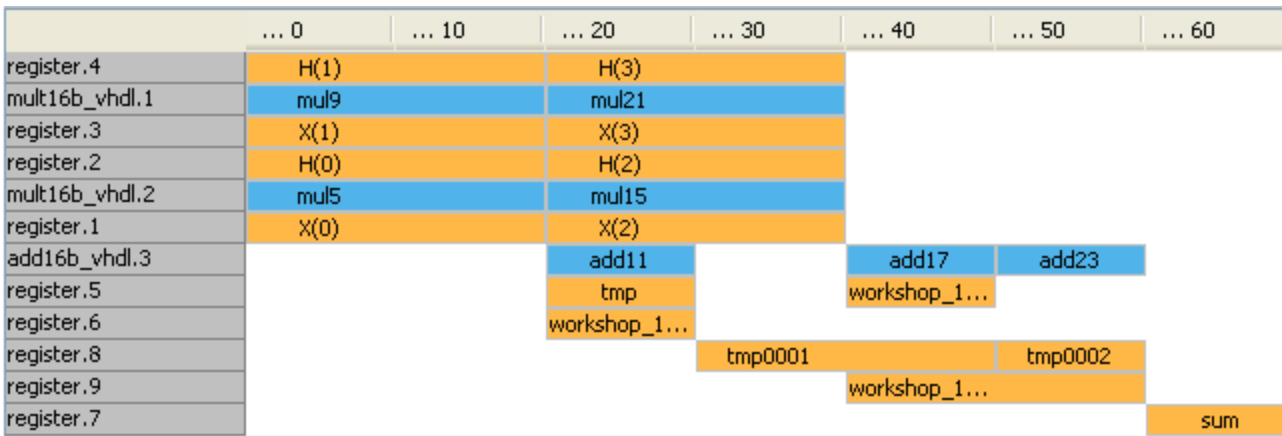
Constant : H[N] // in memory

Without memory constraints

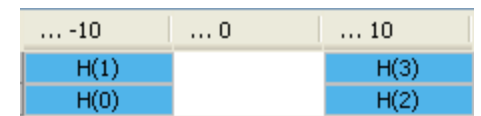
Iteration\_period = 60ns

Nb\_opr(\*) = 2

Nb\_opr(+) = 1

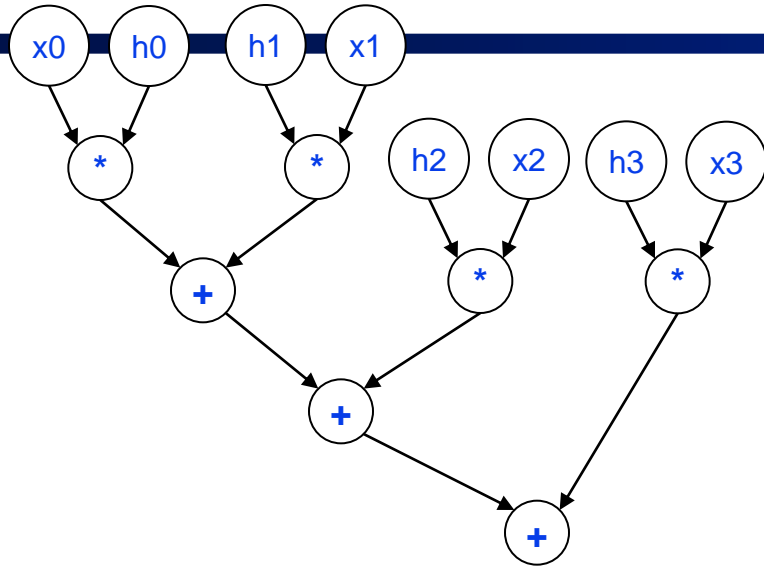


=> 2 memory banks  
Latency(arch) = 60 ns



The memory mapping has to be done by the user

# Memory constraints



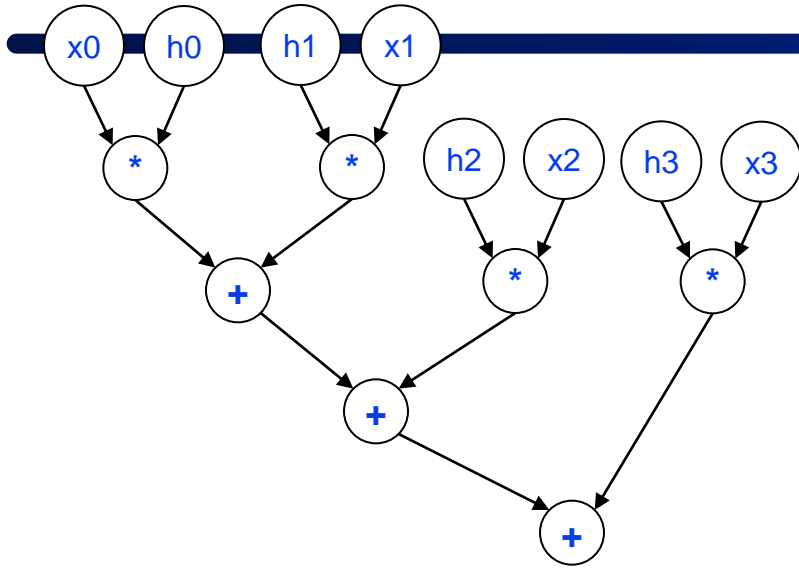
With memory constraints  
 => 1 memory bank  
 Iteration\_period = 100ns  
 Nb\_opr(\*) = 1  
 Nb\_opr(+) = 1



Latency(arch) = 90 ns

The memory access is the bottleneck !

# Impact on the I/O interface



Input :  $X[N]$

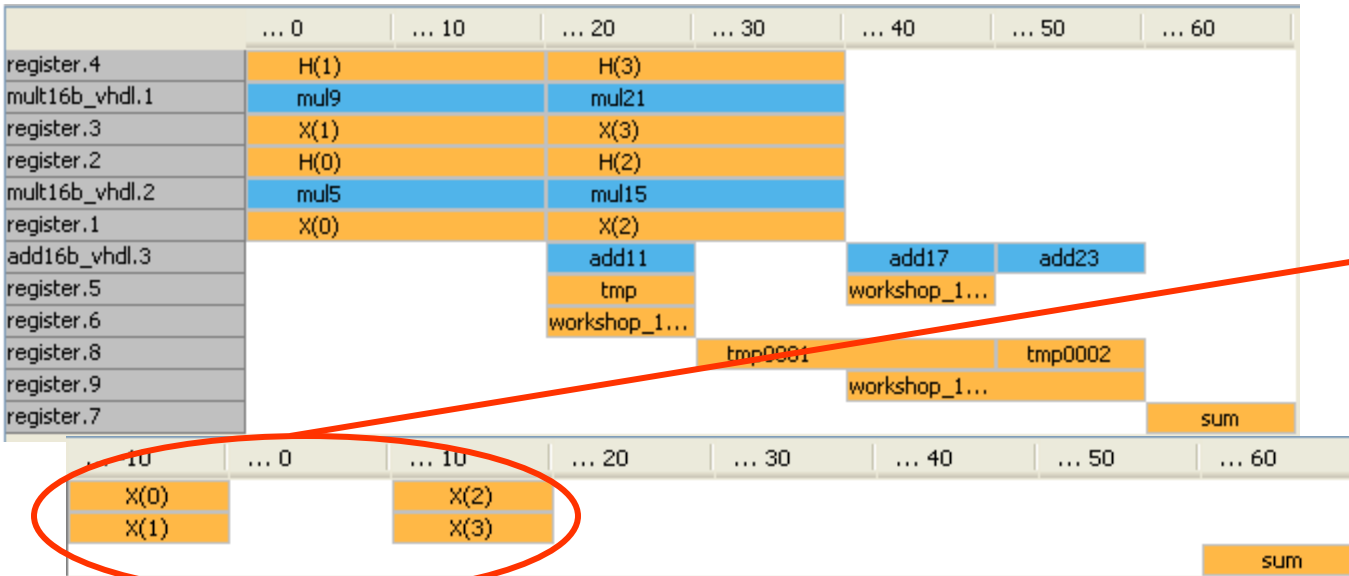
Constant :  $H[N]$  // in memory

Without I/O constraints

Iteration\_period = 60ns

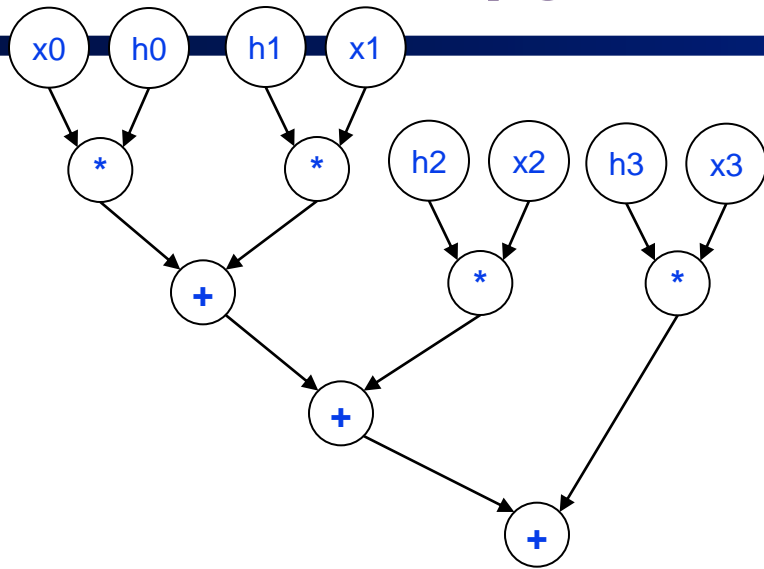
Nb\_opr(\*) = 2

Nb\_opr(+) = 1



=> 2 input ports  
Latency(arch) = 60 ns

# I/O timing constraints



With I/O constraints

4 input data in parallel

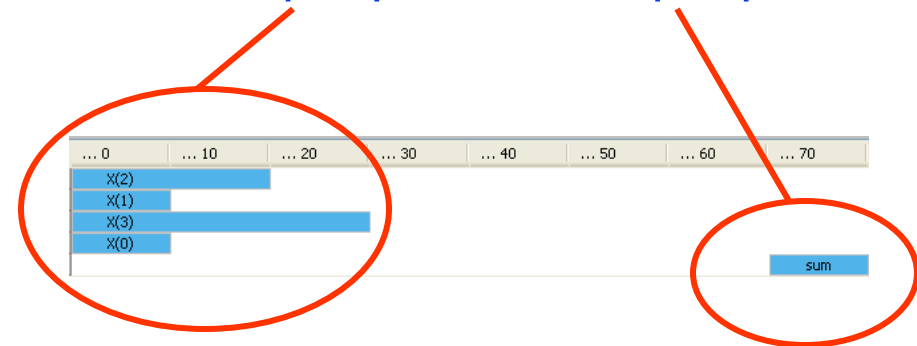
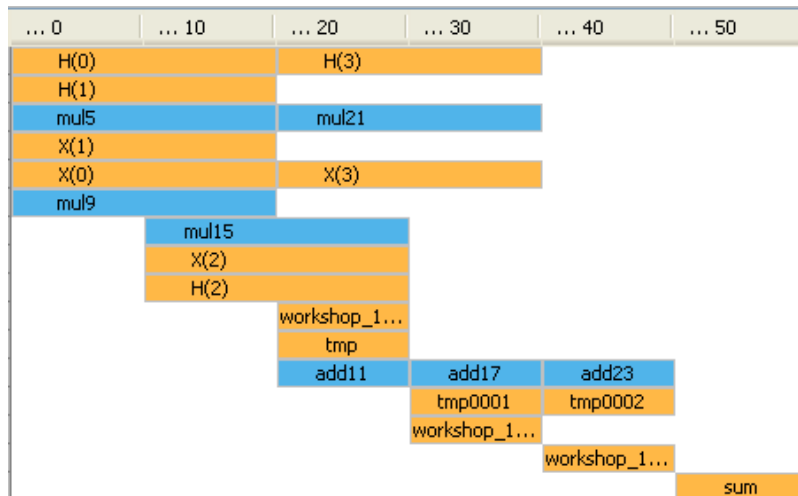
Latency = 50ns

Iteration\_period = 60ns

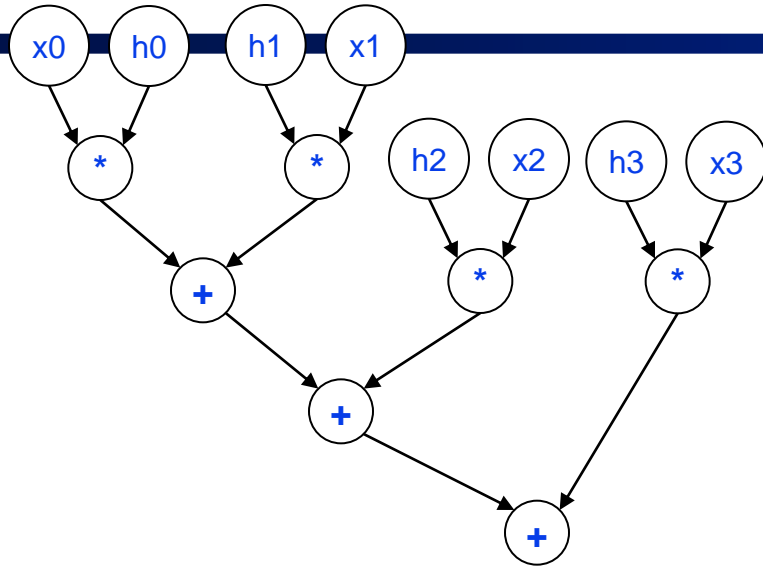
Nb\_opr(\*) = 3 (and not 4)

Nb\_opr(+) = 1

=> 4 input ports / 1 output port



# I/O timing constraints



With I/O timing constraints

1 data per 4 cycles

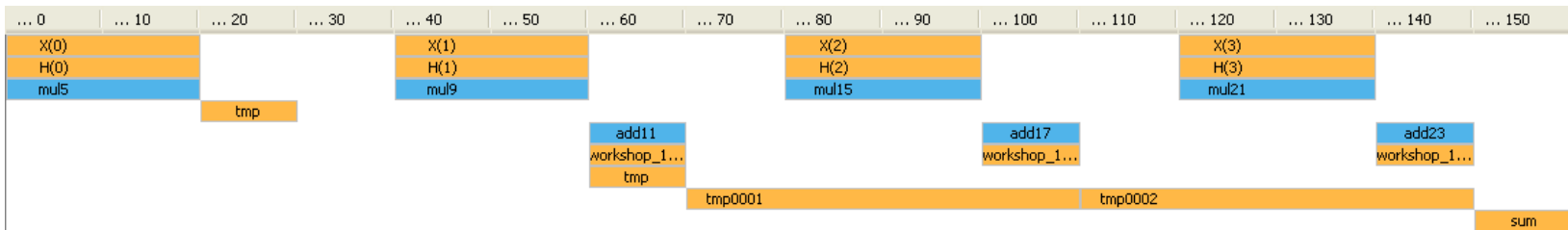
Latency (arch) = 150 ns

Iteration\_period = 170ns

Nb\_opr(\*) = 1

Nb\_opr(+) = 1

=> 1 input port / 1 output port



# Synthesis steps

---

## Compilation

- Generates a formal modeling of the specification

## Selection

- Chooses the architecture of the operators

## Allocation

- Defines the number of operators for each selected type

## Scheduling

- Defines the execution date of each operation

## **Binding (or Assignment)**

- Defines which operator will execute a given operation
- Defines which memory element will store a data

## Architecture generation

- Writes out the RTL source code in the target language e.g. VHDL



# Specification

---

```
Void example (int a, int b, int g, int c, int d, int h)
```

```
{
```

```
    int e,f;
```

```
    e = a+b;
```

```
    g = a+e;
```

```
    f = c+d;
```

```
    h = f+d;
```

```
}
```

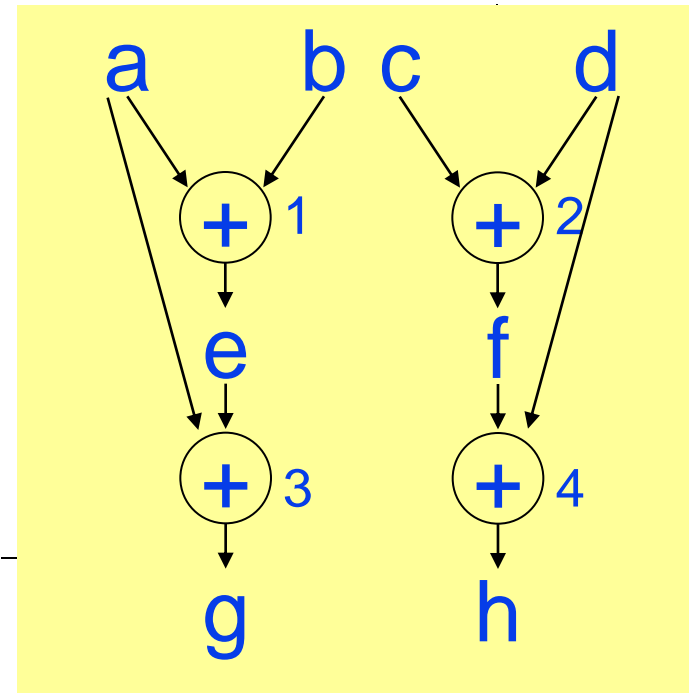
# Compilation => DFG

```
Void example (int a, int b, int g, int c, int d, int h)
```

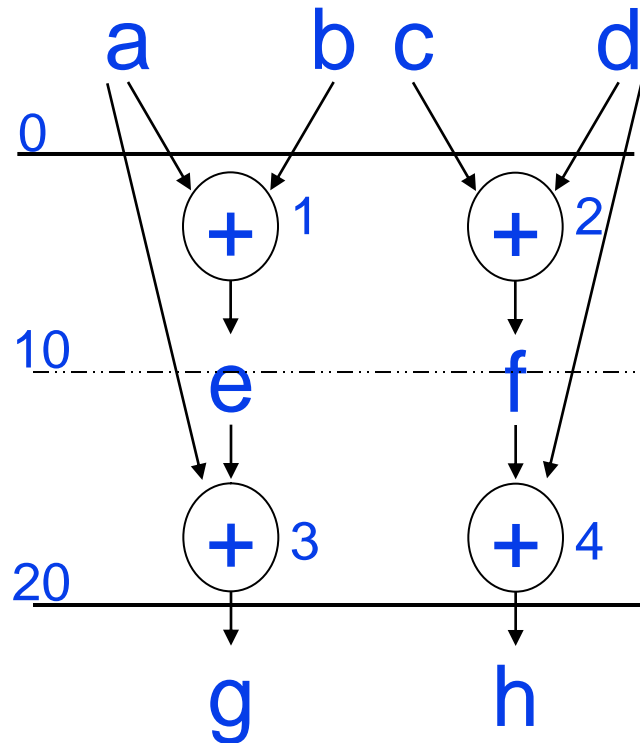
```
{
```

```
    int e,f;  
    e = a+b;  
    g = a+e;  
    f = c+d;  
    h = f+d;
```

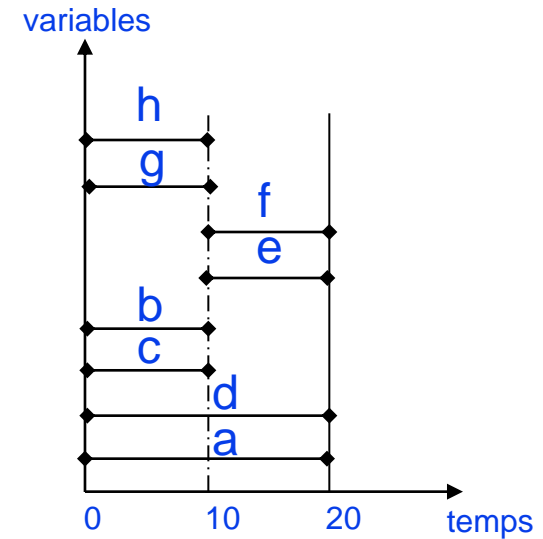
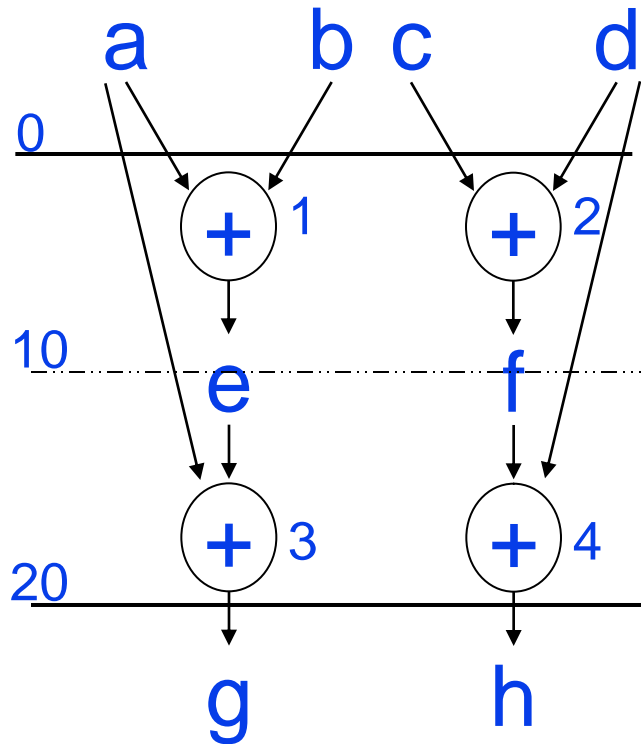
```
}
```



# Scheduling

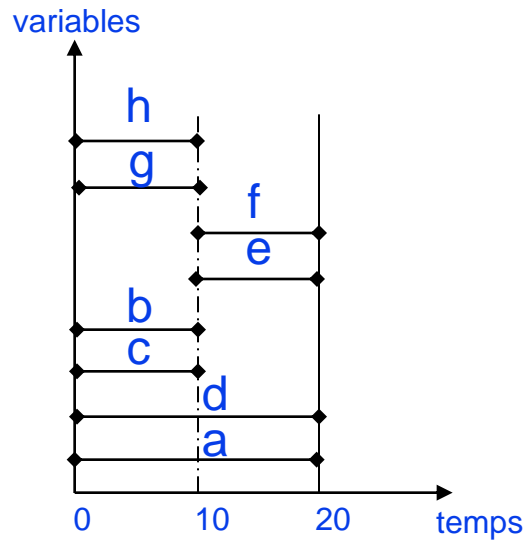


# Timing information

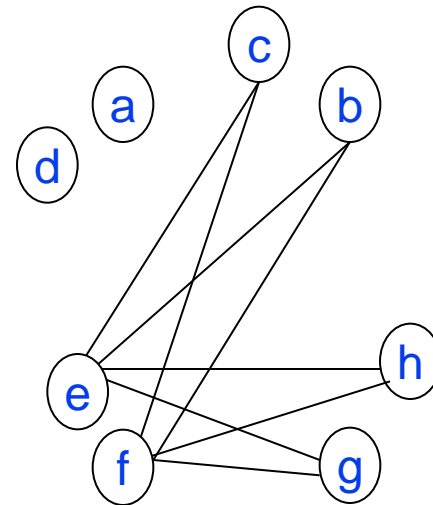


Data lifetimes

# Formal model for variable binding

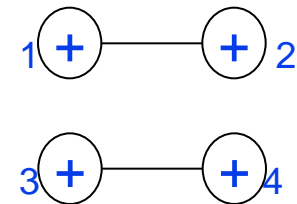
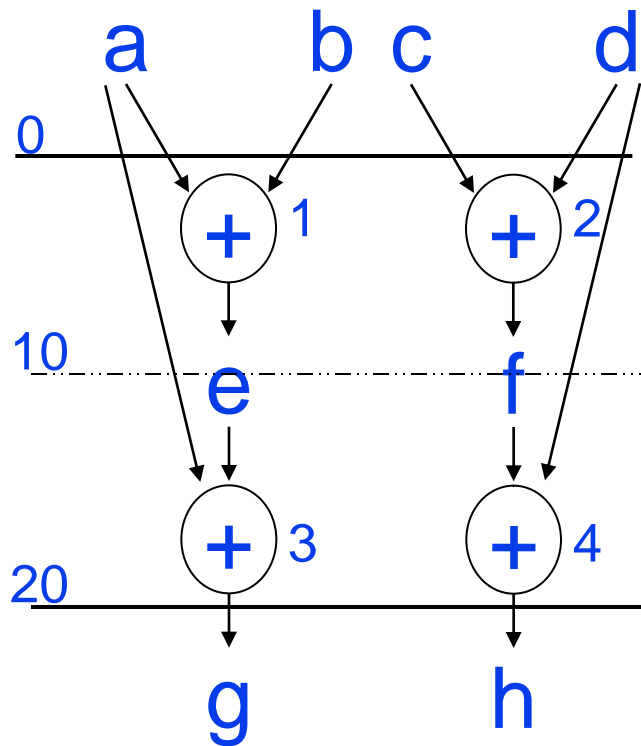


(a) Data lifetimes



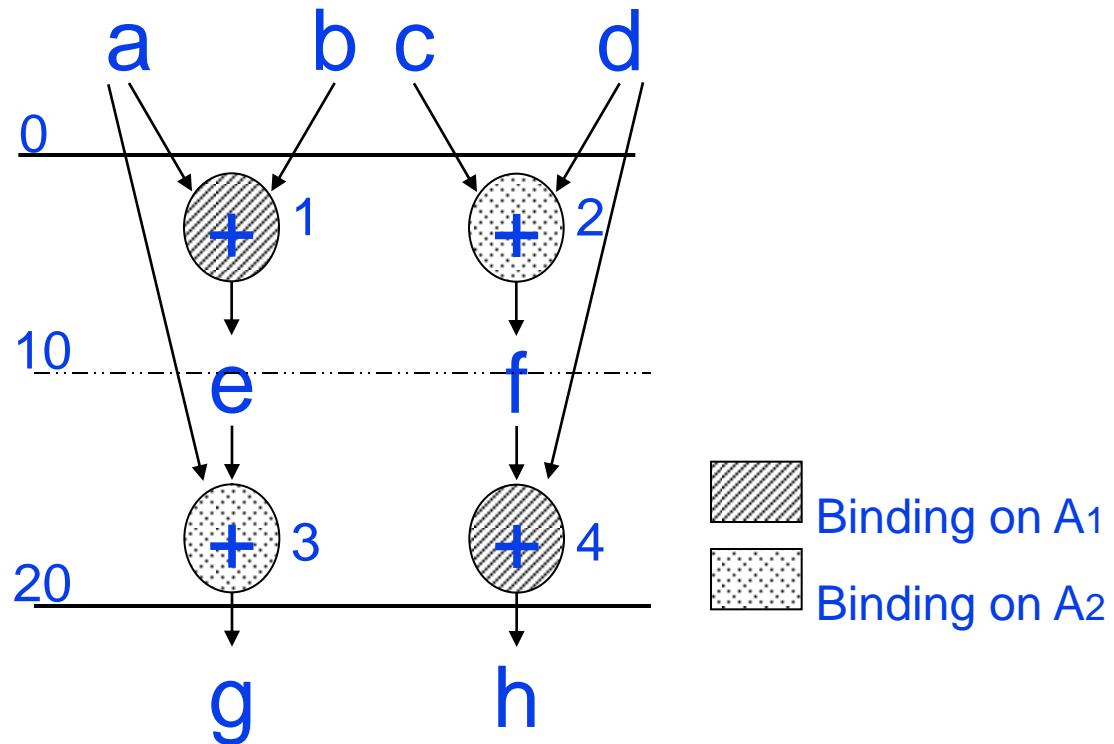
(b) Compatibility graph

# Timing information and formal model



(c) Compatibility graph

# Operation binding

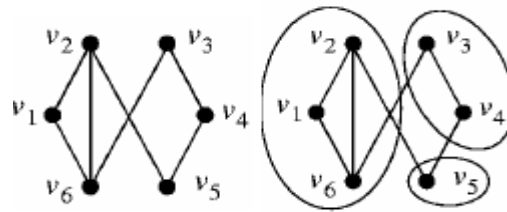


# Compatibility and conflict graphs

Clique partitioning : Binding based on a compatibility graph.

Edge exists between two data which lifetimes are not overlapping: they can share the same register.

Compatibility graph

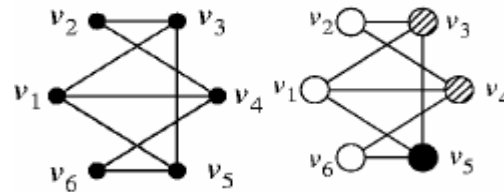


clique (sub-graph)

Graph coloring: Binding based on a conflict graph.

Edge exists between two data which lifetimes are overlapping: they can not share the same register.

Incompatibility graph

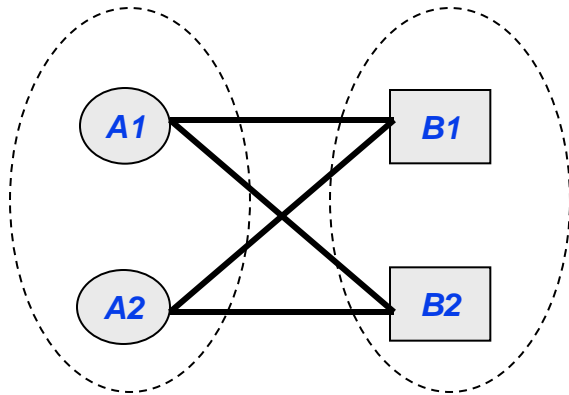


Graph coloring



# (weighted) Bipartite Graph

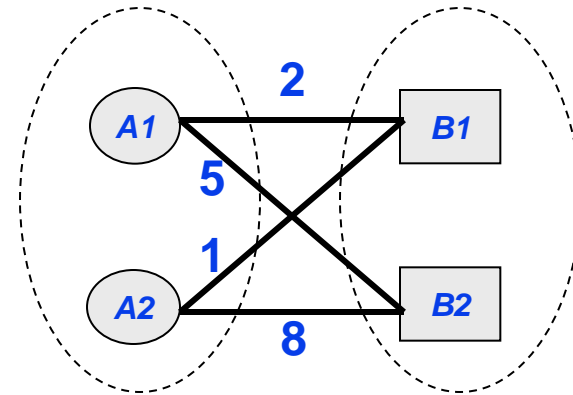
- A bipartite graph is a graph whose vertices can be divided into two disjoint sets A and B such that every edge connects a vertex in A to one in B



A

B

Bipartite Graph



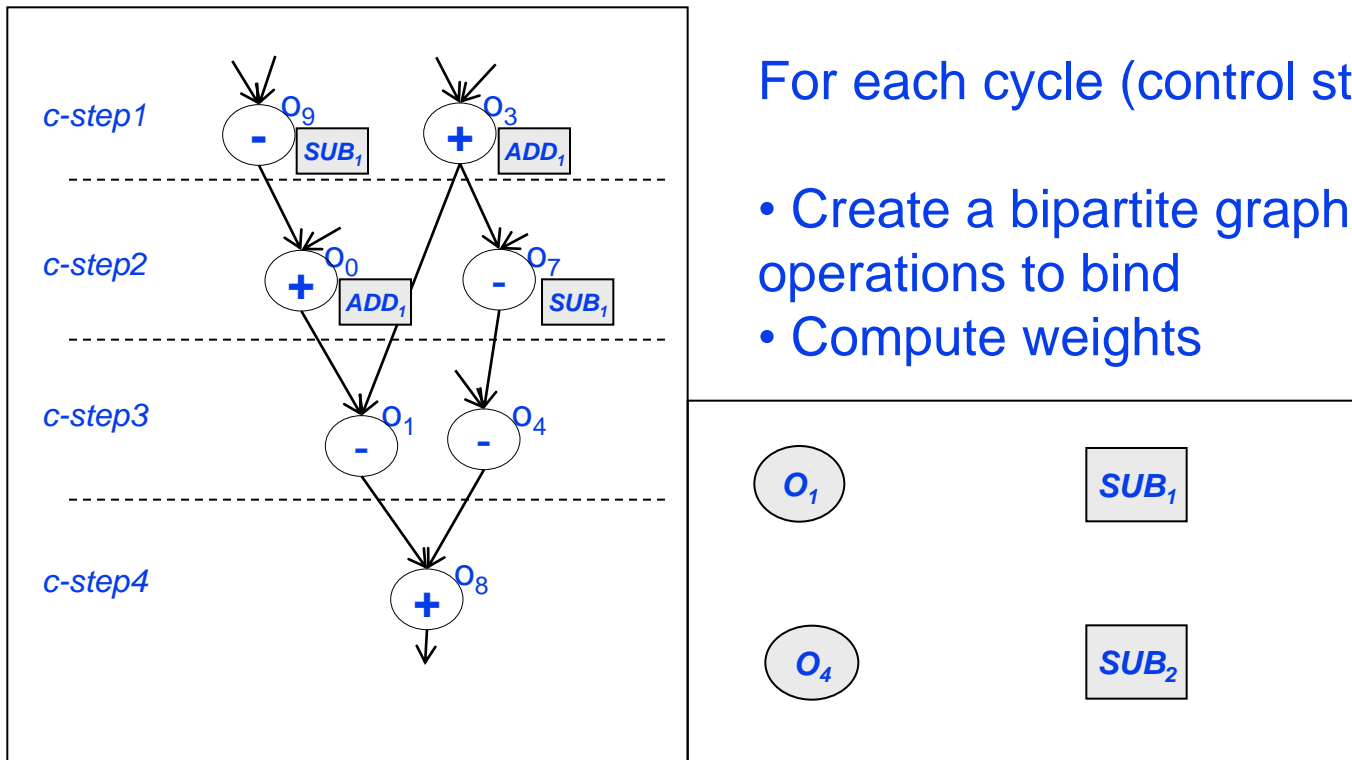
A

B

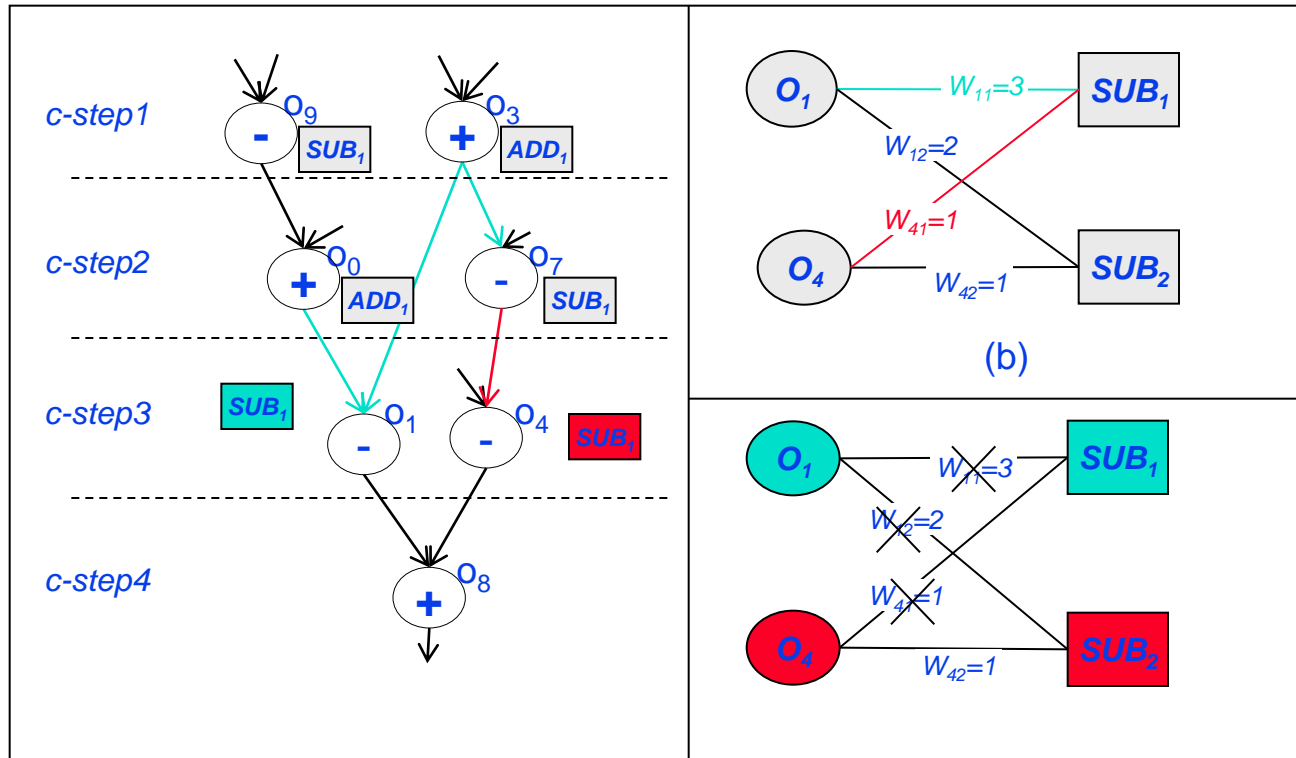
Weighted Bipartite Graph

# Example

Goal: maximize the use of existing connections between operators  
(Muxes optimization) while minimizing their size  
weight = combination between the size and the number of connection

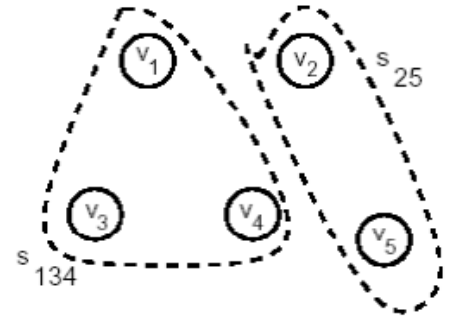
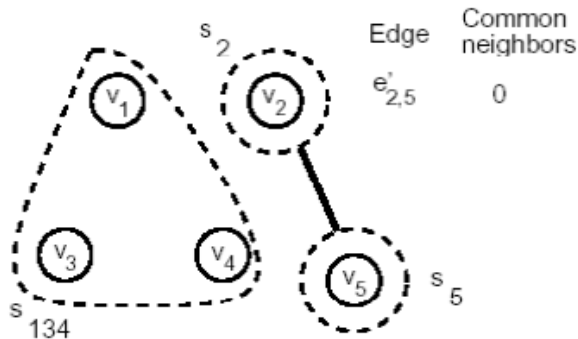
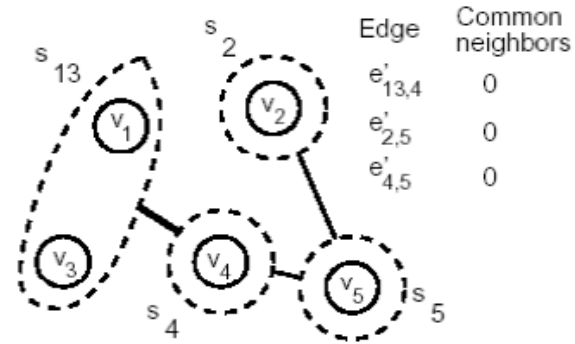
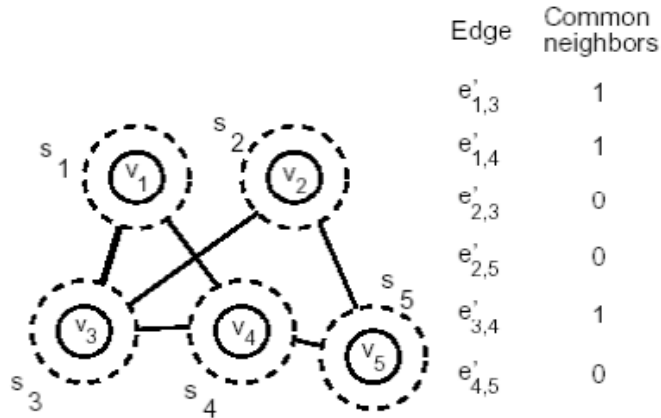


# Bipartite Weighted Matching



Maximum Weighted Bipartite Matching : Hungarian method (munkres algorithm)

# Clique partitioning algorithm: Tseng's Algorithm

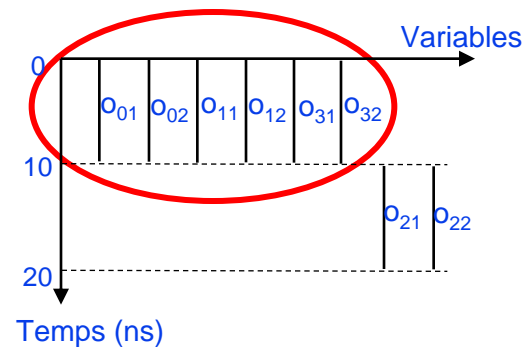
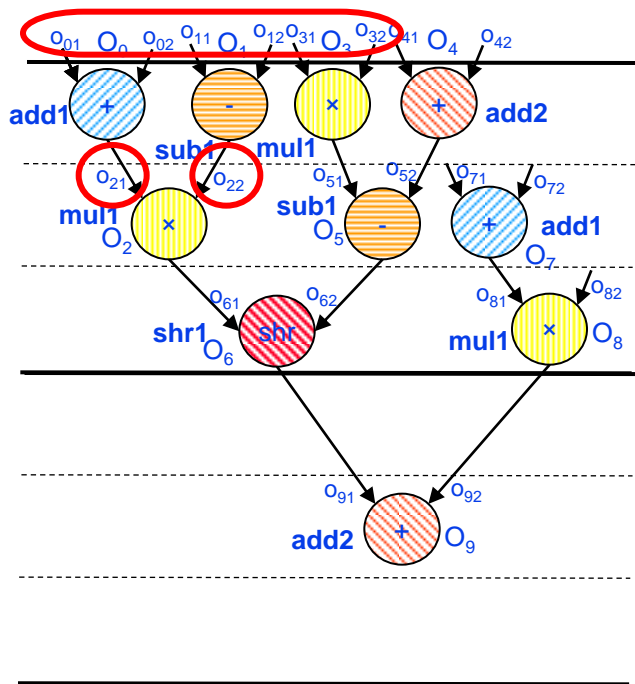


Cliques:  
 $s_{134} = \{v_1, v_3, v_4\}$   
 $s_{25} = \{v_2, v_5\}$

1. Group nodes which have the greatest common neighbor number
2. Repeat until all the edges are removed
3. Each clique corresponds to a storage unit

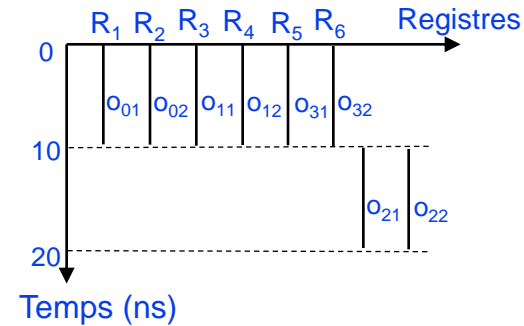
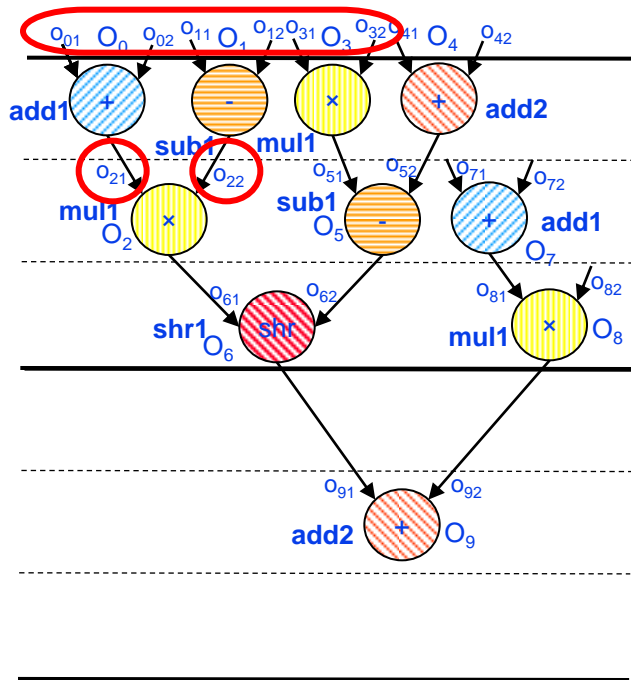
# Data binding : the Left Edge algorithm

- Data are ordered by increasing birth date
- Leftmost data are bound to distinct registers

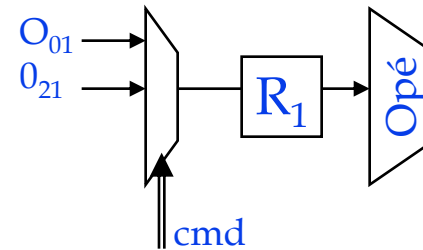
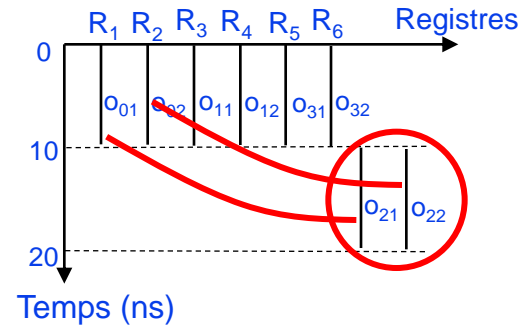
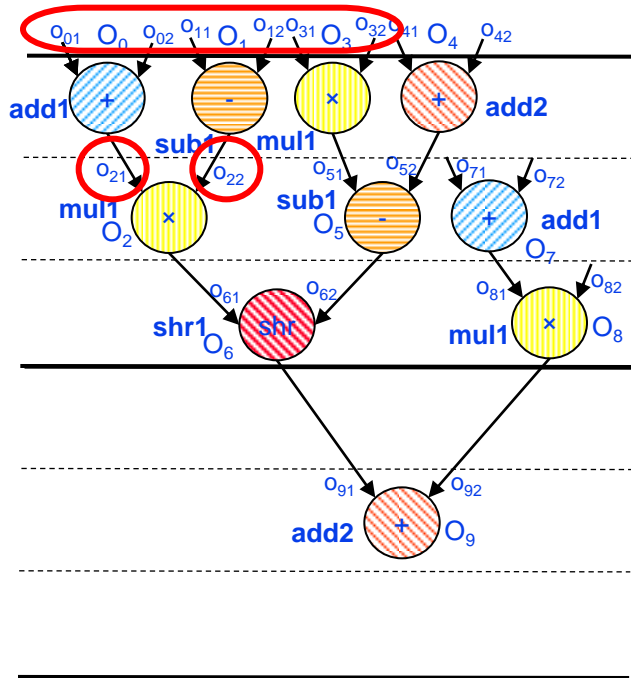


# Data binding : the Left Edge algorithm

- Data are ordered by increasing birth date
- Leftmost data are bound to distinct registers



# Data binding : the Left Edge algorithm



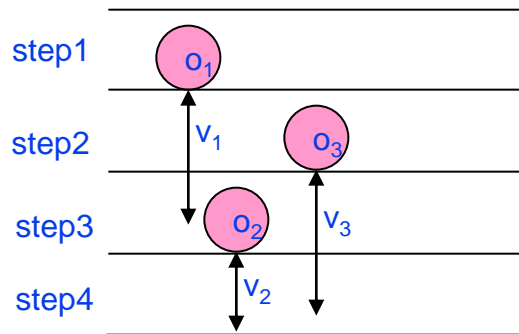
Left-edge algorithm does not take into account multiplexor cost

# Resource Binding

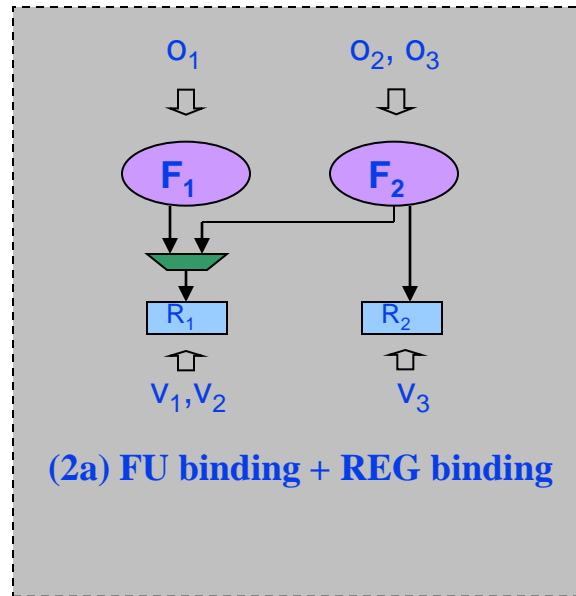
**Multiplexer and interconnect costs are significant.**

**Cyclic inter-dependency exists between FU binding and register binding**

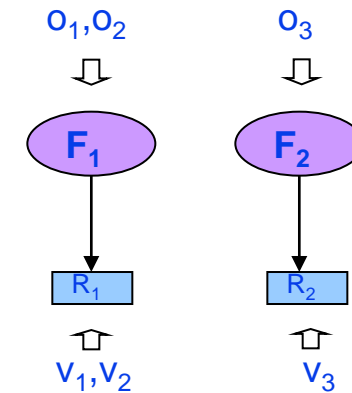
To minimize interconnection, one task needs the other's result to make accurate decision



(1) A scheduling example



(2a) FU binding + REG binding



(2b) REG binding + FU binding

Resource constraints: 2 FUs, 2 REGs

The inter-dependency is far more complicated in real designs

Use « manual allocation » to change FU binding around the best point

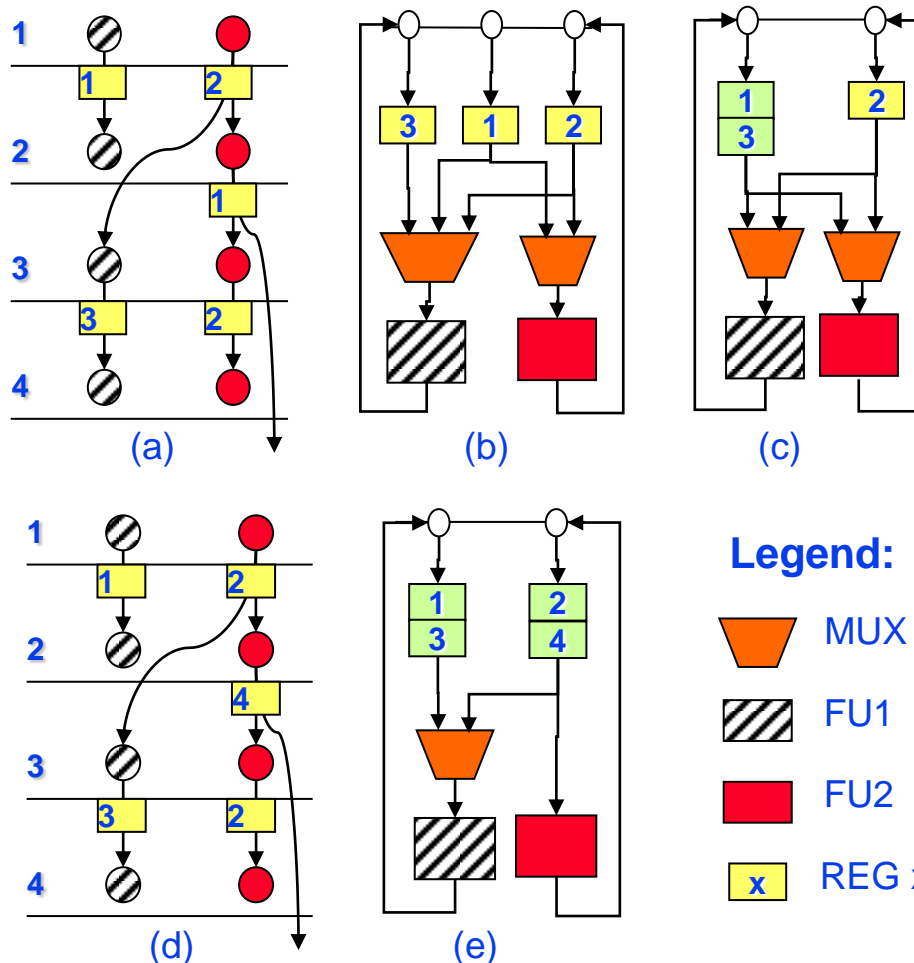
Use a metaHeuristic: Variable Neighborhood Search, simulated annealing or Tabu search



# Resource Binding

Register files may be used to hide the multiplexers, which are replaced by dedicated decoders

Merge registers with non-overlapping access dates



# Outline

---

- Lab-STICC
- General context
- High-Level Synthesis
  - Brief introduction
  - “In details”
- **GAUT**
  - Overview
  - Results
- **Conclusion**
- **References**

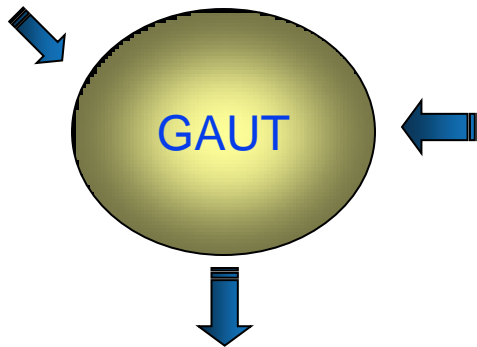
# GAUT

---

- ❑ **An academic, free and open source HLS tool**
- ❑ **Dedicated to DSP applications**
  - Data-dominated algorithm
    - ❑ *1D, 2D Filters*
    - ❑ *Transforms (Fourrier, Hadamar, DCT...)*
    - ❑ *Channel Coding, source coding algorithms*
- ❑ **Input : bit-accurate C/C++ algorithm**
  - bit-accurate integer and fixed-point from Mentor Graphics
- ❑ **Output : RTL Architecture**
  - VHDL
  - SystemC
    - ❑ *CABA: Cycle accurate and Bit accurate*
    - ❑ *TLM: Transaction level model*
    - ❑ *Compatible with both SocLib and MPARM virtual prototyping platforms*
- ❑ **Automated Test-bench generation**
- ❑ **Automated operators characterization**

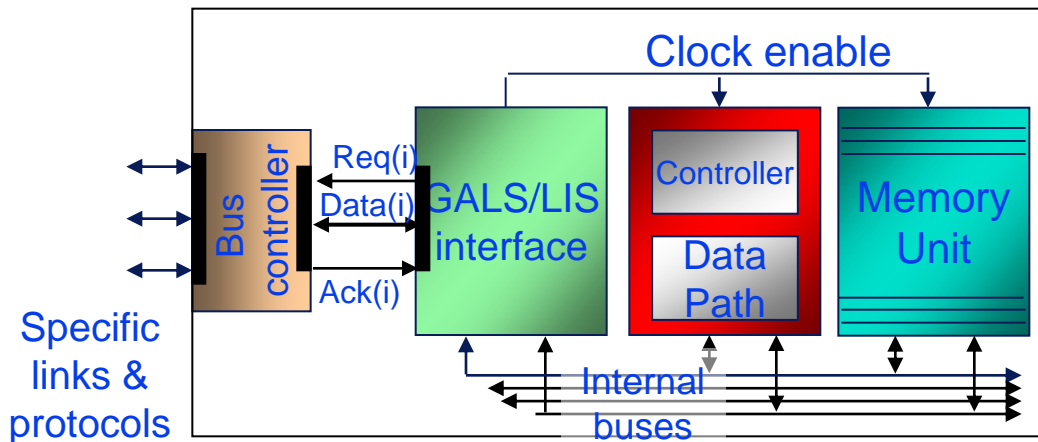
# GAUT: Constraints

Bit accurate  
Algorithm in bit-accurate C/C++



## Synthesis constraints

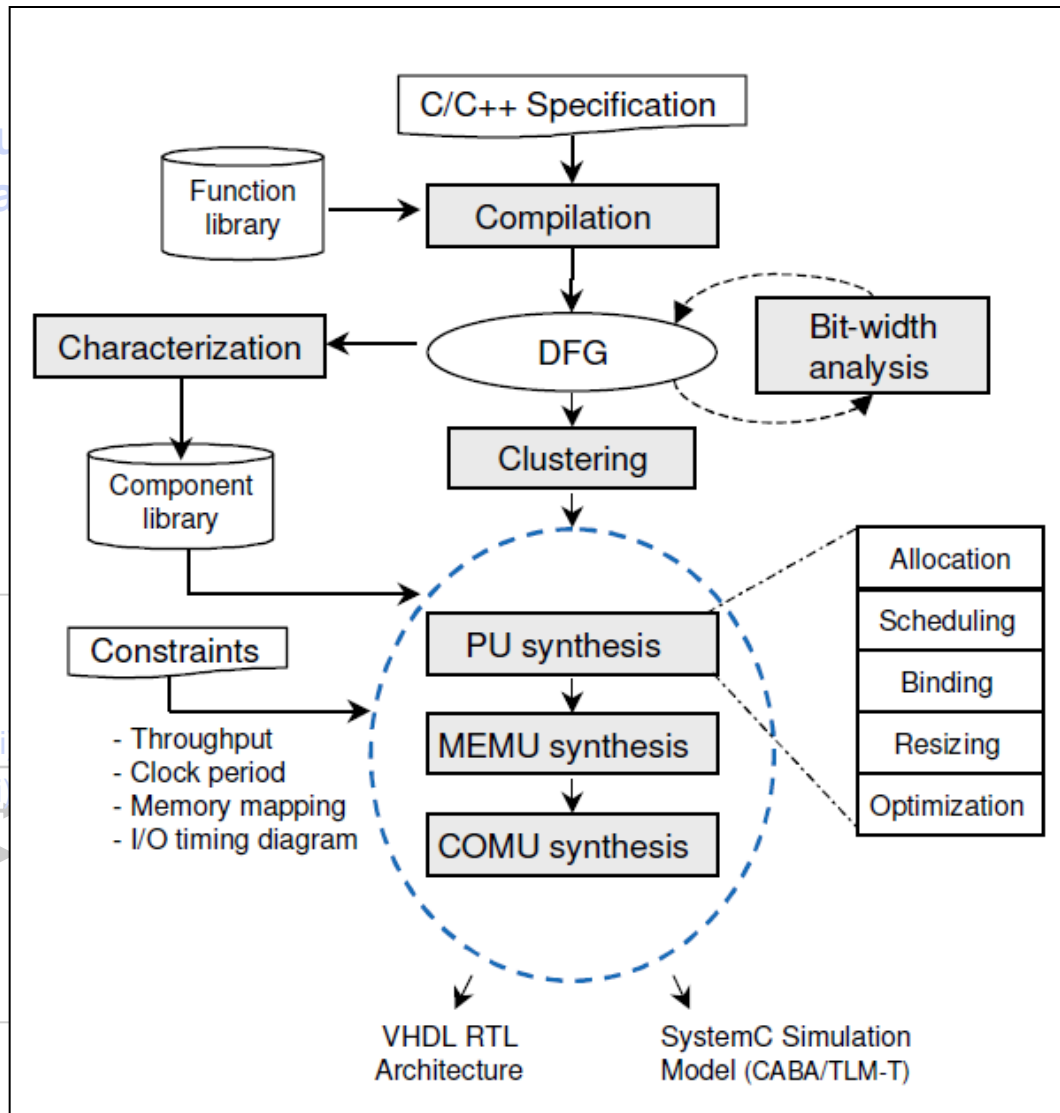
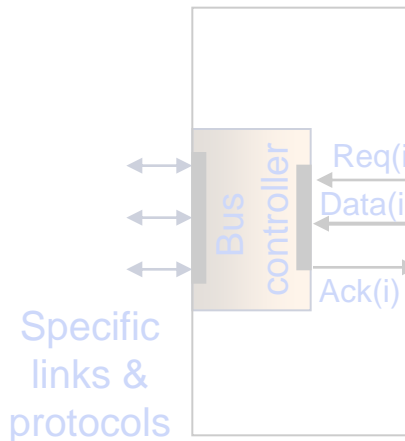
- Initiation Interval (Data average throughput )
- Clock frequency
- FPGA/ASIC target technology
  
- Memory architecture and mapping
- I/O Timing diagram (scheduling + ports)
- GALS/LIS Interface (FIFO protocol)



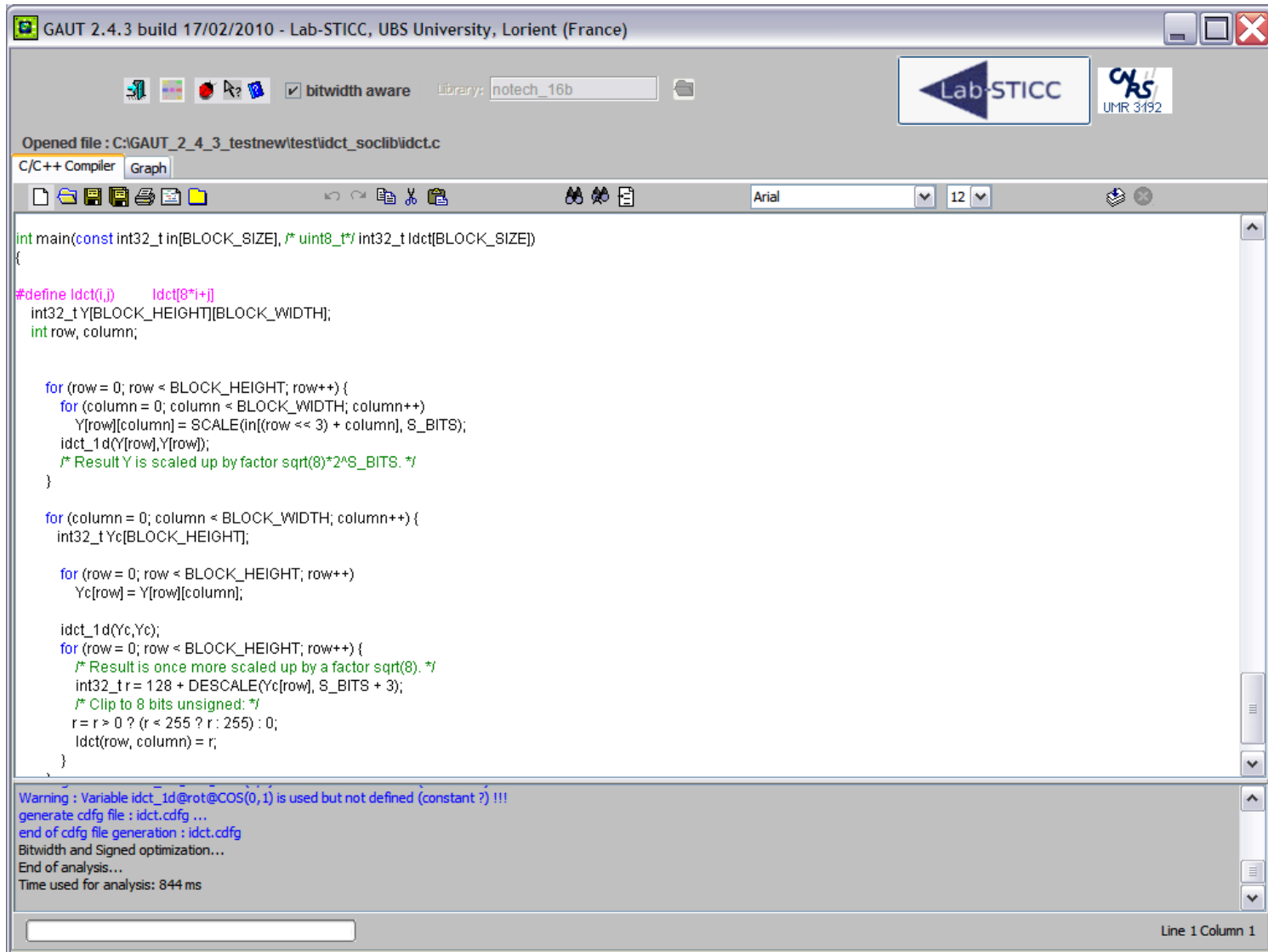
# GAUT: Design flow

Bit accurate  
Algorithm in bit-accurate

Constraints  
(average throughput)  
Technology  
Mapping  
(routing + ports)  
Protocol



# GAUT: Compilation



The screenshot displays the GAUT 2.4.3 IDE interface. The title bar reads "GAUT 2.4.3 build 17/02/2010 - Lab-STICC, UBS University, Lorient (France)". The main window shows a C++ source file named "C:\GAUT\_2\_4\_3\_testnewtest\idct\_soclib\idct.c". The code defines a function "main" that processes a 2D array of integers. It includes a macro "Idct" and several nested loops for scaling and clipping. The compilation output at the bottom shows a warning about an undefined variable and the completion of CDFG file generation and bitwidth optimization.

```
int main(const int32_t in[BLOCK_SIZE], /* uint8_t*/ int32_t Idct[BLOCK_SIZE])
{
#define Idct(i,j)  Idct[8*i+j]
  int32_t Y[BLOCK_HEIGHT][BLOCK_WIDTH];
  int row, column;

  for (row = 0; row < BLOCK_HEIGHT; row++) {
    for (column = 0; column < BLOCK_WIDTH; column++)
      Y[row][column] = SCALE(in[(row << 3) + column], S_BITS);
    idct_1d(Y[row],Y[row]);
    /* Result Y is scaled up by factor sqrt(8)*2^S_BITS. */
  }

  for (column = 0; column < BLOCK_WIDTH; column++) {
    int32_t Yc[BLOCK_HEIGHT];

    for (row = 0; row < BLOCK_HEIGHT; row++)
      Yc[row] = Y[row][column];

    idct_1d(Yc,Yc);
    for (row = 0; row < BLOCK_HEIGHT; row++) {
      /* Result is once more scaled up by a factor sqrt(8). */
      int32_t r = 128 + DESCAL(Yc[row], S_BITS + 3);
      /* Clip to 8 bits unsigned: */
      r = r > 0 ? (r < 255 ? r : 255) : 0;
      Idct(row, column) = r;
    }
  }
}

Warning : Variable idct_1d@rot@COS(0,1) is used but not defined (constant ?) !!!
generate cdfg file : idct.cdfg ...
end of cdfg file generation : idct.cdfg
Bitwidth and Signed optimization...
End of analysis...
Time used for analysis: 844 ms
```

# GAUT: DFG viewer

GAUT 2.4.3 build 17/02/2010 - Lab-STICC, UBS University, Lorient (France)

bitwidth aware Library: notech\_16b

Opened file : C:\GAUT\_2\_4\_3\_testnew\test\dct\_soclib\dct.cdfg

C/C++ Compiler Graph

add1007 4 Keep

Time used for fit size action : 15 ms

# GAUT: Operators characterization

GAUT 2.4.3 build 17/02/2010 - Lab-STICC, UBS University, Lorient (France)

Library: notech\_16b

Lab-STICC UMR 3192

Library Viewer Library Characterizing

Configuration Mode: graph to library (slow)

Target: xc5vfx110-2

Nb Bits (Top Level): 16

Graph: idct.cdfg

Optimization: speed

```
add_op_s32_s15_s33 ; function add ; combinational path delay : 1.575 ;
add_op_s32_s15_s33 ; function add ; minimum period : 2.256 ;
add_op_s32_s32_s32 ; number of slice(s) : 32 ;
add_op_s32_s32_s32 ; function add ; combinational path delay : 1.575 ;
add_op_s32_s32_s32 ; function add ; minimum period : 2.256 ;
add_op_s32_s7_s33 ; number of slice(s) : 32 ;
add_op_s32_s7_s33 ; function add ; combinational path delay : 1.575 ;
add_op_s32_s7_s33 ; function add ; minimum period : 2.256 ;
add_op_s32_s9_s32 ; number of slice(s) : 32 ;
add_op_s32_s9_s32 ; function add ; combinational path delay : 1.575 ;
add_op_s32_s9_s32 ; function add ; minimum period : 2.256 ;
gtmux_op_s32_s2_s32_s2_s32 ; number of slice(s) : 50 ;
gtmux_op_s32_s2_s32_s2_s32 ; function gtmux ; combinational path delay : 2.099 ;
gtmux_op_s32_s2_s32_s2_s32 ; function gtmux ; minimum period : 3.162 ;
lt_op_s32_s2_s32 ; number of slice(s) : 19 ;
lt_op_s32_s2_s32 ; function lt ; combinational path delay : 1.803 ;
lt_op_s32_s2_s32 ; function lt ; minimum period : 3.339 ;
ltmux_op_s32_s10_s32_s10_s32 ; number of slice(s) : 54 ;
ltmux_op_s32_s10_s32_s10_s32 ; function ltmux ; combinational path delay : 2.053 ;
ltmux_op_s32_s10_s32_s10_s32 ; function ltmux ; minimum period : 3.098 ;
mul_op_s16_s32_s32 ; number of slice(s) : 554 ;
mul_op_s16_s32_s32 ; function mul ; combinational path delay : 6.248 ;
mul_op_s16_s32_s32 ; function mul ; minimum period : 6.831 ;
mul_op_s17_s32_s32 ; number of slice(s) : 610 ;
mul_op_s17_s32_s32 ; function mul ; combinational path delay : 6.301 ;
mul_op_s17_s32_s32 ; function mul ; minimum period : 6.894 ;
mul_op_s32_s14_s32 ; number of slice(s) : 514 ;
mul_op_s32_s14_s32 ; function mul ; combinational path delay : 5.123 ;
mul_op_s32_s14_s32 ; function mul ; minimum period : 5.768 ;
mul_op_s32_s16_s32 ; number of slice(s) : 765 ;
mul_op_s32_s16_s32 ; function mul ; combinational path delay : 6.612 ;
mul_op_s32_s16_s32 ; function mul ; minimum period : 7.195 ;
mul_op_s32_s17_s32 ; number of slice(s) : 855 ;
mul_op_s32_s17_s32 ; function mul ; combinational path delay : 6.687 ;
mul_op_s32_s17_s32 ; function mul ; minimum period : 7.282 ;
s11_op_s32_s4_s32 ; number of slice(s) : 74 ;
s11_op_s32_s4_s32 ; function s11 ; combinational path delay : 1.765 ;
s11_op_s32_s4_s32 ; function s11 ; minimum period : 2.658 ;
sra_op_s33_s6_s32 ; number of slice(s) : 158 ;
sra_op_s33_s6_s32 ; function sra ; combinational path delay : 2.537 ;
sra_op_s33_s6_s32 ; function sra ; minimum period : 3.235 ;
sra_op_s34_s5_s32 ; number of slice(s) : 118 ;
sra_op_s34_s5_s32 ; function sra ; combinational path delay : 2.380 ;
sra_op_s34_s5_s32 ; function sra ; minimum period : 2.699 ;
sub_op_s32_s32_s32 ; number of slice(s) : 32 ;
sub_op_s32_s32_s32 ; function sub ; combinational path delay : 1.575 ;
sub_op_s32_s32_s32 ; function sub ; minimum period : 2.260 ;
sub_op_s33_s32_s34 ; number of slice(s) : 33 ;
sub_op_s33_s32_s34 ; function sub ; combinational path delay : 1.597 ;
sub_op_s33_s32_s34 ; function sub ; minimum period : 2.283 ;
=> generate : C:\GAUT_2_4_3_testnew\test\idct_soclib\idct.lib
end.
```

Script and logic

Area : operator only (nb slice)

Propagation time : reg+tri+ope+reg

Database, interpolation...

Status: Starting : carac -cdfg C:\GAUT\_2\_4\_3\_testnew\test\idct\_soclib\idct.cdfg -builtins C:/GAUT\_2\_4\_3\_testnew/GautC/lib/CDFG\_fun



# GAUT: Synthesis steps

The screenshot shows the GAUT 2.4.3 synthesis interface. The main window displays the synthesis configuration and results. The configuration panel on the left includes options for Graph, Lib, Cadency, Clock, IO constraints, Mem constraints, Verbose, Operator optimization, Bitwidth, Chaining, Allocation strategy, Scheduling strategy, Register Allocation, Vhdl output, Output, and Control. The central panel shows a table of operators and their usage, along with a summary of synthesis statistics. The status bar at the bottom indicates "End of synthesis".

**Annotations:**

- Initiation Interval II**: Points to the "Cadency" field, which is set to 940.
- Clock period**: Points to the "Clock" field, which is set to 5 ns.
- I/O timing & memory constraints**: Points to the "IO constraints" and "Mem constraints" radio buttons.
- Data Assginment (Left Edge, MWBM...)**: Points to the "Register Allocation" dropdown menu, which is set to "MWBM".
- HDL coding style: FSMD, FSM+reg, FSM\_ROM+reg...**: Points to the "Vhdl output" dropdown menu, which is set to "fsm\_sigs".

Operator	Number	Use
add_op_s_s_s	0	
add_op_s_s_s	1	
add_op_s_s_s	2	
add_op_s_s_s	14	64
gtmux_op_s_s_s_s	3	34
lt_op_s_s_s	4	34
ltmux_op_s_s_s_s	5	34
mul_op_s_s_s	6	61
mul_op_s_s_s	7	61
mul_op_s_s_s	8	58
mul_op_s_s_s		
sll_op_s_s_s	9	
sra_op_s_s_s	1	
sra_op_s_s_s	1	
sub_op_s_s_s		
sub_op_s_s_s	13	65

Time used for Scheduling : 610 ms  
WARNING: CDFG latency(935) < cadency(940)  
Resizing operators...  
Resizing Mode 2 : input widths are put to the max for each position  
Expected area = 3147  
Time used for resizing operators : 15 ms  
Registers allocation ...  
Number of hardwired constant false register : 15  
Number of fifo register : 0  
Minimal Number of simple register : 100  
Optimal number of simple register for mux optimiz : 4819  
Number of flip flop : 4819  
Number of mux 2 to 1 : 6911  
Time used for registers allocation : 2422 ms  
Data Bus allocation ...

# GAUT: I/O and memory constraints

GAUT 2.4.3 build 17/02/2010 - Lab-STICC, UBS University, Lorient (France)

Library: notech\_16b

Opened file : null

Input/Output Constraints Memory Constraints Synthesis Multi Mode

from PU point of view

Name	Mode	Port	Time
in(0)	Input	1	0
in(1)	Input	1	5
in(2)	Input	1	10
in(3)	Input	1	15
in(4)	Input	1	20
in(5)	Input	1	25
in(6)	Input	1	30
in(7)	Input	1	35
in(8)	Input	1	40
in(9)	Input	1	45
in(10)	Input	1	50
in(11)	Input	1	55
in(12)	Input	1	60
in(13)	Input	1	65
in(14)	Input	1	70
in(15)	Input	1	75
in(16)	Input	1	80
in(17)	Input	1	85
in(18)	Input	1	90
in(19)	Input	1	95
in(20)	Input	1	100
in(21)	Input	1	105
in(22)	Input	1	110
in(23)	Input	1	115
in(24)	Input	1	120
in(25)	Input	1	125
in(26)	Input	1	130
in(27)	Input	1	135
in(28)	Input	1	140
in(29)	Input	1	145
in(30)	Input	1	150

Time used for creating io constraints table : 16 ms

# GAUT: Gantt viewer

GAUT 2.4.3 build 17/02/2010 - Lab-STICC, UBS University, Lorient (France)

Library: notech\_16b

Opened file: C:\GAUT\_2\_4\_3\_testnew\testidct\_soclibidct\_UT.gantt

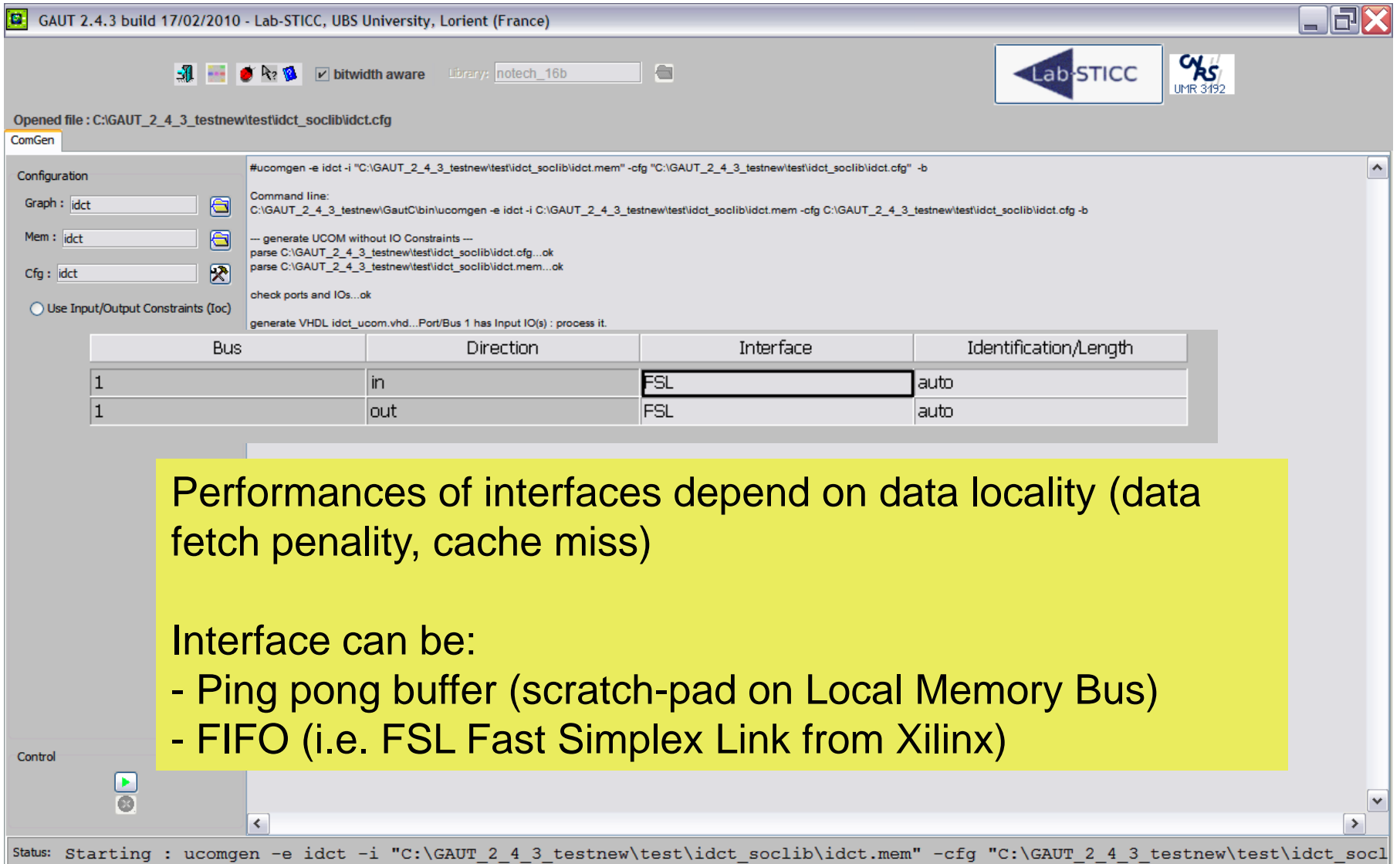
Gantt

Keep

The Gantt chart displays the execution timeline of various operations. The horizontal axis represents time in clock cycles, with major ticks every 5 cycles (0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70). The vertical axis lists operations and registers. Operations are represented by colored bars: purple for constants, yellow for idct operations, blue for multiplication, and orange for addition and subtraction. A specific operation, Y(0,0), is highlighted with a white box and a tooltip showing its duration from cycle 15 to 30. The chart shows a complex sequence of operations, including data path operations like 'in(0)' through 'in(14)', and various arithmetic and control logic blocks.

Time used for creating gantt diagram : 281 ms

# GAUT: Interface synthesis



GAUT 2.4.3 build 17/02/2010 - Lab-STICC, UBS University, Lorient (France)

Library: notech\_16b

Opened file : C:\GAUT\_2\_4\_3\_testnew\test\idct\_soclib\idct.cfg

ComGen

Configuration

Graph : idct

Mem : idct

Cfg : idct

Use Input/Output Constraints (Ioc)

```
#ucomgen -e idct -i "C:\GAUT_2_4_3_testnew\test\idct_soclib\idct.mem" -cfg "C:\GAUT_2_4_3_testnew\test\idct_soclib\idct.cfg" -b
```

Command line:  
C:\GAUT\_2\_4\_3\_testnew\GautC\bin\ucomgen -e idct -i C:\GAUT\_2\_4\_3\_testnew\test\idct\_soclib\idct.mem -cfg C:\GAUT\_2\_4\_3\_testnew\test\idct\_soclib\idct.cfg -b

--- generate UCOM without IO Constraints ---  
parse C:\GAUT\_2\_4\_3\_testnew\test\idct\_soclib\idct.cfg...ok  
parse C:\GAUT\_2\_4\_3\_testnew\test\idct\_soclib\idct.mem...ok

check ports and IOs...ok

generate VHDL idct\_ucom.vhd...Port/Bus 1 has Input IO(s) : process it.

Bus	Direction	Interface	Identification/Length
1	in	FSL	auto
1	out	FSL	auto

Control

Status: Starting : ucomgen -e idct -i "C:\GAUT\_2\_4\_3\_testnew\test\idct\_soclib\idct.mem" -cfg "C:\GAUT\_2\_4\_3\_testnew\test\idct\_soclib\idct.cfg" -b

Performances of interfaces depend on data locality (data fetch penalty, cache miss)

Interface can be:

- Ping pong buffer (scratch-pad on Local Memory Bus)
- FIFO (i.e. FSL Fast Simplex Link from Xilinx)

# GAUT: Test-bench generation

The screenshot displays the GAUT 2.4.3 build 17/02/2010 interface. The title bar indicates the location: Lab-STICC, UBS University, Lorient (France). The main window shows the configuration of a testbench for a VHDL design. The configuration panel on the left includes fields for Mem (idct), Impl (bitwidth\_notech), Simulate (PU), Graph (idct), Stimuli (FILE), and Stimulation (idct). The main area shows the compilation and simulation process, including the generation of a testbench and the execution of a ModelSim script. The simulation results are displayed in a waveform viewer, showing the timing of various signals. The waveform shows a clock signal (clk) and several data signals (bus\_1, bus\_2, s0, s...). The simulation time is 990 ns, and the cursor is at 1021 ns.

Configuration

- Mem : idct
- Impl : bitwidth\_notech
- Simulate : PU
- Graph : idct
- Stimuli : FILE
- Sti : idct
- Nb Iteration : 1
- Result File
- Warning
- Comp lib
- Choose vsim directory : E:/modeltech\_6.4f/win32/

Compilation Log:

```
- Compiling package idct_pack
- Loading package idct_pack
- Compiling entity idct_um
- Compiling architecture idct_um_arch of idct_um
- Compiling entity idct_top
- Compiling architecture idct_top_arch of idct_top
- Compiling entity idct_stimuli
- Compiling architecture idct_stimuli_arch of idct_stimuli
- Loading package textio
- Compiling entity idct_probe
- Compiling architecture idct_probe_arch of idct_probe
** Warning: testbench.vhd(1485): (vcom-1194) FILE declaration was written using VHDL_1993
** Warning: testbench.vhd(1486): (vcom-1236) Shared variable
- Compiling entity idct_test
- Compiling architecture idct_test_arch of idct_test
```

Simulation Log:

```
#E:/modeltech_6.4f/win32/vsim -c -do script_compil.do > C:\Program Files\Xilinx\ISE10.1\bin\nt\vsim.exe
Reading E:/modeltech_6.4f/tcl/vsim/pref.tcl
# 6.4f
# do script_compil.do
# Loading project idct

#E:/modeltech_6.4f/win32/vsim -c -do script_modelsim.val -q
Reading E:/modeltech_6.4f/tcl/vsim/pref.tcl
# 6.4f
# vsim -do script_modelsim.val -c -quiet work.idct_test
# ** Note: (vsim-3812) Design is being optimized...
# // ModelSim SE 6.4f Oct 22 2009
# //
# // Copyright 1991-2009 Mentor Graphics Corporation
# // All Rights Reserved.
# //
# // THIS WORK CONTAINS TRADE SECRET AND
# // PROPRIETARY INFORMATION WHICH IS THE PROPERTY
# // OF MENTOR GRAPHICS CORPORATION OR ITS LICEN
# // AND IS SUBJECT TO LICENSE TERMS.
# //
# do script_modelsim.val
# 1
```

Waveform Viewer:

Signal	Value
clk	0
bus_1	0
bus_2	0
s0	0
s...	0

Status: You can see the results here or in the file

Test-bench Generation  
Modelsim Script Generation  
Result File Generation

# GAUT: more than 100 downloads each year



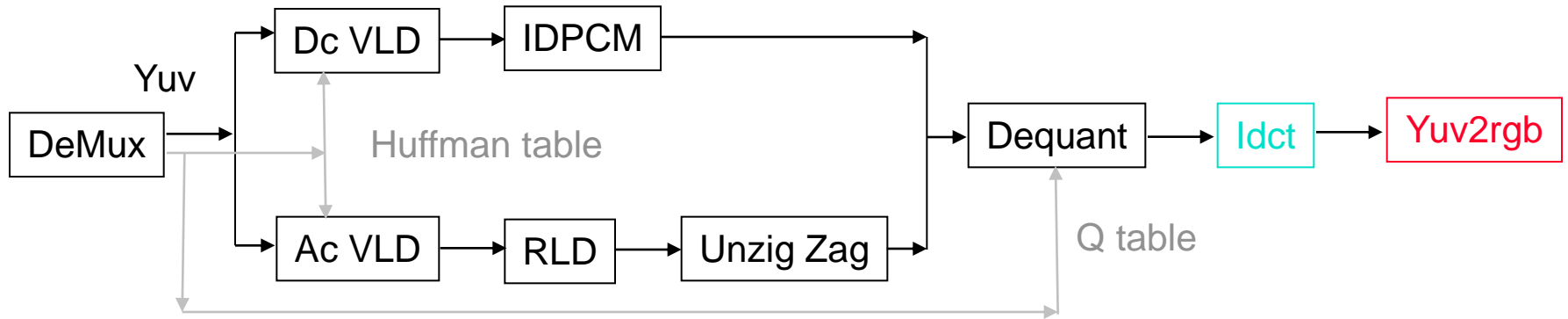
From more than 60 countries

# Outline

---

- Lab-STICC
- General context
- High-Level Synthesis
  - Brief introduction
  - “In details”
- **GAUT**
  - Overview
  - Results
- **Conclusion**
- **References**

# Experimental results: MJPEG decoding



Block Diagram of mjpeg baseline decoder

Function	Time ratio
IDCT	43,41%
yuv2rgb	15,07%
Entropy decoding	8,41%
DeQuantization	5,10%
others (each function is <5%)	28,01%

Execution time ratio for software MJPEG decoding  
(by using gprof)



# Resource estimation for IDCT

```

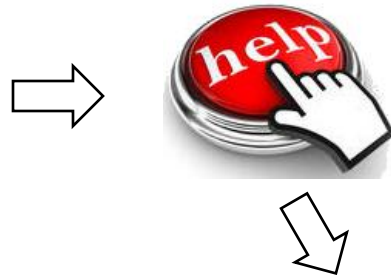
C/C++ Compiler Graph
int main(const int32_t in[BLOCK_SIZE], /* uint8_t */ int32_t Idct[BLOCK_SIZE])
{
#define Idct(i,j) Idct(8*i+j)
int32_t Y[BLOCK_HEIGHT][BLOCK_WIDTH];
int row, column;

for (row = 0; row < BLOCK_HEIGHT; row++) {
for (column = 0; column < BLOCK_WIDTH; column++)
Y[row][column] = SCALE(in((row << 3) + column), S_BITS);
Idct_1d(Y[row], Y[row]);
/* Result Y is scaled up by factor sqrt(8)*2^S_BITS. */
}

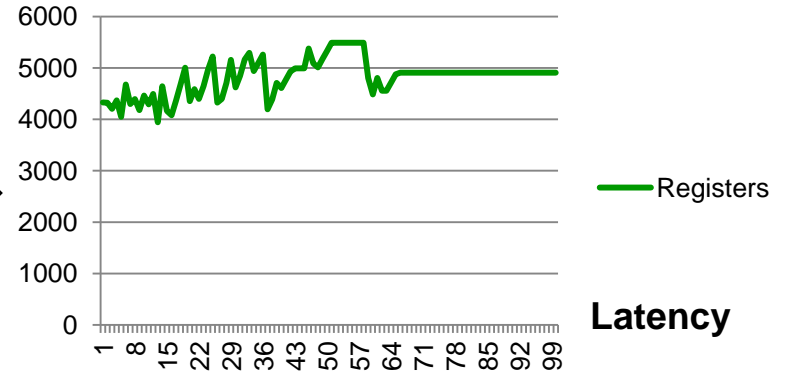
for (column = 0; column < BLOCK_WIDTH; column++) {
int32_t Yc[BLOCK_HEIGHT];

for (row = 0; row < BLOCK_HEIGHT; row++)
Yc[row] = Y[row][column];

Idct_1d(Yc, Yc);
for (row = 0; row < BLOCK_HEIGHT; row++) {
/* Result is once more scaled up by a factor sqrt(8). */
int32_t r = 128 + DESCALE(Yc[row], S_BITS + 3);
/* Clip to 8 bits unsigned. */
r = r > 0 ? (r < 255 ? r : 255) : 0;
Idct(row, column) = r;
}
}
Warning : Variable idct_1d@rot@COS(0,1) is used but not defined (constant ?) !!!
generate cfg file : idct.cfg ...
    
```

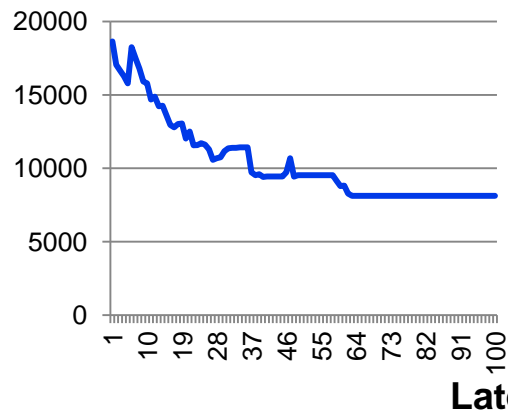


## Registers



## LUTs

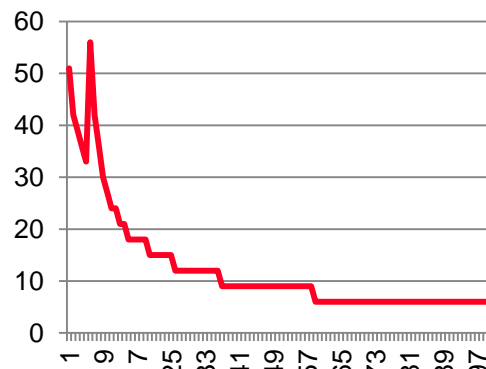
— LUTs



Latency

— DSP

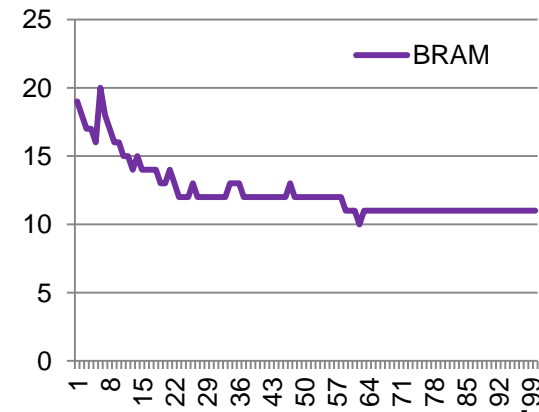
## DSP Block



Latency

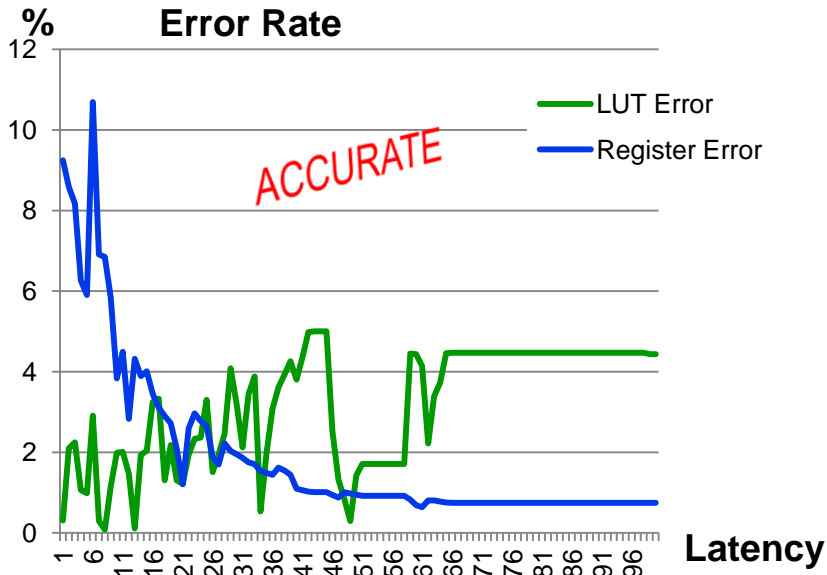
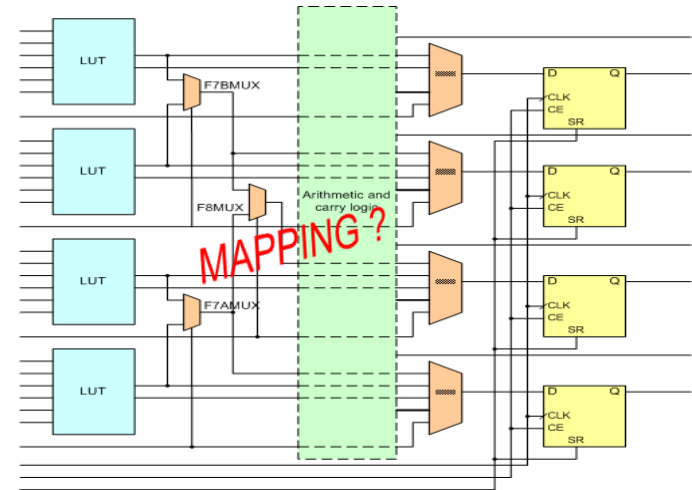
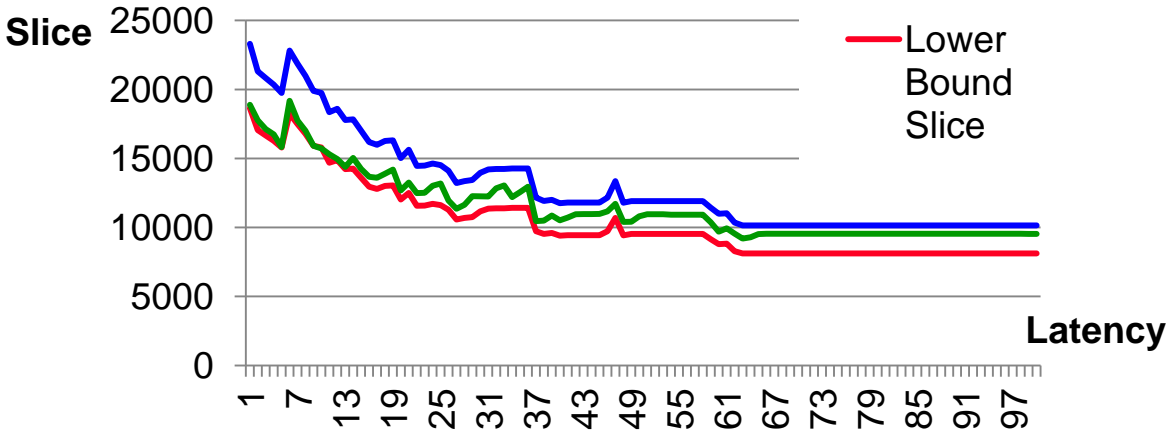
## BRAM

— BRAM

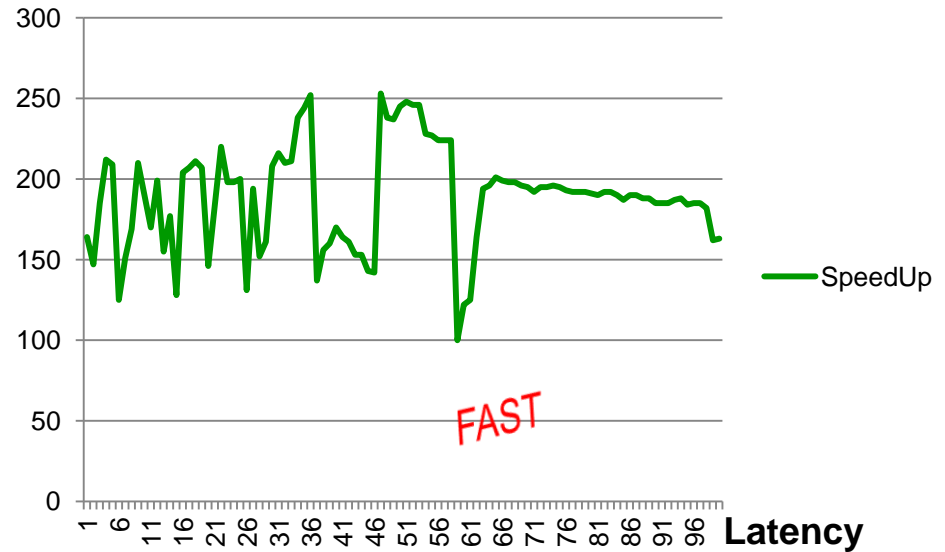


Latency

# Resource estimation for IDCT



SpeedUp = Logic Synthesis Time/HLS Synthesis Time



# Synthesis results

Parallelism 1 (read/write 32 bits)								
Operators				Reg (1 bit flip flop)	Mux 2:1	Area (slices)	Latency (cycles)	Freq (Mhz)
<i>add</i>	<i>mult</i>	<i>sra</i>	<i>sub</i>					
2	3	1	1	2653	3236	7033	129	123,5
2	2	1	1	2818	3525	6948	188	128,1
1	2	1	1	3304	3905	6988	228	124
1	1	1	1	2876	3858	6192	348	123,7
1	1	1	1	2421	3938	6422	448	125,7

Parallelism 2 (read/write 64 bits)									
Operator				Reg (1 bit flip flop)	Mux 2:1	Area (slices)	Latency (cycles)	Freq (Mhz)	
<i>add</i>	<i>mult</i>	<i>sra</i>	<i>sub</i>						
4	7	3	2	2904	2965	9409	97	126	
2	3	1	1	2942	3268	7863	156	120,2	
2	3	1	1	3112	3300	8101	196	128,9	
1	2	1	1	3429	3969	7529	316	128,4	
1	1	1	1	2880	3106	6498	416	121,9	

Parallelism 4 (read/write 128 bits)								
Operator				Reg (1 bit flip flop)	Mux 2:1	Area (slices)	Latency (cycles)	Freq (Mhz)
<i>add</i>	<i>mult</i>	<i>sra</i>	<i>sub</i>					
9	15	6	5	3459	2694	12070	33	138,9
3	4	2	2	3021	2947	9282	92	132,1
2	3	1	1	2917	3091	7812	132	128,7
1	2	1	1	3462	4257	7846	252	122,5
1	1	1	1	2850	3314	6719	352	120,8

IDCT

Parallelism 1 (read/write 32 bits)			
Reg (1 bit flip flop)	Area (slices)	Latency (cycles)	Freq (Mhz)
388	525	12	249,18
362	524	13	282,11
272	462	14	188,96
238	460	15	188,96

YUV2RGB

# Synthesis results

Parallelism 1 (read/write 32 bits)					Reg (1 bit flip flop)	Mux 2:1	Area (slices)	Latency (cycles)	Freq (Mhz)
Operators									
<i>add</i>	<i>mult</i>	<i>sra</i>	<i>sub</i>						
2	3	1	1	2653	3236	7033	129	123,5	
2	2	1	1	2818	3525	6948	188	128,1	
1	2	1	1	3304	3905	6988	228	124	
1	1	1	1	2876	3858	6192	348	123,7	
1	1	1	1	2421	3938	6422	448	125,7	

Parallelism 2 (read/write 64 bits)					Reg (1 bit flip flop)	Mux 2:1	Area (slices)	Latency (cycles)	Freq (Mhz)
Operator									
<i>add</i>	<i>mult</i>	<i>sra</i>	<i>sub</i>						
4	7	3	2	2904	2965	9409	97	126	
2	3	1	1	2942	3268	7863	156	120,2	
2	3	1	1	3112	3300	8101	196	128,9	
1	2	1	1	3429	3969	7529	316	128,4	
1	1	1	1	2880	3106	6498	416	121,9	

Virtual prototyping

IDCT

Parallelism 4 (read/write 128 bits)								
Operator				Reg (1 bit flip flop)	Mux 2:1	Area (slices)	Latency (cycles)	Freq (Mhz)
<i>add</i>	<i>mult</i>	<i>sra</i>	<i>sub</i>					
9	15	6	5	3459	2694	12070	33	138,9
3	4	2	2	3021	2947	9282	92	132,1
2	3	1	1	2917	3091	7812	132	128,7
1	2	1	1	3462	4257	7846	252	122,5
1	1	1	1	2850	3314	6719	352	120,8

Hardware prototyping

YUV2RGB

Parallelism 1 (read/write 32 bits)			
Reg (1 bit flip flop)	Area (slices)	Latency (cycles)	Freq (Mhz)
388	525	12	249,18
362	524	13	282,11
272	462	14	188,96
238	460	15	188,96

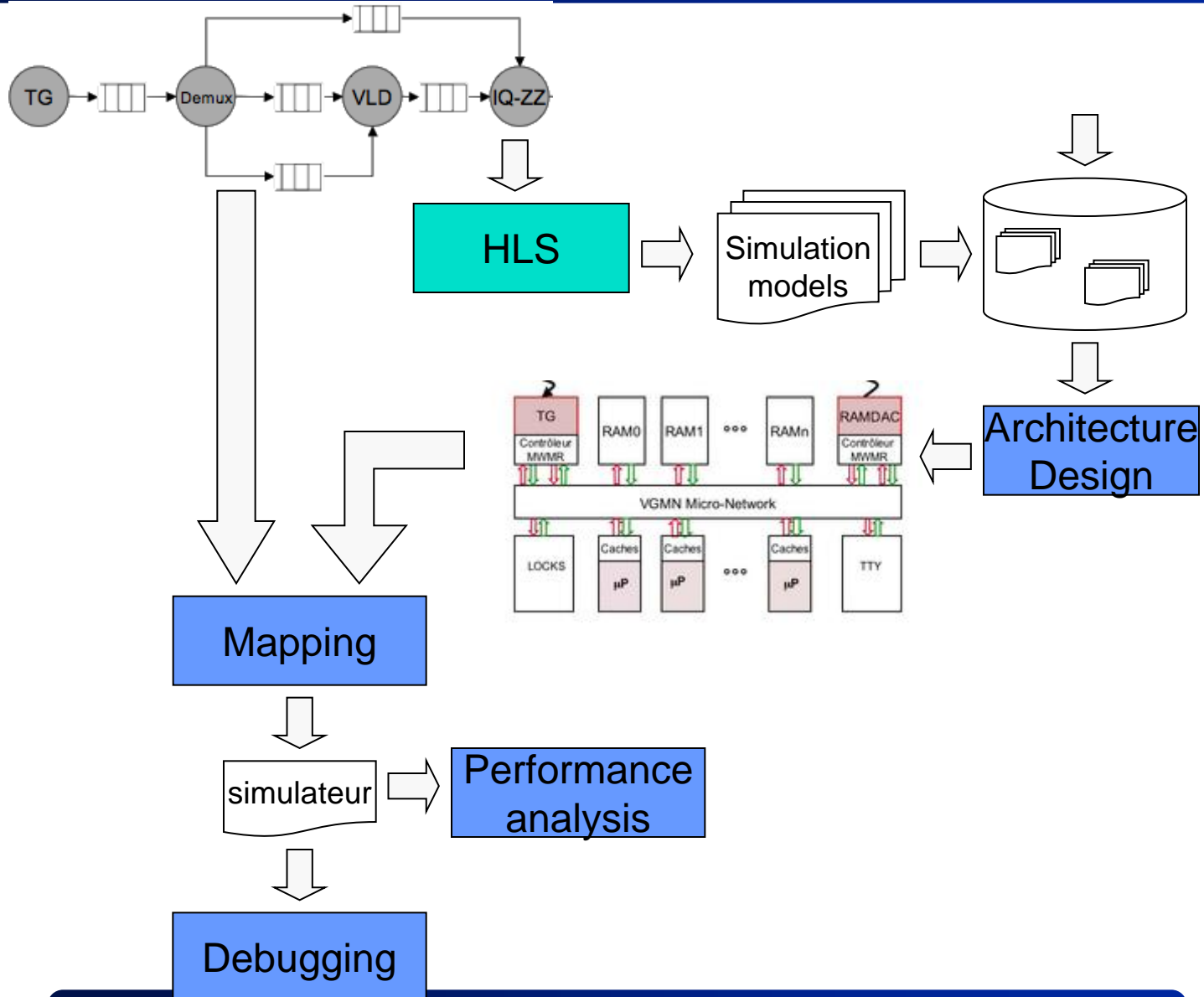
# SoCLib: a virtual prototyping platform

- French National Research Project (ANR)
- Free and open source virtual prototyping environment
  - **Library of SystemC simulation models**
  - **Hardware components**
    - CPUs, HW-ACCs, memories, busses
    - VCI/OCP interface protocol is used
  - **Two types of model are available for each HW component**
    - CABA (*Cycle Accurate / Bit Accurate*)
    - TLM-DT (*Transaction Level Modeling with Distributed Time*)
  - **Software components**
    - OS, API...
  - **Associated tools**
    - Simulation, configuration, debug
    - Automatic generation of simulation models
- GAUT is used, to generate simulation models of HW-ACC
  - **CABA and TLM-DT**

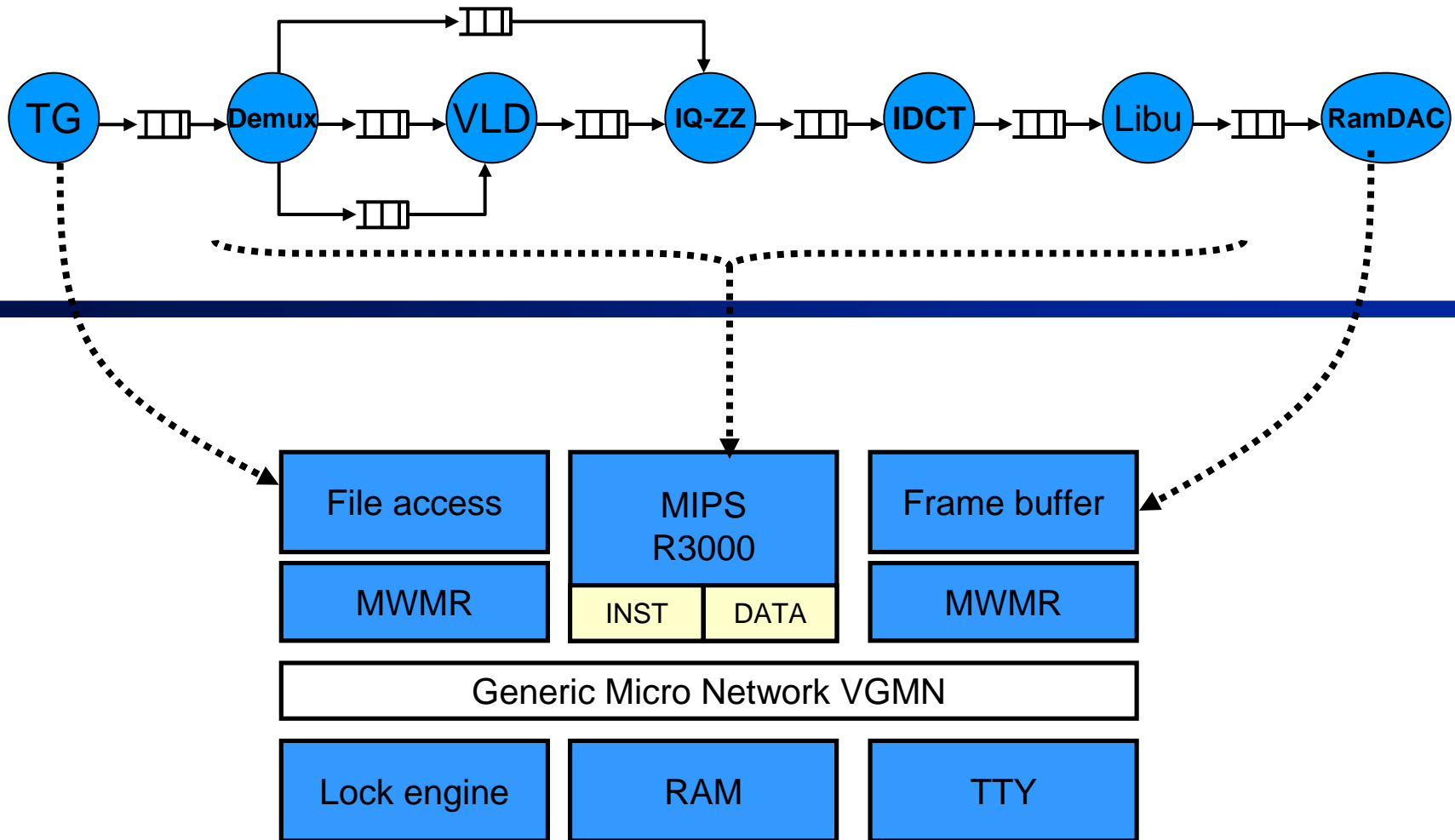


[www.soclib.fr](http://www.soclib.fr)

# SoCLib: Design flow

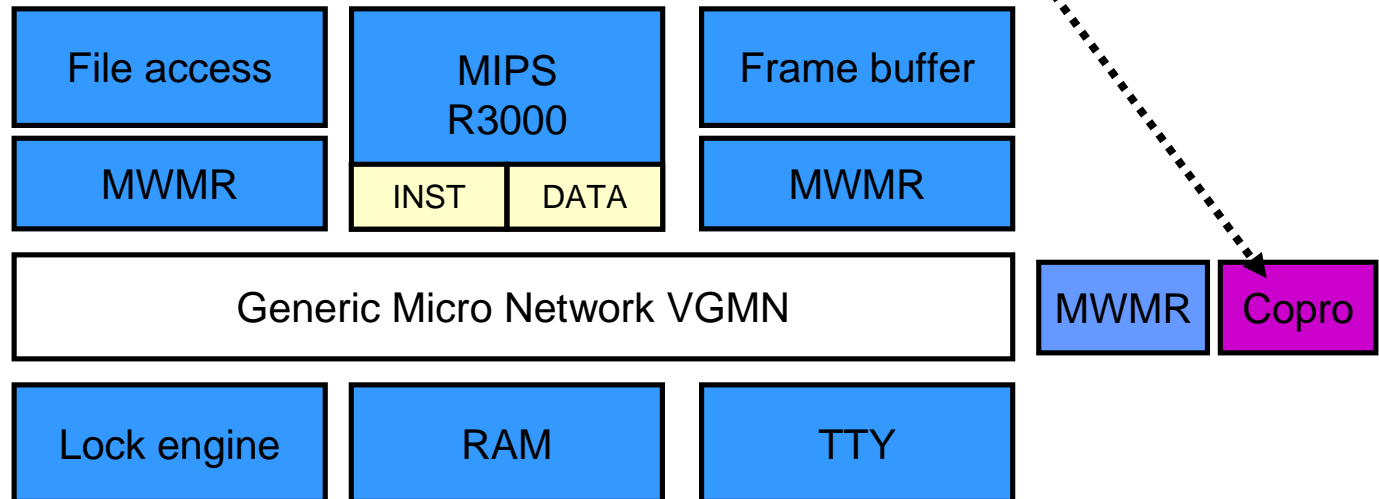
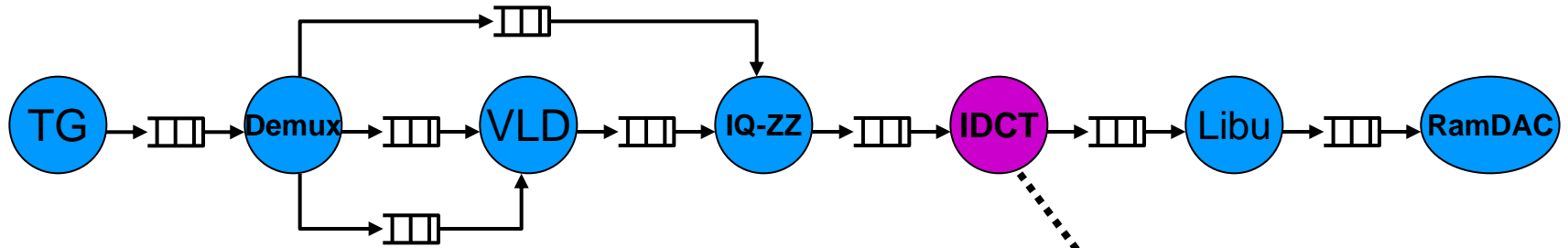


# Experiments: architecture #1



Pure software implementation on a mono-processor architecture

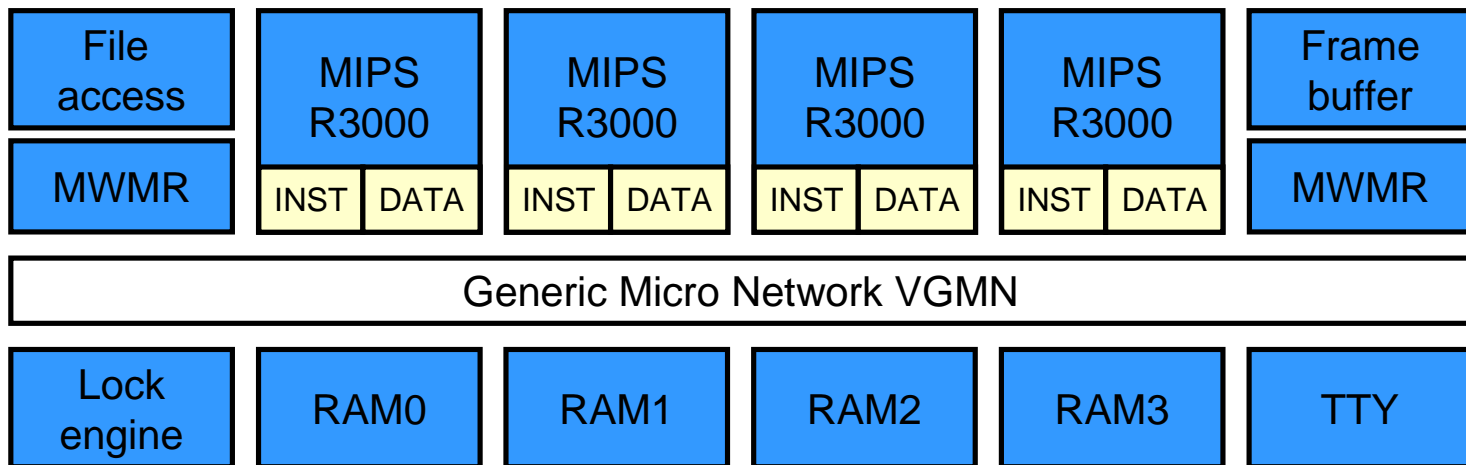
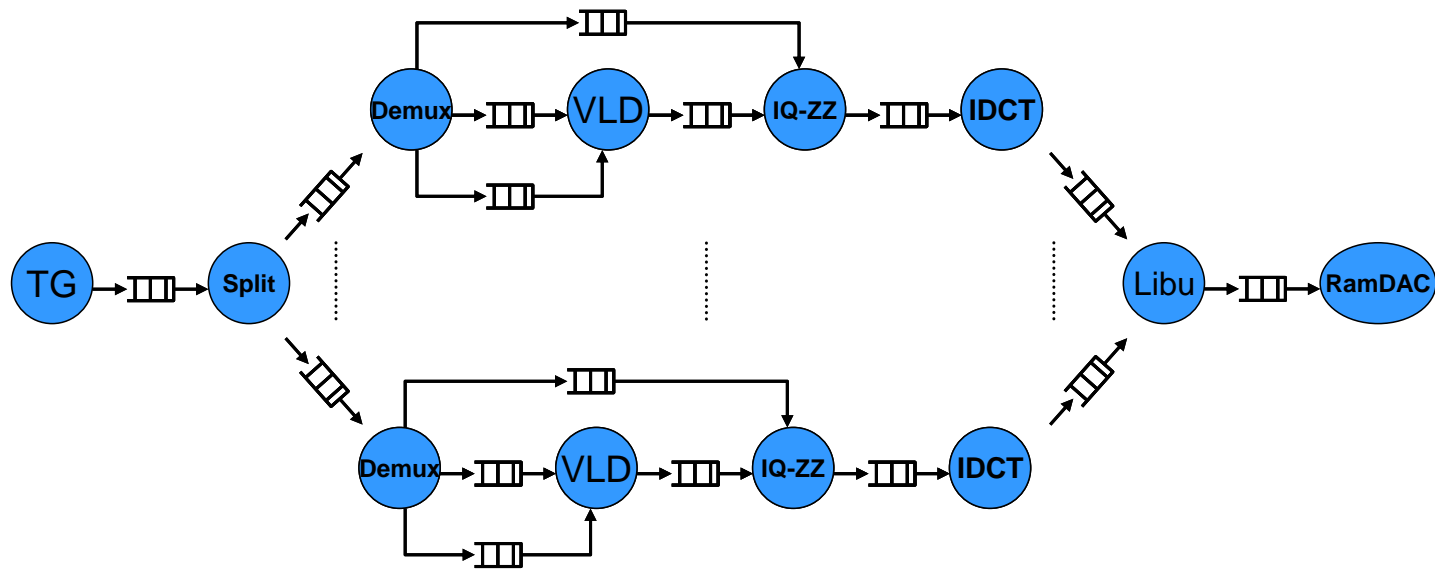
# Experiments: architecture #2



Software implementation on a mono-processor architecture  
+ IDCT as HW accelerator

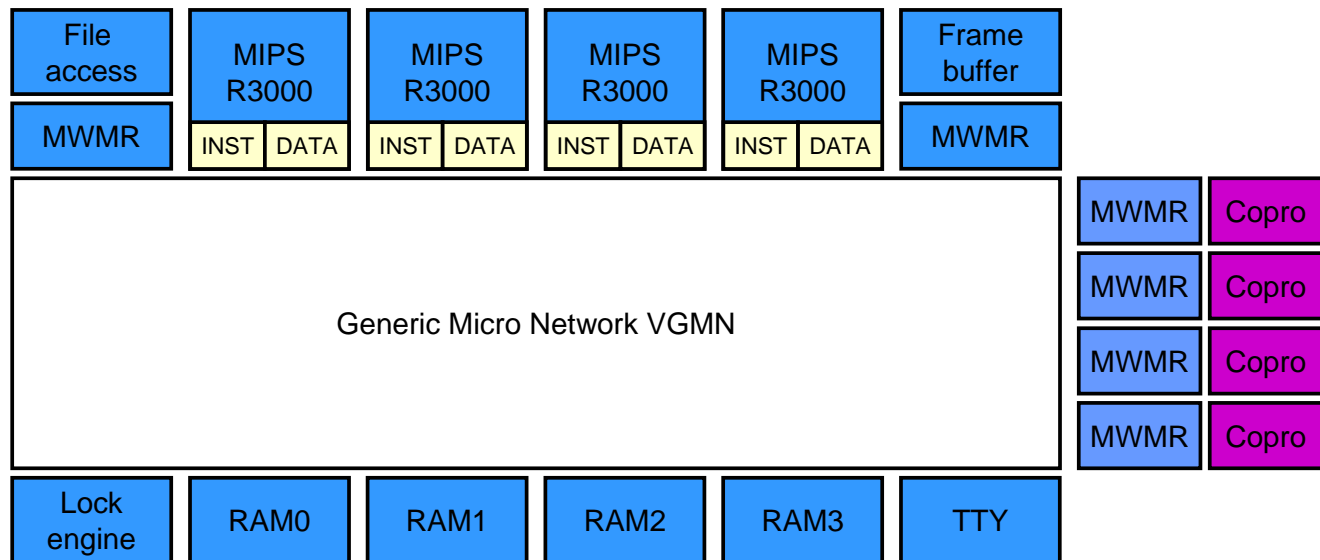
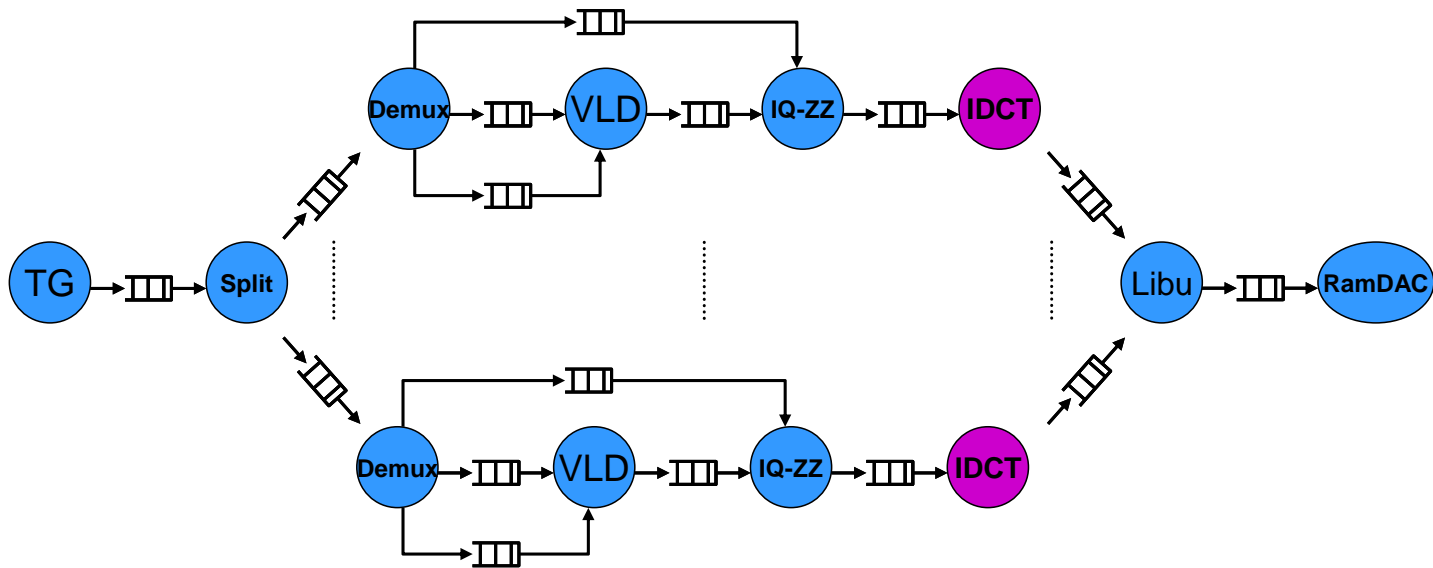


# Experiments: architecture #3

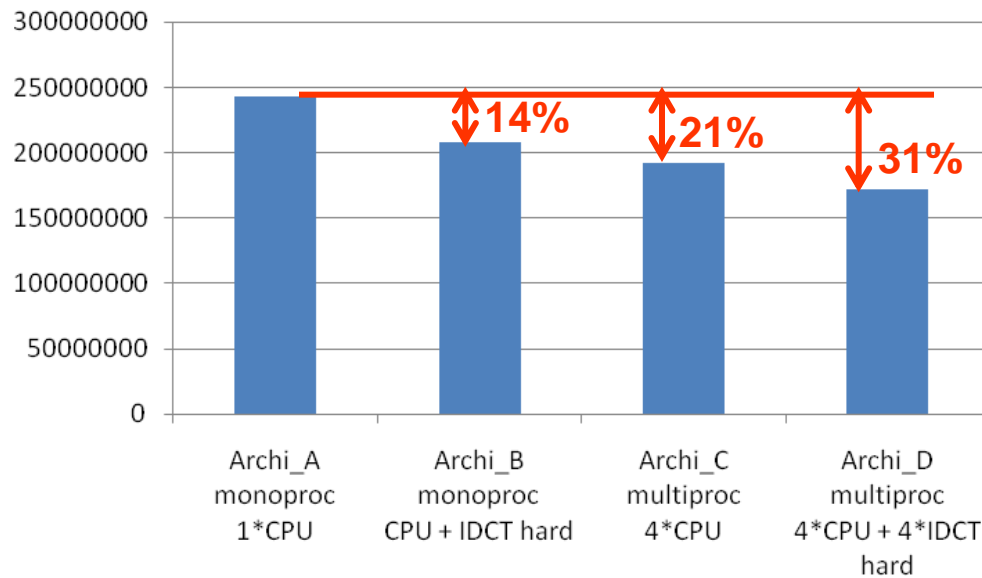


Parallelized software implementation on a multiprocessor architecture

# Experiments: architecture #4



# MJPEG Results

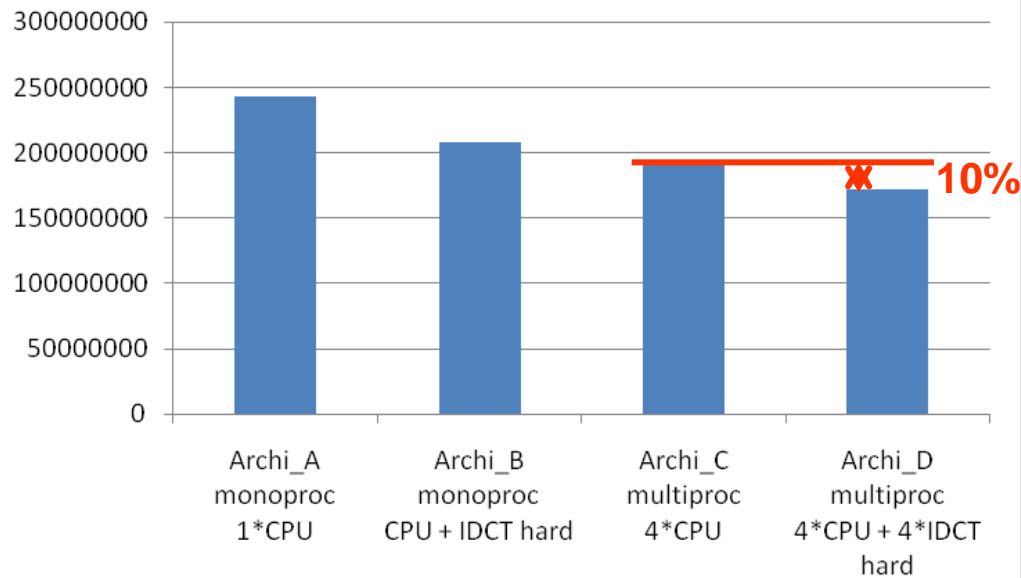


IDCT generated by GAUT reduces the application latency by 14%

Parallelization of the application on 4 CPUs reduces the latency by 21%

Execution time of the application (in cycles)  
to process 50 images of 48\*48 pixels

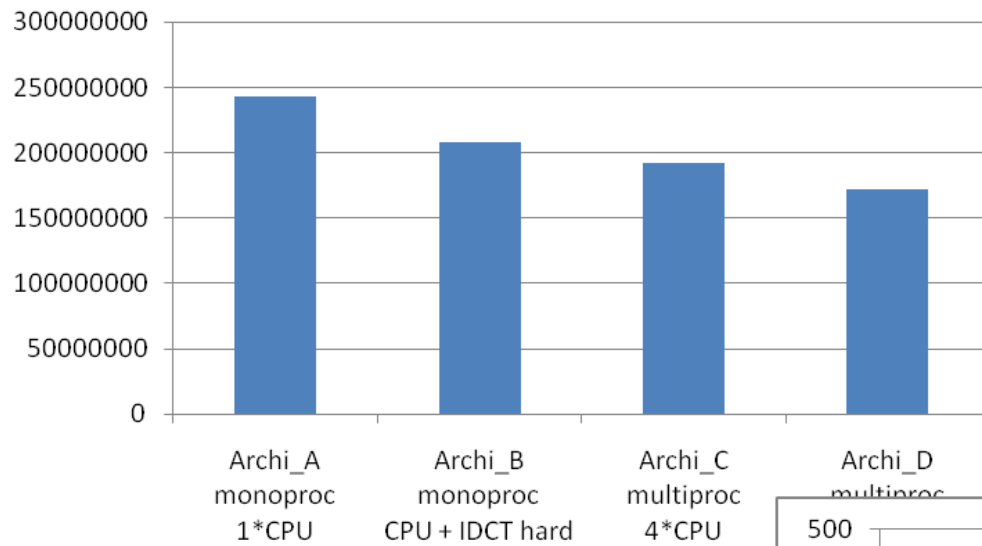
# MJPEG Results



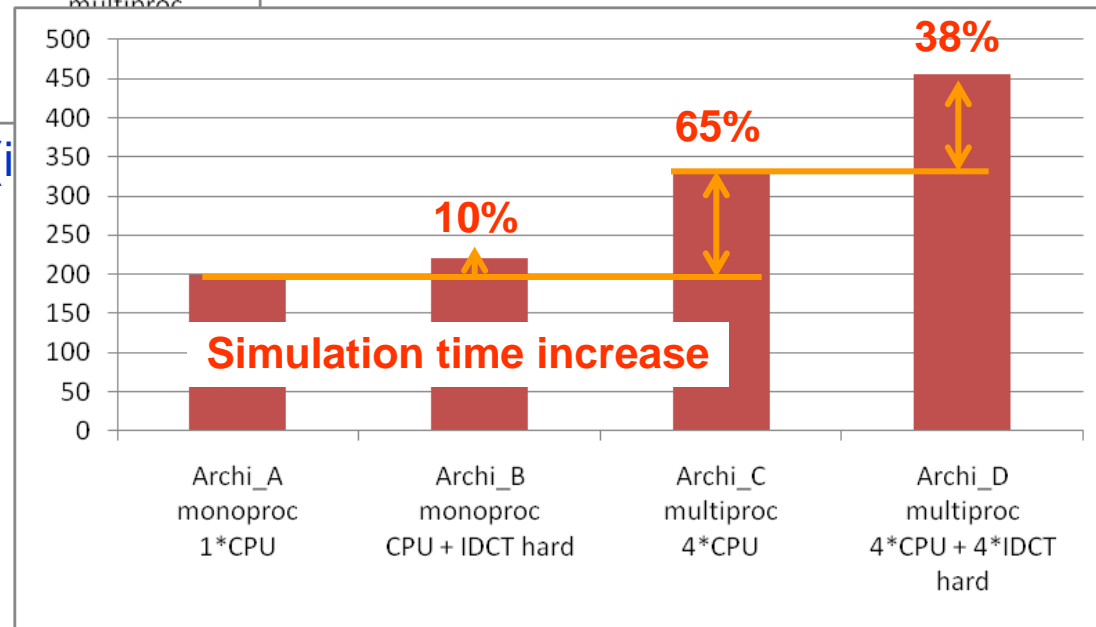
The 4 HW IDCT in the multiprocessor architecture further reduce the latency by 10%

Execution time of the application (in cycles)  
to process 50 images of 48\*48 pixels

# MJPEG Results



Execution time of the application (i to process 50 images of 48\*48



Simulation time (in secondes)

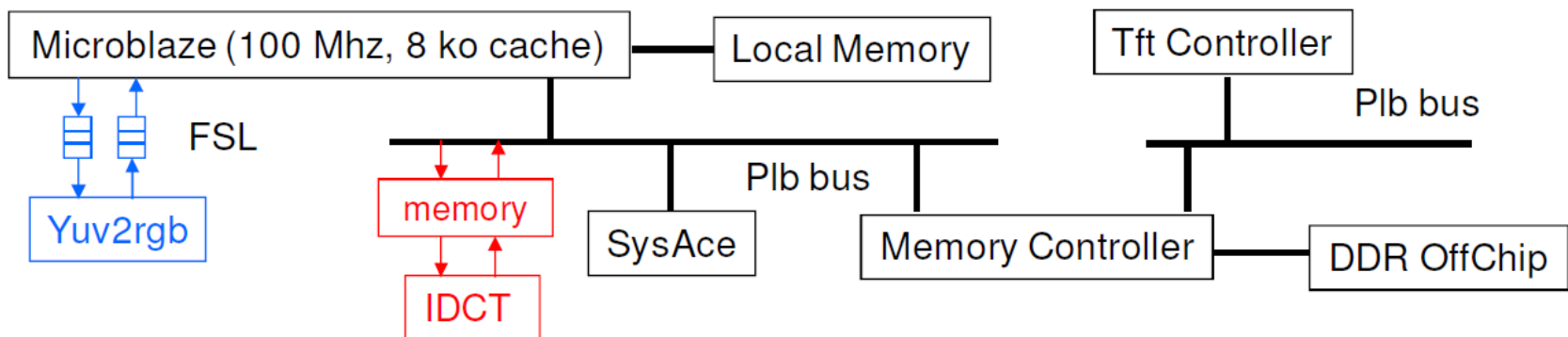
# MJPEG: Hardware prototyping

## □ Real time decoding: 24 QCIF images/sec

- IDCT: maximum I/O bandwidth (4 parallel input ports) and the lower latency (33 cycles, Freq. 138,9Mhz)
- YUV2RGB: minimum latency (12 cycles, Freq. 249,18Mhz)

## □ Compared to a pure SW implementation

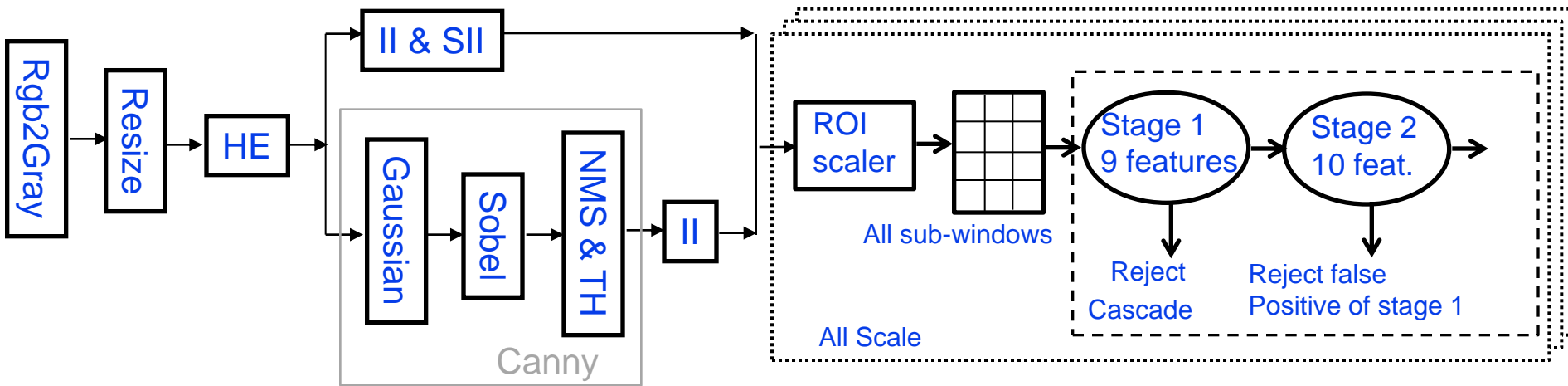
- 10x speed-up for the IDCT function
- 5x speed-up for the yuv2rgb function



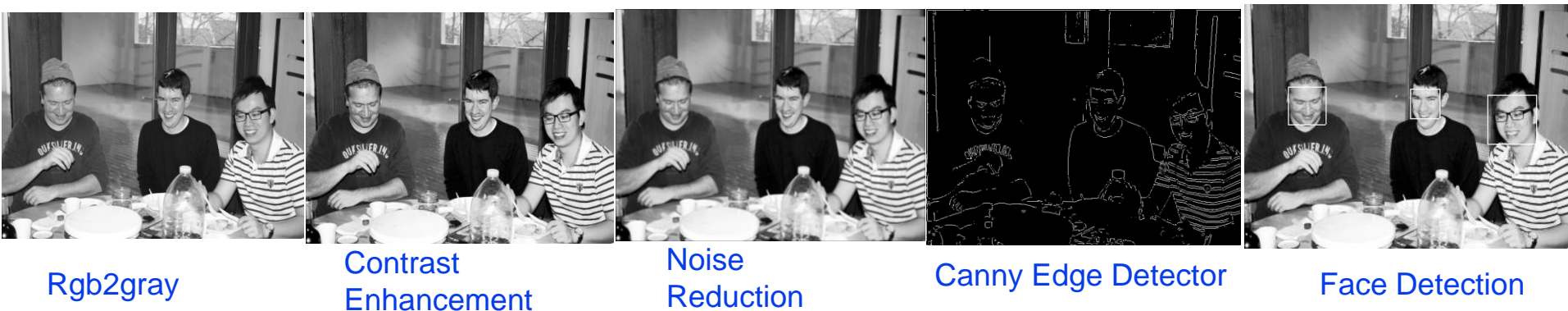
**SoC design on a FPGA Xilinx Virtex 5 LX110 (XUPV5) board**

# Viola Jones: Hardware prototyping

## ■ Block Diagram of a Viola Jones Face detector



## ■ 7x speed-up compared to a pure sw implementation



# HLS for Hardware prototyping

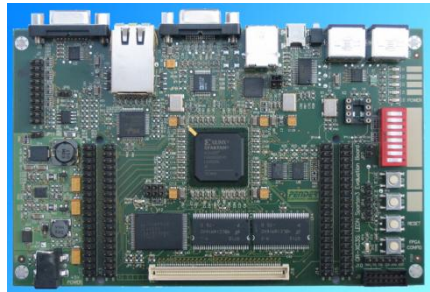
Slope detection : acos (cordic) hwpu

Texture detection: gaussian filter and square root hwpu

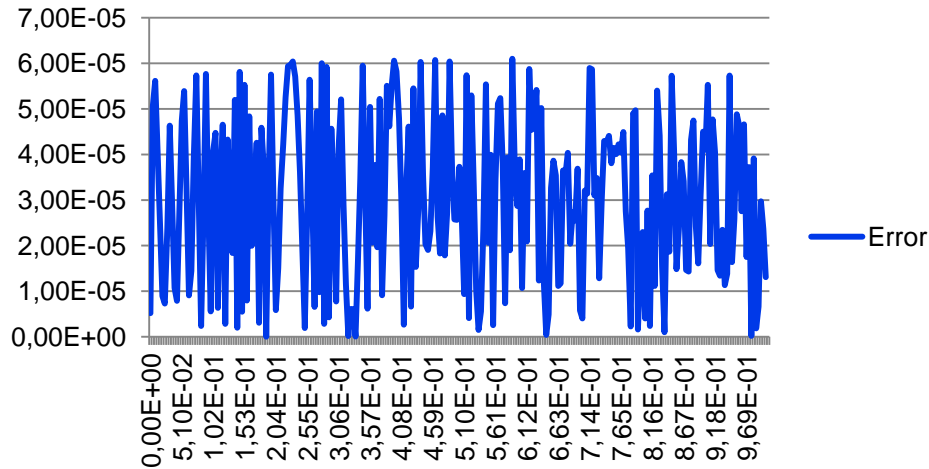
Soc Leon3 interface (AHB, Glib)

SpeedUp >= 140

Error <= 0.00006



**Error**



$$2^{-n} \approx 0.00001 \leq \text{error} \leq 2^{-(n-2)} \approx 0.00006$$

n = 16, number of rotation

**SpeedUp**



```

/* __ieee754_acos(x)
 * For |x|<=0.5 acos(x) = pi/2 - (x + x*x^2*R(x^2))
 * where
 * R(x^2) is a rational approximation
 * of (asin(x)-x)/x^3
 * For x>0.5
 * acos(x) = pi/2 - (pi/2 - 2*asin(sqrt((1-x)/2)))
 */
    
```



# Prototyping platform

---

## Sundance platform

Mother board

Daughter boards

*DSP C62 C67 (Texas Instrument)*

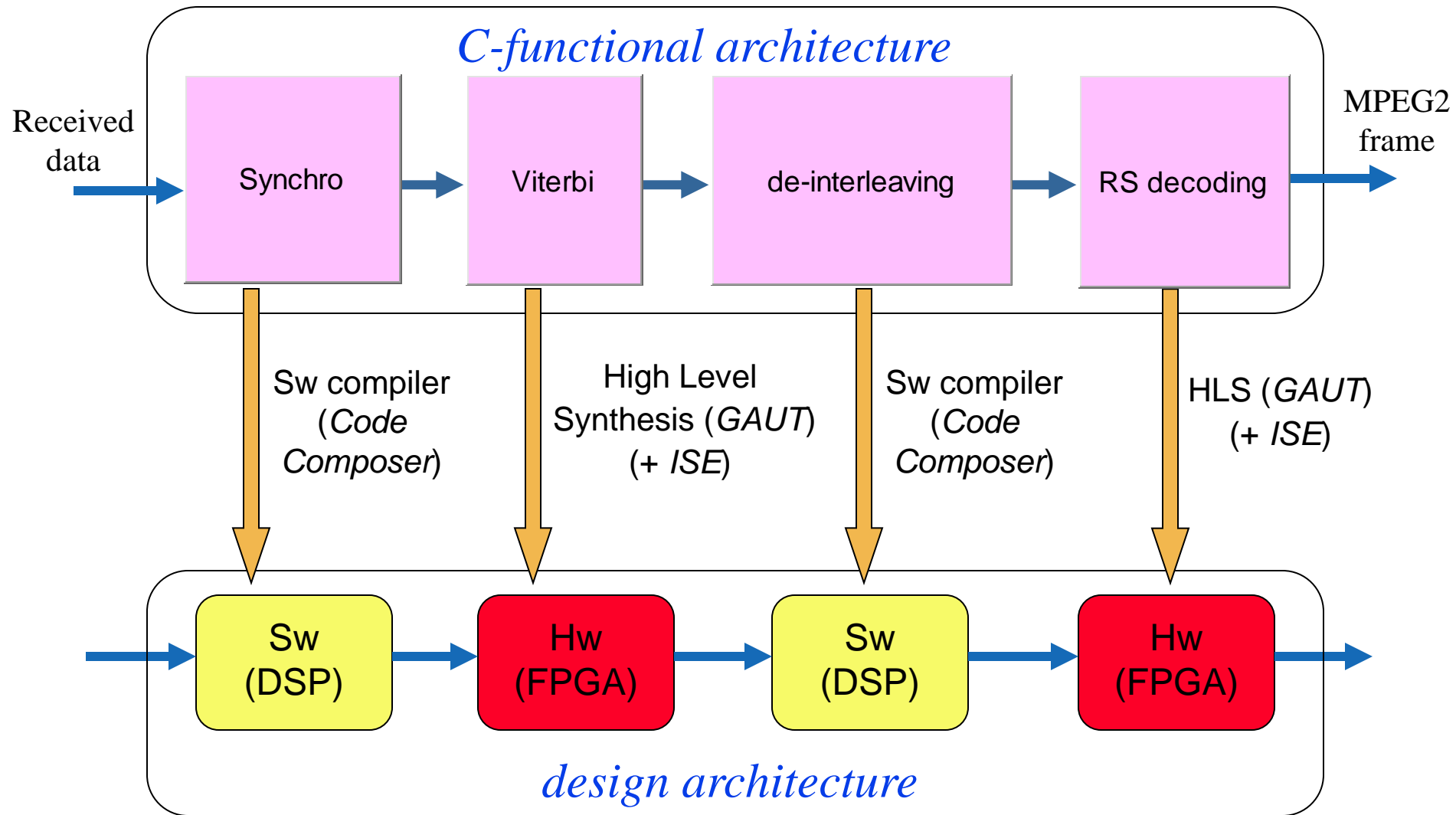
*FPGA Virtex 1000E (Xilinx)*

Interconnection matrix

*Point to point links : Com Port (CP, up to 20 Mbytes/sec) and Sundance Digital Bus (SDB, up to 200 Mbytes/sec)*



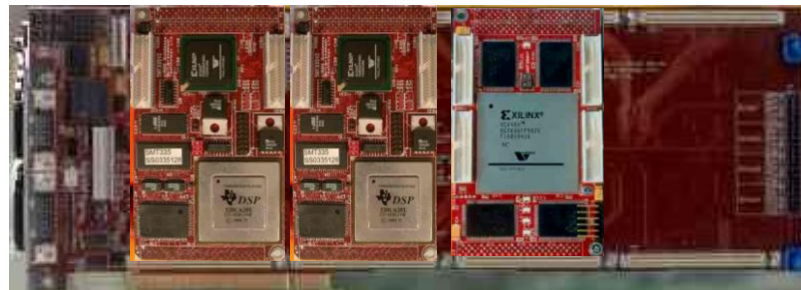
# DVB-DSNG receiver architecture mapping



# DVB-DSNG receiver

---

- Synchronization and interleaving : Sw : C62 DSP
- Viterbi and Reed Solomon decoders : Hw : Virtex-1000E FPGA
- 4 SDB links
- 26 Mbps throughput (limited by the synchronization bloc...C64 for higher throughputs)

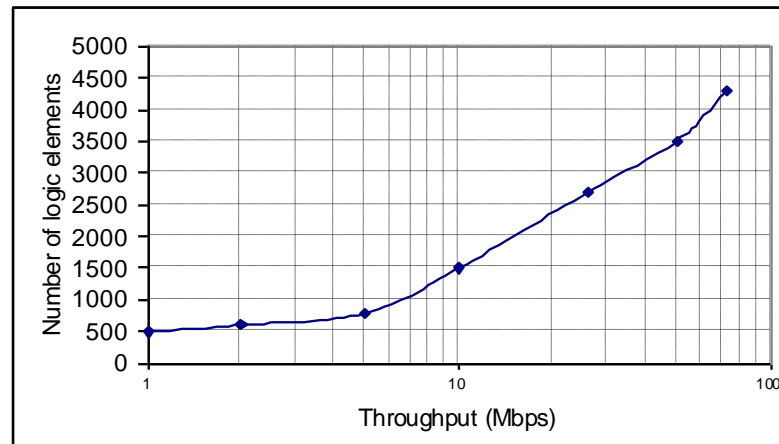


# Viterbi decoding

- functional/application parameters : state number, throughput

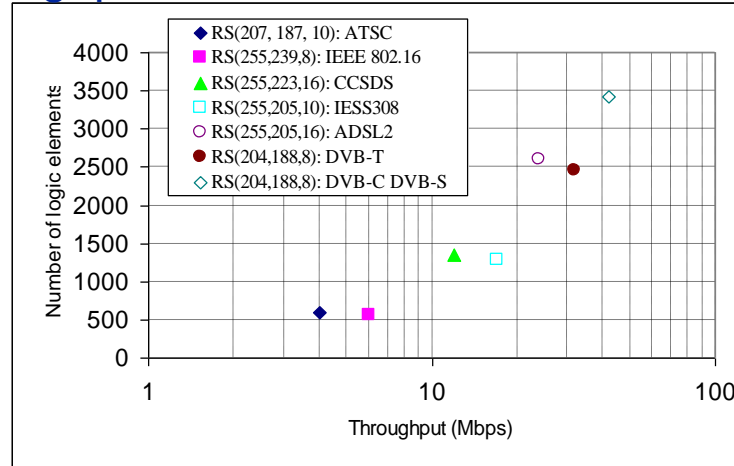
State Number	8	16	32	64	128
Throughput (Mbps)	44	39	35	26	22
Synthesis Time (s)	1	1	3	9	27
<i>Number of logic elements</i>	223	434	1130	2712	7051

- DVB-DSNG standard : throughput : 1.5 to 72 Mbps, 64 states Viterbi decoder

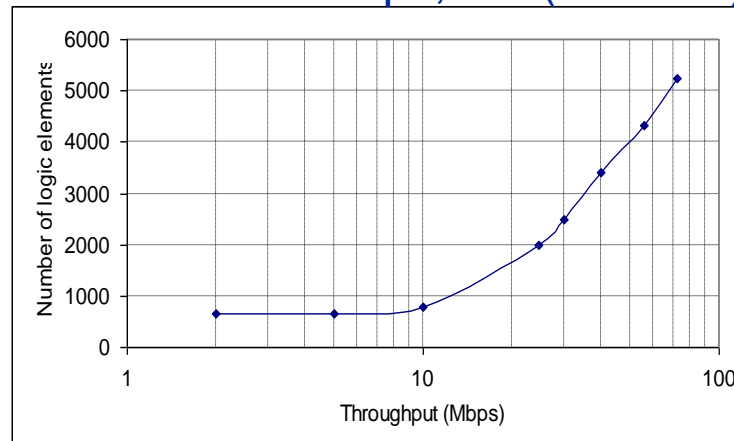


# Reed Solomon decoding

- functional/application parameters : number of input symbols, data symbols, throughput

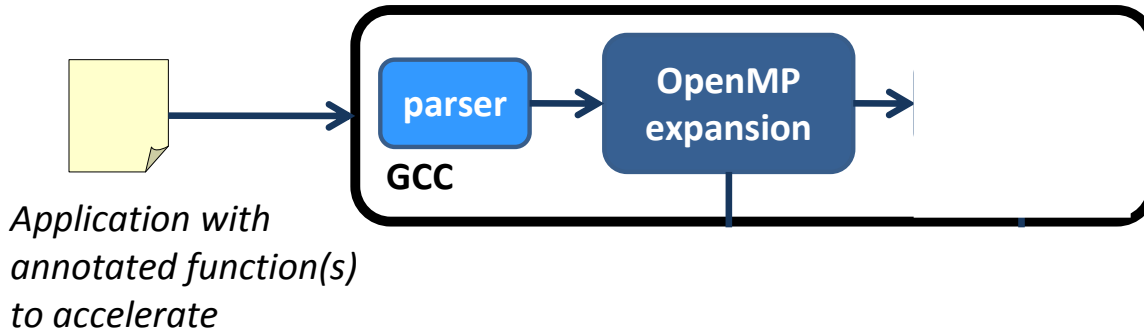


- DVB-DSNG standard : 1.5 to 72 Mbps, RS (204/188) decoder

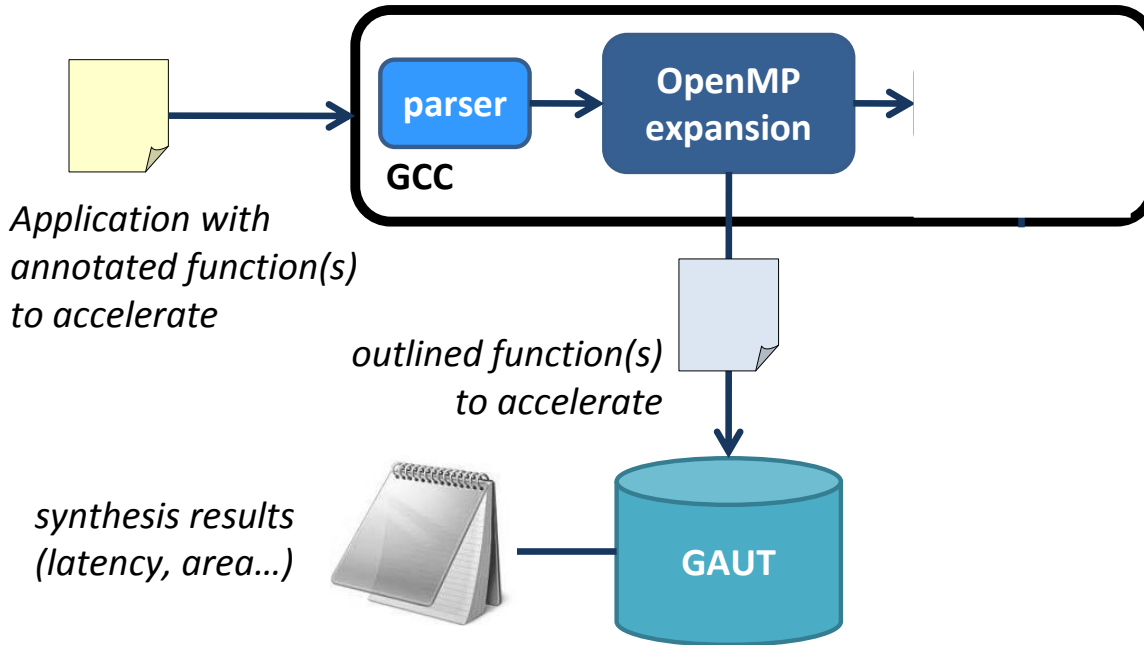


# OpenMP/HLS & manycores/HWPU

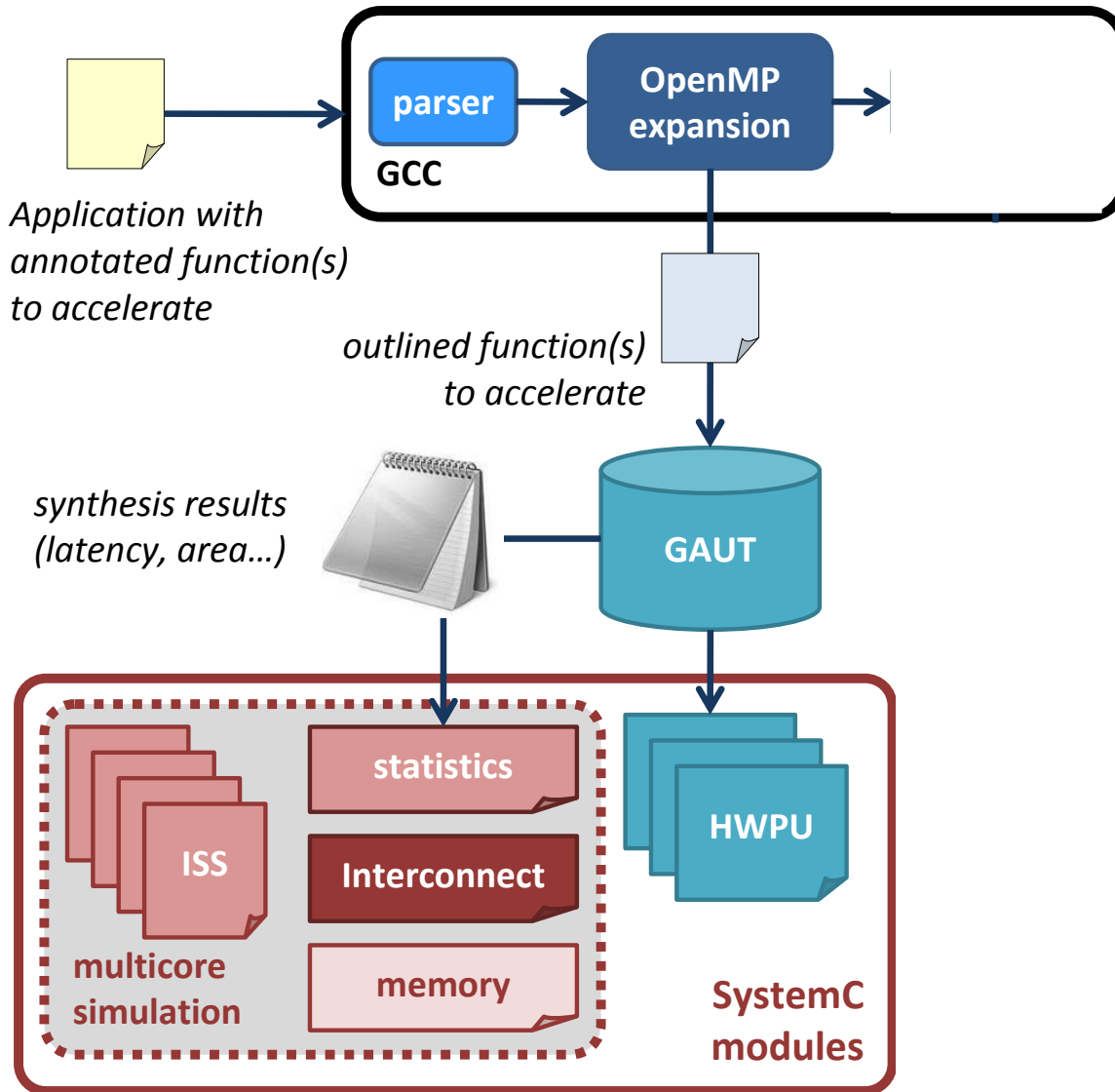
---



# OpenMP/HLS & manycores/HWPU

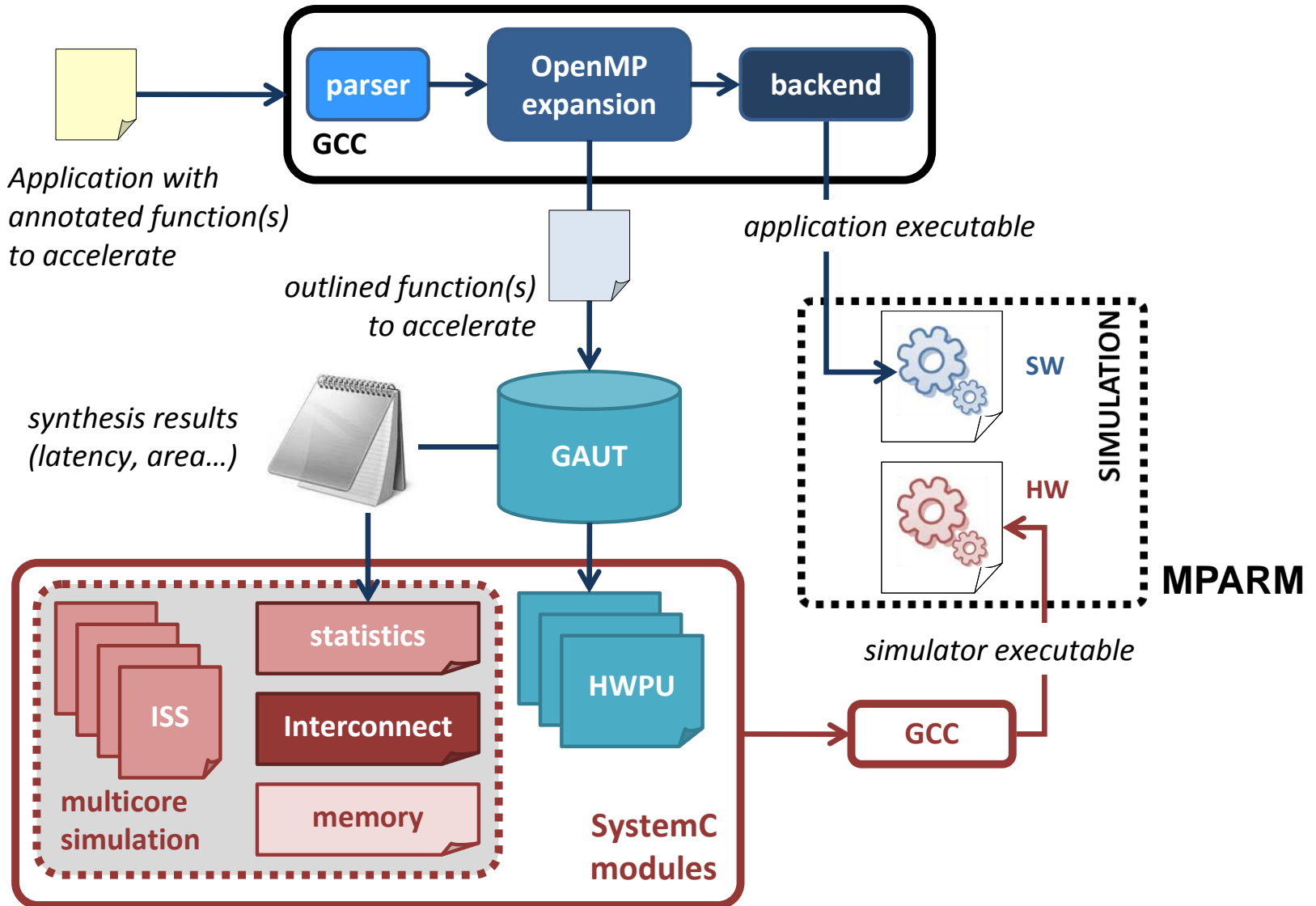


# OpenMP/HLS & manycores/HWPU

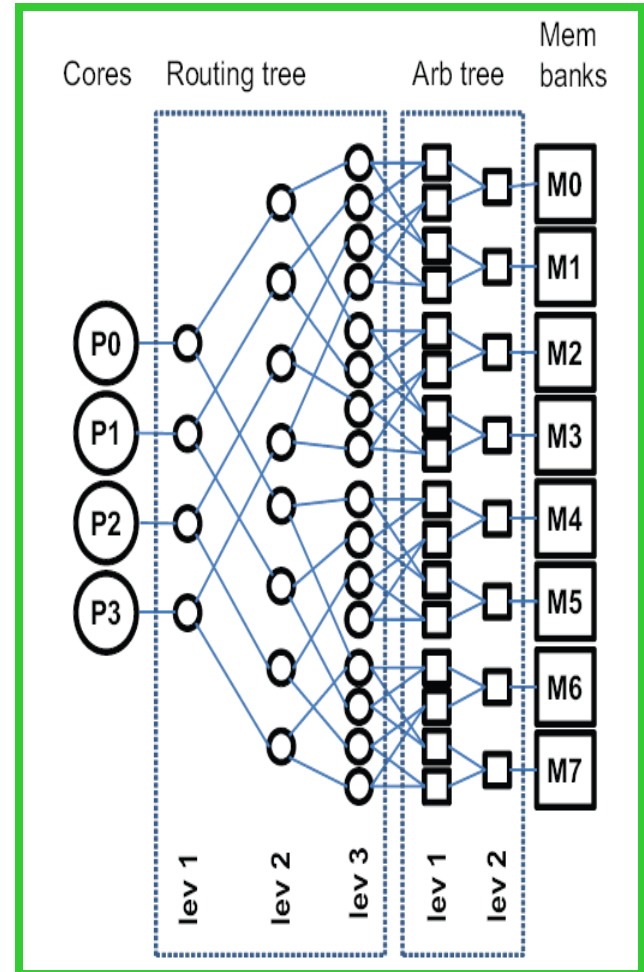
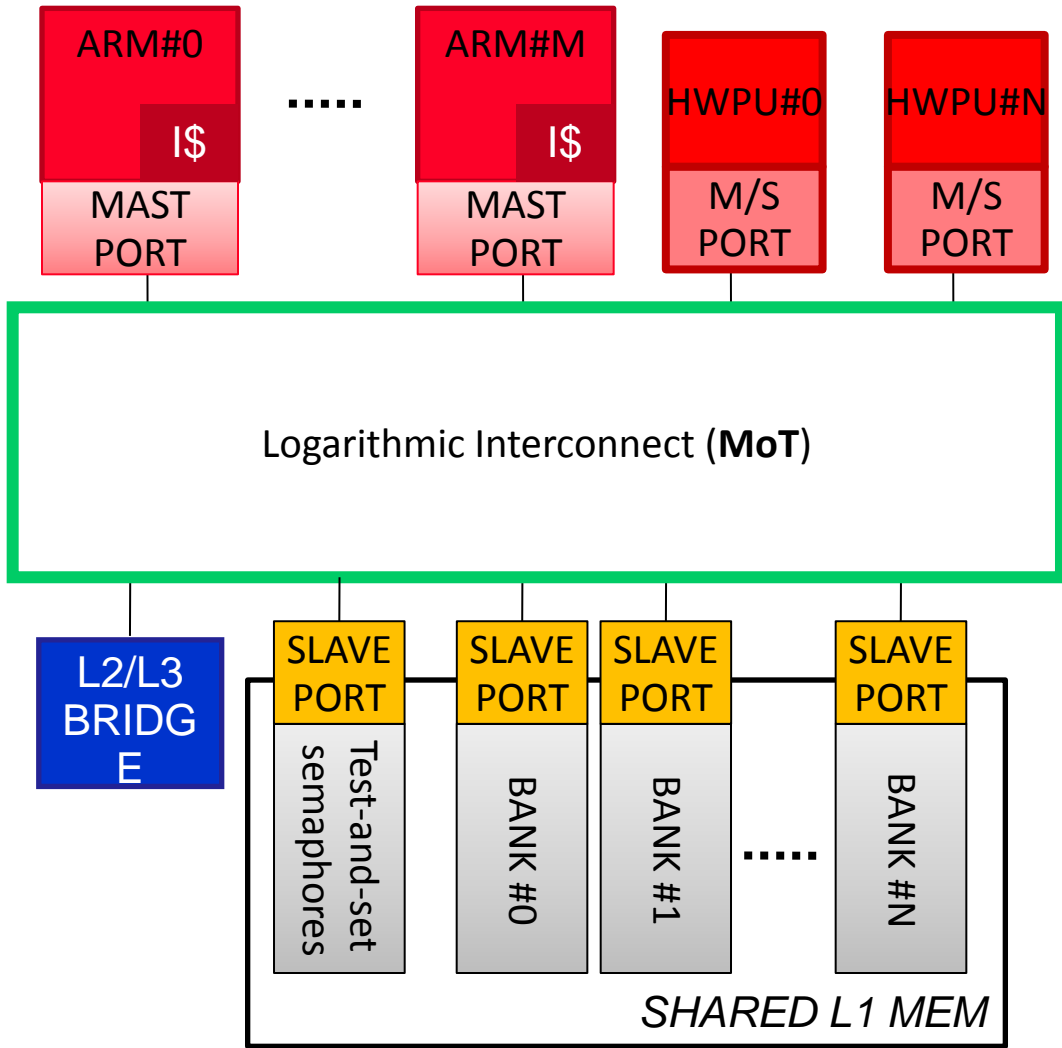




# OpenMP/HLS & manycores/HWPU

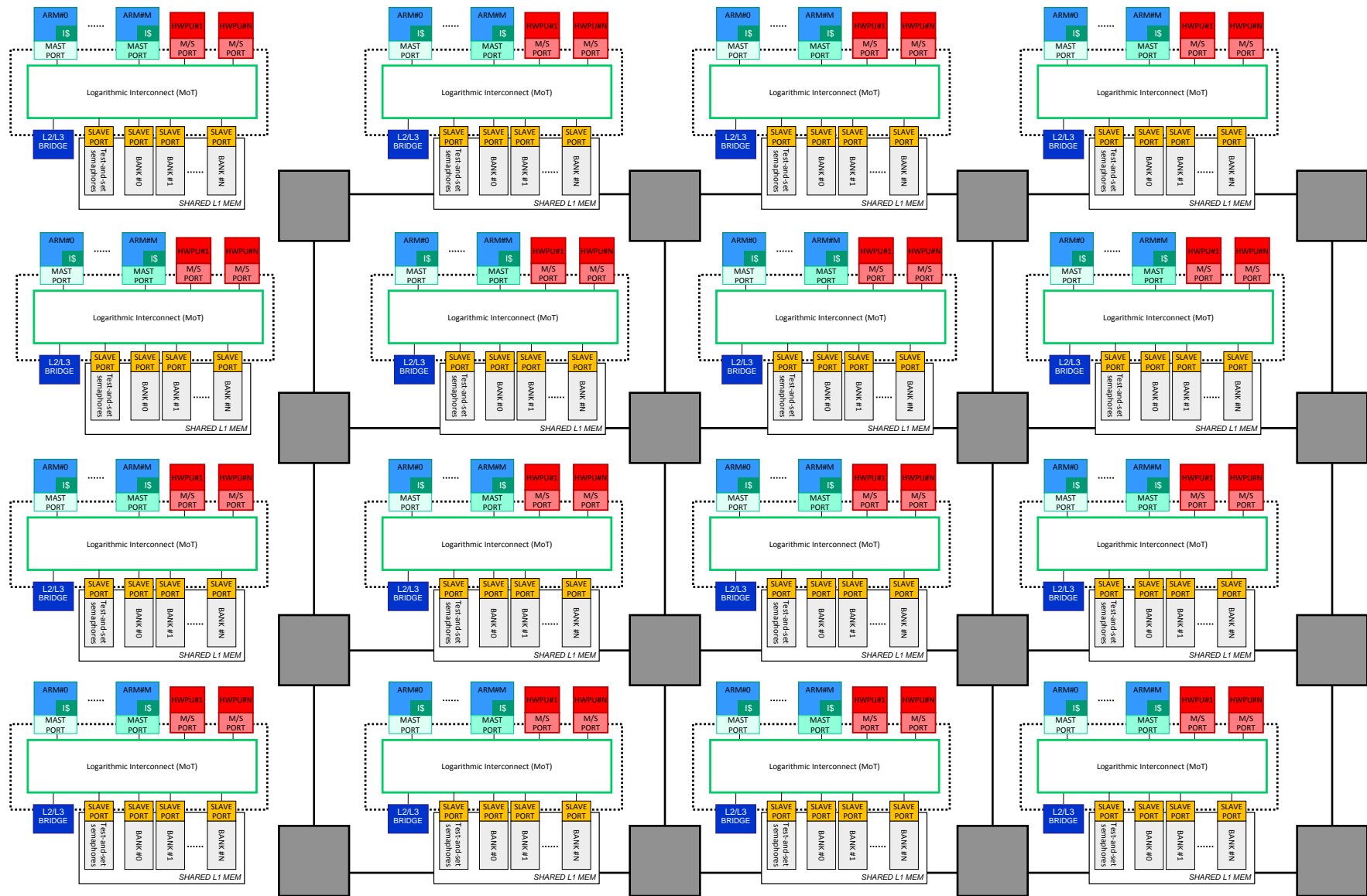


# MPARM Architecture



Mesh Of Tree

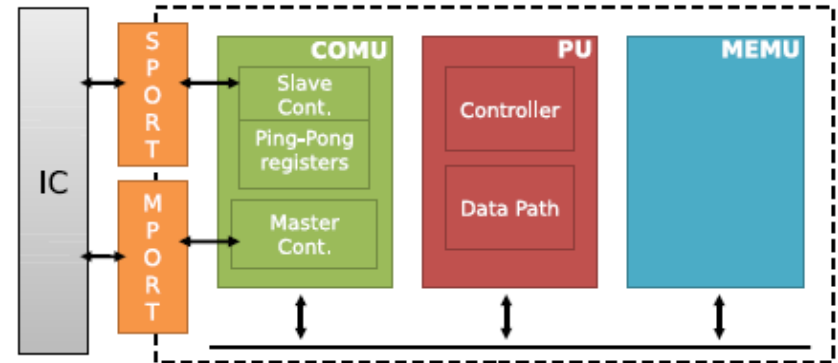
# Target architecture



# HWPU Integration

## □ Interface de communication

- Maître / Esclave
- Registres de configuration
  - *Nombre d'entrées*
  - *Nombre de sorties*
  - *Emplacements des entrées*
  - *Emplacements des sorties*
  - *Mode*
  - *Etat du HWPU*
  - *Démarrage*
- Les registres de configuration peuvent être doublés
  - *Recouvrement de la configuration et du calcul*



## □ Interface de programmation

Function name	Brief description
<code>bool acc_busy ()</code>	Returns TRUE if no programming channel is available
<code>void acc_reset ()</code>	Once a channel has been granted resets programming registers
<code>void acc_set_input_count (int count)</code>	Sets number of inputs
<code>void acc_set_output_count (int count)</code>	Sets number of outputs
<code>void acc_set_in_addr (int addr)</code>	Sets current input parameter's address
<code>void acc_set_out_addr (int addr)</code>	Sets current input parameter's address
<code>void acc_trigger ()</code>	Initiates execution
<code>void acc_wait ()</code>	Waits for the HWPU to complete execution

# Example


---

```
void foo()  
{  
    int A, B, C;  
  
    #pragma omp accelerate input(A, B) output(C)  
    C = A + B;  
}
```



# Example

```
void foo()  
{  
    int A, B, C;  
  
    #pragma omp accelerate input(A, B) output(C)  
    C = A + B;  
}
```



**Compilation**

```
void newfunc_acc_0 (const int * in1, const int * in2,  
                   int * out1)  
{ *out1 = *in1 + *in2; }
```

# Example

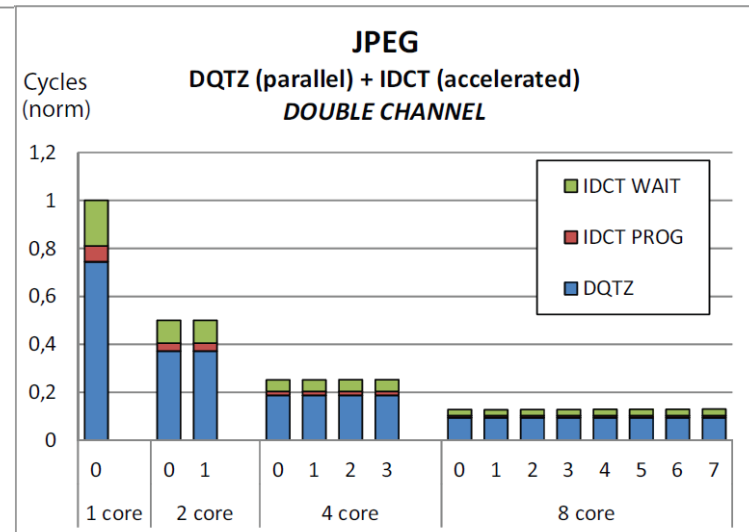
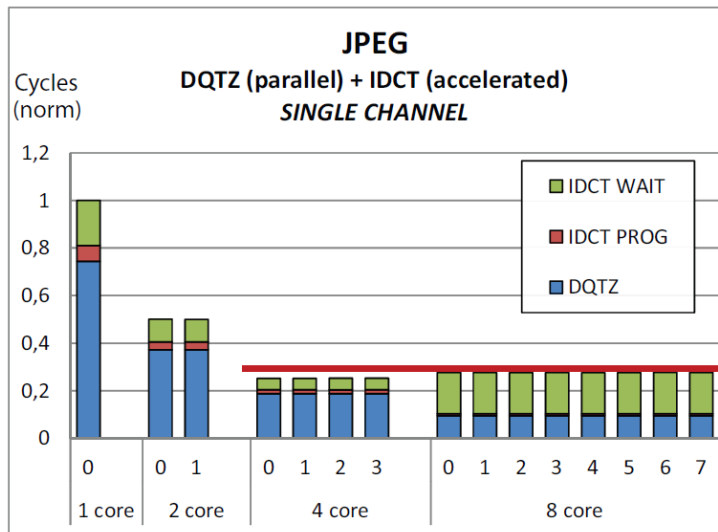
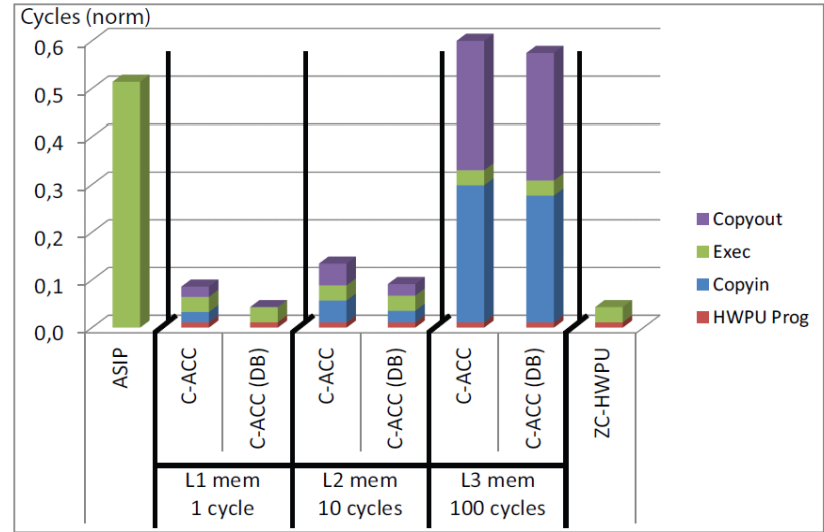
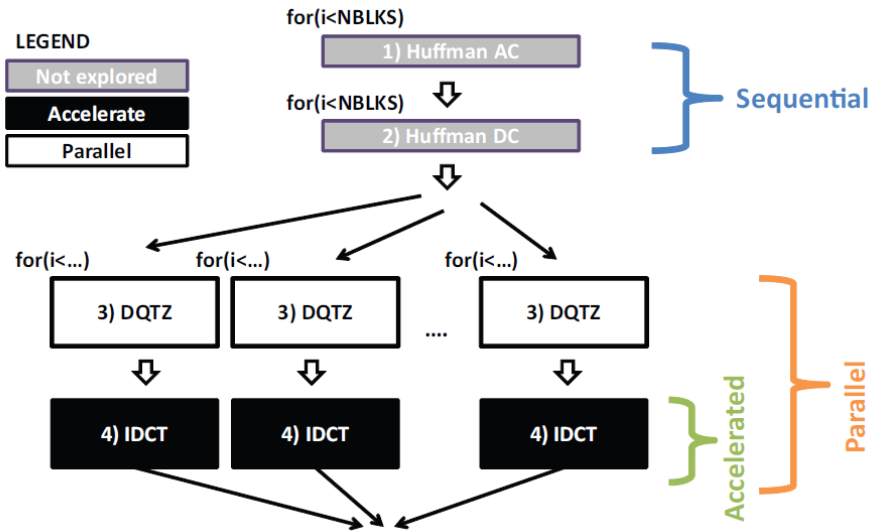
```
void foo()  
{  
    int A, B, C;  
  
    #pragma omp accelerate input(A, B) output(C)  
    C = A + B;  
}
```

Compilation

```
void newfunc_acc_0 (const int * in1, const int * in2,  
                   int * out1)  
{ *out1 = *in1 + *in2; }
```

```
void foo() {  
    int A, B, C;  
    int *in_addr, *out_addr;  
  
    in_addr = {&A, &B};  
    out_addr = {&C};  
    while (!omp_program_hwp_u (2, 1,  
                               in_addr, out_addr));  
    omp_acc_wait (); }  
→  
→  
→  
→
```

# Results





# GAUT 4 (not yet available, but soon...)

---

## **An open source HLS tool**

- For both data and control-dominated algorithms (CDFG)

## **Input :**

- C/C++ bit-accurate integer and fixed-points from Mentor Graphics
- SystemC : C and C++ lack the constructs and semantics to represent design hierarchy, timing, synchronization/concurrency
- Floating point

## **Output : RTL Architecture**

- VHDL , Verilog
- SystemC (CABA + TLM)
- Resource and timing estimation

## **Automated Test-bench generation**

## **Automated operators characterization**

## **Automated interface generation**

- AXI, AHB, FSL, ...

# GAUT 4 (not yet available, but soon...)

## □ Constraints

- Clock, I/O protocols, loop transformations (unrolling, merging, loop pipelining with Initiation Interval), memory mapping, function inlining, resource constraints

## □ Objectives

- Minimization: area i.e. resources, latency, power consumption...
- Maximization: throughput

## □ Keys features

- *Used robust and state of the art compilation technology to extract instruction-level (Vectorization) and loop level parallelism (Polyhedral model: graphite for GCC, Polly for LLVM)*
- *Many scheduling strategies : modulo scheduling (SMS,IMS) , Force Directed List Scheduling (FDLS), System of difference constraint (SDC)...*
- *Memory analysis and optimizations: automatic partitioning of array elements to reduce conflicts and increase throughput*
- *Pattern mining for efficient resource sharing*
- *Hierarchy synthesis and function level parallelism/pipelining*
- *Design Space Exploration with directives (Loop transformation, memory partitioning) and constraints (script): one body of code, many hardware outcomes*

# Conclusion

---

- **HLS allows to automatically generate several RTL architectures**
  - From an algorithmic/behavioral description and a set of constraints
  
- **HLS allows to generate**
  - VHDL models for synthesis purpose
  - SystemC simulation models for virtual prototyping
  
- **HLS allows to explore the design space of**
  - Hardware accelerators
  - MPSoC architectures including HW accelerators
  
- **GAUT is free downloadable at**
  - <http://lab-sticc.fr/www-gaut>

# References

---



---

## High Level Synthesis of ASICs Under Timing and Synchronization Constraints

---

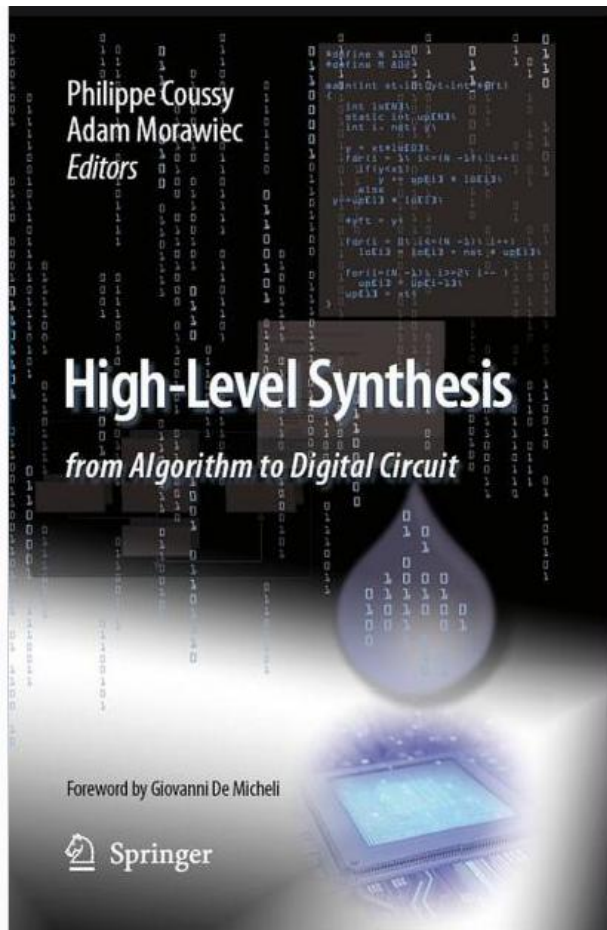
David C. Ku  
Giovanni De Micheli



---

Kluwer Academic Publishers

# References



# Academic tools

---

- Streamroller (Univ. Mich.)**
- SPARK (UCSD)**
- xPilot (UCLA)**
- UGH (TIMA+LIP6)**
- MMALPHA (IRISA+CITI+...)**
- ROCCC (UC Riverside)**
- GAUT (UBS / Lab-STICC)**
- ...**

# Commercial tools

---

- ❑ **CatapultC (Mentor Graphics => Calypto)**
- ❑ **PICO (Spin-off HP => Synfora => Synopsys)**
- ❑ **Cynthecizer (Forte design)**
- ❑ **Cyber (NEC)**
- ❑ **AutoPilot (AutoESL => Xilinx)**
- ❑ **C to Silicon (Candence)**
- ❑ **Synphony (Synopsys)**
- ❑ **...**

# ***Une introduction à la synthèse de haut-niveau***

***(ou comment générer des architectures  
matérielles à partir du langage C)***

**Université de Bretagne-Sud  
Lab-STICC**

**Philippe COUSSY**  
**[philippe.coussy@univ-ubs.fr](mailto:philippe.coussy@univ-ubs.fr)**