Architecture Synthesis for Linear Time-Invariant Filters

Antoine Martinet

CITI lab,
INRIA's SOCRATE Team,

Under the supervision of
Florent de Dinechin

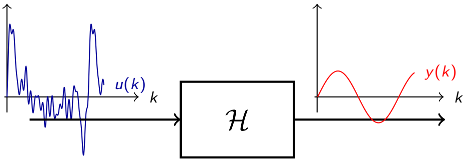
2 February - 31 July, 2015

Table of Contents

Table of Contents

Introduction: LTI Filters

Signal: temporal variable $x(k)$ with $\{x(k)\}_{k \geq 0} \in \mathbb{R}$



LTI (Linear Time Invariant) filters are particular filters that are:

- Linear: outputs are linear combinations of inputs (allows to use linear algebra definitions)
- Time-Invariant: all coefficients are constant

Purpose of this work

- Lopez's PhD thesis studies how to compute LTI filters in software:
  - scheduling issues
  - fixed size
- The goal here is to do such a work in hardware, where we have more flexibility:
  - full parallelism
  - arbitrary size

In this context, constraints become degrees of freedom.

FIR, IIR

The most simple LTI Filter: the FIR (Finite Impulse Response) filter

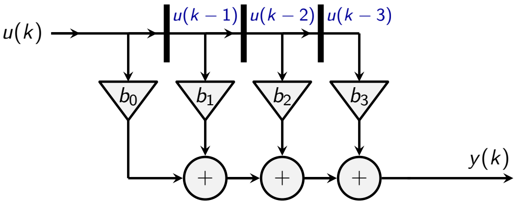
$$y(k) = \sum_{i=0}^{n} b_i u(k-i) \tag{1}$$

FIR, IIR

The most simple LTI Filter: the FIR (Finite Impulse Response) filter

$$y(k) = \sum_{i=0}^{n} b_i u(k-i) \tag{1}$$

Abstract architecture for the direct form realization of a FIR filter:

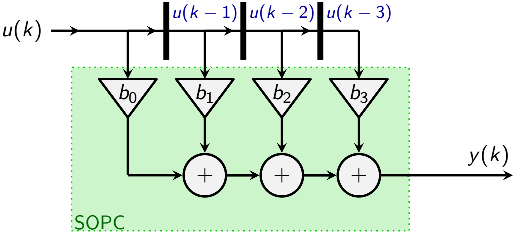


FIR, IIR

The most simple LTI Filter: the FIR (Finite Impulse Response) filter

$$y(k) = \sum_{i=0}^{n} b_i u(k-i) \quad (1)$$

Abstract architecture for the direct form realization of a FIR filter:



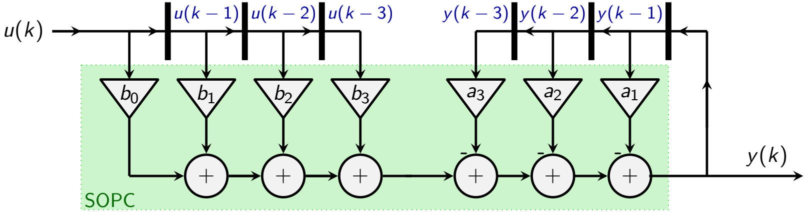
SOPC (Sum Of Products by Constants):
basic block of LTI filters implementation.

FIR, IIR

Full LTI complexity: the IIR (Infinite Impulse Response) filter:

$$y(k) = \sum_{i=0}^{n} b_i u(k-i) - \sum_{i=0}^{n} a_i y(k-i) \qquad (1)$$

Abstract architecture for the direct form realization of an IIR filter:



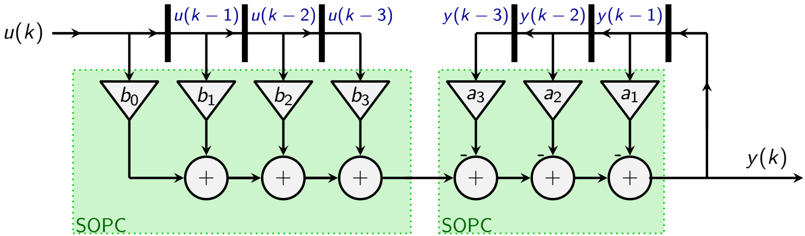
SOPC (Sum Of Products by Constants):
basic block of LTI filters implementation.

FIR, IIR

Full LTI complexity: the IIR (Infinite Impulse Response) filter:

$$y(k) = \sum_{i=0}^{n} b_i u(k-i) - \sum_{i=0}^{n} a_i y(k-i) \quad (1)$$

Abstract architecture for the direct form realization of an IIR filter:



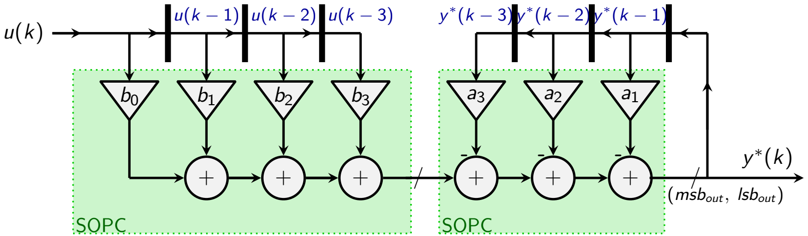
SOPC (Sum Of Products by Constants):
basic block of LTI filters implementation.

FIR, IIR

Full LTI complexity: the IIR (Infinite Impulse Response) filter:

$$y(k) = \sum_{i=0}^{n} b_i u(k-i) - \sum_{i=0}^{n} a_i y(k-i) \quad (1)$$

Abstract architecture for the direct form realization of an IIR filter:



SOPC (Sum Of Products by Constants):
basic block of LTI filters implementation.

State-Space representation

The "ABCD" form

Let's define $\boldsymbol{x}(k)$ a state vector (hardware register)

$$\begin{cases} \boldsymbol{x}(k+1) = \boldsymbol{Ax}(k) + \boldsymbol{Bu}(k) \\ \boldsymbol{y}(k+1) = \boldsymbol{Cx}(k) + \boldsymbol{Du}(k) \end{cases} \tag{2}$$

With:

$$\boldsymbol{A} \in \mathbb{R}^{n_x \times n_x}, \ \boldsymbol{B} \in \mathbb{R}^{n_x \times n_u}, \\ \boldsymbol{C} \in \mathbb{R}^{n_y \times n_x}, \ \boldsymbol{D} \in \mathbb{R}^{n_y \times n_u}$$

Equivalent matrix formulation:

$$\begin{pmatrix} \boldsymbol{x}(k+1) \\ \boldsymbol{y}(k+1) \end{pmatrix} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}(k) \\ \boldsymbol{u}(k) \end{pmatrix} \tag{3}$$

The SIF: a unified realization representation

Problems with ABCD:

- this doesn't gives an explicit order in the operations
- the whole analysis on precisions has to be rebuilt for each new realization.

The SIF (Speciallized Implicit Form) generalizes the state-space.
Addition: $\boldsymbol{t}(k)$ describes explicit intermediate computations:

$$\begin{pmatrix} \boldsymbol{J} & \boldsymbol{0} & \boldsymbol{0} \\ -\boldsymbol{K} & \boldsymbol{I}_{n_x} & \boldsymbol{0} \\ -\boldsymbol{L} & \boldsymbol{0} & \boldsymbol{I}_{n_y} \end{pmatrix} \begin{pmatrix} \boldsymbol{t}(k+1) \\ \boldsymbol{x}(k+1) \\ \boldsymbol{y}(k) \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{M} & \boldsymbol{N} \\ \boldsymbol{0} & \boldsymbol{P} & \boldsymbol{Q} \\ \boldsymbol{0} & \boldsymbol{R} & \boldsymbol{S} \end{pmatrix} \begin{pmatrix} \boldsymbol{t}(k) \\ \boldsymbol{x}(k) \\ \boldsymbol{u}(k) \end{pmatrix} \quad (4)$$

The SIF as an algorithm

for *int i = 0 ; i ≤ n_t; i++* **do**

$$\boldsymbol{t}_i(k+1) \leftarrow -\sum_{j<i} \boldsymbol{J}_{ij}\boldsymbol{t}_j(k+1) + \sum_{j=1}^{n_x} \boldsymbol{M}_{ij}\boldsymbol{x}_j(k) + \sum_{j=1}^{n_u} \boldsymbol{N}_{ij}\boldsymbol{u}_j(k)$$

end

for *int i = 0 ; i ≤ n_x; i++* **do**

$$\boldsymbol{x}_i(k+1) \leftarrow \sum_{j=1}^{n_t} \boldsymbol{K}_{ij}\boldsymbol{t}_j(k+1) + \sum_{j=1}^{n_x} \boldsymbol{P}_{ij}\boldsymbol{x}_j(k) + \sum_{j=1}^{n_u} \boldsymbol{Q}_{ij}\boldsymbol{u}_j(k)$$

end

for *int i = 0 ; i ≤ ny; i++* **do**

$$\boldsymbol{y}_i(k) \leftarrow \sum_{j=1}^{n_t} \boldsymbol{L}_{ij}\boldsymbol{t}_j(k+1) + \sum_{j=1}^{n_x} \boldsymbol{R}_{ij}\boldsymbol{x}_j(k) + \sum_{j=1}^{n_u} \boldsymbol{S}_{ij}\boldsymbol{u}_j(k)$$

end

Algorithm 1: Computation of SIF outputs from inputs

A more common notation

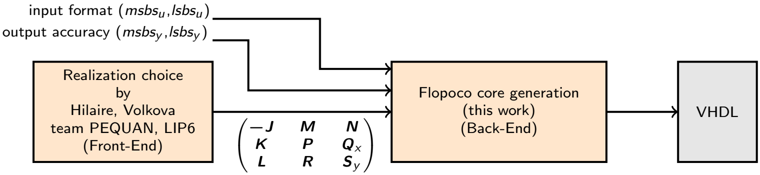
A non-compact notation:

$$\begin{pmatrix} \boldsymbol{J} & \boldsymbol{0} & \boldsymbol{0} \\ -\boldsymbol{K} & \boldsymbol{I}_{n_x} & \boldsymbol{0} \\ -\boldsymbol{L} & \boldsymbol{0} & \boldsymbol{I}_{n_y} \end{pmatrix} \begin{pmatrix} \boldsymbol{t}(k+1) \\ \boldsymbol{x}(k+1) \\ \boldsymbol{y}(k) \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{M} & \boldsymbol{N} \\ \boldsymbol{0} & \boldsymbol{P} & \boldsymbol{Q} \\ \boldsymbol{0} & \boldsymbol{R} & \boldsymbol{S} \end{pmatrix} \begin{pmatrix} \boldsymbol{t}(k) \\ \boldsymbol{x}(k) \\ \boldsymbol{u}(k) \end{pmatrix} \tag{5}$$

An easier way to communicate SIFs:
The Z matrix:

$$Z = \begin{pmatrix} -\boldsymbol{J} & \boldsymbol{M} & \boldsymbol{N} \\ \boldsymbol{K} & \boldsymbol{P} & \boldsymbol{Q} \\ \boldsymbol{L} & \boldsymbol{R} & \boldsymbol{S} \end{pmatrix} \tag{6}$$

Workflow



Workflow overview of tools usage

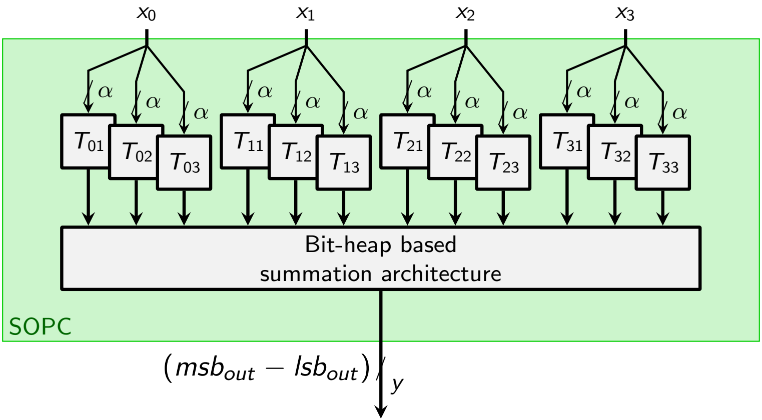
Table of Contents

The FloPoCo Framework: computing just right

FloPoCo:

- C++ framework
- Target: FPGAs
- Work: generating arithmetical cores in VHDL computing just right
- Reference: http://flopoco.gforge.inria.fr/

Basic block: SOPC



SOPC architecture

Architecture generation algorithm

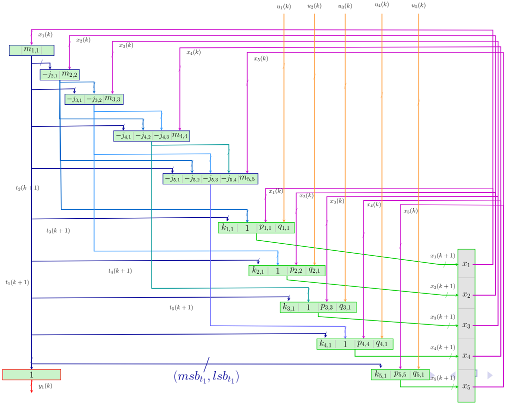
Assuming all the MSBs and LSBs are known.

```
for i=1; i=Z.nbLines(); i++ do
    row[ ]= Z[i][ ] //pick first row of Z
    for j=1; j=1; j=Z.nbCols() j++ do
        assign(SOPC[i], row[j], "T","X","U",[msbs,lsbs][i][j])
    end
    Second pass for wiring.
end
```

Algorithm 2: Architecture Generation Algorithm

Example

$$Z = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & m_{1,1} & 0 & 0 & 0 & 0 & 0 \\
j_{2,1} & 1 & 0 & 0 & 0 & 0 & m_{2,2} & 0 & 0 & 0 & 0 \\
j_{3,1} & j_{3,2} & 1 & 0 & 0 & 0 & 0 & m_{3,3} & 0 & 0 & 0 \\
j_{4,1} & j_{4,2} & j_{4,3} & 1 & 0 & 0 & 0 & 0 & m_{4,4} & 0 & 0 \\
j_{5,1} & j_{5,2} & j_{5,3} & j_{5,4} & 1 & 0 & 0 & 0 & 0 & m_{5,5} & 0 \\
k_{1,1} & 1 & 0 & 0 & 0 & p_{1,1} & 0 & 0 & 0 & 0 & q_{1,1} \\
k_{2,1} & 0 & 1 & 0 & 0 & 0 & p_{2,2} & 0 & 0 & 0 & q_{2,1} \\
k_{3,1} & 0 & 0 & 1 & 0 & 0 & 0 & p_{3,3} & 0 & 0 & q_{3,1} \\
k_{4,1} & 0 & 0 & 0 & 1 & 0 & 0 & 0 & p_{4,4} & 0 & q_{4,1} \\
k_{5,1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{5,5} & q_{5,1} \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}$$

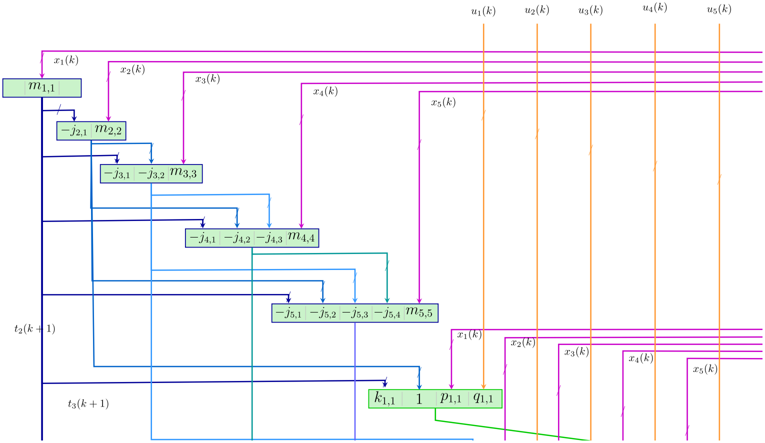


Example

$$
Z = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & m_{1,1} & 0 & 0 & 0 & 0 & 0 \\
j_{2,1} & 1 & 0 & 0 & 0 & 0 & m_{2,2} & 0 & 0 & 0 & 0 \\
j_{3,1} & j_{3,2} & 1 & 0 & 0 & 0 & 0 & m_{3,3} & 0 & 0 & 0 \\
j_{4,1} & j_{4,2} & j_{4,3} & 1 & 0 & 0 & 0 & 0 & m_{4,4} & 0 & 0 \\
j_{5,1} & j_{5,2} & j_{5,3} & j_{5,4} & 1 & 0 & 0 & 0 & 0 & m_{5,5} & 0 \\
k_{1,1} & 1 & 0 & 0 & 0 & p_{1,1} & 0 & 0 & 0 & 0 & q_{1,1} \\
k_{2,1} & 0 & 1 & 0 & 0 & 0 & p_{2,2} & 0 & 0 & 0 & q_{2,1} \\
k_{3,1} & 0 & 0 & 1 & 0 & 0 & 0 & p_{3,3} & 0 & 0 & q_{3,1} \\
k_{4,1} & 0 & 0 & 0 & 1 & 0 & 0 & 0 & p_{4,4} & 0 & q_{4,1} \\
k_{5,1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{5,5} & q_{5,1} \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
$$

Each line will be a SOPC

Example


$u_1(k)$
$u_2(k)$
$u_3(k)$
$u_4(k)$
$u_5(k)$
$x_1(k)$
$x_2(k)$
$x_3(k)$
$x_4(k)$
$x_5(k)$
$m_{1,1}$
$-j_{2,1}$ $m_{2,2}$
$-j_{3,1}$ $-j_{3,2}$ $m_{3,3}$
$-j_{4,1}$ $-j_{4,2}$ $-j_{4,3}$ $m_{4,4}$
$-j_{5,1}$ $-j_{5,2}$ $-j_{5,3}$ $-j_{5,4}$ $m_{5,5}$
$t_2(k+1)$
$x_1(k)$
$x_2(k)$
$x_3(k)$
$x_4(k)$
$x_5(k)$
$k_{1,1}$ 1 $p_{1,1}$ $q_{1,1}$
$t_3(k+1)$

Example

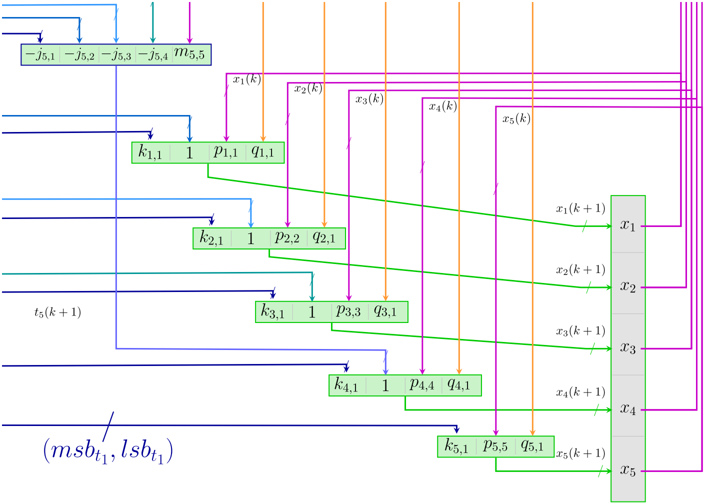
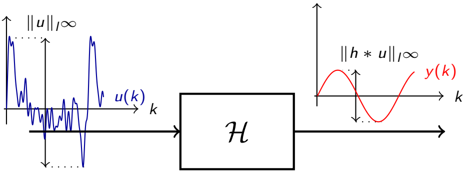

$-j_{5,1}$ $-j_{5,2}$ $-j_{5,3}$ $-j_{5,4}$ $m_{5,5}$
$x_1(k)$
$x_2(k)$
$x_3(k)$
$x_4(k)$
$x_5(k)$
$k_{1,1}$ 1 $p_{1,1}$ $q_{1,1}$
$k_{2,1}$ 1 $p_{2,2}$ $q_{2,1}$
$k_{3,1}$ 1 $p_{3,3}$ $q_{3,1}$
$k_{4,1}$ 1 $p_{4,4}$ $q_{4,1}$
$k_{5,1}$ $p_{5,5}$ $q_{5,1}$
$t_5(k+1)$
(msb_{t_1}, lsb_{t_1})
$x_1(k+1)$
$x_2(k+1)$
$x_3(k+1)$
$x_4(k+1)$
$x_5(k+1)$
x_1
x_2
x_3
x_4
x_5

Table of Contents

Computing the MSB

MSB should allow the full range of the signal.
Worst-Case Peak Gain (WCPG):

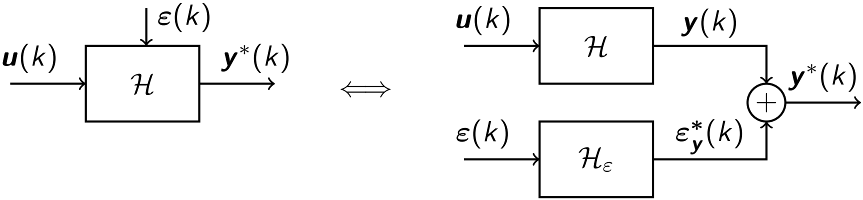


$$\|\mathcal{H}\|_{wcpg} = \sup_{u \neq 0} \frac{\|h * u\|_{l^\infty}}{\|u\|_{l^\infty}} \qquad (7)$$

This computation is done by colleagues in LIP6.

Computing the LSB: Point of view about error

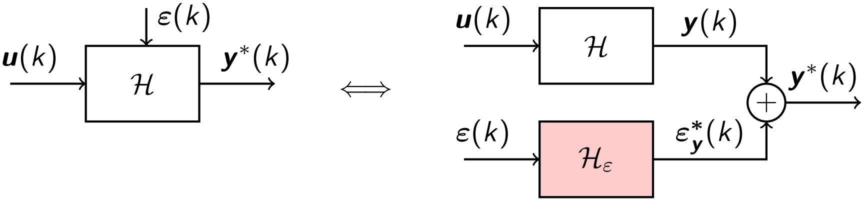
ε is the error introduced by the SOPC (closely related to 2^{lsb_i}).
ε^* is the total error (taking the loopback into account)



Error propagation with respect to the ideal filter

Computing the LSB: Point of view about error

ε is the error introduced by the SOPC (closely related to 2^{lsb_i}).
ε^* is the total error (taking the loopback into account)



Error propagation with respect to the ideal filter

Computing the LSB: Impact of errors

for *int i = 0 ; i ≤ n_t; i++* **do**

$$\boldsymbol{t}_i(k+1) \leftarrow -\sum_{j<i} \boldsymbol{J}_{ij}\boldsymbol{t}_j(k+1) + \sum_{j=1}^{n_x} \boldsymbol{M}_{ij}\boldsymbol{x}_j(k) + \sum_{j=1}^{n_u} \boldsymbol{N}_{ij}\boldsymbol{u}_j(k) + \varepsilon_{t_i}(k)$$

end

for *int i = 0 ; i ≤ n_x; i++* **do**

$$\boldsymbol{x}_i(k+1) \leftarrow \sum_{j=1}^{n_t} \boldsymbol{K}_{ij}\boldsymbol{t}_j(k+1) + \sum_{j=1}^{n_x} \boldsymbol{P}_{ij}\boldsymbol{x}_j(k) + \sum_{j=1}^{n_u} \boldsymbol{Q}_{ij}\boldsymbol{u}_j(k) + \varepsilon_{x_i}(k)$$

end

for *int i = 0 ; i ≤ ny; i++* **do**

$$\boldsymbol{y}_i(k) \leftarrow \sum_{j=1}^{n_t} \boldsymbol{L}_{ij}\boldsymbol{t}_j(k+1) + \sum_{j=1}^{n_x} \boldsymbol{R}_{ij}\boldsymbol{x}_j(k) + \sum_{j=1}^{n_u} \boldsymbol{S}_{ij}\boldsymbol{u}_j(k) + \varepsilon_{y_i}(k)$$

end

Algorithm 3: Computation of SIF outputs from inputs

Computing the LSB: current solution

Let's define:

$$\boldsymbol{v'} = \begin{pmatrix} \boldsymbol{t}(k+1) \\ \boldsymbol{x}(k+1) \\ \boldsymbol{y}(k) \end{pmatrix} \tag{8}$$

Main constraint:

$$\varepsilon_{v'} < \mathbf{2}^{-lsb_{v'}} \tag{9}$$

Transformed into the following constraint:

$$|\langle\langle \mathcal{H}_{\varepsilon} \rangle\rangle| \cdot \mathbf{2}^{lsb_{v'}+1} < \mathbf{2}^{-lsb_{y_i}} \tag{10}$$

Lopez gives a simple solution that matches the fixed size constraint.
The solution matching the hardware context (arbitrary precision) is work in progress.

Table of Contents

Future Work

- Finish implementation of LSB computation (still waiting for the WCPG code)
- Removal of power of two: adapt KCM and SOPC FloPoCo cores
- Sub-filter detection: open question for either Front-End and Back-End
- Better interface, using matlab matrix syntax

Conclusion

To conclude

- Adapting LTI Filters generation from software to hardware
- Reuse of Lopez's calculations in our context
- Implementation for FPGAs in a parametric view

Any question?

Full definition of the SIF

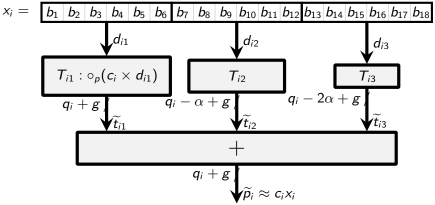
$$\begin{pmatrix} \boldsymbol{J} & \boldsymbol{0} & \boldsymbol{0} \\ -\boldsymbol{K} & \boldsymbol{I}_{n_x} & \boldsymbol{0} \\ -\boldsymbol{L} & \boldsymbol{0} & \boldsymbol{I}_{n_y} \end{pmatrix} \begin{pmatrix} \boldsymbol{t}(k+1) \\ \boldsymbol{x}(k+1) \\ \boldsymbol{y}(k) \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{M} & \boldsymbol{N} \\ \boldsymbol{0} & \boldsymbol{P} & \boldsymbol{Q} \\ \boldsymbol{0} & \boldsymbol{R} & \boldsymbol{S} \end{pmatrix} \begin{pmatrix} \boldsymbol{t}(k) \\ \boldsymbol{x}(k) \\ \boldsymbol{u}(k) \end{pmatrix} \tag{11}$$

With:

$$\begin{aligned} \boldsymbol{J} \in \mathbb{R}^{n_t \times n_t}, \boldsymbol{M} \in \mathbb{R}^{n_t \times n_x}, \boldsymbol{N} \in \mathbb{R}^{n_t \times n_u}, \\ \boldsymbol{K} \in \mathbb{R}^{n_x \times n_t}, \boldsymbol{P} \in \mathbb{R}^{n_x \times n_x}, \boldsymbol{Q} \in \mathbb{R}^{n_x \times n_u}, \\ \boldsymbol{L} \in \mathbb{R}^{n_y \times n_t}, \boldsymbol{R} \in \mathbb{R}^{n_y \times n_x}, \boldsymbol{S} \in \mathbb{R}^{n_y \times n_u}, \end{aligned} \tag{12}$$

$$\tag{13}$$

KCM multiplier



The FixRealKCM method when x_i is split in 3 chunks

Mathematical definition of a filter $\mathcal{H}$

Definition of a filter: $\boldsymbol{y} = \mathcal{H}(\boldsymbol{u})$ With $dim(\boldsymbol{y}) = n_y$ and, $dim(\boldsymbol{u}) = n_u$
Linearity:

$$\mathcal{H}(\alpha \cdot \boldsymbol{u}_1 + \beta \cdot \boldsymbol{u}_2) = \alpha \cdot \mathcal{H}(\boldsymbol{u}_1) + \beta \cdot \mathcal{H}(\boldsymbol{u}_2)$$

Time invariance:

$$\{\mathcal{H}(\boldsymbol{u})(k - k_0)\}_{k \geq 0} = \mathcal{H}(\{\boldsymbol{u}(k - k_0)\}_{k \geq 0})$$

Impulse response

$$u = \sum_{i \geq 0} u(l)\delta_l$$

Where δ_l is a Dirac impulsion centered in l:

$$\delta_l(k) = \begin{cases} 1 & when \;\; k = l \\ 0 & else \end{cases} \tag{14}$$

Time invariance gives: $\mathcal{H}(\delta_l)(k) = h(k - l)$
Computation of the outputs:

$$y_i(k) = \sum_{j=1}^{n_u} \sum_{l=0}^{k} u_j(l) h_{i,j}(k - l), \;\; \forall 1 \leq i \leq n_y$$

From:

$$\boldsymbol{Z} = \begin{pmatrix} -\boldsymbol{J} & \boldsymbol{M} & \boldsymbol{N} \\ \boldsymbol{K} & \boldsymbol{P} & \boldsymbol{Q} \\ \boldsymbol{L} & \boldsymbol{R} & \boldsymbol{S} \end{pmatrix} \tag{15}$$

The ABCD form is deducible from the SIF:

$$\begin{aligned} \boldsymbol{A_Z} = \boldsymbol{KJ}^{-1}\boldsymbol{M} + \boldsymbol{P}, \ \ \boldsymbol{B_Z} = \boldsymbol{KJ}^{-1}\boldsymbol{N} + \boldsymbol{Q}, \\ \boldsymbol{C_Z} = \boldsymbol{LJ}^{-1}\boldsymbol{M} + \boldsymbol{R}, \ \ \boldsymbol{D_Z} = \boldsymbol{LJ}^{-1}\boldsymbol{N} + \boldsymbol{S}, \end{aligned} \tag{16}$$

with:

$$\begin{aligned} \boldsymbol{A_Z} \in \mathbb{R}^{n_x \times n_x}, \boldsymbol{B_Z} \in \mathbb{R}^{n_x \times n_u}, \\ \boldsymbol{C_Z} \in \mathbb{R}^{n_y \times n_x}, \boldsymbol{D_Z} \in \mathbb{R}^{n_y \times n_u}, \end{aligned} \tag{17}$$

Definition of the error filter:

$$\boldsymbol{Z}_\varepsilon = \begin{pmatrix} -\boldsymbol{J} & \boldsymbol{M} & \boldsymbol{M}_t \\ \boldsymbol{K} & \boldsymbol{P} & \boldsymbol{M}_x \\ \boldsymbol{L} & \boldsymbol{R} & \boldsymbol{M}_y \end{pmatrix} \tag{18}$$

with:

$$\boldsymbol{M}_t = (\boldsymbol{I}_{n_t} \; \boldsymbol{0}_{n_t \times n_x} \; \boldsymbol{0}_{n_t \times n_y}), \tag{19}$$

$$\boldsymbol{M}_x = (\boldsymbol{0}_{n_x \times n_t} \; \boldsymbol{I}_{n_x} \; \boldsymbol{0}_{n_x \times n_y}), \tag{20}$$

$$\boldsymbol{M}_y = (\boldsymbol{0}_{n_y \times n_t} \; \boldsymbol{0}_{n_y \times n_x} \; \boldsymbol{I}_{n_y}), \tag{21}$$

Interface specification:

X l c
x_1_1 x_1_2 ... x_1_c
x_2_1 x_2_2 ... x_2_c
.
.
.
x_l_1 x_l_2 ... x_l_c

Where:

- X = name of the matrix
- X $\in$ { J, K, L, M, N, P, Q, R, S, T}
- x_i_j = the coefficient
- $i \in [1, l]$
- $j \in [1, c]$
- l = the number of lines
- c = the numer of columns.

Computing the LSB: Error definition

Errors introduced by SOPCs:

$$\boldsymbol{\varepsilon}_{v'}(k) = \begin{pmatrix} \boldsymbol{\varepsilon}_t(k) \\ \boldsymbol{\varepsilon}_x(k) \\ \boldsymbol{\varepsilon}_y(k) \end{pmatrix} = \begin{pmatrix} \varepsilon_{t_1}(k) \\ \varepsilon_{t_2}(k) \\ \vdots \\ \varepsilon_{t_{n_t}}(k) \\ \varepsilon_{x_1}(k) \\ \varepsilon_{x_2}(k) \\ \vdots \\ \varepsilon_{x_{n_x}}(k) \\ \varepsilon_{y_1}(k) \\ \varepsilon_{y_2}(k) \\ \vdots \\ \varepsilon_{y_{n_y}}(k) \end{pmatrix} \tag{22}$$

$\boldsymbol{\varepsilon}^*_{v'}(k)$ will represent the total error and is defined similarly to $\boldsymbol{\varepsilon}_{v'}(k)$

LSB computation I

$$|\langle\langle \mathcal{H}_{\varepsilon} \rangle\rangle| \cdot \mathbf{2}^{lsb_{v'}+1} < \mathbf{2}^{-lsb_{y_i}} \qquad (23)$$

Expanded line:

$$\sum_{j=1}^{n_t} |\langle\langle \mathcal{H}_{\varepsilon} \rangle\rangle_{i,j}| \cdot \mathbf{2}^{lsb_{t_i}+1} + \sum_{j=n_t}^{n_t+n_x} |\langle\langle \mathcal{H}_{\varepsilon} \rangle\rangle_{i,j}| \cdot \mathbf{2}^{lsb_{x_i}+1} + \sum_{j=n_t+n_x}^{n_t+n_x+n_u} |\langle\langle \mathcal{H}_{\varepsilon} \rangle\rangle_{i,j}| \cdot \mathbf{2}^{lsb_{u_i}+1} < \mathbf{2}^{-lsb_{y_i}} \qquad (24)$$

LSB computation II

Splitting error in $n' = nt + nx + nu$ equal chunks:

$$|\langle\langle \mathcal{H}_\varepsilon \rangle\rangle_{i,j}| \cdot 2^{lsb_{v'}+1} < \frac{2^{-lsb_{y_i}}}{n'} \tag{25}$$

$\Leftrightarrow$

$$2^{lsb_{t_j}+1} < \frac{2^{-lsb_{y_i}}}{|\langle\langle \mathcal{H}_\varepsilon \rangle\rangle_{i,t_j}| \cdot n'} \tag{26}$$

$$2^{lsb_{x_j}+1} < \frac{2^{-lsb_{y_i}}}{|\langle\langle \mathcal{H}_\varepsilon \rangle\rangle_{i,x_j}| \cdot n'} \tag{27}$$

Then:

$$lsb_{x_j} < lsb_{y_i} - log(n' \cdot \|\langle\langle \mathcal{H}_\varepsilon \rangle\rangle\|_{i,x_j}) - 1 \tag{28}$$

$$lsb_{t_j} < lsb_{y_i} - log(n' \cdot \|\langle\langle \mathcal{H}_\varepsilon \rangle\rangle\|_{i,t_j}) - 1 \tag{29}$$

References I

Small bibliography

K.D. Chapman.
Fast integer multipliers fit in FPGAs (EDN 1993 design idea winner).
EDN magazine, 39(10):80, May 1993.

S. Chevillard, M. Joldeș, and C. Lauter.
Sollya: An environment for the development of numerical codes.
In K. Fukuda, J. van der Hoeven, M. Joswig, and N. Takayama, editors, *Mathematical Software - ICMS 2010*, volume 6327 of *Lecture Notes in Computer Science*, pages 28–31, Heidelberg, Germany, September 2010. Springer.

Florent de Dinechin, Matei Istoan, and Abdelbassat Massouri.
Sum-of-product architectures computing just right.
In *Application-Specific Systems, Architectures and Processors (ASAP)*. IEEE, 2014.

Florent de Dinechin and Bogdan Pasca.
Designing custom arithmetic data paths with FloPoCo.
IEEE Design & Test of Computers, 28(4):18–27, July 2011.

Thibaut Hilaire.
Analyse et synthèse de l'implémentation de lois de contrôle-commande en précision finie.
PhD thesis, Université de Nantes, 2006.

References II

P.Chawdhry I.Njabeleke, R.Pannett and C.Burrows.
Design of h-infiny loop-shaping controllers for fluid power systems.
In *IEEE Colloquium Robust Control - Theory, Software and Applications*, 1997.

J.F.Whidborne and R.H.Istepanian.
Reduction of controller fragility by pole sensivity minimization.
In *Int. J.Control*, 2000.

G.Li J.Wu, S.Chen and J.Chu.
Constructing sparse realizations of finite precision digital controllers based on a closed-loop stability related measure.
In *IEEE Proc. Control Theory and Applications*, 2003.

Benoit Lopez.
Implémentation optimale de filtres linéaires en arithmétique virgule fixe.
PhD thesis, Université Paris VI, 2014.

P.Chevel T.Hilaire and J.F.Whidbornz.
A unifying framework for finite wordlength realizations.
In *IEEE*, 2007.

P.Chevel T.Hilaire and Y.Trinquet.
Implicit state-space representation: a unifying framework for FWL implementation of LTI systems.
In *Proc. of the 16th IFAC World Congress.*, 2005.

References III

A. Volkova, T. Hilaire, and C. Lauter.
Reliable evaluation of the Worst-Case Peak Gain matrix in multiple precision.
In *IEEE Symposium on Computer Arithmetic*, 2015.