Determining Fixed-Point Formats using the Worst-Case Peak Gain measure

Anastasia Volkova, Thibault Hilaire, Christoph Lauter

Sorbonne Universités, University Pierre and Marie Curie, LIP6, Paris, France

ASILOMAR 49
November 10, 2015
Motivation

Filter implementation flow:
Motivation

Filter implementation flow:
- Transfer function generation
Motivation

Filter implementation flow:

- Transfer function generation
- State-space, DFI, DFII, …
Filter implementation flow:

- Transfer function generation
- State-space, DFI, DFII, …
- Software or Hardware implementation
Motivation

Filter implementation flow:

- Transfer function generation
  - Coefficient quantization
- State-space, DFI, DFII, ...
- Software or Hardware implementation
Filter implementation flow:

- **Transfer function generation**
  
  ![Coefficient quantization]

- **State-space, DFI, DFII, ...**
  
  ![Large variety of structures with no common quality criteria]

- **Software or Hardware implementation**
Motivation

Filter implementation flow:

- Transfer function generation
  - Coefficient quantization
- State-space, DFI, DFII, ...
  - Large variety of structures with no common quality criteria
- Software or Hardware implementation
  - Constraints: power consumption, area, error, speed, etc.
Motivation: Automatized filter implementation flow

Focus on Fixed-Point realization.

Optimization of wordlengths:
- Take more - pay more
- Take less - overflow risk

What we want in the end:
- Rigorous algorithm for Fixed-Point Formats (FxPF) determination
- Integration into automatized code generator for filters
- Multiple wordlength paradigm
Motivation: Automatized filter implementation flow

Focus on Fixed-Point realization.
Motivation: Automatized filter implementation flow

Focus on Fixed-Point realization.
Optimization of wordlengths:
- Take more - pay more
- Take less - overflow risk

What we want in the end:
- Rigorous algorithm for Fixed-Point Formats (FxPF) determination
- Integration into automatized code generator for filters
- Multiple wordlength paradigm
\( \mathcal{H} := (A, B, C, D) \) is a Bounded Input Bounded Output LTI filter in state-space representation:

\[
\mathcal{H} \begin{cases} 
  x(k+1) &= Ax(k) + Bu(k) \\
  y(k) &= Cx(k) + Du(k)
\end{cases}
\]

with \( q \) inputs, \( n \) states and \( p \) outputs and state matrices
Basic brick: interval through filter

The Worst-Case Peak Gain theorem

Input $u(k)$

$\forall k, \quad |u(k)| \leq \bar{u}$
Basic brick: interval through filter

The Worst-Case Peak Gain theorem

Input $u(k)$

$\forall k, \ |u(k)| \leq \bar{u}$
Basic brick: interval through filter

The Worst-Case Peak Gain theorem

\[ \forall k, \ |u(k)| \leq \bar{u} \]

Input \( u(k) \)

Output \( y(k) \)

amplification/attenuation
Basic brick: interval through filter

The Worst-Case Peak Gain theorem

\[ \forall k, \quad |u(k)| \leq \bar{u} \]

\[ |y(k)| \leq \langle \|\mathcal{H}\| \rangle \bar{u} \]

Worst-Case Peak Gain

\[ \langle \|\mathcal{H}\| \rangle = |D| + \sum_{k=0}^{\infty} |CA^k B| \]
Two’s complement Fixed-Point arithmetic

\[ t = -2^m t_m + \sum_{i=\ell}^{m-1} 2^i t_i \]

- Wordlength: \( w \)
- Most Significant Bit position: \( m \)
- Least Significant Bit position: \( \ell := m - w + 1 \)
Two’s complement Fixed-Point arithmetic

\[ t = -2^m t_m + \sum_{i=\ell}^{m-1} 2^i t_i \]

- quantization step: \(2^\ell\)
- \(t\) represented by integer \(T = t \cdot 2^\ell\)
- \(T \in [-2^m; 2^m - 2^\ell] \cap \mathbb{Z}\)
Two’s complement Fixed-Point arithmetic

\[ t = -2^m t_m + \sum_{i=\ell}^{m-1} 2^i t_i \]

- \( y(k) \in \mathbb{R} \)
- wordlength \( w \) bits
- minimal Fixed-Point Format (FPF) is the least \( m \):

\[ \forall k, \ y(k) \in [-2^m; 2^m - 2^{m-w+1}] \]
Problem statement

Let $\mathcal{H} := (A, B, C, D)$ be a LTI filter:

$$\mathcal{H}\left\{ \begin{array}{c} x(k + 1) = Ax(k) + Bu(k) \\ y(k) = Cx(k) + Du(k) \end{array} \right.$$

Given

- wordlength constraints $w_x$ and $w_y$ for each state and output variable
- input domain $\tilde{u}$

we need to determine the minimal FPF for all variables of filter $\mathcal{H}$, i.e. find the least $m_x$ and $m_y$ such that

$$\forall k, \quad x_i(k) \in [-2^{m_{x_i}}; 2^{m_{x_i}} - 2^{m_{x_i}}w_{x_i} + 1]$$

$$\forall k, \quad y_i(k) \in [-2^{m_{y_i}}; 2^{m_{y_i}} - 2^{m_{y_i}}w_{y_i} + 1]$$
Modification of $\mathcal{H}$

Let $\zeta(k) := \begin{pmatrix} x(k) \\ y(k) \end{pmatrix}$ be a vertical concatenation of state and output vectors.

We seek to determine the least $m$ such that $\forall k, |\zeta_i(k)| \leq 2^m \zeta_i(k) - 2^m \zeta_i(k) - w_{i+1}$. 


Modification of $\mathcal{H}$

Let $\zeta(k) := \begin{pmatrix} x(k) \\ y(k) \end{pmatrix}$ be a vertical concatenation of state and output vectors.

Then the state-space takes the following form:

$\mathcal{H}_\zeta \left\{ \begin{array}{l} x(k+1) = Ax(k) + Bu(k) \\ \zeta(k) = \begin{pmatrix} I \\ C \end{pmatrix} x(k) + \begin{pmatrix} 0 \\ D \end{pmatrix} u(k) \end{array} \right.$
Modification of $\mathcal{H}$

Let $\zeta(k) := \begin{pmatrix} x(k) \\ y(k) \end{pmatrix}$ be a vertical concatenation of state and output vectors.

Then the state-space takes the following form:

$$
\mathcal{H}_\zeta \left\{ \begin{array}{c}
  x(k+1) = Ax(k) + Bu(k) \\
  \zeta(k) = \begin{pmatrix} I \\ C \end{pmatrix} x(k) + \begin{pmatrix} 0 \\ D \end{pmatrix} u(k)
\end{array} \right. 
$$

We seek to determine the least $m_\zeta$ such that

$$
\forall k, \quad |\zeta_i(k)| \leq 2^{m_\zeta_i} - 2^{m_\zeta_i-w_i+1}.
$$
Computing the MSB using the WCPG theorem

Applying the WCPG theorem on filter $\mathcal{H}_\zeta$ gives

$$\forall k, \quad |\zeta_i(k)| \leq (\langle H_\zeta \rangle \bar{u})_i$$
Computing the MSB using the WCPG theorem

Applying the WCPG theorem on filter $\mathcal{H}_\zeta$ gives

$$\forall k, \quad |\zeta_i(k)| \leq (\langle H_\zeta \rangle \bar{u})_i$$

Therefore, the smallest $m_\zeta$ satisfying

$$(\langle H_\zeta \rangle \bar{u})_i \leq 2^{m_\zeta} - 2^{m_\zeta} - w_i + 1,$$

will satisfy the wordlength constraints.
Computing the MSB using the WCPG theorem

Applying the WCPG theorem on filter $\mathcal{H}_\zeta$ gives

$$\forall k, \quad |\zeta_i(k)| \leq (\langle H_\zeta \rangle \bar{u})_i$$

Therefore, the smallest $m_{\zeta_i}$ satisfying

$$(\langle H_\zeta \rangle \bar{u})_i \leq 2^{m_{\zeta_i}} - 2^{m_{\zeta_i} - w_i + 1},$$

will satisfy the wordlength constraints.

We can compute $m_{\zeta_i}$ with

$$m_{\zeta_i} = \left\lceil \log_2 (\langle H_\zeta \rangle \bar{u})_i - \log_2 (1 - 2^{1 - w_{\zeta_i}}) \right\rceil.$$
Taking the quantization errors into account

The exact filter $\mathcal{H}_\zeta$ is:

$$
\mathcal{H}_\zeta \left\{ 
\begin{array}{l}
\mathbf{x}(k+1) = \mathbf{Ax}(k) + \mathbf{Bu}(k) \\
\zeta(k) = \left( \begin{array}{c} \mathbf{I} \\
\mathbf{C} \end{array} \right) \mathbf{x} + \left( \begin{array}{c} 0 \\
\mathbf{D} \end{array} \right) \mathbf{u}(k)
\end{array} \right.
$$
Taking the quantization errors into account

The actually implemented filter $\mathcal{H}_\zeta$ is:

$$
\mathcal{H}_\zeta \left\{ \begin{array}{l}
x^\diamondsuit(k+1) = \diamond_m x (Ax^\diamondsuit(k) + Bu(k)) \\
\zeta^\diamondsuit(k) = \diamond_m \zeta \left( \begin{pmatrix} I \\ C \end{pmatrix} x^\diamondsuit(k) + \begin{pmatrix} 0 \\ D \end{pmatrix} u(k) \right)
\end{array} \right.
$$

where $\diamond_m$ is some operator ensuring faithful rounding:

$$
|\diamond_m(x) - x| \leq 2^{m-w+1}.
$$
Taking the quantization errors into account

The actually implemented filter $\mathcal{H}_\zeta$ is:

$$\mathcal{H}_\zeta \left\{ \begin{array}{l} x^\diamond (k+1) = \diamond m_x (A x^\diamond (k) + B u(k)) \\ \zeta^\diamond (k) = \diamond m_\zeta \left( \begin{pmatrix} I \\ C \end{pmatrix} x^\diamond (k) + \begin{pmatrix} 0 \\ D \end{pmatrix} u(k) \right) \end{array} \right. $$

where $\diamond m$ is some operator ensuring faithful rounding:

$$|\diamond m(x) - x| \leq 2^{m-w+1}. $$

It holds

$$\mathcal{H}_\zeta \left\{ \begin{array}{l} x^\diamond (k+1) = A x^\diamond (k) + B u(k) + \varepsilon_x(k) \\ \zeta^\diamond (k) = \begin{pmatrix} I \\ C \end{pmatrix} x^\diamond (k) + \begin{pmatrix} 0 \\ D \end{pmatrix} u(k) + \begin{pmatrix} \varepsilon_x(k) \\ \varepsilon_y(k) \end{pmatrix} \end{array} \right. $$

with

$$|\varepsilon_x(k)| \leq 2^{m_x-w_x+1} \quad \text{and} \quad |\varepsilon_y(k)| \leq 2^{m_y-w_y+1}. $$
Implemented filter decomposition

\[ u(k) \xrightarrow{\mathcal{H}_\zeta} \zeta^\diamond(k) \]
Implemented filter decomposition

\[ H \] depends on the MSBs of filter \( H \).
Implemented filter decomposition

\[ H^\diamond \] depends on the MSBs of filter \( H^\diamond \).
Implemented filter decomposition
Implemented filter decomposition

Filter $\mathcal{H}_\Delta$ depends on the MSBs of filter $\mathcal{H}_\zeta$
Implemented filter decomposition

Filter $H_\Delta$ depends on the MSBs of filter $H_\zeta$
Implemented filter decomposition

Filter $\mathcal{H}_\Delta$ depends on the MSBs of filter $\mathcal{H}_\zeta$
Two step approach

Step 1  Determine the MSBs $m_\zeta$ for the exact filter $H_\zeta$, applying the WCPG theorem;

Step 2  Compute the error-filter, induced by the format $m_\zeta$ and deduce the FPF of the implemented filter $H_\zeta^{\diamond}$
MSB computation error analysis

$$m_{\xi_i} = \left[ \log_2 (\langle H_{\xi} \rangle \bar{u})_i + \log_2 (1 - 2^{1-w_i}) \right]$$
MSB computation error analysis

\[ m_{\zeta_i} = \left[ \log_2 (\langle H_{\zeta} \rangle \bar{u})_i + \log_2 (1 - 2^{1-w_i}) \right] \]

\[ \langle H_{\zeta} \rangle + \varepsilon_{WCPG} \]
MSB computation error analysis

\[ m_{\zeta_i} = \left[ \log_2 (\langle H_\zeta \rangle \bar{u})_i + \log_2 \left( 1 - 2^{1-w_i} \right) \right] \]

\[ \langle \langle H_\zeta \rangle \rangle + \varepsilon_{WCPG} \]

\[ \hat{m}_{\zeta_i} = \left[ \log_2 (\langle H_\zeta \rangle \bar{u})_i + \log_2 \left( 1 - 2^{1-w_i} \right) \right] \]

\[ + \log_2 \left( 1 + \frac{(\varepsilon_{WCPG} \cdot \bar{u})_i}{\langle \langle H_\zeta \rangle \rangle \bar{u})_i} \right) \]
MSB computation error analysis

\[ m_{\zeta_i} = \left[ \log_2 (\langle \mathcal{H}_\zeta \rangle \bar{u})_i + \log_2 (1 - 2^{1-w_i}) \right] \]

\[ \langle \langle \mathcal{H}_\zeta \rangle \rangle + \varepsilon_{WCPG} \]

\[ \hat{m}_{\zeta_i} = \left[ \log_2 (\langle \mathcal{H}_\zeta \rangle \bar{u})_i + \log_2 (1 - 2^{1-w_i}) \right. \]

\[ + \log_2 \left( 1 + \frac{(\varepsilon_{WCPG} \cdot \bar{u})_i}{\langle \langle \mathcal{H}_\zeta \rangle \bar{u} \rangle_i} \right) \]

\[ = m_{\zeta_i} + \left[ \ldots \right] \]

\[ \in \{0,1\} \]
MSB computation error analysis

\[ m_{\zeta_i} = \left[ \log_2 \left( \langle H_\zeta \rangle \bar{u}_i \right) + \log_2 \left( 1 - 2^{1-w_i} \right) \right] \]

\[ \langle \langle H_\zeta \rangle \rangle + \varepsilon_{WCPG} \]

\[ \hat{m}_{\zeta_i} = \left[ \log_2 \left( \langle H_\zeta \rangle \bar{u}_i \right) + \log_2 \left( 1 - 2^{1-w_i} \right) \right] \]

\[ + \log_2 \left( 1 + \frac{\varepsilon_{WCPG} \cdot \bar{u}_i}{\langle \langle H_\zeta \rangle \rangle \bar{u}_i} \right) \]

\[ = m_{\zeta_i} + \left[ \ldots \right] \]

\[ \varepsilon_{WCPG} \in \{0,1\} \]

Adjust error term \( \varepsilon_{WCPG} \) in order to be at most off by one.
Algorithm

**Step 1** Determine the MSBs $m_\zeta$ for the exact filter $\mathcal{H}_\zeta$, applying the WCPG theorem;

**Step 2** Compute the error-filter $\mathcal{H}_\Delta$, induced by the format $m_\zeta$ and deduce the MSB $m_\hat{\zeta}$ of the implemented filter;

**Step 3** If $m_\hat{\zeta}_i = m_\zeta_i$ then return $m_\hat{\zeta}_i$ otherwise $m_\zeta_i \leftarrow m_\zeta_i + 1$ and go to Step 2.
Numerical examples

Example:

- Random filter with 6 states, 1 input, 3 outputs
- $\bar{u} = 3.7776$, wordlengths set to 7 bits

\[ m_\zeta = (4, 4, 4, 4, 2, 3, 6, 5, 5) \]

\[ m_\zeta^{\diamond} = (4, 5, 4, 4, 2, 3, 6, 5, 5) \]
Numerical examples

Example:

- Random filter with 6 states, 1 input, 3 outputs
- $\bar{u} = 3.7776$, wordlengths set to 7 bits

\[
\begin{align*}
    m_\zeta &= (4, 4, 4, 4, 2, 3, 6, 5, 5) \\
    m_\zeta^{\diamond} &= (4, 5, 4, 4, 2, 3, 6, 5, 5) \\
    \text{After 3 iterations:} \\
    m_\zeta^{\diamond} &= (4, 5, 5, 4, 3, 4, 6, 5, 5)
\end{align*}
\]
Numerical examples

After 3 iterations:

\[ \bar{x}_2(k) \]
Conclusion and Perspectives

Conclusion

- Rigorous procedure for Fixed-Point Formats determination
- Filter computation errors are taken into account, ensuring that no overflow occurs
- Multiple-wordlength paradigm

Perspectives

- Integrate into optimization procedures in automatic workflow
- Solve off-by-one problem
Thank you!
Questions?