Fast Channel Simulation via Waveform Over-Relaxation

Original

Availability:
This version is available at: 11583/2422519 since: 2015-07-15T09:36Z

Publisher:
IEEE

Published
DOI:10.1109/SPI.2011.5898850

Terms of use:
openAccess
This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)
Abstract
This paper presents a fast simulation method for long coupled multi-chip interconnects. The channel is represented through a time-domain passive delay-rational macromodel, which is identified from tabulated scattering samples. A two-level Waveform Relaxation (WR) framework is then applied in order to perform fast transient simulation of the terminated interconnect via an iterative process. A new over-relaxation scheme is introduced for improving the convergence of the WR iterations.

1 Introduction
The electrical verification of point-to-point links connecting different subsystems requires extensive simulation [1, 2]. This analysis is quite demanding, since multiple coupled channels must be simulated concurrently in the time domain, in order to investigate all possible sources of signal degradation including intersymbol interference, crosstalk, jitter, and nonlinear distortions induced by driver and receiver circuitry. Long bit sequences need to be analyzed for each sensitive channel, including multiple aggressor patterns to assess worst-case scenarios.

The most direct simulation approach involves a linear analysis based on convolution, combined with a statistical post-processing. This technique allows very fast and comprehensive analyses. Unfortunately, nonlinear effects of drivers and receivers are only approximated, so that the results may not be fully representative. Alternative approaches based on direct circuit-based (SPICE) simulations allow direct treatment of nonlinear transceiver models. In these simulations, suitable macromodels of the channel are first identified from native tabulated scattering responses and then synthesized as SPICE-compatible netlists. Main drawback of this approach is runtime, which can be considerable for complex channels.

In this work, we propose an alternative approach that is able to preserve SPICE (or better) accuracy, with dramatic speedup in runtime. The technique we pursue in this work is based on a two-level Waveform Relaxation (WR) [3, 8, 9, 11, 12, 13]. One relaxation loop is based on transverse partitioning, aimed at decoupling individual channels through suitable relaxation processes. A second relaxation loop is based on a longitudinal partitioning, aimed at decoupling each individual channel from its terminations. The basic scheme was first presented in [10].

The main contribution of this paper is a new formulation based on a successive over-relaxation [14, 15]. The new formulation is able to fix possible convergence issues of the basic scheme through the introduction of an over-relaxation parameter, which is tuned in a preprocessing phase in order to guarantee the best convergence rate. Although we cannot prove that the over-relaxed scheme will be able to guarantee convergence in the general case, we show its practical effectiveness on a set of industrial benchmarks.

2 Problem statement
We consider a generic chip-to-chip link with \( P \) fully-coupled electrical ports (\( P \) even), terminated by possibly nonlinear single-ended termination circuits (drivers and receivers). A graphical illustration of the system topology is depicted in the left panel of Fig. 4. The mathematical formulation of the interaction between channel and terminations can be obtained as

\[
\begin{align*}
\{ b(t) = h(t) * a(t) \} \\
F_{q}(a_q(t); b_q(t); t; \frac{d}{dt}) = 0, \quad q = 1, \ldots, P 
\end{align*}
\]

where the first row represents the convolution between transient scattering waves \( a(t) \) entering the channel and the channel impulse response matrix \( h(t) \), and second row collects all nonlinear and dynamic equations representing drivers and receivers. For convenience, also the terminations are represented in the transient scattering form. Using a more compact operator notation, we have

\[
\begin{align*}
\{ b = \mathcal{H} a \} \\
\mathcal{A} = \mathcal{F}(b)
\end{align*}
\]

where the linear operator \( \mathcal{H} \) represents the channel, and where the explicit nonlinear operator \( \mathcal{F} \) is diagonal.

We assume that a Delay-Rational Macromodel (DRM) is available for the channel [5, 4]. In the Laplace-domain, a DRM can be expressed as

\[
H^{i,j}(s) = \sum_{m=0}^{M_{i,j}} Q_{i,j}^{l,m}(s) e^{-s\tau_{i,j}^{l,m}} + D^{i,j},
\]

where \( \tau_{i,j}^{l,m} \) are suitable delays, and \( Q_{i,j}^{l,m}(s) \) are rational coefficients. The identification of (3) can be performed from tabulated scattering matrix samples \( \tilde{S}_l \in \mathbb{C}^{P\times P} \) available at the discrete frequencies \( \omega_l \), \( l = 1, \ldots, L \) through Delayed Vector Fitting (DVF) or the Delayed Sanathanan-Koerner (DSK) iterations, see [5, 4]. Model passivity can also be checked enforced [6, 7]. The main advantage of DRM operators is that their time-domain application can be cast as a delayed recursive convolution, whose numerical evaluation has a cost that scales only linearly with the number of time steps to be computed [10]. Thus, the time-domain evaluation of the first row in (2) is extremely efficient.

3 Inner and Outer Waveform Relaxation

The main disadvantage of (1) or (2) is that, for each time step, there is an instantaneous coupling between channel and terminations. Within a time-stepping framework, this would
require the concurrent solution of the fully-coupled channel together with its terminations using a nonlinear solver. This is what standard SPICE solvers do. We use a completely different approach here. Our unknowns are vectors collecting all time samples of all port variables, rather than individual time samples. These unknowns are computed collectively through an iterative scheme that refines an initial estimate via a Waveform Relaxation (WR) approach.

We start by reviewing the two-level WR scheme presented in [10]. A first level exploits a decoupling process between different channels, by decomposing the channel operator as \( \mathcal{H} = \mathcal{D} + \mathcal{C} \), where \( \mathcal{D} \) collects all transfer matrix entries representing direct transmission and reflection coefficients, and \( \mathcal{C} \) collects all near and far end crosstalks. A transverse partitioning and "outer" relaxation loop with index \( \mu \) is obtained by delaying the application of the coupling operator \( \mathcal{C} \) by one iteration. This can be expressed as

\[
\begin{align*}
\{ \mathbf{b}_\mu &= \mathcal{D} \mathbf{a}_\mu + \theta_{\mu-1}, \\
\mathbf{a}_\mu &= \mathcal{F} (\mathbf{b}_\mu),
\end{align*}
\]

(4)

to be solved for \( \mu = 1, 2, \ldots \) with a suitable initial condition, e.g., \( \theta_0 = 0 \). The evaluation of the relaxation source \( \theta_{\mu-1} \) is performed via recursive convolutions. The above outer relaxation is motivated by the small correction that is expected at each iteration as an effect of the inter-channel couplings \( \mathcal{C} \), which for a well-designed link are small. This condition should also ensure fast convergence of the WR iterations.

A second "inner" relaxation loop is also introduced in order to avoid the concurrent solution of individual (decoupled) channels and their terminations. A longitudinal partition is applied in order to separate channels from terminations, and port variable estimates are refined through a longitudinal relaxation with index \( \nu \), whereas channel and termination equations perform updates to the interface variables alternating in time. Formally, this can be represented as

\[
\begin{align*}
\{ \mathbf{b}_{\mu,\nu} &= \mathcal{D} \mathbf{a}_{\mu,\nu} + \theta_{\mu-1}, \\
\mathbf{a}_{\mu,\nu} &= \mathcal{F} (\mathbf{b}_{\mu,\nu}),
\end{align*}
\]

(5)

to be solved for any fixed \( \mu \) for \( \nu = 1, \ldots, \mathcal{I}_\mu \) with a suitable initial condition, e.g., the available solution estimate at previous transverse relaxation step \( \mathbf{a}_{\mu,0} = \mathbf{a}_{\mu-1,\mathcal{I}_{\mu-1}} \). Further details on this scheme, denoted as WR-LPTP (Waveform Relaxation via Longitudinal and Transverse Partitioning) are available in [10].

Figure 1: Spectral radius of operator \( P_I \) for \( I = 2, 4, \infty \) plotted versus frequency for Cases A and B.

Convergence of inner and outer loops is detected by monitoring the respective residual norms

\[
\begin{align*}
\xi (\mu, \nu) &= \| \mathbf{a}_{\mu,\nu} - \mathbf{a}_{\mu,\nu-1} \| \\
\delta (\mu, I) &= \| \mathbf{a}_{\mu, I} - \mathbf{a}_{\mu-1, I} \|.
\end{align*}
\]

(6) (7)

When below a prescribed tolerance \( \epsilon \), these norms are used as stopping conditions for the iterations.

Two industrial benchmarks will be used to illustrate the proposed WR schemes. Case "A" is a fully-coupled 18-port channel providing the electrical link between CPU and an IO card. Case "B" is a similar 18-port channel connecting two boards through a connector. Both cases are courtesy of IBM. Figure 1 reports the frequency-dependent spectral radius of iteration operator \( P_I \) for the two cases. We see that case A is expected to converge, since condition (10) holds. Case B is instead expected to have problems, since there are some frequencies for which condition (10) is violated. The results confirm these Propositions. Figure 2 reports the inner and outer error estimates (6)-(7) through iterations (with \( I = 4 \)) for the two cases. The error for case A converges below the adopted stopping threshold, whereas the outer error for case B blows up. For case A, the transient results are in perfect agreement with SPICE, as Fig. 3 shows. However, a 1000-bit SPICE simulation requires about 23 minutes, whereas the same deck is solved by WR in only 31 seconds, with a \( 44 \times \) speedup. No validation is possible for case B, since the transient waveforms are not even bounded through iterations.
4 Successful Over-Relaxation

In Section 3, we observed that there may be cases where the standard WR-LPTP algorithm does not converge, since the spectral radius of the iteration operator exceeds one and condition (10) does not hold. We now construct a new relaxation scheme that overcomes these difficulties. This new algorithm is denoted as WR-SOR and uses a Successful Over-Relaxation technique, as described in [14, 15].

The starting point is system (2), where operator $\mathcal{H}$ is split into its (block-) diagonal part $\mathcal{D}$ and remainder $\mathcal{C}$

$$
\begin{cases}
    b - \mathcal{D}a - \mathcal{C}a &= 0 \\
    a - \mathcal{F}(b) &= 0.
\end{cases}
$$

We multiply both equations by an over-relaxation parameter $\eta$ (to be determined) and we add $(b - \mathcal{D}a)$ and $(a - \mathcal{F}(b))$ to both sides of first and second equation, respectively, obtaining

$$
\begin{cases}
    \eta b - \eta \mathcal{D}a - \eta \mathcal{C}a + b - \mathcal{D}a &= b - \mathcal{D}a \\
    \eta a - \eta \mathcal{F}(b) + a - \mathcal{F}(b) &= a - \mathcal{F}(b).
\end{cases}
$$

Rearranging the various terms and introducing the outer relaxation index $\mu$, we obtain

$$
\begin{cases}
    b_\mu = \mathcal{D}a_\mu + \theta_{\mu-1} \\
    a_\mu = \mathcal{F}(b_\mu) + \varphi_{\mu-1},
\end{cases}
$$

where $\theta_\mu$ and $\varphi_\mu$ are the outer relaxation sources, defined as

$$
\begin{cases}
    \theta_\mu = (1 - \eta)(b_\mu - \mathcal{D}a_\mu) + \eta \mathcal{C}a_\mu, \\
    \varphi_\mu = (1 - \eta)(a_\mu - \mathcal{F}(b_\mu)).
\end{cases}
$$

System (13) is a generalization of (4), which can be obtained by setting $\eta = 1$ in (13)-(14). The introduction of the overrelaxation adds an outer relaxation source $\varphi_\mu$ also to the equation corresponding to the channel terminations, causing a negligible overhead both in computation and storage. The scheme is consistent, since setting $\mu \to \infty$ leads to the original system (2).

At any outer iteration, the actual solution is performed by applying a second level of (longitudinal) partitioning, as for the WR-LPTP scheme, with suitable relaxation sources and inner iteration index $\nu$. The resulting WR-SOR scheme reads

$$
\begin{cases}
    b_{\mu,\nu} = \mathcal{D}a_{\mu,\nu-1} + \theta_{\mu-1} \\
    a_{\mu,\nu} = \mathcal{F}(b_{\mu,\nu}) + \varphi_{\mu-1},
\end{cases}
$$

The right panel of Fig. 4 provides a graphical illustration of this scheme. At the initialization stage $\mu = 0$ we set

$$
\begin{cases}
    \theta_0 = 0 \\
    \varphi_0 = \mathcal{F}(0).
\end{cases}
$$

The convergence of inner and outer loops is detected as for the WR-LPTP algorithm, by monitoring the respective residual norm estimates $\xi_{\mu,\nu}$, as in (6)-(7).

The convergence of the WR-SOR scheme can be assessed by a frequency-domain analysis as for the WR-LPTP scheme. The resulting expression for the iteration operator is

$$
P_{\omega,\eta} = \mathbf{I} - \eta \left[\mathbf{I} - \left(\mathcal{D} \mathcal{F}\right)^{\mathcal{X}}\right] \left(\mathbf{I} - \mathcal{P}\right).
$$

The error of the iterative solution with respect to the exact solution can also be formally computed as

$$
\mathcal{E}_{\omega,\eta} = \mathbf{A}_{\omega,\eta} - \mathbf{A}_{\text{exact}} = -\eta \mathcal{P}_{\omega,\eta} \mathbf{A}_{\text{exact}}.
$$

Therefore, convergence and consistency of the WR-SOR scheme are guaranteed if $\rho_{\text{max}}(P_{\omega,\eta}) < 1$.

Operator $P_{\omega,\eta}$ is parameterized by $\eta$. It is thus possible to find a value $\eta_{\text{opt}}$ such that the spectral radius of $P_{\omega,\eta}$ is minimized and the convergence rate is optimal. We first provide the feasibility conditions for this optimization. For any fixed frequency $\omega$, let us denote

$$
\lambda_q(\omega) \in \text{eig}\{\Lambda(\omega)\} \quad q = 1, \ldots, P
$$

the generic eigenvalue of matrix

$$
\Lambda = \left\{\left[\mathbf{I} - \left(\mathcal{D} \mathcal{F}\right)^{\mathcal{X}}\right] \left(\mathbf{I} - \mathcal{P}\right)\right\},
$$

such that $P_{\omega,\eta} = \mathbf{I} - \eta \mathbf{A}$ from (17). The WR-SOR scheme will converge if

$$
|1 - \eta \lambda_q(\omega)| < 1 \quad \forall \omega, q.
$$

A straightforward derivation leads to the feasibility condition

$$
0 \leq \eta \leq \eta_{\min} = \min_{\omega, q} \frac{2 \cos \phi_q(\omega)}{\rho_q(\omega)}
$$
with $\rho_\eta$ and $\phi_\eta$ denoting magnitude and phase of $\lambda_\eta$. Condition (22) is feasible only if $\Re \lambda_\eta > 0 \ \forall \eta, \omega$, leading to a non-vanishing range for the overrelaxation parameter. The optimal value for $\eta$ is then found by minimizing the spectral radius of the iteration operator, i.e.,

$$\eta_{\text{opt}} = \arg \min_{\omega, q} \left| 1 - \eta \lambda_q(\omega) \right|.$$  \hspace{1cm} (23)

5 Results

In this section, we show that the WR-SOR scheme is able to fix the convergence problems that were observed in Section 3 when applying the WR-LPTP scheme to case B. The optimal over-relaxation parameter was first computed through a preprocessing analysis by minimizing the spectral radius of the iteration operator throughout the frequency range of interest. The resulting $\rho_{\text{max}}$, depicted in the left panel of Fig. 5, results strictly unitary bounded. Running the WR-SOR scheme on case B leads to the evolution of the iteration error depicted in the right panel of Fig. 5. As expected this error converges below the prescribed stopping threshold. Finally, we validated the WR-SOR scheme with a direct SPICE simulation. The results for one of the termination voltages of case B are depicted in Fig. 6, showing excellent correlation.

In conclusion, the proposed scheme is able to improve or guarantee convergence in some cases that are problematic for standard preexisting WR schemes. A general methodology for the optimization of convergence rate of the over-relaxed WR scheme has been presented. The numerical results are indeed encouraging, although the proposed methodology is still not able to guarantee the convergence in the general case. Future work will be devoted to further generalizations and to on-the-fly adjustment of the over-relaxation parameter.