Balanced Stochastic Truncation of Coupled 3D Interconnect

Amir Zjajo, Nick van der Meijs, Rene van Leuken
Circuits and Systems Group
Delft University of Technology
Mekelweg 4, 2628 CD Delft, The Netherlands
E-mail: amir.zjajo@ieee.org

Abstract— **Model order reduction techniques have been used extensively to reduce the complexity of extracted interconnect circuits and to expedite fast and accurate circuit simulation. In this paper, we introduce a balanced stochastic truncation in model order reduction of a coupled 3D interconnect to provide uniform approximation of the frequency response of the original system over the whole frequency domain. As the experimental results indicate, the proposed approach can significantly reduce the complexity of interconnect, while retaining high accuracy in comparison to the original model.**

I. Introduction

In the nanometer regime, the transistor scaling has been slowing down due to the challenges and hindrances of increasing variability, short-channel effects, power/thermal problems and the complexity of interconnect. The 3D integration has been proposed as one of the alternatives to overcome the interconnect restrictions [1]. In this context, deriving an efficient model order reduction (MOR) techniques that can provide parameterized coupled 3D interconnects in a reduced parameter space and facilitate efficient delay calculation is one of the critical tasks. In an asymptotic waveform evaluation (AWE) model order reduction [2] explicit moment matching was used to compute the dominant poles via Padé approximation. As the AWE method is numerically unstable for higher-order moment approximation, a more efficient solution is to use a projection-based Krylov subspace MOR methods, such as the Padé via Lanczos (PVL) method [3], or PRIMA [4]. However, these methods endure accuracy loss of the reduced model in order to guarantee stability. Additionally, PRIMA-like methods do not preserve structure properties like reciprocity of a network. Alternatively, MOR can be performed by means of singular-value-decomposition (SVD) based approaches such as control-theory-based truncated balance realization (TBR) methods, where the state variables are truncated to achieve the reduced models [5]-[11]. The major advantage of SVD-based approaches over Krylov subspace methods lies in their ability to ensure the errors satisfy an *a priori* upper bound [8] while preserving stability and passivity for asymmetric and indefinite systems [9]. Additionally, SVD-based methods typically lead to optimal or near optimal reduction results as the errors are controlled in a global way, although, for large scale problems such as interconnect, iterative methods have to be used to find an adequate balanced approximation (truncation). Accordingly, several SVD approaches approximate the dominant Cholesky factors (dominant eigensubspaces) of controllability and observability Gramians [6],[10]-[11] to compute the reduced model.

In this paper, we introduce a balanced *stochastic* truncation [12] in model order reduction of coupled 3D interconnect to include variability and provide a uniform approximation of the frequency response of the original system over the whole frequency domain. The approach presented here produces orthogonal basis sets for the dominant singular subspace of the controllability and observability Gramians, which significantly reduce the complexity and computational costs of SVD, while preserving model order reduction accuracy and the quality of the approximations of the TBR procedure.

II. Adjusted balanced Stochastic Truncation

The performance of 3D integrated circuit can be enhanced to exceed the performance of planar 2D ICs by improving interconnect delay, mainly by increasing the wiring pitch, which causes a reduction in resistance and line-to-line capacitance per unit length. For each performance condition applied, the tier boundaries are necessarily shifted in the wire length distribution towards shorter wires such that the longest wire in each tier can satisfy the new delay condition. Consequently, wires that no longer satisfy the new delay condition are routed to higher tiers where they have larger cross sections and pitches. These wires may have numerous features: bends, crossings, vias, etc., and are modeled by circuit extractors in terms of a large number of connected circuit elements: capacitors, resistors and more recently inductors. The propagation of signals in wires and through silicon vias (TSV) that satisfy delay conditions and signal integrity is evaluated with the partial element equivalent circuit (PEEC) method as it provides quasi-static circuit equivalent models easily linked to traditional circuit simulators [13]. Written in a state space form, such a model can be expressed as

$$\begin{aligned} D_j(dx_j / dt) &= G_j x_j(t) + B_j u_j(t) \\ y_j(t) &= E_j^T x_j(t) \end{aligned} \tag{1}$$

where $D_j, G_j \in \mathscr{R}^{n_j \times n_j}$ are matrices describing the reactive (capacitive and inductive) and dissipative parts of the model, respectively, expressed as a function of TSV geometry and material properties, $B_j \in \mathscr{R}^{n_j \times m_j}$ is a matrix that defines the input

This research was supported in part by the CATRENE program under the Computing Fabric for High Performance Applications (COBRA) project CA104.

978-1-4673-4743-3/13/$31.00 ©2013 IEEE

ports, $E_j \in \mathcal{R}^{p_j \times n_j}$ is matrix that defines the outputs, $x_j(t) \in \mathcal{R}^{n_j}$ are internal state vectors, $u_j(t) \in \mathcal{R}^{m_j}$ are internal inputs and $y_j(t) \in \mathcal{R}^{p_j}$ are internal outputs. The impact of thermo-mechanical stresses induced during TSV formation on the coupling signal integrity is analyzed in [14]. For TSVs arranged in a row (border) or in a bundle, the stress components add up and thus propagate larger distances into the surrounding silicon, implying the need for a larger keep out zones (KOZ). As a consequence, the coupling signal integrity and the electrical performance of vertical interconnect is highly dependent on its structure and placement. In a state space terms, we express signal integrity coupling of (1) through the relations

$$\begin{aligned} u_j(t) &= K_{j1} y_1(t) + \ldots + K_{jk} y_k(t) + Q_j u(t), \quad j = 1, \ldots, k \\ y(t) &= P_1 y_1(t) + \ldots + P_k y_k(t) \end{aligned} \tag{2}$$

where $K_{jk} \in \mathcal{R}^{m_j \times p_k}$, $Q_j \in \mathcal{R}^{m_j \times m}$, $P_j \in \mathcal{R}^{p \times p_j}$ are coupling matrixes and $y(t) \in \mathcal{R}^p$ and $u(t) \in \mathcal{R}^m$, are the vectors of external outputs and inputs, respectively. If I-$H(s)K$ is invertible, the input-output relation of the coupled system (1), (2) can be written as $y(s) = \Gamma(s) u(s)$, where $y(s)$ and $u(s)$ are the Laplace transforms of the external output $y(t)$ and the external input $u(t)$, respectively, and the closed-loop transfer function $\Gamma(s)$ has the form

$$\begin{aligned} &\Gamma(s) = P(I - H(s)K)^{-1} H(s) Q \\ &H(s) = diag(H_1(s), \ldots, H_k(s)) \quad H_j(s) = E_j^{\,T} (sD_j - G_j)^{-1} B_j \end{aligned} \tag{3}$$

A generalized state space realization of $\Gamma(s)$ is given by

$$\begin{aligned} &\mathcal{D}(dx/dt) = \mathcal{G}x(t) + \mathcal{B}u(t) \quad \mathcal{D} = D \in \Re^{n,n} \quad \mathcal{G} = G + BKE^T \in \Re^{n,n} \\ &y(t) = \mathcal{E}^T x(t) \qquad \mathcal{B} = BQ \in \Re^{n,m} \qquad \mathcal{E}^T = PE^T \in \Re^{p,n} \end{aligned} \tag{4}$$

To guarantee the passivity of the reduced model and simplify the computational procedure, we first convert original descriptor systems into standard state-space equations by mapping $\mathcal{D} \rightarrow I$, $\mathcal{G} \rightarrow \mathcal{D}^{-1}\mathcal{G}$ and $\mathcal{B} \rightarrow \mathcal{D}^{-1}\mathcal{B}$. Unlike balanced truncation methods [5]-[11], the *stochastic* balancing algorithm requires solving one Lyapunov and one Riccati equation. If we define $\Phi(s) = \Gamma(s)\Gamma^T(-s)$, and let W be a square minimum spectral factor of Φ, satisfying $\Phi(s) = W^T(-s)W(s)$, a state space realization $(\mathcal{D}_W, \mathcal{G}_W, \mathcal{B}_W, \mathcal{E}_W)$ of $W(s)$ can be obtained as

$$\mathcal{D}_W = \mathcal{D} \quad \mathcal{G}_W = \mathcal{G} \quad \mathcal{B}_W = \mathcal{B} + Y\mathcal{E} \quad \mathcal{E}_W^T = \mathcal{E}^T - \mathcal{B}_W^T X \tag{5}$$

where Y is the controllability Gramian (e.g. the low rank approximation to the solution) of Γ given by Lyapunov equation

$$\mathcal{G}Y + Y\mathcal{G}^T + \mathcal{B}\mathcal{B}^T = 0 \tag{6}$$

and X is the observability Gramian of W, being solution of the Riccati equation

$$X\mathcal{G} + \mathcal{G}^T X + \mathcal{E}F\mathcal{E}^T + X\mathcal{B}_W M^{-1} \mathcal{B}_W^T X = 0 \tag{7}$$

where $F \in \mathcal{R}^{p \times p}$ is symmetric, positive semi-definite and $M \in \mathcal{R}^{m \times m}$ is symmetric, positive definite. The model order reduction system

$$(\widehat{\mathcal{D}}, \widehat{\mathcal{G}}, \widehat{\mathcal{B}}, \widehat{\mathcal{E}}^T) = (T^{-1}\mathcal{D}T, T^{-1}\mathcal{G}T, T^{-1}\mathcal{B}, \mathcal{E}^T T) \tag{8}$$

where $\widehat{\mathcal{D}}, \widehat{\mathcal{G}} \in \mathcal{R}^{l \times l}$, $\widehat{\mathcal{B}} \in \mathcal{R}^{l \times m}$ and $\widehat{\mathcal{E}} \in \mathcal{R}^{p \times l}$ are of order l much smaller than the original order n, but for which the output y(s) and $\widehat{y}(s)$ are approximately equal for inputs $u(s)$ of interest, is *stochastically* balanced in transfer function $\Gamma(s)$, if $Y = X = \Sigma$ $= \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$, where $1 = \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$.

The balancing transformation matrix T tends to be highly ill-conditioned. As a consequence, the square root method [11] avoids explicit balancing of (8) by calculating the Cholesky factors of the Gramians instead of the Gramians themselves. Recently, in [6] and [10] it has been observed that solutions often have low numerical rank, which means that there is a rapid decay in the eigenvalues of the Gramians. In the original implementation this step is the computation of exact Cholesky factors, which may have full rank. We formally replace these (exact) factors by (approximating) low rank Cholesky factors [6],[10]. The iterative procedure approximates the low rank Cholesky factors S and R with $r_S, r_R \ll n$, such that $R^T R \approx X$ and $S^T S \approx Y$. The observability Gramian X is obtained by solving the Riccati equation (7) with a Newton double step iteration

$$\begin{aligned} &(\mathcal{G}^T - Z^{(z-1)}\mathcal{B}_W^T) X^{(z)} + X^{(k)} (\mathcal{G} - \mathcal{B}_W Z^{(z-1)^T}) = \\ &\qquad\qquad - \mathcal{E}^T F \mathcal{E} = Z^{(z-1)} M Z^{(z-1)^T} \\ &Z^{(z)} = X^{(z)} \mathcal{B}_W M^{-1} \end{aligned} \tag{9}$$

where the feedback matrix $Z = X\mathcal{B}_W M^{-1}$, for $z = 1,2,3,\ldots$, which generates a sequence of iterates $X^{(z)}$. This sequence converges towards the stabilizing solution X if the initial feedback Z_0 is stabilizing, i.e., G-$BZ^{(0)\mathrm{T}}$ is stable. If we partition T and T^{-1} as $T = [J\ U]$ and $T^{-1} = [L\ V]^{-1}$ then $I_l = LJ$ is the identity matrix, $\Pi = JL$ is a projection matrix, and L and J are truncation matrices. In the related balancing model reduction methods, the truncation matrices L and J can be determined knowing only the Cholesky factors of the Gramians Y and X. Let

$$S^T R = U \Sigma V^T \tag{10}$$

where $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_l)$, be singular value decomposition (SVD) of $S^T R$ of dimension $N \times m$. The cost of this decomposition including the construction of U is $14Nm^2 + O(m^3)$ [15]. To avoid this, in this paper we perform eigenvalue decomposition

$$(S^T R)^T S^T R = U \Lambda U^T \tag{11}$$

Comparing (11) with (10) shows that the same matrix U is constructed and that

$$(S^T R U)^T S^T R U = \Lambda = \Sigma^T \Sigma \tag{12}$$

This algorithm requires Nm^2 operations to construct $(S^T R)^T$ $S^T R$ and $Nmn + O(m^3)$ operations to obtain $S^T R U \Sigma^{-1}$for a $n \times n$ Σ. The balancing transformation matrix L and J can be determined as

$$L = \Sigma^{-1/2} V^T R \qquad J = S^T U \Sigma^{-1/2} \tag{13}$$

then, under a similarity transformation of the state-space model, both parts can be treated simultaneously after a transformation of the system $(\widehat{\mathcal{D}}, \widehat{\mathcal{G}}, \widehat{\mathcal{B}}, \widehat{\mathcal{E}}^T)$ with a nonsingular matrix $T \in \mathcal{R}^{n \times n}$ into a balanced system

$$\widehat{\mathcal{D}} = L\mathcal{D}J^T \quad \widehat{\mathcal{G}} = L\mathcal{G}J^T \quad \widehat{\mathcal{B}} = J^T\mathcal{B} \quad \widehat{\mathcal{E}} = \mathcal{E}L \tag{14}$$

In this algorithm we assume that $k \leq r$ (*rank* S^TR). Note that SVDs are arranged so that the diagonal matrix containing the singular values has the same dimensions as the factorized matrix and the singular values appear in non-increasing order.

III. Experimental results

All numerical examples, which demonstrate the accuracy, stability and efficiency of the proposed algorithm are conducted in MatLab and carried out on a PC with an Intel Core 2 Duo CPU running at 2.66 GHz and with 3 GB of memory. Four geometries that form the limits of the TSV dimension range for 3D-SOC [16] are considered in order to derive the set of TSV parameters that result in the minimum and the maximum delay through the TSV. The TSV model trends observed by sweeping the material properties and the frequency indicated a large percentage change [17]. However, the impact of this change on the path delay is not significant once the overall path circuit is considered. The dimensions and TSV RLC values computed for the four TSV geometries as per [18] are given in Table I. Driver resistance and load capacitance are estimated by assuming a 40x buffer driver and a 1x buffer load in PTM 65nm technology [19]. The transition times (T_r) of victim and aggressor inputs vary from 10ps to 100ps. Due to the variation is TSV width, height and the liner oxide thickness, TSV RLC values vary considerably. The sensitivity of each given data to the sources of variation is chosen randomly, while the total σ variation for each data is chosen in the range of 10% to 30% of their nominal value. The scaled distribution of the sources of variation is considered to have a skewness of 0.5, 0.75, and 1. Figure 1 shows the dependence of the victim delay on the input skew of a pair of coupled 200µm intermediate wires. If the victim and aggressor inputs switch in the same direction, coupling effects can speedup the victim transition and reduce the interconnect delay of the victim wire, which changes the best-case delay. On the other hand, if victim and aggressor inputs switch in an opposite direction, victim transitions slow down thus affecting the worst-case victim delay.[1] The interconnect delay decreases when the coupling effect occurs if two input signals switch in the same direction, and the coupling effects are apparent when the input skew is within approximately $[-0.6T_r, 0.6T_r]$. The figure of merit selected for this benchmark is to capture the cross-talk between lines, i.e., the transfer function between the input of the first line, and the output of the second line. Each line and the couplings are modeled via distributed elements, with a total model order of 6006 states. The convergence history for solving the Lyapunov equation (6) with respect to the number of iteration steps is plotted in Figure 2. Convergence is obtained after 36 iterations. The *cpu*-time needed to solve the Lyapunov equation according to the related tolerance for solving the shifted systems inside the iteration is 0.27 seconds.

	Thick TSV		*Thin TSV*	
	t_{diel}	$2\times t_{diel}$	t_{diel}	$2\times t_{diel}$
Dimensions				
Diameter (µm)	8	8	4	4
Length (µm)	40	40	40	40
Aspect ratio (L/D)	5	5	10	10
Pitch (µm)	16	16	8	8
t_{diel} *(µm)*	0.5	1	0.5	1
Parasitic Components				
R_{TSV} *(Ω)*	0.20	0.20	0.50	0.50
L_{TSV} *(pH)*	15.00	15.00	20.20	20.20
C_{TSV} *(fF)*	52.30	26.10	30.90	15.40
C_c *(fF)*	5.88	5.88	5.87	5.87
Performance				
50% Delay (ps)	4	3	3	2
Corner Case	*Worst*			*Best*

TABLE I– TSV GEOMETRIES CONSIDERED, VALUES OF THEIR PARASITIC COMPONENTS AND PERFORMANCE CORNERS

Note further that saving iteration steps means that we save large amounts of memory-especially in the case of multiple input and multiple output systems where the factors are growing by p columns in every iteration step. The convergence history of the Newton double step iteration (9) for solving the Riccati equation (7) is illustrated in Figure 3. Due to symmetry, the matrices F and M can be factored by a Cholesky factorization. Hence, the equations to be solved in (9) have a Lyapunov structure similar to (6).

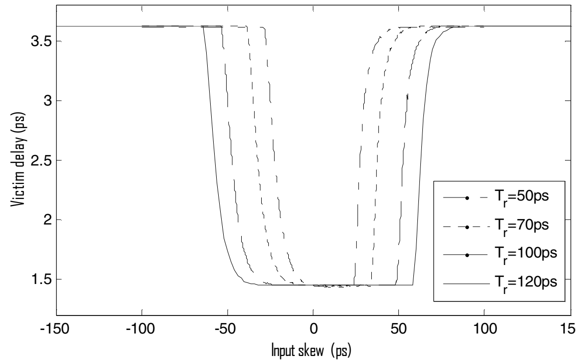

Figure 1: Delay change curve of a pair of 200 μm coupled intermediate interconnects in vertical tier-to-tier path (PTM 65nm technology). Linear driver model with 50Ω driver resistance is used. The load capacitance of each wire is 3*fF*.

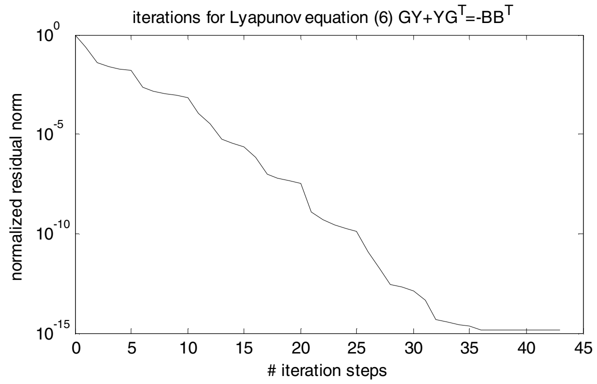

Figure 2: Convergence history of residual form. Convergence is obtained after 36 iterations.

[1] Since RLC interconnect is a linear system, a system of multiple aggressors can be analyzed by superposition for efficient timing and noise analysis. This is also an accepted approximation in the case of nonlinear driver model. As a result, we illustrate two coupled interconnects in this paper.

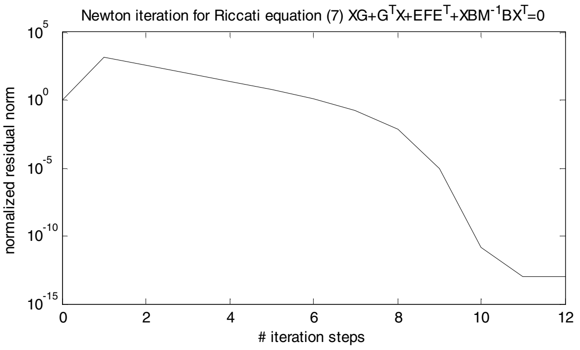

Figure 3: Convergence history of the normalized residual form of the Newton double step iteration (9) for solving the Riccati equation (7).

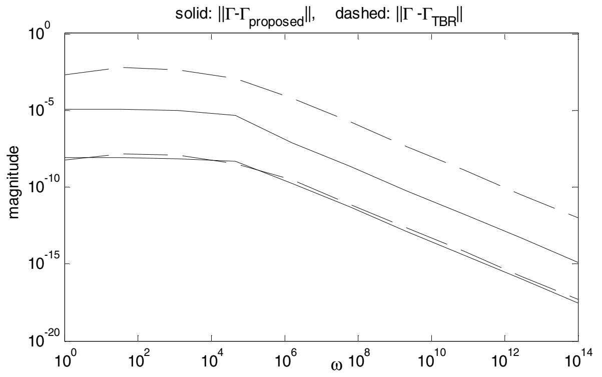

Figure 4: The Bode magnitude plot of the approximation errors.

In this algorithm the (approximate) solution of the Riccati equation is provided as a low rank Cholesky factor product [20] rather than an explicit dense matrix. The algorithm requires much less computation compared to the standard implementation, where Lyapunov is solved directly by the Bartels-Stewart or the Hammarling method. Note that the number of iteration steps needs not to be fixed *a priori*. However, if the Lyapunov equation should be solved as accurate as possible, correct results are usually achieved for low values of stopping criteria that are slightly larger than the machine precision [21]. The *cpu*-time needed to solve the Riccati equation inside the iteration is 0.77 seconds. Figure 4 illustrates a comparison with the commonly used truncated balance realization (TBR) method [6]. When very accurate Gramians are selected, the approximation error of the reduced system is very small compared to the Bode magnitude function of the original system. The lower two curves correspond to the highly accurate reduced system; the proposed model order reduction technique delivers a system of lower order. The *cpu* time of the proposed method is 11.47 seconds versus 19.64 seconds for the TBR method. The upper two denote k=15 reduced orders; the proposed technique delivers two orders of magnitude better accuracy. The *cpu* time of the proposed method for k=15 reduced order is 8.35 seconds. On the other hand, the TBR method requires 14.64 seconds *cpu* time.

IV. Conclusions

By adopting parameter dimension reduction techniques, 3D interconnect model extraction can be performed in a reduced parameter space, thus providing significant reductions on the required simulation samples for constructing accurate models. In this paper, within this framework, we present an efficient methodology for coupled 3D interconnect model reduction based on adjusted balanced stochastic truncation. Extensive experiments are conducted on a large set of random test cases, showing very accurate results.

References

[1] W. Topol, et al., “Three-dimensional circuits”, *IBM Jour. Res. Devel.,* vol. 50, n. 4/5, pp. 491-506, 2006.

[2] L.T. Pillage, R.A. Rohrer, “Asymptotic waveform evaluation for timing analysis,” *IEEE Trans. CAD*, vol. 9, no. 4, pp. 352-366, 1990.

[3] P. Feldmann, R.W. Freund, “Efficient linear circuit analysis by Pade approximation via the Lanczos process,” *IEEE Trans. on CAD*, vol. 14, no. 5, pp. 639-649, 1995.

[4] A. Odabasioglu, M. Celik, L. Pileggi, “PRIMA: Passive reduced-order interconnect macromodeling algorithm,” *IEEE Trans. CAD*, vol. 17, no. 8, pp. 645-654, 1998

[5] B.C. Moore, “Principal component analysis in linear systems: controllability, observability, and model reduction,” *IEEE Trans. Auto. Contr.*, vol. 26, no. 1, pp. 17-31, 1981.

[6] J. Li, J. White, “Efficient model reduction of interconnect via approximate system Grammians,” *Proc. IEEE ICCAD*, pp. 380-384, 1999.

[7] A. Zjajo, Q. Tang, M. Berkelaar, N. van der Meijs, “Balanced truncation of a stable non-minimal deep-submicron CMOS interconnect,” *Proc. IEEE ICICDT*, pp. 1-4, 2011.

[8] N. Wong, V. Balakrishnan, “Fast positive-real balanced truncation via quadratic alternating direction implicit iteration,” *IEEE Trans. CAD*, vol. 26, no. 9, pp. 1725-1731, 2007.

[9] J.R. Phillips, L. Daniel, L.M. Silveira, “Guaranteed passive balancing transformations for model order reduction,” *IEEE Trans. CAD*, vol. 22, no. 8, pp. 1027-1041, 2003.

[10] T. Penzl, “A cyclic low-rank Smith method for large sparse Lyapunov equations,” *SIAM Jour. on Sc. Comp.*, vol. 21, pp. 1401-1418, 2000.

[11] M.G. Safonov, R.Y. Chiang, “A Schur method for balanced-truncation model reduction,” *IEEE Trans. Auto. Cont.*, vol. 34, no. 7, pp. 729-733, 1989.

[12] M. Green, “Balanced stochastic realizations,” *Lin. Algebra Appl.*, vol. 98, pp. 211-247, 1988.

[13] A.E. Ruehli, A.C. Cangellaris, “Progress in the methodologies for the electrical modeling of interconnects and electronic packages,” *Proc. of IEEE*, vol. 89, no. 5, pp. 740-771, 2001.

[14] A. Mercha et al., “Comprehensive analysis of the impact of single and arrays of through silicon vias induced stress on high-k / metal gate cmos performance,” *Proc. IEEE IEDM*, pp. 2.2.1-2.2.4., 2010.

[15] G. Golub, C. van Loan, *Matrix computations*, Johns Hopkins University Press, Baltimore MD, 1996.

[16] “International technology roadmap for semiconductors”, available at http://www.itrs.net/links/2010itrs/home2010.htm

[17] R. Jagtap, “A methodology for early exploration of TSV interconnects in 3D stacked ICs,” MSc Thesis, *Delft University of Technology*, 2011.

[18] I. Savidis, E. Friedman, “Closed-form expressions of 3-d via resistance, inductance, and capacitance,” *IEEE Trans. Electron Dev.*, vol. 56, no. 9, pp. 1873-1881, 2009.

[19] “Predictive technology model” available at http://ptm.asu.edu/

[20] T. Reis, T. Stykel, “PABTEC: Passivity-preserving balanced truncation for electrical circuits,” *IEEE Trans. CAD*, vol. 29, no. 9, pp. 1354-1367, 2010.

[21] The Numerics in Control Network, available at http://www.win.tue.nl/wgs/niconet.html