

Balancing-Based Model Reduction for Fast Power Integrity Verification

Original

Balancing-Based Model Reduction for Fast Power Integrity Verification / Carlucci, A., Grivet-Talocia, S., Mongrain, S., Kulasekaran, S., Radhakrishnan, K.. - ELETTRONICO. - (2023), pp. 1-3. (2023 IEEE 32nd Conference on Electrical Performance of Electronic Packaging and Systems (EPEPS) Milpitas, CA, USA 15-18 October 2023) [10.1109/EPEPS58208.2023.10314870].

Availability:

This version is available at: 11583/2985822 since: 2024-02-09T10:49:56Z

Publisher:

IEEE

Published

DOI:10.1109/EPEPS58208.2023.10314870

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

IEEE postprint/Author's Accepted Manuscript

©2023 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

as $\mathbf{w} = \mathbf{\Delta}(d)\mathbf{z}$. During transient analysis, system (1) is augmented with models of the controllers that sense the output voltage and return per-core duty cycle signals in a closed-loop configuration (not shown, see [1] for details).

Considering a nominal duty cycle value as the operating point, the dynamics of the PDN can be approximated with a small-signal linearization as described in [2]. Hence, for the purposes of the formulation, we can focus on a generic linear system described by a state-space realization $(\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D})$.

III. FORMULATION

Balanced Truncation (BT) [3] is a well known method to obtain a simplified model of reduced state-space dimension n that preserves the input-output behavior of a given full-order system. BT finds principal directions in the state space that are simultaneously most controllable and observable using a pair of special matrices, namely the controllability Gramian \mathbf{P} and the observability Gramian \mathbf{Q} . Based on these, projection matrices \mathbf{S}_r and \mathbf{S}_l can be constructed [3], [4], which are used to remove the least important states. The most appealing feature of BT is the availability of an explicit error bound on the approximation error, computed from the truncated *Hankel singular values* [5].

A. Approximate Gramians via Vector Fitting

Computation of system Gramians \mathbf{P} , \mathbf{Q} can be carried out by solving two associated Lyapunov equations. For very large-scale systems, this direct solution becomes too computationally expensive, so that alternative methods have been devised, e.g. iterative methods [6]. A simpler way of approximating \mathbf{P} and \mathbf{Q} , proposed in [7], is based on the following identities

$$\begin{aligned} \mathbf{P} &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \mathbf{X}(j\omega) \mathbf{X}^T(-j\omega) d\omega \\ \mathbf{Q} &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \mathbf{X}_d(j\omega) \mathbf{X}_d^T(-j\omega) d\omega \end{aligned} \quad (2)$$

with $\mathbf{X}(j\omega) = (j\omega\mathbb{I} - \mathcal{A})^{-1}\mathcal{B} \in \mathbb{C}^{N \times P}$ and $\mathbf{X}_d(j\omega) = (j\omega\mathbb{I} - \mathcal{A}^T)^{-1}\mathcal{C}^T$. In [7], it is suggested that these integrals can be evaluated by quadrature rules, so that \mathbf{P} is approximated as a suitably weighted sum of the snapshots $\mathbf{X}(j\omega_k)$ computed at K nodes $\{\omega_k\}_{k=1}^K$, and similarly for \mathbf{Q} . Here, we make a different use of these snapshots as we propose an alternative way of approximating the Gramians. The key observation is that, if an (approximate) pole-residue expansion

$$\mathbf{X}(s) \approx \check{\mathbf{X}}(s) = \sum_{i=1}^{\nu} \frac{\mathbf{R}_i}{s - q_i} \quad (3)$$

is available in terms of ν stable poles $\{q_i\}_{i=1}^{\nu}$, then the computation of the integral can be carried out analytically. First, we replace $\mathbf{X}(s)$ in (2) with the approximation $\check{\mathbf{X}}(s)$. Then the residue of the integrand at each left-half plane pole q_i is given by $\mathbf{R}_i \check{\mathbf{X}}^T(-q_i)$. At this point, the Residue Theorem is invoked [8, App. E.2] to express the integral along the imaginary axis in terms of these residues,

$$\mathbf{P} \approx \sum_{i=1}^{\nu} \mathbf{R}_i \check{\mathbf{X}}^T(-q_i).$$

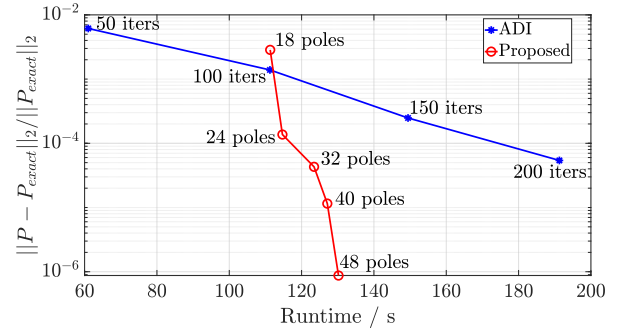


Fig. 2. Convergence of the proposed method compared to the ADI method [10]. The error norm is reported as a function of overall runtime.

The pole-residue approximation (3) is here determined using the Vector Fitting algorithm to find suitable basis poles q_i , starting from a limited number K of evaluations $\mathbf{X}(j\omega_k)$. Once the basis poles are known, the residues \mathbf{R}_i in (3) are computed by exploiting the optimality condition given in [9], which in our setting reads

$$\sum_{i=1}^{\nu} \frac{\mathbf{R}_i}{-q_j - q_i} = \mathbf{X}(-q_j), \quad j = 1, \dots, \nu,$$

where $\mathbf{X}(-q_j)$ are ν additional evaluations of $\mathbf{X}(s)$ at the mirror images of the basis poles. For each matrix entry (h, m) of \mathbf{R}_i , this condition is a $\nu \times \nu$ linear system giving the ν unknown residue entries $(\mathbf{R}_i)_{hm}$, $i = 1, \dots, \nu$. The coefficient matrix for this linear system is derived from the Cauchy matrix $\mathbf{\Phi}$ defined by $(\mathbf{\Phi})_{ij} = (-q_i - q_j)^{-1}$ as follows

$$(\mathbf{R}_1 \quad \dots \quad \mathbf{R}_\nu) (\mathbf{\Phi} \otimes \mathbb{I}_P) = (\mathbf{X}(-q_1) \quad \dots \quad \mathbf{X}(-q_\nu))$$

IV. NUMERICAL RESULTS

A. Validation of MOR via approximate Gramians

We assess the proposed algorithm for Gramian computation via VF. Results are presented for the input network of the PDN of an Intel-based enterprise server (see Fig. 1). The starting point is a passive LTI network with 181 ports and dynamic order 6170. The rational approximation (3) is determined using $K = 70$ log-spaced frequency values in $[0, 10]$ GHz. Figure 2 shows that, as ν increases, the error between the approximate Gramian \mathbf{P} and the exact solution computed via Lyapunov equation (direct method) decreases. Moreover, proposed method takes a shorter time to produce a more accurate solution with respect to ADI. Figure 3 (bottom panel) compares two selected responses of the full-size system and the reduced-order model. The latter is basically indistinguishable from the reference even if the number of states was reduced to 700, as confirmed by the Hankel singular values (top panel), which provide the explicit bound on the model approximation.

B. Modeling a full-system Power Distribution Network

We now consider a complete Power Distribution Network including multi-phase FIVRs ($N_p = 3$) for $N_c = 60$ cores,

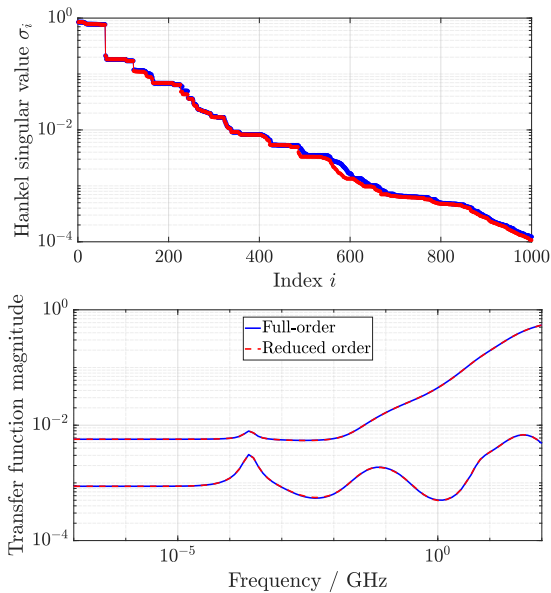


Fig. 3. The top panel shows the (normalized) system singular values of the input network computed with exact Gramians (blue) and approximate ones (red). The bottom panel shows two matrix entries of the input network transfer function computed using a reduced model of order 700, much lower than the initial 6170 states.

extracted from an Intel-based enterprise server platform. Each core has $N_o = 57$ outputs, leading to a total number $N_c N_o = 3420$ of output voltages to be stabilized. For this system, the proposed approach was applied to compute Gramians of the locally linearized approximation around the nominal duty cycle. The resulting bases were used in a structured projection with passivity preservation similarly to [2]. Figure 4 shows the results of a transient analysis where excitation currents range between 0 and 20 A/core. All cores are persistently excited starting from $t = 0.1 \mu\text{s}$ with diversified current profiles. For this test case, we applied MOR to reduce the dynamic order from more than $5 \cdot 10^4$ down to 400. Transient simulation of the reduced system takes only 66 s, corresponding to a speedup of more than $15 \times$ with respect to the full-order simulation (1038 s), both carried out through a simple MATLAB solver based on the Backward Euler method. The maximum error among all output voltages is lower than 2 mV at all times.

In order to enable a comparison with less efficient reference approaches, a simplified version of the same structure was realized by including only 16 cores. The results of this comparison are reported in Table I, which confirms that proposed method provides clear improvements with respect to [2] in terms of accuracy for a given order, and with respect to [1] in both accuracy and efficiency.

V. CONCLUSIONS

This paper presented a method derived from balanced truncation that is applicable to large-scale systems, with a particular focus on acceleration of power integrity verification analyses for multi-core microprocessors. Besides proposing a novel procedure to compute approximate Gramian matrices

TABLE I
COMPARATIVE NUMERICAL ANALYSIS BASED ON 16-CORES BENCHMARK

| Method | Order | Max. error | Runtime |
|--------------------------------|-------|------------|---------|
| Full order (HSPICE) | - | - | 1410 s |
| Full order (MATLAB) | 18074 | - | 139 s |
| Approx. balancing (this paper) | 250 | 0.9 mV | 6 s |
| Moment-matching as in [2] | 250 | 1.6 mV | 6 s |
| Parametric fitting as in [1] | - | 24 mV | 181 s |

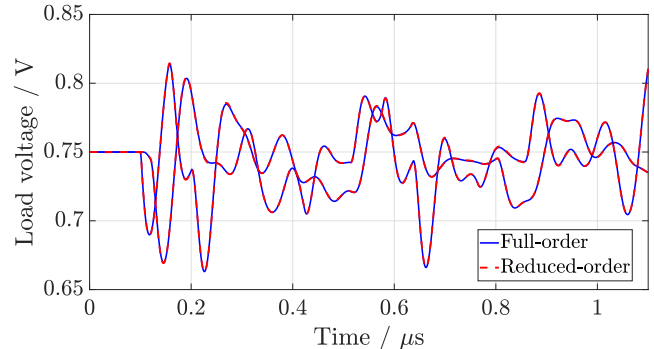


Fig. 4. Transient analysis comparing load voltage waveforms at two particular load ports obtained from the full-order model and the reduced-order model of the PDN of a 60-core Intel-based enterprise server.

using rational fitting, the presented approach improves on existing work [2] by providing higher accuracy and guidance in choosing the reduced model order through the concept of system singular value, inherited from balanced truncation. Major speedup factors in transient simulation are observed.

REFERENCES

- [1] A. Carlucci, T. Bradde, S. Grivet-Talocia, S. Mongrain, S. Kulasekaran, and K. Radhakrishnan, "A compressed multivariate macromodeling framework for fast transient verification of system-level power delivery networks," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, 2023. [Online]. Available: <https://doi.org/10.1109/TCPMT.2023.3292449>
- [2] A. Carlucci, S. Grivet-Talocia, S. Mongrain, S. Kulasekaran, and K. Radhakrishnan, "A structured Krylov subspace projection framework for fast power integrity verification," in *2023 IEEE 27th Workshop on Signal and Power Integrity (SPI)*, 2023, pp. 1–4.
- [3] B. Moore, "Principal component analysis in linear systems: Controllability, observability, and model reduction," *IEEE Transactions on Automatic Control*, vol. 26, no. 1, pp. 17–32, Feb. 1981.
- [4] M. Safonov and R. Chiang, "A Schur method for balanced-truncation model reduction," *IEEE Transactions on Automatic Control*, vol. 34, no. 7, pp. 729–733, Jul. 1989.
- [5] K. Zhou, J. C. Doyle, and K. Glover, *Robust and optimal control*. Prentice Hall Upper Saddle River, NJ, 1996.
- [6] J.-R. Li and J. White, "Low rank solution of Lyapunov equations," *SIAM Journal on Matrix Analysis and Applications*, vol. 24, no. 1, pp. 260–280, 2002.
- [7] J. R. Phillips and L. M. Silveira, "Poor man's TBR: a simple model reduction scheme," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 24, no. 1, pp. 43–55, 2005.
- [8] S. Grivet-Talocia and B. Gustavsen, *Passive macromodeling: Theory and applications*. John Wiley & Sons, 2015.
- [9] S. Gugercin, A. C. Antoulas, and C. Beattie, " \mathcal{H}_2 model reduction for large-scale linear dynamical systems," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 2, pp. 609–638, 2008.
- [10] J. Saak, M. Köhler, and P. Benner, "M-M.E.S.S.-2.1 – the matrix equations sparse solvers library," Feb. 2022, see also: <https://www.mpi-magdeburg.mpg.de/projects/mess>.