GPU-Accelerated Non-Linear Analog and Mixed-Signal Circuit Transient Simulation

Chi-Un Lei, Ka Lok Man, Nan Zhang and Yanyan Wu

***Abstract*—Deep sub-micron technology enables a high integration of (non-linear) circuit systems for electronic products. In order to design non-linear circuits with guaranteed reliability, a sophisticated simulation procedure is required. However, in existing GPU-enhanced circuit simulation methodologies, numerous data transfers between cores hamper the benefits of GPU computations. In this project, we focus on the parallelization of computations for nonlinear analog/mixed-signal circuit transient simulation. By applying multi-dimensional inverse Laplace Transform, each time-sampled points of the transient response can be computed independently on the GPU architecture. Power-efficient hardware architecture is also explored for a further speedup. With the developed platform, circuit designers can verify circuit's dynamic properties quickly in order to meet the need of time-critical and high-yield designs.**

***Index Terms*—Graphic processing unit, circuit simulation, computer-aided design (CAD)**

I. Introduction

Deep sub-micron technology enables a high integration of circuit systems for very-large-scale integration (VLSI) electronic products. These systems contain high-performance and sensitive analog/mixed-signal (AMS) circuits with nonlinear characteristics. Moreover, due to the continuous shrinking size of process technologies, unreliability effects affect performance and production yield of circuits. In order to develop circuits with guaranteed reliability, large-scale circuit-level reliability analysis is required. In order to meet the time-critical design requirement, the designed circuit should be verified within a reasonable time. Therefore, transient simulation of circuits is one of the key component in VLSI circuit design process [1]–[4]. The key challenge faced by simulator designers is how to accelerate the simulation while taking modeling applicability into account.

Emerging simulation techniques have been proposed recently. Among those techniques, computation using graphics processing unit (GPU) architectures is one of the promising candidates [5], [6]. It is because GPU is good at handling computation-intensive operations. However, existing time-domain solvers approximate AMS circuits by linear models and use Newton-Raphson iteration to compute transients. As a result, accuracy has been sacrificed. Furthermore, such computation always requires numerous data transfers between CPU processors and GPU processors through the relatively slow communication channel. Furthermore, the issue of high power dissipation becomes apparent, especially for systems with a high integration of processors. These bottlenecks hamper the benefits of GPU computations. Therefore, it remains being a hot area for electronic design automation (EDA) researchers to develop a GPU-based solver for AMS circuit simulation.

In this project, we aim to develop a frequency domain nonlinear transient simulator for AMS circuits. In particular, we will develop highly parallelized frequency domain computation algorithms for time domain transient simulations, as well as to develop hardware architectures for large scale GPU computations. With the developed simulator, we can verify the circuit's dynamic properties quickly in order to meet the requirement of time-critical and high-yield designs.

II. Proposed Solutions

Given the schematic or model of a circuit, time-domain transient responses of the circuit will be computed. An overview of the proposed simulation flow is shown in Fig. 1. We will develop algorithms as well as micro-architectures to speed up the simulation.

A. Parallelized Multi-dimensional Inverse Laplace Transform

We will apply multi-dimensional (MD) Inverse Laplace Transform (ILT) and Laguerre Functions for time-domain transient response computations. MD ILT is a generalized 2D ILT [7] which effectively constructs time-sampled responses from its frequency-sampled responses [8]. The MD Laplace Transform of a generalized time-domain function $f\left(t_1, t_2, \ldots, t_n\right)$ can be represented as

$$F\left(s_1, s_2, \ldots, s_n\right) = \int_0^\infty \int_0^\infty \cdots \int_0^\infty f\left(t_1, t_2, \ldots, t_n\right) e^{-s_1 t_1 - s_2 t_2 \ldots - s_n t_n} dt_1 dt_2 \ldots dt_n. \tag{1}$$

Similarly, the time-domain transient construction, i.e., the the numerical MD ILT computation, can be determined by the approximation of $f\left(t_1, t_2, \ldots, t_n\right)$ when numerical values of the MD function $F\left(s_1, s_2, \ldots, s_n\right)$ are known:

$$f\left(t_1, t_2, \ldots, t_n\right) = \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \cdots \sum_{n_N=0}^{\infty} q_{n_1, n_2, \ldots, n_N} l_{n_1}\left(t_1\right) l_{n_2}\left(t_2\right) \ldots l_{n_N}\left(t_N\right), \tag{2}$$

Manuscript received January 13, 2012.

C.-U. Lei is with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong. Email: culei@eee.hku.hk

K.L. Man is with the Department of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University, China, Myongji University, South Korea and Baltic Institute of Advanced Technology, Lithuania. Email: ka.man@xjtlu.edu.cn

N. Zhang is with the Department of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University, China. Email: nan.zhang@xjtlu.edu.cn

Y. Wu is with the Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, China. Email: yanyan.wu@xjtlu.edu.cn

The research work presented in this paper is partially sponsored by SOLARI (HK) CO (www.solari-hk.com) and KATRI (www.katri.co.jp and www.katri.com.hk).

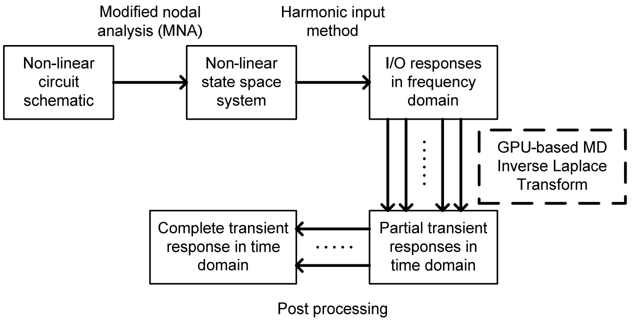
Modified nodal analysis (MNA)
Harmonic input method
Non-linear circuit schematic
Non-linear state space system
I/O responses in frequency domain
GPU-based MD Inverse Laplace Transform
Complete transient response in time domain
Partial transient responses in time domain
Post processing

Fig. 1. Overview of an AMS circuit simulation flow.

where $t_1 = t_2 = \cdots = t_n = t$, t is the sampling time, $\{q_{n_1,n_2,\ldots,n_N}\}$ and $\{l_{n_i}(t_i)\}$ are functions of $F(s_1, s_2, \ldots, s_n)$, and can be determined by computing Laguerre Functions.

Comparing to the matrix-valued computation in existing algorithms, computing MD ILT on a GPU architecture is efficient because MD ILT can be converted into computationally favorable Fast Fourier Transform (FFT) easily. Furthermore, MD ILT for each time-sampled point can be determined independently. Therefore, all sampling points can be computed simultaneously on a multi-processor system.

Currently, preliminary frequency-domain simulation has been developed in the Matlab environment, and has been verified by a non-linear CMOS circuit and a non-linear RC ladder. In order to leverage fast computation using parallelized processors, the work will be proceeded by reducing the computation load, developing an adaptive time-step transient simulation with arbitrary input responses.

B. Architecture for Parallelized Multi-Processor GPU Computations

Data transfer between CPU and GPU is one of the bottlenecks in GPU computations. Therefore, we will develop a CPU-GPU architecture with a low-overhead data flow, as shown in Fig. 2. At the beginning of every iteration, all frequency-sampled data is computed by CPU, and then is transferred to GPU. Given frequency-sampled data as input, the time-domain response is computed distributedly, by computing Laguerre Function via FFT. Obtained results are then transferred from GPU to CPU. Finally, post-processing is done by CPU. This distributed computation can significantly reduce data transmission between cores.

We have built a prototype on a CPU-GPU architecture. The work will be preceded by developing algorithms for program parallelization, in order to i) move the sequential computation and parallelizable computation to CPU and GPU, respectively, ii) utilize the memory usage in cores, and iii) utilize the transmission bandwidth between cores.

C. Power-Performance Tradeoff of the GPU-CPU Architecture

Power dissipation (and heat generation) of the hardware architecture will limit the utilization and scalability in many-core computation. Therefore, power efficiency become a new concern in the computation research community. We plan to find out the power-performance tradeoff for an improved hardware configuration. We will analyze the power throughput ratio ($\frac{\text{energy}}{\text{cycle}}$) and energy throughput ratio ($\frac{\text{power}}{(\text{throughput})^2}$) of single GPU systems, multi-GPU systems and (hybrid) CPU-GPU systems. Meanwhile, we will also analyze the power dissipation of the data movement and control flow movement. Based on the analysis, we can develop a computation partition strategy for power-efficient simulations.

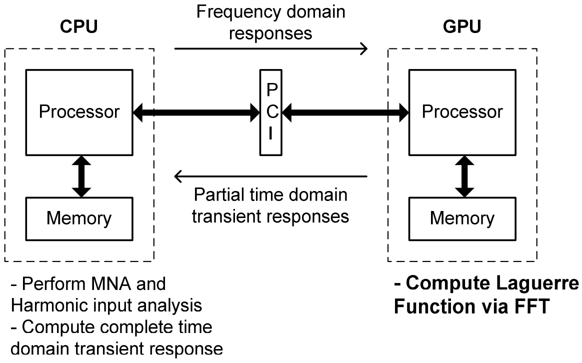
CPU
Frequency domain responses
GPU
Processor
P C I
Processor
Memory
Partial time domain transient responses
Memory
- Perform MNA and Harmonic input analysis
- Compute complete time domain transient response
- Compute Laguerre Function via FFT

Fig. 2. Data flow and operations in the GPU-CPU architecture.

III. Conclusion

In this paper, we have proposed a GPU-CPU architecture to accelerate the circuit simulation process. A prototype has been developed, and will be used for power-performance analysis. By developing parallelized algorithms and dedicated micro-architectures, we aim to obtain a "10X speed-up" simulation and make large-scale AMS circuit simulations practical. Results of this study can be extended to several interesting directions for future research, namely:

1) development of algorithms and data structures for thread scheduling, memory allocation and pointer operation for the architecture;
2) development of a GPU-FPGA-CPU architecture for a low-overhead computation;
3) development of a real-time circuit simulation methodology for seamless circuit design.

References

[1] F. N. Najm, *Circuit Simulation.* John Wiley & Sons, Inc., 2010.

[2] C. U. Lei, Y. Wang, Q. Chen, and N. Wong, "On vector fitting methods in signal/power integrity applications," in *Proc. IAENG International MultiConference of Engineers and Computer Scientists*, Mar. 2010, pp. 1407–1412.

[3] C. U. Lei, K. Man, and Y. Wu, "VLSI macromodeling and signal integrity analysis via digital signal processing techniques," in *Proc. IAENG International MultiConference of Engineers and Computer Scientists*, Mar. 2011, pp. 1031–1032.

[4] C. U. Lei and N. Wong, "WISE: Warped impulse structure estimation for time-domain linear macromodeling," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 2, no. 1, pp. 131–139, Jan. 2012.

[5] K. Gulati, J. Croix, S. P. Khatri, and R. Shastry, "Fast circuit simulation on graphics processing units," in *Proc. IEEE Asia and South Pacific Design Automation Conference*, Jan. 2009, pp. 403–408.

[6] F. Croix, "Introduction to GPU programming for EDA," in *Proc. IEEE International Conference on Computer-Aided Design*, Nov. 2009, pp. 276–280.

[7] P. Li and L. Pileggi, "Compact reduced-order modelling of weakly nonlinear analog and RF circuits," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 23, no. 2, pp. 184–203, Feb. 2005.

[8] J. Abate, G. Choudhury, and W. Whitt, "Numerical inversion multidimensional Laplace transforms by the Laguerre method," *Performance Evaluation*, vol. 31, pp. 229–243, 1998.