Communication Centric On-Chip Power Grid Models for Networks-On-Chip

Nizar Dahir  
School of EECE,  
Newcastle University  
Newcastle Upon Tyne, UK  
nizar.dahir@ncl.ac.uk

Terrence Mak  
School of EECE,  
Newcastle University  
Newcastle Upon Tyne, UK  
terrence.mak@ncl.ac.uk

Alex Yakovlev  
School of EECE,  
Newcastle University  
Newcastle Upon Tyne, UK  
alex.yakovlev@ncl.ac.uk

Abstract— Adverse effects of unreliable on-chip power supply delivery are exacerbated due to the rapid shrinking of device dimensions and the ever increasing power consumptions in nanometre-scale integration. Power supply integrity becomes a critical concern. Particularly, on-chip communication networks, such as networks-on-chip (NoC), dictates the power dissipations and overall system performance in multi-core systems and emerging embedded computing architectures. These new communication centric architectures require dedicated power grid model that embeds distinctive communication characteristics and spatial parameters for analysing impacts of IP-drop and power supply noise. In this paper, we present a new on-chip power delivery model that captures the on-chip communication patterns and power grid dynamics. This model integrates cycle-accurate simulation of networks-on-chip to analyze the impact of different design entities on power supply noise. The model has been rigorously evaluated using SPICE simulation. Novel observations of power delivery integrity due to communication network design are presented. This model provides a unique and communication-centric perspective to analyse power supply integrity that leads to future robust and reliable multi-core system design.

Keywords: Networks-on-chip, power supply noise, power grid simulation, on-chip routing.

I. INTRODUCTION

Power supply noise has an adverse effect on digital circuit performance and the reliability. It could result in signal deterioration and create soft errors. Recently, it has been reported that variation in power supply would have significant impact in operational frequency and system power dissipation [1, 2]. Both resistive and inductive voltage drop are sources of power supply noise. The on-chip resistive voltage drop occurs mainly due to the variation of power delivery in the power grid network. The inductive drop is mainly due to wire inductance in the package and is proportional to the current demand.

Technology scaling exacerbates the problem of power supply noise. Wire in the power network is shrinking and resistance is substantially increasing. On the other hand, demand of power delivery is rapidly increasing due to the large-scale integration and higher switching frequency. For example, the maximum voltage drop as a percentage of supply voltage is rapidly increasing as a result of technology scaling (See Figure 1). Mitigating power supply noise becomes a grand challenge for the sustainability of future large-scale integration development.

Different approaches have been developed for mitigating the on-chip power supply noise problem. In [3, 4], techniques are proposed to determine the optimal values and positions of decoupling capacitors for minimizing the power noise during floor-planning.

Optimal power-gating scheduling [6, 7] is also proposed to minimize the voltage drop caused by the switching activities in the gated block. For multi-core systems, workload assignment can have significant impact on the power noise dissipation in the system. In [8], a simulated-annealing approach is employed to optimize assignment of workloads to the cores, such that the resulting power noise can be minimized.

The emerging multi-core system requires a dedicated and high-performance on-chip communication system. Networks-on-chip (NoC) had been proposed as a promising infrastructure to deliver scalable and high-performance on-chip inter-core communication [11]. Power budget of NoC takes up a significant overall proportion in a chip. For example, the routers at the MIT RAW on-chip network [10] consumes 40% of the tile power, and the communication network takes up to 35% of the overall system power. A similar measurement is found for the Intel’s 80-tile TeraFLOPS prototype where the communication power budget is about 28% [9]. Despite the aforesaid studies and methods developed for the on-chip power grid, power delivery and load for on-chip communication systems have not been studied. In this paper, we present a computational model for on-chip power delivery network and this model is
dedicated for networks-on-chip. The major contributions of this paper are summarized as follows:

1. Present an integrated model of power supply grid, in which detailed on-chip communication dynamics including communication traffic pattern and communication link bandwidth are considered.
2. Introduce a coarse-grained power grid model that can simplify the computational complexity for voltage distribution evaluation. The trade-off between numerical accuracy and computational time is evaluated.
3. Verification of the bus model, which is employed for modelling the communication link, thorough evaluation using SPICE model, has been carried out and an insignificant error of the model is found.
4. The model has been employed to analyse the power supply noise in networks-on-chip. Various NoC specific parameters are thoroughly investigated. Novel observations about power supply noise hot-spot distribution and variation of noise due to routing algorithms and traffic patterns are found.

II. BACKGROUND
On-chip power grid system comprises of two fundamental models, which are models of workload and power grid. The workload model determine the electrical load that is produced by the computational capacity and the power grid model determines the grid node voltages with respect to the workload performance. This section surveys the state-of-the-art in these two areas.

A. Power Grid Model
To analyze the power delivery grid integrity in VLSI circuits, the grid is modelled as an RLC network while loads are often modelled as independent current sources or equivalent passive elements. By solving this model to determine the node voltages is difficult due to the enormous number of elements and nodes in the grid. This problem has been considered by many researchers and several solutions have been proposed to solve the power grid size problem during both simulation and modelling.

Examples such as simulating large power grid as multigrid-like [10], hierarchical [11, 12] or partition-based [13] were proposed. These methods are based on iterative computation and they are not suitable for an analysis that involves solving the model several times. Alternative direct methods are preferable in this case. In [14] a fast and direct model to determine the peak voltage noise induced by switching cores is proposed. The minimum value of a node voltage is given by

$$V_j^{\text{min}} = \frac{1}{\lambda_j} \left( \sum_{i=1}^{k} x_{ij} V_i^{\text{min}} + \frac{1}{2} \sum_{i=1}^{k} C_{ij} V_{dd} \right)$$ (1)

where \(\lambda_j = \sum_{i=1}^{k} x_{ij} \lambda_i + (1/2) \sum_{i=1}^{k} C_{ij} \) and \(x_{ij} = \frac{t_i^2}{(6L_{ij})^2 + 3R_{ij} \phi_j}, R_{ij}, L_{ij} \) and \(C_{ij} \) are the resistance inductance and capacitance between nodes \(i \) and \(j \) in the power grid, respectively, and \(C_{ij}^\text{Load} \) is the load at node \(j \). This model provides an accurate estimation and a maximum error of 5% was reported [14].

B. Work Load Model

Considerable amount of literatures have been published on workload model for power grid analysis. Techniques that represent the workload with equivalent passive elements were reported. In [15], a macro model based on the effective impedance of the current consumption is proposed. Other techniques that are based on independent current source model for the active blocks are developed. For example, macro-model can be used to determine the waveforms of these current sources. In [16], frequency domain current macro-model is proposed where the current wave from is determined where the input vector pairs of the modelled circuits are partitioned according to the hamming distance and a current macro-model is built for each distance using regression.

In this paper, we focus on the interconnects workload in networks-on-chip. Previously developed workload models will be elaborated and modified to model the of on-chip bus load that generated by the communication traffic.

C. On-Chip Bus Model
Independent current sources or static coupling capacitors are proposed to represent the on-chip computational load. This can be a reasonable representation for power consumption of simple logic or, to an extent, for microprocessors. For complex on-chip communication systems, such as networks-on-chip, dynamic power consumption through global interconnects imposes a significant variation in power and a load. Therefore, effective model that captures the communication dynamic is required.

A model for on-chip bus is essential for determining on-chip communication load. In the literature, a number of current [17], energy [18] and power [4, 7] based models for on-chip bus have been proposed. On-chip interconnects are modelled as capacitively and inductively coupled distributed RLC lines. In [17], an analytical model for on-chip bus current based on decoupling techniques is presented. On-chip bus wires are driven by exponential voltage source \( (V_o) \) and loaded by capacitor \( (C_j) \). Two port network with source and load impedances is used to derive a closed form for the wire current the wire. By employing decoupling transformation, the bus currents can be obtained.

III. METHODOLOGY
Power supply noise in networks-on-chip is introduced due to on-chip communication traffic. In contrast to conventional model for logic or microprocessors, the engendered noise or IR drop would have an interesting correlation to the spatial distribution of the traffic load and also a temporal pattern of the packet injection rate. In this paper, we integrate the power supply grid model with networks-on-chip traffic load, such as routing and communication patterns, to introduce a novel computational power supply grid model dedicated for NoCs. The proposed model will be detailed in this section.

A. Networks-on-Chip
Networks-on-chip (NoCs) have been proposed as an alternative to the bus-based on-chip interconnectivity to solve many problems associated with the latter such as high wire length and poor scalability [8, 19]. NoCs are used to connect components on the same chip. The transfer of data is achieved in a way similar to conventional packet switched computer networks. NoC uses packet switching and packets are routed...
from the source to the destination. The packet is split into smaller data units called flits. The interconnected components can be general purpose microprocessors, memory blocks or control circuitry. Each component (IP) is attached to a router which is used as a gateway to connect the IP to other IPs and to route information for the overall system. The term tile is often used to stand for the IP core and router.

Many design parameters can affect the spatial and temporal distribution of communication loads in NoCs. Routing algorithm, traffic pattern, and packet injection rate are examples of these parameters. Routing algorithms can be deterministic, semi-adaptive or fully-adaptive. Examples of deterministic routing are XY. DyAD [20] and odd-even are examples of semi-adaptive routing algorithms.

Communication load has spatial and temporal distributions both are determined by the design entities of the system. This distribution in time and space of communication loads is reflected as spatial and temporal voltage drop in the power delivery grid.

B. Model Overview

To determine the effect of each of these parameters (or combination thereof) on the power grid, traffic resulting from these parameters need to be employed with an on-chip bus and power grid models.

Figure 2 illustrates the NoCs communication load effect on the power grid. When a sending router sends data flit the bus drivers draw current that is delivered through the power grid causing a Weibull shaped pulse of current. This pulse of current affects the supply voltage in the grid due to the resistive and inductive voltage drops in the power grid interconnects.

The on-chip bus RLC equivalent electrical circuit is shown in Figure 3 with the bus drivers and loads. The bus drivers are modelled as exponential voltage sources while the bus loads are modelled as capacitances. The NoC router functional block diagram is also illustrated. In this work we consider the bus switching load, i.e. the power delivered to the bus drivers through the power grid. Figure 4 summarizes the components and the parameters considered this work.

C. Compartamental Modelling for Communication Fabrics

Routing the data packets in systems-on-chip (SoCs) that use NoCs to communicate is performed using routers. Routers include a cross-bar switch (see Figure 3) with four input four output channels (assuming mesh topology) for global communication (north, east, south and west) and on input and one output for local communication. In our model we assume that the bus is driven by data sending circuitry this implies that the load considered is the flit sending load for each router channel. Also we neglect the sending of flits to local channels since the main concern of this work is global communication load.

Bearing the above assumptions in mind and the fact that the power grid as whole is a linear network which implies that the effect of each load may be accumulated linearly we compute the bus driving load at node j in the power grid at the kth simulaton cycle \( c_{j}^{load}(k) \) as follows:

\[
C_{j}^{load}(k) = \sum_{d\in\{N,E,S,W\}} f(sw(k,ch_{i,j}^{d}))
\]

where \( j \) is a node in the power grid. \( ch_{i,j}^{d} \) is channel \( d \) \((d \in \{N,E,S,W\}) \) of the router that is supplied with power through node \( j \) in the power grid. \( k \) is simulation cycle and \( sw \) is a vector of length \( n \), the bus size, containing elements resulting from the bus switching required to send a flit:

\[
sw(w) \in \{0,1,-1\} \quad w=1,2,... n
\]  

where \( sw(w) \) is 0 when the wire is quiet , 1 when the wire is switching up (0→1) and 1 when the wire is switching down (1→0). The function \( f(sw) \) is the function that computes the load equivalent capacitance \( C_{load} \) and can be expressed as:

\[
f(sw) = Q_{tot}/V_{DD} = \int I_{bus}(t)dt)/V_{DD}
\]

where \( Q_{tot} \) is the total charge delivered during bus switching and \( I_{bus} \) is the bus current that is computed using the bus model proposed by [17] after modifying the formula for the total bus current \( I_{bus} \) to include the direction of wire switching as follows:
\[ I_{bus} = \sum_{k=1}^{n} \sum_{i=1}^{n} M_{k,i} \delta w(i) \sum_{j=1}^{n} M_{i,j}^T l_i(t) \quad (5) \]

where \( I_i \) is wire current and \( M \) is the decoupling transformation matrix. The equivalent load capacitance resulting from Eq. (2) is used to characterize the load in the power grid model described in section II.

D. Power Grid Granularity

In power grid simulation there are two techniques for model solution [12] iterative and direct. The iterative is suitable for DC analysis and is used to obtain a single system solution when the time variation of workloads is not important. On the other hand the direct technique is used when multiple system solutions are necessary. This is the case in this work since we need to solve the system for each simulation cycle. We used coarse grained synthetic power grid in order to achieve significant result in a practical simulation time. Power grids designed for real VLSI circuits [21] may contain tens of thousands of nodes (neglecting the vias between different metal layers [1] which, if considered, will increase the number of nodes by a factor of two).

A coarse grid approach was used by [10] where the number of nodes in the power grid is reduced by creating a coarse-grained grid. The analysis is performed on the coarse grid and solution is mapped back to the original (fine) grid using linear interpolation taking into account the values of conductances between the nodes. Based on these assumptions and assuming a uniform regular mesh power grid, the voltage at certain node \( j \) in the grid with load \( L_j \) degrades by a factor of \( \alpha \) when the grid granularity increases while preserving the same grid structure. Voltage at node \( j \) for the fine grained grid is given by:

\[ V_{j}^{fi} = \alpha V_{j}^{lc} \quad (6) \]

where \( V_{j}^{fi} \) is the voltage at node \( j \) in the grid with segment length \( l_i \) and \( l_g \) are the fine and coarse segment lengths which is obtained as \( l_g = \frac{l_i}{g} \) where \( g \) is the granularity multiple. \( \alpha \) is the granularity effect factor. Experimental results show that \( \alpha \sim 1 \) for \( g > 1 \) as we will see in the results section.

IV. EXPERIMENTAL RESULTS

To evaluate our model, an NoC with size of a 4x4 mesh, 32-bit communication links, 1 mm inter-router distance and 1 GHz bus frequency is considered. Switching activities at data line follow a Gaussian random distribution and the node packet injection follows a Poisson distribution in line with [8]. For wire dimensions of the power grid and the bus interconnects, 65 nm technology with \( V_{DD} = 1 \text{V} \) are considered. The topology of the power grid is a uniform mesh with size 4 mm× and grid segment length of 100 μm. The RLC parasitic of the communication interconnects and grid are determined using the PTM model [22]. The router switching data are extracted using Noxim [23], which is a cycle accurate NoCs simulator based on SystemC.

Figure 6 shows the verification of the decoupling method used to determine the current drawn by the on-chip interconnects against SPICE simulation. The relative root mean square error is 5.8%. This model is reasonably accurate and can be employed to model the power delivery in a networks-on-chip.

To determine the accuracy of our model in terms of power grid granularity, we compared the resulting voltage drop caused by a range of loads in a fixed position for different power grid granularities. In line with [10, 24], wire-segment length and width is varied in order to to keep a linear proportionality with the total resistance. The resulting values of load degrading factor, \( \alpha \) (See Eq. (6) ) are shown in Table 1. Results showing that the coarse-grained model demonstrates voltage estimations have a close-resemblance to the fine-grain model. For example, an insignificant error, less than 0.2%, can be found using a coarse-grained model with a granularity of 256 times larger. Most importantly, the simulation time, which is the computational delay of the model, is significantly reduced from 10 minutes to less than 10 milliseconds. To evaluate the accuracy of the coarse-grained model, relative error of the coarse-grained voltage is evaluated with respected to the \( V_{DD} \) and the fine-grained voltage. Note that the error, at which the load is attached and the maximum voltage drop, is considered here. The error is expressed as \( \text{error} = [(V_{DD} - V_{lc})/(V_{DD} - V_{fi})] \times 100\% \). Table 1 also shows that the relative error increases with the granularity of the model. This is also in line with the results from [24].
The $V_{DD}$ variation of cycle-based simulation is presented in Figure 7. Two nodes are considered here. One node is busy node, which implies the highest switching activity, and the other node is the quietest in the network. A XY routing algorithm with a packet injection rate (PIR) of 0.01 packet per cycle per node are considered. The model demonstrates a capability of capturing cycle voltage variation under different scenarios in a networks-on-chip. In Figure 8, impacts of different routing algorithms on average and minimum supply voltages are presented. The average and minimum voltages for 10,000 cycle simulations of networks-on-chip are shown. The worst case and average values of $V_{DD}$ are plotted for each routing algorithm for a range of packet injection rates.

In Figure 9, the effect of traffic pattern is shown. From these results it can be seen, as is expected, that both average and worst case (minimum) supply voltage decreases, as the packet injection rate increases. Also, it can be noticed that the impact of traffic pattern on supply voltage is slightly higher than that of routing algorithm and the impact of both parameters (routing algorithms and traffic pattern) becomes more distinguishable at higher packet injection rates.
other algorithms in terms of the amount and the number of nodes affected by the voltage drop. Similarly, it can be observed that XY routing is the worst in such scenario.

V. CONCLUSIONS

In this work, a comprehensive model to capture the effect of on-chip communication workloads on power grid is proposed. The proposed model uses decoupling technique to determine the bus switching load during NoC simulation and determines the effect of this load on power grid for each simulation cycle. The coarse-grained models are compared with the fine-grained model and relative error is evaluated which increases with the granularity of the model. Results show that this model has the capability to demonstrate the effect of on-chip communication parameters on power supply noise. In the future, the model can be employed to combine with a current macro models of the IP cores and to identify potential hot-spot of power supply noise in the system. This model provides a unique and communication-centric perspective to analyse power supply integrity that leads to future robust and reliable multi-core system design.

REFERENCES


