Modeling and Analysis of Simultaneous Switching Noise for Full Wafer Scale Chip Core

Hyunwoo Kim, Seonguk Choi, Joonsang Park, Haeyeon Kim, Keeyoung Son, Junghyun Lee, Jiwon Yoon, Jonghyun Hong, Boogyo Sim, Keunwoo Kim, Taein Shin, and Joungho Kim
School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, Republic of Korea
E-mail: kimhyunwoo@kaist.ac.kr

***Abstract*—In this paper, we model and analyze the simultaneous switching noise (SSN) for the full wafer scale chip (FWSC) core. The FWSC has emerged as a potential solution in the artificial intelligence (AI) accelerator market with its high performance and power efficiency. However, the enormous switching operations of FWSC result in a huge simultaneous switching current (SSC), which leads to a high SSN and degrades the power integrity (PI) in the FWSC. Therefore, the SSN should be accurately evaluated to guarantee the PI in the FWSC. We model and analyze the SSN within the hierarchical FWSC core PDN comprising voltage regulator modules (VRMs), PCB PDN, multiarray power/ground (P/G) silicone rubber socket-based PDN, and chip PDN. The hierarchical FWSC core PDN is modeled into equivalent circuit models, and the SSC spectrum of the FWSC core is extracted by a chip power model (CPM), respectively. Then, we fully analyze SSN in the time domain using both the modeled PDN impedance and SSC spectrum. As a result, the high SSN corresponding to 66 % of VDD is induced by overlapping of the impedance peak of PDN and the current peak of SSC spectrum.**

***Index Terms*—Full wafer scale chip, power distribution network, power integrity, simultaneous switching noise**

I. Introduction

The exponential growth of generative artificial intelligence (AI) models such as ChatGPT has dramatically increased the complexity of AI models and the amount of data required for training these models. This has led to significant demand for high bandwidth and low latency systems. To deal with the challenges, the AI accelerator based on the graphics processing unit (GPU) with high bandwidth memory (HBM) has been widely adopted. These approaches, however, have limitations such as insufficient on-chip memory capacity as well as enormous energy consumption [1].

Full wafer scale chip (FWSC) is a promising solution to overcome these limitations of conventional AI accelerators. FWSC is a novel AI accelerator with numerous AI processing elements (PEs) consisting of a router and a core with compute and memory unit, on a single wafer as shown in Fig. 1(a). FWSC has larger on-chip memory capacity and higher bandwidth than conventional AI accelerators. This is because additional cores and expanded memory capacity can be integrated due to the huge size of FWSC. Also, FWSC has high power efficiency as dies and PEs within FWSC are connected by on-chip interconnections instead of off-chip interconnections [2].

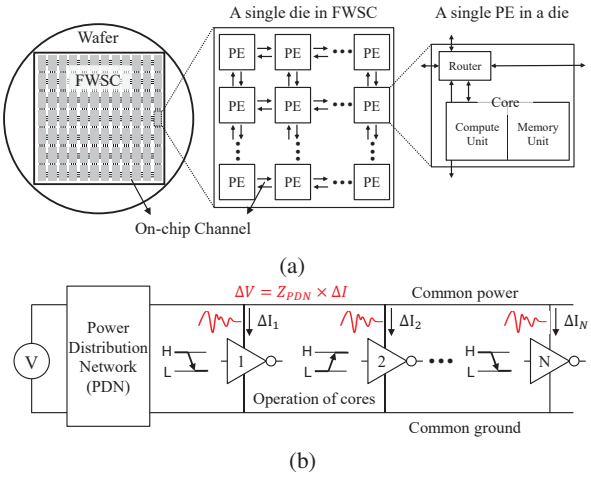


Fig. 1. (a) Conceptual view of the FWSC with numerous PEs composed of a router and a core with compute unit and memory unit (b) SSN generation caused by switching operation of the FWSC cores

Despite significant performance improvements of FWSC, the tremendous number of cores operations in FWSC causes a huge simultaneous switching current (SSC). This leads to a high simultaneous switching noise (SSN) as shown in Fig. 1(b), which degrades power integrity (PI) in the FWSC. Several studies have been conducted for the FWSC, however, most of the studies focus on power distribution network (PDN) design and analysis [3]. Therefore, SSN modeling and analysis for FWSC core are essential for ensuring the PI of FWSC.

In this paper, we model and analyze SSN for the FWSC core. To model SSN, the hierarchical FWSC core PDN is designed and modeled into equivalent circuit models. Also, the SSC spectrum of FWSC core extracted by the chip power model (CPM) is required. Finally, SSN is modeled by multiplying both the modeled PDN impedance and SSC spectrum in the frequency domain and analyzed in the time domain.

II. PDN and SSC Spectrum Modeling of FWSC Core

In this section, we design and model FWSC structure and the hierarchical FWSC core PDN. Each PDN component is modeled into equivalent circuit models, and the overall hierarchical PDN is modeled by cascading all the modeled PDN components. Then, the SSC spectrum is extracted by the CPM using a piece-wise linear (PWL) current source.

979-8-3503-1798-5/23/$31.00 ©2023 IEEE

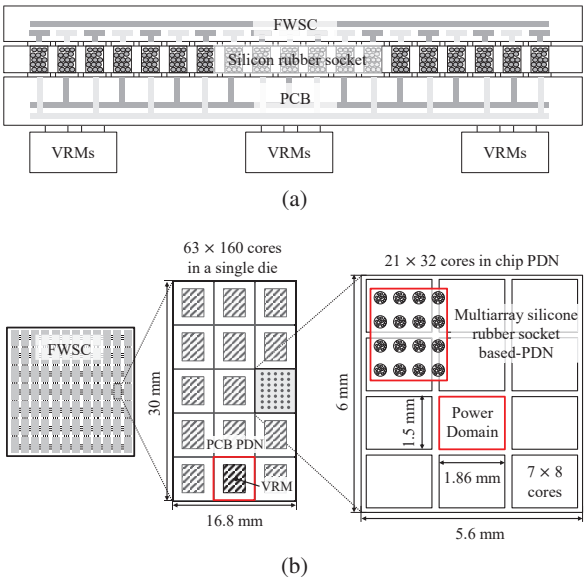


Fig. 2. (a) Side view of the FWSC structure consisting of VRMs, PCB, silicone rubber socket, and the FWSC (b) Top view of the hierarchical FWSC core PDN including VRMs, PCB PDN, multiarray P/G silicone rubber socket-based PDN, and chip PDN

A. Design of FWSC Structure and Core PDN

As shown in Fig 2(a), the FWSC structure is composed of VRMs, PCB, silicone rubber socket, and the FWSC. In order to supply reliable power across the FWSC, it is designed with a 3D structure that can be directly interconnected from VRMs to the FWSC to effectively minimize the current path [3].

The hierarchical FWSC core PDN, as illustrated in Fig. 2(b), incorporates several PDN components including VRMs, PCB PDN, multiarray power/ground (P/G) silicone rubber socket-based PDN and chip PDN. Each die of the FWSC is designed to have the dimensions of 16.8 mm $\times$ 30 mm and integrates 63 $\times$ 160 cores, resulting in a total of 10,080 cores per die. Each die of the FWSC has an independent PDN separated by 15 PCB PDNs depending on the number of VRMs. Accordingly, the dimensions of PCB PDN are set to 5.6 mm $\times$ 6 mm with 21 $\times$ 32 cores. Then, to address the SSN generated within the FWSC core PDN, the chip PDN is structured into 12 power domains for each PCB PDN. Each of these domains is designed to accommodate 56 (7 $\times$ 8) cores within the same power domain. Accordingly, the dimensions of each power domain are set to 1.86 mm x 1.5 mm.

B. Modeling of FWSC Core PDN

To estimate the overall hierarchical PDN impedance profile, we need to model each PDN component described in Section II-A. Table I summarizes the physical dimensions of the FWSC core PDN components. Given the physical dimensions as Table I, each PDN component is modeled into equivalent circuit models using the transmission line modeling (TLM) method, and the overall hierarchical PDN is modeled by cascading all the modeled PDN components with the segmentation modeling method [3-5]. Herein, the VRM is modeled as a two-element RL model (0.1 mΩ, 20 nH, respectively).

TABLE I
PHYSICAL DIMENSIONS OF FWSC CORE PDN COMPONENTS

	Paramter	Description	Value
Chip P/G Meshed Plane	L_{chip}	Length of chip unit cell	30 [μm]
	W_{chip}	Width of chip unit cell	13 [μm]
	S_{chip}	Space of chip unit cell	17 [μm]
	t_{chip}	Metal (Copper) Thickness in chip	0.5 [μm]
	H_{Si}	Height of silicon substrate	30 [μm]
	H_{SiO_2}	Height of silicon dioxide	0.5 [μm]
Multiarray P/G Silicone Rubber Socket	D_{socket}	Diameter of socket	240 [μm]
	H_{socket}	Height of socket	400 [μm]
	P_{socket}	Pitch of socket	400 [μm]
PCB P/G Solid Plane and P/G Via	L_{PCB}	Length of PCB unit cell	800 [μm]
	S_{PCB}	Space of PCB unit cell	800 [μm]
	t_{PCB}	Metal (Copper) Thickness in PCB	17 [μm]
	D_{via}	Diameter of via	200 [μm]
	D_{pad}	Diameter of pad	300 [μm]
	$D_{clearance}$	Diameter of clearance	350 [μm]
	P_{via}	Pitch of via	400 [μm]
	S_{via}	Space between the pad to P/G plane	100 [μm]

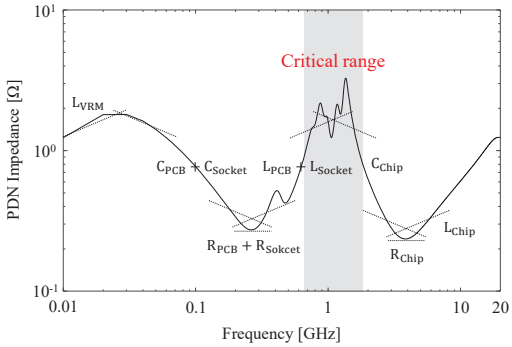


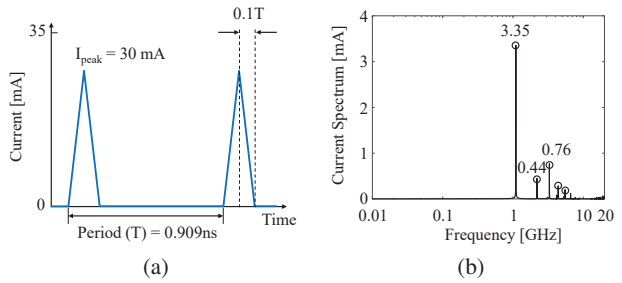
Fig. 3. Modeled hierarchical PDN impedance profile probed at the center of chip PDN

We analyze the hierarchical FWSC core PDN composed of a PCB PDN with 56 unit cells, 112 P/G vias, 112 multiarray P/G silicone rubber sockets, and a chip PDN with 3100 unit cells per one power domain. Fig. 3 shows the modeled hierarchical PDN impedance profile probed at the center of unit chip PDN. The inductance and capacitance of each PDN component dominate the hierarchical PDN impedance at the specific frequency range. Especially, there is a critical frequency range with a high impedance peak around 1.36 GHz, which causes severe P/G noise, such as SSN.

C. Modeling of FWSC Core SSC Spectrum

The SSC spectrum as well as PDN impedance are necessarily required to model the SSN. The CPM is a conventional method to model the SSC spectrum in the form of PWL current source to consider the chip operation [6].

To model the total SSC spectrum of the FWSC cores using CPM, we assume that a single core consumes 30 mW peak power with 1V VDD, and its clock frequency (f_{clock}) is set to 1.1 GHz [2]. Using the simplified CPM methodology of [6], the current profile of a single FWSC core is depicted as Fig. 4(a). The period of the current profile is 0.909 ns (T = $1/f_{clock}$), and the rise time and fall time are set to a tenth of one clock period. Also, The duration of the current profile is set to 1000ns, and the values of I_{peak}, I_B are 30 mA, and 0 mA,



(a) (b)

Fig. 4. (a) Current profile of a single FWSC core using the simplified CPM methodology (b) SSC spectrum of a single FWSC core extracted by FFT

respectively. The SSC spectrum is extracted by fast Fourier transformation (FFT) as shown in Fig. 4(b). The current peaks in the SSC spectrum indicate 2.6, 1.6, and 0.57 mA at 1.1 GHz, 2.2 GHz, and 3.3 GHz, respectively, which is the fundamental frequency (f_{clock}) and its harmonics.

The next step to model the total SSC spectrum of the FWSC core is to determine the switching scenario of cores. Based on Section II-A, since there are 56 cores for one power domain, we consider the worst-case switching scenario of all 56 cores operating simultaneously. Hence, the total SSC spectrum of the FWSC core is extracted by superposing 56 SSC spectrum of a single FWSC core.

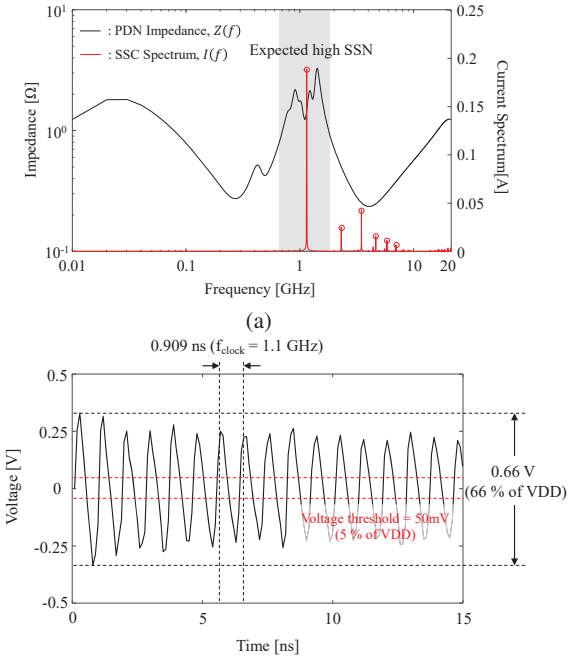
III. Modeling and Analysis of SSN for FWSC Core

In this section, we model and analyze the SSN for the FWSC core based on the PDN impedance $Z(f)$ and the SSC spectrum $I(f)$. The SSN Spectrum $V(f)$ is derived by the multiplication of $Z(f)$ and $I(f)$: $V(f) = Z(f) \times I(f)$. Then, the SSN in the time domain $V(t)$ is derived by inverse FFT.

Fig. 5(a) shows the hierarchical PDN impedance profile of the FWSC core PDN $Z(f)$ and the total SSC spectrum of the FWSC core $I(f)$ modeled in Section II. The peak point in the SSN spectrum $V(f)$ is expected to occur around 1.1 GHz because the critical frequency range of $Z(f)$ and the peak point of $I(f)$ are located close to each other. To confirm the effect of $Z(f)$ and $I(f)$, we analyze the SSN $V(t)$ in the time domain as shown in Fig. 5(b). As a result, the frequency and period of $V(t)$ are 1.1 GHz and 0.909 ns, respectively. This means that the frequency component of 1.1 GHz most dominantly affects the SSN as predicted in Fig. 5(a). Also, the peak-to-peak voltage of $V(t)$ is about 0.66 V, which is 66 % of VDD. It is higher than the required voltage threshold, which is typically set to less than 5 % of VDD for the system to operate properly. This high voltage fluctuation should be appropriately suppressed because it can cause system reliability issues, such as signal distortion and logic failure. Thus, the SSN suppression methodology based on the decoupling capacitor (decap) is required to ensure reliable PI in the FWSC.

IV. Conclusion

In this paper, we modeled and analyzed the SSN within the hierarchical FWSC core PDN including VRMs, PCB PDN, multiarray P/G silicone rubber socket-based PDN, and chip PDN. Based on the TLM method and simplified CPM, the hierarchical FWSC core PDN and SSC spectrum of the FWSC



(a) (b)

Fig. 5. (a) Hierarchical PDN impedance profile of the FWSC core PDN and total SSC spectrum of the FWSC core modeled in Section II (b) Modeled SSN for the FWSC core

core are modeled, respectively. Then, the SSN is modeled and analyzed in terms of PDN impedance and SSC spectrum. Consequently, the high SSN is generated by the peak of PDN impedance and SSC spectrum, which may lead to PI issues for the system. In accordance, a proper decap strategy is required for suppressing the high SSN and ensuring reliable PI of the FWSC.

Acknowledgment

We would like to acknowledge the technical support from ANSYS Korea. This research was supported by National RD Program through the National Research Foundation of Korea (NRF) funded by Ministry of Science and ICT (NRF-2022M3I7A4072293). This work was supported by Samsung Electronics Co., Ltd (IO201207-07813-01)

References

[1] Chen, Yu-Hsin, et al. "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks." IEEE journal of solid-state circuits 52.1 (2016): 127-138.

[2] Lie, Sean. "Cerebras architecture deep dive: First look inside the hardware/software co-design for deep learning." IEEE Micro 43.3 (2023): 18-30.

[3] Kim, Hyunwoo, et al. "Design and Analysis of Hierarchical Power Distribution Network (PDN) for Full Wafer Scale Chip (FWSC) Module." 2022 IEEE Electrical Design of Advanced Packaging and Systems (EDAPS). IEEE, 2022.

[4] Kim, Jaemin, et al. "Modeling and measurement of interlevel electromagnetic coupling and fringing effect in a hierarchical power distribution network using segmentation method with resonant cavity model." IEEE transactions on advanced packaging 31.3 (2008): 544-557.

[5] He, Jiayi, et al. "Extracting characteristic impedance of a transmission line referenced to a meshed ground plane." 2016 IEEE International Symposium on Electromagnetic Compatibility (EMC). IEEE, 2016.

[6] Ko, Baekseok, et al. "Simplified chip power modeling methodology without netlist information in early stage of soc design process." IEEE Transactions on Components, Packaging and Manufacturing Technology 6.10 (2016): 1513-1521.