2023 IEEE Symposium on Electromagnetic Compatibility & Signal/Power Integrity (EMC+SIPI) | 979-8-3503-0976-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/EMCSIPI50001.2023.10241680

Power Supply Induced Jitter (PSIJ) Modeling, Analysis, and Optimization of High Bandwidth Memory (HBM) I/O Interface

Hyunwook Park[1)], Taein Shin[1)], Seongguk Kim[1)], Keeyoung Son[1)], Keunwoo Kim[1)], Boogyo Sim[1)], Hyungmin Kang[1)], Seonguk Choi[1)], Jiwon Yoon[1)], Hyunwoo Kim[1)], Chulsoon Hwang[2)], and Joungho Kim[1)]
[1)]*School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST),* Daejeon, South Korea
[2)]*Department of Electrical and Computer Engineering, Missouri University of Science and Technology (MST),* Rolla, MO, USA
E-mail: hyunwookpark@kaist.ac.kr

***Abstract*—Power supply induced jitter (PSIJ) in high bandwidth memory (HBM) I/O interface is modeled, analyzed, and optimized for different HBM generations. Precise models for VDDQ power distribution networks (PDNs), simultaneous switching current (SSC), and jitter sensitivities of the clock and I/O buffers are implemented for PSIJ estimation. Compared to the SPICE, the average error rate of the estimated PSIJ is 4.26 %. The critical frequency bands in the jitter spectrum where large jitters occur are derived by comparing the relative impact of the modeled interface factors in the frequency domain. For the optimization, on-chip and on-interposer decoupling capacitor (decap) placement strategies using machine learning (ML) are applied. The decap effects in the critical ranges are analyzed. Finally, based on the integrated analysis of the limitation of the decap solution and all the I/O interface factors, the major challenges of high-frequency PSIJ are characterized.**

Keywords— I/O interface, High bandwidth memory, Power Supply Induced Jitter

I. Introduction

Recently, HBM-graphic processing unit (GPU) module has become an indispensable solution for high-performance computing systems [1]. The HBM-GPU module provides TB/s scale bandwidth in the I/O interface where 1024 I/Os are integrated [2]. However, huge SSC drawn by the I/O buffers causes simultaneous switching noise (SSN) as shown in Fig. 1. As a result, PSIJs generated both in clock buffers and I/O drivers are accumulated so that severely degrade the eye opening. Moreover, due to the increasing data rate as generations, switching power has increased while the timing margin becomes tighter [3]. Hence, PSIJ needs to be precisely modeled, analyzed, and optimized to ensure signal integrity (SI) and power integrity (PI) in the HBM I/O interface.

The PSIJ in the HBM I/O interface has been actively studied [4]–[7]. The PSIJ analysis based on the combined SI-PI co-simulation is investigated [4], [5]. In [4], the impact of on-chip decap placement and their corresponding parasitic inductance on PSIJ is studied with distributed PDN models. However, only jitters that occurred in the I/O driver are simulated, excluding the clock buffers. In addition, only 3 I/O drivers are considered for simplicity which is not practical. To consider the entire switching I/Os in the physical layer (PHY), a super driver is introduced to emulate worst-case SSC and SSN. However, both Paulis *et al.* [4] and Shi *et al.* [5] focus on analyzing PSIJ only in terms of the PDN impedance. Detailed analysis of the impact on PSIJ of all components that configure the I/O interface lacks. The system-level PSIJ is modeled and the analysis of interface factors including PDNs,

This research was supported by National R&D Program through the National Research Foundation of Korea(NRF) funded by Ministry of Science and ICT (NRF-2022M3I7A4072293).

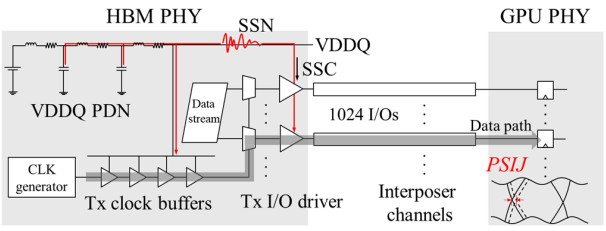

Fig. 1. HBM I/O Interface including VDDQ PDN, Tx clock buffers, and I/O drivers. PSIJ by 1024 switching I/Os degrades eye openings.

channels, clock buffers, and I/O drivers is investigated [6]. Depending on the variation of design parameters of the interface factors, the relative effects on the PSIJ are compared. However, only 32 I/Os of 1 channel (CH) in the DWORD of the HBM PHY are considered. Moreover, since a simplified lumped model is used for the PDN, it cannot consider the position of the victim I/O and the decap placement. In other words, only the PSIJ that occurred in a portion of the HBM PHY was analyzed, not the entire region.

In this paper, as the extension of our previous work [7], we model, analyze, and optimize the PSIJ in HBM I/O interfaces. Compared to [7], detailed modeling, verification, and analysis of the PSIJ are conducted for different HBM generations with varying data rates. VDDQ PDNs, SSC, and jitter sensitivities of the clock and I/O buffers are modeled and integrated into the jitter spectrum. The modeled PSIJ is verified to be precise, having 4.26 % error rate compared to the SPICE simulation. The relative effects of the modeled interface factors are compared to characterize the critical frequency bands in the jitter spectrum. On-chip and on-interposer decap solutions are applied using ML, and their effects in the critical bands are analyzed. Based on the integrated analysis of the limitation of the decap solution and all the interface factors, the major challenges of high-frequency PSIJ are identified.

II. PSIJ Modeling of HBM I/O Interface

A. Target HBM I/O interface

Target HBM I/O interface for this work includes Tx clock, Tx I/O buffers, HBM VDDQ PDN, and interposer channels as depicted in Fig. 1. Fig. 2 shows the top view of the HBM-GPU module including HBM logic die with ballout and decap areas, and VDDQ PDN. A total of 1024 I/O buffers are in the PHY since 8 DWORDS have 128 I/Os respectively. Each DWORD consists of 4 CHs with 32 I/Os. To deal with PSIJ generated in the entire VDDQ domain, the PHY is divided into 4 regions by grouping 2 DWORDS into 1 I/O region. Hence, each region denoted in red box has 256 I/Os and 1 observation port. The 4 ports are located in DWORD0 and DWORD2 for CHs 0, 1, 4, and 5 and CHs 2, 3, 6, and 7 in this work.

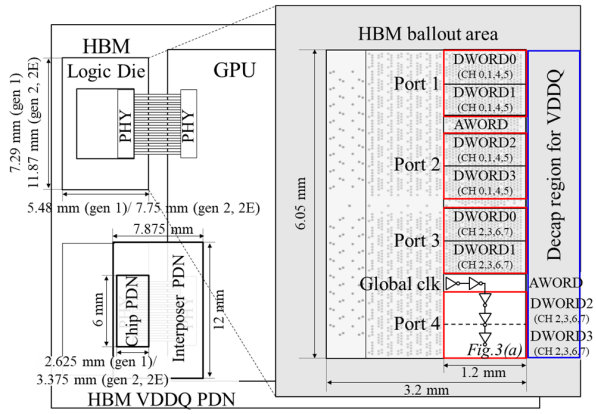

Fig. 2. Top view of the target HBM-GPU module including HBM logic die with HBM ballout and decap areas, and HBM VDDQ PDN.

A global clock network for the VDDQ domain is implemented on the PHY as shown in Fig. 2. Clock signals are assumed to be distributed from AWORD (Address/command buffers) clock generator to Tx drivers in the DWORDS [6], [7]. The distance between 2 clock buffer stages is designed to be 0.41 mm. Therefore, depending on the position of the DWORDS, the number of clock stages varies from 3 to 5. The drivers in DWORDs 1 and 2 are connected to the 3-stage clock buffers, and those in DWORDs 0 and 3 are to the 5-stage.

Hierarchical VDDQ PDN including on-chip, on-interposer, and on-package (PKG) PDNs is considered. As shown in Fig. 2, the size of on-chip PDNs is set as 2.625×6 mm^2 for HBM1, and 3.375×6 mm^2 for HBM2 and 2E. The size of on-interposer PDN is 7.875×12 mm^2 for all the gen 1, 2, and 2E [8]. The on-interposer PDN is designed to short both the HBM and GPU VDDQ domains [5]. The decap region available for on-chip PDN is denoted as a blue box. On-interposer decaps are assumed to be available on the entire on-interposer PDN.

B. Modeling of PSIJ

The total PSIJ spectrum at an observation port $i \in \{1:4\}$ of each I/O region is modeled. The PSIJ spectrum $J_i(f)$ at port i is the sum of jitters contributed from the clock buffers $J_{\text{clk},i}(f)$ and the I/O drivers $J_{\text{driver},i}(f)$:

$$\begin{aligned} J_i(f) &= J_{\text{clk},i}(f)+J_{\text{driver},i}(f) \\ &= S_{\text{clk},i}(f)\times V_i(f)+S_{\text{driver}}(f)\times V_i(f) \\ &= S_{\text{clk},i}(f)\times Z_{\text{sum},i}(f)\times I(f)+S_{\text{driver},i}(f)\times Z_{\text{sum},i}(f)\times I(f). \end{aligned} \quad (1)$$

Each of them is expressed as the multiplication of their jitter sensitivities $S(f)$ and SSN spectrum $V_i(f)$ at port i. The same SSC $I(f)$ is assumed to be drawn by each I/O region. Then, $V_i(f)$ can be calculated by the multiplication of $I(f)$ and the total HBM VDDQ PDN impedance $Z_{\text{sum},i}(f)$. $Z_{\text{sum},i}(f)$ is sum of all the self- and transfer impedances seen at port i: $Z_{\text{sum},i}(f)=Z_{ii}(f)+\sum_{j\neq i} Z_{ij}(f)$ where $j \in \{1:4\}$. This is because both the self-switching noise and transferred noises need to be taken into account for precise modeling [8]. Finally, PSIJ in time domain $J_i(t)$ is derived by inverse Fourier transformation ($IFFT$): $J_i(t)= IFFT(J_i(f))$. Then, the peak-to-peak PSIJ $J_{\text{pk-pk},i}$ can be represented as $max(J_i(t))$-$min(J_i(t))$.

Detailed circuit implementation of 1-stage Tx clock buffer and I/O driver is depicted in Fig. 3(a). 4 tapered inverter

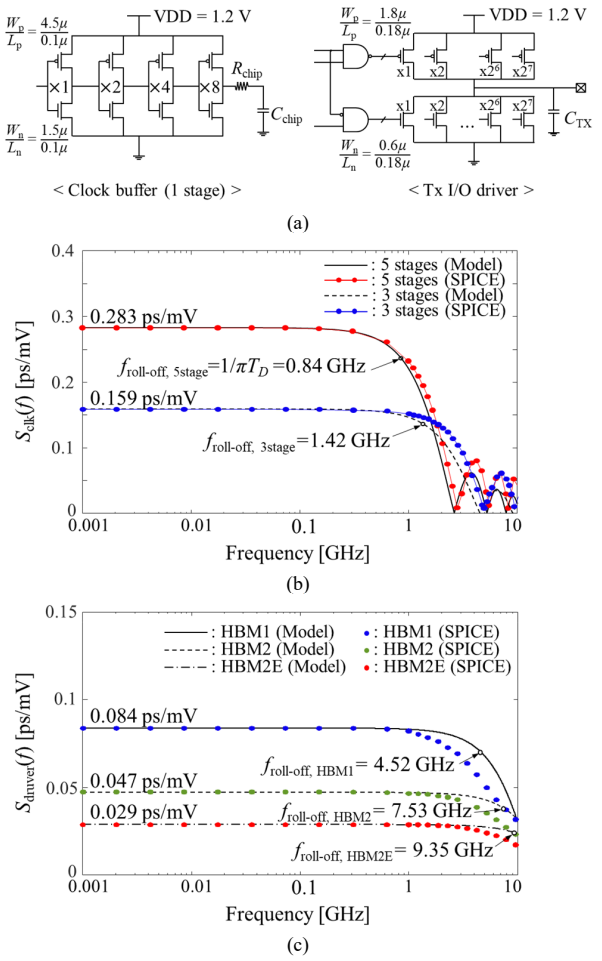

Fig. 3. (a) Schematic of clock buffer and I/O driver. (b) Modeled jitter sensitivity of 5 and 3 stage clock buffers. (c) Modeled jitter sensitivity of I/O driver depending on the HBM generations.

chains are designed for 1 stage based on TSMC 65 nm process [6]. On-chip RDL resistance and capacitance are 28.88 Ω and 0.155 pF respectively [1]. The I/O driver operates at 1 Gbps, 2 Gbps, and 3.2 Gbps for HBM 1, 2, and 2E respectively. The I/O drivers are also designed with TSMC 65 nm process based on the HBM JEDEC Standard [2]. The drivers are designed to have 11~12 % UI of rise/fall time with C_{TX}=0.4 pF – 110 ps for HBM1, 60 ps for HBM2, and 38 ps for HBM2E.

Both $S_{\text{clk}}(f)$ and $S_{\text{driver}}(f)$ are modeled using propagation delay-based methods [9]:

$$\begin{aligned} S(f) &= S_0\times sinc(\pi f T_{\text{D}}) = \frac{j}{2\pi f T_{\text{D}}} S_0(1-e^{-j2\pi f T_{\text{D}}}) \\ S_0 &= \frac{T_{\text{p,max}}-T_{\text{p,min}}}{VDD_{\text{max}}-VDD_{\text{min}}},\ T_{\text{D}} = \frac{T_{\text{p,max}}+T_{\text{p,min}}}{2}. \end{aligned} \quad (2)$$

where VDD_{max} and VDD_{min} are the max and min DC supply voltage which are set as 1.3 V and 1.1 V respectively. $T_{\text{p,max}}$ and $T_{\text{p,min}}$ are open loop delays with VDD_{max} and VDD_{min} respectively. $T_{\text{p,max}}/T_{\text{p,min}}$ of clock buffers are 405.4/ 348.9 ps for 5 stages and 239.8/208.1 ps for 3 stages. Those of I/O drivers are 78.74/ 62.04 ps, 51.97/ 42.58 ps, and 36.9/ 31.2 ps for HBM1, 2, and 2E respectively.

The modeled S_{clk} is depicted in Fig. 3(b). Both the 3-stage and 5-stage models are well correlated with SPICE simulation

TABLE I. PSIJ COMPARISION RESULTS WITH SPICE SIMULATION

Case	5-stage clk		3-stage clk		IO	5-stage clk+IO	3-stage clk+IO
	w/o aggr	*w/ aggr*	*w/o aggr*	*w/ aggr*	*w/ aggr*	*w/ aggr*	*w/ aggr*
Model [ps]	4.9	13.3	2.3	9.4	7	16.2	13.9
SPICE [ps]	5.2	14	2.4	9.8	7.4	16.2	13.2
Error [%]	5.8	5	4.2	4.1	5.4	0	5.3

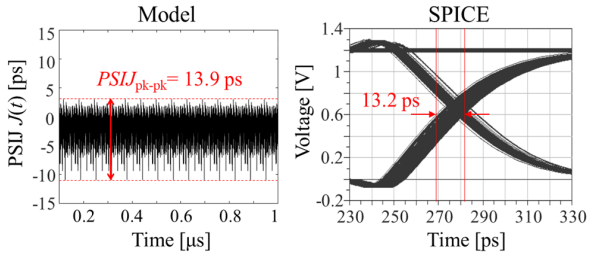

Fig. 4. PSIJ comparison results of 3-stage clock buffers and I/O driver with 128 I/O aggressors.

results. Constant jitter sensitivity profiles are shown at relatively low frequencies under the roll-off frequency $f_{\text{roll-off}}$. This is because same-phase jitters are accumulated when the noise wavelength is longer compared to the propagation delay of the clock buffers. However, the sensitivity rapidly decreases over $f_{\text{roll-off}}$ because the phase variation is compensated rather than accumulated. Therefore, the 5-stage with the longer delay has higher jitter sensitivity at low frequencies, but it rolls off earlier than the 3-stage – $S_{\text{clk,5stage}}(f=0)=0.283$ ps/mV> $S_{\text{clk,3stage}}(f=0)=0.159$ ps/mV, and $f_{\text{roll-off,5stage}}=0.84$ GHz <$f_{\text{roll-off,3stage}}=1.42$ GHz. Fig. 3(c) shows the modeled S_{driver} of HBM1, 2, and 2E I/O drivers. Since the rise/fall time is designed to be shorter in the order of gen 2E, 2, and 1, the driver strength is larger to realize a higher slew rate. Therefore, propagation delay becomes shortest in gen 2E, which leads to the smallest jitter sensitivity at low frequencies and the highest $f_{\text{roll-off}}$. $S_{\text{driver}}(f=0)$ is 0.084 ps/mV, 0.047 ps/mV, and 0.029 ps/mV for gen 1, 2, and 2E respectively. $f_{\text{roll-off}}$ is 4.52 GHz, 7.52 GHz, and 9.35 GHz for gen 1, 2, and 2E respectively – higher bandwidth for the higher generation.

$I(f)$ is extracted by the SPICE simulator using the designed I/O drivers. Since 256 I/Os are in 1 I/O region, 128 switching I/Os considering data bus inversion (DBI) are simulated as the worst-case scenario. The total loading capacitor for a single driver is 1.69 pF: $C_{\text{load}}=C_{\text{Tx}}+C_{\text{channel}}+C_{\text{Rx}}=0.4$ pF+0.89 pF+0.4 pF [1]. The hierarchical PDN model $Z(f)$ includes on-chip grid P/G planes, P/G μbump array, on-interposer meshed P/G planes, multiarray P/G TSVs, and PKG PDN [8]. All the components of on-chip and on-interposer PDNs are implemented as precise distributed models. Therefore, Z contains impedance profiles seen at 4 observation ports and decap ports for the decap assignment. Details of the PDN modeling are explained in our previous work [8, Section III].

C. Verification of the modeled PSIJ

The modeled PSIJ is verified by comparing $J_{\text{pk-pk}}$ with the SPICE simulation. Table I shows the comparison results of 7 cases in HBM gen 2 with and without 128 I/O aggressors. It is verified in only 1 I/O region due to the computational limitation in the SPICE simulator. In all the cases, the estimated PSIJ results by the model are well-correlated. The

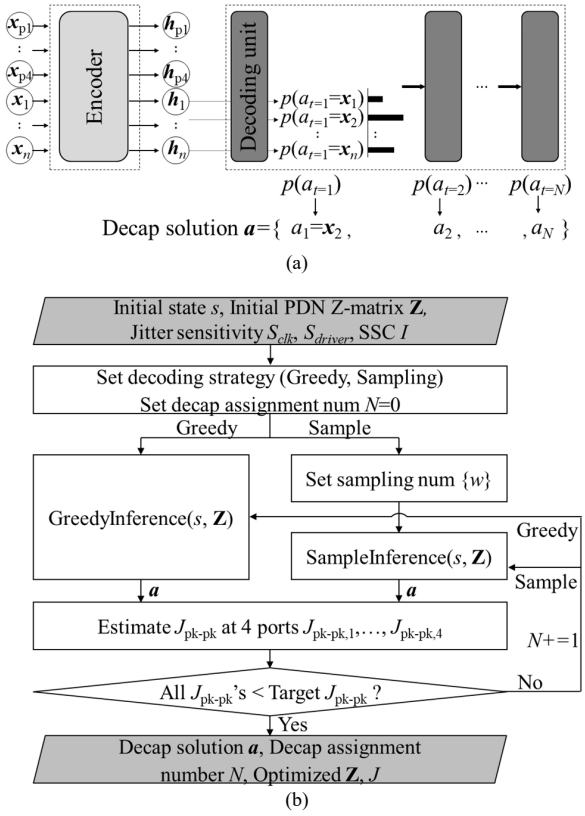

Fig. 5. (a) Transformer network for on-chip and on-interposer decap placement [8]. (b) Overall flow chart of the PSIJ optimization [7].

average error rate is 4.26 %. One case of 3-stage clock buffers and I/O driver with 128 aggressors is plotted in Fig. 4.

III. PSIJ ANALYSIS AND OPTIMIZATION OF HBM I/O INTERFACE

A. Decap strategy for PSIJ optimization using ML

Since I/O switching noises occur majorly in the frequency range over hundreds of MHz, one of the most important processes to reduce PSIJ is the on-chip and on-interposer decap strategies. Therefore, the PSIJ optimization in this work focuses on the placement of the on-chip and on-interposer NMOS unit decaps. The on-chip and on-interposer VDDQ PDNs in Fig. 2 are represented in a 3-D grid world as same as [8, Fig. 3(a)]. The size of 1 grid is 375×375 μm^2 which is the same as that of the unit decap. The 4 observation ports and decap ports are represented in vectors $\boldsymbol{x}_{\text{p1:p4}}$ and $\boldsymbol{x}_{1:n}$ respectively. $\boldsymbol{x}$ contains the 3-D coordinates, p1:p4 indicates the 4 observation ports, and n is the number of the decap assignment candidates. The unit decap is modeled to have 1.055 nF capacitance and 0.7 mΩ ESR [8].

The PSIJ optimization method proposed in our previous work is used for this work [7]. In the method, the pre-trained transformer network for the decap placement is introduced [8]. As shown in Fig. 5(a), the network is trained to derive a decap solution $\boldsymbol{a}=\{a_1, a_2, \ldots, a_N\}=\{\boldsymbol{x}_{a1}, \boldsymbol{x}_{a2}, \ldots, \boldsymbol{x}_{aN}\}$ for given input state $s=\{\boldsymbol{x}_{\text{p1:p4}}, \boldsymbol{x}_{1:n}\}$ to maximize reward r. N is the number of decap assignments and $\boldsymbol{x}_{a1:aN}$ is a subset of $\boldsymbol{x}_{1:n}$. r is defined as the reduction of 10 self- and transfer impedances Z_{11}, Z_{22}, Z_{33}, Z_{44}, Z_{12}, Z_{13}, Z_{14}, Z_{23}, Z_{24}, and Z_{34} at $\boldsymbol{x}_{\text{p1:p4}}$ from 10 MHz to 20 GHz [8, eq. (5)]. Details of the reinforcement learning algorithm for training are described in [8, Section II-C].

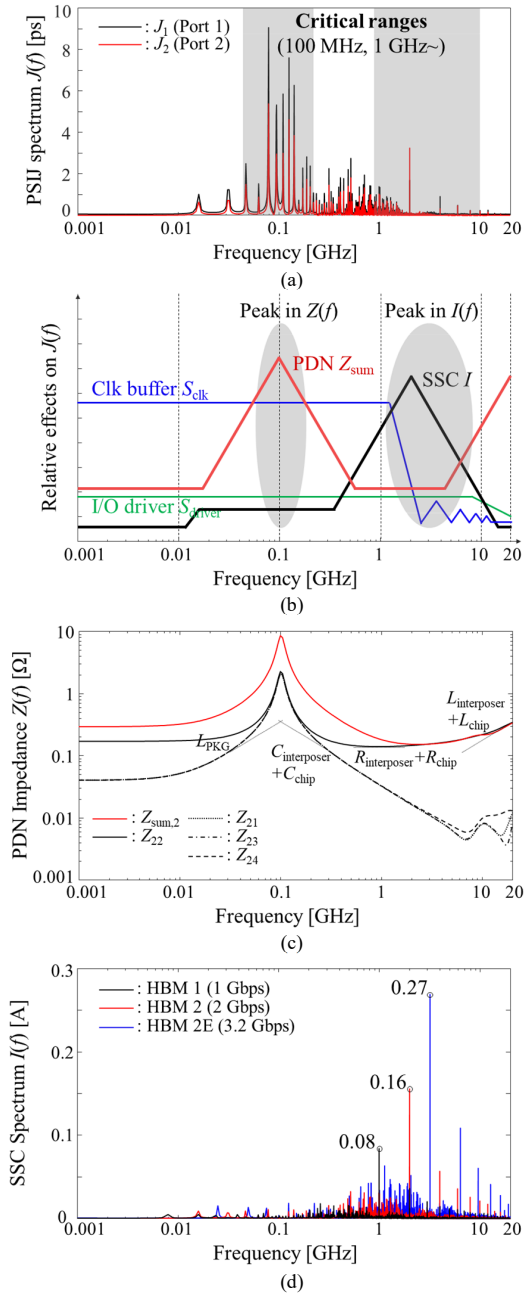

Fig. 6. (a) Initial $J(f)$ at port 1 and 2 of HBM2. (b) Relative effect profiles of Z_{sum}, I, S_{clk}, and S_{driver} on $J(f)$. (c) Initial PDN impedance profiles seen at port 2 – Z_{sum}, Z_{22}, Z_{21}, Z_{23}, and Z_{24}. (d) $I(f)$ depending on the HBM generation.

Fig. 5(b) shows the overall flow chart of the PSIJ optimization method using the transformer network. The initial s, the modeled impedance matrix $\mathbf{Z}$, S_{clk}, S_{driver}, and I are given as inputs. The decoding strategy is set to either greedy or sampling. Starting from N=0, the network outputs a decap solution $\boldsymbol{a}$ by inference. For every end of the loop, $J_{pk\text{-}pk}$'s at 4 ports of the PDN with $\boldsymbol{a}$ are computed by (1) and check if all the $J_{pk\text{-}pk}$'s meet the given target $J_{pk\text{-}pk}$. If not, increase N and repeat the loop until all the $J_{pk\text{-}pk}$'s are satisfied.

The inference consists of computations performed by the encoder and the decoder of the network as shown in Fig. 5(a). The encoder outputs high-dimensional information $\boldsymbol{h}$ from s,

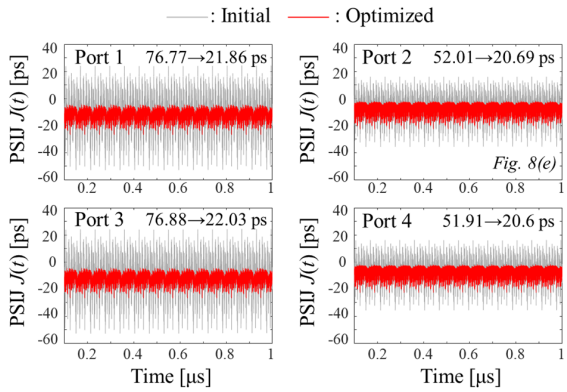

Fig. 7. Initial and optimized $J(t)$ at the 4 observation ports in HBM2. Peak-to-peak 22.05 ps, which is 4.4 % of UI, is given as the PSIJ specification.

which is called embedding [8, eqs. (7)–(12)]. With the $\boldsymbol{h}$ by the encoder, the decoding unit sequentially computes the probability distribution function $p(a_t)$ for selecting the position of the unit decap among $\boldsymbol{x}_{1:n}$, which is represented in the black histogram in Fig. 5(a). $p(a_t)$ is estimated by attention. The attention is a parallel computation for capturing the relationship between nodes [8, eqs. (13)–(16)]. For the greedy selection, a_t is selected as the position where the probability is maximized for every t. However, in the sampling selection, a_t is chosen according to $p(a_t)$. The sampling number w of inferences are performed and the best solution $\boldsymbol{a}$ is selected. The sampling selection with w=100 is used for this work.

B. Analysis of PSIJ in the initial HBM I/O interface

Fig. 6(a) shows initial PSIJ spectrums $J(f)$ at port 1 and 2 of HBM2. There are 2 critical frequency ranges around 100 MHz and over 1 GHz, where large jitters occur. This can be explained by overlapping the relative effects of Z_{sum}, I, S_{clk}, and S_{driver} as shown in Fig. 6(b). The red line indicates Z_{sum}, the black line for I, and the blue and green lines for S_{clk} and S_{driver} respectively. The 2 critical ranges are mainly determined where peaks in Z_{sum} and I occur.

Fig. 6(c) shows initial VDDQ PDN impedances at port 2 including $Z_{sum,2}$, Z_{22}, Z_{21}, Z_{23}, and Z_{24}. High anti-resonance between L_{PKG} and $C_{interposer}$+C_{chip} occurs around 100 MHz in $Z_{sum,2}$, which is the main factor of one of the critical ranges in $J(f)$. Compared to the self-impedance, note that the transfer impedances also critically contribute to the anti-resonance in $Z_{sum,2}$. Therefore, transferred noises also need to be decoupled carefully. $Z_{sum,2}$ increases over 1 GHz where self loop inductance $L_{interposer}$+L_{chip} in Z_{22} dominates the profile.

Fig. 6(d) shows the SSC spectrum $I(f)$ depending on the generation. Since all the data rates are over 1 Gbps, most of the spectrums are concentrated in over 1 GHz ranges. In other words, high peaks are made at fundamental frequencies and their harmonics. The fundamental frequencies are 1 GHz, 2 GHz, and 3.2 GHz for gen 1, 2, and 2E respectively. Not only does the frequency range of the peaks becomes higher, but the magnitude of those also increases. The magnitude of the peak at the fundamental frequency is 0.08 A, 0.16 A, and 0.27 A for gen 1, 2, and 2E respectively. This is because the I/O driver has larger driver strength for realizing the higher slew rate.

As shown in Fig. 6(b), the effect of S_{clk} is dominant in the frequency range under $f_{roll\text{-}off}$ of the clock buffers, which is roughly 1 GHz. The effect under $f_{roll\text{-}off}$ becomes larger according to the number of stages as explained in Fig. 3(b). It is also verified in the comparison between J_1 and J_2 in Fig.

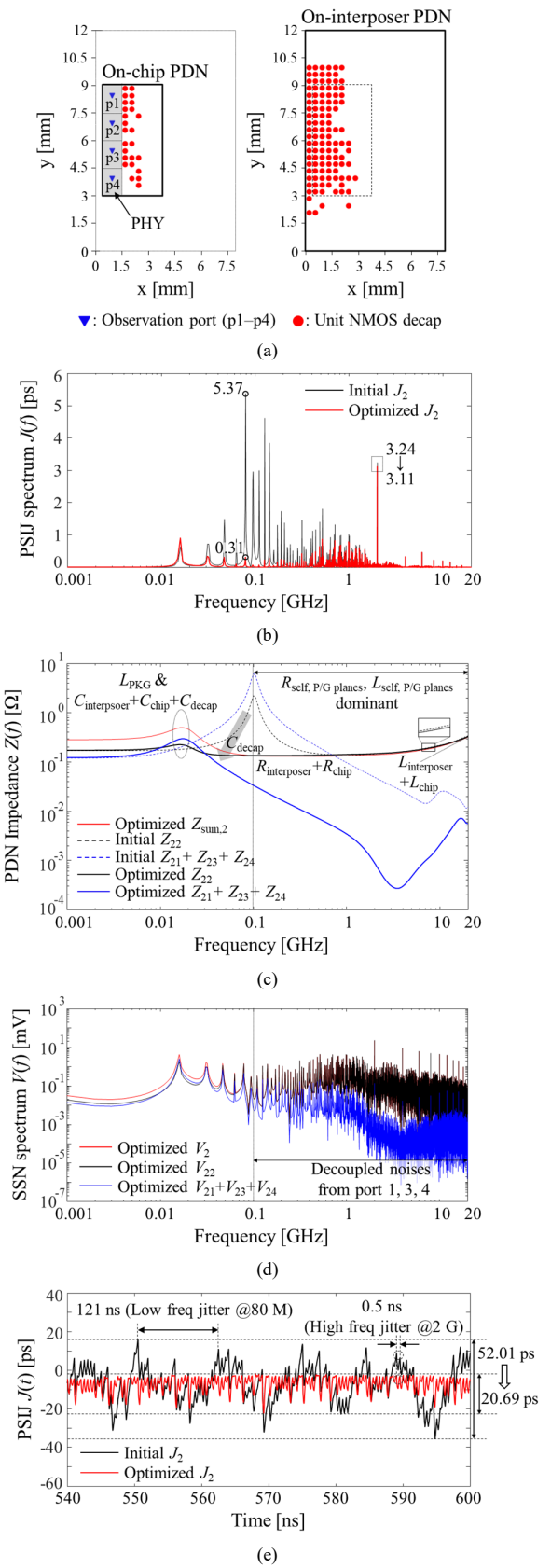

Fig. 8. (a) Optimized on-chip and on-interposer decap assignment to meet target PSIJ in HBM2. (b) Initial and optimized $J(f)$ at port 2. (c) Optimized $Z(f)$. (d) Optimized $V(f)$. (e) Initial and optimized $J(t)$.

6(a). J_1 with 5-stage clock buffers shows larger jitter spectrum under 1 GHz. Compared to the S_{clk}, the effect of S_{driver} is 1.89 to 9.75 times smaller under 1 GHz since the propagation delay is much smaller in the driver as mentioned in Section II-B. However, it becomes similar over 1 GHz after S_{clk} rolls off.

C. Analysis of the optimized PSIJ results

Fig. 7 shows the initial and optimized $J(t)$ at the 4 observation ports in HBM2. 22.05 ps is given as a peak-to-peak PSIJ specification which is about 4.4 % UI. All the optimized $J_{pk\text{-}pk}$'s satisfy the given target specification. The initial $J_{pk\text{-}pk}$'s are 76.77 ps, 52.01 ps, 76.88 ps, and 51.91 ps at ports 1, 2, 3, and 4 respectively. The values of $J_{pk\text{-}pk}$'s seen at ports 1 and 3 are about 24.9 ps larger than those at ports 2 and 4 due to the larger S_{clk} profile around 100 MHz where the anti-resonance occurs in Z. $J_{pk\text{-}pk}$'s are reduced to 21.86 ps, 20.69 ps, 22.03 ps, and 20.6 ps respectively by the assigned decaps. The optimized decap assignment is depicted in Fig. 8(a). A total of 138 EA unit decaps are assigned, which are denoted in red circles. The total capacitance is 145.6 nF and the layout area is 19.4 mm^2. To minimize all the jitters at the 4 ports, the decaps are placed near and between the ports.

Fig. 8(b) depicts the initial and optimized $J(f)$ at port 2. The PSIJ is suppressed by the decaps both in the 2 critical regions. However, the amount of the reduced PSIJ components over 1 GHz is relatively lower than that of around 100 MHz. The PSIJ is reduced from 5.37 ps to 0.31 ps at 80 MHz, but only from 3.24 ps to 3.11 ps at 2 GHz. As shown in the optimized $Z_{sum,2}$ denoted as the red line in Fig. 8(c), the capacitance of the decaps C_{decap} significantly reduces the anti-resonance and shifts the resonant frequency from 100 MHz to 17 MHz. However, the magnitude of $Z_{sum,2}$ is slightly decreased over 1 GHz compared to the initial $Z_{sum,2}$ in Fig. 6(c).

The initial and optimized self-impedance and the sum of transfer impedances at port 2 are also shown in Fig. 8(c). The corresponding optimized SSN spectrum V's are plotted in Fig. 8(d). When comparing the optimized Z_{22} and Z_{21}+Z_{23}+Z_{24}, both contribute to Z_{sum} under 100 MHz. In addition, around 100 MHz noises are easily reduced by the decaps. However, over 100 MHz, the profile of $Z_{sum,2}$ follows that of Z_{22}. Hence, the total spectrum V_2 follows the self-switching noise spectrum V_{22} as shown in Fig. 8(d). This is because most of the noise coupling paths from ports 1, 3, and 4 to port 2 are decoupled by the decaps as shown in the blue lines in Fig. 8(c). In other words, V_{22} is more dominant than the coupled noises V_{21}+V_{23}+V_{24} in the optimized total PSIJ.

As shown in the optimized Z_{22} in Fig. 8(c), $R_{interposer}$+R_{chip} and $L_{interposer}$+L_{chip} determine the impedance profile sequentially over 100 MHz. $R_{interposer}$+R_{chip} and $L_{interposer}$+L_{chip} are the resistance and inductance of the current loops seen at the port. Hence, the positions of the decaps are the most important factor to determine those. This is the reason why the decaps are optimized to be located near the 4 ports both in the on-chip and on-interposer PDNs. However, even though the decaps shorten the current paths, the impedance reduction is limited by intrinsic self-resistance and inductance of on-chip and on-interposer P/G planes themselves.

Fig. 8(e) shows the detailed initial and optimized $J(t)$ at port 2. In the initial $J(t)$, the low-frequency jitter with a period of 121 ns is more dominant than the high-frequency jitter with a period of 2 ns to determine $J_{pk\text{-}pk}$. This is consistent with the comparison of the jitter components at 80 MHz and 2 GHz of the initial $J(f)$ in Fig. 8(b). However, in the optimized $J(t)$, 80

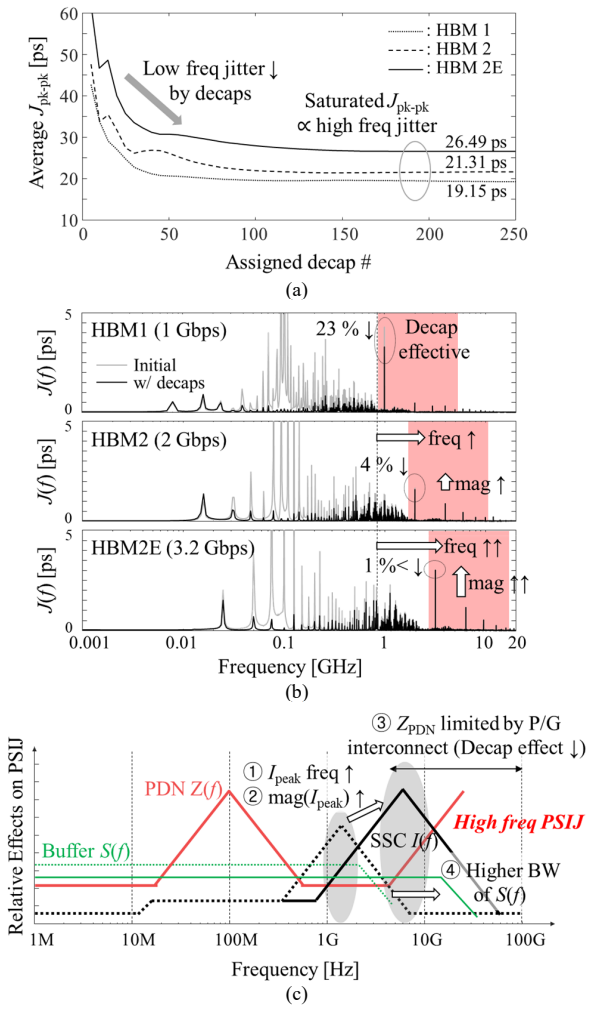

Fig. 9. (a) The average J_{pk-pk} of 4 ports according to the number of decaps in HBM1, 2, and 2E. (b) Initial and optimized $J(f)$ at port 1 depending on the generation. The same 138 EA unit decaps are assigned. (c) Emerging high-frequency PSIJ issues.

MHz jitters are almost suppressed by the decaps. However, 2 GHz jitters are still remained due to the intrinsic self-inductance of on-chip and on-interposer P/G planes.

D. High-frequency PSIJ issues depending on generations

Fig. 9(a) shows the average J_{pk-pk} of the 4 ports according to the assigned number of decaps in HBM1, 2, and 2E. The average J_{pk-pk}'s rapidly decrease in the beginning since the low-frequency jitters in the ~MHz range are easily mitigated by the decaps. However, when exceeding a certain number, the average J_{pk-pk}'s become saturated to 19.15 ps, 21.31 ps, and 26.49 ps respectively. The saturated values are determined by high-frequency jitters in the ~GHz range and increase as generations.

Fig. 9(b) shows the initial and optimized $J(f)$ at port 1 depending on the generation. The PSIJ results with the same 138 EA decaps for a fair comparison are plotted. The larger jitter components are formed at the higher frequency range as a generation gets higher. However, on-chip and on-interposer decap solutions are limited to the range within ~3 GHz. The reduction rates at the fundamental frequency are 23 %, 4 %, and less than 1 % in gen 1, 2, and 2E respectively.

For the next-generation HBM I/O interfaces, to meet the demands on the higher bandwidth, the data rate, driver strength, the number of I/Os, etc are expected to keep increasing [1], [3], [6]. Therefore, as shown in Fig. 9(c), both the frequency band and the magnitude of I_{peak}'s in the SSC spectrum will keep increasing. But the decap effects will be more limited by on-chip and on-interposer interconnects. Also, the bandwidth of the sensitivity of buffers will become higher. As a result, the high-frequency PSIJ will become one of the most challenging issues to ensure SI/PI in the I/O interface.

The PSIJ reduction over the GHz frequency range requires both the silicon and P/G interconnect designs to be further robust. By separating the power domain for Tx I/O drivers and lowering the VDD for those, the high-frequency jitters can be mitigated. This is because it reduces the magnitude of the SSC by lowering the driver strength like lower-power double data rate (LPDDR). Using a thicker top metal for the on-chip P/G plane, making on-chip and on-interposer P/G planes as multi-layers over 2 or more layers, and increasing the metal density of those can be solutions by reducing intrinsic resistance and inductance. Also, further shortening the distance between the Tx I/O drivers and decap layout can help to reduce PSIJ.

IV. Conclusion

As the timing margin becomes tighter in the HBM I/O interface, PSIJ needs to be accurately predicted and analyzed in the pre-design stages. In this paper, PSIJ in the HBM I/O interface is modeled, analyzed, and optimized. Based on the integrated analysis of all the modeled interface factors and the limitation of the decap solution, the major challenges of high-frequency PSIJ in the HBM I/O interface are characterized.

References

[1] S. Kim *et al*., "Processing-in-memory in High Bandwidth Memory (PIM-HBM) Architecture with Energy-efficient and Low Latency Channels for High Bandwidth System," in *Proc. IEEE 28th Conf. Elect. Perform. Electron. Packag. Syst. (EPEPS)*, 2019, pp. 1–3.

[2] *High Bandwidth Memory DRAM (HBM1, HBM2)*, Standard JESD235D, 2021.

[3] K. Son *et al*., "Thermal and Signal Integrity Co-Design and Verification of Embedded Cooling Structure With Thermal Transmission Line for High Bandwidth Memory Module," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 12, no. 9, pp. 1542-1556, Sept. 2022.

[4] F. de Paulis *et al*., "Impact of chip and interposer PDN to eye diagram in high speed channels," in *Proc. IEEE 22nd Workshop Signal Power Integrity (SPI)*, Brest, France, May. 2018, pp. 1–4.

[5] W. Shi, Y. Zhou and S. Sudhakaran, "Power delivery network design and modeling for High Bandwidth Memory (HBM)," in *Proc. IEEE 25th Conf. Elect. Perform. Electron. Packag. Syst. (EPEPS)*, 2016, pp. 3-6.

[6] T. Shin *et al*., "Modeling and Analysis of System-Level Power Supply Noise Induced Jitter (PSIJ) for 4 Gbps High Bandwidth Memory (HBM) I/O Interface," in *Proc. IEEE Elect. Design Adv. Packag. Syst. (EDAPS)*, 2021, pp. 1–3.

[7] H. Park *et al*., "Scalable Transformer Network-based Reinforcement Learning Method for PSIJ Optimization in HBM," in *Proc. IEEE 31st Conf. Elect. Perform. Electron. Packag. Syst. (EPEPS)*, 2022, pp. 1–3.

[8] H. Park *et al.*, "Transformer Network-based Reinforcement Learning Method for Power Distribution Network (PDN) Optimization of High Bandwidth Memory (HBM)," *IEEE Trans. Microw. Theory Techn.*, vol. 70, no. 11, pp. 4772-4786, Nov. 2022.

[9] X. J. Wang and T. Kwasniewski, "Propagation Delay-Based Expression of Power Supply-Induced Jitter Sensitivity for CMOS Buffer Chain," in *IEEE Trans. Electromagn. Compat.*, vol. 58, no. 2, pp. 627–630, April 2016.