Power Delivery Network Design and Modeling for High Bandwidth Memory (HBM)

Wenjun Shi, Yaping Zhou, and Sunil Sudhakaran

NVIDIA Corporation
Santa Clara, California 95050
wshi@nvidia.com

Abstract --- A modeling method to consider simulation switching noise of HBM and its impact on HBM timing is described. This method combines partial element equivalent circuit model for power delivery network and S-parameters based HBM channel model together in HBM studies.

Keywords—High Bandwidth Memory, simultanuous switching noise, power supply induced jitter

I. INTRODUCTION

As graphics chips become faster and faster, the demand for memory bandwidth and speed continues to increase. Even today's fastest speed GDDR5 is beginning to reach the limits of bandwidth, efficiency and cost-effectiveness. High Bandwidth Memory (HBM) is a new type of memory. It uses vertically stacked memory dies interconnected by "through silicon vias (TSV)", and the memory chips communicate to CPU or GPU through ultra-fine and wide parallel bus in silicon interposer.

The second generation high bandwidth memory (HBM2) [1] runs at 2Gbps over single-ended, non-terminated channels. Each memory chip has 1024 data, 112 address, 72 pairs of strobes/clock, 128 data-bit-inversion (DBI), 128 DM/CB, and 80 other signals. The number of signals is very large, and to make the situation even worse, more than one HBM chips are usually integrated into one system. Simultaneous switching noise (SSN) with so many single ended signals is a huge challenge to power delivery network (PDN) designs. This paper focuses on modeling such PDNs and studying the impact from SSN.

II. HBM2 PDN Modeling and Design

The GPU system that we're studying integrates four HBM2 chips with GPU on a silicon interposer. Figure 1 shows the top view of the system.

Figure 1: Top view of GPU with four HBM2 chips

To consider the worst case simultaneous switching noises, we need to accurately find out the maximum number of switching signals. HBM memory implements a DBI AC signal per byte to reduce the number of signals simultaneously switching. Table 1 shows how this counting is done. In our system with four HBM chips, there are potentially 3328 signals switching at the same time.

Table 1: Number of simultaneous switching signals

Signals	Per Channel	Per HBM	Current System	Counts towards SSN	Comments
DQ	128	1024	4096	2048	Half due to DBI
WDQS	4 pairs	32 pairs	128 pairs	**256**	
RDQS	4 pairs	32 pairs	128 pairs	0	When writing
DM_CB	16	128	512	512	
DBI	16	128	512	0	1 DBI per 8 DQs. When N=4 switch, DQ SSN=4 and no DBI. When N=5 switch, DQ SSN=3 and DBI switches, total 4. Considered in DQ count
PAR	4	32	128	0	Not enabled
ADD	14	112	448	448	
CLK	1 pairs	8 pairs	32 pairs	**64**	
CKE	1	8	32	0	Rarely toggle
AERR	1	8	32	0	Rarely toggle
DERR	4	32	128	0	Rarely toggle
Total	**202**	**1616**	**6464**	**3328**	

HBM signal channels are in silicon interposers. These channels are dominated by the effect from channel resistance and capacitance, and the channels can be approximated as RC to consider the first order effect. To effectively drive these RC dominated channels, strong driver strength is required. In HBM2, the maximum transmit driver current specification is

18mA [1]. High switching current needs a robust PDN to mitigate simultaneous switching noise impact, and requires good PDN modeling to assist PDN design decisions.

We devised a modeling method (as shown in Figure 2) to simulate HBM2 simultaneous switching noises (SSN) and its impact on signal eye diagrams due to power noise induced jitter,. This method considers the limitation of electromagnetic EDA modeling tools, and it consists of two main parts:

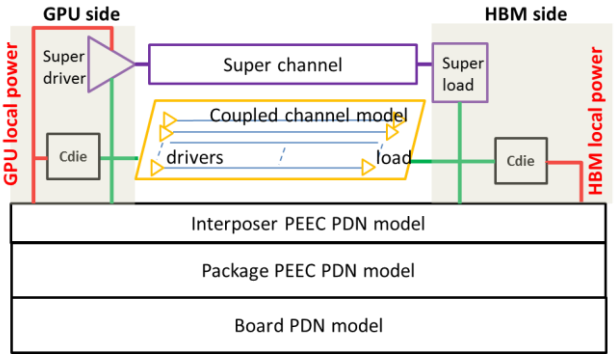


Figure 2: Method to model SSN and power supply induced jitter

1) Partial element equivalent circuit (PEEC) [2] based model is used to model silicon interposer and package PDN noises. To be able to model the worst case switching of 3328 signals in four HBM2 chips, we need to model power, ground and all signals together in PEEC modeling to correctly consider all possible current switching and return loops. Since HBM signals in silicon interposer and power/ground in interposer and package are too complicated, there are no electromagnetic EDA tools that can extract PEEC models for the full system. In our modeling, we extract signals and power/group PEEC models for a portion of the system, and use this model multiple times in parallel to represent the full system PEEC model. In this model, there is a super channel [3] representing 3328 signal channels in parallel, and this channel is driven by that many driver transistor models in parallel (super driver) and loaded with that many receivers in parallel (super receiver) too. To excite the worst SSN, the super driver switches with patterns including 101010…1010000000…0000 to excite chip-package resonance in the PDN. This model provides SSN with explicit ground bounce effect, but transistors in super driver and super receiver always see the local power and ground.

2) Once SSN is modeled, its impact on signal jitter is assessed by including small number of individual driver models and the accurate channel model for these signals. Eleven adjacent signals are chosen in our simulations so that we can consider the worst crosstalk noise in the channels. The channel model for these eleven signals is a loop-based S-parameters model. Care is taken to make sure that the eleven driver/receiver transistor models see their local power and ground, and the ground nodes for the S-parameters models don't short ideal ground in SPICE simulations. Since switching current from these eleven signals is much smaller than the 3328 signals generating SSN, the addition of these signals to above PEEC modeling is not expected to have large impact on SSN from PEEC modeling.

This method models HBM PDN with PEEC model and model a small number of HBM signals with S-parameters model. This method is used to assist decision making on PDN design and decoupling solution, and to study pre-silicon timing budget.

Package capacitors can be added to the top and bottom layers for PDN decoupling. We decide to use low profile capacitors on package bottom layer mainly for middle-frequency decoupling and use larger capacitance capacitors on package top layer for lower frequency decoupling. Another decoupling capacitor related decision is capacitor type. There are standard capacitors as well as low ESL ones (reverse geometry, 3-terminal, 8 terminal LICA etc.). We eventually choose standard two-terminal capacitors.

There are three IO and memory related power rails in HBM2 (Figure 3): GPU VDDQ, HBM VDDQ, and HBM core power VDDC. All power rails are specified to be 1.2V for HBM2. Inside HBM2 chips, VDDQ and VDDC are separated, and HBM VDDC has much higher on-die capacitance than HBM VDDQ. In the meantime, HBM VDDQ has higher capacitance than GPU VDDQ on-die capacitance. Whether we can short these rails together and how to short them are also important design decisions. After many studies with simulations, we decide to short all three power rails together in the package, utilizing the large HBM VDDC on-die capacitance to help address IO power noise from simultaneous switching. Since HBM chips and GPU are very close to each other, and power and ground are very well connected in silicon interposer and package, this decision also makes intuitive sense. Figure 4 shows impedance profiles from GPU VDDQ point of view in different cases. When all three rails are separated, chip-package resonance has a high impedance peak. This resonance peak is reduced when GPU VDDQ and HBM VDDQ are shorted on interposer/package, and it's further reduced greatly if HBM VDDC is also shorted together.

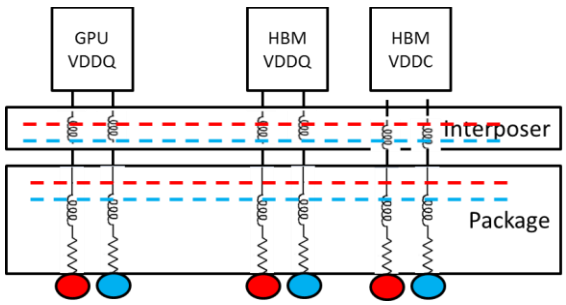


Figure 3: Three main ways to short HBM2 related power rails

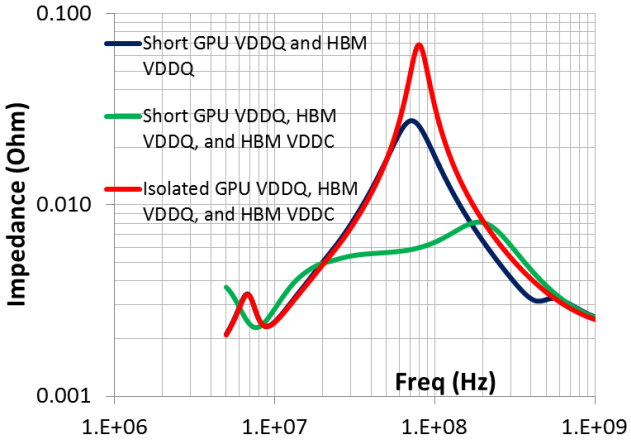


Figure 4: Impedance profiles from GPU HBM VDDQ point of view in different cases

III. SSN and Power Supply Induced Jitter (PSIJ)

SSN is simulated with above modeling method. Figure 5 shows an example. Here the super driver is switching 1010…100000…00 type data pattern. With all the power delivery design optimization, middle frequency peak-to-peak noise is about 40mV. Noise waveform also has high frequency components from individual toggling of super driver. This large noise is due to high driver strength (18mA) used in HBM2.

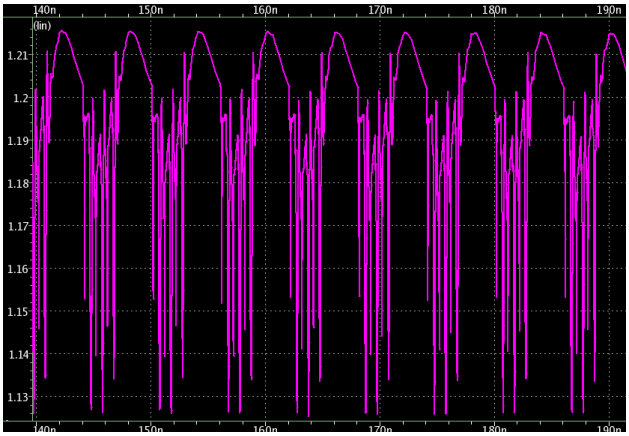
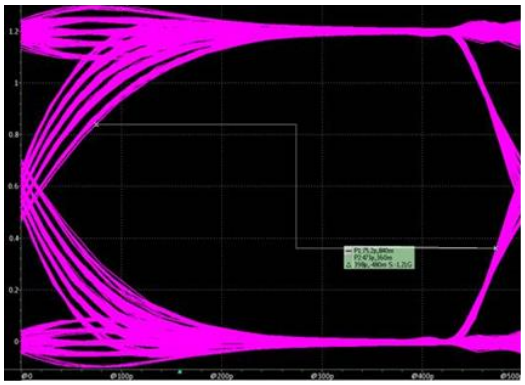


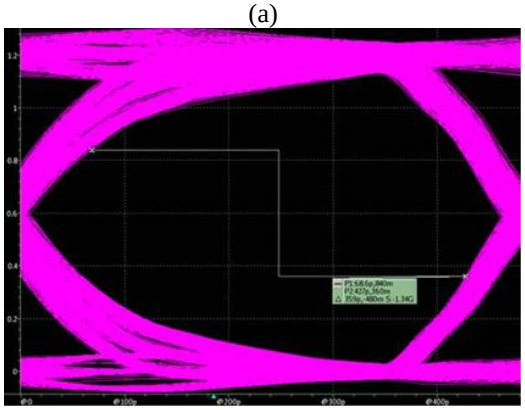
Figure 5: Simultaneous switching noises

Once HBM SSN is modeled, its impact on HBM signal timing budget can then be assessed. Post-layout transistor model of a signal brick can be placed in this noisy power environment, and power supply induced jitter on clock buffer distribution and IO pads can be simulated. This is an important part in pre-silicon design phase to determine HBM2 signal timing budget.

When HBM channel models for eleven signals and their driver and receiver models are added, as described in previous section, we can simulated receiver eye diagrams. The eye diagrams consider channel effect dominated by RC, crosstalk in the channels due to very dense signal routings in silicon interposer, and SSN impact on driver and receivers. Fig. 6 shows eye diagrams with (a) and without (b) considering SSN effect. Eye diagram is closed by ~40ps from SSN. Note that HBM2 eye width is measured from VIH/VIL[1].



(a)



(b)

Figure 6: Eye diagrams with (a) and without (b) PSIJ

IV. Summary

A method is described to model HBM simultaneous switching noises and power induced jitter on HBM signals. This method combines PEEC modeling for PDN and accurate S-parameters based HBM channel model together. Examples are used to demonstrate how this method is used to make HBM power supply design decisions and help pre-silicon timing budget study.

V. Reference:

[1] High Bandwidth Memory Specifications from Jedec website (https://www.jedec.org/news/pressreleases/jedec-updates-groundbreaking-high-bandwidth-memory-hbm-standard)

[2] A. E. Ruehli: Equivalent Circuit Models for Three-Dimensional Multiconductor Systems, IEEE Transactions on Microwave Theory and Techniques, Vol. 22 (1974)

[3] M. Ha, et al, A study of reduced-terminal models for system level SSO noise analysis, EPEPS 2010.