

Design, Packaging, and Architectural Policy Co-optimization for DC Power Integrity in 3D DRAM

Yarui Peng¹, Bon Woong Ku¹, Younsik Park²,
Kwang-II Park², Seong-Jin Jang², Joo Sun Choi², and Sung Kyu Lim¹
¹School of ECE, Georgia Institute of Technology, Atlanta, GA, US
²Samsung Electronics, Hwaseong-si, Gyeonggi-do, Korea
yarui.peng@gatech.edu, limsk@ece.gatech.edu

ABSTRACT

3D DRAM is the next-generation memory system targeting high bandwidth, low power, and small form factor. This paper presents a cross-domain CAD/architectural platform that addresses DC power noise issues in 3D DRAM targeting stacked DDR3, Wide I/O, and hybrid memory cube technologies. Our design and analysis include both individual DRAM dies and a host logic die that communicates with them in the same stack. Moreover, our comprehensive solutions encompass all major factors in design, packaging, and architecture domains, including power delivery network wire sizing, redistribution layer routing, distributed, and dedicated TSV placement, die bonding style, backside wire bonding, and read policy optimization. We conduct regression analysis and optimization to obtain high quality solutions under noise, cost, and performance tradeoff. Compared with industry standard baseline designs and policies, our methods achieve up to 68.2% IR-drop reduction and 30.6% performance enhancement.

Categories and Subject Descriptors

B.3.2 [Memory Structures]: Design Styles

Keywords

3D DRAM, design, packaging, architectural policy, IR drop

1. INTRODUCTION

Modern computer systems require ever-increasing memory bandwidth and capacity. By stacking multiple DRAM dies and using through-silicon-vias (TSVs) as vertical connections, 3D DRAM becomes a promising solution that provides high memory bandwidth and capacity with low power consumption. One challenge in 3D DRAM is unreliable power delivery, the result of more devices requiring current while the number of bumps that can fit is smaller. In addition, DRAM dies are mounted on top of a processor, resulting in longer paths to the power supply.

To mitigate power delivery issues in 3D DRAM, several studies have proposed design and packaging techniques. Edge TSVs

This work is supported by Samsung Electronics Co., Ltd.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org. DAC'15 June 07 - 11, 2015, San Francisco, CA, USA
Copyright 2015 ACM 978-1-4503-3520-1/15/06 ... \$15.00
<http://dx.doi.org/10.1145/2744769.2744819>

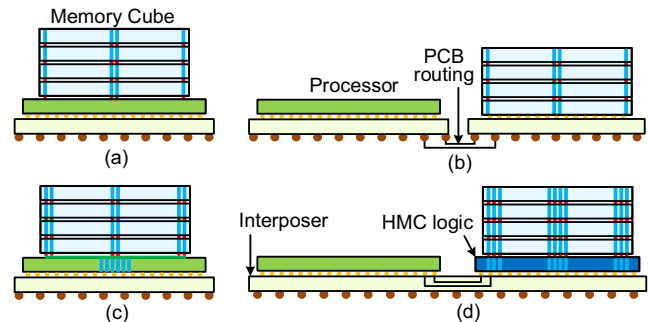


Figure 1: Default configurations of four 3D DRAM designs. (a) on-chip stacked DDR3, (b) off-chip stacked DDR3, (c) Wide I/O, and (d) HMC.

are used in a stacked DDR3 design [2] to reduce power noise. Sub-bank partitioning with local decoupling capacitors is proposed in [5] to maintain DRAM regularity. Another study [6] found TSV alignment to be effective at reducing IR drop and current crowding. To achieve low power distribution network (PDN) impedance, a redistribution layer (RDL) is added between memory and logic die in [1]. From the memory controller perspective, the relationship between bank activity and IR drop in a hybrid memory cube (HMC) is characterized in [4], which proposed an optimized request scheduling policy that addresses the bank starvation problem. This policy is appropriate for designs with high vertical IR drop but has little impact on designs with many TSVs when horizontal IR drop dominates.

Most studies focusing on a single isolated solution are limited to face-to-back (F2B) bonding. Our goal is to conduct comprehensive research covering many key solutions from multiple domains. To accomplish this goal, we develop a cross-domain CAD platform that accurately models and evaluates DC power integrity in 3D DRAM. This work investigates the impact of logic/memory interaction, TSV and RDL optimization, wire bonding, face-to-face (F2F) bonding, and read scheduling policy on IR drop and performance. We use four modern 3D DRAM benchmarks: off-chip stacked DDR3, on-chip stacked DDR3, Wide I/O, and HMC shown in Figure 1. Our design, packaging, and architectural domain solutions are co-optimized to achieve the best solutions under IR drop, performance, and cost tradeoffs. To the best of our knowledge, this study is the first to comprehensively analyze and optimize the power integrity of modern 3D DRAMs across multiple domains.

2. SIMULATION INFRASTRUCTURE

2.1 3D DRAM Benchmarks

Table 1: Benchmark specifications

Benchmark	Stacked DDR3 [2]	Wide I/O [3]	HMC [5]
Capacity	4Gb × 4 dies = 16Gb		
Stand-alone?	yes/no	no	yes
Stacked logic die	T2 (or none)	T2	HMC logic
Logic size (mm ²)	9.0×8.0	9.0×8.0	8.8×6.4
DRAM size (mm ²)	6.8×6.7	7.2×7.2	7.2×6.4
# banks per die	8	16	32
# channel	1	4	16
Speed (Mbps/pin)	1600	200	2500
Data width	8	512	512
3D IC benefit	capacity	low power	bandwidth
Target app	PC & laptop	mobile	GPU & server

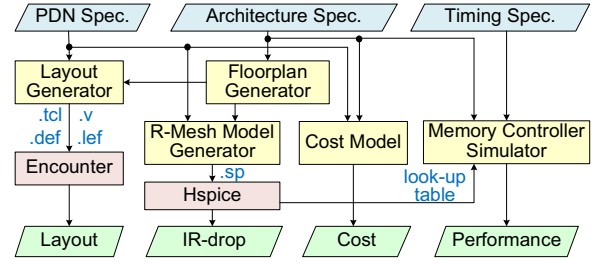
To provide wide coverage of various 3D DRAM applications, we choose stacked DDR3 [2], Wide I/O [3], and HMC [5] as benchmarks and assume that the stacked DDR3 can be configured as a separate chip (off-chip) or mounted on logic (on-chip). We use published designs references and scale power measurement results from Samsung and Micron into 20nm-class DRAM technology. To ensure that our study is both realistic and up-to-date, we obtain detailed DDR3 power maps through our industry collaborations.

We choose the following three benchmarks because of their unique target applications: (1) Stacked DDR3 provides a low-cost and backward-compatible 3D DRAM solution. With no footprint increase, memory capacity can be easily extended. However, 3D stacking is not considered in the original 2D DDR3 design. As a result, performance and power in its 3D implementation are not fully optimized. (2) Wide I/O, with a large number of I/O connections, can be mounted directly on top of logic processors reducing both power and footprint. JEDEC specifications dictate that micro-bumps be located in the center of both memory and logic dies. (3) Micron proposes HMC [5] as the next-generation high-performance memory solution with large power consumption. Mounted on top of its own control logic die, HMC handles communication with the processor through a silicon interposer. Moreover, we use a full-chip OpenSPARC T2 processor in 28nm technology as the host chip. The design specifications of our benchmarks are listed in Table 1. We use stacked DDR3 as an example in sections 3 to 5 and results for all four benchmarks are provided in Section 6.

2.2 CAD Platform and Validation

We propose an integrated CAD and architectural simulation platform shown in Figure 2. Our CAD solution combines commercial tools with in-house tools implemented in C++ and includes a series of scripts for design automation. Our floorplan generator produces a block-level 3D DRAM floorplan based on the given design and architectural specifications. Then our PDN layout generator produces design files for PDN routing. Our memory die floorplan consists of blocks such as arrays, row/column decoders, and peripheral circuits. Then, our tool reads the corresponding power map. Next, we perform special routes and produce a combined floorplan with both globally and locally-routed PDN using Cadence Encounter. Figure 3 presents two examples of our auto-generated layouts which are used for pre-design analysis of routing congestion and early-stage routing planning. Lastly, we calculate the cost of design and packaging solutions that include metal usage, TSV count and location, RDL, and bonding style. We also use our memory controller simulator to obtain performance data.

For the IR-drop calculation, we build a resistive mesh model (R-Mesh) for each metal layer based on design and technology information. As we focus on power supply noise, the R-Mesh is built for VDD only. However, the ground net can be analyzed in complementary fashion as well. PDN wire resistance is modeled depend-

**Figure 2: Our integrated architecture/CAD platform**

ing on the metal layer usage which is defined as the area percentage of VDD PDN on one layer. Local PDN supplies power within each block, while global PDN is used to connect them. The resistivity of each metal layer as well as its routing direction is read from the technology file. PG rings, vias, and inter-die connections are generated automatically. We use HSPICE to simulate it and calculate the IR drop. Figure 3 shows two R-Mesh model examples.

Since each row activation contains a write-back operation when the row is closed, we focus on read operation only. We generate a 2D DDR3 design using the aforementioned CAD method. For one bank operation, the max IR drop is 22.5mV for read and 22.4mV for write, and their IR-drop distributions are similar. However, in 3D DRAM, the maximum IR drop depends on both single die operation and inter-die coupling. For naming convenience, the 3D DRAM memory state is represented as " R_1 - R_2 - R_3 - R_4 ," where R_1 to R_4 are the number of active banks from the bottom DRAM die (DRAM1) to the top die (DRAM4). The default state is 0-0-0-2 assuming zero-bubble interleaving read (IDD7) in our stacked DDR3.

To verify our R-Mesh model, we compare IR-drop results with commercial tools shown in Figure 4. Using Encounter Power System (EPS) on the generated 2D DDR3 design, we perform IR-drop simulation assuming that the left two banks are in the interleaving read mode. The max IR drops are 32.6mV and 32.2mV using EPS and R-Mesh, respectively. Our R-Mesh model shows only 1.3% error and achieves 517x speed up because it does not perform parasitic extraction from the layout and reduces the total resistor count.

2.3 Memory Controller Simulator

For 3D DRAM operations, the IR-drop constraint is a critical factor that affects the memory performance. In the standard JEDEC DDR3 specifications, two timing parameters, used to limit the maximum IR drop, are row to row delay (tRRD) and four active window (tFAW). Without considering detailed 3D stacking properties, these timing parameters limit the maximum number of banks that can be read in parallel. Thus, less parallelism reduces the maximum performance of the 3D DRAM.

To study memory performance, we build a 3D DRAM memory controller simulator that performs cycle-by-cycle simulations for each DRAM bank and memory channel. Major DRAM read operation timing parameters such as tCL, tRCD, tRP, tRAS, and tCCD are modeled. If an active bank does not receive further read requests in a few cycles, the bank is closed to reduce IR drop. We generate 10,000 read requests with temporal and spatial locality under a row hit rate of 80%. For stacked DDR3, each read request arrives every five DRAM cycles with a burst length of eight, assuming a heavy work load. Our memory controller has a priority queue of size 32 so that it can smartly schedule the requests for the best performance. Interleaving mode reads two banks per die in maximum to avoid current overdrawn from charge pump.

3. DESIGN SOLUTIONS

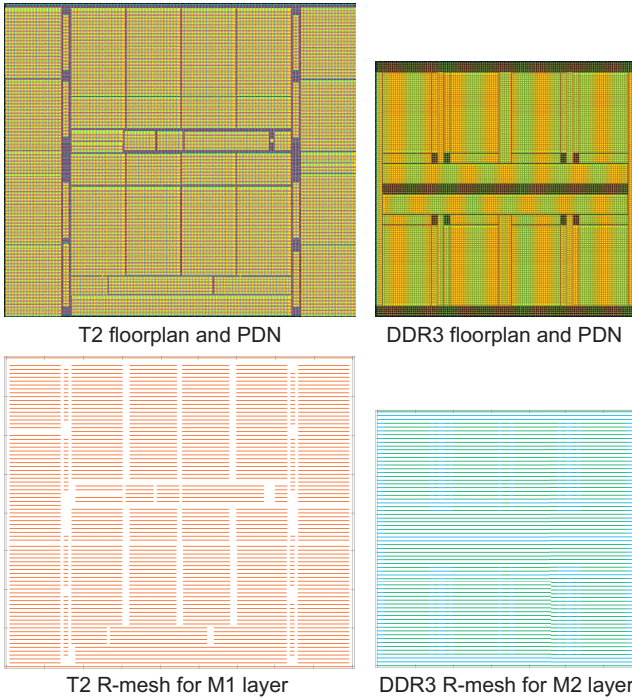


Figure 3: Automatically generated layouts and their R-Mesh models for T2 full-chip and stacked DDR3

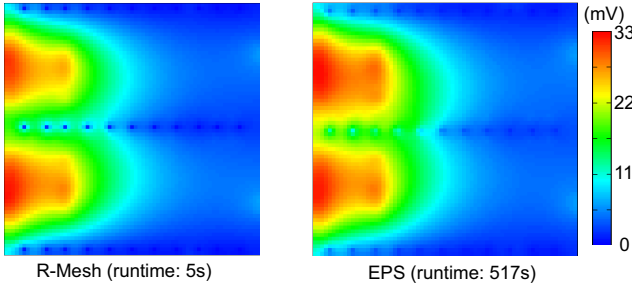


Figure 4: Validation of R-Mesh against Cadence EPS

A traditional design technique for IR-drop reduction is to increase metal usage which also applies to 3D IC. Assuming a 10% M2 usage and 20% M3 usage for VDD as baseline, with 2x PDN metal usage, IR drop is reduced more than 40% for stacked DDR3. However, the vertical IR drop becomes more significant in 3D IC. Thus, we explore unique design solutions in 3D IC.

3.1 Stand-alone vs. Mounted on a Logic Die

Depending on the application, 3D DRAM can be mounted on logic (on-chip) or separated as a stand-alone chip (off-chip). For mounted memory, one solution for stable power supply is to add dedicated PG TSVs on the logic die. Dedicated TSVs can be fabricated through via-last technology, which reduces TSV resistance and provides a clean power supply directly to memory dies. However, these dedicated TSVs penetrate the bottom die, occupy extra silicon area, and become routing blockages on logic, increasing design complexity and logic die cost dramatically. Assuming the same supply voltages of the logic and the DRAM die, power and ground nets from both dies can be connected together, thus their power noises are coupled. As results show, with a 50.05mV logic die power noise, the DRAM IR drop increases from 30.03mV in the off-chip stacked DDR3 design to 64.41mV in the on-chip design.

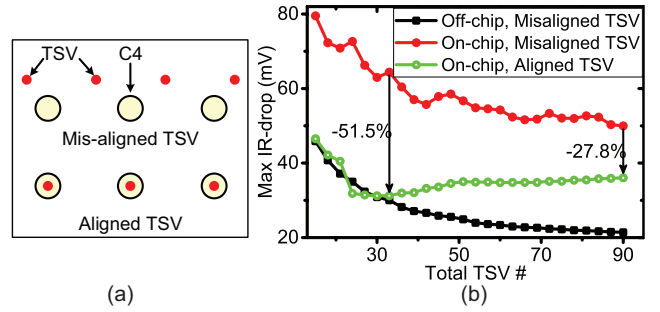


Figure 5: (a) C4-TSV alignment, and (b) TSV count and alignment impact in stacked DDR3

Table 2: Comparison of TSV and RDL options in Figure 6

Design option	(a)	(b)	(c)	(d)
Logic die cost	High	Low	Medium	Medium
DRAM die cost	High	Low	High	Medium
Overall cost	Highest	Lowest	High	Medium
IR drop(mV)	30.03	50.76	38.46	49.36

3.2 Impact of TSV Count and Alignment

Another intuitive design solution is to increase the PG TSV count. More PG TSVs reduce vertical IR drop and current crowding. However, if a uniform TSV pitch is assumed, not all TSVs can perfectly align with C4 bumps on the logic die. The misaligned TSV increases the inter-die coupling resulting in a higher IR drop on the DRAM die. Figure 5 compares the on- and off-chip designs with various TSV numbers. The results show that using more TSVs reduces IR drop, but the reduction saturates with many TSVs. By carefully placing TSVs near C4 bumps on the logic die and reducing average C4-to-TSV distance, IR drop reduces by as much as 51.5% in on-chip stacked DDR3 while logic IR drop merely increases by 0.2%. More TSVs do not always guarantee a lower IR drop because of TSV misalignment, especially when the TSV count is small. For on-chip designs, increasing the TSV count leads to larger coupling from T2. Thus, the IR drop increases slightly on memory dies.

3.3 Impact of TSV Location and RDL

Various TSV design considerations affect the max IR drop. Edge TSVs [2] can significantly reduce the IR drop by shortening the power supply path. However, dedicated edge TSVs introduce much higher cost to both logic and DRAM because large keep-out zones (KOZs) must be inserted around TSVs to avoid stress and noise issues. A low-cost solution called “center TSV” groups all TSVs into the center of the die and does not block routing on the logic die. To alleviate the high IR drop, the RDL can be added as a backside routing layer. Unlike routing layers fabricated using the silicon process, the RDL is much thicker and allows non-manhattan routing. With a much lower resistivity, the RDL is easy for fabrication. Thus, it is suitable to deliver power to the edge of DRAM chips at lower cost. A RDL can be inserted only between logic and bottom DRAM die or on all dies. Figure 6 shows four design options, and Table 2 compares their tradeoffs between cost and IR drop. Center TSV without a RDL has the lowest cost but highest IR drop. Replacing edge TSVs with a RDL reduces cost but introduces higher power noise because of additional RDL resistance.

4. PACKAGING SOLUTIONS

4.1 Impact of Dedicated TSVs and Wire Bond

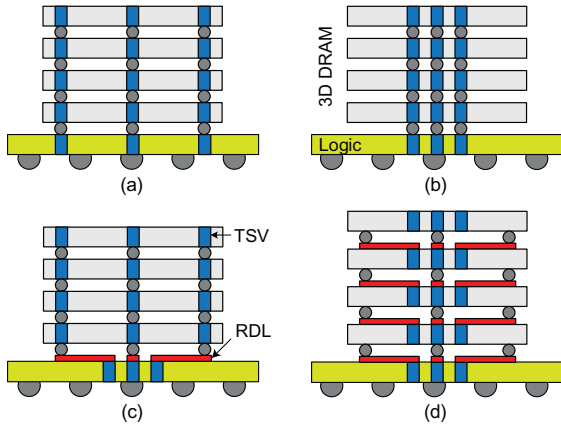


Figure 6: TSV locations in 3D DRAM vs. logic and their RDL needs. (a) edge (memory) + non-center (logic), (b) center + center, (c) edge + center + RDL, and (d) center + center + RDL

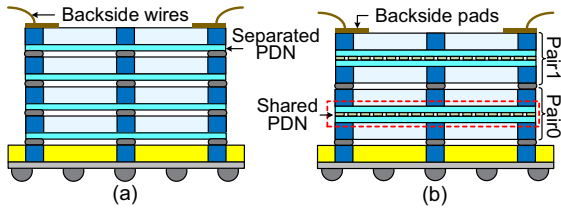


Figure 7: Wire-bonding cross-section view: (a) F2B, (b) F2F

In addition to design techniques, advanced packaging solutions also help improve power integrity in 3D DRAM. To alleviate the inter-die impact shown in Section 3.1, dedicated TSVs can be used to directly deliver power to the DRAM dies. With this packaging solution, the logic and the DRAM PDNs are fully decoupled, which results in an IR drop similar to that of the off-chip design.

In a 3D DRAM design, layouts of all DRAM dies are kept identical so that all memory dies share the same fabrication process, which improves the yield and cost. By taking advantage of the backside metallization process, additional metal pads for wire connections are formed on the backside. Figure 7 (a) shows the proposed packaging solution with wire bonding. Signal TSVs are used for low-power and high performance, and PG TSVs are used to supply power between memory dies. However, with backside wire bonding, an extra power delivery path is built from the top to the bottom die. With this method, the maximum IR drop reduces, and bonding wires can directly connect to large off-chip decoupling capacitors, which provide better AC power integrity. Table 3 summarizes impact of dedicated and wire bonding on the stacked DDR3 design. Both dedicated TSVs and wire bonding reduce the IR drop as much as 50% for on-chip designs. However, since both wire bonding and dedicated TSVs provide direct power supply, a combination of both technologies provides only marginal additional benefits.

4.2 Impact of PDN Sharing with F2F Bonding

Another packaging technique also takes advantage of layout regularities in 3D DRAM. Traditional DRAM technology uses three metal layers: M1 for signal routing, M2 for mixed signal/power routing, and M3 for power routing. Since memory has a highly regular layout, the PDN is usually designed symmetrically. Thus, by changing the die orientation of DRAM1 and DRAM3, F2F bonding can form between the two bottom dies and the two top dies. F2F vias can be placed almost everywhere, thus, PDNs of two

Table 3: Impact of dedicated TSVs and wire bonding

Design	Dedicated TSV?	IR drop (mV)		
		Baseline	Wire-bonded	$\Delta\%$
On-chip	no	64.41	30.04	-53.4%
On-chip	yes	31.18	27.18	-12.8%
Off-chip	yes	30.03	27.10	-9.76%

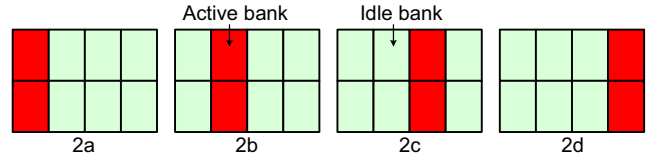


Figure 8: Four cases of the two-bank interleaving read state

F2F-bonded dies are tightly connected. In this way, a pair of F2F-bonded dies share their PDNs together. B2B bonding is used to connect two DRAM pairs and the bottom DRAM die with the logic die. For signal pads on the top metal layer, it is preferable that they be placed on the symmetry axis. Even if asymmetrical I/O pins are placed on the top metal layer, no re-design for F2F bonding is needed. Simply mirroring all masks or using multiplexers to select from a pair of symmetrical pads maintains the same layouts for all dies in F2F design. Thus, F2F remains as a low-cost solution for 3D DRAM. F2F bonding can also be used in combination with wire bonding, shown in Figure 7 (b), and provide even larger IR-drop benefits.

Unlike the F2B design, in which each DRAM die uses two metal layers for PDNs, a pair of DRAM dies in the F2F design can use four metal layers together. This feature, called PDN sharing, provides additional IR-drop benefits. If one die in a pair is idle while another is active, the active die can use all four PDN layers. With PDN sharing, the IR drop of the idle die increases but leads to a significant IR-drop reduction for the whole system. For example, under the 0-0-0-2 memory state, the overall maximum IR drop with F2F bonding decreases by 42.8% and 41.1% compared with F2B bonding in off-chip and on-chip stacked DDR3, respectively.

4.3 Impact of Inter-Die Spatial Locality

The memory state has a large impact on F2F benefits as well. For example, Figure 8 shows four cases from the top-down view for the two-bank interleaving read mode, and Table 4 shows IR-drop results. If two dies of a pair have active banks in the same location, it is called “intra-pair overlapping.” With intra-pair overlapping, the current is congested in a small area, and both dies do not have extra PDN resources to share. Results also show that if the active regions on two dies are separated further, the IR-drop reduction is larger with less current congestion. If active banks overlap in different pairs, the impact on the IR drop is small since PDNs between pairs are separated. Thus, F2F provides IR-drop benefits over F2B, especially for designs with low bank activities and low probability of intra-pair overlapping. To avoid inter-pair overlapping, IR-drop-aware read scheduling policies can rearrange bank activities so that the probability of inter-pair overlapping remains low.

5. ARCHITECTURAL SOLUTIONS

5.1 Impact of Memory State and I/O Activity

Assuming zero-bubble reading, if more DRAM dies are activated, I/O activity per die decreases, which leads to lower power consumption per active die. Table 5 lists IR-drop simulations for various cases. For simplicity, the detailed active bank location

Table 4: Impact of intra-pair overlapping in stacked DDR3 for the cases shown in Figure 8

Memory state	Intra-pair overlapping	Max IR drop (mV)		
		F2B	F2F+B2B	$\Delta\%$
0-0-2a-2a	yes	28.14	27.21	-3.3%
0-0-2b-2b		18.06	17.42	-3.5%
0-2a-0-2a	no	27.32	15.24	-44.2%
2a-0-0-2a		26.51	15.24	-42.5%
0-0-2b-2a	no	27.38	17.98	-34.3%
0-0-2c-2a		27.04	17.10	-36.8%
0-0-2d-2a		26.86	15.27	-43.1%

Table 5: Impact of Memory state and I/O activity in off-chip stacked DDR3

Memory state	IO activity per die	Power (mW)		IR drop (mV)	
		active die	total	F2B	F2F+B2B
0-0-0-2	100%	229.5	310.5	30.03	17.18
2-0-0-0			310.5	26.26	14.61
0-0-0-2	50%	175.5	256.5	26.42	15.15
0-0-2-2			405.0	28.14	27.21
0-0-0-2	25%	126.9	207.9	22.93	13.23
2-2-2-2			507.6	24.82	23.57

is not considered, and active banks are assumed to be located on the edge, which is the worst case of a certain memory state. For the 0-0-0-2 state, 25% I/O activity reduces die power by 44.7%, which leads to 23.64% and 22.99% IR-drop reductions for F2B and F2F+B2B designs, respectively. Moreover, if the read activity is balanced among dies (e.g., the 2-2-2-2 state), more banks can be activated in parallel, and the maximum IR drop of that state is even smaller than the 0-0-0-2 state with 100% I/O activity. In addition, worst-IR-drop cases for F2B and F2F differ. For F2F design with PDN sharing, the 0-0-0-2 state does not cause high IR-drop. However, because of the intra-pair overlapping effect, the 0-0-2-2 state becomes the worst case. Compared with F2B, F2F reduces the worst-case IR drop by 9.4%.

5.2 Impact of the Read Scheduling Policy

From the perspective of performance, if the IR drop is not considered during memory operations, the memory controller can activate as many banks as possible if there is no timing violation or bus conflict. However, parallel reading is always limited for power integrity concerns, especially in 3D DRAM. However, since the standard read policy is not aware of 3D stacking, simply limiting row activation pessimistically constrains parallel operations. As shown in Section 5.1, impact of unique memory and I/O activity requires a detailed IR-drop-aware policy for optimum performance. Moreover, as balanced reads increase parallelism in 3D DRAM without IR-drop overhead, distributing read requests evenly achieves the best tradeoff between the IR drop and performance.

Considering detailed 3D DRAM IR drops, we propose IR-drop-aware read policies based on a look-up table. With our fast and accurate R-Mesh model, the max IR drops of each memory state with various I/O activities are saved in a look-up table read by the memory controller for read request scheduling. For each cycle, the memory controller checks all read requests in the priority queue and tries to send a request to each DRAM channel. Under a given IR-drop constraint, the read request that can be sent to memory must satisfy the following conditions: (1) timing specifications are met; (2) sending the request causes no conflict on I/O buses; and (3) the IR-drop constraint is met. This read policy is compared to JEDEC DDR3 standard policy with a tRRD of 8 and a tFAW of 32. Moreover, two request scheduling policies are implemented. One is called first-come-first-served (FCFS), and another is called distributed-read (DistR). For FCFS, the memory

Table 6: Impact of architectural policy in stacked DDR3. Standard policy uses tRRD and tFAW. First-come-first-served and distributed-read are denoted as FCFS and DistR, respectively.

IR-drop policy	Standard Scheduling policy	Our IR-drop-aware policy	
		FCFS	DistR
IR-drop constraint	none	24mV	24mV
Runtime (us)	109.3	84.68 (-22.6%)	75.85 (-30.6%)
Bandwidth (read/clock)	0.114	0.148 (+29.2%)	0.165 (+44.2%)
Max IR drop (mV)	30.03	23.98 (-20.2%)	23.98 (-20.2%)

Table 7: Case study for impact of IR-drop on DRAM performance in off-chip stacked DDR3 design

Mounting style Case #	off-chip			on-chip		
	1	2	3	4	5	6
Bonding style	F2B	F2B	F2F	F2B	F2B	F2F
PDN metal usage	1x	1.5x	1x	1x	1x	1x
Wire bonding	no	no	no	no	yes	no
Max IR drop (mV)	30.03	22.15	17.18	64.41	30.04	65.43

controller assigns a higher priority to the read request which comes in first. For DistR, the memory controller tries to balance the read across multiple DRAM dies to increase die-level parallelism under the IR-drop constraint. Thus, the read request, whose target die has the least number of active banks, has the highest priority.

Table 6 compares the performance of three read scheduling policies based on the F2B stacked DDR3 design. We set the IR-drop constraint for our IR-drop-aware policies to 24mV. With bank activation constraints, the standard policy results in a longer runtime and a lower average bandwidth. With a detailed IR-drop look-up table, the memory performance improves by 22.6%. Furthermore, by taking advantage of DistR and balanced workloads, the performance improves by 30.63%. The maximum IR drop of our policy also decreases by 20.15% compared to the standard policy since memory states with high IR drops are avoided. Note that scheduling policy has a small impact if the IR-drop constraint is high or the bank activity is low. In both cases, not the IR drop but single-bank performance becomes the system bottleneck.

5.3 Impact of IR-drop on DRAM Performance

Since design and packaging optimizations reduce the IR drop, allowed memory states differ for various designs under the same IR-drop constraint. Table 7 lists a few examples. With our memory simulator, impact of various IR-drop optimization methods on performance is studied. Figure 5.3 shows runtime needed to finish all read requests. If the IR-drop constraint is too tight, it allows no memory state. With a relaxed IR-drop constraint, more states are allowed. Therefore, the memory controller can send more parallel read requests. As results show, all IR-drop optimization methods are able to improve memory performance under a certain IR-drop constraint. Interestingly, although the F2F design (Case 3) reduces the worst-case IR drop only by 9.4%, it outperforms the F2B design with 1.5x PDN (Case 2) with an IR-drop constraint smaller than 18mV because PDN sharing shows larger benefits when bank activities are low. Therefore, F2F has a higher tolerance to low IR-drop constraints.

6. CROSS-DOMAIN CO-OPTIMIZATION

6.1 Cost and IR-drop Model

An intuitive way to lower the IR-drop is using every solution available. However, this approach leads to a very expensive design with marginal IR-drop benefits. Therefore, co-optimization of the IR drop, performance, and cost is critical to provide overall guidelines. We propose a cost estimation model with every technology

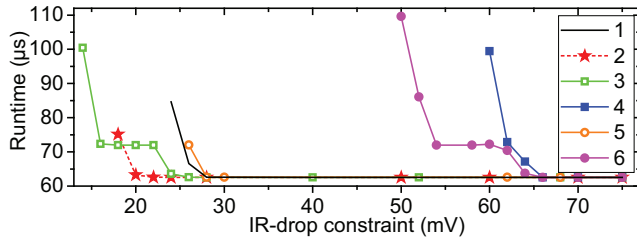


Figure 9: Performance results for the cases shown in Table 7

Table 8: Cost model summary for four benchmarks

Solution	Abbreviation	Input Range	Cost Range
M2 VDD usage	M2	10%-20%	0.025-0.05
M3 VDD usage	M3	10%-40%	0.025-0.10
Power TSV #	TC	15-480	0.078-0.44
Dedicated TSV	TD	Yes/No	0.06/0
Bonding style	BD	F2B/F2F	0.045/0.06
RDL layer	RL	Yes/No	0.05/0
Wire bonding	WB	Yes/No	0.03/0
		Center only (C)	0
TSV location	TL	Edge and center (E)	0.5×TC
		Distributed (D)	TC

parameter included as a cost term. Table 8 lists these cost terms. Except for the TSV count (TC), the cost of which is calculated by a square root function, other terms are proportional to inputs. An input range ensures a realistic solution. For the Wide I/O design, the power TSV count is fixed at 160, which matches specifications. For stacked DDR3 and Wide I/O designs, only center and edge TSVs are options. For HMC, resulting from a high power consumption, PG TSVs are placed between banks. We call this TSV location style “distributed TSV.” The minimum power TSV count is 160 for sufficient supply current.

For technology co-optimization, brute-force searching for every combination in one benchmark takes 4637 hours on a four-core system. To reduce runtime, we choose a few sample cases for M2, M3, and TC, because they are continuous variables. For other optimization options, we search all valid combinations. After performing R-Mesh simulations on the sample cases, we use MATLAB regression analysis to obtain an IR-drop model with a root mean square error (RMSE) of less than 0.135 and an R^2 of larger than 0.999. With the regression analysis, total runtime decreases to ten hours. Combined with total cost estimation, we define an IR-cost term by

$$IR-cost = IR-drop^\alpha \times Cost^{1-\alpha}, \quad (1)$$

where $\alpha \in [0, 1]$. We perform MATLAB global optimization to obtain the best solutions. With $\alpha=0$, we found the lowest cost solution, while $\alpha=1$, the lowest IR-drop solution.

6.2 Putting it Altogether: Best Solutions

Table 9 summarizes the best solutions for all four 3D DRAM designs. As expected, using no optimization option results in the lowest cost but the highest IR drop. By gradually increasing α , results show the priority of each optimization option. We achieve optimal tradeoff with $\alpha=0.3$. Since packaging solutions such as wire bonding and F2F bonding are low-cost solutions but able to reduce IR drop significantly, they have higher priority. Because increasing the TSV count yields only a marginal gain but increases the cost significantly, placing more TSVs on a DRAM chip is unnecessary. The RDL is not a good option for the lowest IR drop. However, for Wide I/O design, since the specifications require that all PG pumps be located in the center, edge TSVs must be paired with RDL for interface connections. With edge TSVs, the IR drop can decline

Table 9: Best options for four benchmarks (see Table 8 for the meaning of abbreviations). α is from Equation (1).

α	M2 (%)	M3 (%)	TC	TL	TD	BD	RL	WB	IR drop (mV)		Cost
									Matlab	R-Mesh	
Stacked DDR3, off-chip											
0	10	10	15	C		F2B	N	N	88.73	88.73	0.23
0.3	20	22	24	E		F2F	N	N	22.75	23.01	0.37
1	20	40	360	E	Y	F2F	N	Y	9.733	9.540	0.87
Baseline	10	20	33	E		F2B	N	N	30.03	30.03	0.35
Stacked DDR3, on-chip											
0	10	10	15	C	N	F2B	N	N	117.6	117.6	0.17
0.3	20	22	21	E	N	F2B	N	Y	25.51	27.09	0.32
1	20	40	420	E	Y	F2F	N	Y	9.864	9.843	0.92
Baseline	10	20	33	E	Y	F2B	N	N	31.18	31.18	0.35
Wide I/O											
0	10	10		C	N	F2B	N	N	110.1	110.2	0.35
0.3	20	40	160	E	Y	F2F	Y	Y	4.864	4.841	0.73
1	20	40	160	E	Y	F2F	Y	Y	4.864	4.841	0.73
Baseline	10	20		E	Y	F2B	Y	N	13.56	13.62	0.62
HMC											
0	10	10	160	C	N	F2B	N	N	459.7	459.7	0.35
0.3	20	25	160	D	Y	F2B	N	Y	18.63	18.65	0.76
1	20	40	480	D	Y	F2B	N	Y	13.76	13.84	1.17
Baseline	10	20	384	E	Y	F2B	N	N	47.90	47.90	0.77

to below 20mV for the stacked DDR3 and the Wide I/O designs. However, only with distributed TSVs for HMC can the same IR drop be achieved. Because of the likelihood of inter-die overlapping, the F2F benefit declines in HMC. However, distributed TSVs are preferable for the stacked DDR3 and the Wide I/O designs.

7. CONCLUSION

This paper investigated impact of various design, packaging, and architectural policy options on 3D DRAM DC power integrity. Based on our CAD/architectural platform and four 3D DRAM benchmarks, results showed that inter-die coupling, the TSV count, location, and alignment strongly affected the IR drop. We used the RDL to replace edge TSVs at the cost of a higher IR drop. Packaging solutions such as backside wire bonding and F2F bonding reduced the IR drop significantly with low cost overhead. With regard to performance, our IR-drop-aware policies optimized performance as much as 30.6%. Distributing activity to multiple DRAM dies reduced the IR drop and increased performance under a tight IR-drop constraint. Based on the regression analysis, we proposed best co-optimization solutions for the stacked DDR3, Wide I/O, and HMC designs.

8. REFERENCES

- [1] W. Beyene et al. Signal and power integrity analysis of a 256-GB/s double-sided IC package with a memory controller and 3D stacked DRAM. In *Electronic Components and Technology Conference*, pages 13–21, May 2013.
- [2] U. Kang et al. 8 Gb 3-D DDR3 DRAM Using Through-Silicon-Via Technology. *Journal of Solid-State Circuits*, 45(1):111–119, Jan 2010.
- [3] J.-S. Kim et al. A 1.2 V 12.8 GB/s 2 Gb Mobile Wide I/O DRAM With 4×128 I/Os Using TSV Based Stacking. *Journal of Solid-State Circuits*, 47(1):107–116, Jan 2012.
- [4] M. Shevgoor et al. Quantifying the Relationship Between the Power Delivery Network and Architectural Policies in a 3D-stacked Memory Device. In *International Symposium on Microarchitecture*, MICRO-46, pages 198–209, 2013.
- [5] Q. Wu and T. Zhang. Design Techniques to Facilitate Processor Power Delivery in 3-D Processor-DRAM Integrated Systems. *Transactions on Very Large Scale Integration Systems*, 19(9):1655–1666, Sept 2011.
- [6] X. Zhao, M. Scheuermann, and S. K. Lim. Analysis and Modeling of DC Current Crowding for TSV-Based 3-D Connections and Power Integrity. *Transactions on Components, Packaging and Manufacturing Technology*, 4(1):123–133, Jan 2014.