Understanding the Role of the Power Delivery Network in 3D-Stacked Memory Devices

Manjunath Shevgoor[†], Jung-Sik Kim[‡], Niladrish Chatterjee[†], Rajeev Balasubramonian[†],
Al Davis[†], Aniruddha N. Udipi[⋆]
[†]University of Utah, [‡]DRAM Design Team, Memory Division, Samsung Electronics , [⋆]ARM R&D

Abstract

Many of the pins on a modern chip are used for power delivery. If fewer pins were used to supply the same current, the current would have to travel farther on the chip to reach the same destination. This results in an "IR-drop" problem, where some of the voltage is dropped across long resistive wires and a lower voltage is supplied to the circuits. The same problem also manifests if the pin count is the same, but the current draw is higher. IR-drop is especially problematic in 3D DRAM devices because (i) low cost (few pins) is a high priority, and (ii) DRAM devices are the first to embrace 3D-stacking which increases current draw without providing proportionate room for more pins.

This paper is the first to characterize the relationship between the power delivery network and the maximum supported activity in a 3D-stacked DRAM memory device. The design of the power delivery network determines if some banks can handle less activity than others. It also determines the combinations of bank activities that are permissible. Both of these attributes can feed into architectural policies. For example, if some banks can handle more activities than others, the architecture benefits by placing data from high-priority threads or data from frequently accessed pages into those banks. The memory controller can also derive higher performance if it schedules requests to specific combinations of banks that do not violate the IR-drop constraint.

We first define an IR-drop-aware scheduler that encodes a number of activity constraints. This scheduler, however, falls short of the performance of an unrealistic ideal PDN that imposes no scheduling constraints by 4.6x. By addressing starvation phenomena in the scheduler, the gap is reduced to only 1.47x. We present a case study, where we profile the application and use this information to place pages. This Profile Page Placement scheme brings the performance to 15.3% of the unrealistic ideal PDN. We thus show that architectural polices can help mitigate the limitations imposed by a cost constrained design.

1. Introduction

DRAM supply voltages have been dropping every generation in order to improve power efficiency in DRAM. However, as supply voltage decreases, circuits become increasingly more sensitive to power supply noise. A 100 mV supply noise on a 1 V system is a much greater threat to correctness than on a 2.5 V system. This has resulted in an increase in the importance afforded to the power delivery network in DRAMs, not traditionally a major area of focus [13].

Of the hundreds of pins on a chip, more than half are used to supply power and ground. These power pins are scattered across the chip so that the supply current need not travel very far on the chip. Some of the supplied voltage is dropped across the PDN; this is a function of the supplied current *I* and the resistance of the PDN *R*. This is commonly referred to as *"IR-drop"*. If the IR-drop is very high, a lower supply voltage is delivered to the chip's circuits, possibly leading to incorrect operation. For example, in commercial DDR3 DRAM chips [10] (page 111), if the supply voltage is rated at 1.5 V, the minimum allowed voltage at the circuits is specified to be 1.425 V, i.e., up to 75 mV can be dropped across the PDN.

The IR-drop becomes unacceptable if the DRAM chip is either drawing too much power, or if the PDN's resistance is too high. The latter is kept in check by using many pins for power delivery and ensuring that current travels on relatively short wires. The former is kept in check by imposing limits on the maximum activity on the chip. For example, DRAM chips allow a maximum of four row activations within the timing parameter *tFAW*. Other examples also exist, such as the timing parameter *tRRD* [8] (page 429), which imposes a minimum gap between consecutive DRAM Activates (These constraints are in place to preserve the integrity of the Power Delivery, and the Charge Pumps). The floorplan and pin layout are design-time decisions made at the circuit level, and as architects we have little control over this. We focus on controlling the activity on the device.

However, because of technology and market forces, the values of *I* and *R* are being raised. First, the onset of 3D-stacking will increase the current draw per package. Micron is slated to release its 3D-stacked memory+logic device, the Hybrid Memory Cube (HMC), next year. Such devices not only have to feed current to 4 or 8 DRAM chips on the stack, but also to high-power SerDes circuits on the logic layer. Second, DRAM memory devices are highly cost sensitive. The packaging cost of the device is a linear function of the number of pins. This is nicely illustrated by Dong et al. [5] (Figure 7 of their paper). They show that for a 3D-stacked device, increasing the pin count from 600 to 900 leads to approximately a 1.5X increase in packaging cost. To reduce cost, there is a push towards reducing pin count. The HMC also attempts to boost memory bandwidth by using many narrow links. Future HMC devices are projected to use as many as 512 pins for data input/output [17]. The push to reduce cost and the push to allocate more pins for data will reduce the pins available for power delivery, thus increasing *R*.

With such a future 3D-stacked memory device in mind, we carry out a detailed circuit-level IR-drop analysis. We then show that without additional current limiting constraints, the level of activity (current draw) can lead to IR-drop violations. We also characterize how IR-Drop varies based on how the activity is distributed across banks on the 3D device. This particular observation is key – it shows that *architectural policies can play a role in dictating the maximum IR-drop, and hence the performance and the packaging cost of a device.*

We show that most of the loss of performance due to the activity constraints can be made up using architectural policies that are aware of these constraints. We present a case study which places the most accessed OS pages in banks that can support the highest activity.

2. Background

2.1. DRAM Chips and 3D Stacks

A modern-day memory system is implemented with DIMMs that contain commodity 2D DRAM chips that comply with the DDR3/DDR2 standard. Each DRAM chip typically organizes its data arrays into 8 banks. Each bank can be simultaneously processing a different transaction. However, because of limitations on current draw, we are restricted to issuing no more than four row activations within a time period defined by the tFAW timing constraint. Additionally successive Activates to any chip must have a minimum spacing of tRRD. This current draw limitation is in turn defined by the charge pumps provisioned on the chip and the power delivery network that feeds these charge pumps.

Conventional DDR3 systems are facing capacity, power, and bandwidth challenges. This has motivated a shift towards 3D-stacked memory+logic devices that can simultaneously address all three challenges. Micron's Hybrid Memory Cube (HMC) is an example of such a device [9]. We therefore use the HMC as a target platform in this study. Such a device stacks 4 or 8 DRAM chips on a logic layer, thus providing high capacity in a package. It provides high internal bandwidth with many TSVs and high external bandwidth by implementing high-speed signaling circuits on the logic layer.

The HMC architecture implements 32 banks on each DRAM die. An HMC with 8 DRAM dies has 256 independent banks. These 256 banks are organized into 16 *Vaults*. A vault is a vertical pillar of data that contains 2 banks from each of the 8 dies. The banks in a vault share a single set of TSVs for data transfer.

Future generation HMCs are expected to have 8 links per cube for a total peak bandwidth of 320 GBps. To support the much higher bandwidth, the future HMC will be clearly provisioned with many more pins. Of these, 512 will be used for the data links. Of course, cost will play a significant role in the commoditization of these devices and there will be a push to lower pin count, while still supporting high activity levels within the 3D-stack. We assume that each die must continue to respect the tFAW and tRRD constraints. In addition, it must also respect per-stack current draw constraints, dictating what can and cannot be scheduled in a given cycle.

2.2. Power Delivery Networks

The current drawn by a 3D stacked memory+logic device is expected to be much higher than that of a 2D DRAM chip [9, 17]. High peak currents can have many adverse effects, such as IR-drop, power supply noise, electromigration, and higher temperatures. Of these, we focus on IR-drop in this study.

Power is delivered through pins on the package and C4 bumps on the device. Each circuit receives its supply voltage from the nearest C4 bump that is connected to the power supply. If the C4 bumps allocated for power and ground are few and far between, a non-trivial amount of the supply voltage is dissipated across the long wires that carry power to individual circuits. Based on the length of these on-chip power delivery wires, and based on the maximum allowed voltage drop that can be tolerated, a maximum current draw specification is computed.

There is a linear relationship between packaging cost and pin count [5, 7, 8]. Packaging costs have already started exceeding silicon IC fabrication costs [7]. Increasing the number of pins on the chip to reduce the IR Drop will lead to increased cost of production. In a highly cost sensitive industry like the DRAM industry, this increased packaging cost [5] can prove to be prohibitive.

IR drop analysis can be divided into static and dynamic IR drop analysis. In static analysis, static current loads are assumed to be driven by the PDN. The PDN is reduced to a resistive network and the voltage drop across this resistive network is calculated based on a given current source. Dynamic IR drop analysis takes circuit switching as well as the capacitive and inductive nature of the PDN and the package into account. We focus on static IR drop analysis in this study.

3. Methodology- Modelling IR-Drop

We first explain in detail our methodology to model IR-drop within a 3D stack. This methodology takes into account the impact of TSVs, C4 bumps, and bank activities on voltage drops within the PDN.

We use the layout of Samsung's 4-stacked 3D design as a starting point [19]. That package includes 4 2 Gb chips. We extrapolate it to an 8 GB HMC style design. The 2 Gb chip has 8 banks; the HMC design has 32 independent banks in each die. So our layout replicates each bank four times. We also consider a shrink factor of 0.8 in the linear dimension (0.64 for area) because of moving from a 50 nm technology to a 40 nm technology. The estimated chip area is $13.52 \times 16.72 mm^2$, which is about 2.3 times larger than the 2 Gb DDR3 chip at 50 nm. The final layout is shown in Fig-

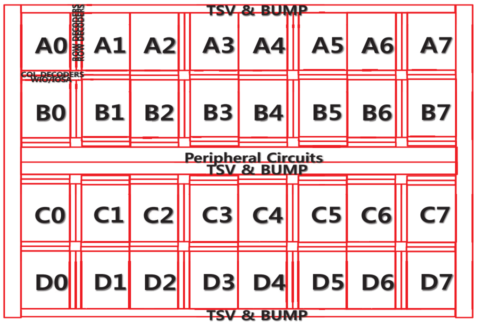

Figure 1: Layout of each DRAM die

ure 1, with the 32 banks organized as four rows of 8 banks each; the banks in each row are referred to as $A_0 - A_7$, $B_0 - B_7$, $C_0 - C_7$, and $D_0 - D_7$.

Most commodity DRAM chips assume C4 bumps along the center stripe. Kang et al. [19] show that C4 bumps and TSVs along the center can lead to a severe IR-drop problem. They overcome this problem by introducing rows of bumps/TSVs at the top and bottom of the chip (see the strips at the top and bottom of the layout in Figure 1). This is a relatively costly method to combat the problem because it requires more bumps/TSVs that impact area, yield, and packaging cost. We address the problem with low cost architectural solutions. Our model retains only the bumps/TSVs at the center stripe, but assumes 2X wider wires for just power and ground signals to reduce their resistances. The assumption is that routing tools will be able to accommodate the wider wires with a minimal impact on area. An analysis of cost for the two approaches is beyond the scope of this paper.

The layout also shows that the top of the center stripe accommodates peripheral circuits and the TSVs and bumps occupy the bottom of the center stripe. Because of this, the banks in the bottom half of the chip are closer to the power source and exhibit a lower IR-drop. As we show later, this has a small impact on the pattern of parallel activations allowed.

This work doesn't focus on IR-drop within the logic die as a logic process has other orthogonal approaches to combat IR-drop. Also, the IR Drop within the logic die has minimal impact on the IR Drop in the DRAM layers. We model the power of the logic die based on values provided for the Micron HMC [9] and assume that the power is uniformly distributed across the logic chip. Although the total current drawn by the logic die affects the IR Drop in the other dies, the distribution has little effect.

We use Synopsys HSPICE Version C-2009.09-SP1 64-BIT to model voltage drops. We model a 3D mesh of wire resistances, similar to models used in prior work [16]. The mesh includes 3 metal layers each for 9 different dies. Capacitances are not required because this is a static-IR model. We therefore only provide resistance values per wire and current draw values based on the number of activations occurring in a bank. The netlist was created using a Perl script. The grid of resistances which forms the PDN is connected to the VDD and VSS bumps on one side and to the circuit elements on the other side. Circuit elements connected to the PDN are modeled as current sources which draw constant current.

The values of resistances of metal wires, TSVs, and bumps are adopted from measured values in prior work [20, 12, 21]. These values are 0.031, 0.196, and 0.224 $\Omega/\square$ (read as Ohms per *square*, which is the unit of sheet resistance) for the three metal layers, and 0.25 Ω for C4+TSV.

External power (VDD) is supplied at 1.5 V, the same as the DDR3 specification. We could have also used the HMC's 1.2 V specification, but other parameters, such as current draw and resistances are not known. Hence, we restrict ourselves to the DDR3 model where more parameters are known. The specification requires that the voltage at the circuits (VDD-VSS, effective drain-to-source voltage) not drop below 1.425 V, i.e., we can tolerate a maximum IR-drop of 75 mV. Values for power consumed within the DRAM chip are calculated with Micron's power calculator for DDR3 chips [15].

Every DRAM operation will introduce a voltage drop in the PDN. According to Micron data sheets, the highest current is drawn by the Column Read command, followed by Column Write, and Activate/Precharge. We do not model the IR Drop caused by Refresh in this work. We simulate the IR Drop caused by Column Read, Column Write, Activate, and Precharge. Using the results from these simulations, we create constraints for each of these commands. These constraints ensure that at no time does the IR Drop go above 75 mV. These constraints are similar in spirit to today's DDR3 specification that disallows more than 4 ACTs within a tFAW time window.

We validate our Power Delivery Network model by making sure that the IR Drop does not exceed the 75 mV constraints when a 2D 8Gb, 8-bank chip, is executing 4 Activates and a Col-Rd. The 4 Activate limit is imposed by tFAW, and at any time a 2D DRAM chip can only execute a single Column Read (unlike the 3D dies used in our design). Therefore, this combination gives the highest activity that can be seen on a 2D DRAM chip. We locate the Activates and the Column Read in banks that are most susceptible to IR-Drop to model the worst case.

4. Quantifying the Impact of IR Drop

We start by performing some simple analysis on a 3D memory stack under specific sequences of memory accesses or bank activations. We observe the IR-drop in each case, focusing in particular on worst-case access patterns that cause IR-drop to exceed the 75 mV limit or best-case access patterns that yield acceptable IR-drop. We then draw upon these observations to develop a broad set of guidelines that can be used to influence the behavior of the memory controller. We also present some ideas on how the memory controller and operating system would exploit these guidelines to improve performance. We consider a layout that uses bumps and TSVs at the center stripe.

Voltage Map We first illustrate the IR-drop phenomenon with a Voltage Map. This is shown in Figure 2 (due to space constraints, we only show the Voltage Map of the top DRAM die, where the effects of non-uniform IR-drop are most starkly visible). This is an illustrative experiment, where we assume that an activate is happening in each of the 256 banks on the 3D stack. This is an unrealistic scenario and the IR-drop is unusually high because of the high current draw. The figure is only meant to highlight the banks that experience lower voltages than others and are therefore more prone to IR-drop violations.

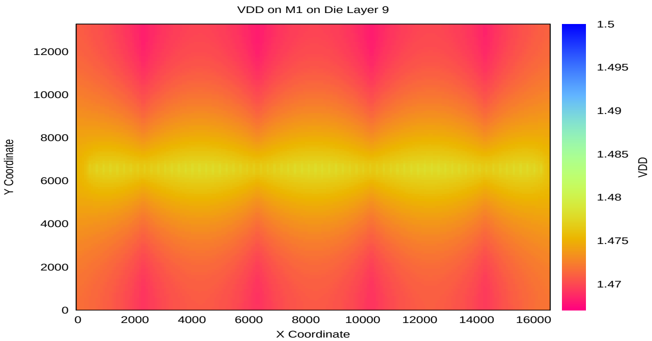

Figure 2: IR-drop phenomenon on the Top DRAM die. (Best viewed in color)

We observe that as we move up the various layers in the stack, IR-drop becomes worse since we traverse the TSVs all the way up. Note that even though TSVs are low resistance, they are relatively small in number, and are responsible for carrying significant amounts of current to the upper dies, resulting in a large IR-drop. So, in general, bottom dies are more favorable than top dies. Similarly, as we move laterally away from the row of power pins in the center of each die, IR-drop becomes progressively worse. The middle two rows of banks (banks $B_0 - B_7$ and $C_0 - C_7$, as shown in Figure 1) experience much higher voltages than the top and bottom banks (banks $A_0 - A_7$ and $D_0 - D_7$). We observe that vertical IR-drop variations between dies are much more significant than horizontal IR-drop variations within a die. This is because the TSVs are shared by all the dies, resulting in higher current densities in these TSVs. Because the bump/TSV row is in the bottom half of the center stripe, the bottom two rows of banks (*C* and *D*) are slightly closer to the power source than the top two rows of banks (*A* and *B*).

IR-Drop Regions It is clear from these maps that there are distinct regions in the chip with widely varying susceptibilities to IR-Drop. In the interest of simplicity, we divide the stack into eight IR-drop regions, as shown in Figure 3, to separate out the vulnerable regions. For example, the region A-Top refers to 32 banks in the A row in the top 4 dies, and the region C-Bottom refers to 32 banks in the C row in the bottom 4 dies. A-Top has the worst IR-drop characteristics, while C-Bottom has the best. Section 6.2 presents a case-study of a page mapping policy designed to exploit the differences in characteristics of these eight regions. For example, hot (frequently accessed) pages of an application can be placed in lower 4 dies, so that they enjoy the highest bandwidth.

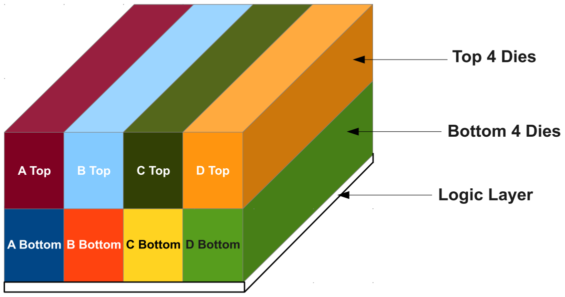

Figure 3: The eight IR-drop regions in the stack

Best and Worst Case Activation Patterns We next examine the impact on IR-drop if the 3D-stack is asked to service *N* simultaneous activation requests. Two requests are said to be *simultaneous* if the second request occurs within tRAS of the first request. For the purposes of this study, we assume that command bandwidth is not a constraint – this is a reasonable assumption to make given that an HMC part will likely have multiple channels talking to the processor, and the logic die in the HMC will likely buffer requests and can issue them as required. These *N* activates can be distributed among the 256 DRAM banks in $^{256}C_N$ ways, ruling out the possibility of an exhaustive study. Later, we develop some guidelines for the combinations of activates that tend to behave well or poorly. The high-level insight from that analysis is as follows.

- For any operation, moving to higher die layers or moving away from the center TSV strip causes higher IR Drop.
- Banks at the edge of the die experience higher IR Drops, especially banks A0, D0, A7, D7.
- Since the row decoders of the 2 banks in a vault lie right next to each other, activating both banks causes large IR Drops.
- Simultaneous operations in banks that share PDN wires (A0 and B0 for example) yield higher IR-drops.
- Lastly, having operations in the same bank in adjacent dies increases the current density in the shared power TSVs.

Based on this insight, we are able to estimate the best-case and worst-case scenarios when reading from banks.

IR-Drop Specific Timing Constraints

Since Col-Rd consumes the most current of all DRAM commands, we create detailed constraints for Reads and then model Writes, Activates and Precharges in terms of Reads. Having detailed constraint for each command would make for an infeasible memory controller.

With the assumed PDN, we measure the worst voltage in each region when that region performs the worst-case pattern of *N* reads. When one region is receiving reads, we assume that the other regions are idle. The data shows that regions A-Top, B-Top, D-Top, and C-Top can only safely handle a single read at a time. With a worst-case pattern, just two reads can lead to a voltage under 1.425 V. Thus, regardless of what else is happening on the 3D-stack, the memory controller must enforce that these regions never service more than 1 read at a time. This rule is especially restrictive because these 4 regions are the furthest from the power sources at the center stripe. For each of the other 4 regions, we can safely service as many as 4 reads even with the worst-case patterns, without violating IR-drop. Note that 4 is the upper-bound for a region because each vault can handle only one Read at a time. In other words, the four regions A-Bottom, B-Bottom, C-Bottom, and D-Bottom, are relatively unconstrained by IR-drop because of their proximity to the power source. The previous discussion assumed that all reads were being serviced by a single region and all other regions were idle. Next, we must estimate the maximum allowed activity in each region while other regions are also servicing requests. To simplify the rules for the memory controller, we first consider groups of two regions at a time. We find that A-Bottom and B-Bottom can handle 8 requests at a time; A-Top and B-Top can only handle 1 read; C-Bottom and D-Bottom can handle 8 combined requests; C-Top and D-Top can handle 1 combined request. Therefore, data placement in banks has a significant impact on request parallelism.

The process is then continued. We notice that the constraints for the bottom regions is markedly different from the constraints for the top regions. We group 4 regions together and find their worst-case allocation. We find that A-Top, B-Top, C-Top, and D-Top can together handle no more than 1 request, while A-Bottom, B-Bottom, C-Bottom, and D-Bottom can together handle 16 requests, one in each vault. When all 8 regions are grouped together, we find that no more than 8 simultaneous reads can be supported in the worst-case. The multi-region constraints assume that the rules before them have been satisfied.

Thus, a series of rules (20 rules in this case) are generated for the memory controller and a request is issued only if none of the 20 conditions are violated. If we consider and allow best-case scenarios, the number of rules would be much larger.

While the rules are expressed in terms of Reads, each Read can be substituted with 6 precharges, or two activates, or one write.

Based on the rules explained above, if a request to A-Top and B-Top were to be scheduled, the following rules would need to be satisfied: (i) schedule no more than 1 request to A-Top, (ii) schedule no more than 2 requests to B-Top, (iii) schedule no more than 1 request to A-Top and B-Top if there is a request to A-Top. In short, if A-Top is servicing a request, B-Top cannot handle a request; but if A-Top is idle, B-Top can service 2 requests. So in this case, the Read request to B-Top would have to wait until the Read in A-Top is completed.

5. Architecture Simulation Methodology

We conduct performance studies using a modified version of the USIMM simulation infrastructure [3]. While the version of USIMM used in the Memory Scheduling Championship used memory traces as inputs, we plug the USIMM framework into Simics so that the memory requests are generated by a cycle-accurate out-of-order processor model. We also modify the USIMM framework so that the communication protocol represents that of an HMC, instead of DDR3. The memory controller on the processor receives requests from the last level cache and issues them to the 3D-stacked HMC device in FCFS fashion. We also assume an FR-CFS scheduling policy on the HMC, along with closed page management. The HMC scheduler obeys various DDR3-style timing constraints [1]. Only one bank in a vault can receive a command or transfer data in any cycle. Reads and Writes to different Vaults can take place in parallel. The scheduler respects the tFAW and tRRD constraints and not issue more than four activates to a die at a time, where activates to any particular die are separated by tRRD. It also obeys the rules formulated by the IR-drop analysis in Section 4. We assume multi-programmed workloads constructed out of SPEC2k6 benchmarks. We run 8 instances of each benchmark on a processor with 8 out-of-order cores.

Processor	
ISA	UltraSPARC III ISA
CMP size and Core Freq.	8-core, 3.2 GHz
Re-Order-Buffer	64 entry
Fetch, Dispatch, Execute, and Retire	Maximum 4 per cycle
Cache Hierarchy	
L1 I-cache	32KB/2-way, private, 1-cycle
L1 D-cache	32KB/2-way, private, 1-cycle
L2 Cache Coherence Protocol	8MB/64B/8-way, shared, 10-cycle Snooping MESI
DRAM Parameters	
DRAM configuration	2 16-bit uplinks, 1 16-bit downlink @ 6.4 Gbps 32 banks/DRAM die, 8 DRAM dies/3D-stack
Total DRAM Capacity	8 GB in 1 3D-DRAM

Table 1: Simulator parameters.

6. Case Study

A naive memory controller would not allow more than one Read or two Activates at a time on the device. We therefore introduced a smarter memory controller that is IR-drop-aware, which tries to limit the impact of IR Drop vulnerable regions, on the rest of the die stack.

6.1. Controlling Starvation

Some regions in the die stack can support higher levels of activity than others. As a result, some pathological situations can arise that lead to starvation and lower throughput. Consider the following example: If there exists a Read in the top regions, the bottom regions can support at most seven reads. However, if there are no reads in the top regions, the bottom regions can support 16 reads. If the bottom regions are currently handling (say) 10 reads, the scheduler can safely issue reads to the bottom region, but not to the top region. As a result, the requests to the top region can get starved. Eventually every thread will be waiting on a pending memory request to the top region. At this time, the requests to the top region will be slowly drained (at the rate of 1 or 2 reads at a time). During this drain, there are no other pending requests to the bottom regions, so they remain idle. This leads to long stall times for every thread and memory bandwidth underutilization.

To prevent such disparities in the Quality of Service, we prioritize any request that is older than *P* times the average read latency. This pushes the scheduler to a steady state where the top regions are constantly draining 1 or 2 requests while the bottom regions are draining up to 8 requests. We empirically determined that performance is optimized when *P* has a value 1.2.

6.2. PDN-Quality Aware Page Placement

Since different regions in the 3D DRAM have different power delivery characteristics and constraints, we map the most accessed pages to the the regions that have the best power characteristics. We profile each application for 2 Million DRAM accesses to find the most accessed pages. We understand that the profile base data-placement relies on future events and cannot be implemented in reality. Sudan et al [18] describe a dynamic scheme to identify and relocate most used data. We leave a detailed implementation of a dynamic *PDN-Aware Page Placement Scheme* for future work. After profiling, the pages are sorted according to the number of accesses and then divided into eight sections. Since each of the bottom four regions in the die stack allow the maximum number of reads to happen we map the most accessed 4 sections to these regions. The rest are mapped to *C_TOP, B_TOP, D_TOP, A_TOP*, in that order.

Figure 4 shows the results for PDN Aware Page Placement. *Real PDN* shows the performance with all IR Drop constraints imposed on the 3D DRAM; *Real PDN Starv Ctrl* shows the performance when starvation control is implemented; *Real PDN Starv Ctrl PPP* shows the performance of the system which uses the data placement technique just described ; *Ideal PDN Starv Ctrl* shows the performance when the 3D DRAM has an ideal PDN (No IR Drop constraints), and tried to minimize starvation.

The performance of *Real PDN Starv Ctrl* is 3.1x better than *Real PDN*. On average, *Real PDN Starv Ctrl PPP* can improve performance by 24%, relative to the Real PDN with starvation control. The Ideal PDN design with starvation control can yield a performance improvement of 47%, so there is still room for improvement. It must be noted that even a single Read being performed in the Top regions can reduce the instantaneous memory bandwidth by 50%. Therefore to completely recover all the performance lost to IR Drop, almost all Reads and Writes need to be serviced by the Bottom regions.

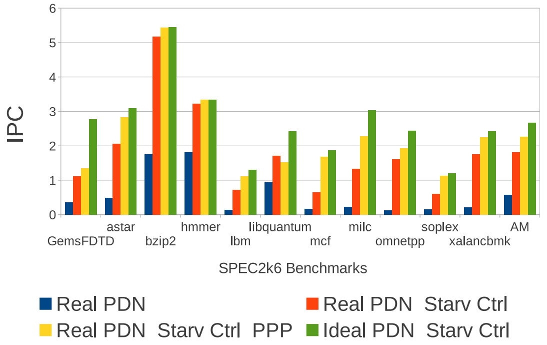

Figure 4: Impact of PDN Aware Page Placement, on IPC

6.3. Future Work

Prioritizing requests to Bottom regions to limit the detrimental effects of Top regions could overcome the effects of the upper regions and alleviate performance degradations that are seen in some benchmarks. We leave the implementation of such a scheduler to future work.

The page placement scheme proposed in this paper relies on information about the future. Page placement schemes which use other metrics to identify the most important pages are needed.

This paper proposes placing most used data into the lower banks, where as other schemes such as placing the pages with the highest peak rate of access could lead to better utilization of the higher activity threshold of the inner banks.

7. Related Work

Current Aware Memory Architectures. Recent papers [6, 11] have tried to address the high write current needs of PCM by evaluating how much current is needed by each write and not being constrained by the worst case. Kim et al. [14] address the t_{FAW} constraint in DRAM stacked over the processor by dynamically allocating Activates to every memory channel connected to a particular DRAM die. This is a part of our baseline.

Page Placement. Many prior works have influenced page placement in the memory hierarchy to handle NUMA latencies [4, 2], increase row buffer hit-rate [18], etc. Our work borrows the key ideas in these techniques and shows that they can be highly effective to address the IR-drop problem.

8. Conclusion

This paper presents the problem of IR-Drop in 3D DRAM, a problem that has to the best of our knowledge not been explored by the architecture community. We quantify the impact of this problem and present a case study which alleviates the impact of IR Drop.

Regions which have poor IR drop characteristics not only have low bandwidth, but also reduce the effective bandwidth of the entire 3D stack when servicing a request. To truly overcome the problems posed by IR-Drop, the solution must address both the spatial and temporal challenges. By reducing starvation of requests to poorly supplied regions and by intelligently placing data in the DRAM stack, this paper achieves performance that is very close to that of the unrealistic Ideal PDN. Much work remains to define the microarchitecture and OS mechanisms that can achieve this level of performance in a complexity-effective manner.

References

[1] Micron DDR3 SDRAM Part MT41J256M8, 2006.

[2] R. Chandra, S. Devine, B. Verghese, A. Gupta, and M. Rosenblum. Scheduling and Page Migration for Multiprocessor Compute Servers. In *Proceedings of ASPLOS*, 1994.

[3] N. Chatterjee, R. Balasubramonian, M. Shevgoor, S. Pugsley, A. Udipi, A. Shafiee, K. Sudan, M. Awasthi, and Z. Chishti. USIMM: the Utah SImulated Memory Module. Technical report, University of Utah, 2012. UUCS-12-002.

[4] J. Corbalan, X. Martorell, and J. Labarta. Page Migration with Dynamic Space-Sharing Scheduling Policies: The case of SGI 02000. *International Journal of Parallel Programming*, 32(4), 2004.

[5] X. Dong, J. Zhao, and Y. Xie. Fabrication Cost Analysis and Cost-Aware Design Space Exploration for 3-D ICs. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, Dec. 2010.

[6] A. Hay, K. Strauss, T. Sherwood, G. H. Loh, and D. Burger. Preventing PCM Banks from Seizing Too Much Power. In *Proceedings of MICRO-44*, 2011.

[7] ITRS. International Technology Roadmap for Semiconductors, 2007 Edition, Assembly and Packaging, 2007.

[8] B. Jacob, S. W. Ng, and D. T. Wang. *Memory Systems - Cache, DRAM, Disk*. Elsevier, 2008.

[9] J. Jeddeloh and B. Keeth. Hybrid Memory Cube – New DRAM Architecture Increases Density and Performance. In *Symposium on VLSI Technology*, 2012.

[10] JEDEC. *JESD79-3E: DDR3 SDRAM Specification*, 2009.

[11] L. Jiang, Y. Zhang, B. R. Childers, and J. Yang. FPB: Fine-grained Power Budgeting to Improve Write Throughput of Multi-level Cell Phase Change Memory. In *Proceedings of MICRO*, 2012.

[12] J.S. Kim at al. A 1.2 V 12.8 GB/s 2 Gb mobile wide-I/O DRAM with 4X128 I/Os using TSV-based stacking. In *Proceedings of ISSCC*, 2011.

[13] B. Keeth, R. J. Baker, B. Johnson, and F. Lin. *DRAM Circuit Design - Fundamental and High-Speed Topics*. IEEE Press, 2008.

[14] D. Kim, S. Yoo, S. Lee, J. H. Ahn, and H. Jung. A quantitative analysis of performance benefits of 3D die stacking on mobile and embedded SoC. In *Proceedings of DATE*, 2011.

[15] Micron Technology Inc. *Calculating Memory System Power for DDR3 - Technical Note TN-41-01*, 2007.

[16] S. R. Nassif. Power grid analysis benchmarks. In *ASP-DAC*. IEEE, 2008.

[17] G. Sandhu. DRAM Scaling and Bandwidth Challenges. In *NSF Workshop on Emerging Technologies for Interconnects (WETI)*, 2012.

[18] K. Sudan, N. Chatterjee, D. Nellans, M. Awasthi, R. Balasubramonian, and A. Davis. Micro-Pages: Increasing DRAM Efficiency with Locality-Aware Data Placement. In *Proceedings of ASPLOS-XV*, 2010.

[19] U. Kang at al. 8 Gb 3-D DDR3 DRAM Using Through-Silicon-Via Technology. *Solid-State Circuits, IEEE Journal of*, Jan. 2010.

[20] Q. Wu and T. Zhang. Design Techniques to Facilitate Processor Power Delivery in 3-D Processor-DRAM Integrated Systems. *VLSI Systems, IEEE Transactions on*, Sept. 2011.

[21] H. Y. You, Y. Hwang, J. W. Pyun, Y. G. Ryu, and H. S. Kim. Chip Package Interaction in Micro Bump and TSV Structure. In *Proceedings of 62nd IEEE ECTC*, 2012.