Quantifying the Relationship between the Power Delivery Network and Architectural Policies in a 3D-Stacked Memory Device

Manjunath Shevgoor
University Of Utah
shevgoor@cs.utah.edu

Jung-Sik Kim
Memory Division, Samsung Electronics
jungsik1.kim@samsung.com

Niladrish Chatterjee
University Of Utah
nil@cs.utah.edu

Rajeev Balasubramonian
University Of Utah
rajeev@cs.utah.edu

Al Davis
University Of Utah
ald@cs.utah.edu

Aniruddha N. Udipi
ARM R&D
ani.udipi@arm.com

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 wires and pins used for power delivery would have to carry larger currents over longer distances. This results in an "IR-drop" problem, where some of the voltage is dropped across the long resistive wires making up the power delivery network, and the eventual circuits experience fluctuations in their supplied voltage. The same problem also manifests if the pin count is the same, but the current draw is higher. IR-drop can be especially problematic in 3D DRAM devices because (i) low cost (few pins and TSVs) is a high priority, (ii) 3D-stacking increases current draw within the package without providing proportionate room for more pins, and (iii) TSVs add to the resistance of the power delivery network.

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. Finally, by adding a dynamic page placement policy, performance is within 1.2x of the unrealistic ideal PDN. We thus show that architectural polices can help mitigate the limitations imposed by a cost constrained design.

MICRO-46 December 7-11, 2013, Davis, CA, USA
Copyright 2013 ACM 978-1-4503-2638-4/13/12 ...$15.00
http://dx.doi.org/10.1145/2540708.2540726 .

Categories and Subject Descriptors

B.3.2 [**Memory Structures**]: Design Styles Primary Memory

General Terms

Design, Performance

Keywords

Power delivery network; 3D-stacked memory devices; IR-drop; Memory scheduling; Data placement

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. Traditionally, Power Delivery Networks (PDNs) in DRAMs have not received much attention, but with the move towards high performance and low-voltage DRAM, managing power supply noise becomes increasingly critical for correctness and performance [28].

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; by Ohm's Law, 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 [23], 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 relatively short distances. 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 window $tFAW$. Other examples also exist, such as the timing window $tRRD$ [19], which imposes a minimum gap between consecutive DRAM Activates[1].

Technology and market forces are raising the values of I and R. First, the onset of 3D-stacking will increase the current draw I per package. Micron has announced the imminent release of its 3D-stacked memory+logic device, the Hybrid Memory Cube (HMC). There will likely be other similar products, including some that only stack multiple DRAM dies [47]. Second, 3D stacks introduce a vertical resistive component (e.g., through silicon vias or TSVs) within the PDN, thus increasing R. Third, 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. [11] (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 pressure to reduce pin count. Similarly, to improve data bandwidth, there is pressure to allocate more pins for data signals. Both will reduce the pins available for power delivery, thus potentially increasing R.

With such future 3D-stacked memory devices in mind, we carry out a detailed circuit-level static IR-drop analysis. We then show that without additional current limiting constraints, the level of activity (current draw) can lead to IR-drop violations. The activity on the device must be throttled to avoid these IR-drop violations. We make the key observation that IR drop not only depends on the number of banks that are servicing requests, but also on the location of these banks and the DRAM commands being executed. We characterize how IR-Drop varies with activity distribution across banks on the 3D device. Thus, *architectural policies can play a role in dictating the maximum IR-drop, and hence the performance and the packaging cost of a device.* These observations lead us to introduce a number of IR-drop-aware rules within the memory controller. However, this basic design yields performance that is 4.7x lower than a memory device with an unrealistic over-provisioned PDN that never has IR-drop violations.

We show that most of this steep performance loss can be recovered with smarter architectural policies implemented in the memory scheduler and in the OS page manager. The memory scheduler is designed to better handle frequent starvation scenarios. We also introduce a dynamic page migration scheme that identifies critical pages and places them in the regions with the highest immunity to IR-drop. With these policies in place, the new design has performance that is only 1.2x lower than the unrealistic ideal PDN.

A few caveats are worth noting: (i) There are potentially many ways to tackle the IR-drop problem (more pins, more TSVs, fatter wires/TSVs, new materials, voltage regulators, higher supply voltage, in-package decaps, etc.) and the magnitude of the problem in future technologies is yet unclear. The goal of this paper is to explore an architectural approach to the problem. If successful, this approach may obviate the need for more expensive approaches, or it may be one of many solutions that are deployed to handle voltage problems. (ii) There are many possible sources of voltage noise and this work only focuses on analyzing static IR-drop. Note that other voltage noise sources may eat into the votage margins, resulting in even lower tolerance for static IR-drop. A holistic architectural solution that can cope with several voltage noise sources is left as future work. This paper therefore represents an initial stab at a complex problem.

2. BACKGROUND

2.1 2D DDR3 Memory Systems

A modern-day memory system is implemented with DIMMs that contain commodity 2D DRAM chips that comply with the DDR3 or DDR2 standard. The processor socket typically has up to four memory controllers that are used to drive four DDR3 memory channels. These channels are wide (64 or 72 bits for data) and run at frequencies that are roughly half that of the processor frequency. The channel is essentially a bus that connects to multiple DIMMs. If more DIMMs are placed on the channel, the higher load forces the channel to operate at a lower frequency. This leads to a capacity-bandwidth trade-off. Some recent high-capacity systems have tried to provide high capacity and high bandwidth by introducing buffer chips on the board [7]. In such systems, the processor memory controllers drive narrow high-speed buses that each connect to a single buffer chip. This buffer chip then uses wide and slow DDR3 channels to connect to multiple DIMMs [37]. The buffer-on-board solution does incur a steep power penalty.

Each DDR3 DRAM chip typically organizes its data arrays into 8 banks. Each bank can be concurrently processing a different memory transaction. To access data in a bank, the memory controller first issues a row activate (ACT) command that brings data in a row of cells to the row buffer. Individual cache lines in the row are read and written with COL-RD and COL-WR commands. Before accessing a different row, the bitlines are equalized with a precharge (PRE) command.

Even though the banks can all be busy at the same time, because of limitations on current draw, the memory controller is restricted to issuing no more than four row activations within a time period defined by the tFAW timing constraint. Further, the tRRD timing parameter enforces a gap between activations to different banks. 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.

2.2 3D-Stacked Memory

3D stacking is being widely employed within prototype memory devices [42, 38, 46, 12]. Of these devices, we use Micron's Hybrid Memory Cube (HMC) as an evaluation platform because it will soon be commercially available and several design details are already available in the public domain [38, 43, 20]. The ideas and analyses in this paper will apply to almost any 3D-stacked memory device. In fact,

[1] Some of these constraints are influenced not just by the PDN, but also by the charge pumps. We expand on this in Section 4.

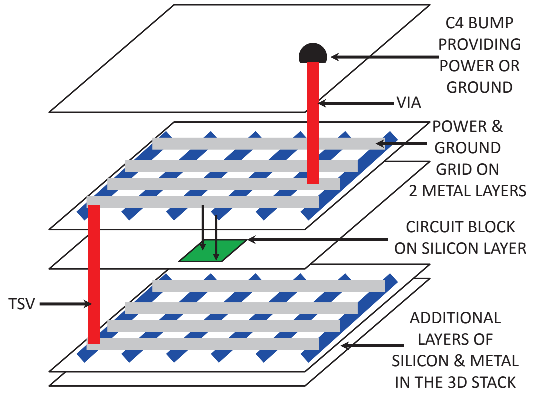

C4 BUMP PROVIDING POWER OR GROUND
VIA
POWER & GROUND GRID ON 2 METAL LAYERS
CIRCUIT BLOCK ON SILICON LAYER
TSV
ADDITIONAL LAYERS OF SILICON & METAL IN THE 3D STACK

Figure 1: Illustrative cross-section of a portion of the power delivery network. VDD and VSS are supplied through C4 bumps and fed to the circuit block with vias/TSVs and horizontal power/ground grids on metal layers.

the proposed approach is a better fit for cost-constrained 3D-stacked DRAM devices that do not include a logic layer. Most of our analysis is therefore focused on the IR-drop caused by the current drawn by the DRAM stack.

The HMC stacks 4 or 8 DRAM chips on a logic layer, thus providing high capacity in a package. It replaces several on-board interconnects with power-efficient through-silicon vias (TSVs). 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. An entire cache line can be accessed from a single bank in a single HMC, similar to single subarray access [48] for low energy and limited overfetch.

The first generation HMC uses 1866 total TSVs at 60μm pitch and 256 signal pins [20]. The external links are driven by high-frequency SerDes circuits on the logic chip. The HMC is a high-power, high-bandwidth, and high-cost design point. 3D-stacked DRAM packages that exclude a logic layer and high-speed SerDes links will likely be constructed with much fewer TSVs and external pins for power and ground.

Like most memory products, there will be a push to lower cost by reducing TSVs and pin counts, while still supporting high activity levels within 3D-stacked DRAM. The power delivery network for the package will dictate various timing constraints (similar to tFAW and tRRD) that will throttle the peak current drawn by the package.

2.3 Power Delivery Networks

The aggregate current drawn by a 3D stacked memory device is expected to be much higher than that of a 2D DRAM chip [20, 43]. High peak currents can have many adverse effects, such as static IR-drop, dynamic IR-drop, power supply noise, and higher temperatures. Of these, we focus on static IR-drop in this paper.

Power is delivered through pins on the package and C4 bumps on the device. A number of TSVs are used to carry power/ground signals from the C4 bumps to each chip on the stack. The metal layers for the chip implement a horizontal grid of power/ground wires that carry these signals to each circuit block. Figure 1 shows an example PDN, illustrating the entire path from bump to destination circuit. A portion of the supply voltage is dropped across the PDN – this *IR-drop* is a function of the effective resistance of the PDN R and the current I that it carries. If the C4 bumps and TSVs allocated for power and ground are few and far between, the lengths of the on-die resistive wires is longer, increasing the value of R. This increases the voltage drop across the PDN. Based on the length of these on-chip power delivery wires, and based on the maximum voltage drop that can be tolerated, a maximum current draw specification is computed. The memory controller is then provided various timing parameters that prevent the current draw from exceeding this maximum.

Zhang et al. [51] show that IR-drop in processors will increase three-fold as we move from 45 nm to 16 nm technology. This trend is driven by various factors: (i) non-increase in the number of C4 bumps, (ii) slightly lower supply voltages in future generations, (iii) narrower wires with higher resistances, and (iv) higher current densities.

A 3D PDN is inherently more resistive than a 2D PDN because of the presence of power and ground TSVs. A 3D package also draws higher aggregate currents than a 2D package. Khan et al. [29] report that when moving from 2D to 3D ICs, the IR-drop is greater than the Ldi/dt voltage droop. Thus, there are many indications that the IR-drop problem will be significant in future 3D devices.

Some prior work [17, 47] has attempted to design a better TSV network to reduce IR-drop. However, these typically introduce more TSVs, which impacts cost [27], while not eliminating the IR-drop problem. Voltage regulators [31] can also help alleviate the IR-drop problem, but may not be viable because of their negative impact on density and cost.

Assuming that the IR-drop can be tolerated, there is a strong motivation to reduce the number of pins, C4 bumps, and TSVs allocated for power/ground. There is a linear relationship between packaging cost and pin/C4 count [11, 18, 19]. Dong et al. [11] shows that for a 3D-stacked device, increasing the pin count from 600 to 900 leads to approximately a 1.5X increase in packaging cost. Packaging costs have already started exceeding silicon IC fabrication costs [18]. Routing many C4 bumps through the Redistribution Layer (RDL) inside the package incurs additional cost. Increased package routing density can lead to decreased packaging yield and lead to increased packaging cost [45]. This steers the cost-sensitive DRAM industry towards lower pin/C4 counts. Similarly, a high TSV count also negatively impacts area, routing, yield, and cost.

IR-drop analysis in the PDN can be broken down into static and dynamic components. In static IR-drop 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. When dynamic current consumption is simulated, PDN noise such as ground and power bounce can be analyzed. In 2D DRAM chips, dynamic IR-drop is alleviated with decoupling capacitors (Decaps) [19]. While a 3D package can provision more Decaps than a 2D package, it is not clear how dynamic IR-drop will scale in future technologies. As a first step, this paper focuses on the static IR-drop problem.

3. METHODOLOGY

We first explain in detail our methodology to simulate IR-drop within an HMC-style 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 [47]. That package includes 4 2 Gb chips. We extrapolate it to an 8 Gb design by quadrupling the number of banks. 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 (Frugal) is shown in Figure 2a. Unlike a 2D DRAM floor plan, which only has one row of banks on either side of the C4 bumps, the 32 bank floor plan will have 2 rows of banks on each side of the C4 bumps. The 32 banks are 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 low-cost commodity DRAM chips assume C4 bumps along the center stripe. Kang et al. [47] 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 Expensive layout in Figure 2b). This is a relatively costly method to combat the problem because it requires more bumps/TSVs that impact area, yield, and packaging cost. We therefore restrict ourselves to the Frugal layout in this study and attempt to address the IR-drop problem with architectural solutions.

The power grid specifications used in our model are adopted from Wu et al. [49]. Due to the back to back arrangement of banks, we assume 2X wider wires for power and ground signals to reduce their resistances. We increase the wire width from 2 μm to 4 μm, while keeping the pitch of the supply wires fixed at 12 μm. The assumption is that the pitch of the supply wires is wide enough for signal wires and that routing tools may be able to accommodate the wider wires with a minimal impact on area.

In our evaluations, we model 536 C4 bumps and 536 TSVs for power and ground. The C4 bumps have a pitch of 120 μm. The TSVs in our design are placed with a pitch of 40 μm [26]. We also assume an additional 256 signal C4 bumps and 992 signal TSVs. Similar to the floorplan used by Kang et al. [47], the layout assumes that the top of the center stripe accommodates peripheral circuits, while the bottom of the center stripe accommodates TSVs and bumps. 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 level of activity allowed in each bank.

We also confirmed that our TSV count is large enough to provide the necessary current in the DRAM stacks. Assuming a migration density threshold of 7400 A/cm^2 [34], 50% derate, 5 W requirement in the DRAM stacks, and 25 μm microbump diameter for TSVs, we would need a minimum of 229 power and 229 ground TSVs. If we make pessimistic assumptions regarding the current-carrying capability of lead-free solder or the size of the bumps, it is possible to hit the current wall before a possible IR-drop violation, i.e., for acceptable operation, we would provision enough TSVs that static IR-drop would not be a problem.

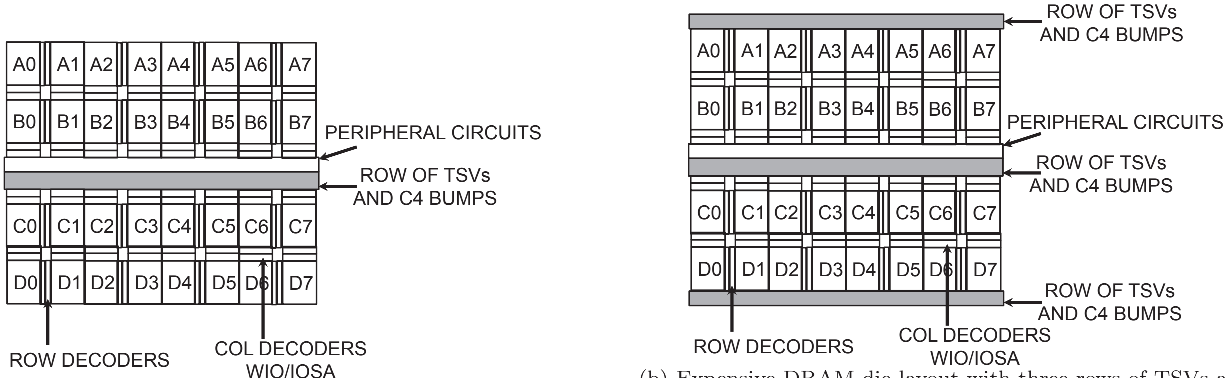
This work doesn't focus on IR-drop within the logic die as a logic process has other orthogonal approaches to combat IR-drop (more metal layers for example). The logic die also doesn't suffer from IR-drop across TSVs. Also, a logic chip will typically be absent in a cost-constrained memory device. We model the power of the logic die based on values provided for the Micron HMC [20] and assume that the power is uniformly distributed across the logic chip. We note that the assumptions for the logic chip have a small impact on the IR-drop within the DRAM chips. This is because the DRAM die and the logic die only share the resistance of a small segment of C4 bumps, so a high current draw in the logic chip only exposes the DRAM chip to a negligible amount of IR-drop.

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 [36]. 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 activity in a bank. The netlist was created using a Perl script. The grid of resistance which forms the PDN is connected to the VDD and VSS bumps on one side and is connected to circuit elements on the other side. Circuit elements connected to the PDN are modeled as current sources which draw a fixed amount of current. The values of resistances of metal wires, TSVs, and bumps are adopted from measured values in prior work [49, 26, 50]. 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 current consumed within the DRAM chip are from Micron's data sheets [1]. Note that regardless of the assumed supply voltage, DRAM arrays will have small margins for IR-drop. This is because DRAM arrays are designed to operate at as high a supply voltage as possible. If DRAM arrays were designed to operate at lower supply voltages, they would suffer from higher leakage currents and high refresh overheads (another emerging bottleneck in future DRAM cells).

Every DRAM operation will introduce a voltage drop in the PDN. According to Micron data sheets, the highest current is drawn by the COL-RD command, followed by COL-WR, and ACT/PRE. This is discussed in more detail in Section 4. 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.

Because modern 2D devices do not allow other commands



(a) Frugal DRAM die layout with one row of TSVs and bumps.

(b) Expensive DRAM die layout with three rows of TSVs and bumps.

Figure 2: Two possible layouts for DRAM dies. Both layouts have four rows of eight banks each.

to be issued during a refresh cycle, we do not model IR-drop caused by refresh. Future 3D devices may allow activities in some banks while other banks are refreshing. Such a model would require more sophisticated IR-drop analyses and memory controllers.

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 AND MANAGING IR-DROP

We start by performing an analysis on a 3D memory stack under specific sequences of bank operations. 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 show how the memory controller and operating system would exploit these guidelines to improve performance. The methodology for constructing the PDN is validated by first creating the PDN for an 8-bank, 8Gb 2D DRAM die. We see that in the 2D case, the PDN easily accommodates 4 Acts in parallel, as well as a Column Read.

Multiple factors make IR-drop worse in the 32-bank 8-die case. The TSVs introduce a new source of IR-drop. The lateral wiring on each die also sees a higher current. This is because there are four rows of banks and multiple banks (e.g., A0 and B0) receive their power from the same set of lateral wires. In the 8-bank 2D case, every bank has its dedicated set of wires within the power grid. To alleviate this problem in the 32-bank 8-die design, the power and ground wires have to be made 2x wider.

4.1 Voltage Map

We first illustrate the basic IR-drop phenomenon with a voltage map across all eight DRAM dies (die layers 2-9). This is shown in Figure 3. In this experiment, we assume that activates are happening in all 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 illustrate the banks that experience lower voltages than others, and are therefore more prone to IR-drop violations.

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 larger 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. 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), and hence experience lower IR-drop.

4.2 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 4, 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. This motivates the design of page placement policies that can exploit this inherent difference between banks. For example, the most accessed pages can be placed in banks capable of higher activity levels.

4.3 Best and Worst Case Operations

Next, we examine the impact on IR-drop if the 3D-stack is asked to service N simultaneous operations; an operation can be any of Read, Write, Activate or Precharge. 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 communicating to the processor and a request buffer. These N operations can be distributed among the

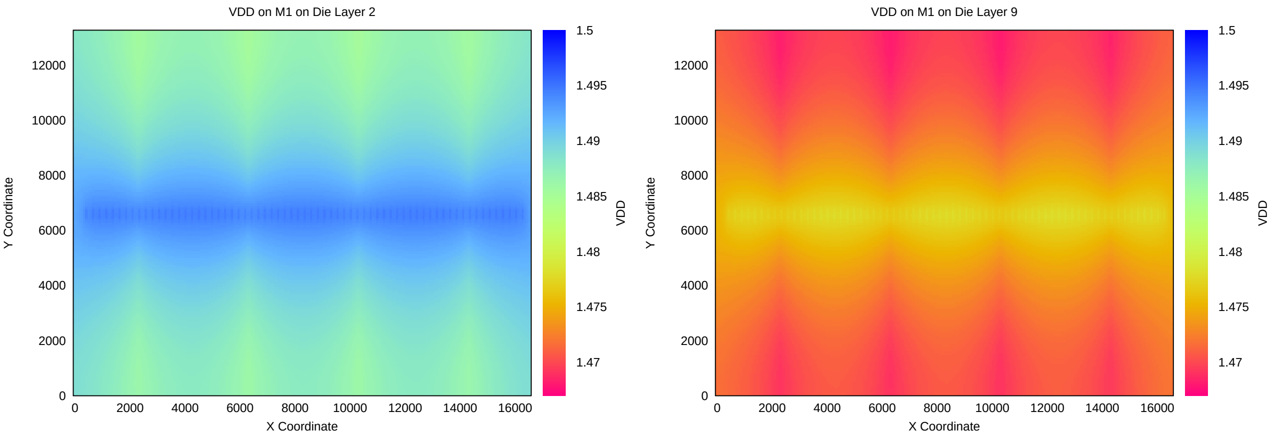


Figure 3: Basic IR-drop phenomenon on the top and bottom dies when all Banks are Activating (Best viewed in color). The vertical structures with high IR-drop are the Row Decoders.

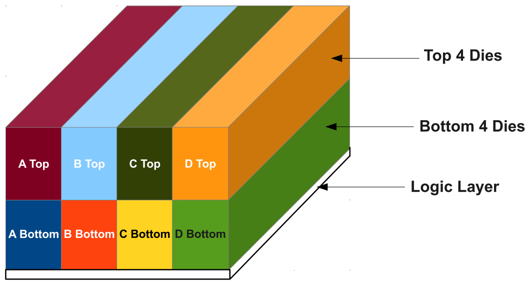


Figure 4: The eight IR-drop regions in the stack

256 DRAM banks in $\binom{256}{N}$ ways, ruling out the possibility of an exhaustive study. Later, in the results section, we develop some guidelines for the combinations of operations 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, because of the longer distances that the current needs to travel.
- Banks at the edge of the die experience higher IR-drops, especially banks A0, D0, A7, D7. This is because those banks are not supplied from all sides.
- Since the row decoders of the 2 banks in a vault lie right next to each other, activating both banks causes large IR-drops. Row decoders are placed adjacent to each other so that some control circuits, DC generators, and decoupling caps can be shared.
- 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.

All the patterns mentioned here lead to increased current density in either the wires or the TSVs, leading to possible IR-drop violations.

Based on this insight, we are able to estimate the best-case and worst-case scenarios when activating banks. For example, if asked to do 8 activates in the B-top region, minimum IR-drop is experienced by placing 4 activates in the B_0 vault (one in each of the 4 top dies) and 4 activates in the B_2 vault (one in each of the 4 top dies). The maximum IR-drop is experienced when placing 4 activates in the top die at banks B_0, B_1, B_2, and B_3, and 4 more activates in the same banks directly below. In all of our allocations, we ensure that a single die is never asked to perform more than 4 simultaneous activates because, similar to the tFAW constraint, the charge pumps on a single die are only provisioned to handle at most 4 simultaneous activates.

4.4 Col-Rd/Col-Wr Commands

In 2D DRAM chips, violations are either caused when the charge pumps are depleted or when IR-drop is high. In 2D DRAM chips, the charge pump violations typically happen before IR-drop violations. Hence, a larger focus is placed on Activates. Activates consume more charge and dissipate higher average power than Column Rd/Wr. Activates occur for the duration of tRAS, which is much longer than the duration for a Column Rd/Wr (tDATA_TRANS). This is why timing constraints ($tFAW$, $tRRD$) in 2D DRAM chips refer to the rate at which Activates can be performed.

For the reasons mentioned earlier, IR-drop is much more severe in 3D stacks and IR-drop violations are encountered before charge pump depletions. IR-drop is influenced more by peak power than average power. Column Rd/Wr instantaneous current (IDD4R/IDD4W) is 3x the instantaneous current for Activates (IDD0). As a result, the focus must shift from Activates to Column Rd/Wr.

The following is a brief explanation for why Column Rd/Wr has higher peak power than an Activate. The data sensing during an Activate is done by the Bit Line Sense Amps (BLSA, referred to as Local and Global sense amps in [32]). During a Column Rd, the data has to be moved from the BLSAs, which are adjacent to the arrays, to the IO-Sense Amps (IOSA), which are in the center stripe. Also, the data transfer needs to happen at the speed of the channel (Vault) clock, which is in the range of Gigahertz. These factors make IDD4R very high.

While it is possible to come up with rules for every possible combination of Read, Write, Act, and Pre, such a list for the 256 banks in the die stack would make the controller intractably complex. In order to simplify the rules for the memory controller, we define the impact of each operation in terms of the impact of a Column Read. For example, we define that 2 Activates correspond to one Column Read. This means that the worst IR-drop caused by 2 Activates cannot be greater than the least IR-drop caused by a Column Read. Even though IDD4W is less than IDD4R, we find that 2 banks cannot perform writes in parallel, without exceeding the IR-drop caused by a Column Read. So one Column Write is deemed equivalent to one Column Read. Finally, six Precharges are equivalent to a single Column Read.

4.5 IR-drop Specific Timing Constraints

To keep the memory controller simple, it must only encode the worst-case guideline. For example, in a given region, in the best case, IR-drop may not be violated with 8 reads. But in the worst case, IR-drop may be violated with just 5 reads. To reduce complexity, we may want to enforce the rule that the region can safely accept only 5 reads. To accept any more reads, the memory controller would have to maintain a very large table of safe read combinations. Hence, for each region, we do a number of Spice simulations to find the worst-case read combinations and the minimum number of reads that lead to an IR-drop violation. With the assumed PDN, we simulate the 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, and D-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. B-Top and C-Top can service up to 2 Reads at any time. 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 there are only 4 vaults available per region. In other words, the four regions A-Bot, B-Bot, C-Bot, and D-Bot, 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 are 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. These rules are summarized in Table 1. If we consider and allow best-case scenarios, the number of rules would be much larger.

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.

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

Note that a conventional simple memory controller is unaware of IR-drop and regions. Such a memory controller would disallow 2 parallel reads because in the worst case, both reads may be destined for A-Top, thus causing an IR-drop violation. Such a naive baseline will have very poor performance and is not considered further in this study. Our baseline model adopts the novel constraints we introduce in Table 1. In the next section, we introduce additional mechanisms to improve the memory device's throughput.

4.6 Future DRAM Generations

A DDR3 DRAM chip cannot activate more than 4 banks in a chip within a time period specified by t_{FAW}. The reason for this is that the Wordlines on the chip need to be driven by a voltage greater than the supply voltage of the chip. By over-driving the Access Transistors on the word lines, the sense-amps are able to see the true voltage that is present on cell capacitors. This increased voltage (V_{PP}) is provided by charge pumps which are present on the chip. Performing successive activates depletes these charge pumps, following which they need time to recover. Doing no more than 4 activates within a window of t_{FAW} ns ensures that the output of the charge pumps stays within the required voltage.

Future generations of DRAM like DDR4 have V_{PP} (2.5 V) [22] supplied externally, hence replacing internal charge pumps [35]. By doing this, an 8Gb DDR4 device is able to lower its t_{FAW} to as low as 20 ns [9], with the eventual goal of eliminating the t_{FAW} constraint [21] altogether.

As described in Section 4.3, IR-drop worsens when the activity on the chip increases. The DRAM design described in this paper tries to stay faithful to today's DDR3 design as much as possible. We conservatively assume that just like DDR3 DRAM, the dies on the 3D stack will also be subjected to the t_{FAW} constraint. If the t_{FAW} constraint is indeed reduced or eliminated in the future, the IR-drop problem reported in this paper becomes even greater because of the increased activity.

Constraint Type	**Description**	**Constraint for Region(s)**	**Num. Parallel Column Read units allowed**
Single Region Constraints	Reads taking place only in that One region	A_TOP	1
		B_TOP	2
		C_TOP	2
		D_TOP	1
		A_BOT	4
		B_BOT	4
		C_BOT	4
		D_BOT	4
Two Region Constraints (Reads happening only in these two regions)	At least one in A_TOP	A_TOP and B_TOP	1
	No Reads in A_TOP	A_TOP and B_TOP	2
	At least one Read in A_BOT	A_BOT and B_BOT	8
	No Reads in A_BOT	A_BOT and B_BOT	8
	At least one Read in D_TOP	C_TOP and D_TOP	1
	No Reads in D_BOT	C_TOP and D_TOP	2
	At least one Read in D_TOP	C_BOT and D_BOT	8
	No Reads in D_BOT	C_BOT and D_BOT	8
Four Region Constraints (Reads happening in) only these four regions)	No Reads in Bottom Regions	A_TOP, B_TOP, C_TOP, D_TOP	1
	No Reads in Top Regions	A_BOT, B_BOT, C_BOT, D_BOT	16
Die-Stack wide Constraint	At least one Read in Top Regions	All Regions	8
	Reads only in Bottom Regions	All Regions	16

Table 1: Max. parallel Column Reads allowed in each Region

5. OVERCOMING THE CONSTRAINTS IMPOSED BY IR-DROP

In the previous section, we showed that IR-drop imposes new and severe constraints on device activity. 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 and obeys the 20 rules we introduce in Table 1 to support higher activity levels on the memory device. However, even this smarter memory controller is restrictive and falls well short of the performance of an unconstrained memory device. This section introduces additional optimizations to bridge this gap.

5.1 Handling Throughput Oscillations

According to the rules in Table 1, some regions 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 that is based on the rules defined in Table 1.

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.

Instead, it is more effective to be in a steady state where the top regions are dealing with 1 request, while the bottom regions are dealing with 8 requests. While the threads waiting for the top region are stalled briefly, other threads continue to make progress in the meantime. This yields a higher aggregate throughput than the default design that frequently oscillates between high and low throughput phases.

To prevent such oscillations, 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.

5.2 Smart Page Placement

Some regions can drain requests at a faster rate than others and therefore yield much lower queuing delays and memory access latencies. To optimize throughput, most memory requests should be steered towards these regions that are more immune to IR-drop violations. This can be achieved with OS policies that carefully select the regions where pages are placed.

To estimate the potential for improvement, we first implement a profile-based oracular scheme. Our benchmarks are profiled for 2 million DRAM accesses. The pages are sorted according to access count and split into eight sections. Starting with the most accessed section, they are mapped to A_Bot,B_Bot,C_Bot, D_Bot, C_Top, B_Top, D_Top, A_Top, in that order. The benchmarks are simulated again with these page-to-region assignments.

In a realistic implementation, page activities must be tracked at the memory controller or on the logic die of the 3D stack. Page activities from the recent past must dictate page migrations and page placements in the future. We assume that the base layer of the 3D stack keeps track of all pages touched in the last epoch (a pre-defined time interval). For these touched pages, we track the average queuing delay for the blocks in that page. Pages with the highest queuing delays are moved from the top regions to the bottom regions. Note that access count in the last epoch is not an effective metric. If an application's critical pages are placed in the top region, the core will be starved and it will register few page accesses in any epoch. This is why we use queuing delay to identify pages that are introducing the most stall cycles. Any page that has an average queuing delay greater than Hot_Page_Migration_Threshold*Average_queuing_Delay is migrated to the Bottom regions.

The metric for demotion of cool pages to the Top regions is the number of page accesses in the last epoch. Any page that has less than *Cold_Page_Migration_Threshold* number of page accesses in the last epoch is migrated to the Top regions.

Pages which are not candidates for migration to Top or Bottom regions are not moved. At the end of every epoch, the DRAM stack is unresponsive to the CPU for *Migration_Penalty* number of cycles, similar to a refresh cycle. All migrations happen during this window. The Hot-Page and Cold-Page migration thresholds are dynamically modified such that all the migrations can happen within *Migration_Penalty* number of cycles.

Hot_Page_Migration_Threshold is initialized (HMT) to 1.2 and *Cold_Page_Migration_Threshold* (CMT) to 0, such that during the initial epoch, there are many page migrations. If the number of migrations is more than can be accommodated in the penalty window, then HMT is incremented by 0.05 and CMT is decremented by 1. A negative value for CMT means that no pages are demoted. However, if the number of migrations are less than can be handled in the penalty window, then HMT is decremented by 0.05 and CMT is incremented by 1.

We assume an epoch length of 15 M cycles. After every epoch, the DRAM system incurs a *Migration-Penalty* of 10K cycles (an overhead of less than 0.1%). If a shorter epoch is used, then a large portion of the pages in the Bottom area go untouched, potentially yielding unwanted migrations to Top regions. We observe that after an initial warm up period, the number of migrations per epoch stabilizes and is easily accommodated in the migration penalty window. We assume that a single page migration takes 4184 cpu cycles, and that 8 Rd/Wr can be happening in parallel. This migration penalty is based on the number of cache lines in a page and the DRAM timing parameters tRCD and tCAS (tRCD + tCAS + tDATA_TRANS * Num_lines). This migration is efficient because it only engages the high-bandwidth TSVs and does not engage off-chip links. In terms of array access overhead, page migration increases the average number of memory accesses by 0.6%

6. ARCHITECTURE SIMULATION METHODOLOGY

We conduct performance studies using a modified version of the USIMM simulation infrastructure [5]. 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-FCFS scheduling policy on the HMC, along with closed page management, where a DRAM row is kept open till there are no more requests to that row in the read queue. The HMC scheduler obeys various DDR3-style timing constraints, summarized in Table 2. The TSVs in a Vault are shared by all the banks in the Vault, only one bank in a vault can perform a read or a write in any cycle. Reads and Writes to different Vaults can take place in parallel. The scheduler must not issue more than four activates to a die at a time. It also obeys the rules formulated by the IR-drop analysis in Section 4. We use multi-programmed workloads constructed out of SPEC2k6 benchmarks. We run 8 instances of each benchmark on a processor with 8 out-of-order cores. All relevant simulation parameters are summarized in Table 2.

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
DRAM Timing Parameters	t_{RC} = 48.75 ns t_{RCD} = 13.75 ns t_{RAS} = 35 ns t_{FAW} = 50 ns t_{WTR} = 7.5 ns t_{RP} = 13.75 ns

Table 2: Simulator and DRAM parameters [1].

7. RESULTS

As an upper bound, we present performance results when the DRAM stack is constructed with an Ideal PDN. With the Ideal PDN, t_{FAW}, and t_{RRD} are the only constraints that limit activity in the die stack. No IR-drop based Read constraints are imposed on the system with Ideal PDN.

Figure 5 shows the impact of starvation and throughput oscillations on the IPC. Starvation becomes severe when different regions of the 3D stack have different permissible maximum activity. The bar labeled *RealPDN* shows the performance of the system where the DRAM stack has a realistic PDN. *RealPDN Starv Ctrl* and *IdealPDN Starv Ctrl* show the performance of the Real PDN and Ideal PDN with the starvation control mechanism described in Section 5.1. Compared to the *IdealPDN Starv Ctrl* scheme (the unrealistic upper bound), the *RealPDN* scheme is 4.6x worse. By adding starvation control, we see that in the Real PDN case, the performance improves by 213%, while with the ideal PDN, the performance improves by 10.2%. By identifying

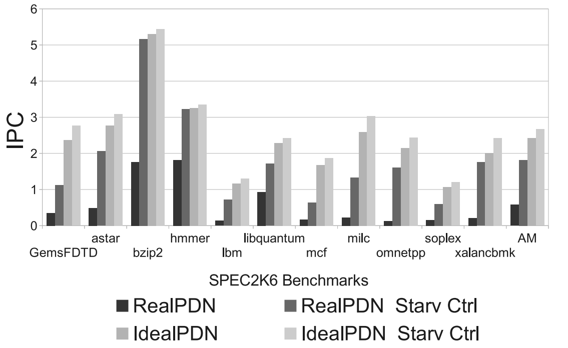


Figure 5: Effect of starvation on performance

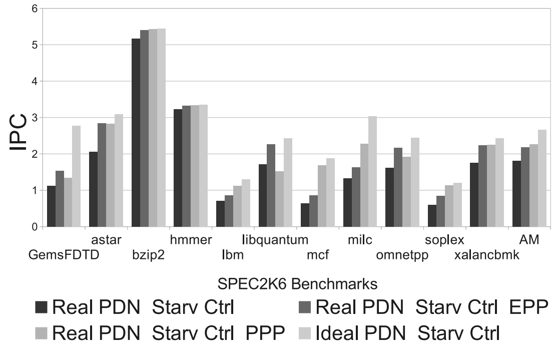


Figure 7: Effect of Page Placement schemes on performance

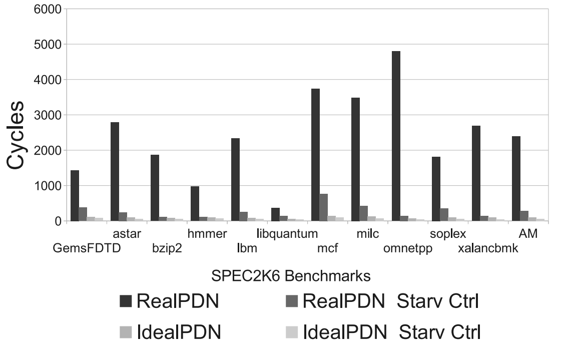


Figure 6: Effect of starvation on Read-Queue

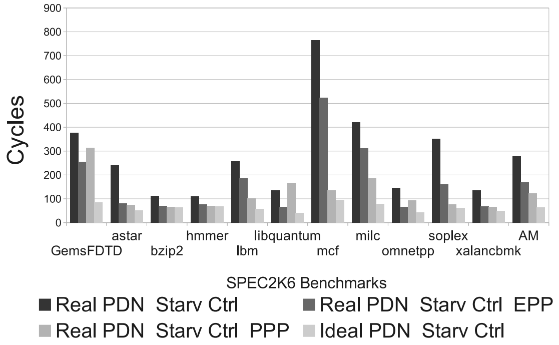


Figure 8: Effect of Page Placement on Read-Queue

and prioritizing requests to pages that are suffering starvation, the scheduler is able to prevent the suffering thread (and eventually all other threads) from stalling altogether.

Figure 6 shows the increase in the average read queue latency of a request when there is no starvation control. There is an 8.6X increase in Average Read Queue Latency of the Real PDN, when starvation is not addressed. With an Ideal PDN, the Average Read Queue Latency increases by 51.1%.

Figure 7 shows the performance improvement with Profiled Page Placement (PPP) and with Epoch based Page Placement (EPP). *RealPDN StarvCtrl PPP* represents the system with a Real PDN, with starvation control, and with PPP. *RealPDN StarvCtrl EPP* represents the system with a real PDN, with starvation control, and with EPP. On average, PPP can improve performance by 24%, while the EPP scheme improves performance by 20%, 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. We do not attempt this as this would halve memory capacity and would worsen overall performance by impacting page fault rates. However, if the system is not utilizing its full memory capacity, it argues for moving all free pages to the Top layers. The PPP scheme decreases the read queue delay by 55%, while EPP decreases the average queuing delay by 38% (shown in Figure 8).

The PPP scheme is not a true upper bound as it captures activity over the entire simulation to classify hot and cold pages. The EPP scheme can occasionally out-perform the PPP scheme by taking advantage of temporal locality. For example, if all pages are accessed equally over the entire simulation, PPP has little to offer. But the EPP scheme will try to move the pages most recently active into the Bottom Regions; this is advantageous if those pages continue to remain hot for a few more epochs. Of course, EPP incurs other penalties – the cost of migration, and inaccuracies in predicting the hot pages in the next epoch.

8. RELATED WORK

Voltage Aware Processor Architectures. A recent workshop paper by Zhang et al. [51] is the first to articulate the importance of IR-drop from an architectural perspective. The paper focuses on IR-drop within a processor, develops a tool called VoltSpot, and argues that if more pins are dedicated for the PDN, fewer pins are made available for data I/O, thus impacting the architecture. A number of prior papers have examined voltage noise from activity fluctuations (Ldi/dt) and developed architectural solutions to smooth out activity [25, 16, 40, 39, 14, 13]. Our work differs from these prior voltage-aware architecture papers because of our focus on DRAM chips and the very different architectural techniques that it leads to.

Current Aware Memory Architectures. Phase Change Memory, which requires large write currents, also requires current-aware scheduling at the memory controller. Hay et al. [15] address the high current needs of PCM banks by evaluating the current needed by each write. They use the concept of *Power Tokens* to keep track of the PCM power usage.

Another recent paper targets the same problem while performing fine-grain power allocations and introducing global charge pumps on the DIMM [24]. The above schemes relate to variation in write activities in PCM, whereas our work focuses on variations in IR-drop based on DRAM bank activity. These works therefore target different problems and develop different solutions. A key contribution here is our demonstration that it is not only enough to just track current consumption – it is also important to track *where* current is being consumed. Depending on which banks are currently active, it may or may not be possible to consume more current. Kim et al. [30] 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.

Page Placement. Many prior works have influenced page placement in the memory hierarchy to handle NUCA latencies [6, 3], NUMA latencies [2, 8, 4], conflict misses in caches [44], DRAM power modes [33], DRAM/PCM hybrids [41, 10], etc. Our work borrows the key ideas in these techniques and shows that they can be highly effective to address the emerging IR-drop problem.

Circuit Efforts to Reduce IR-drop. Wu et al. [49] study the impact of a stacked processor-DRAM device on power delivery and propose using Decaps on the DRAM layer to reduce dynamic IR-drop. Kang et al. [47] suggest the addition of TSVs to provide more power/ground lines. While they foresee only a modest area overhead of 0.5%, the package routing and extra package pins required to accomplish this will increase cost. Healy et al. [17] compare the power supply noise caused by different power TSV topologies. Their results suggest the use of a topology where the Power/Ground TSVs are spread evenly rather than clustered over the Power/Ground C4 bumps. Jung et al. [27] illustrate that higher TSV counts can increase routing congestion because power/ground TSVs can be larger than standard cells, thus exacerbating IR-drop issues. They propose power consumption aware TSV placement.

9. CONCLUSIONS

In this work, we highlight an emerging important problem. We show that while 2D DRAM chips are rarely exposed to IR-drop violations, 3D-stacked memory devices can be highly prone to such violations. For acceptable performance, the memory controller must encode a number of rules to handle the non-uniform vulnerability of each bank to IR-drop. We show that even such a smart memory controller falls short of the performance of an unconstrained memory device by 4.6x. A large fraction of this gap is bridged by introducing a smarter scheduler and a page migration mechanism. The starvation-aware scheduler brings the gap to 1.47x. By further adding page migration, the gap is brought to 1.2x. We thus show that effective architectural policies can yield high performance at low pin/TSV counts.

Several future technology parameters are currently unknown and it is not clear how commercial designs will cope with IR-drop constraints. Our work shows that architectural policies can represent an important approach, possibly mitigating the need for some costly approaches.

Our work can be extended in many different ways. For example, our IR-drop analysis can be extended to handle refresh, and more sophisticated schedulers can be designed to bridge the 1.2x performance gap. While more scheduling rules can improve performance, they also increase power, area, and design complexity. The definition of an optimal memory controller remains an open problem.

Acknowledgments

We thank the reviewers and our shepherd, Ron Ho, for several suggestions that helped improve the paper. This work was supported in parts by NSF grants CCF-0916436, CNS-1302663, and Samsung.

10. REFERENCES

[1] Micron DDR3 SDRAM Part MT41J1G4, 2009.

[2] M. Awasthi, D. Nellans, K. Sudan, R. Balasubramonian, and A. Davis. Handling the Problems and Opportunities Posed by Multiple On-Chip Memory Controllers. In *Proceedings of PACT*, 2010.

[3] M. Awasthi, K. Sudan, R. Balasubramonian, and J. Carter. Dynamic Hardware-Assisted Software-Controlled Page Placement to Manage Capacity Allocation and Sharing within Large Caches. In *Proceedings of HPCA*, 2009.

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

[5] 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.

[6] S. Cho and L. Jin. Managing Distributed, Shared L2 Caches through OS-Level Page Allocation. In *Proceedings of MICRO*, 2006.

[7] E. Cooper-Balis, P. Rosenfeld, and B. Jacob. Buffer On Board Memory Systems. In *Proceedings of ISCA*, 2012.

[8] 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.

[9] David Wang. Why migrate to DDR4? `http://www.eetimes.com/design/memory-design/4409694/Why-migrate-to-DDR4`.

[10] G. Dhiman, R. Ayoub, and T. Rosing. PDRAM: A Hybrid PRAM and DRAM Main Memory System. In *Proceedings of DAC*, 2009.

[11] 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.

[12] Elpida Memory, Inc. News Release: Elpida, PTI, and UMC Partner on 3D IC Integration Development for Advanced Technologies Including 28nm. `http://www.elpida.com/en/news/2011/05-30.html`, 2011.

[13] M. S. Gupta, J. Oatley, R. Joseph, G. Y. Wei, and D. Broooks. Understanding Voltage Variations in Chip Multiprocessors using a Distributed Power-Delivery Network. In *Proceedings of DATE*, 2007.

[14] M. S. Gupta, K. Rangan, M. D. Smith, G. Y. Wei, and D. Broooks. Towards a Software Approach to Mitigate

Voltage Emergencies. In *Proceedings of ISLPED*, 2007.

[15] 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.

[16] K. Hazelwood and D. Brooks. Eliminating Voltage Emergencies via Microarchitectural Voltage Control Feedback and Dynamic Optimization. In *Proceedings of ISLPED*, 2004.

[17] M. Healy and S. K. Lim. Power-supply-network design in 3d integrated systems. In *Quality Electronic Design (ISQED), 2011 12th International Symposium on*, march 2011.

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

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

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

[21] JEDEC. DDR4 in an Enterprise Server. `http://www.jedec.org/sites/default/files/Art_Kilmer.pdf`.

[22] JEDEC. DDR4 Mini Workshop. `http://www.jedec.org/sites/default/files/JS_Choi_DDR4_miniWorkshop.pdf`.

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

[24] 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.

[25] R. Joseph, D. Brooks, and M. Martonosi. Control Techniques to Eliminate Voltage Emergencies in High Performance Processors. In *Proceedings of HPCA*, 2003.

[26] 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.

[27] M. Jung and S. K. Lim. A study of IR-drop noise issues in 3D ICs with through-silicon-vias. In *3DIC*. IEEE, 2010.

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

[29] N. H. Khan, S. M. Alam, and S. Hassoun. System-level comparison of power delivery design for 2D and 3D ICs. In *3DIC*. IEEE, 2009.

[30] 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.

[31] W. Kim, M. S. Gupta, G. Y. Wei, and D. Broooks. System Level Analysis of Fast, Per-core DVFS Using On-Chip Switching Regulators. In *Proceedings of HPCA*, 2008.

[32] Y. Kim, V. Seshadri, D. Lee, J. Liu, and O. Mutlu. A Case for Exploiting Subarray-Level Parallelism (SALP) in DRAM. In *Proceedings of ISCA*, 2012.

[33] A. Lebeck, X. Fan, H. Zeng, and C. Ellis. Power Aware Page Allocation. In *Proceedings of ASPLOS*, 2000.

[34] M. Lu, D.-Y. Shih, S. K. Kang, C. Goldsmith, and P. Flaitz. Effect of Zn doping on SnAg solder microstructure and electromigration stability. *Journal of Applied Physics*, 106, 2009.

[35] Micron Inc. DDR4 SDRAM FAQs. `http://www.micron.com/products/dram/ddr4-sdram`.

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

[37] S. Pawlowski. Intelligent and Expandable High-End Intel Server Platform, Codenamed Nehalem-EX. Intel Developer Forum, `http://blogs.intel.com/technology/Nehalem-EX_Steve_Pawlowski_IDF.pdf`, 2009.

[38] T. Pawlowski. Hybrid Memory Cube (HMC). In *HotChips*, 2011.

[39] M. Powell and T. N. Vijaykumar. Pipeline Damping: A Microarchitecture Technique to Reduce Inductive Noise in Supply Voltage. In *Proceedings of ISCA*, 2003.

[40] M. Powell and T. N. Vijaykumar. Exploiting Resonant Behavior to Reduce Inductive Noise. In *Proceedings of ISCA*, 2004.

[41] L. Ramos, E. Gorbatov, and R. Bianchini. Page Placement in Hybrid Memory Systems. In *Proceedings of ICS*, 2011.

[42] Samsung. Samsung to Release 3D Memory Modules with 50% Greater Density, 2010. `http://www.computerworld.com/s/article/9200278/Samsung_to_release_3D_memory_modules_with_50_greater_density`.

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

[44] T. Sherwood, B. Calder, and J. Emer. Reducing Cache Misses Using Hardware and Software Page Placement. In *Proceedings of SC*, 1999.

[45] K. Sheth, E. Sarto, and J. McGrath. The importance of adopting a package-aware chip design flow. In *Proceedings DAC*, 2006.

[46] Tezzaron Semiconductor. 3D Stacked DRAM/Bi-STAR Overview, 2011. `http://www.tezzaron.com/memory/Overview\_3D\_DRAM.htm`.

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

[48] A. N. Udipi, N. Muralimanohar, N. Chatterjee, R. Balasubramonian, A. Davis, and N. Jouppi. Rethinking DRAM Design and Organization for Energy-Constrained Multi-Cores. In *Proceedings of ISCA*, 2010.

[49] 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.

[50] 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.

[51] R. Zhang, B. H. Meyer, W. Huang, K. Skadron, and M. R. Stand. Some Limits of Power Delivery in the Multicore Era. In *Proceedings of WEED*, 2012.