CACTI-3DD: Architecture-level Modeling for 3D Die-stacked DRAM Main Memory

Ke Chen[‡][†], Sheng Li[†], Naveen Muralimanohar[†], Jung Ho Ahn[§] , Jay B. Brockman[‡] , Norman P. Jouppi[†]
[‡]University of Notre Dame, [†]Hewlett-Packard Labs, [§]Seoul National University
[‡]{kchen2, jbb}@nd.edu, [†]{kec, sheng.li4, naveen.muralimanohar, norm.jouppi}@hp.com, [§]gajh@snu.ac.kr

Abstract—**Emerging 3D die-stacked DRAM technology is one of the most promising solutions for future memory architectures to satisfy the ever-increasing demands on performance, power, and cost. This paper introduces CACTI-3DD, the first architecture-level integrated power, area, and timing modeling framework for 3D die-stacked off-chip DRAM main memory. CACTI-3DD includes TSV models, improves models for 2D off-chip DRAM main memory over current versions of CACTI, and includes 3D integration models that enable the analysis of a full spectrum of 3D DRAM designs from coarse-grained rank-level 3D stacking to bank-level 3D stacking. CACTI-3DD enables an in-depth study of architecture-level tradeoffs of power, area, and timing for 3D die-stacked DRAM designs. We demonstrate the utility of CACTI-3DD in analyzing design trade-offs of emerging 3D die-stacked DRAM main memories. We find that a coarse-grained 3D DRAM design that stacks canonical DRAM dies can only achieve marginal benefits in power, area, and timing compared to the original 2D design. To fully leverage the huge internal bandwidth of TSVs, DRAM dies must be re-architected, and system implications must be considered when building 3D DRAMs with redesigned 2D planar DRAM dies. Our results show that the 3D DRAM with re-architected DRAM dies achieves significant improvements in power and timing compared to the coarse-grained 3D die-stacked DRAM.**

Keywords: 3D architecture, DRAM, TSV, Main memory, Modeling

I. Introduction

Modern computer systems demand ever-increasing performance, power-efficiency, and capacity from Dynamic Random Access Memories (DRAMs) to meet system performance requirements. As Moore's Law drives CMOS technology into the deep nanoscale regime, DRAM scaling faces serious challenges in speed, bandwidth, capacity, and cost. For example, historically CPU performance has improved at an annual rate of 55% while the memory access time has only improved by 10%, resulting in the well-known memory wall problem [6]. Moreover, for decades DRAM capacity had increased 4× every 3 years, but is now scaling much more slowly [6], resulting in a memory capacity wall problem. Power and cost of DRAMs are also facing similar challenges [6].

The DRAM industry has continued to innovate both technologies and architectures in order to scale the performance, power, capacity and cost of DRAMs as shown in Table I. New materials and fabrication processes have been steadily introduced. Hierarchical wordlines and datalines were incorporated to maintain a steady trend of increasing DRAM capacity [7]. Motivated to lower the cost and increase the density of DRAM, the DRAM cell size is decreasing from 8F^2 to 6F^2 and then to 4F^2 [14]. The memory interface itself has seen numerous advances from SDRAM to DDR through the upcoming DDR4 standard. The wide adoption of multicore processors renders memory bandwidth and capacity even more critical. Indicative of this trend, conservative DRAM manufacturers are adopting more radical technologies such as 3D die-stacked DRAM [10], [17].

Innovations	Tech Gens (nm)
Hierarchical wordline [7]	200
Interface from DDR to DDR2, DDR3 [5], [8]	100 & 65
Varying number of cells per bitline [18]	90
Cell size from 8F^2, to 6F^2 and 4F^2 [14]	65 & 36
3D stacking [17]	50
Copper metallization [14]	44
High-k dielectric gate oxide [14]	31

TABLE I
DRAM TECHNOLOGY/ARCHITECTURE INNOVATIONS AND THE TECHNOLOGY GENERATIONS WHEN THE INNOVATION WERE(WILL BE) WIDELY USED. THE LAST TWO ROWS ARE FUTURE MILESTONES ACCORDING TO ITRS [14]. THE SELECTED REFERENCES ARE REPRESENTATIVE DESIGNS BUT NOT NECESSARILY THE FIRST DESIGN.

Despite technology advances from the DRAM industry, the ability to propose and evaluate new DRAM designs and their system implications is currently limited by the availability and quality of appropriate system-level tools. CACTI-D [15] built a solid foundation for modeling DRAM technologies, including cells and DRAM subarrays. However, since its peripheral circuit models including the control path and data path were inherited from SRAM models, CACTI-D's overall DRAM model is more appropriate for embedded DRAM than off-chip DRAM main memory. Correct modeling of the peripheral circuits including hierarchical wordlines and datalines are critical, since peripheral circuits play a critical role in determining the overall power, area, and timing. For example, modern DRAM designs usually achieve an area efficiency of approximately 50% [14]. Vogelsang [18] partially addressed this problem by developing a power model with detailed peripheral circuit models for DRAM main memories. However, this model is only for power and does not estimate timing and area of a DRAM design, which is insufficient since power, area, and timing are inseparable for modern DRAMs.

Especially important today is the ability to model emerging 3D die-stacked DRAM technology [10], [17], which shows tremendous promise for addressing performance, power, and capacity challenges in the near future. Tsai et. al [16] extended an earlier CACTI version to model 3D die-stacked SRAMs. However, 3D DRAM is substantially different from 3D SRAM in the memory cell physics, fabrication technology, circuit implementation, memory organization, and peripheral circuit arrangements. Thus, a 3D SRAM model provides an inadequate basis for modeling 3D die-stacked DRAM [10], [17].

978-3-9810801-8-6/DATE12/©2012 EDAA

Ke Chen and Jay Brockman are partially supported by the C2S2 Focus Center, one of six research centers funded under the Focus Center Research Program (FCRP), a Semiconductor Research Corporation entity.

This material is based upon work supported by the Department of Energy under Award Number DE - SC0005026. The disclaimer can be found at http://www.hpl.hp.com/DoE-Disclaimer.html

Jung Ho Ahn was supported by the Smart IT Convergence System Research Center funded by the Ministry of Education, Science and Technology as Global Frontier Project.

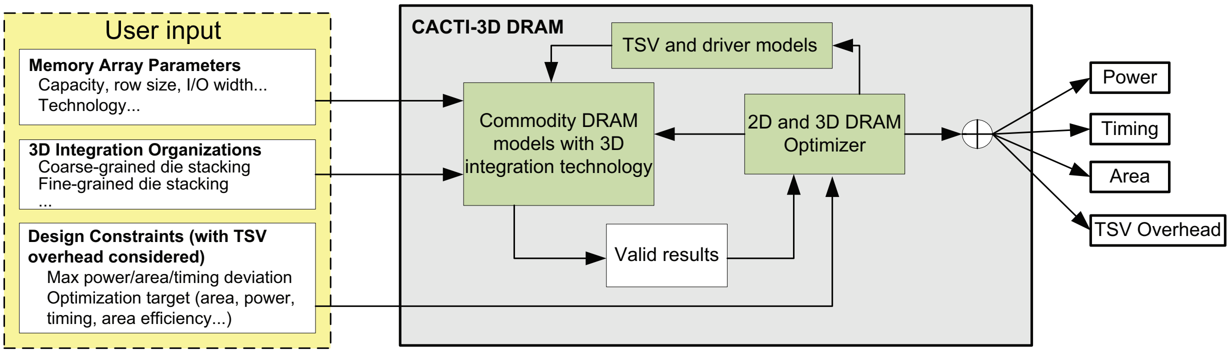
User input
Memory Array Parameters
Capacity, row size, I/O width...
Technology...
3D Integration Organizations
Coarse-grained die stacking
Fine-grained die stacking
...
Design Constraints (with TSV overhead considered)
Max power/area/timing deviation
Optimization target (area, power, timing, area efficiency...)
CACTI-3D DRAM
TSV and driver models
Commodity DRAM models with 3D integration technology
2D and 3D DRAM Optimizer
Valid results
Power
Timing
Area
TSV Overhead

Fig. 1. Block diagram of the CACTI-3DD framework.

In this paper, we introduce CACTI-3DD, a major extension to CACTI, designed to support architecture-level modeling of modern and future DRAM main memories. CACTI-3DD captures all the major innovations shown in Table I including hierarchical wordlines and datalines and emerging 3D DRAM technology. It simultaneously models power, area, and timing of commodity 2D and 3D DRAM designs, a requirement for accurate architecture-level tradeoffs. To demonstrate its capabilities, we use CACTI-3DD to compare different 3D DRAM architectures.

II. CACTI-3DD

CACTI-3DD is the first dedicated framework for simultaneously modeling power, area, and timing of 2D planar DRAM main memory and 3D DRAM designs. Figure 1 shows the block diagram of CACTI-3DD. The key enhancements of CACTI-3DD include: 1) power, area, and timing models of TSVs, including optimization of the driver size for each TSV; 2) DRAM models that capture all the technology and architecture innovations shown in Table I; and 3) an optimizer that performs design space exploration for both 2D planar DRAMs and 3D die stacked DRAMs to find the best design that fits in the design constraints of area, power, and timing.

Besides the new models of 3D DRAMs and TSVs, CACTI-3DD also significantly improved the models of 2D planar DRAMs. Compared to the current state-of-the-art models for 2D planar DRAMs in CACTI-D [15] and the work from Vogelsang [18], CACTI-3DD's major changes include: 1) modeling a complete center stripe including memory bus and controllers; 2) hierarchical wordlines and datalines; 3) separation of row/column address and data paths including separation of row/column decoders for independent column select and read/write operations; 4) simultaneously modeling power, area, and timing of a DRAM device.

Once CACTI-3DD collects the input parameters including the memory array configurations, 3D DRAM configurations, and the design constraints, it performs an automatic design space exploration to find the best design. CACTI-3DD also considers temperature as an input when computing final results, especially for leakage power since it is a strong function of temperature. CACTI-3DD sweeps the per-die DRAM configurations and the associated TSV configurations to find the optimal design based on the design constraints. Although floorplan is an important design considerations in 3D circuits and architectures, it is simpler in DRAM structures than in processors. For 3D DRAMs, the floorplan for an individual DRAM die is mostly regular and similar to that of 2D DRAM dies. The peripheral functional circuits can be shared by all the DRAM dies and placed on a single DRAM die or a interface logic die (as shown in Section IV) in the stack. CACTI-3DD considers these situations during the optimization process and places the address and data TSVs on the edge of banks/subarrays. In addition to the power, area, and timing of the best DRAM designs, CACTI-3DD also reports the TSV overhead in area, timing, and power. The TSV overhead is critical for performance and cost trade-offs, especially for 3D DRAMs with bank-level partitioning that require a large number of TSVs.

A. Modeling of 2D and 3D DRAM

CACTI-3DD models a full spectrum of 3D DRAM designs from coarse-grained rank-level die-stacking to bank-level die-stacking to support comprehensive design space exploration for emerging 3D DRAM designs. Figure 2 illustrates the range of DRAM architectures covered by CACTI-3DD. The first row ((a), (b), (c)) shows the transformation of 2D DRAM architecture to coarse-grained and fine-grained rank-level 3D DRAM architectures. The second row shows the zoomed-in details of the circuits inside a planar bank ((d), (e)) and a 3D stacked bank ((f), (g)). In order to accurately model 3D DRAMs, CACTI-3DD first significantly improves the power, area, and timing models of commodity 2D planar DRAMs, then builds new models for the emerging 3D DRAM designs.

As shown in Figure 2 (a), CACTI-3DD models modern designs of 2D planar commercial DRAMs from major vendors. A 2D planar DRAM device usually comprises multiple arrays (e.g. DDR3 has 8 to 16 arrays), with each array or two serving as a bank. Between all arrays, the center stripe encompasses logic such as voltage regulators, control units, data buffers, I/O pads, and the burst logic. The center stripe and memory buses before row/column predecoders are shared by all banks, while the banks themselves are independent. A "spine-like" memory bus is used to connect the banks with the center stripe. The bus extends to the vertical edges of the banks and sends row addresses to the row predecoders/decoders. All the activation/read/write operations can take place concurrently on multiple banks, while the command/address/data bits are transported sequentially within each bank. This scheme enables bank interleaving without much overhead from the memory bus side.

As shown in Figure 2 (e), the row logic (pre-decoders, decoders, and drivers) and column logic (similar to the row logic) are placed at the edge of the banks. Figure 2 (d) shows the details inside a bank. Each bank is composed of multiple subarrays. Modern DRAM designs leverage hierarchical wordlines in the banks, the global wordlines (metal layer 2) that route across the entire bank, and the local wordlines (polysilicon) that route only across a subarray. This design relaxes the pressure on metal pitch scaling. It also speeds up row activation by reducing the wordline latency, but at the cost of extra area and power overhead for local wordline logic. Although global wordlines are repeated metal wires and route over a bank, they do not interfere with cell subarrays inside a bank since their repeaters are only placed inside the local wordline logic stripes between subarrays. As shown in Figure 2 (d) and (e), the data lines are also hierarchical. Global data lines (metal layer 3) are connected

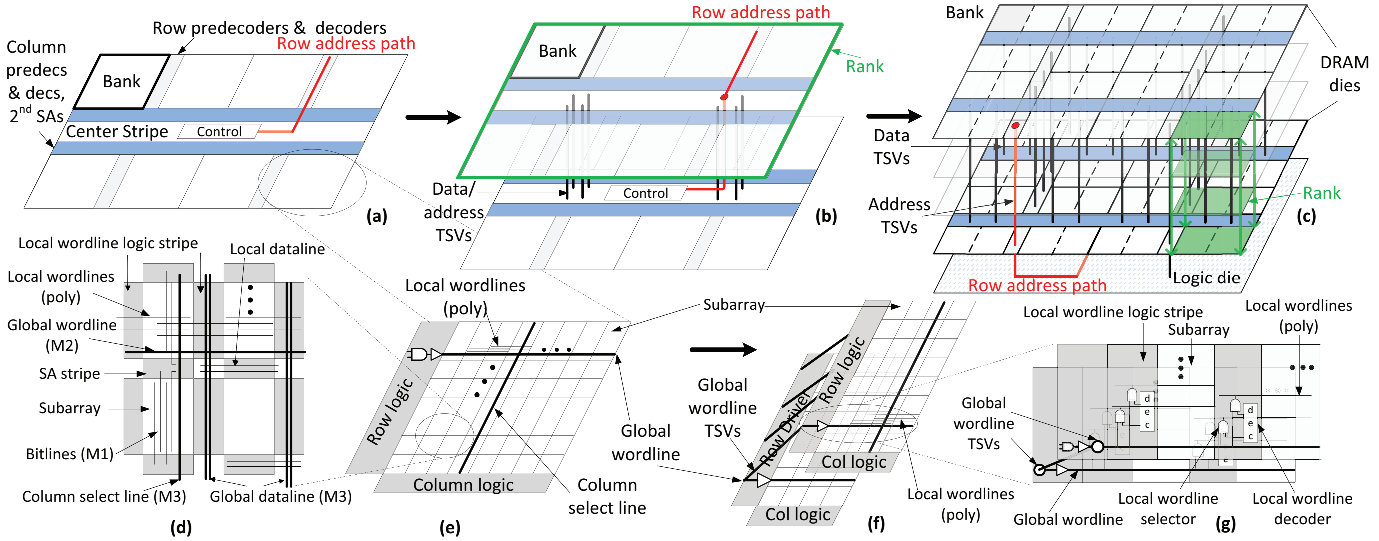

Fig. 2. 2D planar DRAM architecture and 3D DRAM architectures with different levels of partitioning. The TSV height is exaggerated for the purpose of clear illustration. (a) Floorplan of a 2D offchip DRAM die; (b) Coarse-grained rank-level die-stacking; (c) Fine-grained rank-level partitioning and stacking with vertical rank organization, smaller bank size, and a large number of TSVs; (d) Subarrays inside a bank with global/local wordlines and column select lines highlighted; (e) 2D planar bank structures; (f) Bank-level partitioning and die-stacking; (g) Folded global wordlines for bank-level die-stacking.

to the multiplexors and other DQ circuits outside the banks, and are also placed inside the logic stripes between subarrays. A local data line (metal layer 2) spans a subarray along the same direction as the wordlines, and connects to all the sense amplifiers (SAs) within that subarray. All the metal wires form dense on-pitch interconnects above the array.

CACTI-3DD models a suite of 3D DRAM designs with different partitioning strategies. Figure 2 (b) demonstrates a coarse-grained rank-level die-stacking that leverages canonical DRAM dies (such as DDR3). It uses TSVs only as inter-rank interconnects for identical dies to increase DRAM capacity per stack. [1] In the coarse-grained design, the rank still consists of the planar die as in a 2D design. The TSVs are placed in the center stripe, and function as the data and control buses. An example of this design can be found in the 3D DRAM design from Samsung [17].

While it has the advantage of simplicity, the coarse-grained rank-level die stacking does not fully utilize the internal bandwidth that the TSVs can offer. Fine-grained rank-level die stacking involves a major re-partitioning of DRAM arrays, combining portions from multiple stacked dies to form a 3D vertical rank as in Figure 2 (c). This strategy requires a large number of TSV interconnects to transport more address bits and data bits to support a finer granularity of banks. The larger number of banks can provide a higher level of concurrency, and the smaller bank size yields better access latency and power (with shorter wordlines and bitlines), compared to the coarse-grained rank-level die-stacking as in Figure 2 (b). However, the large number of banks on each DRAM die results in high contention on the memory bus that limits the concurrency. Moreover, since the interconnect technologies of DRAM dies are optimized for cost, performance becomes a serious issue for the heavily used on-die memory bus. Hence a logic die is introduced to solve this problem as shown in Figure 2 (c). The ranks are now organized vertically. The inter-rank interconnects are then placed on the logic die so that the interconnect topology and fabrication technology can be optimized for power, area, and/or timing. Within a rank, the inter-bank interconnects are implemented using TSVs, which consume minimal area, latency, and power. Moreover, peripheral circuitry including the entire center stripe can also be brought down to the bottom logic die and optimized for performance, power, and/or area using a logic fabrication process. A design similar to this can be found in the Hybrid Memory Cube (HMC) [10] from Micron.

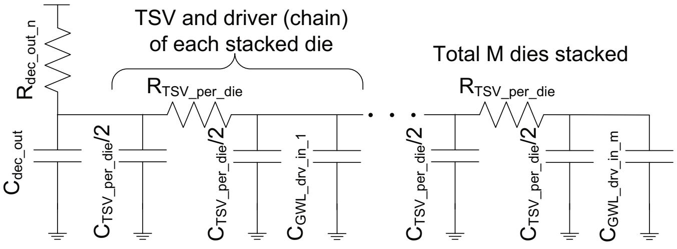

Fig. 3. RC circuit models of drivers and TSV connections of stacked global wordlines as shown in Figure 2 (g), assuming the global wordlines are distributed on M number of stacked DRAM dies.

Bank-level die stacking involves breaking the structure within a bank as shown in Figure 2 (f). While the row decoders remain the same, an individual global wordline splits into multiple sections. Thus, by connecting fewer local wordlines to the global wordline, the wordline latency can be reduced. Global wordline drivers are duplicated in each die. The duplication overhead is offset by reducing driver sizes to match the smaller capacitive load on the partitioned global wordline. The column logic is distributed onto different dies. Since a TSV pitch can be 32× as large as a local wordline pitch, this design requires increasing the ratio of local wordlines to TSV-connected global wordlines. In each subarray, one global wordline needs to connect multiple local wordlines through the local wordline selector as shown in Figure 2 (g). The ratio of the number of global wordlines to the number of local wordlines is determined by the ratio of the TSV pitch to the local wordline pitch (or cell height). Note that the local wordline decoders/selectors also exist in contemporary 2D planar DRAM designs [7] to relax the pitch of the global wordlines, with each global wordline driving 2 or 4 local wordlines. Although distributing one subarray to different dies is applicable for SRAMs [16], it is impractical for DRAMs due to

[1] A rank is the combination of banks that output the same width of data as the data bus, and is formed by banks from different DRAM chips for 2D planar DRAM main memories.

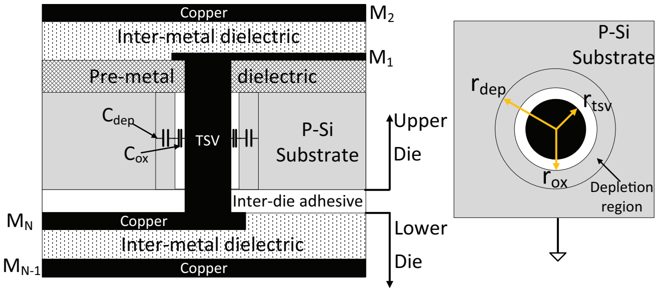

Fig. 4. Cross-section and top-down view of a TSV integrated with via-first and face-to-back process (Figure is based on [2]). Both the upper die (with the TSV in the silicon substrate) and the lower die are extracted from a multi-die stack. Via-first TSVs connect metal layer 1 of the upper die and the top metal layer of the lower die, and the connections between the top metal layer and metal layer 1 of the lower die are made with normal inter-metal-layer vias.

the requirement of partitioning the polysilicon local wordlines and the associated drivers/contacts, which decreases the DRAM density and area efficiency significantly.

Figure 3 shows the RC circuit model of drivers and TSV connections of stacked global wordlines as shown in Figure 2 (g). The global wordline decoder/driver in the row logic stripe connects the driver chain of the global wordline segment on the same die directly, and those on the other dies through a TSV. As shown in Figure 3, the product of the total number of drivers on all the stacked dies and the number of stacked DRAM dies (M) is added to the total resistance and capacitance of the TSV. With the RC model, the delay (as derived in [3]) and power are computed using Equations 1 and 2. The area of the circuit is calculated by summing the area of all transistors, wiring and TSV overhead.

$$T = \tau \ln \frac{V_{end}}{V_{start}} \quad (1)$$

$$P_{total} = \underbrace{\alpha f C_{ckt} \Delta V V_{dd}}_{Dynamic} + \underbrace{V_{dd} I_{sub} + V_{dd} I_g}_{Leakage} \quad (2)$$

Here τ is the time constant computed with the Elmore delay method; V_{start} and V_{end} are the beginning voltage and the voltage when the circuit is considered to have "switched", respectively; C_{ckt} is the total capacitance of the circuit; ΔV is the voltage swing; V_{dd} is the supply voltage; α is the activity factor of the circuit; and f is the clock frequency. $C_{ckt} \Delta V V_{dd}$ gives the energy per access. I_{sub} and I_g are total average subthreshold leakage current and gate leakage current respectively in the circuit block and are used to compute total leakage power. Both I_{sub} and I_g are proportional to transistor size.

B. TSV Modeling

TSVs are critical for enabling 3D die stacking. With their small geometry size, TSVs help realize high density, high bandwidth, and low-inductance vertical interconnection. CACTI-3DD accurately models TSV resistance and capacitance, then analytically models the power, area, and timing of TSV based vertical interconnects (including both TSVs and their drivers) in 3D DRAM designs. CACTI-3DD models "via-first" TSVs that are fabricated before the Si front-end of line (FEOL) device fabrication processing [2]. The geometry parameters used in CACTI-3DD are from ITRS projections [14] including both global interconnect level TSVs (larger and sparser TSVs that travel across larger number of dies) and intermediate interconnect level TSVs (smaller and denser TSVs that travel across a smaller number of dies).

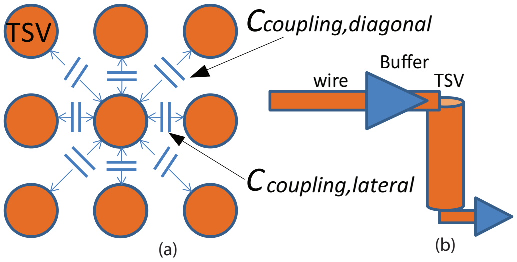

Fig. 5. TSV modeling scheme illustration. (a) TSV 3×3 bundle illustrating coupling capacitance. (b) TSV with a driving buffer.

The resistance, capacitance and inductance are functions of physical parameters and material characteristics of TSVs. Figure 4 shows the cross-sectional view of the TSV structure between two stacked dies, along with key parameters. The impact of inductance on delay and power dissipation for clock frequencies at the gigahertz scale can be ignored for the 3D TSVs [2].

The resistance of a TSV is analytically modeled in CACTI-3DD using Equation 3.

$$R_{tsv} = \frac{\rho l_{tsv}}{\pi r_{tsv}^2} \quad (3)$$

where ρ is the resistivity of the conducting material, r_{tsv} and l_{tsv} represent the radius and length of the TSV, respectively. As shown in Figure 4, for via-first TSVs, the TSVs only need to travel the silicon substrate without the metal layers. Thus, l_{tsv} is only determined by the thickness of a die after thinning and the thickness of the pre-metal dielectic. The impact of skin effect on resistance is negligible for clock frequencies at the gigahertz scale [2].

The capacitance of TSVs is analytically modeled in CACTI-3DD by Equations 4 - 8. The parameters are the same as in Figures 4 and 5,

$$C_{tsv} = C_{intrinsic} + C_{coupling} \quad (4)$$

$$C_{intrinsic} = \frac{C_{ox} C_{dep}}{C_{ox} + C_{dep}} \quad (5)$$

$$C_{ox} = \frac{2\pi \varepsilon_{ox} l_{tsv}}{\ln(r_{ox}/r_{tsv})} \quad (6)$$

$$C_{dep} = \frac{2\pi \varepsilon_{si} l_{tsv}}{\ln(r_{dep}/r_{ox})} \quad (7)$$

$$C_{coupling} = \alpha \frac{\varepsilon_{si}}{S} \pi d_{tsv} l_{tsv} \quad (8)$$

where C_{tsv} is the total capacitance of a TSV, which is the summation of the intrinsic capacitance ($C_{intrinsic}$) of a single TSV and the coupling capacitance ($C_{coupling}$) of the TSV to its neighbors. $C_{intrinsic}$ of a single TSV is the series combination of the capacitance in oxide region (C_{ox}) and the capacitance in depletion region (C_{dep}), which can be derived by solving Poisson's equation in cylindrical coordinates with a full depletion approximation and is expressed by Equations 5, 6, and 7 as in [2]. C_{ox} is computed using Equation 6, where r_{ox} is the radius of the oxide region and ε_{ox} is the oxide dielectric constant. C_{dep} is computed using Equation 7, where r_{dep} is the radius of the depletion region and ε_{si} is the silicon dielectric constant.

As in Equation 4, the other major component of C_{tsv} is the coupling capacitance. The coupling capacitance between two adjacent 3D TSVs can be computed using the closed-form Equation 8 as derived in [12], where S is the space between the vias, d_{tsv} is the diameter and l_{tsv} is the length of the TSV, α is a fitting number accounting for technology and nonlinearity of the coupling capacitance (e.g. for TSVs with aspect ratio of 6 and pitch to diameter

Term	D / L / P (μm)	Reported	Modeled	Error (%)
R ($m\Omega$)	4 / 25 / 30	140	180	28.6
	1.2 / 15 / 20	350	339	3.14
	7.5 / 50 / 60	230	220	-4.5
$C_{intrinsic}$ **(fF)**	4 / 15 / 30	7.0	7.82	11.4
	1.2 / 15 / 20	3.0	3.31	10.3
	7.5 / 50 / 60	47.4	52.6	10.9
$C_{coupling}$ **(fF)**	1.75 / 12 / 4	1.16	1.34	13.4
	1.75 / 12 / 6	0.60	0.51	14.3
Area (μ**m**^{2})	7.5 / 50 / 40	1895	1629	-14.0

TABLE II
VALIDATION OF TSV MODELS IN CACTI-3DD AGAINST SIMULATION AND MEASUREMENT RESULTS FROM INDUSTRIAL DESIGNS AND LABORATORIAL EXPERIMENTS [4], [9], [12], [13], [17]. D, L, AND P REPRESENT DIAMETER, LENGTH, AND PITCH, RESPECTIVELY.

ratio of 2, α is approximately 0.37 [12]). Unlike a horizontal wire that has coupling capacitance with neighbors in the same metal layer and the metal layers above and below, the TSVs are laid out in regular matrixes. As demonstrated in [11], for a TSV in the matrix, the coupling capacitance to the immediate neighbors is much more significant than those to non-adjacent TSVs. Thus, a 3×3 bundle is sufficient to model the coupling capacitance of a TSV. The coupling capacitance of the center TSV equals the sum of the coupling capacitance to 4 lateral vias and to 4 diagonal vias because of its symmetry as shown in Figure 5 (a).

CACTI-3DD also models TSV driving circuits. Unlike repeated wires where repeaters/buffers can be inserted anywhere in the wire to reach the optimal performance, repeater chains have to be inserted instantly before and after a TSV to ensure the driving strength as demonstrated in Figure 5 (b). We use the logical effort method to compute the size and number of buffer stages that give the minimum delay. When evaluating 3D DRAM designs, CACTI-3DD accounts for the RC of both the buffer chains and the TSVs to obtain the power, delay and area. The area of a TSV includes both the TSV area itself and the total area of the buffer chain. If the buffer devices can fit inside the inter-TSV space, then the TSV area is assumed to be no more than the square of the TSV pitch. Otherwise, extra area is needed to place the buffer, and the pitch has to be extended.

III. VALIDATION

The primary focus of CACTI-3DD is the accurate modeling of power, area, and timing of DRAM structures in both 2D planar and 3D stacked DRAM designs. We first validate the TSV models and then validate DRAM systems for both 2D planar and 3D stacked DRAMs.

A. TSV validation

Our validations for TSVs consider resistance, intrinsic capacitance, coupling capacitance, and area. The validation targets include both finite-element simulated data [12], [13] and measured data from real TSV fabrications [4], [9], [17]. The modeled results generated by CACTI-3DD are well aligned with the reported results across a large span of geometry sizes as shown in Table II, which contributes to the modeling accuracy of entire 3D DRAM designs.

B. Overall 2D and 3D DRAM Validation

We validate CACTI-3DD against several industrial DRAM chips for both 2D planar and 3D stacked DRAM designs with different technologies and different DRAM generations. These commercial memory validation targets include a Micron 1Gb 78 nm DDR3 memory [1], [8], a Samsung 2Gb 80 nm DDR2 memory [5], and a Samsung 8Gb 3D DDR3 DRAM [17]. The configurations for the validations are based on the published data of the target DRAMs in [5], [8], [17] including the DRAM density, technology, number of banks, IO width, and burst length. The tool outputs key DRAM metrics, including timing ($t_{RCD}, t_{RAS}, t_{CAS}, t_{RP}$), area, and power (activate/read/write power consumption).

Table III shows the comparison of CACTI-3DD model results against the reported numbers of the target 2D and 3D DRAM designs. Note that the results from the Micron DDR3 data sheet are based on 1.5V 1Gb ×8 DRAM with a 1066 Mb/s data rate. The Micron DDR3 DRAM is assumed to be fabricated using 78 nm technology since it is the mainstream DRAM technology to provide a competitive DRAM die size with low fabricate cost for DDR3 DRAMs [1]. The area of the Micron 78 nm DDR3 DRAM is measured from its die photo. The results of latency, power, and area generated by CACTI-3DD yield high accuracy, with errors within a range of 10% compared to reported data for most of the key parameters. As seen in Table III, the model for 3D memory is especially accurate, with the error percentages of the modeling results being 0.9% and 2.2% for timing and area respectively.

The validation targets cover different DRAM memories at technology generations from 80nm to 50nm, from designs of major commercial vendors, and across DRAM generations from DDR2, DDR3 to the emerging 3D DRAM. Thus, the validation stresses CACTI-3DD in a comprehensive and detailed way as well as tests its accuracy for modeling memories across multiple technology generations and different DRAM architectures.

IV. 3D DRAM MAIN MEMORY DESIGN TRADE-OFFS

CACTI-3DD models a full spectrum of 3D DRAM design options to support the architecture-level design space exploration of emerging 3D off-chip DRAM main memories. In this section, we demonstrate the utility of CACTI-3DD by applying it to a design trade-off study between coarse-grained rank-level (Figure 2 (b)) and fine-grained rank-level (Figure 2 (c)) 3D DRAM designs. For both design schemes, we consider 8 dies (2Gb per DRAM die) in the 3D DRAM stack, with 50 nm technology, and an inter-die data width of 32 bits. Additionally, the fine-grained design has an additional logic die.

The coarse-grained rank-level 3D DRAM leverages the canonical planar designed DRAM dies in a package, and uses TSVs only as the inter-die/rank buses. Because of the minimal re-design efforts for the DRAM dies and the limited usage of TSVs, this design significantly increases the memory capacity per package, with small non-recurring engineering (NRE) cost and little extra fabrication cost for the TSVs. However, the benefit of 3D integration at this level is still marginal since it does not fully utilize the internal bandwidth that the TSVs can provide. The fine-grained rank-level 3D DRAM (Figure 2 (c)) involves re-architecting the DRAM dies and introducing an additional logic die. Thus, this design incurs much greater NRE cost for both the DRAM dies and the logic dies than does the coarse-grained design, not to mention the fabrication cost of the extra TSVs. However, it can bring significantly more benefits to overall performance and power compared to the coarse-grained alternative. The additional logic die not only offers the freedom of exploring different interconnect options for the data path on the logic die such as “H-tree”, but also enables sharing common DRAM logic among DRAM dies to increase DRAM area efficiency.

Table IV shows the modeling results of CACTI-3DD for both coarse-grained and fine-grained rank-level 3D DRAM designs. Compared to the coarse-grained design, the activation energy and latency of the fine-grained design are reduced by 48.5% and 46.9% respectively because of the reduced bank size and the optimized

Validation Targets	Latency (ns)	Real / Model / Err	Power (mW)	Real / Model / Err	Area (mm^2) Real / Model / Err
Samsung 2Gb 80nm DDR2	t_{RCD}	8.8 / 8.39 / -4.6%		N/A	195.64 / 189.89 / -2.9%
	t_{RP}	9.0 / 9.04 / 0.4%			
Micron 1Gb 78nm DDR3	t_{CAS}	15 / 13.9 / -7.3%	P_{ACT}	90.6 / 93.9 / 3.6%	100.22 / 90.45 / -9.8%
	t_{RAS}	36 / 32.1 / -10.9%	P_{RD}	57.1 / 62.4 / 9.4%	
	t_{RC}	51 / 50.6 / -0.9%	P_{WR}	39.0 / 37.5 / -3.9%	
Samsung 8Gb 3D 60nm DDR3	t_{CAS}	15 / 14.9 / -0.9%		N/A	98.1×4 / 100.3×4 / 2.2%

TABLE III
VALIDATION OF DYNAMIC POWER, AREA AND TIMING RESULTS OF CACTI-3DD AGAINST INDUSTRIAL DRAM MEMORY DESIGNS. THE REPORTED DATA ARE FROM MICRON AND SAMSUNG [5], [8], [17]. 'N/A' MEANS THE SPECIFIED DATA IS NOT AVAILABLE IN THE CORRESPONDING PUBLICATION.

Strategy	Chip stack				TSV overhead		Bank		Bus	
	Area	Energy	Latency	Concurrency	Area	Latency	Energy	Latency	Area	Latency
Coarse-grained	100%×8	100%	100%	100%	0.6%	1.5 ns	100%	100%	0.3%	100%
Fine-grained	65.1%×8 + 48.2%	51.5%	53.1%	400%	4.3%	1.5 ns	50.5%	58.3%	5.3%	37.2%

TABLE IV
DESIGN TRADEOFF ANALYSIS OF COARSE-GRAINED AND FINE-GRAINED RANK-LEVEL 3D DRAMS WITH RESULTS FROM CACTI-3DD. THE ENERGY AND LATENCY TERMS ARE EVALUATED ON A PER ACCESS BASIS. AT THE CHIP LEVEL, WE SHOW DRAM DIE AND LOGIC DIE SIZES SEPARATELY FOR THE FINE-GRAINED STRATEGY. FOR TSVS, WE ASSUME A 40 μm PITCH AND SHOW THE WORST CASE LATENCY. THE AREA TERMS ARE NORMALIZED TO THE ENTIRE DRAM CHIP OF THE COARSE-GRAINED STRATEGY. VALUES FOR BANK AND BUS ARE NORMALIZED TO THE COARSE-GRAINED CASE.

interconnection. In particular, the bank has lower energy (50.5%) and latency (58.3%) primarily because of the reduced wordline and bitline lengths (smaller time constant and less precharge/restore energy). In addition, the fine-grained design has 4× as high concurrency as coarse-grained design (32-banks vs. 8-banks per die).

In order to fully utilize the increased number of ranks and banks in the fine-grained design, we assume that the fine-grained design needs to use an interconnect different from the "spine" bus. We choose an H-tree interconnect in this paper. Were the H-tree interconnect fabricated on the 32-bank DRAM dies, it would cause an area overhead of as much as 5.3%, compared to the 0.3% negligible overhead of the memory bus on the canonical 8-bank DRAM die. However, by implementing the interconnects and the common logic in the logic die, the total die size of the 8 DRAM dies can be reduced by 35.9%, whereas the logic die consumes area as much as 48.2% of a single DRAM die in the coarse-grained design. The performance of the interconnect on the logic die improves by 62.8% from using a logic process and high performance devices on the logic die.

CACTI-3DD also provides an analysis of the TSV overhead for 3D DRAMs. As shown in Table IV, the TSV area overhead for the coarse-grained design is marginal, only 0.6% for a 40 μm pitch. However, the fine-grained design yields a 4.3% area overhead (considering signal, power, and redundant TSVs) as the cost of high bandwidth. In both designs, the worst case TSV latency (to the 8th DRAM die) is 1.5 ns, or 1 cycle for a 667 MHz bus frequency.

V. CONCLUSION

Ever-increasing demands on performance, power, and cost of DRAMs drive significant innovational changes in every technology generation. Emerging 3D die-stacked DRAMs can provide the most promising solution for future DRAM architectures. CACTI-3DD is the first architecture-level integrated power, area, and timing tool to model future 3D die-stacked DRAMs. It accurately models TSVs, the critical enabler of 3D integration, and considers all major innovations for both 2D planar DRAM and 3D die-stacked DRAM designs. By providing these capabilities, CACTI-3DD bridges the gap between circuit, device technologies, and system organizations, enabling architects to perform quantitative research on a broad design space for both commodity 2D planar DRAMs and future 3D die-stacked DRAMs. Using CACTI-3DD, we show that coarse-grained 3D DRAMs can significantly increase DRAM capacity per package with minimal extra cost. However, to fully leverage the huge internal bandwidth of TSVs, DRAM dies must be re-architected, and system implications must be considered when building 3D DRAMs using a re-designed 2D planar DRAM dies. Our results show that a 3D DRAM design with re-architected DRAM dies can achieve a significant improvement in power, area, and timing compared to a coarse-grained 3D DRAM design.

REFERENCES

[1] F. Fishburn, *et al.*, "A 78nm 6F^2 DRAM Technology for Multigigabit Densities," *in VLSIT*, 2004.

[2] G. Katti, *et al.*, "Electrical Modeling and Characterization of Through Silicon via for Three-Dimensional ICs," *IEEE Trans. on Electron Devices*, no. 1, pp. 256–262, January 2010.

[3] M. A. Horowitz, "Timing Models for MOS Circuits," Stanford University, Tech. Rep., 1984.

[4] J. Kim, *et al.*, "A 1.2V 12.8GB/s 2Gb Mobile Wide-I/O DRAM with 4x128 I/Os Using TSV-Based Stacking," in *ISSCC*, 2011.

[5] K-H. Kyung, *et al.*, "A 800Mb/s/pin 2Gb DDR2 SDRAM using an 80nm Triple Metal Technology," in *ISSCC*, 2005, pp. 468–469.

[6] P. Kogge, *et al.*, "Exascale computing study: Technology challenges in achieving exascale systems," 2008.

[7] M. Nakamura, *et al.*, "A 29ns 64Mb DRAM with Hierarchical Array Architecture," in *ISSCC*. IEEE, 1995, pp. 246–247.

[8] Micron Technology Inc., "DDR3 Datasheet," 2007.

[9] R. S. Patti, "Three-Dimensional Integrated Circuits and the Future of System-on-Chip Designs," in *Proceedings of the IEEE*, 2006.

[10] J. T. Pawlowski, "Hybrid Memory Cube (HMC)," in *Proceedings of Hot Chips 23*, 2011.

[11] R. Weerasekera, *et al.*, "Compact Modelling of Through-Silicon Vias (TSVs) in Three-Dimensional (3-D) Integrated Circuits," in *3DIC*, 2009.

[12] I. Savidis and E. Friedman, "Closed-Form Expressions of 3-D Via Resistance, Inductance, and Capacitance," *IEEE Trans. on Electron Devices*, vol. 56, no. 9, September 2009.

[13] I. Savidis and E. G. Friedman, "Electrical Modeling and Characterization of 3-D Vias," in *ISCAS*. IEEE, 2008, pp. 784–787.

[14] Semiconductor Industries Association, "International Technology Roadmap for Semiconductors."

[15] S. Thoziyoor, J. Ahn, M. Monchiero, J. B. Brockman, and N. P. Jouppi, "A Comprehensive Memory Modeling Tool and Its Application to the Design and Analysis of Future Memory Hierarchies," in *ISCA*, 2008.

[16] Y.-F. Tsai, Y. Xie, V. Narayanan, and M. J. Irwin, "Three-Dimensional Cache Design Exploration Using 3D Cacti," in *ICCD*, 2005.

[17] U. Kang, *et al.*, "8 Gb 3-D DDR3 DRAM Using Through-Silicon-Via Technology," *JSSC*, vol. 45, no. 1, pp. 111–119, 2010.

[18] T. Vogelsang, "Understanding the Energy Consumption of Dynamic Random Access Memories," in *MICRO*, 2010, pp. 363–374.