DESTINY: A Tool for Modeling Emerging 3D NVM and eDRAM caches

Matt Poremba[¶], Sparsh Mittal[⋆], Dong Li[⋆], Jeffrey S. Vetter[⋆§], Yuan Xie[♯]
[¶]Pennsylvania State University [⋆]Oak Ridge National Laboratory
[§]Georgia Institute of Technology [♯]University of California at Santa Barbara
Email: mrp5060@psu.edu,{mittals,lid1,vetter}@ornl.gov,yuanxie@ece.ucsb.edu

***Abstract*—The continuous drive for performance has pushed the researchers to explore novel memory technologies (e.g. non-volatile memory) and novel fabrication approaches (e.g. 3D stacking) in the design of caches. However, a comprehensive tool which models both conventional and emerging memory technologies for both 2D and 3D designs has been lacking. We present DESTINY, a microarchitecture-level tool for modeling 3D (and 2D) cache designs using SRAM, embedded DRAM (eDRAM), spin transfer torque RAM (STT-RAM), resistive RAM (ReRAM) and phase change RAM (PCM). DESTINY facilitates design-space exploration across several dimensions, such as optimizing for a target (e.g. latency or area) for a given memory technology, choosing the suitable memory technology or fabrication method (i.e. 2D v/s 3D) for a desired optimization target etc. DESTINY has been validated against industrial cache prototypes. We believe that DESTINY will drive architecture and system-level studies and will be useful for researchers and designers.**

***Keywords*—Cache, SRAM, eDRAM, STT-RAM, ReRAM, PCM, non-volatile memory (NVM or NVRAM), modeling tool, validation.**

I. Introduction

Recent trends of increasing system core-count and memory bandwidth bottleneck have necessitated use of large size on-chip caches. For example, Intel's Ivytown processor has 37.5MB SRAM LLC [1]. To overcome the limitations of SRAM, viz. high leakage power and low density, researchers and designers have explored alternate memory technologies, such as eDRAM, STT-RAM, ReRAM and PCM. These technologies enable design of large size caches, for example, IBM's 22nm POWER8 processor uses 96MB L3 eDRAM cache [2]. In parallel, research has also been directed to novel fabrication techniques such as 3D integration that enables vertical stacking of multiple layers [3]. 3D stacking provides several benefits such as high density and higher flexibility in routing signals, power and clock.

Lack of comprehensive and validated modeling tools, however, hinders full study of emerging memory technologies and design approaches. The existing modeling tools model only a subset of memory technologies, for example CACTI [4] and its extension can model SRAM, eDRAM and DRAM but not NVMs, and NVSim [5] models only 2D designs of SRAM and NVMs but not eDRAM. As an increasing number of commercial designs utilize 3D stacking [6, 7], architecture and system level research on 3D stacking has become even more important. A few 3D modeling tools exist such as CACTI-3DD [8] and 3DCacti [9], however, they do not model NVMs. Also, 3DCACTI has not been updated to support technology nodes below 45nm. Since different tools use different modeling frameworks and assumptions, comparing the estimates obtained from different tools may be incorrect. Thus, a single, validated tool which can model both 2D and 3D designs using prominent memory technologies is lacking. Due to this, several architecture-level studies on 3D caches derive their parameters using a linear extrapolation of 2D parameters which may be inaccurate or sub-optimal.

In this paper, we present DESTINY[1], a 3D design-space exploration tool for SRAM, eDRAM and non-volatile memory. DESTINY utilizes the 2D circuit-level modeling framework of NVSim for SRAM and NVMs. Also, it utilizes the coarse- and fine-grained TSV (through silicon via) models from CACTI-3DD. Further, DESTINY adds the model of eDRAM (Section II-A) and two additional types of 3D designs (Section II-B). Overall, DESTINY enables modeling of both 2D and 3D designs of five memory technologies (SRAM, eDRAM and three NVMs). Also, it is able to model technology nodes ranging from 22nm to 180nm. Clearly, DESTINY provides *comprehensive* modeling and design space exploration capability which is not provided by any of the existing tools.

We have compared the results from DESTINY against several commercial prototypes [6, 7, 10–14] to validate 2D design of eDRAM and 3D designs of SRAM, eDRAM and ReRAM in DESTINY (Section III). We observe that the modeling error is less than 10% for most cases and less than 20% for all cases. This can be considered reasonable for an academic modeling tool and is also in range with the errors produced by previous tools [5].

DESTINY provides the capability to explore a large design space which provides important insights and is also useful for early stage estimation of emerging memory technologies. For example, while it may be straightforward to deduce the optimal memory technology for some parameters (e.g. the technology

M. Poremba and S. Mittal are co-first authors.
Support for this work was provided by U.S. Department of Energy, Office of Science, ASCR. This manuscript has been authored by UT-Battelle, LLC, under Contract No. DE-AC0500OR22725 with the U.S. Department of Energy. M. Poremba and Y. Xie are supported in part by NSF 1218867, 1213052, 1409798 and DoE DE-SC0005026. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for the United States Government purposes. This work was performed when M. Poremba was working as a graduate student intern at ORNL.

[1] The source-code of DESTINY can be downloaded from the following `git` repository: https://code.ornl.gov/3d_cache_modeling_tool/destiny.git

978-3-9815370-4-8/DATE15/©2015 EDAA

with smallest cell size is likely to have lowest area), this may not be easy for other parameters such as read/write EDP (energy delay product), since they depend on multiple factors. Clearly, use of a tool such as DESTINY is imperative for full design space exploration and optimization

II. Modeling Framework

DESTINY is designed to be a comprehensive tool able to model multiple memory technologies. Figure 1 shows a high-level diagram of DESTINY. DESTINY framework utilizes the 2D circuit-level model of NVSim, which was extended to model 2D eDRAM and 3D design of SRAM, eDRAM and monolithic NVMs. For a given memory technology, the device-level parameters (e.g. cell size, set-voltage, reset voltage) are fed to DESTINY. Then, possible configurations are generated which are passed to the circuit-level modeling code. For 3D designs, 3D modeling is also done and the configurations include different number of 3D layers (e.g. 1, 2, 4, 8, 16 etc.). The designs which are physically infeasible are considered as invalid and are discarded, for example, if the refresh period of an eDRAM design is greater than its retention period, it is considered invalid. This reduces the number of possible options. The remaining configurations are passed through an optimization filter which selects the optimal configuration based on a target such as least read latency or least area etc.

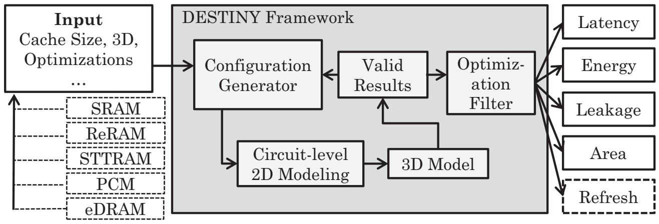

Fig. 1: High-level overview of DESTINY framework.

DESTINY also allows design space exploration across multiple memory technologies, for example, finding an optimal technology for a given metric/target. For such use cases, device-level parameters for multiple memory technologies can be fed as input to DESTINY (shown in left of Figure 1). Using these, the best results for each technology are found which are further compared to find the optimal memory technology. Similar approach is also used for finding the optimal layer count for a given target. More details can be found in our extended technical report [15].

For details on the data storage mechanism of each memory technology, we refer the reader to previous work [5, 16]. We now discuss the specific extensions made in DESTINY modeling framework.

A. eDRAM Model

NVSim provides an incomplete eDRAM model which has also not been validated against any prototype. For enabling modeling of eDRAM, we separate the peripheral and device logic to simulate multiple types of technologies. It is well-known that eDRAM requires refresh for maintaining data integrity and typical retention periods range from 40μs – 100μs [6, 7] for temperature in the range of 380K. We implemented a refresh model based on Kirihata et al. [17], whereby all subarrays are refreshed in parallel, row-by-row. This approach provides the benefit that the refresh operations do not significantly reduce the availability of banks to service requests. DESTINY can also be extended to model other refreshing schemes such as refreshing the mats in parallel. From performance and feasibility perspective, eDRAM cache designs which provide bank availability (i.e., the percentage of time where the bank is not refreshing) below a threshold are not desired and hence, they are discarded by DESTINY. Since retention period of eDRAM varies exponentially with the temperature [18], DESTINY scales the retention period accordingly to model the effect of the temperature. DESTINY computes the refresh latency, energy, and power which are provided as the output of the tool.

B. 3D Model

Several flavors of 3D stacking have been explored in the literature, such as face-to-face, face-to-back, and monolithic stacking [3]. In all these approaches except monolithic stacking, dies are bonded using various techniques (e.g., wafer-to-wafer, die-to-wafer, or die-to-die bonding). These approaches differ in terms of their effect on die testing and yield. Wafer-to-wafer can potentially reduce yield by bonding a dysfunctional die anywhere in the stack. Die-to-wafer and die-to-die can reduce this by testing individual dies, although it has adverse effect on alignment.

The most common 3D stacking is known as face-to-back bonding. In this form, through silicon vias (TSVs) are used to penetrate the bulk silicon and connect the first metal layer (the back) to the top metal layer (the face) of a second die. In face-to-face bonding, the top metal layer of one die is directly fused to the top metal layer of a second die. Monolithic stacking does not use TSVs at all. Instead, monolithically stacked dies build devices on higher metal layers connected using normal metal layer vias if necessary.

Each approach has its own advantages and disadvantages. Face-to-back must carefully consider placement and avoid transistors when being formed through the bulk silicon while face-to-face does not. Therefore, face-to-face has the potential for higher via density. The downside of face-to-face is that only two layers can be formed in this manner. Monolithic stacking benefits from the highest via density, however this technique cannot be applied in a design which requires transistors to be formed on higher layers, since this can destroy previously formed transistors.

The 3D model of DESTINY facilitates all of the aforementioned types of 3D stacking. Separate from this, the granularity at which TSVs are placed can be either coarse- or fine-grained similar to the approach in CACTI-3DD [8]. This granularity will define how many TSVs are placed and what portions of a cache (e.g., peripheral circuits or memory cells) will reside on different dies. We utilize these models in our validation. First, a model for direct die-to-die stacking with face-to-face bonding is provided [14]. Secondly, the monolithic stacking model for 3D horizontal ReRAM is provided [12]. Face-to-back bonding using TSVs is utilized elsewhere [10, 13].

A few previous works (e.g. [8]) assume that TSVs in face-to-back are buffered, which may lead to redundant buffering in some designs and increases the latency and energy overhead of the TSVs. This overhead may be acceptable in large-sized DRAMs which are modeled in CACTI-3DD, but is unacceptable in caches which are relatively smaller in size and becomes increasingly obvious with smaller memory macro designs. Further, several memory peripheral components already provide full-swing logic signals which do not require extra buffering. In our work, we provide a TSV model which may act as a buffered or unbuffered TSV as well as vias used in face-to-face bonding.

With increasing number of layers, the number of memory mats in each layer is reduced and hence, we need to select a scheme for folding of the memory banks. Our coarse- and fine-grained models assume simplistic folding scheme, where the mats are equally divided in all the layers. In coarse-grained model, TSVs are used to broadcast *undecoded* row and column select signals to all layers at once. One logic layer is assumed to provide output in this model over a shared TSV bus spanning all layers. The fine-grained model differs by broadcasting *decoded* row and column signals to all layers. It is assumed that a dedicated logic layer is used for all pre-decoder units. The resulting design uses more TSVs, but its area and latency may be reduced.

Monolithically stacked horizontal ReRAM (HReRAM) [12] uses a concept similar to cross-point designs (see Figure 2). The limitations of the cross-point designs, however, are the increased sneak current and voltage drop associated with increasing subarray size. In 3D design, this limitation becomes even more prominent and it further limits the subarray size of 3D-stacked ReRAM as the sneak current can potentially flow into upper layers as well. To address this limitation, we extended the cross-point model in NVSim to account for the increased number of wordlines and bitlines in the third dimension.

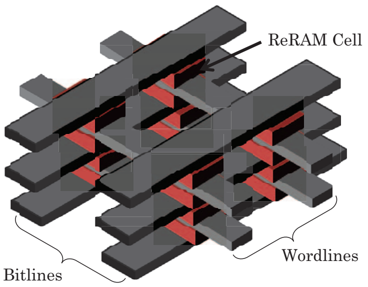

Fig. 2: 4-layer monolithically stacked ReRAM. Storage elements are sandwiched directly between layers of wordlines (south-east orientation) and bitlines (south-west orientation).

An example of HReRAM is shown in Figure 2 with 4 layers monolithically stacked. This monolithic stacking implies that there are no TSVs between memory layers. Instead, the memory cells are sandwiched between wordlines and bitlines and additional layers are added similar to adding more metal layers. Our 3D ReRAM model considers current flow between all inactive bitlines and wordlines when a single cell is read. This model dramatically reduces the number of valid designs when considering ReRAM, so we cannot simply stack cross-point arrays which are considered optimal for a 2D design. Typically this effect can be slightly mitigated using diode accessed cells as in [12] and hence, we provide the ability to model diodes or no access device.

III. Validation Results

To show the accuracy of DESTINY, we validate it against several industrial prototypes. We first obtain the cache/macro configuration from the corresponding prototype papers and use them to set the device-level input parameters for DESTINY. Finally, we compare the results from actual value and projected value from DESTINY and obtain percentage modeling error. As we show below, the modeling error is less than 10% in most cases and less than 20% in all cases.

A. 3D SRAM Validation

We validate the 3D SRAM model of DESTINY against two previous works [13, 14] which utilize hSpice models to simulate latency and energy of 3D-stacked SRAM caches. Hsu and Wu [13] sweep over various cache sizes ranging from 1MB to 16MB. Their work assumes stacking at the bank level, that is a 2D planar cache containing N banks can be stacked up to N-layers. Since NVSim does not model banks, we only compare against the smallest cache size. Our proposed design assumes shared vertical bitlines, which corresponds to the *fine-grained* model in DESTINY. Analogous to their bank folding method, we assume a fixed configuration for two layers and fold along a single dimension in the bank layout to estimate four layer latency and energy. Our two layer design assumes a 4×32 bank layout[2]. Based on the aspect ratio of our SRAM cells and the size of the subarray, this design attempts to keep area square, which is likely the configuration of an hSpice model. The four layer design folds along the number of mats per column assuming a 4×16 bank layout. Table I shows the validation results of DESTINY against the 3D SRAM design in Hsu and Wu [13]. Notice that all errors are less than 4%.

TABLE I: Validation for 3D SRAM model.

Design	Metric	Actual	Projected (DESTINY)	Error
1MB [13]	Latency	1.85 ns	1.91 ns	+3.54%
2 layers	Energy	5.10 nJ	5.05 nJ	-0.98%
1MB [13]	Latency	1.75 ns	1.80 ns	+2.68%
4 layers	Energy	4.5 nJ	4.51 nJ	+0.18%
4MB [14]	Latency	7.85 ns	7.23 ns	-7.91%
2 layers	Energy	0.13 nJ	0.13 nJ	-2.59%
4MB [14]	Latency	6.10 ns	6.95 ns	+14.03%
4 layers	Energy	0.12 nJ	0.13 nJ	+4.75%
2MB [14]	Latency	5.77 ns	5.78 ns	+0.05%
2 layers	Energy	0.12 nJ	0.13 nJ	+2.74%
2MB [14]	Latency	4.88 ns	5.53 ns	+13.5%
4 layers	Energy	0.12 nJ	0.13 nJ	+8.46%
1MB [14]	Latency	3.95 ns	3.90 ns	-1.11%
2 layers	Energy	0.11 nJ	0.11 nJ	-0.13%
1MB [14]	Latency	3.07 ns	3.04 ns	-0.85%
4 layers	Energy	0.11 nJ	0.11 nJ	-0.89%

Puttaswamy and Loh [14] explore the design space of 3D SRAM for 65nm technology node. Their work considers a range of cache sizes from 16KB to 4MB. As explained above, we assume a fixed cache dimension for each cache size and fold the four layer design in half to measure results. These validation results are also shown in Table I. Clearly, all errors are less than 15% and thus, DESTINY is reasonably accurate in modeling 3D SRAM caches.

[2] The bank layouts are specified as mat×mat, and subarray layouts are specified as wordline×bitline.

B. 2D and 3D eDRAM Validation

As stated before, the eDRAM model in NVSim is incomplete and has not been validated. Hence, we validate both 2D and 3D model of eDRAM. The prototype works referenced below typically provide information at the macro level, rather than a full bank. Macros are well suited for verification since they are a memory dense unit (i.e., there is no test circuitry, ECC logic, etc.) and are closest to the modeling assumptions of DESTINY and hence we compare against a macro. Note that for fair comparison, we remove ECC, spare, and parity wordlines and bitlines as these are not modeled in DESTINY.

TABLE II: Validation of 2D and 3D eDRAM.

Design	Metric	Actual	Projected (DESTINY)	Error
2D 2Mb [7] 65nm	Latency	<2 ns	1.46 ns	—
	Area	0.665 mm^2	0.701 mm^2	+5.42%
2D 1Mb [6] 45nm	Latency	1.7 ns	1.73 ns	+1.74%
	Area	0.239 mm^2	0.234 mm^2	-2.34%
2D 2.25Mb [11] 45nm	Latency	1.8 ns	1.75 ns	-2.86%
	Area	0.420 mm^2	0.442 mm^2	+5.31%
3D 1Mb [10] 2-layers	Latency	<1.5 ns	1.42 ns	—
	Area	0.139 mm^2	0.149 mm^2	+9.32%

Barth and Reohr et al. [7] present a 65nm 2D eDRAM prototype. To validate against it, we use the 2Mb macro layout with total 8 subarrays and thus, use the organization of a 1024×2048 bank layout. Klim et al. [11] and Barth and Plass et al. [6] present 45nm 2D eDRAM designs. We validate against them using a subarray layout of 256×1024 as used by them. From Table II, it is clear that the modeling errors in 2D eDRAM validation for all cases is less than 6%.

Golz et al. present a 3D eDRAM prototype with 2 layers in 32nm technology [10]. We use the 1Mb array as our validation target. Based on the 16Kb μArray size of 32×512 and 1Mb layout, we assume two 1024×512 subarrays. From Table II, the modeling error in area is less than 10% and thus, DESTINY can be considered as reasonably accurate.

C. 3D ReRAM Validation

Our final validation target is a monolithically stacked ReRAM memory [12], also known as 3D *horizontal* ReRAM. In monolithically stacked designs, additional wordlines and bitlines are stacked directly by fabricating extra metal layers with NVM cells used in place of vias. This type of design does not use TSV or flip-chip style bonding. Our validation therefore considers our more detailed model of cross-point architecture spanning multiple layers.

We design the simulated memory as a RAM 8Mb in size. The design consists of 4 subarrays each accessed in parallel with a 64-bit input bus. We again remove the ECC logic and specify two monolithically stacked layers per die with one die total. The results of validation are shown in Table III.

TABLE III: Validation of 3D ReRAM.

Metric	Actual [12]	Projected (DESTINY)	Error
Read Latency	25 ns	24.16 ns	-3.37%
Write Latency	17.20 ns	20.13 ns	+17.05%

It is clear that the read latency projection of DESTINY is very close to the value reported in [12] while the error in write latency is higher. This can be attributed to the fact that Kawahara et al. [12] use a write optimization to reduce sneak current, which is not modeled in DESTINY. Furthermore, the range of acceptable write pulse times according to their shmoo plot is very wide, ranging from 8.2ns – 55ns. Our prediction falls in the lower end of the plot which is closer to the 8.2ns write pulse for a total of 17.2ns write time at the bank level.

IV. Future Work and Conclusion

Due to the emerging nature of these memory technologies, only a limited number of prototypes have been demonstrated. Due to the lack of prototypes, we could not validate 3D STT-RAM and 3D PCM, although based on our validation results with 3D ReRAM, we expect that DESTINY will be accurate in modeling them also. We plan to perform these validations as these prototypes become available. Further, we plan to extend DESTINY to model MLC (multi level cell) support for NVMs and also model other emerging memory technologies such as race track memory [16]. Furthermore, we plan to integrate DESTINY in a performance simulator (such as Gem5) to enable architecture/system-level study of these technologies at different levels in cache hierarchy.

In this paper, we presented DESTINY, a comprehensive, validated tool for modeling both 2D and 3D design of prominent conventional and emerging memory technologies. We described the modeling framework of DESTINY and also performed validations against a large number of industrial prototypes. We believe that DESTINY will be useful for architects, CAD designers and researchers.

References

[1] S. Rusu *et al.*, "Ivytown: A 22nm 15-core enterprise Xeon® processor family," in *IEEE ISSCC*, 2014, pp. 102–103.

[2] E. J. Fluhr *et al.*, "POWER8: A 12-core server-class processor in 22nm SOI with 7.6 Tb/s off-chip bandwidth," in *ISSCC*, 2014, pp. 96–97.

[3] B. Black *et al.*, "3D processing technology and its impact on IA32 microprocessors," in *ICCD*, 2004, pp. 316–318.

[4] S. J. E. Wilton *et al.*, "Cacti: an enhanced cache access and cycle time model," *IEEE JSSC*, vol. 31, no. 5, pp. 677–688, 1996.

[5] X. Dong *et al.*, "NVSim: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *IEEE TCAD*, 2012.

[6] J. Barth *et al.*, "A 45 nm SOI Embedded DRAM Macro for the POWER Processor 32 MByte On-Chip L3 Cache," *IEEE JSSC*, 2011.

[7] J. Barth *et al.*, "A 500 MHz Random Cycle, 1.5 ns Latency, SOI Embedded DRAM Macro Featuring a Three-Transistor Micro Sense Amplifier," *IEEE JSSC*, vol. 43, no. 1, pp. 86–95, 2008.

[8] K. Chen *et al.*, "CACTI-3DD: Architecture-level modeling for 3D die-stacked DRAM main memory," in *DATE*, 2012, pp. 33–38.

[9] Y.-F. Tsai *et al.*, "Three-dimensional cache design exploration using 3DCacti," in *ICCD*, 2005, pp. 519–524.

[10] J. Golz *et al.*, "3D stackable 32nm High-K/Metal Gate SOI embedded DRAM prototype," in *VLSIC*, 2011, pp. 228–229.

[11] P. Klim *et al.*, "A one MB cache subsystem prototype with 2GHz embedded DRAMs in 45nm SOI CMOS," in *VLSIC*, 2008.

[12] A. Kawahara *et al.*, "An 8Mb multi-layered cross-point ReRAM macro with 443MB/s write throughput," in *ISSCC*, 2012, pp. 432–434.

[13] C.-L. Hsu *et al.*, "High-performance 3D-SRAM architecture design," in *IEEE APCCAS*, 2010, pp. 907–910.

[14] K. Puttaswamy *et al.*, "3D-Integrated SRAM Components for High-Performance Microprocessors," *IEEE TC*, 2009.

[15] S. Mittal *et al.*, "Exploring Design Space of 3D NVM and eDRAM Caches Using DESTINY Tool," Oak Ridge National Laboratory, USA, Tech. Rep. ORNL/TM-2014/636, 2014.

[16] S. Mittal *et al.*, "A Survey Of Architectural Approaches for Managing Embedded DRAM and Non-volatile On-chip Caches," *IEEE Transactions on Parallel and Distributed Systems (TPDS)*, 2014.

[17] T. Kirihata *et al.*, "An 800MHz embedded DRAM with a concurrent refresh mode," in *ISSCC*, 2004, pp. 206–523.

[18] A. Agrawal *et al.*, "Mosaic: Exploiting the Spatial Locality of Process Variation to Reduce Refresh Energy in On-Chip eDRAM Modules," in *HPCA*, 2014.