Design Space Exploration for 3D-stacked DRAMs

Christian Weis†, Norbert Wehn†, Loi Igor⋆ and Luca Benini⋆
†Microelectronic Systems Design Research Group, University of Kaiserslautern
67663 Kaiserslautern, Germany
⋆DEIS, University of Bologna, Bologna, Italy
{weis, wehn}@eit.uni-kl.de {igor.loi, luca.benini}@unibo.it

***Abstract*—3D integration based on TSV (through silicon via) technology enables stacking of multiple memory layers and has the advantage of higher bandwidth at lower energy consumption for the memory interface. As in mobile applications energy efficiency is key, 3D integration is especially here a strategic technology. In this paper we focus on the design space exploration of 3D-stacked DRAMs with respect to performance, energy and area efficiency for densities from 256Mbit to 4Gbit per 3D-DRAM channel. We investigate four different technology nodes from 75nm down to 45nm and show the optimal design point for the currently most common commodity DRAM density of 1Gbit. Multiple channels can be combined for main memory sizes of up to 32GB. We present a functional SystemC model for the 3D-stacked DRAM which is coupled with a SDR/DDR 3D-DRAM channel controller. Parameters for this model were derived from detailed circuit level simulations. The exploration demonstrates that an optimized 1Gbit 3D-DRAM stack is 15x more energy efficient compared to a commodity Low-Power DDR SDRAM part without IO drivers and pads. To the best of our knowledge this is the first design space exploration for 3D-stacked DRAM considering different technologies and real world physical commodity DRAM data.**

I. Introduction

Mobile applications like video processing and graphics are characterized by ever increasing demands on the memory bandwidth and size. As a consequence, the numbers of IOs of the memory subsystem is continuously increasing. The energy per bit consumed for going off-chip is many times higher than the one required for on-chip accesses, as complex and power hungry IO transceiver circuits are needed to deal with the electrical characteristics of interconnections between chips in a conventional package. Moreover, the random access latencies and the internal cycle times of DRAMs are not decreasing at the same rate as the microprocessor cycle time. This problem is known in high-performance computing as the *Memory Wall* [1], but it is even more daunting in mobile platforms, because power and cost constraints are much more stringent.

3D integration and *3D-stacked memories* have been proposed as a promising solution to the power versus bandwidth dilemma and the Memory Wall. 3D-stacked memories reduce the distance between CPU and external RAM from centimeters to micrometers and improve the random access times - but more importantly, they provide a major boost in energy efficiency in comparison to standard SDR or DDR/2/3 DRAM devices. The pairing of high bandwidth communication with the lower power consumption of 3D integrated memory is an ideal fit for mobile devices. In the last years 3D integration of ICs, especially of DRAMs, received tremendous attention [2]–[8]. There are basically two competing approaches using TSV technology for 3D-DRAMs:

- *Commodity DRAMs* with wide-IO interfaces [3]. Process technologies for commodity DRAMs are optimized for low leakage and minimum cost/bit. This implies optimization for density and cheap technology. DRAM technologies strongly differ from logic technologies, which are optimized for transistor performance and interconnect. As a consequence, random access times in commodity DRAMs are still in the order of 25-30ns.
- *Embedded DRAMs* (eDRAMs) mainly used as SRAM cache replacement [9] in logic technologies. Due to the performance of logic technologies and relaxed density requirements, random access times of less than 2ns are reported [10]. eDRAMs are limited in density and not as economical, especially when targeting memory sizes greater than 72Mb.

In this paper we focus on 3D-stacked solutions using commodity DRAMs to benefit maximally from the large progress in commodity DRAM development w.r.t. density, low leakage, yield, availability and especially low cost. Our 3D-stacked DRAM architecture exploration is based on a sophisticated SystemC model which is linked to a cycle-accurate channel controller to perform subsystem simulations for mobile computing systems. The architecture evaluation for an optimized 3D-stacked DRAM is driven by area, performance and most importantly energy efficiency metrics. Multi-dimensional metrics for a fair comparison are difficult to identify. Thermal issues because of 3D stacking are out of scope of this paper. The main steps forward with respect to similar explorations reported in the literature are: (i) accurate models for DRAM area, power and speed based on detailed circuit-level information on commercial DRAM chips, (ii) cycle-accurate model of a channel controller specifically tuned to 3D-integrated DRAMs, (iii) focus on mobile systems, both in terms of modeling and in terms of design specification and constraints.

The paper is structured as follows. Section II gives a brief overview of the state-of-the-art. Models of the 3D-DRAM stack and the subsystem are presented in Section III. In Section IV the 3D-DRAM organization is discussed. Section V presents an overview of the 3D-channel controller architecture. Section VI finally provides the results of the design space exploration.

978-3-9810801-7-9/DATE11/©2011 EDAA

II. Related Work

Recent investigations [4]–[8] have shown the performance, energy and form factor advantages of 3D integration by using wide-IO buses and TSV interconnects. Facchini et al. [4] mainly focus on the optimization of the interface between processor and memory (e.g., DRAM). The internal structure of the memory is untouched and not optimized. In contrast to [4] we put emphasis on the DRAM itself and we see further improvements if DRAMs are redesigned to take advantage of the high vertical interconnect density. In [8] Loh uses a so-called "true-3D" configuration based on Tezzaron's 3D technology [11] for performance and power evaluations. However Tezzaron's 3D-stacked DRAM approach requires an additional chip layer in logic technology to speed up the interface and also to reduce the access time to the DRAM. The widely used CACTI [12][1] tool is often used to evaluate the performance and the energy consumption of 3D-RAM architectures [5], [8]. Random access times for DRAMs below 2ns were published in [9]. However such low access times are only feasible for small eDRAM macros, e.g. 2.39Mb in SOI technology are reported in [10], and not for commodity DRAMs. Moreover CACTI assumes an equal shrinking for all devices in the DRAM which is not true for advanced commodity DRAMs in which the memory cells shrink faster than the periphery.

Following the 3D-DRAM integration taxonomy proposed by [4], we focus on scenario 3/4 (CMOS IOs). Thus we use CMOS IOs for the connection to the channel controller. Therefore we removed the complex IOs (SSTL) for area, performance and power calculations. Our analysis and exploration are based on state-of-the-art technology data from Inotera, Qimonda and Winbond. These data sets enable us to accurately predict power, performance and area for 3D-DRAM architectures (i), and also to optimize the internal structure, organization and technology of 3D-DRAMs. Together with a cycle-accurate 3D-DRAM channel controller model (ii), which creates realistic traces for the simulations of the 3D-DRAM, a complete optimization of the 3D-DRAM subsystem integrated on mobile computing systems (iii) is done.

Memory controllers manage the architectural and circuit level interface between processors and DRAM. State-of-the-art DRAM controllers [13] are still designed for narrow off-chip interfaces. They are quite complex components and deploy many complex features to maximize the exploitable interface bandwidth. A DRAM controller can be coarsely split in two parts, *front end* and *back-end*, also called *channel controller*. The front end includes a multiport arbitration interface and IO queues with reordering capabilities to improve power consumption and access latency. Many published works describe advanced DRAM front-ends, see [14] for a survey. Front-end design is out of the scope of this paper: we assume a state-of-the-art front-end which delivers memory transactions to the channel controller, and we focus on adaptation of the channel controller to wide 3D-DRAM interfaces. To the best of our knowledge, ours is the first 3D-DRAM-specific channel controller model presented in the open literature.

[1]CACTI was originally developed to evaluate the performance and energy consumption of different caching systems (mainly SRAM) at HP Labs.

III. Modeling the 3D-DRAM subsystem

In a 3D-DRAM subsystem the 3D-DRAM stack is always connected to a 3D-DRAM channel controller [4], [7]. Figure 1 shows the 3D subsystem architecture model used for our exploration.

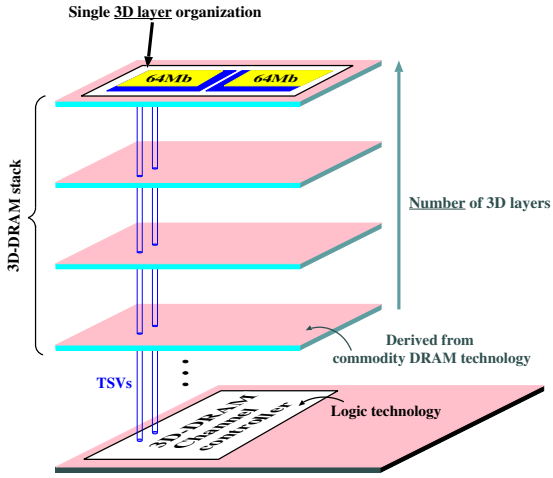

Fig. 1. 3D-DRAM subsystem architecture: A single vertical channel

A. 3D-DRAM stack

The development of a new model from the ground up was forced because models provided by the memory vendors are not flexible enough on the internal DRAM structure to estimate the benefit of 3D-stacked DRAM. The architecture exploration of 3D-stacked DRAM requires a sophisticated modeling approach:

1) Extensive *circuit level simulations* with SPICE were performed to calculate the basic data (e.g., wiring delays or energy per activation). This information together with architectural parameters, geometrical and electrical data are input parameters to the *3D-DRAM generator model* which calculates timing, power and area data, see Figure 2.

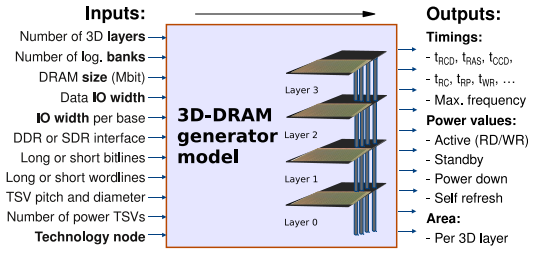

Fig. 2. Inputs/Outputs of the 3D-DRAM generator model

2) The outputs of the generator model are input parameters to the *functional SystemC model* of the 3D-DRAM stack for simulations.

In addition to the parametrizable functional SystemC model for the 3D-DRAM stack we also implemented a SystemC channel controller to simulate the whole system under realistic traffic scenarios. Table I shows the settings for the aggregated

workload for our 3D-DRAM simulations. This setting (similar to the MPEG-4, as shown in [4]) is typical and used by many companies for DRAM traffic scenarios.

TABLE I
TYPICAL WORKLOAD SPECIFICATION

Time	DRAM states & command usage	
30 %	*Idle*	
	10%	Idle without a bank open
	20%	Idle with min. a bank open
60 %	*Bandwidth for IO* (Write/Read)	
	45% Read	*50%* page hit rate, here every 2^{nd} burst
	15% Write	is applied to a already opened bank (page)
10 %	*CKE low* - clock enable is deasserted	
	1%	Power down
	9%	Self-refresh mode

The technology nodes, the cell sizes, the DRAM interface and the voltage settings used for our exploration are shown in Table II. For comparison we also added the Micron Low-Power DDR commodity device (the power for this device is calculated with the DDR SDRAM power calculator available from Micron's website) and the eDRAM macro data from IBM.

TABLE II
POWER SUPPLY VOLTAGE SCALING

Techn. node	Cell area	DRAM interface	VDD [V]	VPP [V]	VWL- [V]
75nm	$8F^2$ a)	3D SDR/DDR	1.3	2.7	-0.5
65nm	$6F^2$	3D SDR/DDR	1.25	3.0	-0.2
58nm	$6F^2$	3D SDR/DDR	1.2	3.0	0
45nm	$4F^2$	3D SDR/DDR	1.1	3.0	0
Micron [15]	$6F^2$	2D LPDDR	1.5	n.a.	n.a.
IBM [10]	$33F_l{}^2$ b)	eDRAM	1.0	1.6	-0.4

a) F = min. feature size in DRAM technology (cell contact spacing)
b) F_l = min. feature size in logic technology (gate length)

IV. DRAM ORGANIZATION FOR 3D-STACKS

Exploring the design space of 3D-DRAMs is challenging due to a large number of options and configurations. Referring to Figure 1, we have to explore two dimensions: horizontal for a single 3D layer and vertical for the stack. We explore these two dimensions w.r.t. energy efficiency, area efficiency and performance. Before going into detail of the design space exploration, the technologies, the basic 3D-DRAM tile, wiring and TSV issues are discussed in more detail.

A. Technologies and 3D-DRAM tile

Figure 3 shows the structure of a 3D-DRAM core tile which is the base to compose a single 3D layer. The tile size is 64Mbit. We will show later that this is the optimum tile size. The tile contains the memory cell array itself, the column area with the secondary sense-amplifiers and data bus IO drivers, the control unit including the command decoder and voltage control, the row decoder and wordline drivers, and the extra space needed for the TSVs (vias for IO signals and power).

The resulting memory cell area, total area and timing data for this tile are shown in Table III for the different technology nodes. Due to space limitations different redundancy options

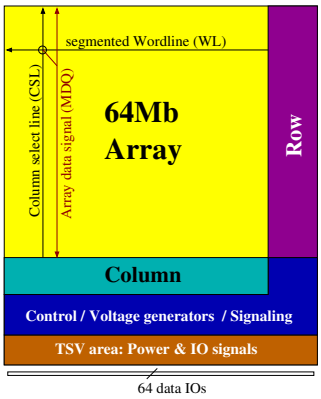

Fig. 3. Layout view of the 3D-DRAM core tile - 64Mbit

for the cell array are not considered. We also indicate the cell type in this table. The numbers for the 75nm and 65nm nodes are extracted from real measurements and simulations from commodity devices, the corresponding data for the 58nm and 45nm were extrapolated.

TABLE III
TECHNOLOGY AND PERFORMANCE DATA OF A 3D-DRAM CORE TILE

Tech. node	Cell area	Cell type	Area [mm^2]	Row t_{RAS}[ns]	t_{RCD} a) [ns]	Column t_{CCD}[ns]
75nm	$8F^2$	Trench	5.20	39.00	9.30	6.05
65nm	$6F^2$	bWL c)	3.54	27.10	7.54	5.42
58nm	$6F^2$	Stack	3.00	31.90	7.31	4.70
45nm	$4F^2$ b)	bWL c)	1.92	26.00	5.98	2.76

a) t_{RCD} = Row to column access delay
b) Using the latest available feasibility studies for this node from 2009
c) Buried WL technology developed at Qimonda

B. Wiring and TSV considerations

Wiring: As already mentioned commodity DRAMs are based on cost-optimized technologies. Thus, the number of interconnect layers and their electrical performance strongly differs from logic technologies. 2-3 Aluminum layers are available for routing with larger crosstalk and higher resistance compared to logic technologies. This strongly impacts the column performance (see CSL in Figure 3). Data bus, column forward (CSL) and backward wiring (MDQ - array data signal after sensing) are routed on the same metal layer (top-level metal in Aluminum). Active shielding is used for CSL and MDQ. Thus, power supply signals are routed in between. Active shielding can not be used for peripheral data buses due to density requirements. Table IV shows our parameters for the routing. They are identical for all technology nodes.

TABLE IV
WIRE DIMENSIONS: DATA BUS, CSL, MDQ ROUTING

Wire usage	Active shield	Width [μm]	Pitch [μm]	Heigth [nm]	Coverage of level below	C_{wire} 1 mm [fF]
Data bus	No	0.5	1.0	900	50%	404
MDQ	Yes	0.35	0.7	900	60%	252
CSL	Yes	0.35	0.7	900	60%	252

Wiring model of the TSV: Π-RC-elements are used to model the TSV and the connections between the stacked 3D-DRAM

layers. Investigations have shown that the accuracy of a Π-model for the TSV is sufficient and that inductance and capacitive coupling could be neglected. In addition to the TSV, wires to and from the DRAM tile have to be considered. The output of the DRAM tile is a tri-state buffer with 100 Ω output resistance. The input capacitance of a data input buffer of the tile is $10\,fF$. The lumped capacitance of a single data IO signal is $120\,fF$. Both wires are routed on top-level metal with a length of L=100 μm. The complete model is shown in Figure 4. Different TSV parameters (extrapolated from IMEC

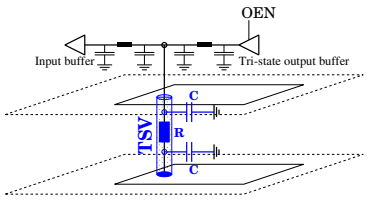

Fig. 4. The vertical connection: TSV wiring model

[4], [16] and Qimonda data), their electrical values and delays according to the model Figure 4 are shown in Table V. In contrast to [4] we used a TSV diameter value of 8 μm and 16 μm pitch for our exploration (marked bold in Table V). This diameter size is a good compromise between reported yield and density. Thus we do not consider redundant TSVs or yield losses.

TABLE V
ESTIMATED TSV PARAMETERS

Diameter [μm]	Pitch [μm]	Lumped Cap. [fF]	Resistance [mΩ]	Height [μm]	Delay τ [ps]
4	8	23	68	50	14.3
6	12	52	30	50	17.2
8	**16**	**94**	**17**	**50**	**21.4**
10	20	144	11	50	26.4

V. 3D-CHANNEL CONTROLLER

This section presents a short description of the 3D channel controller, with emphasis on the 3D features. As introduced in Section III the 3D channel controller handles all DRAM tasks, including memory initialization, refresh cycles and low power modes. In a multicore platform, the controller is interposed between the front-end and the stacked DRAM and optimized to take full advantage of the large DRAM data bus. The 3D channel controller design follows the JEDEC guidelines for DDR and SDR memory types. The main differences between the state-of-the-art and a 3D oriented controller are located mainly in the wide IO data bus, and in a simple and power efficient memory interface. The three stages of the controller provide respectively, data caching, synchronization/buffering and finally DRAM handling. Our controller is parametrizable to support different DRAM sizes and types (e.g. SDR or DDR). The objective of the first stage is to provide data-width adaptation between the front-end buses (processor side) and the synchronization queues, and to cache the portion of the row that has been accessed in the previous command. Since 3D-DRAM allows moving a big amount of data with a single command, this small cache acts like a temporary register to store parts of the write data, or to keep large chunk of data that has been read previously. This goal is achieved by fowarding in case of a "hit" the access to these registers. So the DRAM channel is bypassed. In case of a "miss" the command is transferred to the buffering stage for further elaboration. The caching operations run at the system clock. Therefore, a good speed up is achieved in case of cache hits. The synchronization stage is composed by dual clock FIFOs, which provide both buffering and an asynchronous communication approach between the DRAM and the front-end clock domain. Finally the last stage handles the front-end command translation and data preparation, sending to the DRAM the right command sequences. A status register block is used to track internal DRAM timings, avoiding timing violation during memory bank accesses. A thin and light physical interface is used to provide safe sampling and data shift during read and write command. In SDR mode, due low access time, the physical layer can be omitted, using the DRAM clock directly to perform safe data sampling.

VI. RESULTS

Referring to Figure 1, the complete 3D-DRAM stack is composed of different layers. Each *layer* is composed of so called *banks* and each bank is composed of *tiles*. A tile is the basic memory block which is cut out of the cell array of a commodity DRAM (see Section IV).

A. 3D-DRAM bank exploration

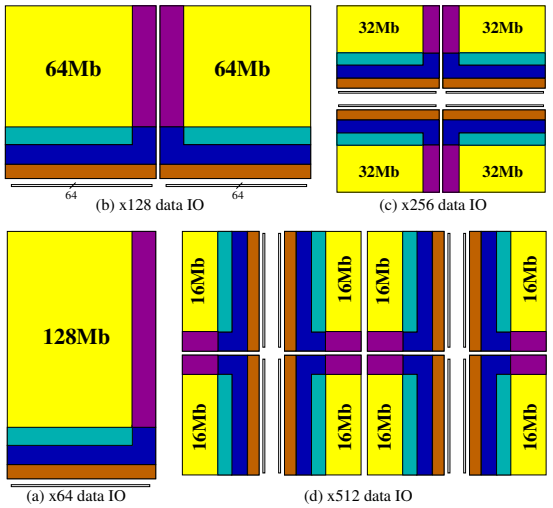

Fig. 5. Bank organization options for a 128Mbit bank

We limited the size of a single bank to 128Mbit, since this size corresponds to a typical bank in a commodity DRAM. Thus, it is well suited for comparison with commodity DRAMs. We consider tile sizes from 16Mb to 128Mb to compose the bank. 4 different organizations for a bank with a capacity of 128Mbit were investigated. These bank organizations are (see also Figure 5):

(a) 1x 128Mb tile with 64 data IOs
(b) 2x 64Mb tiles with 128 data IOs
(c) 4x 32Mb tiles with 256 data IOs
(d) 8x 16Mb tiles with 512 data IOs

Note that each tile has 64 data IOs. Thus the total number of IOs of a single bank is the number of tiles multiplied by 64. The IOs are modeled as described in Figure 4. We calculated the maximum possible frequencies f_{max} for each bank organization. The results are shown in Table VI. Due to space limitations only the maximum frequencies for the 45nm technology are listed. We see a similar trend for the other technology nodes. The throughput (= $f_{max} \cdot$ IOs) for the different bank organizations is also shown in this table.

An important and common metric for the area efficiency of a commodity DRAM device is the *cell efficiency* (CE), which is defined as the ratio of the cell area to the total chip area. This metric is also used by CACTI, but stays constant when scaling the technology. For actual commodity DRAMs, CE is in the range of 45% to 55% for a 1Gbit memory. We calculated the cell efficiencies of the different bank organizations, see Table VI. Obviously, the cell efficiency decreases with an increasing number of tiles in a bank since each tile adds an area overhead for the peripheral circuitries.

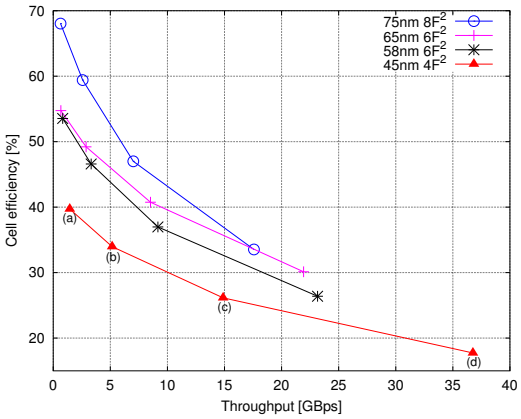

Fig. 6. Cell efficiency vs. throughput for different bank organizations

TABLE VI
THROUGHPUT AND CELL EFFICIENCIES FOR DIFFERENT BANK ORGANIZATIONS IN 45NM TECHNOLOGY

No.	Tile size [Mbit]	IO width (SDR)	CE 45nm [%]	f_{max} 45nm [MHz]	Throughput 45nm [GBps]
(d)	16	x512	17.73	588	36.76
(c)	32	x256	26.15	476	14.88
(b)	64	x128	33.97	357	5.58
(a)	128	x64	39.71	185	1.45

Figure 6 shows the resulting design space for the different bank organizations. In this graph we plotted the throughput versus cell efficiency for the four technology nodes. It can be seen that all technology nodes show the same trend. Obviously organization (d) has the highest throughput but the lowest cell efficiency. However the throughput of a single bank is not the primary optimization criteria. More important is the *energy efficiency* (EE=throughput/power=bandwidth/energy). Thus, we investigated an additional design space, in which we plot the energy efficiency versus cell efficiency.

This graph is shown in Figure 7. We omitted the 58nm and 65nm technology for reasons of clarity. Interesting in this graph is the trade-off between energy and area efficiency. We

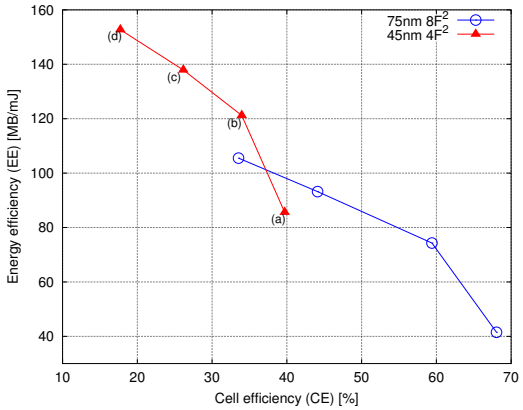

Fig. 7. Energy efficiency vs. cell efficiency for different bank organizations

can see that the product of cell efficiency and energy efficiency, i.e. $EE \cdot CE$, is maximal for bank organization (b) in all technologies.

B. 3D-DRAM Stack

A layer of the 3D-DRAM stack is composed of several banks and the complete stack of multiple layers (see Figure 1). We fix the size of the total stack to 1Gbit to permit a comparison with a state-of-the-art commodity DRAM. Five options exist to configure the layer organization and the number of layers of a 1Gbit stack:

- 1 layer with 8 banks
- 2 layers with 4 banks
- 4 layers with 2 banks
- 8 layers with 1 bank
- 16 layers with 0.5 banks (1 bank splitted on 2 layers)

For each bank - note that a bank size is 128Mbit - four different organizations exist as discussed in the previous chapter. Here, the important and interesting trade-off is horizontal wiring against vertical wiring via TSVs. E.g. the 1 layer/8 bank configuration consists of horizontal wiring only between the banks on the single layer. This configuration resembles most closely a 1Gbit commodity DRAM device. On the other extreme, the 8 layers/1 bank configuration has only vertical TSV wiring and no horizontal wiring. This trade-off impacts energy efficiency as well as cell efficiency. Thus, we use again the product of cell and energy efficiency to quantify the design space.

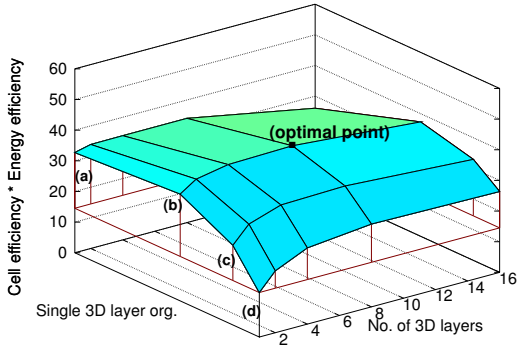

Fig. 8. $CE \cdot EE$ for various organizations of a 1Gbit 3D-stack in 45nm technology

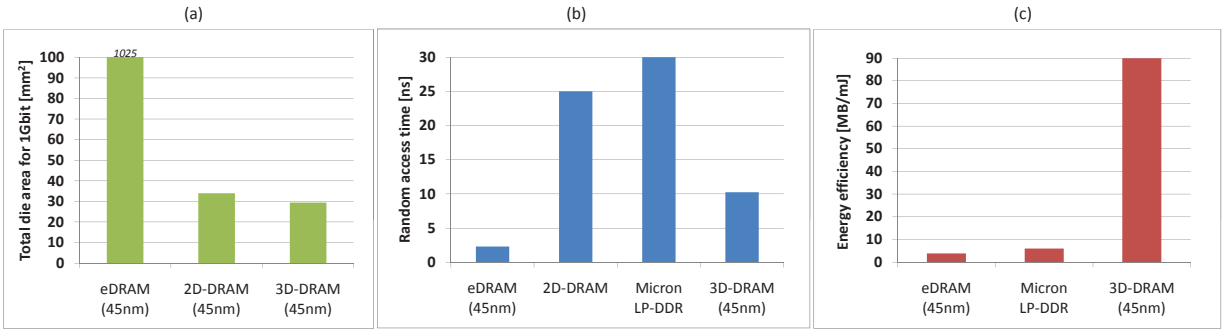

Fig. 9. Comparison to 1Gb commodity 2D-DRAMs, eDRAM and 1Gb, 4 banks, Mobile Low-Power DDR x16 SDRAM

The result of the exploration is shown in Figure 8 for the 45nm technology node. We see that the optimal number of layers is eight for each of the four different bank organizations. Solution (b) is the optimum one w.r.t. bank organization. It is important to mention that this organization was also the optimum in the design space exploration of a single bank only (see Figure 7). Again, the result for other technology nodes is the same.

C. Comparison to standard DRAMs and eDRAM

For the comparison of the optimized 3D-DRAM stack to a 3D-DRAM stack based on eDRAM, a commodity DRAM and a low power DRAM we chose the following memories:

- A commodity 1Gb, 8 bank DRAM from Qimonda.
- A 1Gb Low-Power DDR x16 SDRAM from Micron [15]. However the die area is not reported.
- A stack based on the eDRAM approach from IBM [10]. We extrapolated and scaled the published 2.39Mb macro to 1Gb. For the EE we assumed that one macro is active at the highest reported frequency and the others are in standby mode (at least hidden refresh).

We compared area, random access time and energy efficiency. The results are shown in Figure 9. This comparison shows the advantages of the 3D-DRAM approach. The 3D-DRAM stack is 15x more energy efficient than the 1Gb Low-Power SDRAM from Micron. Note that we subtracted the IO driver and termination power for this device before comparison. It also demonstrates that an eDRAM based approach is only feasible for moderate memory sizes.

VII. Conclusion

In this paper we presented a design space exploration for 3D-stacked memories. We investigated different layer organizations and number of layers for various DRAM technology nodes with most advanced DRAM cells To the best of our knowledge this is the first approach which combines circuit level simulations based on commodity DRAM devices and high level 3D-DRAM architecture exploration for various technologies. A 1Gbit 3D-stack was selected to allow a comparison with commodity memories. A parametrizable SystemC model was implemented which was linked to a 3D-DRAM memory channel controller for realistic IO interface traffic characterization. We demonstrated that an optimized 1Gbit 3D-DRAM stack has a 15x higher energy efficiency compared to a commodity 1Gbit Low-Power DDR SDRAM.

Acknowledgments

This work was supported in part by an EU FP7 Project Pro3D (GA n. 248776).

References

[1] W. A. Wulf and S. A. McKee, "Hitting the Memory Wall: Implications of the Obvious," *Computer Architecture News*, vol. 23(1), pp. 20–24, March 1995.

[2] D. Dutoit and A. Jerraya. (2010, July) 3d integration opportunities for memory interconnect in mobile computing architectures. Future Fab Issue 34. CEA-Leti MINATEC. pp. 38-45. [Online]. Available: http://www.future-fab.com/

[3] Anigundi, R. et al., "Architecture design exploration of three-dimensional (3D) integrated DRAM," in *Proc. Quality Electronic Design Quality of Electronic Design ISQED 2009*, 2009, pp. 86–90.

[4] M. Facchini, T. Carlson, A. Vignon, M. Palkovic, F. Catthoor, W. Dehaene, L. Benini, and P. Marchal, "System-level power/performance evaluation of 3d stacked drams for mobile applications," in *Proc. DATE '09. Design, Automation & Test in Europe Conf. & Exhibition*, 2009, pp. 923–928.

[5] D. H. Woo, N. H. Seong, D. L. Lewis, and H.-H. S. Lee, "An optimized 3D-stacked memory architecture by exploiting excessive, high-density TSV bandwidth," in *Proc. IEEE 16th Int High Performance Computer Architecture (HPCA) Symp*, 2010, pp. 1–12.

[6] EuroCloud Project. (2010, June) Energy-conscious 3D-Server-on-Chip for Green Cloud Services. Press Release. [Online]. Available: http://www.eurocloudserver.com/

[7] I. Loi and L. Benini, "An efficient distributed memory interface for many-core platform with 3D stacked DRAM," in *Proc. Design, Automation & Test in Europe Conf. & Exhibition (DATE)*, 2010, pp. 99–104.

[8] G. H. Loh, "3D-Stacked Memory Architectures for Multi-core Processors," in *Proc. 35th Int. Symp. Computer Architecture ISCA '08*, 2008, pp. 453–464.

[9] Hongbin Sun et al., "3D DRAM Design and Application to 3D Multicore Systems," *IEEE Design & Test of Computers*, vol. 26, no. 5, pp. 36–47, 2009.

[10] Klim, P. J. et al, "A 1 MB Cache Subsystem Prototype With 1.8 ns Embedded DRAMs in 45 nm SOI CMOS," vol. 44, no. 4, pp. 1216–1226, 2009.

[11] Tezzaron Semiconductor Corp. (2010, May) Octopus^{TM} 8-Port DRAM for Die-Stack Applications. Data sheet. [Online]. Available: http://www.tezzaron.com/memory/datasheets/

[12] S. J. E. Wilton and N. P. Jouppi, "CACTI: an enhanced cache access and cycle time model," *IEEE Journal of Solid-State Circuits*, vol. 31, no. 5, pp. 677–688, 1996.

[13] Denali Software Inc, "Databahn dram memory controller ip," 2009. [Online]. Available: https://www.denali.com/en/products/databahndram.jsp

[14] B. Jacob, with contributions by S. Srinivasan and D. T. Wang, *The Memory System: You Can't Avoid It; You Can't Ignore It; You Can't Fake It.* Morgan & Claypool Publishers, Jun. 2009, ISBN 978-1598295870.

[15] Micron Technology, Inc. (2009, July) 1Gb: x16, x32 Mobile Low-Power DDR SDRAM - Rev. K 07/09 EN. Data sheet. [Online]. Available: http://www.micron.com/

[16] Van Olmen, J. et al, "3D Stacked IC demonstrator using Hybrid Collective Die-to-Wafer bonding with copper Through Silicon Vias (TSV)," in *Proc. IEEE Int. Conf. 3D System Integration 3DIC 2009*, 2009, pp. 1–5.