Wide I/O or LPDDR? Exploration and Analysis of Performance, Power and Temperature Trade-offs of Emerging DRAM Technologies in Embedded MPSoCs

Mohammad Hossein Hajkazemi, Mohammad Khavari Tavana, Houman Homayoun
Department of Electrical and Computer Engineering
George Mason University
Fairfax, VA, USA
Email: {mhajkaze, mkhavari, hhomayou}@gmu.edu

***Abstract*—Wide I/O, the recent JEDEC DRAM standard, has created an opportunity for architects to overcome the "memory wall" challenge. 2.5D/3D integration enables Wide-IO to deliver high memory bandwidth and low latency for mobile applications. On the other hand, LPDDR3 was introduced to mainly address the power budget challenge in embedded MPSoCs. Employing either Wide I/O or LPDDR3 leads to different power, performance and thermal behavior of the system that necessitates a thorough analysis of both technologies. In this paper, we present a comprehensive analysis of the latency, performance, power and thermal behavior of these two emerging DRAM technologies in embedded MPSoCs. We conduct a comprehensive study to understand the impact of core count, core micro-architecture, and core-memory integration technology on the trade-offs that these two memory technologies offer. We show that while stacked Wide I/O outperforms LPDDR3 by as much as 7%, it increases the power consumption by 14%. To improve the power efficiency, we evaluate stacked LPDDR3, a DRAM design that is as high performance as Wide I/O - yet it is as power efficient as LPDDR3.**

Keywords—Wide I/O; LPDDR3; Performance; Power; Temperature

I. INTRODUCTION

In recent years, there have been significant innovations in portable platform industry. Smart handheld devices including tablets and Smartphones have taken over the electronic market [1]. These platforms are mainly relying on portable SoC computers such as TI OMAP, Nvidia Tegra and Samsung Exynos, which are low power – yet high performance. To keep up with uninterrupted rapid growth in mobile platform industry, mobile application processors (APs), which are the heart of embedded MPSoCs, are becoming more complex, equipped with larger number of processing cores to respond to the higher performance demand of mobile applications. For example, *Exynos 7 octa,* embedded in *Galaxy Note4,* uses a big and a little quad-core processor [3], TI OMAP, embedded in Amazon Kindle Fire, employs a dual-core processor with 1.7 GHz frequency, and Nvidia Tegra K1 in Microsoft Surface uses four ARM A9. While core count is almost doubling every two years in these platforms, main memory DRAM capacity scales at a slower pace and doubles only every three years [25]. In addition, DRAM bandwidth required by applications is growing rapidly, almost doubles every four years [4] [10]. On the other hand, increasing core counts in embedded MPSoCs further increases the bandwidth demand. As a result, memory bandwidth per core is insufficient to meet the demands of future mobile applications. Therefore, in portable computer era, main memory performance is becoming an increasingly important contributor to the overall system performance, especially due to the so-called memory wall challenge.

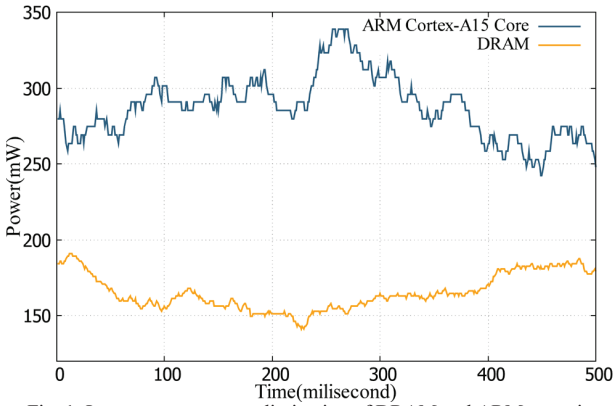


Fig. 1. Instantaneous power dissipation of DRAM and ARM cores in OMAP5 running KNN benchmark.

Besides higher performance, lower power dissipation is another factor that should be considered in portable platforms. A device with high power consumption results in shorter battery lifetime. In addition, since using an expensive and large heats-ink is not feasible in such platforms, high power dissipation can lead to thermal issues as well. Therefore, it is necessary to make portable platforms as low power as possible.

Among embedded MPSoC's components, DRAM is one of the major power consumers. For example, our measurements on OMAP5 SoC platform with DDR3L DRAM reported in Fig. 1 demonstrate that DRAM's power consumption is comparable to core power. Therefore using low-power DRAM can reduce the total power consumption of the embedded MPSoC considerably.

The increasing demand for low power DRAM has driven JEDEC to standardize LPDDRx for mobile platforms. LPDDR2/LPDDR3, compared with older generation of low-power DRAMs including DDR3L, removes on-die termination

978-1-4673-7166-7/15/$31.00 ©2015 IEEE

(ODT) and applies lower supply voltage [5]. This results in 10% and 67% typical idle and self-refresh power reduction [5].

Using stronger APs leads to higher performance (i.e. higher resolution display or shorter response time for handheld devices) only if the memory can keep up with core processing speed and not becomes a bottleneck [1]. However, more complex cores result in memory congestion and therefore exacerbate the memory-wall problem [18]. Using 3D-integration, JEDEC has addressed this issue by standardizing Wide I/O DRAM protocol [6-7]. Wide I/O consists of four DRAM dies vertically connected using Through Silicon Via interconnect (TSV) and micro bumps [22]. More number of channels, wider data bus and faster interconnects makes Wide I/O DRAM a high-bandwidth memory. Moreover, Wide I/O uses similar low-power techniques as LPDDR2/LPDDR3 does and is expected to be as low power as LPDDR3 [5][10]. In addition, it removes delay lock loop (DLL) to achieve lower static power [8]. Furthermore, due to low-capacitance TSVs, the I/O power consumption reduces drastically [9].

Unlike high-performance platforms, due to high-cost and large footprint, using high-end thermal management technologies (e.g. heat-sink) in portable platforms is not practical. Therefore, despite consuming low power, the high power density of Wide I/O could potentially create thermal issues. Stacking Wide I/O on the processor can even exacerbate the power-density problem.

Recent works have mainly focused on power and performance optimization of Wide I/O and LPDDR3 [8][26]. There are also a number of recent research on LPDDRx and Wide I/O power and performance modeling [9][23][24]. However, to the best of our knowledge, this is the first paper that conducts a comprehensive analysis on a wide range of architectures to compare the two technologies from performance, power and temperature points of view in portable platforms. We consider three different packaging to study the two technologies including, off-chip LPDDR3, 2.5D Wide I/O which integrates the memory and the processor through the use of interposer, and 3D Wide I/O where stacks the memory on the processor. We also propose and analyze stacking LPDDR3 on processor die using 3D-integration (stacked LPDDR3). We observe the following key findings:

- For future embedded MPSoCs with simple in-order core micro-architecture (e.g. ARM Little, Cortex-A7), stacked LPDDR3 technology is more power and performance-efficient compared to Wide I/O, however Wide I/O becomes more competitive for MPSoCs with complex out-of-order core (e.g. ARM Big, Cortex-A9) micro-architectures. Therefore, considering the integration cost, in embedded MPSoCs with complex micro-architectures, 2.5D Wide I/O is a more preferable choice over 3D Wide I/O.
- We observe a marginal performance improvement by using eight channels over four channels in Wide I/O, across all designs. Therefore, as in MPSoCs the cost is one of the major concerns, employing 8-channel DRAM is not a cost-effective solution.
- Across almost all studied applications and microarchitectures, Wide I/O consumes higher power compared to LPDDR3, contrary to what has been shown in latest industry reports [5][10].
- LPDDR3 has noticeable advantages over Wide I/O in terms of PDP (Power Delay Product) for in-order cores. However, Wide I/O outperforms LPDDR3 for out-of-order cores, as more memory requests are generated and can be serviced simultaneously.

In summary the main contributions of this paper are as follows:

- We explore and analyze the performance, power and thermal behavior of LPDDR3 and Wide I/O memory technologies in embedded MPSoCs with various processor configurations, core micro-architectures, core count and different memory packaging, to compare the two technologies.
- Exploiting the low power consumption of LPDDR3 and high bandwidth of 3D-integration we carefully study and analyze stacked LPDDR3 in which the performance is becoming comparable to Wide I/O while resulting in lower power consumption and improved thermal behavior.
- We show that unlike high-end platforms where the core is the dominant power consumer, in embedded MPSoCs with Wide I/O the DRAM memory subsystem and the processor have almost comparable power footprint contributing to a similar extent to the total power of the MPSoC. Therefore, to overcome the thermal challenges in stacked Wide I/O, the conventional techniques that mainly focus on processor core power are insufficient, calling for new thermal management techniques targeting the DRAM and processor, simultaneously.

The rest of this paper is organized as follows: Section II presents the background. Section III introduces the methodology we use for DRAM technology analysis. Section IV presents the analysis results. Section V introduces our proposed stacked LPDDR3 and Section VI analyzes the studied architectures in detail. Section VII introduces some related works. Finally, Section 7 concludes the paper.

II. BACKGROUND

LPDDR DRAM was introduced in embedded platforms to overcome the power challenge caused by DDR family. The power efficiency proposed by LPDDR is mainly achieved by using lower supply voltage. In addition, other techniques are employed to save more energy including deep power down mode, low frequent refreshing due to lower temperature, DLL elimination, and optional ODT [12].

Although LPDDR family addresses the power challenges of DDR family for mobile devices, it still suffers from the pin count constraint [8]. In addition, latest works have shown that bandwidth requirement in handheld devices increases rapidly [10] [25]. Therefore, the pin constraint can create a bottleneck for high bandwidth-demanding applications in portable devices. Using 3D-integration, the emerging Wide I/O technology was proposed to address the bandwidth challenge.

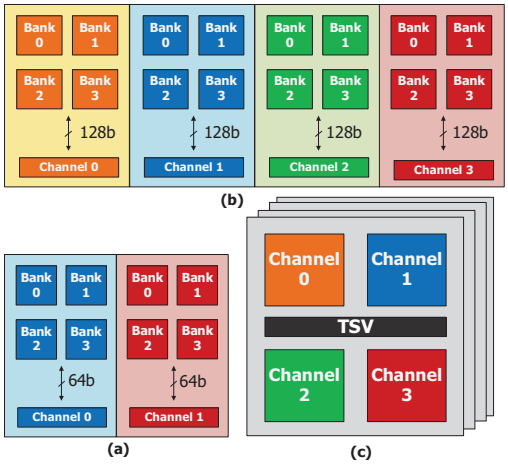


Fig. 2. (a) LPDDR3 architecture, (b) Wide I/O architecture, (c) Top view of a 4-Channel Wide I/O.

TSV is the key enabler for 3D-integration [32]. TSV improves the capacitance per connection and consequently reduces connection power consumption by as much as 6X [11]. Moreover, it shortens the connection length by 200X [11]. Also, TSV provides higher number of connections for Wide I/O, leading to higher number of memory channels and therefore higher throughput. However, all of these benefits do not come for free. In fact, using TSVs and 3D-integration in Wide I/O comes with almost three times more cost per bit than LPDDR3 [21].

Fig. 2 (a) and (b) show the LPDDR3 and Wide I/O architectures that we study in this paper. Fig. 2 (c) shows the top view of a 4-channel Wide I/O including channels and TSV areas location. As Fig. 2 (a) and (b) report, Wide I/O has two times more channels than LPDDR3. In addition, the data bus width per channel in Wide I/O is two times higher compared with the bus width in LPDDR3. As shown in Fig. 2 (c), Wide I/O has four dies stacked vertically. Each die consists of memory cells and TSV area. Memory cells are accessed by the memory controller through the channels. The memory controller is implemented in a separate logic die placed beneath the DRAM cell layers.

III. Methodology

A. Studied Architectures

Based on the current trend of increasing core count in portable SoCs [3], in this work we model a 4-core and an 8-core embedded MPSoC with a 4 GB memory as the target platform. The processor is modeled using 32nm process technology. For each studied MPSoC we model two different architectures: an in-order core similar to ARM Cortex-A7, and a 2-way out-of-order superscalar core similar to ARM Cortex-A9. Table I summarizes the detailed micro-architectural parameters we use in this work.

TABLE I. Application Processor Parameters

	Config 1	Config 2
Core Clock	2GHz	2GHz
Core Count	4,8	4,8
Issue, Commit Width	1	2
INT & FP Instruction Queue	1	16
INT Reg, FP Reg	16	32
L1 Cache	16KB, 2-way	32KB, 4-way,
L2 Cache	256KB	512KB

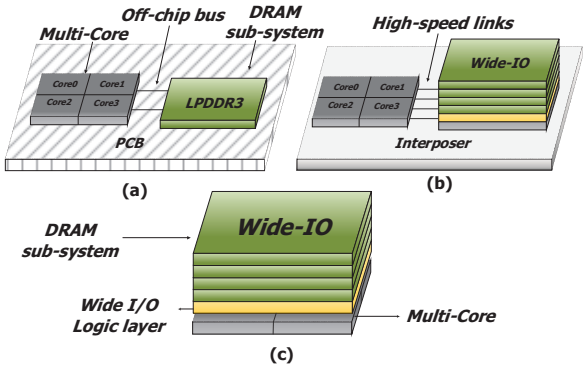


Fig. 3. Packaging and integration technologies for LPDDR3 and Wide I/O studied in this work.

In this work we study three different packaging and integration technologies for LPDDR and Wide I/O, as is illustrated in Fig. 3. As Fig. 3 shows, while LPDDR3 is off-chip, Wide I/O is embedded with the processor cores using either an interposer (2.5D) or 3D TSV interconnects (stacked). For LPDDR3, the memory controller resides on the same die as the processor core does. However, for Wide I/O, in both 2.5D and 3D, a separate logic die is integrated beneath the DRAM cell layers as a memory controller.

We consider two channels for LPDDR3 [5]. However, as 3D-integration allows more number of channels, for Wide I/O we assume a four and an eight-channel design [8]. We consider four banks per channel in all studied designs. The data bus width for LPDDR3 and Wide I/O is set to 64 and 128 bits respectively.

Fig. 4 gives an overview of our simulation methodology. For simulation we integrated SMTSIM [13] with DRAMSim2 [14]

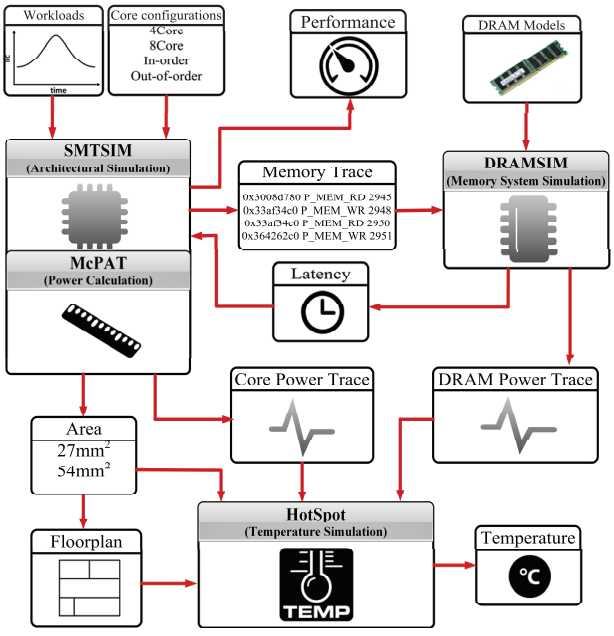


Fig. 4. Methodology overview.

TABLE II. WORKLOAD'S APPLICATIONS

Type	#W	Workload
4-core	1	ammp, applu, art, bwave_06
	2	bzip2_06, cactusADM_06, parser, facerec
	3	gcc_06, gobmk_06, lbm_06, leslie3d_06
	4	libquantum_06, mcf, mg.big, omnetpp_06
	5	perim_big, soplex_06, sp.big, swim
	6	vortex_3, perlbench_06_diffmail, crafty, namd_06
8-core	1	ammp, applu, libquantum_06, bwaves_06, bzip2_06, cactusADM_6, parser, facerec
	2	gcc_06, gobmk_06, lbm_06, leslie3d_06, libquantum_06, mcf, mg.big, omnetpp_06
	3	perim_big, soplex_06, sp.big, swim, treeadd, twolf, vpr_place, zeusmp_06
	4	ammp, gcc_06, perim_big, applu _ gobmk_06, soplex_06, parser, lbm_06
	5	sp.big, bwaves_06, leslie3d_06, swim, bzip2_06, libquantum_06, treeadd, cactusADM_06
	6	mcf, twolf, cg.big, mg.big, vpr_place, facerec, omnetpp_06, zeusmp_06

and McPAT 0.8 [15]. SMTSIM is used to simulate the processor architectures. McPAT is used for processor power simulation. DRAMsim2 is extensively modified to simulate different memory architectures including LPDDR3 and Wide I/O and different packaging technologies. DRAMsim2 is also equipped with a power profiler to generate the memory subsystem power trace. Table II shows the 12 workloads we use for 4-core and 8-core platforms. The benchmarks used in the workloads are selected randomly. We fast-forward benchmarks for 1B instructions and then run for another 1B.

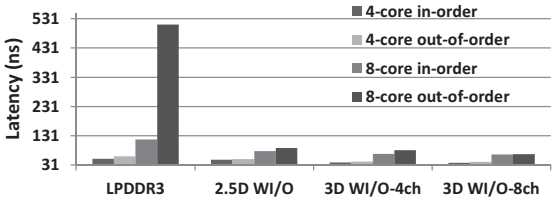
B. Performance, Power and Thermal Model

In this paper, we carefully model LPDDR3 and Wide I/O architectures to account for their architectural as well as timing and power characteristics. We use the timing and power model parameters reported in [9]. Table III reports the main timing and power model parameters we use for LPDDR3 and Wide I/O. Some important parameters are defined as follows: CL, BL, RAS and TAW account for column access strobe, burst duration, row access strobe and two bank activation window time. Moreover, iddxs are different current consumption per access in different conditions. For example idd4w and idd4r are writing and reading energy per access respectively.

Since DRAMSim2 does not report the memory controller power consumption, similar to [29] and [30], we use the reported data in [16] to estimate the controller power in 2.5D and stacked Wide I/O 3D. According to the data reported in [16], the controller dissipates 1.8 more power than DRAM cells.

TABLE III. LPDDR3 AND WIDE I/O MODEL PARAMETERS

Timing model parameters								
	Clk (MHz)	CL (ns)	BL (ns)	RAS (ns)	RCD (ns)	RRD (ns)	RC (ns)	TAW (ns)
LPDDR3	800	12	8	36	15	8	48	40
Wide I/O	200	3	8	12	5	3	16	14
Power model parameters								
	idd0 (mA)	Idd02 (mA)	idd2n (mA)	idd3n (mA)	idd4r (mA)	idd4w (mA)	idd52 (mA)	vdd (V)
LPDDR3	15	80	2	2	5	10	160	1.2
Wide I/O	6.06	21.82	0.16	0.58	1.82	1.82	48.34	1.2



5. Average memory access latency of the studied DRAM technologies used in studied architectures.

Table IV reports the thermal configuration parameters used for HotSpot simulation [17]. We assume that in our studied memory sub-system, the power dissipation is distributed evenly among the DRAM layers. Based on McPAT results, the size of a 4-core and an 8-core architecture are found to be 27.84 mm^2 and 55.68 mm^2, respectively.

IV. ANALYSIS

In this section, we report the performance, power and thermal behavior of the studied architectures. We analyze performance-power tradeoffs of the entire system as well as main memory for different DRAM technologies. We also study the impact of increasing the core count and DRAM channels (in 3D Wide I/O) on performance, power and temperature of the DRAM as well as the processor. We report the results for four distinct memory technologies and packaging including LPDDR3, 2.5D Wide I/O, 3D stacked Wide I/O with four channels and 3D stacked Wide I/O with four and eight channels.

A. Performance Analysis

1) Memory access time: The average memory access latency for different memory technologies including Wide I/O and LPDDR3 is shown in Fig. 5. The results are averaged across all the workloads.

As expected, stacked Wide I/O has the shortest memory access latency while LPDDR3 has the longest. 2.5D Wide I/O stands in between due to longer and shorter CL latency compared to 3D Wide I/O and LPDDR3, respectively.

As Fig. 5 shows, designs with out-of-order cores have noticeably higher memory access latency compared to designs with in-order cores, for both 4-core and 8-core cases and across all memory technologies and configurations. Moreover, the difference between the memory latencies becomes larger by increasing the number of cores from four to eight; especially when we compare Wide I/O to LPDDR3 for 8-core design. This

TABLE IV. THERMAL CONFIGURATION PARAMETERS USED IN HOTSPOT

Parameters	Values
Processor and DRAM Die thickness	0.15 mm, 0.05mm
Silicon thermal conductivity	100 W/mK
Silicon specific heat	1750 KJ/m^3K
Spreader thickness	0.5 mm
Spreader thermal conductivity	400W/mK
Interface material conductivity	4 W/mK
Heatsink thickness	0.1mm
Heatsinkconductuvity	400W/mK
Ambient and initial temperature	45°C
Sampling interval	1 ms

is because a larger number of cores generate greater number of memory requests resulting in higher bandwidth demand. Higher bandwidth utilization leads to longer memory latency [18] [29]. The relation between the memory latency and the bandwidth utilization is somewhat linear for low bandwidth utilization [18] [29]. However, for high bandwidth utilization the latency grows exponentially due to queuing effect [18] [29]. Therefore, in LPDDR3, for which the requests are serviced slowly, higher utilization results in queuing effect. As a result, higher latency growth is observed.

We also observe that Wide I/O does not benefit much from doubling the number of channels in 4-core design; however as the number of cores increases to 8, the difference between 4 and 8 channels is becoming larger as more memory requests are generated.

2) Processor performance analysis: Fig. 6 presents the performance improvement of various designs using Wide I/O with different packaging technologies as compared to LPDDR3.

Core complexity is one of the main factors that affect the processor performance. While out-of-order processors can hide DRAM latency due to the out-of-order execution capability, they generate more memory requests and consequently could result in higher DRAM congestion and longer memory access latency. In contrast, due to in-order execution, the DRAM latency cannot be hidden in in-order processors; however memory congestion occurs less frequently as a result of lower number of memory requests. Therefore, the high bandwidth of Wide I/O has less impact on in-order processors' performance and is more beneficial in out-of-order processors. As Fig. 6 shows, out-of-order design gains more benefits in terms of IPC from Wide I/O compared to in-order design.

Besides core complexity, core count is another competing factor that affects performance. The more cores the embedded MPSoC employs, the higher DRAM bandwidth it requires to satisfy higher number of generated DRAM requests. On the other hand, as discussed in Section I, Wide I/O has been introduced for bandwidth-demanding applications. Therefore, as shown in Fig. 6, the performance improvement is higher for the 8-core design compared to the 4-core design across all memory technologies. Moreover, comparing 4-core out-of-order processor with 8-core in-order processor in Fig. 6, we observe that Wide I/O is more beneficial for designs with more complex cores than for design with larger number of cores.

Fig. 6. Embedded MPSoC performance improvement, using different DRAM technology compared to LPDDR3.

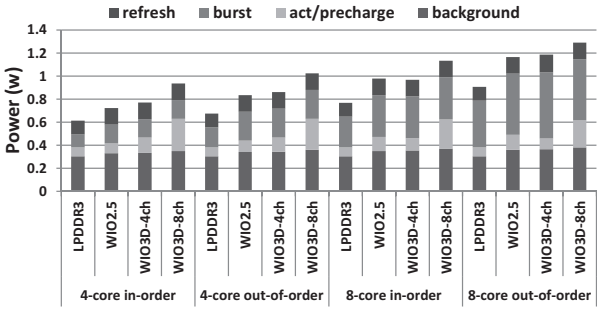


Fig. 7. Power breakdown for different DRAM technologies.

We also observe only a slight performance improvement by using eight channels over four channels, across all designs. As the number of channels increases, DRAM can service more parallel accesses and therefore the chances of queuing and memory congestion is reduced. However this does not comes for free and in fact incurs significant cost to DRAM, in particular in embedded systems where cost is a prime concern. Therefore, employing 8-channel DRAM is not a cost-effective solution.

B. Power Analysis

In this section we evaluate power consumption of various memory technologies. To have a better understanding of different DRAM technologies' impact on power consumption, we break down the DRAM power to background, active/pre-charge, burst and refresh power. The background power shows the energy dissipation used to operate the control circuitry [19]. Depending on whether or not the device is in low power mode, the background power varies. The active/precharge power demonstrates the power consumed to activate and precharge a row within the data array. The burst power accounts for data transmission on the data bus, and finally the refresh power shows the energy dissipated for refreshing the rows. The results are reported in Fig. 7.

As shown in Fig. 7, for both in-order and out-of-order designs with four and eight cores, LPDDR3 achieves the lowest power consumption while stacked Wide I/O with eight channels results in the highest power. 2.5D and stacked Wide I/O with four channels consume almost similar power and lies in the middle of the spectrum. Although the reported results in [5] and [10] predict less power consumption for Wide I/O, and also Table III reports smaller values for Wide I/O's power model parameters such as read and write consuming current compared to LPDDR3's, due to shorter access time (see Fig. 5), Wide I/O services more number of requests in a given time which results in higher power consumption.

We also observe that for more complex core micro-architecture and for larger number of cores, a higher power is consumed. This is in part, due to larger number of memory requests and consequently larger burst power. Moreover, As Fig. 7 illustrates, under these circumstances, the refresh power remains almost constant, while the background power increases slightly. This is happening because of higher DRAM utilization that causes DRAM to stay in low-power mode less frequently [19]. Moreover, high DRAM utilization causes more

activation/precharge operation and thus results in higher act/precharge power consumption [19].

C. Temperature Analysis

In this section, we study the thermal behavior of the DRAM technologies discussed in this paper. The steady state temperature of the hottest spot of the studied DRAM technologies are reported in Fig. 8.

As Fig. 8 illustrates, LPDDR3 has the lowest temperature among all DRAM technologies, core micro-architecture, and core counts, as it is located 'off the chip'. 2.5D Wide I/O comes next. As 2.5D Wide I/O is integrated with the core die using interposer, it benefits from the heat-sink. However, stacked layers and higher power consumption of the 2.5D Wide I/O results in up to 6.9ºC higher temperature compared to LPDDR3.

As shown in Fig. 8, the temperature of stacked Wide I/O is up to 5.4ºC higher than that of 2.5D Wide I/O. This temperature rise is due to the heat generated by the processor layer. Moreover, Fig. 8 shows that stacked wide I/O with eight channels has the highest temperature among the studied DRAMs due to its high power consumption. The temperature rises up to 5.7ºC compared to stack Wide I/O with four channels.

As Fig. 8 illustrates, employing processor with more complex micro-architecture, i.e. out-of-order, results in temperature rise in all DRAM designs due to higher power density. However, increasing the number of cores has negligible impact on temperature. In fact, as the number of cores increases, the area also increases, and therefore the power density remains almost constant.

Furthermore, comparing out-of-order 4-core design with in-order 8-core design shows a slight temperature rise in both LPDDR3 and 2.5D Wide I/O. However the temperature drops by 2ºC and 4ºC for 4-channel and 8-channel stacked Wide I/O DRAM, respectively. That is because in contrast to LPDDR3 and 2.5D Wide I/O cases where the temperature is decided by the DRAM, the power density of the embedded MPSoC is decided by both the DRAM power as well as the processor power when Wide I/O is stacked. In this case, while the DRAM layer size is doubled, the memory accesses and therefore the burst and active/recharge power remains almost the same. As a result, while the power density of the cores almost remains constant, the power density of the Wide I/O decreases compared to LPDDR3 and 2.5D Wide I/O. Therefore the power density of the entire embedded MPSoC decreases.

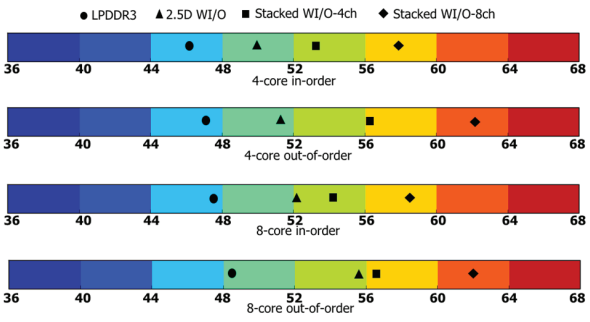


Fig. 8. Steady state temperature of different DRAM technologies.

Another interesting observation when DRAM is stacked with embedded core is, unlike the high-end core, the thermal behavior is decided by both the core power as well as DRAM power, as DRAM power is comparable to core power. In high-end MPSoCs with stacked DRAM, the DRAM's power consumption has a small contribution in MPSoC's total power, which causes the DRAM temperature to be mainly decided by the core power [20]. Our observation in embedded MPSoCs, and across various micro-architectures (in-order and out-of-order) shows that the stacked Wide I/O power is comparable with core power. Therefore, applying techniques that are proposed recently to reduce the processor temperature to reduce DRAM temperature such as the one in [20] and [7] are not sufficient to significantly affect the stacked DRAM temperature.

V. Stacked LPDDR3

As shown in Fig. 7, regardless of core complexity and core count in embedded MPSoCs, LPDDR3 consumes lower power compared to Wide I/O. Moreover, the relative cost per bit of LPDDR3 is three times less than Wide I/O [21]. However, as Fig. 5 shows, Wide I/O offer higher performance compared to LPDDR3.

In order to benefit from both the lower power of LPDDR3 and the higher performance of 3D stacking (for Wide I/O), we evaluate stacked LPDDR3 where LPDDR3 memory is integrated with the core in 3D. Die-stacked DRAM has been explored and validated in many prior works [18][20][31]. However, to the best of our knowledge, this is the first work that proposes and analyzes stacked LPDDR3 in embedded MPSoCs. While we consider the same power model parameter of LPDDR3 reported in Table III for the stacked LPDDR3, similar to [18] and [29] which assume smaller read and write latency for the stacked DRAM, we reduce the column access strobe to model fast interconnects. We assume the proposed stacked LPDDR3's CL is equal to that of stacked Wide I/O. Since we do not modify the LPDDR3' architecture, we assume the rest of the timing parameters remain constant. The physical interface, bus width, capacity and other basic specifications of stacked LPDDR3 are the same as LPDDR3. Fig. 9 compares stacked LPDDR3 and stacked Wide I/O in terms of power, performance and temperature.

As shown in Fig. 9, compared to stacked Wide I/O, stacked LPDDR3 has the least performance loss in in-order 4-core design. The performance loss increases as the core count and the micro-architecture complexity of the cores increases. The results show that for large number of cores and for out-of-order superscalar cores, stacked Wide I/O still provides noticeable

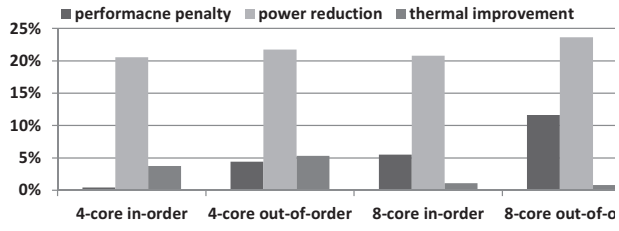


Fig. 9. Performance, power and thermal comparison of stacked LPDDR3 relative to Wide I/O with eight channels.

benefits in terms of performance compared to stacked LPDDR3. However, comparing the power consumption results show that, as the number of cores and micro-architecture complexity of cores increases, stacked LPDDR3 provide higher power reduction benefits compared to stacked Wide I/O.

Also, as Fig. 9 shows, despite having an almost equal power reduction for 4-core and 8-core designs, the thermal improvement is noticeably higher in stacked LPDDR3 compared to Stacked Wide I/O for 4-core design. According to the results in Fig. 8, off-chip LPDDR3 temperature is more comparable to Wide I/O for 8-core design. Our observation reports that the stacked LPDDR3 steady state temperature for 4-core in-order, 4-core out-of-order, 8-core in-order and 8-core out-of-order designs are 50.2oC, 51.4oC, 53.4oC and 56.4oC respectively.

VI. Qualitative Analysis

In this section we present a qualitative analysis of all studied DRAM technologies. We compare DRAM power, latency, and temperature, core power and total application performance. In addition, we report the processor PDP (Power Delay Product) to find the best DRAM technology for the studied architectures. In Fig. 10, we use spider graph to present this comparison. To have a better understanding, the graph parameter values are scaled to five equal levels where the first level shows the smallest and the fifth represents the largest. Polygons inside the spider chart show different DRAM technologies. Since smaller values of power, latency, temperature and PDP are desired, the closer the polygons get to the center of the chart, the better design they present. Figs. 10 (a), 10 (b), 10 (c) and 10 (d) show the results for 4-core in-order, 4-core out-of-order, 8-core in-order, and 8-core out-of-order MPSoCs. We present our qualitative comparison of results in form of key finding as follows:

Key Finding 1: Using stacked LPDDR3 results in the lowest PDP meaning that this design is the most power and performance-efficient for embedded MPSoCs with four in-order cores. As the core count increases to 8 and the micro-architecture complexity increases to out-of-order superscalar, the stacked Wide I/O becomes the most efficient design.

Key Finding 2: For 8-core out-of-order MPSoCs, 3D and 2.5D Wide I/O polygons almost overlap, meaning that both offer almost similar power, performance and thermal efficiency. Moreover, both 3D and 2.5D Wide I/O comes with the lowest PDP. Since using interposer is a more cost-effective choice than expensive 3D stacking, 2.D wide I/O is more preferable as core complexity and core count increases in future embedded MPSoCs.

Key Finding 3: A marginal performance boost is obtained by using eight channels over four channels in Wide I/O. Consequently, in platforms where the cost is one of the major concerns, employing 8-channel DRAM is not a cost-effective solution.

Key Finding 4: Contrary to our expectation and latest industry reports [5][10], 2.5D and stacked wide I/O are shown to have the highest power consumption in all studied configurations. However, LPDDR3 (2D and 3D) always offer the lowest power consumption across all core configurations (see Table III).

VII. Related Works

LPDDR3 [5] and Wide I/O [6] are the two state-of-the-art memory technologies introduced by JEDEC to improve the energy-efficiency and bandwidth of mobile DRAMs. Several recent works have proposed different implementation of LPDDR3 and Wide I/O [27][28]. There are also a number of recent works focused on system and architecture level optimization of power, performance and temperature in LPDDR3 and Wide I/O [8][26]. For example [8] proposes a new architecture for Wide I/O referred as 3D-SWIFT to achieve high access parallelism. [26] improves the Wide I/O refresh power consumption by adapting the refresh period based on the temperature variation.

Furthermore, some research have been done on modeling and characterization of emerging DRAM technologies. Chandrasekar et al. propose a comprehensive system and circuit level power model for both LPDDR3 and stacked Wide I/O [9]. Matthias et al. suggest a new methodology to model the backend of stacked Wide I/O memory systems using TLM modeling. Hansson et al. propose an event-base memory controller model capable of simulating emerging memory technologies including LPDDR3 and Wide I/O [2].

Manil et al. analyze different DRAM technologies with worst-case bandwidth scenario in real-time systems [24]. Based on their analysis they propose a methodology to select an appropriate DRAM technology. However, [24] has not investigated the impact of different DRAM technologies on the system performance and power. Moreover it has not studied the thermal challenges in MPSoCs with stacked DRAM.

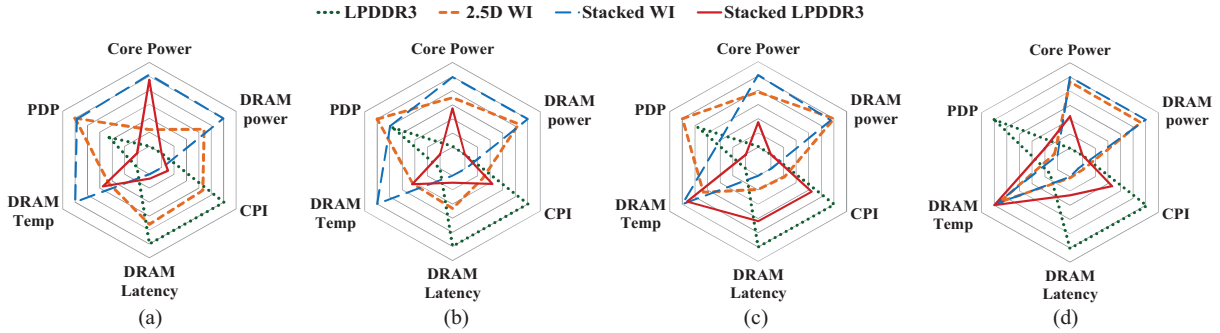


Fig. 10. Qualitative analysis of studied DRAM technologies with different core configuration and core count.

To the best of our knowledge this research is the first work that investigates performance, power and temperature trade-offs of the two important emerging DRAM technologies, LPDDR3 and Wide I/O, across a wide range of core and micro-architecture configurations and various applications, to understand the implication of emerging DRAM technologies choice in future embedded MPSoCs.

VIII. Conclusion

In this work, we present a comprehensive analysis of the latency, performance, power and thermal behavior of the two emerging DRAM technologies including LPDDR3 and Wide I/O in embedded MPSoCs across a wide range of core configurations, micro-architectures and packaging technologies. We demonstrate that in contrast to our expectation and latest industry reports [5][10], LPDDR3 has low-power consumption advantage over Wide I/O. However, it suffers in performance. Therefore, we propose and analyze 3D stacked LPDDR3 to benefit from both the lower power feature of LPDDR3 and the higher performance of 3D stacking (for Wide I/O). Our analysis shows that while 3D stacked LPPDR3, compared to 3D Wide I/O, is more performance and power-efficient for embedded MPSoCs with simple in-order core micro-architecture, it becomes less competitive for more complex our-of-order micro-architecture. Furthermore, we show that stacked Wide I/O has noticeable advantage over LPDDR3 as the number of cores increases in future MPSoCs.

References

[1] S. Yang, J. Pyo, Y. Shin, and J. C. Son, "A 1.6 GHz quad-core application processor manufactured in 32 nm high-k metal gate process for smart mobile devices." Communications Magazine, IEEE 51, no. 4 (2013): 94-98.

[2] A. Hansson, N. Agarwal, A. Kolli, T. Wenisch, and A. N. Udipi, "Simulating DRAM controllers for future system architecture exploration." In Performance Analysis of Systems and Software (ISPASS), 2014 IEEE International Symposium on, pp. 201-210. IEEE, 2014.

[3] G. Blake, R. G. Dreslinski, and T. Mudge, "A survey of multicore processors." Signal Processing Magazine, IEEE 26.6 (2009): 26-37.

[4] O. Mutlu, "Memory scaling: A systems architecture perspective." (IMW), pp. 21-25. IEEE, 2013.

[5] M. Greenberg, "LPDDR3 and Wide I/O DRAM: Interface Changes that give PC-Like Memory Performance to Mobile Devices", Memcon 2012, San Jose.

[6] JEDEC Solid State Technology Association, "JEDEC Standard: Wide I/O Single Data Rate Specification", https://www.jedec.org/sites/default/files/docs/JESD229.pdf, Dec. 2011.

[7] J. Meng, K. Kawakami, and A. K. Coskun, "Optimizing energy efficiency of 3-D multicore systems with stacked DRAM under power and thermal constraints.", (DAC), pp. 648-655. 2012.

[8] T. Zhang, C. Xu, K. Chen, G. Sun, and Y. Xie, "3D-SWIFT: a high-performance 3D-stacked wide IO DRAM.", (GLS-VLSI), pp. 51-56. ACM, 2014.

[9] K. Chandrasekar, C. Weis, B. Akesson, N. Wehn, and K. Goossens, "System and circuit level power modeling of energy-efficient 3D-stacked wide I/O DRAMs." (DATE), pp. 236-241, 2013.

[10] M. Greenberg, "How do I Decide? Is LPDDR3 or Wide I/O the right memory technology for my next smartphone and tablet?", Mobile Forum Taiwan and Korea, 2012.

[11] S Bansal, M Greenberg, "3D-IC is Now Real: Wide-IO is Driving 3D-IC TSV", EDPS, California, 2012.

[12] JEDEC Solid State Technology Association," JEDEC Standard: Low Power Double Data Rate 3 (LPDDR3)", http://www.jedec.org/sites/default/files/docs/JESD209-3B.pdf, Aug 2013.

[13] D. M. Tullsen, "Simulation and modeling of a simultaneous multithreading processor." In The 1996 22 nd International Conference for the Resource Management & Performance Evaluation of Enterprise Computing Systems, CMG. Part 2(of 2), pp. 819-828. 1996.

[14] R. Paul, E. Cooper-Balis, and B. Jacob, "Dramsim2: A cycle accurate memory system simulator." Computer Architecture Letters 10, no. 1 (2011): 16-19.

[15] L. Sheng, et al., "McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures.". MICRO-42, pp. 469-480. IEEE, 2009.

[16] J. Jeddeloh, B. Keeth., "Hybrid memory cube new DRAM architecture increases density and performance." (VLSIT), pp. 87-88. IEEE, 2012

[17] K. Skadron, et al.,"Temperature-aware microarchitecture." In ACM SIGARCH Computer Architecture News, vol. 31, no. 2, pp. 2-13. ACM, 2003.

[18] T. Le-Nguyen, F. J. Kurdahi, A. M. Eltawil, and H. Homayoun, "Heterogeneous memory management for 3D-DRAM and external DRAM with QoS." (ASP-DAC), pp. 663-668. IEEE, 2013.

[19] E. Cooper-Ballis, "Buffer-on-board memory system", PHD dissertation, UMUC, 2012.

[20] Z. Dali, H. Homayoun, and A. V. Veidenbaum, "Temperature aware thread migration in 3D architecture with stacked DRAM.", (ISQED), pp. 80-87. IEEE, 2013.

[21] W. Elsasser, 5 Emerging DRAM Interfaces You Should Know for Your Next Design, Cadence white paper, 2013

[22] K. Jung-Sik, et al., "A 1.2 V 12.8 GB/s 2 Gb Mobile Wide-I/O DRAM With 4 128 I/Os Using TSV Based Stacking." Solid-State Circuits, IEEE Journal of 47, no. 1 (2012): 107-116.

[23] M. Jung, et al. "TLM modelling of 3D stacked wide I/O DRAM subsystems: a virtual platform for memory controller design space exploration." In Proceedings of the 2013 Workshop on Rapid Simulation and Performance Evaluation: Methods and Tools, p. 5. ACM, 2013

[24] M. Gomony, et al., "DRAM selection and configuration for real-time mobile systems." (DATE), 2012, pp. 51-56. IEEE, 2012.

[25] Atwood, Greg. "Current and emerging memory technology landscape." Flash Memory Summit (2011).

[26] M. Sadri, M. Jung, C. Weis, N. Wehn, and L. Benini,"Energy optimization in 3D MPSoCs with wide-I/O DRAM using temperature variation aware bank-wise refresh." (DATE), pp. 1-4, 2014.

[27] Y. Bae, et al. "A 1.2 V 30nm 1.6 Gb/s/pin 4Gb LPDDR3 SDRAM with input skew calibration and enhanced control scheme." (ISSCC), pp. 44-46. IEEE, 2012.

[28] J. Kim, et al., "A 1.2 V 12.8 GB/s 2 Gb Mobile Wide-I/O DRAM With 4 x128 I/Os Using TSV Based Stacking," Solid-State Circuits, IEEE Journal of , vol.47, no.1, pp.107,116, Jan. 2012.

[29] M. H. Hajkazemi, M. Chorney, R. J. Behrouz, M. Khavari Tavana, and H. Homayoun, "Adaptive Bandwidth Management for Performance-Temperature Trade-offs in Heterogeneous HMC+DDRx Memory", In Proceedings of the 25th edition of the great lakes symposium on VLSI, ACM 2015.

[30] M. J. Khurshid, and M. Lipasti. "Data compression for thermal mitigation in the hybrid memory cube." In Computer Design (ICCD), 2013 IEEE 31st International Conference on, pp. 185-192. IEEE, 2013.

[31] G. H. Loh, "3D-stacked memory architectures for multi-core processors." In ACM SIGARCH computer architecture news, vol. 36, no. 3, pp. 453-464. IEEE Computer Society, 2008.

[32] H. Homayoun, V. Kontorinis, T. W. Lin, A. Shayan and D. M. Tullsen, "Dynamically Heterogeneous Cores Through 3D Resource Pooling", International Symposium on High-Performance Computer Architecture, HPCA, pp. 1-12, 2012.