

Decoupled DIMM: Building High-Bandwidth Memory System Using Low-Speed DRAM Devices

Hongzhong Zheng¹, Jiang Lin², Zhao Zhang³ and Zhichun Zhu¹

¹Dept. of ECE
University of Illinois at Chicago
{hzheng2, zzhu@uic.edu}

² Austin Research Lab
IBM Corp.
{linji@us.ibm.com}

³Dept. of ECE
Iowa State University
{zzhang@iastate.edu}

ABSTRACT

The widespread use of multicore processors has dramatically increased the demands on high bandwidth and large capacity from memory systems. In a conventional DDR x DRAM memory system, the memory bus and DRAM devices run at the same data rate. To improve memory bandwidth, we propose a new memory system design called *decoupled DIMM* that allows the memory bus to operate at a data rate much higher than that of the DRAM devices. In the design, a synchronization buffer is added to relay data between the slow DRAM devices and the fast memory bus; and memory access scheduling is revised to avoid access conflicts on memory ranks. The design not only improves memory bandwidth beyond what can be supported by current memory devices, but also improves reliability, power efficiency, and cost effectiveness by using relatively slow memory devices. The idea of decoupling, precisely the decoupling of bandwidth match between memory bus and a single rank of devices, can also be applied to other types of memory systems including FB-DIMM.

Our experimental results show that a decoupled DIMM system of 2667MT/s bus data rate and 1333MT/s device data rate improves the performance of memory-intensive workloads by 51% on average over a conventional memory system of 1333MT/s data rate. Alternatively, a decoupled DIMM system of 1600MT/s bus data rate and 800MT/s device data rate incurs only 8% performance loss when compared with a conventional system of 1600MT/s data rate, with 16% reduction on the memory power consumption and 9% saving of memory energy.

Categories and Subject Descriptors: B.3.2 [Primary Memory]: Design Styles

General Terms: Design, Performance, Power, Cost

Keywords: DRAM Memories, Decoupled DIMM, Bandwidth Decoupling

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ISCA'09, June 20–24, 2009, Austin, Texas, USA.

Copyright 2009 ACM 978-1-60558-526-0/09/06 ...\$5.00.

1. INTRODUCTION

The widespread use of multicore processors has put high demands on off-chip memory bandwidth and memory capacity. The memory bandwidth has been improved dramatically in recent years, with the data transfer rate passing 800MT/s (Mega-Transfers per second) and 1333MT/s and reaching 1600MT/s in standard DDR3 (Double Data Rate) memory systems with 1.5V power supply. The proposed DDR4 memory may reach 3200MT/s in the near future. Nevertheless, the race to high data transfer rate is putting high pressure on DRAM devices. The current JEDEC compatible DRAM devices that can support 1600MT/s data rate are not only expensive but also of low density. Some DDR3 devices have been pushed to run at data rates of 1600MT/s or higher by using a supply voltage higher than the JEDEC recommendation. However, they consume substantially more power, overheat easily and thus sacrifice their lifetime reliability.

We propose a new design, called *decoupled DIMM*, that makes it possible to construct high-bandwidth memory systems using DRAM devices of relatively low data rate. The design is based on an assumption, which is valid for memory systems in most multicore systems, that a memory channel has more than one DIMM (Dual Inline Memory Module) mounted, and/or each DIMM has more than one memory rank. In such a system, the memory bandwidth from all DRAM devices is at least double of the bus bandwidth. In other words, DRAM devices can operate at a much lower data rate and their combined bandwidth can still match the bus bandwidth. In the decoupled DIMM design, we introduce a chip called *synchronization buffer* (*sync-buffer* in short) that relays data between the memory bus and DRAM devices, so that DRAM devices can operate at a lower data rate. The devices can run at half of the bus data rate by default; and other data rates are also possible. Furthermore, we revise memory access scheduling to avoid any potential access conflict that may be introduced by the difference of data rate. We show that this issue can be elegantly addressed.

The design offers a number of benefits by decoupling the data rate of the DRAM devices and that of the bus and memory controller. First of all, memory bandwidth can be improved beyond what can be supported by current memory devices. In other words, DRAM device will no longer be the bottleneck in improving memory bandwidth. Memory cost can be reduced and capacity be increased, as high-speed devices tend to be costly and of low density. Memory reliability can also be enhanced by allowing DRAM devices to

operate at a conservative data rate. Furthermore, because DRAM memory devices consume less power at a lower data rate, memory power efficiency can be improved.

The use of sync-buffer incurs an extra delay in data transfer, and reducing device data rate slightly increases data burst time, both contributing to a slight increase of memory idle latency. Nevertheless, our analysis and performance evaluation show that the overall performance penalty is small when compared with a conventional DDR x memory system of the same high data rate on the bus and at the devices (which may not even be available). Although the sync-buffer consumes a certain amount of extra power, our experimental results show that it is more than offset by the power saving from reducing device data rate. The use of sync-buffer also has the advantage of reducing the electrical load on bus, which helps increase the number of DIMMs that can be installed in a system and therefore increase memory capacity.

Several DIMM designs and products, such as Registered DIMM [19], MetaRAM [16], Fully-buffered DIMMs [29] and Mini-Rank [33], use some form of bridge chip to improve capacity, performance and/or power consumption of DDR x ¹ memory system. The key aspect of this design is the decoupling of bandwidth match between memory bus and a single rank of memory devices, which none of the previous studies did. The idea of bandwidth decoupling may be applied to other types of memory systems; for example, to Fully-Buffered DIMM which has a narrow channel bus running at a frequency much higher than the devices, but still with a bandwidth match between the channel bus and a single memory rank. The idea is not complicated, but it can bring significant improvements in performance, cost effectiveness, power efficiency, capacity and/or reliability.

We have studied two design variants of the sync-buffer and thoroughly evaluated the performance of decoupled DIMM. The simplest design of sync-buffer is to always use 1 : 2 ratio of data rate conversion between devices and bus; and a slightly more complex design allows an $m : n$ ratio. Our simulation using memory-intensive workloads shows that a decoupled DIMM system of 2667MT/s bus data rate and 1333MT/s device data rate improves performance by 50.9% (up to 77.0%) and 26.9% (up to 43.1%) on average over conventional memory systems of 1333MT/s and 1600MT/s data rates, respectively. When compared with a conventional system of 1600MT/s data rate, a decoupled DIMM system of 1600MT/s bus data rate and 800MT/s device data rate only incurs 8.1% performance loss with 15.9% memory power reduction and 9.2% memory energy saving. The decoupled DIMM also provides opportunities to build memory systems of the same overall bandwidth but with fewer channels than conventional memory system designs, so as to reduce the cost on motherboard and the processor pin count. Our results show that a decoupled DIMM system of two 2133MT/s channels and with 1066MT/s devices only incurs 4.8% performance loss when compared with a conventional memory system of four 1066MT/s channels.

The rest of the paper is organized as follows. Section 2 presents the organization of contemporary DRAM memory systems and the DRAM technology trends. Section 3 discusses our proposed schemes and design alternatives in details. After presenting the experimental environment in Section 4, we analyze the performance and power impact of our

¹In the subsequent discussions, we use DDR x as a general term to represent DDR, DDR2, DDR3 or DDR4.

proposed scheme in Section 5, discuss the related work in Section 6 and summarize the work in Section 7.

2. MEMORY SYSTEM ORGANIZATION AND DRAM TECHNOLOGY TRENDS

Organization of Memory Systems. A typical, conventional DDR x memory system in a workstation or server system may consist of two to three memory channels, with one to four DIMMs in each channel. Each physical channel can work independently to maximize concurrency, or be ganged with others as a logical channel for higher combined channel bandwidth. Each DIMM consists of a number of DRAM devices organized into one or more ranks to support the 64-bit data path. For example, a 4GB DIMM without ECC may have sixteen 2Gbit x4 devices organized into a single rank, or sixteen 2Gbit x8 devices organized into two ranks. The data rate of the devices must match the desired bus data rate, e.g. the bus runs at 1600 MT/s in a memory system with DDR3-1600 devices.

DRAM Bandwidth and Latency Trends. The current DDR x DRAM technology evolves from Synchronous DRAM (SDRAM) to DDR, DDR2 and DDR3, with DDR4 being planned. Table 1 gives a comparison². Memory bandwidth has been improved dramatically; for instance, the data transfer rate increases from 133MT/s of SDRAM-133 to 1600MT/s of DDR3-1600³. Thus, data burst time (data transfer time) has been reduced significantly from 60ns to 5ns for transferring a 64-byte data block. By contrast, the internal DRAM device operation delay, such as precharge time (T_{pre}), row activation time (T_{act}) and column access time (T_{col}), only decreases moderately. As a consequence, the data transfer time only counts for a small portion of the overall memory idle latency (without queuing delay) now.

Power Consumption of DRAM Device. The power consumption of a DRAM chip can be classified into four categories: background power, operation power, read/write power and I/O power [20]. The background power is consumed all the time with or without operations. Today's DRAMs support multiple low power modes that can reduce the background power when a chip has no operations. The operation power is consumed when a chip performs activation or precharge operations; the read/write power is consumed when data are read out or written into a chip; and the I/O power is used for driving the data bus and terminating data from other ranks if necessary. For DRAM modules like DDR3 DIMMs, multiple ranks and chips are involved for each DRAM access; and the power consumed by an access is the sum of power consumed by all ranks/chips involved. Table 2 gives the power parameters of Micron 1Gbit devices, including background power (non-operating power in the table) for different power states, read/write power, and operation power for activation and precharge [17, 18]. I/O power is not listed in the table. It is estimated using typical on-die termination schemes and parameters for our evaluations [17].

²The price of 1GB unbuffered non-ECC DIMM is based on Samsung products. The price of DDR3-800 is estimated for comparison; and the price of SDRAM-1333 and DDR-400 is high because they are not mass productions.

³It is projected that the data transfer rate of future DDR4 can reach 3200MT/s.

	SDRAM-133	DDR-400	DDR2-800	DDR3-800	DDR3-1066	DDR3-1333	DDR3-1600
Voltage (V)	3.3	2.5	1.8	1.5	1.5	1.5	1.5
Max Capacity (per chip)	512Mb	1Gb	2Gb	4Gb	4Gb	2Gb	1Gb
Price (1GB UDIMM non-ECC)	\$49	\$49	\$28	\$75	\$99	\$129	\$175
Bus Freq. (MHz)	133	200	400	400	533	667	800
BW (MB/s/channel)	1066	3200	6400	6400	8533	10666	12800
t_{CK} (ns)	7.5	5	2.5	2.5	1.87	1.5	1.25
Timing(t_{CL}) (memory cycle)	3	3	6	6	8	9	11
Burst Length (memory cycle)	8	4	4	4	4	4	4
$T_{pre}, T_{act}, T_{col}$ (ns)	22.5	15	15	15	15	13.5	13.75
T_{bl} (ns)	60	20	10	10	7.5	6	5

Table 1: Comparison of DRAM Technologies.

Parameters (voltage/current)	Values (DDR3-800/1066/1333/1600)	Parameters (current)	Val. (DDR3-800/1066/1333/1600)
Normal voltage	1.5V	Operating burst read	130/160/200/250mA
Operating active-precharge	90/100/110/120mA	Operating burst write	130/160/190/225mA
Active standby	50/55/60/65mA	Burst refresh current	200/220/240/260mA
Precharge standby current	50/55/60/65mA	Active power-down current	25/30/35/40mA
Self refresh fast mode	7mA	Precharge power-down fast mode	25/25/30/35mA
Self refresh slow mode	3mA	Precharge power-down slow mode	10mA

Table 2: Parameters for calculating the power consumption of Micron 1Gbit devices, DDR3-800/1066/1333/1600.

3. DECOUPLED DIMM: CONCEPTS, DESIGNS AND IMPLEMENTATIONS

3.1 Overview of Decoupled DIMM Design

In general, the memory system performance can be measured as access latency and sustained throughput. Decoupled DIMM is based on two observations. First, DRAM devices can run at a data rate lower than that of the bus; and the combined bandwidth of all memory ranks in a channel can still match the bus bandwidth in a typical memory system. Second, reducing the device data rate only incurs little penalty on the memory access latency. The advantages of decoupled DIMM, as discussed in Section 1, are improvements on memory bandwidth (with higher bus frequency), memory reliability, and/or memory power efficiency.

Figure 1 compares the organizations of a decoupled DIMM with 2133MT/s bus data rate and 1066MT/s device data rate and a conventional DDR3-1066 DIMM, both with x8 devices. The organizations with x4 or x16 devices are similar. A key component in this design is the *synchronization buffer* (*sync-buffer*) labeled as SYB in the figure. It relays data between the higher speed bus and the lower speed devices, and makes the devices appear to be of the high data rate to the memory controller. There can be a single or multiple sync-buffers on a DIMM with more than one ranks, which is an implementation choice. The details of the sync-buffer will be discussed in Section 3.2.

The decoupled DIMM design decouples DRAM devices from the bus and memory controller in the race to higher data rate. The benefits include:

(1) **Performance.** The bus frequency is now only limited by the memory controller and bus implementation, with DRAM devices removed from being the bottleneck of frequency increase. Therefore, memory systems with higher bandwidth per-channel can be built with relatively slower DRAM devices.

(2) **Power Efficiency.** With the decoupled DIMM, the DRAM devices can operate at a low frequency, which will save memory power *and* energy. The power is reduced because the required electrical current to drive DRAM devices decreases with the data rate (as shown in Table 2). Although the energy spent for data read/write is not reduced, the energy spent on background, I/O and activations/precharges drops significantly. Our experimental results show that,

when compared with a conventional memory system with a faster data rate, the power reduction and energy saving from the devices are larger than the extra power and energy consumed by the sync-buffer. In overall, the decoupled DIMM is more power-efficient and consumes less energy.

(3) **Reliability.** In general, DRAM devices with higher data rates cause more concerns on reliability. In particular, it has been shown in various tests that increasing the data rate of DDR3 devices by increasing their operation voltage beyond the suggested 1.5V will cause memory data errors. The decoupled DIMM design allows DRAM devices to operate at a relatively slow speed with little performance penalty.

(4) **Cost Effectiveness.** Standard DRAM devices with the highest available data rate incurs a high cost premium. For example, the unit price of 1GB non-ECC DIMM is \$99 for DDR3-1066 and \$175 for DDR3-1600 [15] (for standard 1.5V DIMM).

(5) **Device Density.** The decoupled DIMM allows the use of high-density and low-cost devices such as DDR3-1066 devices to build a high-bandwidth memory system. By contrast, the conventional DIMM has to use low-density and high-cost devices, e.g. DDR3-1600 at this moment.

(6) **DIMM Count per Channel.** Similar to the chipset in MetaRAM [16], the sync-buffer in decoupled DIMM hides the devices inside the ranks from the memory controller, providing smaller electrical load for the controller to drive. This in turn makes it possible to mount more DIMMs in a single channel.

In summary, by decoupling DRAM devices from the bus and memory controller, the decoupled DIMM can “magically” improve the memory bandwidth by one or more generations and in the meanwhile improve the memory cost, reliability and power efficiency.

3.2 Design and Implementation

Figure 2 shows the structure of a single decoupled DIMM and its sync-buffer, using x8 devices as an example. The designs with x4 and x16 devices are the same except on the wire connections. The sync-buffer caches and relays the commands/addresses from the DDR x bus with one cycle delay at the bus clock frequency plus the clock synchronization overhead. The data from devices (for read requests) or from the memory controller (for write requests) are buffered and relayed with multi-cycle latency plus clock synchronization

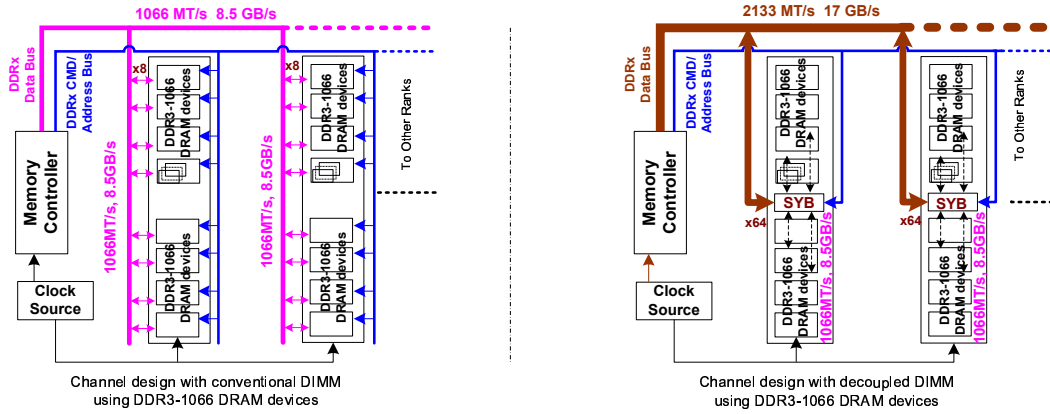


Figure 1: Conventional DIMM Organization vs. decoupled DIMM Organization with DDR3-1066, x8 devices as example.

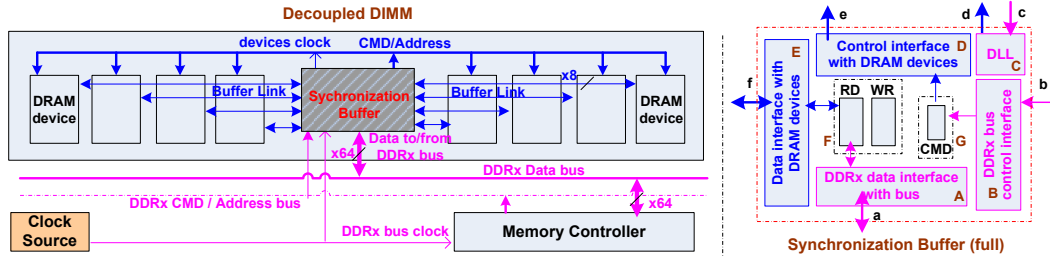


Figure 2: The decoupled DIMM design using x8 device. In the right part: (a) Data to/from DDRx bus, (b) Command/address from DDRx bus (c) Clock from data bus, (d) Clock to DRAM devices, (e) Command/address to DRAM devices, and (f) Data to/from DRAM device.

overhead.

There is at least one sync-buffer on each DIMM. If a DIMM has more than one ranks, two sync-buffers can be used. On those DIMMs, all ranks can be connected to a single sync-buffer through the on-DIMM DDR3 bus, or be organized as two groups connecting to a sync-buffer each. Using two sync-buffers on those DIMMs make it possible for a single DIMM to match the bus bandwidth if the device frequency is at least half of the bus frequency. Our experiments show that, if the aggregate bandwidth of all DIMMs matches the bus bandwidth, using single or double sync-buffers only has small (less than 1%) performance difference. Using more than two sync-buffers is possible, but is not considered in our study as any performance gain will be negligible.

As shown in Figure 2, the sync-buffer has seven components: (1) the data interface with DDRx bus, (2) a DDRx control interface, (3) a delayed loop lock (DLL) logic, (4) a control interface with DRAM devices, (5) a data interface with DRAM chips, (6) a data buffer for read/write data, and (7) a command/address buffer. The command/address buffer entry (32-bit wide) holds a set of command/address, which for DDR3 systems includes BA0-BA2 (banks address), A0-A13 (row/column address), RAS (Row Address Strobe), CAS (Column Address Strobe), WE (Write Enable), CKE (Clock Enable), ODT (On Die Termination) and CS (Chip Select), 23 effective bits in total. The data buffer has a read entry and a write entry, each of 64-byte wide. An incoming data burst is pipelined with the corresponding outgoing data burst, so that the last chunk of the outgoing burst completes one device cycle later than the last chunk of the

incoming burst. The memory controller should ensure the timing constraints of each rank; therefore, access conflicts will never happen on any buffer entry. The DLL is used to reduce the clock skew from the memory bus.

The left part of Figure 2 shows a single clock source to the sync-buffer, which then converts this clock signal into a slower clock signal to the devices. If the conversion ratio is $1 : m$, where m is an integer, a simple frequency divider using shift registers can be used. If the conversion is $n : m$, where n is another integer, then a PLL (Phase Lock Loop) logic is needed, which can be part of the sync-buffer or external to it (as the PLL in a registered DIMM).

Memory Access Scheduling. The memory controller decides the timing of DRAM operations, i.e. precharge, activation, column access and read or write operations, and the data bus usage for read/write requests [24, 30]. It must keep tracking the status of all memory ranks and banks, avoid bus usage conflicts, and maintain all DRAM timing constraints to ensure memory correctness. In addition, a memory controller may enforce some scheduling policy that prioritizes memory requests in certain ways. The decoupled DIMM design will require some extension to the memory access scheduling. However, the extension should be minimum because this function is fairly complex in current memory controllers.

With the decoupled DIMM, one must consider in memory access scheduling that there are two levels of buses, a channel bus and rank buses, with different clock frequencies; and that usage conflicts may appear on a rank bus even if it does not appear on the channel bus. In our design, the memory access scheduling works as if all ranks were directly

attached to the channel bus at the high frequency, with all timing constraints adjusted to the channel bus frequency and with the sync-buffer delay included. It further enforces an extra timing constraint to separate any two consecutive commands sent to memory ranks sharing the same rank bus. With that constraint, no access conflict will appear on any rank bus as long as no access conflict appears on the channel bus.

Figure 3 gives an example to show the difference in memory access scheduling between a conventional design and the decoupled DIMM design. There is a single read request to a precharged rank and the request is transformed to two DRAM operations, an activation (row access) and a data read (column access). The top part of the figure shows the scheduling results for a conventional DDR3-1066 system, and the bottom part shows those for a decoupled DIMM system with 2133MT/s channel data rate and 1066MT/s device data rate. When compared with the conventional system, the sync-buffer of decoupled DIMM increases memory idle latency by two cycles at the device frequency, one cycle for relaying the command/address and another cycle for relaying the data. Nevertheless, if there are multiple pending read requests, then decoupled DIMM may return data much faster because the channel bus has double frequency and the rank buses operate in parallel.

Design Choice of 1 : 2 Data Rate Ratio. The simplest design variant of decoupled DIMM is using 1 : 2 ratio between the channel and the devices’ data rates. In other words, the channel data rate is double of the device data rate. This is a special case of using 1 : m data rate ratio. In this case, the synchronization buffer is the most simplified, as it can use a simple frequency divider to generate the clock signal to the devices from the channel clock signal. The synchronization overhead of the two clocks is also minimized. This design choice is of particular interest in practice because the 1 : 2 ratio is the one between the currently available devices and the projected channel bandwidth for the next generation. The most common devices today are those with data rates of 1066MT/s and 1333MT/s devices, while the data rates of 2133MT/s and 2667MT/s are the next steps projected in the roadmap of DDR x memories.

3.3 Power Modeling

Modeling Methodology. The key component in the power modeling is the power of the sync-buffer, as the power modeling of DRAM devices already exists. We model the sync-buffer using Verilog and then break its power consumption into four portions, including (1) the I/O interface to the DRAM devices, (2) the I/O interface to the channel bus, (3) DLL (delayed loop lock) logic, and (4) non-I/O logics including SRAM data entries, command/address buffer and command/address relay logic. More details of the modeling methodology and the modeling parameters can be found in a previous study on mini-rank [33]. We found that the power consumption of the sync-buffer is small and can be more than offset by the power saving from DRAM devices; the details will be given in Section 5. The power estimation is in line with online sources regarding the power consumption of MetaSDRAM controller chip [16], which has roughly the same pin count as the sync-buffer.

Impacts on Device Power and Energy. The power consumption of DRAM devices increases with data rate (as shown in Table 2) and so does the energy. Consider using a

DDR3-800 device instead of a DDR3-1600 device. When the device is in the precharge standby state, the electrical current for providing the background power drops from 65mA to 50mA. When the device is being precharged or activated, the current to provide the operational power (additional to background) drops from 120mA to 90mA. Both result in energy saving, because the background power is required most time for memory-intensive workloads with little memory idle time, and the precharge and activation times are mostly independent with the data rate. When the device is performing a burst read, the current to provide the read power (additional to the background) drops from 250mA to 130mA (from 225mA to 130mA for write). This part of energy increases slightly because it takes double the time for the DDR3-800 device to perform the burst read. In overall, our experimental results show that the decoupled DIMM is more power-efficient and can save memory energy for memory-intensive workloads.

4. EXPERIMENTAL SETUP

4.1 Simulation Environment

We use M5 [1] as the base architectural simulator and extend its memory part to simulate the conventional DDR3 and our proposed DRAM systems in details. The simulator keeps tracking the states of each memory channel, DIMM, rank and bank. Based on the current memory state, memory commands are issued according to the hit-first policy, under which row buffer hits are scheduled before row buffer misses. Reads are scheduled before write operations under normal conditions. However, when pending writes occupy more than half of the memory buffer, writes are scheduled first until their number drops below one-fourth of the memory buffer size. The memory transactions are pipelined whenever possible and XOR-based address mapping [32, 12] is used as the default configuration. Table 3 shows the major simulation parameters.

To estimate the power consumption of DDR3 DRAM devices, we follow the Micron power calculation methodology [20, 17]. A rank is the smallest power unit. At the end of each memory cycle, the simulator checks each rank state and calculates the energy consumed during the cycle accordingly. The parameters used to calculate the DRAM (with 1Gb x8 devices) power and energy are listed in Table 2 [17, 18]. For those current values presented in manufacturers’ data-sheet that are from the maximum device voltage, they are de-rated by the normal voltage [20].

In all configurations, the default setups use x8 DRAM devices, cache line interleaving, close page mode and auto precharge. We have also done experiments with open page policy with page interleaving and get similar results and conclusions, which are not presented due to space limit. We use simple power management policy by putting a memory rank to a low power mode when there is no pending request to it for 24 processor cycles (7.5ns). The default low power mode is “precharge power-down slow” that consumes 128mW per device with 11.25ns exit latency. From our experiments, it gets the best power/performance trade-offs compared with other low power modes. We use term x CH- y D- z R to represent memory systems with x channels, y DIMMs per channel and z ranks per DIMM. For example, 4CH-2D-2R has four DDR3 channels, two DIMMs per channel, two ranks per DIMM, and nine devices per rank (with ECC).

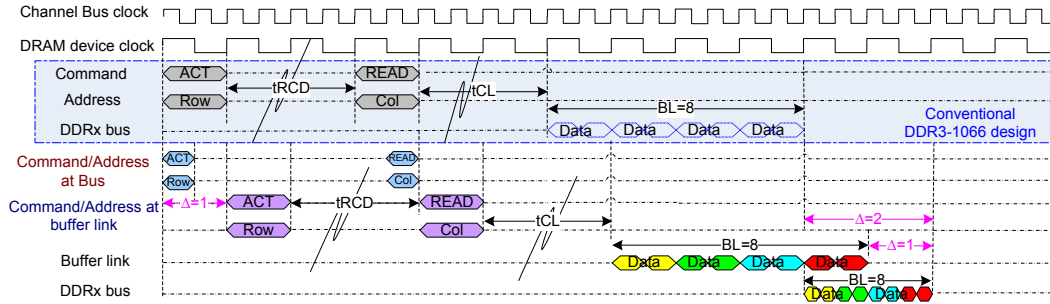


Figure 3: The timing example of decoupled DIMM architecture when the bus data rate is double of the DRAM device data rate.

Parameters	Values
Processor	4 cores, 3.2 GHz, 4-issue per core, 16-stage pipeline
Functional units	4 IntALU, 2 IntMult, 2 FPALU, 1 FPMult
IQ, ROB and LSQ size	IQ 64, ROB 196, LQ 32, SQ 32
Physical register num	228 Int, 228 FP
Branch predictor	Hybrid, 8k global + 2K local, 16-entry RAS, 4K-entry and 4-way BTB
L1 caches (per core)	64KB Inst/64KB Data, 2-way, 64B line, hit latency: 1 cycle Inst/3-cycle Data
L2 cache (shared)	4MB, 4-way, 64B line, 15-cycle hit latency
MSHR entries	Inst:8, Data:32, L2:64
Memory	4/2/1-channels, 2-DIMMs/channel, 2-ranks/DIMM, 8-banks/rank, 9-devices/rank
Memory controller	64-entry buffer, 15ns overhead
DDR3 channel bandwidth	8-byte/channel, 800MT/s (6.4GB/s), 1066MT/s (8.5GB/s), 1333MT/s (10.6GB/s), 1600MT/s (12.8GB/s) (MT: Mega Transfers/second), DDR3-800:6-6-6, DDR3-1066:8-8-8, DDR3-1333:10-10-10, precharge/row-access/column-access:15ns; DDR3-1600:11-11-11, precharge/row-access/column-access:13.75ns
DDR3 DRAM latency	

Table 3: Major simulation parameters.

4.2 Workload Construction

In our experiments, each processor core is single-threaded and runs a distinct application. We classify the twenty-six benchmarks of the SPEC2000 suite into MEM (memory-intensive), MDE (moderate), and ILP (compute-intensive) applications based on their memory bandwidth usage level. The MEM applications are those having memory bandwidth usage higher than 10GB/s when four instances of the application run on a quad-core processor with the four-channel DDR3-1066 memory system. The ILP applications are those with memory bandwidth usage lower than 2GB/s; and the MDE applications are those with memory bandwidth usage between 2GB/s and 10GB/s. Note that program *mcf* is usually classified as a MEM workload. Using the classification method here, it falls into the MDE category because its very low ILP degree causes low memory bandwidth usage despite of its high cache miss rate. Table 4 shows twelve four-core multi-programming workloads randomly selected using these applications.

In order to limit the simulation time while still emulating the representative behavior of program executions, a representative simulation point of 100 million instructions is selected for every benchmark according to SimPoint 3.0 [26]. There are several metrics for comparing performance of multi-core/multithreaded systems [28]. We use both Weighted Speedup [27] and Harmonic mean of normalized IPCs [13] in our study. The weighted Speedup is calculated as $\sum_{i=1}^n (IPC_{multi}[i]/IPC_{single}[i])$ and the harmonic mean of normalized IPCs is calculated as $n/\sum_{i=1}^n (IPC_{single}[i]/IPC_{multi}[i])$, where n is the total number of cores, $IPC_{multi}[i]$ is the IPC value of the application running on the i th core

under the multi-core execution and $IPC_{single}[i]$ is the IPC value of the same application under single-core execution.

5. PERFORMANCE AND POWER ANALYSIS

In this section, we present the overall performance improvement of the decoupled DIMM, compare and analyze its design trade-offs, and evaluate the power efficiency.

5.1 Overall Performance of Decoupled DIMM

Figure 4 compares the performance of two conventional DDR3 memory systems and a decoupled DIMM system: (1) D1066-B1066, a conventional DDR3-1066 memory system; (2) D1066-B2133, a decoupled DIMM system with DDR3-1066 devices and channels of 2133MT/s; and (3) D2133-B2133, a conventional memory system with DDR4-2133 devices and channels of 2133MT/s⁴. The experiments are done with three channel configurations: 1CH-2D-2R, 2CH-2D-2R and 4CH-2D-2R, with single channel, dual channels and four channels, respectively; each channel has two DIMMs and each DIMM has two ranks.

The D1066-B2133 decoupled DIMM system significantly improves the performance of the MEM and MDE workloads over the conventional D1066-B1066 system. It is not a surprise as the decoupled DIMM system doubles the channel bandwidth. The left part of Figure 4 shows the performance using the metric of weighted speedup. Compared with the conventional DDR3-1066 system, the performance gain of

⁴The timing parameters of DDR4-2133 used in the experiment (14-14-14) is projected based on the DDR x performance trend.

Workload	Applications	Workload	Applications	Workload	Applications
MEM-1	swim,applu,art,lucas	MDE-1	ammp,gap,wupwise,vpr	ILP-1	vortex,gcc,sixtrack,mesa
MEM-2	fma3d,mgrid,galgel,quake	MDE-2	mcf,parser,twof,facec	ILP-2	perlbmk,crafty,gzip,eon
MEM-3	swim,applu,galgel,quake	MDE-3	apsi,bzip2,ammp,gap	ILP-3	vortex,gcc,gzip,eon
MEM-4	art,lucas,mgrid,fma3d	MDE-4	wupwise,vpr,mcf,parser	ILP-4	sixtrack,mesa,perlbmk,crafty

Table 4: Workload mixes.

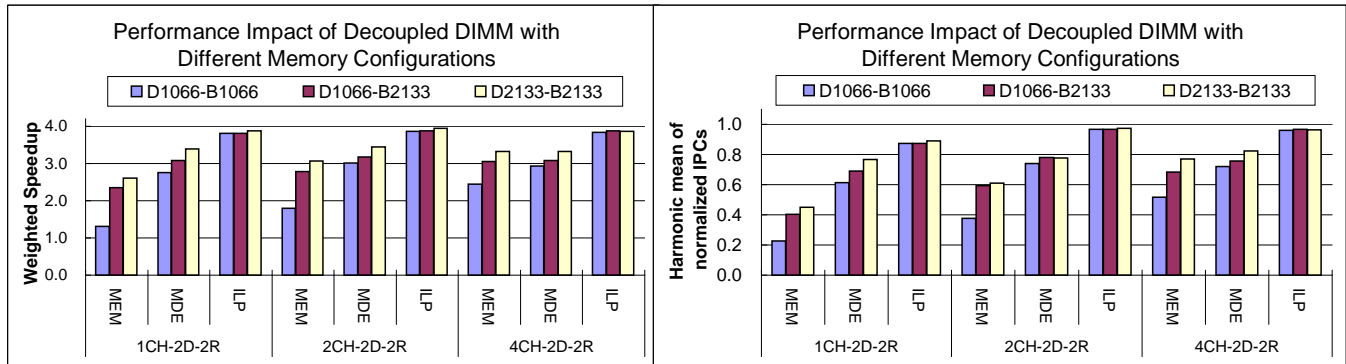


Figure 4: Performance comparison of a decoupled DIMM system (D1066-B2133) and two conventional DDR x memory systems (D1066-B1066 and D2133-B2133).

D1066-B2133 over D1066-B1066 is 79.4% (up to 92.9%), 57.9% (up to 82.7%) and 25.2% (up to 36.3%) on average with the single-, dual- and four-channel configurations, respectively, for the MEM workloads. The performance gain with the four-channel configuration is lower because our experiments only use four-core workloads; and with the four-channel configuration, the memory bandwidth is less a performance bottleneck. Right now, four-core systems normally use two memory channels; and future four-channel systems are expected to run with processors of more cores. For the MDE workloads, since their demands on memory bandwidth are lower, they benefit less from the increase on channel bandwidth; and the performance gain by the decoupled DIMM is 11.7% (up to 20.6%), 5.3% (up to 6.3%) and 5.0% (up to 6.6%) on average with the single-, dual- and four-channel configurations, respectively.

The right part of Figure 4 shows the performance results using the harmonic mean of normalized IPCs as the metric; and the performance trend is similar. Compared with the conventional DDR3-1066 system, the average performance gain of D1066-B2133 is 78.7% (up to 90.5%), 61.4% (up to 87.1%) and 33.2% (up to 53.5%) for the MEM workloads with the single-, dual- and four-channel configurations, respectively; and for the MDE workloads, the performance gain by decoupled DIMM is 12.4% (up to 22.3%), 5.4% (up to 6.4%) and 5.2% (up to 6.0%) on average, respectively. Section 5.2 will further show similar performance trends using other device configurations.

Compared with the conventional D2133-B2133 system, the decoupled DIMM D1066-B2133 system uses slower devices and thus has lower performance. However, the performance difference is small. Their average performance difference is 9.8%, 9.2% and 8.2% for MEM workloads, and 9.4%, 7.2% and 7.1% for MDE workloads, on single-, dual-, and four-channel configurations, respectively. The difference is small because the decoupled DIMM system has the same channel bandwidth; and its aggregate device bandwidth still matches the channel bandwidth. The perfor-

mance loss comes from the slight increase on memory latency as well as a potential imbalance of accesses among all ranks. We will analyze the performance impact of decoupled DIMM in more details in Section 5.2.

5.2 Design Trade-offs with Decoupled DIMM Architecture

The following discussions will use a two-channel configuration called 2CH-2D-2R (with two ranks per DIMM and two DIMMs per channel) as the base configuration. Figure 5 compares the performance of two decoupled DIMM systems, D1066-B2133 and D1333-B2667, with three conventional systems of different data rates, D1066-B1066, D1333-B1333 and D1600-B1600. Compared with the conventional D1066-B1066 system, the D1066-B2133 decoupled DIMM system improves the performance of the MEM workloads by 57.9% on average due to its higher bus bandwidth. Compared with the two conventional DDR3-1333 and DDR3-1600 systems, which use faster DRAM devices but slower buses, the D1066-B2133 decoupled DIMM system can still improve the performance of MEM workloads by 35.5% and 14.4% on average, respectively. This indicates that the bus bandwidth is more crucial to the performance than the device bandwidth; and the decoupled DIMM system is a more balanced design than the conventional system in matching the device and channel bandwidth. Similarly, the decoupled DIMM D1333-B2667 system can improve the performance of MEM workloads by 50.9% (up to 77.0%) and 26.9% (up to 43.1%) on average compared with the conventional DDR3-1333 and DDR3-1600 systems, respectively. As expected, the performance gain of decoupled DIMM on the MDE workloads is lower since those workloads only have moderate demands on memory bandwidth. For instance, the average performance gain of D1333-B2667 over the conventional DDR3-1333 and DDR3-1600 for the MDE workloads is only 4.6% and 3.0%, respectively. As for the ILP workloads, they are hardly affected by memory system configurations, and the decoupled DIMM has no negative performance impact

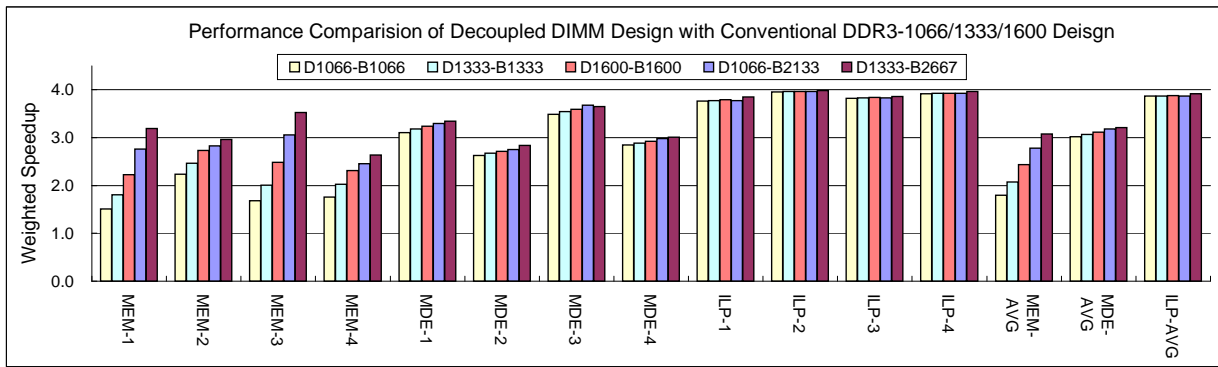


Figure 5: Performance comparison of decoupled DIMM with conventional memory systems of varied data rates for individual workloads.

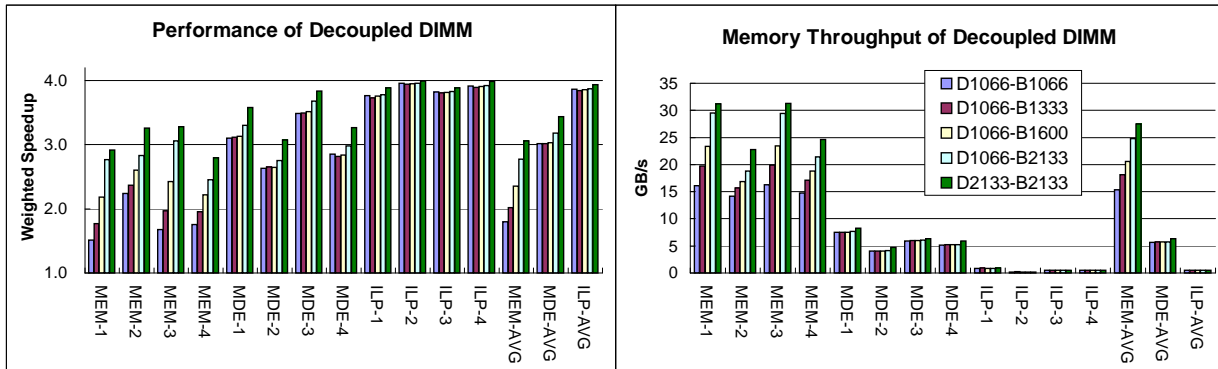


Figure 6: Performance comparison of decoupled DIMM and conventional systems of varied channel bandwidth.

on them.

Figure 6 further compares the performance of decoupled DIMM systems using the DDR3-1066 device and varied channel bandwidth levels of 1066MT/s (8.5GB/s), 1333MT/s (10.6GB/s), 1600MT/s (12.8GB/s) and 2133MT/s (17GB/s). For comparisons, the results on the conventional DDR4-2133 system are also included. In general, for the MEM workloads, the decoupled DIMM can improve performance significantly by using high-bandwidth channels and low-bandwidth (also low-cost) devices; and the performance keeps increasing with the increase of channel bandwidth. The average performance gains of MEM workloads are 13.0% (up to 17.3% on MEM-2), 32.9% (up to 44.4% on MEM-2), and 57.9% (up to 82.7% on MEM-1) for the configurations of D1066-B1333, D1066-B1600 and D1066-B2133, compared with the conventional DDR3-1066 system. The significant performance gain comes from bandwidth increase and improved memory bank utilization; both are critical to memory-intensive workloads. For example, the memory throughput of workload MEM-1 increases from 16.1GB/s with the conventional DDR3-1066 system to 19.7GB/s of the D1066-B1333 system, 23.3GB/s with D1066-B1600, and 29.5GB/s with D1066-B2133, respectively. There is no negative performance impact on the MDE and ILP workloads; and their memory throughput does not change much. The performance gain is still observable. For example, the D1066-B2133 system can improve performance by 5.3% on average for MDE workloads compared with the conventional DDR3-1066.

Figure 7 presents the breakdown of memory read access latency to show the details of performance impact of decoupled DIMM structures. The latency of memory read accesses

is divided into four parts: the memory controller overhead (T_{MCO}), the DRAM operation latency ($T_{operation}$), the additional latency introduced by sync-buffer (T_{SYB}), and the queuing delay ($T_{queuing}$). T_{MCO} is a fixed latency of 15ns (48 processor cycles for our simulation configuration); $T_{operation}$ is the memory idle latency including the DRAM activation, column access and data burst time from DRAM devices under the closed page mode. According to the DRAM device timing and PIN bandwidth configuration, $T_{operation}$ is 120 and 96 processor cycles for DDR3-1066 and DDR4-2133 devices under our simulation configuration, respectively. T_{SYB} is 19, 16, 12 processor cycles for the D1066-B1333, D1066-B1600 and D1066-B2133 configurations, respectively.

In general, the average read latency decreases as the channel bandwidth increases. The additional channel bandwidth from using decoupled DIMM significantly reduces the queuing delay. For instance, $T_{queuing}$ is reduced from 387 processor cycles in the conventional DDR3-1066 system to 310, 217, and 142 processor cycles on average for the MEM workloads in the decoupled DIMM systems of D1066-B1333, D1066-B1600 and D1066-B2133, respectively. The extra latency introduced by the sync-buffer only contributes to a small percentage of the total access latency, especially for the MEM workloads. T_{SYB} is only 3.3%, 3.0% and 2.9% of the overall latency on average for the MEM workloads in the configurations of D1066-B1333, D1066-B1600 and D1066-B2133, respectively, because of the large queuing delay. For the MDE workloads, the queuing delay is less significant than for the MEM workloads. Thus, T_{SYB} has more weight. However, the reduction of queuing delay can more than offset the additional latency from the sync-buffer. For the ILP workloads,

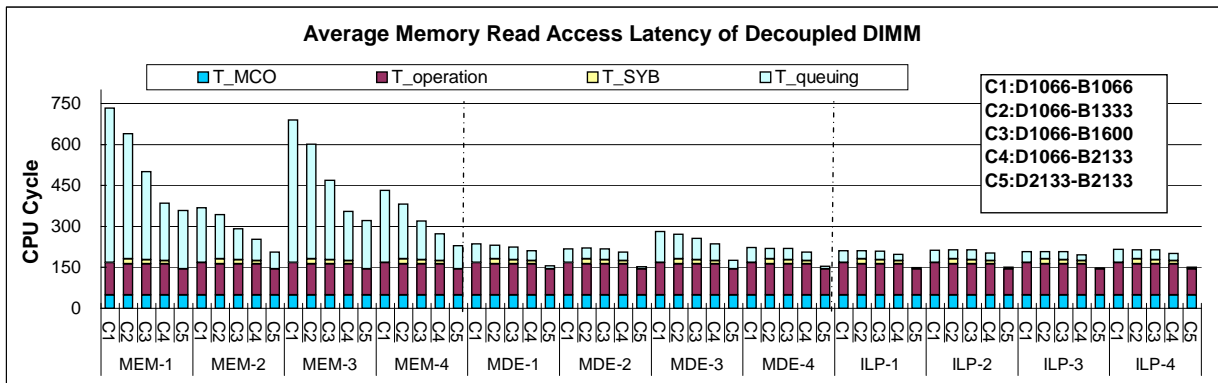


Figure 7: Memory access latency breakdown of decoupled DIMM systems.

the weight of T_{SYB} is even larger. However, the overall performance penalty is very small because the memory stall time is only a small factor of the overall performance.

5.3 Power Saving by Decoupled DIMM Architecture for Given Channel Bandwidth

The decoupled DIMM architecture also provides opportunities of power savings by building memory systems using slow DRAM devices. In this section, we will study design trade-offs for a given channel bandwidth of 1600MT/s, partially because we can only find the power specification for DDR3 devices with a data rate under 1600MT/s. Figure 8 compares the memory power consumption of decoupled DIMM configurations using DDR3-800, DDR3-1066, DDR3-1333 and DDR3-1600 devices. The conventional DDR3-800 and DDR3-1600 systems are also included for comparison. The memory power consumption is divided into five parts: (1) the power consumed by the sync-buffer’s non-I/O logic and its I/O operations with devices, (2) the power by the I/O operations between devices or sync-buffer and DDR x bus, (3) the device read/write power, (4) the device operation power and (5) the device background power. Note that the conventional memory systems do not have the part of power consumed by the sync-buffer.

In general, for a given channel bandwidth and memory-intensive workloads, the memory power consumption decreases with the DRAM device data rate. For instance, the average memory power consumption of the D1333-B1600, D1066-B1600 and D800-B1600 system for the MEM workloads is reduced by 1.6%, 6.7% and 15.9% to 30.3W, 28.7W and 25.8W, respectively, compared with the 30.8W power consumed by the conventional DDR3-1600 system, respectively.

The power reduction comes from the fact that the current to drive DRAM devices decreases with the data rate (see Table 2). For example, the current required for activation is 90mA and 120mA for DDR3-800 and DDR3-1600 devices, respectively. Therefore, the background, operation and read/write powers of a memory system all decrease with the device data rate. The major energy saving comes from the reduction of the operational and power and the background power. Our results show that the DRAM operation power of MEM-1, for example, is reduced from 15.4W of the conventional DDR3-1600 system to 13.2W, 12.4W and 10.6Watt for the D1333-B1600, D1066-B1600 and D800-B1600 systems, respectively.

The power consumed by the sync-buffer is the sum of

part one and part two in the memory power breakdown. However, only part one, the power consumed by the sync-buffer’s non-I/O logic and its I/O operations with devices, is additional when compared with a conventional system. This part goes down with the DRAM device speed because of lower running frequency and less memory traffic passing through the sync-buffer. For instance, the additional power of sync-buffer for the workload MEM-1 is 850mW, 828mW and 757mW per DIMM on the D1333-B1600, D1066-B1600 and D800-B1600 systems, respectively. The second part, the power of I/O operations between the devices or sync-buffer and DDR x bus, is needed by both the conventional systems and the decoupled DIMM systems. The only difference is that the power is consumed by the sync-buffer in the decoupled DIMM design but by the devices in the conventional system. The overall power consumption of the sync-buffer for the workload MEM-1 is 2.54W, 2.51W and 2.32W per DIMM on the D1333-B1600, D1066-B1600 and D800-B1600 systems, respectively, and only about 33% is the extra power. This extra power is more than offset by the power saving at devices.

Figure 9 further presents how the decoupled DIMM impacts the performance and memory power/energy consumption. Compared with the conventional DDR3-1600 system, the D800-B1600 system only causes an average performance loss of 8.1%. The relatively small performance difference is due to the fact that they have the same channel bandwidth. Fixing the channel bandwidth of decoupled DIMM at 1600MT/s, increasing device data rate from 800 MT/s to 1066MT/s and 1333MT/s only helps reduce the conflicts at sync-buffer, which is under 3% for MEM workloads and is not the performance bottleneck. As aforementioned, decoupled DIMM can reduce the memory power consumption by 15.9% on average for the MEM workloads; and thus the average memory energy consumption is reduced by 9.2%. For MDE and ILP workloads, the average power savings are 10.4% and 7.6% with 2.5% and 0.7% performance losses; and the average energy savings are 8.1% and 7.0%, respectively. For ILP workloads, all configurations have similar performance and power results. In summary, the decoupled DIMM allows the use of slow devices, or fast devices running at a low data rate, for improved power efficiency.

5.4 Building Memory Systems with Decoupled DIMM for Given Memory Bandwidth with Fewer Channels

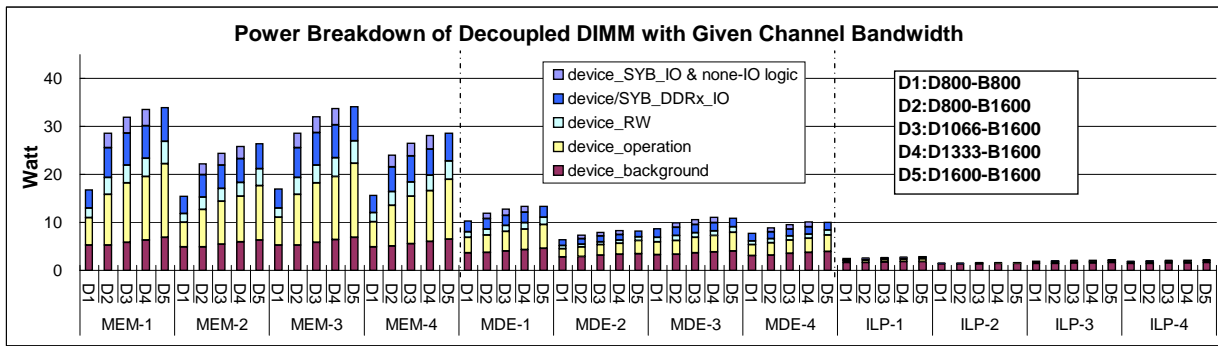


Figure 8: Power breakdown of decoupled DIMM systems for given channel bandwidth.

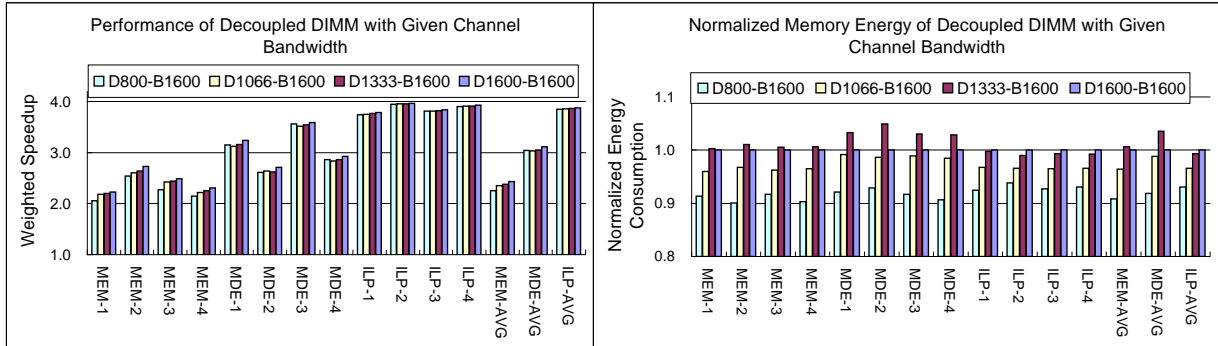


Figure 9: Performance and energy impact of decoupled DIMM for given channel bandwidth.

In this section, we will show that decoupled DIMM can provide opportunities to build memory systems for given system bandwidth with fewer channels and produce similar performance, which can reduce the cost on motherboard and processor or memory controller pin number. Obviously, building more channels in a system will cost more on the design and production as well as increased the design complexity. In addition, sometimes supporting more channels is not feasible due to limits on board space and pin count. As shown in Figure 10, compared with the conventional DDR3-1066 system with four channels, two DIMMs per channel and single rank per DIMM (4CH-2D-1R), the D1066-B2133 system with two channels, two DIMMs per channel and two ranks per DIMM (2CH-2D-2R) only has 4.8%, 4.0% and 2.6% performance loss for MEM, MDE and ILP workloads on average, respectively. Note that the two systems have the same overall system bandwidth (34GB/s) and use devices of the same speed; but the later system only has half the channels with twice the channel speed. Similarly, compared with the conventional DDR3-1066 system with the 2CH-2D-2R configuration, the D1066-B2133 system with the 1CH-2D-4R configuration causes the average performance loss of 5.8%, 6.4% and 6.8% for MEM, MDE and ILP workloads, respectively. Compared with the conventional design with more channels, the performance loss of decoupled DIMM with fewer channels mainly comes from the latency overhead introduced by the synchronization buffer and the increased contention on fewer channels.

6. RELATED WORK

Several DIMM designs and products use some forms of bridge chips to improve capacity, performance and/or power efficiency of DDR x memory systems. Those are the closely

related work to this study. Register DIMM [19] uses a register chip to buffer memory command/address between memory controller and DRAM devices. It reduces the electrical loads on the command/address bus so that more DIMMs can be installed on a memory channel. MetaRAM [16] uses a MetaSDRAM chipset to relay both address/command and data between the memory controller and the devices, so as to reduce the number of externally visible ranks on a DIMM and reduces the load on the DDR x bus. Fully-Buffered DIMM [29] uses very high speed, point-to-point links to connect DIMMs via AMB (Advanced Memory Buffer), which is the key to make the memory system scalable while maintaining the signal integrity on the high-speed channel. The Fully-buffered DIMM channel has fewer wires than the DDR x channel, which means more channels can be put on a motherboard. A recently proposed design called mini-rank [33] uses mini-rank buffer to break each 64-bit memory rank into multiple mini-ranks of narrower width, so that fewer devices are involved in each memory access. The design improves memory power efficiency at the cost of slightly longer data burst time. Our design is different in that it decouples the data rate between the DDR x memory bus and DRAM ranks so as to improve memory system bandwidth. Fully-buffered DIMM also uses different data rates between its channel and DIMMs; however, the channel bandwidth still matches the bandwidth of a single rank. Our work can also improve memory power efficiency by allowing DRAM devices to run at a relatively slow frequency.

There have been many studies on memory system performance evaluation and analysis. Burger et al. evaluate a set of techniques to hide long memory access latency and point out memory bandwidth would be limitation of modern processor design [2]. Cuppu et al. compare various DRAM technolo-

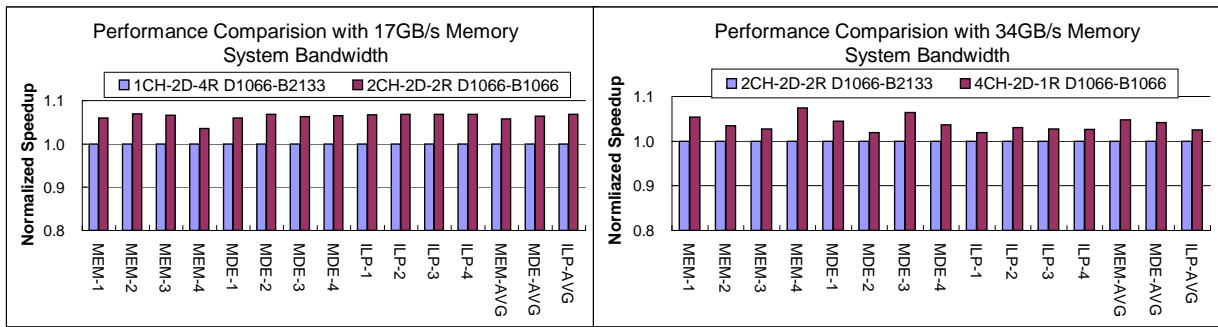


Figure 10: Performance comparison of decoupled DIMM architecture for given system bandwidth. For fair comparison, all configurations have eight 1GB-ranks.

gies in late 1990s, including Fast Page Mode, Extended Data Out, Synchronous Link, SDRAM, Rambus and Direct Rambus for single-threaded processors, focusing on comparisons of different types of DRAM and trade-offs between latency, bandwidth and cost [4]. Another study by Cuppu and Jacob focuses on the performance impact of memory design parameters, such as the number of memory channels, channel bandwidth, burst size, queue size and organization [3]. Ganesh et al. compare Fully-Buffered DIMM and DDR x , focusing on the performance characteristics of FB-DIMM including its scalability in capacity and bandwidth [7].

Memory access scheduling has been a research focus for more than a decade. McKee and Wulf study the effectiveness of five access ordering schemes on a uniprocessor system [14]. Rixner et al. discuss multiple memory access scheduling policies and evaluate their performance impact on media processing applications [24]. Hur and Lin propose adaptive history-based scheduling policies to minimize the expected delay and match the program’s mixture of reads and writes [9]. Zhu and Zhang evaluate memory optimizations for the SMT processors and propose thread-aware scheduling schemes [34]. Shao and Davis propose a burst scheduling mechanism to cluster the accesses on the same row page to maximize the data bus utilization [25]. Nesbit et al. [23] and Mutlu et al. [21] propose fair scheduling policies to balance memory resource usage among the multi cores on chip. Mutlu et al. [22] propose parallelism-aware batch scheduling for improving performance and fairness for multicore memory system. Decoupled DIMM has minimum impact on memory access scheduling as discussed in Section 3.

Most related studies on memory power efficiency focus on the use of DRAM low power modes, particularly for Rambus and Direct Rambus memories [11, 5]. Threaded memory modules [31] splits a conventional DDR x bus into sub-buses to reduce the number of DRAM devices involved in a single memory access. Recently, Diniz et al. propose algorithms to dynamically adjust DRAM power state for Rambus DRAM [6]. Hur and Lin study the power saving potential of a memory scheduling scheme and evaluates a scheme to predict throttling delay for memory power control [10]. Ghosh et al. propose adaptive refresh method to reduce the refresh power [8]. The decoupled DIMM improves memory power efficiency in a different way by allowing DRAM devices to run at a relatively low data rate. It is mostly compatible with existing low-power memory techniques.

7. CONCLUSION

The widespread use of multi-core processors has dramatically increased the demands on high bandwidth and large capacity from memory systems. In this study, we propose a novel design called decoupled DIMM that decouples the data rate of DDR x bus and devices so as to build high-bandwidth memory systems using relatively slow DRAM devices. It not only improves the system performance but also improves the memory capacity, reliability, power efficiency, and cost effectiveness.

Acknowledgment

We appreciate the constructive comments from the anonymous reviewers. This work is supported in part by the National Science Foundation under grants CCF-0541408, CCF-0541366, CNS-0834469 and CNS-0834475.

8. REFERENCES

- [1] N. L. Binkert, R. G. Dreslinski, L. R. Hsu, K. T. Lim, A. G. Saidi, and S. K. Reinhardt. The M5 simulator: Modeling networked systems. *IEEE Micro*, 26(4):52–60, 2006.
- [2] D. Burger, J. R. Goodman, and A. Kagi. Memory bandwidth limitations of future microprocessors. In *Proceedings of the 23rd International Symposium on Computer Architecture*, pages 78–89, 1996.
- [3] V. Cuppu and B. Jacob. Concurrency, latency, or system overhead: Which has the largest impact on uniprocessor DRAM-system performance? In *Proceedings of the 28th International Symposium on Computer Architecture*, pages 62–71, 2001.
- [4] V. Cuppu, B. Jacob, B. Davis, and T. Mudge. A performance comparison of contemporary DRAM architectures. In *Proceedings of the 26th International Symposium on Computer Architecture*, pages 222–233, 1999.
- [5] V. Delaluz, M. Kandemir, N. Vijaykrishnan, A. Sivasubramaniam, and M. J. Irwin. DRAM energy management using software and hardware directed power mode control. In *Proceedings of the 7th International Symposium on High-Performance Computer Architecture*, pages 159–169, 2001.
- [6] B. Diniz, D. Guedes, J. Wagner Meira, and R. Bianchini. Limiting the power consumption of main memory. In *Proceedings of the 34th International Symposium on Computer Architecture*, pages 290–301, 2007.
- [7] B. Ganesh, A. Jaleel, D. Wang, and B. Jacob. Fully-Buffered DIMM memory architectures: Understanding mechanisms, overheads and scaling. In *Proceedings of the 13th International Symposium on High-Performance Computer Architecture*, pages 109–120, 2007.

- [8] M. Ghosh and H.-H. S. Lee. Smart refresh: An enhanced memory controller design for reducing energy in conventional and 3D die-stacked DRAMs. In *Proceedings of the 40th International Symposium on Microarchitecture*, pages 134–145, 2007.
- [9] I. Hur and C. Lin. Adaptive history-based memory schedulers. In *Proceedings of the 37th International Symposium on Microarchitecture*, pages 343–354, 2004.
- [10] I. Hur and C. Lin. A comprehensive approach to DRAM power management. In *Proceedings of the 13th International Symposium on High-Performance Computer Architecture*, pages 305–316, 2008.
- [11] A. R. Lebeck, X. Fan, H. Zeng, and C. Ellis. Power aware page allocation. In *Proceedings of the Ninth International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 105–116, 2000.
- [12] W. Lin, S. K. Reinhardt, and D. Burger. Reducing DRAM latencies with an integrated memory hierarchy design. In *Proceedings of the Seventh International Symposium on High-Performance Computer Architecture*, pages 301–312, 2001.
- [13] K. Luo, J. Gummaraju, and M. Franklin. Balancing throughput and fairness in SMT processors. In *IEEE International Symposium on Performance Analysis of Systems and Software*, pages 164–171, 2001.
- [14] S. A. McKee and W. A. Wulf. Access ordering and memory-conscious cache utilization. In *Proceedings of the First IEEE Symposium on High-Performance Computer Architecture*, pages 253–262, 1995.
- [15] MemoryStore.com. Memory module price. http://www.memorystore.com/config/_generic.asp?cboLevel1=71.
- [16] MetaRAM, Inc. MetaRAM product brief. http://www.metaram.com/pdf/briefs/MetaRAM_DDR3_PB.pdf.
- [17] Micron Technology, Inc. DDR3 SDRAM system-power calculator. http://download.micron.com/downloads/misc/ddr3_power_calc.xls.
- [18] Micron Technology, Inc. MT41J128M8BY-187E. <http://download.micron.com/pdf/datasheets/dram/ddr3/1Gb%20DDR3%20SDRAM.pdf>.
- [19] Micron Technology, Inc. HTF18C64-128-256x72D. http://download.micron.com/pdf/datasheets/modules/ddr2/HTF18C64_128_256x72D.pdf, 2007.
- [20] Micron Technology, Inc. TN-41-01: Calculating memory system power for DDR3. http://download.micron.com/pdf/technotes/ddr3/TN41_01DDR3%20Power.pdf, 2007.
- [21] O. Mutlu and T. Moscibroda. Stall-time fair memory access scheduling for chip multiprocessors. In *Proceedings of the 40th Annual IEEE/ACM International Symposium on Microarchitecture*, pages 146–160, 2007.
- [22] O. Mutlu and T. Moscibroda. Parallelism-aware batch scheduling: Enhancing both performance and fairness of shared DRAM systems. In *Proceedings of the 35th International Symposium on Computer Architecture*, pages 63–74, 2008.
- [23] K. J. Nesbit, N. Aggarwal, J. Laudon, and J. E. Smith. Fair queuing CMP memory systems. In *Proceedings of the 39th International Symposium on Microarchitecture*, pages 208–222, 2006.
- [24] S. Rixner, W. J. Dally, U. J. Kapasi, P. Mattson, and J. D. Owens. Memory access scheduling. In *Proceedings of the 27th International Symposium on Computer Architecture*, pages 128–138, 2000.
- [25] J. Shao and B. T. Davis. A burst scheduling access reordering mechanism. In *Proceedings of the 13th International Symposium on High-Performance Computer Architecture*, pages 285–294, 2007.
- [26] T. Sherwood, E. Perelman, G. Hamerly, and B. Calder. Automatically characterizing large scale program behavior. In *Proceedings of the Tenth International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 45–57, 2002.
- [27] A. Snively, D. M. Tullsen, and G. Voelker. Symbiotic jobscheduling with priorities for a simultaneous multithreading processor. In *Proceedings of the 2002 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 66–76, 2002.
- [28] J. Vera, F. J. Cazorla, A. Pajuelo, O. J. Santana, E. Fernandez, and M. Valero. A novel evaluation methodology to obtain fair measurements in multithreaded architectures. In *Workshop on Modeling Benchmarking and Simulation*, 2006.
- [29] P. Vogt and J. Haas. Fully-Buffered DIMM technology moves enterprise platforms to the next level. <http://www.intel.com/technology/magazine/computing/fully-buffered-dimm-0305.htm>, 2005.
- [30] D. T. Wang. *Modern DRAM Memory Systems: Performance Analysis and a High Performance, Power-Constrained DRAM-Scheduling Algorithm*. PhD thesis, University of Maryland at College Park, Department of Electrical & Computer Engineering, 2005.
- [31] F. A. Ware and C. Hampel. Improving power and data efficiency with threaded memory modules. In *Proceedings of the 24th International Conference on Computer Design*, pages 417–424, 2006.
- [32] Z. Zhang, Z. Zhu and X. Zhang. A permutation-based page interleaving scheme to reduce row-buffer conflicts and exploit data locality. In *Proceedings of the 33rd International Symposium on Microarchitecture*, pages 32–41, 2000.
- [33] H. Zheng, J. Lin, Z. Zhang, E. Gorbatov, H. David and Z. Zhu. Mini-rank: Adaptive DRAM architecture for improving memory power efficiency. In *Proceedings of the 40th International Symposium on Microarchitecture*, pages 210–221, 2008.
- [34] Z. Zhu and Z. Zhang. A performance comparison of DRAM memory system optimizations for SMT processors. In *Proceedings of the 11th International Symposium on High-Performance Computer Architecture*, pages 213–224, 2005.