The DIMM Tree Architecture: A High Bandwidth and Scalable Memory System

Kanit Therdsteerasukdi[1], Gyung-Su Byun[2], Jeremy Ir[3], Glenn Reinman[1], Jason Cong[1], M.F. Chang[3]
[1]Computer Science Department, University of California, Los Angeles
[2]Computer Science and Electrical Engineering Department, West Virginia University, Morgantown
[3]Electrical Engineering Department, University of California, Los Angeles
Email: [1]{therdste, reinman, cong}@cs.ucla.edu, [2]gyungsu.byun@mail.wvu.edu, [3]jeremy.ir@ucla.edu, mfchang@ee.ucla.edu

***Abstract*—The demand for capacity and off-chip bandwidth to DRAM will continue to grow as we integrate more cores onto a die. However, as the data rate of DRAM has increased, the number of DIMMs supported on a multi-drop bus has decreased. Therefore, traditional memory systems are not sufficient to meet both these demands. We propose the DIMM tree architecture for better scalability by connecting the DIMMs as a tree. The DIMM tree architecture is able to grow the number of DIMMs exponentially with each level of latency in the tree. We also propose application of Multiband Radio Frequency Interconnect (MRF-I) to the DIMM tree architecture for even greater scalability and higher throughput. The DIMM tree architecture without MRF-I was able to scale up to 64 DIMMs with only an 8% degradation in throughput over an ideal system. The DIMM tree architecture with MRF-I was able to increase throughput by 68% (up to 200%) on a 64-DIMM system over a 4-DIMM system.**

I. Introduction

The memory wall [25], where DRAM system performance has not been able to scale at the same rate as processor performance, has been an ever-increasing problem for microarchitects. Now, with the emergence of chip multi-processors (CMPs), the problem has become even worse. As we continue to scale further with more and more cores on a chip, we reach a point where overall system performance cannot increase any further due to the limits of the DRAM system. Since the number of concurrent applications and threads increases with the number of cores, so does the working set size and the throughput required by the DRAM system. The larger working set size will increase the number of page faults leading to costly transfers from hard disk to memory. The increased throughput requirements will put an even greater strain on the already scarce DRAM bandwidth. Therefore, a DRAM system for future many-core CMPs will require greater capacity and greater throughput.

The trend in industry to provide greater throughput in DDRx DRAM has been to increase the DRAM clock and data rate over each pin. For example, DDR2 has data rates of 400Mbps/pin, 533Mbps/pin, 677Mbps/pin, and 800Mbps/pin. DDR3 has data rates of 800Mbps/pin, 1066Mbps/pin, 1333Mbps/pin, and 1600Mbps/pin. However, a faster data rate also decreases the number of DIMMs (and overall capacity) that can be connected together on a multi-drop bus in a conventional DDRx DRAM system. Each drop on a multi-drop bus acts as an impendence discontinuity, causing ringing, a longer delay, and slower rise time [11]. We have already seen the maximum number of drops reduce from 8 in DDR2 to 4 in DDR3. If the trend continues, then to support higher data rates in the future for technologies such as DDR4, there must be fewer drops on a multi-drop bus and therefore fewer DIMMs.

The other technique used to connect multiple DIMMs is a point-to-point link. FB-DIMM is an example that uses point-to-point links. Figure 1 shows the difference between connecting DIMMs on a multi-drop bus and point-to-point link. In a point-to-point link, the signals are buffered and repeated at each DIMM. This has the potential to allow an infinite number of DIMMs to be chained together at a high data rate. However, each buffer that is traversed adds latency to the system. DIMM0 in Figure 1b has a latency of 1 hop, while DIMM3 has a latency of 4 hops. In Figure 1a, since all the DIMMs are connected on a multi-drop bus, all the DIMMs have a latency of 1 hop. As the latency increases, throughput will degrade as we will demonstrate later on. Therefore, even with a point-to-point link the number of DIMMs is limited by the added latency.

In this paper, we propose the DIMM tree architecture for scaling the number of DIMMs in a DRAM system, but without severely degrading latency as in a point-to-point linked system. The DIMM tree architecture creates a tree of DIMMs combining the strengths of a multi-drop bus with the strengths of a buffer used in point-to-point links. The multi-drop bus is used to connect all the siblings in the tree. The buffer is used to connect all the children to the parent in the tree. This allows the latency to grow logarithmically with the number of DIMMs rather than linearly as in point-to-point links. We also propose the application of multiband radio frequency interconnect

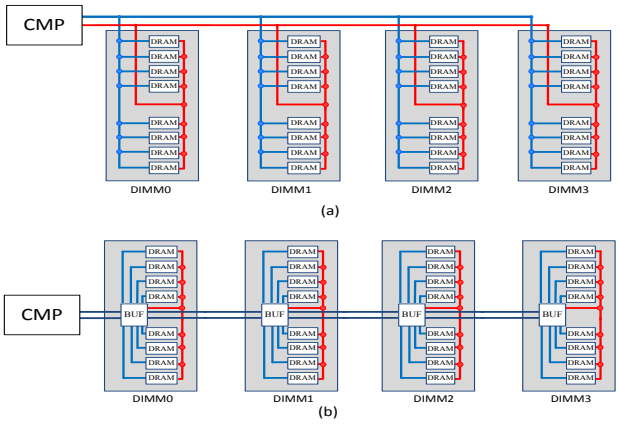

Figure 1. (a) 4 DDRx DIMMs on a multi-drop bus (b) 4 FB-DIMMs connected with point-to-point links.

(MRF-I) [4][6][13] to the DIMM tree architecture to provide even greater scaling and higher throughput. MRF-I in [4] is projected to support up to 4 drops on a multi-drop bus for the very high data rates required for future technologies such as DDR4 and DDR5. MRF-I also allows for multiple channels of data to be transmitted concurrently over a single shared medium, which can be used to increase the aggregate data rate to the DRAM system.

The remainder of this paper is organized as follows. Section 2 gives background for DDRx DIMMs and MRF-I. Section 3 describes the DIMM tree architecture in detail. Section 4 describes our experimental framework. Section 5 describes our results. Section 6 describes related work. Section 7 concludes this paper.

II. BACKGROUND

A. DIMM background

In this section we first give a brief overview of existing DIMM technologies in order to understand the tradeoffs of the DIMM tree architecture. Figure 2a shows a conventional DDRx DIMM consisting of multiple DDRx DRAM chips that are accessed in parallel. In this example, each DIMM contains 8 DRAM chips with each chip containing 8 data pins. A 64-bit data bus is created by aggregating the data signals from each chip together, which means each data line is connected to only one DRAM chip. Each command, address, and control signal, however, is connected to all 8 DRAM chips on the DIMM. Therefore, if there were 4 DIMMs on a single multi-drop bus, then each data line would be connected to 4 DRAM chips, and each command, address, and control signal would be connected to 32 DRAM chips. This demonstrates how quickly the load on a multi-drop bus can increase, thereby degrading the signal integrity.

One technique to reduce the load is to insert a buffer for all the signals on the DIMM between the DRAM chips and the memory controller. This technique is used in load reduced DIMM (LR-DIMM) [10] as shown in Figure 2b. LR-DIMM uses an isolation memory buffer (iMB) to buffer the command, address, control, and data signals to the DRAM chips. There are no changes needed to the DRAM chips themselves, which is very important. Since DRAM chips are commodity parts, and are optimized for high density and low cost, any design changes that may reduce density or increase cost are usually

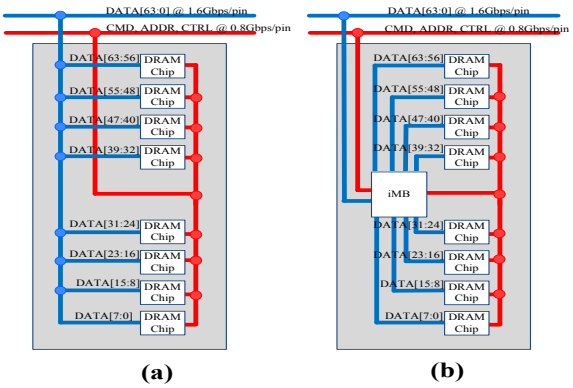

Figure 2. (a) conventional DDRx DIMM (b) LR-DIMM.

avoided. Therefore, our proposal to support the DIMM tree architecture will also not modify the DRAM chips, but just interface to them as is done with the iMB in LR-DIMM.

The data rate of DDRx is always twice the rate of the command, address, and control signals. Thus the name double data rate (DDR). In order for the DRAM chips to deliver such high data rates, internally the DRAM chips fetch the data with a lower data rate and a wider interface, but transfer externally with a higher data rate and a narrower interface. This is known as an n-bit prefetch. For example, DDR3-1600 is an 8n-bit prefetch architecture. Internally, it fetches 8-bits of data in one clock cycle at 800 MHz. Externally it transfers 8-bits of data, one bit at a time, over a single pin at 1.6GHz. Therefore, the 1.6Gbps/pin external data rate of DDR3-1600 is really generated by fetching data at half the clock speed internally. Since the DIMM tree architecture must support more DIMM-to-DIMM interfaces than a conventional DDRx DIMM, we will use this technique to reduce the number of data pins for each DIMM-to-DIMM interface by transmitting at twice the data rate over half the number of pins.

B. Off-Chip Multiband RF-I

Multiband RF-I [4][6][13] is a high aggregate bandwidth and power saving alternative to a traditional interconnect. MRF-I is realized via transmission of electromagnetic waves through multiple carrier channels over a shared transmission line, rather than the transmission of a voltage signal through a single baseband over a wire. In MRF-I, carrier waves are continuously propagated along the transmission line, and data is generated through either the amplitude or phase modulation of the carrier wave. By transmitting independent data streams each over different RF bands, MRF-I can provide simultaneous transmissions of multiple data streams over a shared physical transmission line to improve the aggregate bandwidth and data rates.

There has been much advancement in off-chip MRF-I in recent years [4][6][13]. The most recent advancement, [4], uses ASK modulation with differential signaling, which we refer to as ASK MRF-I. ASK MRF-I in [4] uses differential signaling, which means it uses two lines to propagate a signal. Differential signaling allows for a higher signal integrity, which leads to higher data rates and a higher number of RF bands per pin overall. [4] was successful in demonstrating the high data rate and low power of MRF-I, the low BER, and the feasibility of process integration by implementation in a general-purpose logical CMOS process of 65nm. [4] demonstrated a dual band MRF-I transceiver operating over 10cm on a FR4 board and Roger 4003C board at 8.4Gbps aggregate data rate and 10Gbps aggregate data rate respectively. The power consumption of the dual band MRF-I transceivers on the FR4 and Roger boards were 21mW and 25mW respectively. Both boards operated with less than a 10^{-15} BER.

C. Why use MRF-I for DRAM?

Traditional chip-to-DRAM interconnects are able to support fewer drops as the data rate increases and signal integrity becomes worse. This has become apparent in the reduced number of DIMMs on a multi-drop bus as technology has changed from the slower DDR2 DIMMs, to the faster

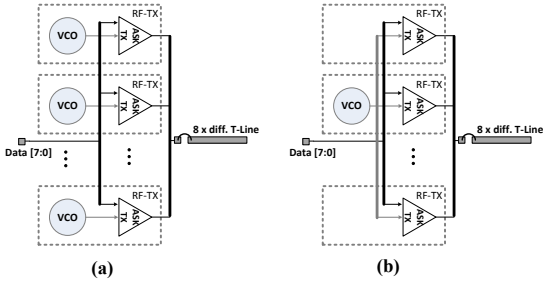

Figure 3. (a) Replicated MRF-I transceiver from [4] (b) optimized with reduced VCO design

DDR3 DIMMs, and in the near future DDR4 DIMMs. Because of the high data rate, DDR4 is projected to only support 1 DIMM on a bus. Even though [4] did not demonstrate multi-drop for MRF-I, multi-drop for MRF-I is very feasible. At the time of submitting this paper, our group is working on a MRF-I mult-drop demo. We project that MRF-I at 4Gbps per RF band (enough to support DDR4) will be able to support 4 DIMMs on a multi-drop bus. Since [4] implements differential signaling using dual bands, there isn't the extra pin overhead usually associated with differential signaling with only the baseband i.e. two pins for two bands with MRF-I instead of two pins for one band for a traditional interconnect.

MRF-I can also be used to reduce pin count or increase bandwidth by supporting more than two bands per pair of differential lines i.e. greater than one band per pin. For example, with 2 RF bands per pin (4 RF bands per pair of differential lines), we could either support the same bandwidth and reduce the number of pins by half, or keep the number of pins the same and double the bandwidth. The latter application is particularly useful from an energy savings perspective as we start to approach data rates of greater than 5Gbps/pin. At about 5Gbps/pin traditional interconnects start to consume power super linearly due to power-hungry circuit techniques of pre-emphasis and equalization that must be used to compensate for the signal loss. Examples include current technologies such as GDDR5 (7Gbps/pin) and future technologies such as DDR4/DDR5 that will reach around 5Gbps/pin. By multiplexing the data over multiple bands, the interconnect power can be kept in the linear power consumption region. For example, transmitting data over 2 RF bands per pin operating at 4Gbps per band will provide 8Gbps/pin. [4] can currently support up to 4 RF bands per pin. However, as MRF-I technology advances, we expect that number to increase even more.

D. Overhead of using MRF-I

There is also an area savings improvement that can be made for multi-bit transceivers. [4] demonstrates a dual band transceiver over a single pair of differential lines. When creating a multiple bit transceiver, the simplest approach would be just to replicate the design. However, this is very area and energy inefficient. The capacitive loading on each ASK transmitter is very low, so each ASK transmitter does not require its own dedicated voltage-controlled oscillator (VCO) in order to produce the RF carrier (as shown in Figure 3a). Instead, a single VCO can be shared among up to 8 ASK transmitters as long as they are using the same RF band (as shown in Figure 3b). This optimization results in both an area and energy savings. We were able to validate these area and energy optimizations by layout and simulation using the Spectre circuit simulator [5] as was done in [14]. The values are shown in TABLE I and Figure 4.

TABLE I. AREA OF 8-BIT TRANSCEIVERS

	BB	2ASK MRF-I	4ASK MRF-I	8ASK MRF-I
Area (mm^2)	0.528	0.372	0.341	0.31
# pins	8	8	4	2
# transceivers	8	4	2	1
# bits	8	8	8	8

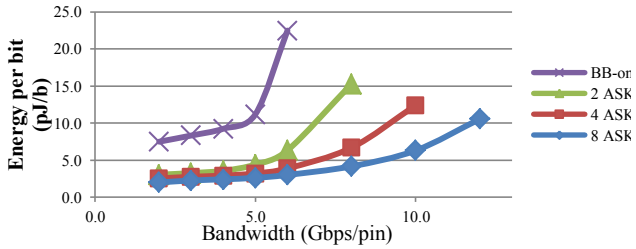

Figure 4. Energy per bit of baseband and ASK MRF-I

The area of 8-bit transceivers, including pads, for baseband (BB) and RF-I transceivers is shown in TABLE I for 65nm technology. We label transceivers for 2, 4, and 8 RF bands per differential lines as 2ASK, 4ASK, and 8ASK respectively. The individual transceiver size can be obtained by taking the "Area" and dividing by "#transceivers." For example, a single 2 ASK transceiver is $0.372mm^2/4$. TABLE I shows that as the number of RF bands per pin increases, the area and number of pins required to transmit 8 bits of data shrinks significantly. We will discuss later on which ASK transceivers are required to support the DIMM tree architecture.

Figure 4 shows the energy per bit as bandwidth is increased for MRF-I versus a traditional interconnect, which is labeled as baseband (BB). The power numbers for the baseband were taken from [8][12][17]. We compare BB against 2ASK, 4ASK, and 8ASK. The figure shows that we can maintain the lower energy per bit at higher data rates by adding more RF bands.

Latencies for the transmitters, receivers, and transmission lines for RF-I versus the baseband are shown in TABLE II for 5cm and 10cm. These latencies fall well within the DDR3-1600 cycle time of 1.25ns, which we use as our level-to-level latency in the DIMM tree architecture. Please note that since [4] was a proof of concept paper to demonstrate the feasibility of off-chip MRF-I, the circuits were not optimized for area or power. One area reducing improvement that can be made without affecting the operation of the design is to place the digital logic circuits directly underneath the passive structures.

TABLE II. TX, RX, AND TRANSMISSION LINE LATENCIES

	TX (ns)	RX (ns)	5cm Line (ns)	10cm Line (ns)
Baseband	0.1	0.21	0.33	0.64
RF-I	0.14	0.22	0.39	0.72

III. THE DIMM TREE ARCHITECTURE

The DIMM tree architecture is designed to increase the capacity of a DRAM system without degradation in throughput. The DIMM tree architecture creates a tree of DIMMs in order to grow the latency logarithmically instead of linearly with the number of DIMMs; this allows the memory system to scale to a many-DIMM DRAM system. The DIMM tree requires a minimum of two DIMMs to be supported on a multi-drop bus. Otherwise, it becomes a chain of DIMMs connected with point-to-point links. Therefore, in the future with much high data rates, a technology such as MRF-I that can support two DIMMs on a multi-drop bus will be required. This section will describe the benefits and organization of a tree of DIMMs, the implementation details of a tree DIMM (T-DIMM) without MRF-I, and the adding of MRF-I to the DIMM tree architecture.

A. Benefits and organization of a DIMM tree

The main benefit of a DIMM tree is the logarithmic increase in latency with the number of DIMMs. This can be seen by comparing the DIMM tree versus a point-to-point and multi-drop bus organization. Figure 5 shows the varying latencies of each level of DIMMs in a point-to-point organization, a multi-drop organization, and the DIMM tree. A multi-drop connection among DIMMs is represented by the DIMMs sharing a common wire. In Figure 5a, the point-to-point organization causes the latency to increase linearly with the number of DIMMs, due to the buffer at each DIMM that acts as a signal repeater. DIMM0 only has a latency of 1 hop from the CMP while DIMM3 has a latency of 4 hops from the CMP. A point-to-point organization of N DIMMs can also be viewed as a tree with branching factor 1 and height N. In Figure 5b, the multi-drop organization causes the latency to be equal among all the DIMMs. In this case all the DIMMs have equal latency of 1 hop away from the CMP. A multi-drop organization of N DIMMs can be viewed as a tree with branching factor N and height 1.

Figure 5c shows a DIMM tree of branching factor 2 and height 2. In a DIMM tree, there are different families of DIMMs connected by a multi-drop bus. Each DIMM contains a buffer in order to generate a new clean signal to its children. This is just as in a point-to-point connection, except each DIMM in the DIMM tree may have multiple children. For example, the CMP and its children, DIMM0 and DIMM1, all share a multi-drop connection (just as the CMPs and their children, do in Figure 5a and b). Likewise, DIMM0 also shares a multi-drop connection with its children, DIMM2 and DIMM4. However, DIMM2 is not on the same multi-drop connection with the CMP, since the buffer on DIMM0 separates the connections. Just as in the point-to-point connection, each buffer represents a connection to an additional hop. Therefore, DIMM0 has a latency of one hop while DIMM2 has a latency of two hops.

B. Tree DIMM (T-DIMM) implementation

A conventional DDR3 DIMM only supports one interface - from the DIMM to the memory controller. In a T-DIMM, however, we must be able to support two interfaces – one to the parent and sibling DIMMs and one to the children DIMMs. In order to support an additional interface on the DIMM without the added pin overhead, we can use a technique similar to the n-bit prefetch used in DDRx described in Section II.A. By transferring some of the signals over half the number of pins but at twice the data rate, we can reduce the overhead of adding a second DIMM interface.

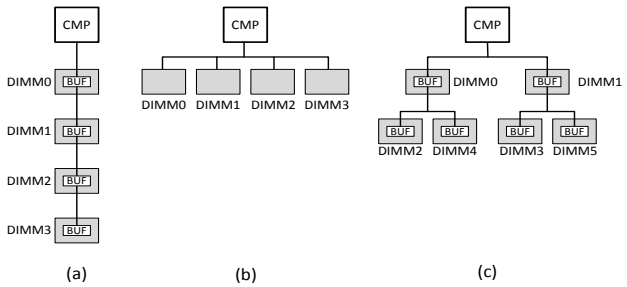

Figure 5. (a) point-to-point links (b) multi-drop bus (c) DIMM tree

Figure 6a shows a single DDR3-1600 T-DIMM with data rates and number of pins for the data, address, command, and control lines. The data and address lines operate at 2X the data rate (3.2Gbps/pin for data, 1.6Gbps for address) of a conventional DDR3 DIMM (1.6Gbps/pin for data, 0.8Gbps for address), but using half the number of pins (32 for data, 7 for address). Therefore, in order to support a second DIMM interface, the number of pins on the DIMM is increased by what amounts to another set of command/control lines plus chip select, which is 10 + log2(number of ranks). We assume there is logic on each DIMM to decode the chip select with "log2(number of ranks)" lines instead of "number of ranks" lines. This design also causes the number of pins needed to interface to the memory controller to be halved. The command and control lines for the T-DIMM operate at the same rate as a conventional DDR3 DIMM (0.8Gbps/pin). All signals to the DRAM chips must go through the DIMM Interface Router (DIR) just as they do for the iMB in LR-DIMM [11].

The DIR, shown in Figure 6b, contains a parent DIMM baseband (BB) transceiver, a router, a buffer, a child DIMM baseband transceiver, several data rate converters, and several baseband transceivers (BB TX/RX). The parent DIMM BB transceiver connects the DIMM to its parent and siblings within the tree hierarchy. The router consists of a lookup table indexed by the rank number specifying four possible routes: the current DIMM, a descendent of the current DIMM, the parent DIMM, or none of the above. The buffer is used to buffer signals that must go to the next level of the tree (i.e., a descendent of the DIMM). The child DIMM BB transceiver connects the DIMM to its children in the tree hierarchy. The memory controller would be the root of the tree, so it would also contain a child DIMM BB transceiver. The data rate converter converts between a data rate of X with Y pins to a data rate of 2X with Y/2 pins and vice versa. This is accomplished by interleaving the values of two signals operating at data rate X onto a single wire at data rate 2X and vice versa.

C. Adding MRF-I to the DIMM tree architecture

The drawback of using a traditional interconnect is that as the DRAM chip data rate is increased, fewer and fewer drops are supported. This directly affects the branching factor of the

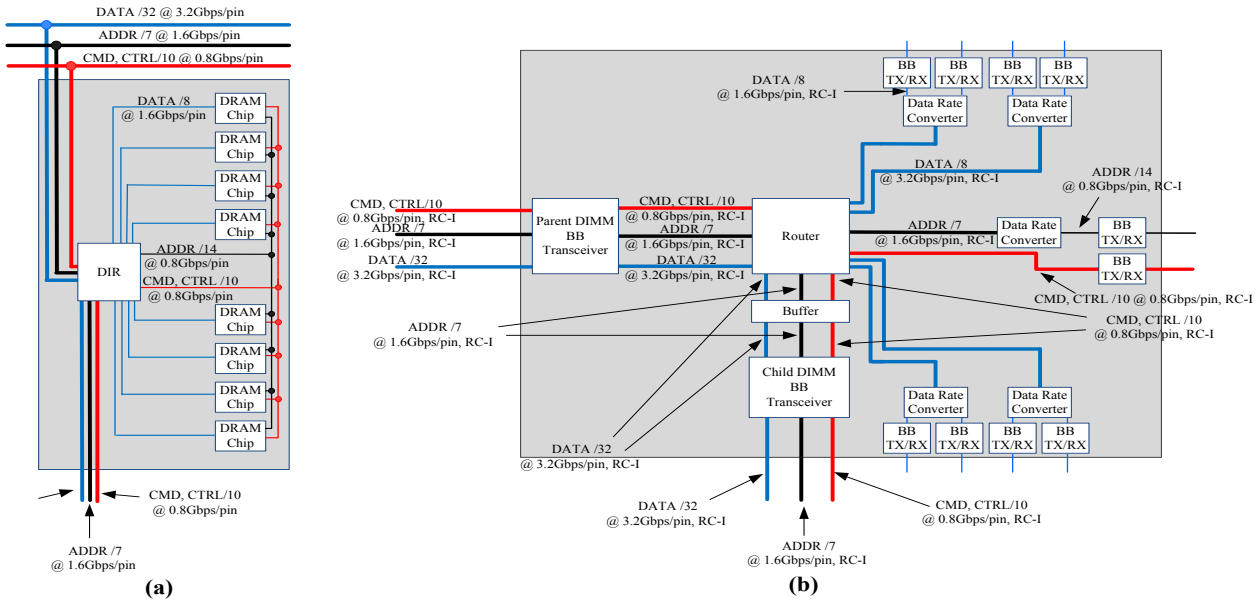

Figure 6. (a) T-DIMM (b) DIMM interface router for T-DIMM

DIMM tree and the rate at which the DIMM tree can grow with each level of latency added. This is especially true since the DRAM chip data signal rates are doubled in order to support an additional DIMM interface. Therefore, as new DRAM chip technologies with much higher data rates such as DDR4 arrive, the scalability of the DIMM tree using a traditional interconnect decreases. Replacing the traditional interconnect with MRF-I will allow the DIMM tree to continue to scale as data rates increase.

MRF-I is projected to support up to 4 drops on a multi-drop bus up to 4Gbps per RF band. That means with 2 ASK MRF-I (1 RF band per pin), a DIMM tree with a branching factor of 4 can support DDRx-2000 DRAM chips (2Gbps/pin). In order to support even higher data rates, the data signals can be multiplexed over more RF bands. Therefore, with 4 ASK MRF-I (2 RF bands per pin) and 8ASK MRF-I (4 RF bands per pin), a DIMM tree with a branch factor of 4 could support DDRx-4000 DRAM chips (4Gbps/pin) and DDRx-8000 (8Gbps/pin) respectively. Currently, [4] is limited to just 8ASK MRF-I. However, as MRF-I technology advances, we expect the number of RF bands per pin to increase.

MRF-I can also be used to provide multiple logical channels over a single physical channel when there is more than 1 RF band per pin. Each RF band on a pin would form a logical channel. By partitioning DIMMs on a multi-drop bus among the logical channels, we can increase the concurrency of DRAM transactions. For example, with 4 DIMMs on a multi-drop bus and two logical channels, there would be two DIMMs per logical channel. All transactions to the DIMMs on the first logical channel can be scheduled independently of the transactions to the DIMMs on the second logical channel. Therefore, we are able to utilize the extra bandwidth provided by MRF-I to current DRAM chip technologies to increase throughput and improve scalability as we shall see in the results section.

Adding MRF-I to the T-DIMM involves replacing the parent and child DIMM baseband transceivers in the DIR with MRF-I transceivers. The parent MRF-I transceiver will always be a 2 ASK MRF-I transceiver regardless of the number of logical channels supported, since each T-DIMM only has one parent. The child MRF-I transceiver, however, will be have a 4 and 8 ASK MRF-I transceiver for 2 and 4 logical channels respectively, since each T-DIMM can have multiple children. Adding MRF-I to the T-DIMM also involves having a set of lines for each logical channel from the router to the buffer, and from the buffer to the child DIMM MRF-I transceiver. These extra lines are required, since the buffer can only buffer a conventional signal, not an RF signal. The BB TX/RX remain the same, so there is no modification needed to interface to the commodity DRAM chips.

IV. Experimental Framework

For our evaluation, we generated memory transaction traces from the SPEC CPU 2006 benchmark suite [9], stream suite [15], and some medical imaging benchmarks. We selected the most memory intensive benchmarks from the SPEC CPU 2006 benchmark suite. The benchmarks included are bzip2, gcc, libquantum, lbm, mcf, milc, and sjeng. Copy and triad are derived from the stream benchmark suite [15], and are streaming benchmarks. Deblur [20], registration [26], and denoise [24] are medical imaging benchmarks. In order to generate memory transaction traffic for a multiprogrammed

TABLE III. WORKLOAD DESCRIPTIONS

	Benchmarks in workload	Memory footprint	Description
Workload1	deblure, registration, denoise	6.6 GB	medical imaging
Workload2	6 denoise, 6 bzip2, 6 sjeng	7.0 GB	mixed spec 2006 and medical imaging
Workload3	mcf, 2 libquantum, 2 milc, 2 gcc, 2 bzip2	5.8 GB	spec2006
Workload4	2 deblure, 2 libquantum, 2 milc, 2 registration, 2 gcc	16.2 GB	mixed spec 2006 and medical imaging
Workload5	2 copy, 2 triad	7.5 GB	stream
Workload6	4 lbm, 4 mcf	4.9 GB	spec2006

TABLE IV. MIXES OF SPEC 2006 BENCHMARKS

	Workload	Benchmark Mixes	Bandwidth (GB/s)	#Transactions (millions)
low_bw_mix_1	Workload1	deblure, registration, denoise	8.2	34.5
low_bw_mix_2	Workload2	6 denoise, 6 bzip2, 6 sjeng	14.3	60.0
med_bw_mix_1	Workload3	mcf, 2 libquantum, 2 milc, 2 gcc, 2 bzip2	30.5	109.2
med_bw_mix_2	Workload4	2 deblure, 2 libquantum, 2 milc, 2 registration, 2 gcc	28.4	119.2
high_bw_mix_1	Workload5	2 copy, 2 triad	59.6	175.0
high_bw_mix_2	Workload6	4 lbm, 4 mcf	748.8	326.5

workload (or a system using virtualization) running on a many-core CMP, we need to model several of the benchmarks on separate cores concurrently. We create 6 workloads shown in TABLE III. Each workload is generated so that the memory footprint is at least 4GB when the benchmarks are run to completion.

Since simulating the benchmarks to completion with a cycle accurate simulator would take several months to complete, we reduce our analysis to a 1 billion instruction phase of each benchmark. The traces were gathered using Pin [19], a dynamic instrumentation tool, with a 2MB 8-way set associative L2 cache model with 64B blocks taken from Simplescalar [3]. The traces were generated by warming up for 1 billion instructions before recording and then running for another 1 billion instructions while recording memory transactions, similar to [18]. We found that warming up for 1 billion instructions was enough to reach beyond the initialization phase of the benchmarks when all the compulsory page faults occur. The traces were taken as input into DRAMsim [22], a detailed cycle accurate memory system simulator. We use DRAMsim's built-in ability to interleave several trace files together in order to create a multiprogrammed CMP workload similar to [7] that will stress the DRAM system. TABLE IV shows the 6 different mixes we use. The mixes are categorized by how much they will stress the DDR3-1600 DRAM system (11.92 GB/s per channel)—i.e. low (0 to 2 channels), med (2 to 4 channels), and high (greater than 4 channels). We use the parameters in TABLE V for the simulations. We modify DRAMsim to support the DIMM tree architecture. For the DRAM chips, we use timing and power parameters from the Micron datasheets for DDR3-1600 [16]. We assume 1 rank on each DIMM.

TABLE V. DRAMSIM PARAMETERS

DRAM type	DDR3-1600
CPU frequency	4GHz
Channel width	8 bytes
Address mapping policy	sdram_close_page_map
Row buffer policy	close_page
Banks per rank	8
Row count	16384
Column count	1024
Rank-to-rank switch time	1 DRAM cycle

V. RESULTS FOR THE DIMM TREE ARCHITECTURE

A. Throughput and scalability

In this section we compare the throughput and scalability of a system of DIMMs connected by point-to-point links, an ideal multi-drop system, the DIMM tree architecture, and MRF-I. We implement all systems using DDR3-1600 DRAM chips for a fair comparison of the architecture. For example, it would not be fair to compare a DDR2-800 FB-DIMM against a multi-drop system of DDR3-1600 DRAM chips, since the data rates are different.

Figure 7 shows the throughput and scalability of a DIMM system using DDR3-1600 DRAM chips connected with point-to-point links. Throughput is measured in GB/s. The number of DIMMs is varied from 4 to 64. As the number of DIMMs is increased from 4 DIMMs, the added latency degrades throughput drastically. With 8 DIMMs, the throughput is degraded on average 15% up to 24%. With 16 DIMMs, the throughput is degraded on average 21% up to 39%. With 32 DIMMs, the throughput is degraded on average 38% up to 56%. The only exception is low_bw_mix_1 where the throughput increases by 34% when going to 16 DIMMs. The reason for this increase is that low_bw_mix_1 has a large amount of rank level parallelism. The gains by exploiting low_bw_mix_1's rank level parallelism exceed the degradation caused by the added latency. However, when increasing the DIMMs from 16 to 32, the throughput begins to decline again with the added latency.

Figure 8 shows the throughput and scalability of an ideal multi-drop system versus the DIMM tree architecture. The ideal multi-drop system models a hypothetical upper bound on the throughput we could achieve if all the DIMMs could be supported on a multi-drop bus. The DIMM tree shown in Figure 8 has a branching factor of 4 (i.e. 4 drops on a multi-drop bus can be supported). The system modeled with the DIMM tree could either be one using a traditional interconnect

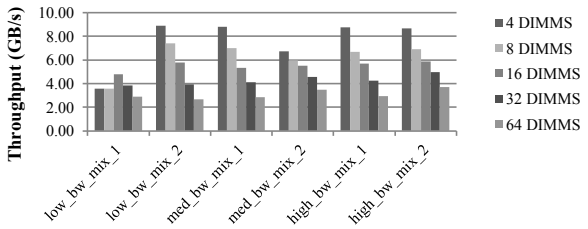

Figure 7. Throughput and scalability of DIMMs connected with point-to-point links

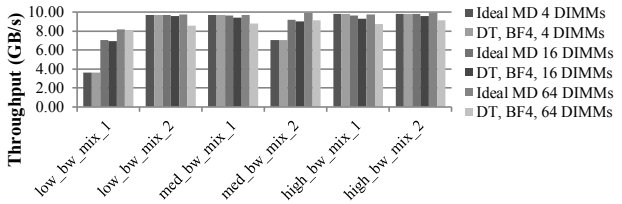

Figure 8. Throughput and scalability of DIMMs connected with DIMM tree versus ideal multi-drop

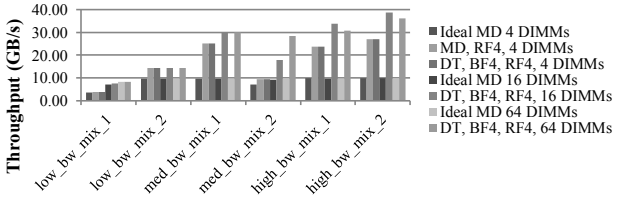

Figure 9. Throughput and scalability with 4 RF bands per pin

or one with one RF band per pin, since they would have the same throughput. The ideal multi-drop system is labeled with "Ideal MD". The DIMM tree architecture with branching factor 4 is labeled with "DT, BF4". We see that the DIMM tree scales much better than a system connected with point-to-point links. With 16 DIMMs, the degradation in throughput with the DIMM tree over the ideal multi-drop is on average 2% up to 4%. With 64 DIMMs, the degradation is on average 8% up to 12%. Mixes low_bw_mix_1 and med_bw_mix_2 both have a large amount of rank level parallelism, and see a large throughput increase as the DIMMs are increased from 4 to 16.

Figure 9 shows the throughput and scalability as we add MRF-I to a DIMM tree of branching factor 4. MRF-I is added with multiple RF bands per pin in order to support multiple concurrent logical channels. Figure 9 shows the throughput of a DIMM tree system with branch factor 4 and 4 RF bands per pin (labeled DT, BF4, RF4) against an ideal multi-drop system (labeled Ideal MD) and a non-ideal multi-drop system with 4 RF bands per pin (labeled MD, RF4). The addition of 4 RF bands per pin in a multi-drop system with 4 DIMMs increases throughput by an average of 93% up to 159%. As the number of DIMMs in the DIMM tree increases, we again see a benefit. The mixes with a high amount of rank level parallelism are able to use the 4 logical channels to schedule transactions to separate ranks concurrently to improve throughput. For example, low_bw_mix_1 with 32 T-DIMMs increases throughput by 124% over a 4-DIMM multi-drop system with 4 RF bands per pin ("MD, RF4, 4 DIMMs"). All the mixes see similar increases in throughput with 4 logical channels. The exception though is low_bw_mix_1, which at 16 DIMMs is already close to its maximum throughput of 8.2GB/s from TABLE IV. With 64 T-DIMMs, we see a throughput increase on average of 68% up to 200% over "MD, RF4, 4 DIMMs".

The 4 logical channels created by the 4 RF bands per pin creates so much more bandwidth than required by the mixes that it offsets any latency caused by additional levels in the DIMM tree. Therefore the DIMM tree with multiple RF bands per pin is able to outperform both an ideal multi-drop system and a non-ideal multi-drop system using multiple RF bands per pin.

B. Power

In this section we compare the power of the DIMM tree architecture versus a conventional DDRx DIMM and a LR-DIMM system from Figure 2. At 4 DIMMs, a DIMM tree with branching factor 4 will perform equivalently to a conventional DDRx DIMM and LR-DIMM, since all 3 systems will have the 4 DIMMs connected with a multi-drop bus. Therefore, we compare all 3 systems with 4 DIMMs using DDR3-1600 DRAM chips so we can compare the power values directly. We obtain the DRAM chip power from DRAMsim [22] and the Micron datasheet for DDR3-1600 [16]. The non DRAM chip numbers are obtained by the highly accurate Spectre circuit simulator [5]. Figure 10 shows the power results in milliwatts. Most of the power is consumed by the DDR3-1600 DRAM chips. The rest of the non DRAM chip power includes the interconnect, baseband or RF transceivers, and any additional structure needed e.g. iMB for LR-DIMM. LR-DIMM adds on average 3% up to 4% more power compared to a conventional DDRx DIMM. The DIMM tree adds on average 5% up to 6% more power compared to a conventional DDRx DIMM. Therefore, unlike past technologies to improve throughput and capacity such as FB-DIMM, the power overhead of the DIMM tree architecture is very small.

VI. RELATED WORK

[21] demonstrated a 10Gbit/s optical link with GaAs based technology to implement an optical multi-drop memory bus. Vantrease et al. [23] proposed Corona, which used photonic links to provide high bandwidth to the DRAM. However, their design required a 3D layout and is incompatible with commodity DRAM parts. Beamer et al. [2] redesigned the entire memory system all the way down to the banks in order to support silicon photonics. While much research is being done with optical interconnect, an optical memory bus still suffers from several critical problems. First the photonic GaAs compound technology is still immature and incompatible with silicon-based DRAM commodity fabrication. Second, critical optical building blocks such as a silicon laser and the Ge p-i-n photo detector [1] are extremely sensitive to temperature/process variations. In contrast, our demonstrated Multiband RF-I is fully compatible with the low-cost CMOS manufacturing and is ready for use now, unlike optical technologies. However, once optical interconnect technology does mature enough, our DIMM architectures can be implemented with optical links instead of RF-I.

Ko et al. [13] demonstrated a MRF-I board using BPSK modulation. The demo achieved a data rate of 3.6Gb/s/pin with 2 RF bands per pin. Therefore, each RF band was able to achieve 1.6Gbps. The RF-I transceivers were manufactured in a 0.18µm 1.8V CMOS technology. However, the BER was too

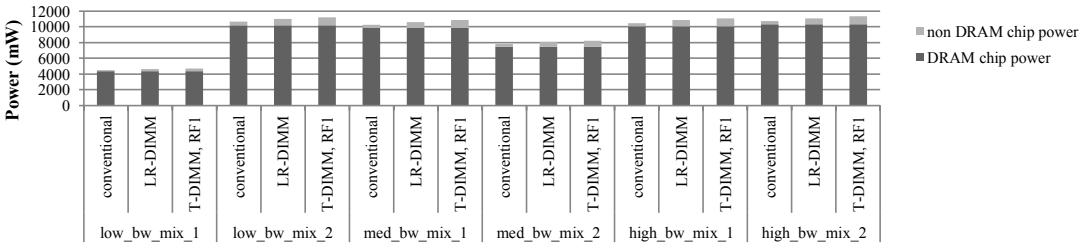

Figure 10. Power comparison of DIMM tree versus a conventional DDRx DIMM and LR-DIMM

high (10^{-7}) to be used for DDR3, which requires a BER of 10^{-12}. Byun et al. [4] demonstrated a MRF-I board using ASK modulation with differential signaling. The demo achieved a data rate of 5Gbps/pin and achieved a BER of less than 10^{-15}.

Fully buffered DIMM (FB-DIMM) [7] was designed to reduce load by interfacing all signals through the advanced memory buffer (AMB), and encoding everything as packets. The AMB connected each FB-DIMM in a point-to-point manner using a high-speed serial link operating at 6 times the DRAM clock. However, FB-DIMM consumed considerably more power than a conventional DDRx DIMM due to its high-frequency serial links and power-hungry AMB used to decode, store, forward, and encode packets.

VII. Conclusion

The DRAM system is one of the most critical components in any modern day computing system. We are reaching a point where we are pushing the limits of traditional interconnect technology for DRAM, sacrificing throughput for capacity. The DIMM tree architecture allows DRAM systems to scale to many DIMMs without sacrificing throughput, while reducing the number of pins to interface with the memory controller. We have shown that the DIMM tree architecture can scale up to 64 DIMMs with only an 8% reduction in throughput over an ideal multi-drop system. By adding MRF-I to the DIMM tree architecture, we are able to scale even further than a system with just the DIMM tree architecture or with just MRF-I. Using 4 RF bands per pin with a DIMM tree of 64 DIMMs, we are able to see an average of 68% (up to 200%) increase in throughput over a 4-DIMM multi-drop system with 4 RF bands per pin. We have also shown that the additional structures required to support the DIMM tree architecture only require 5% more power than a conventional DDRx DIMM and 2% more power than a LR-DIMM. The DIMM tree architecture is a high capacity high throughput DRAM system for future many-core CMPs for running many applications or threads concurrently.

Acknowledgment

This work has been supported in part by the Center for Domain-Specific Computing (CDSC).

References

[1] D. Ahn et al., “High Performance Waveguide Integrated Ge Photo Detectors,” Opt. Express 15, 2007.

[2] S. Beamer et al., “Re-Architecting DRAM Memory Systems with Monolithically Integrated Silicon Photonics,” in Proceedings of ISCA-37, 2010.

[3] D. Burger and T. Austin, “The Simplescalar Tool Set,” version 2.0. Technical Report CS-TR-97-1342, U. of Wisconsin, Madison, 1997.

[4] G. Byun et al., “An 8.4Gb/s 2.5pJ/b Mobile Memory I/O Interface Using Bidirectional and Simultaneous Dual (Baseband and RF-Band) Signaling,” in 2011 IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers, 2011.

[5] Cadence Virtuoso Spectre Circuit Simulator. http://www.cadence.com/products/rf/spectre_circuit/pages/default.aspx

[6] MF. Chang, et al., “Advanced RF/Baseband Interconnect Schemes for Inter- and Intra-ULSI Communications,” IEEE Transactions on Electron Devices, vol. 52, no. 7, Jul 2005, pp. 1271-1285.

[7] B. Ganesh, A. Jaleel, D. Wang, B. Jacob, “Fully-Buffered DIMM Memory Architectures: Understanding Mechanisms, Overheads and Scaling," HPCA-13, 2007.

[8] K. Ha et al., “A 0.13µm CMOS 6 Gb/s/pin Memory Transceiver Using Pseudo-Differential Signaling for Removing Common-Mode Noise Due to SSN”, In IEEE Journal of Solid-State Circuits, vol. 44, no. 11, Nov. 2009.

[9] J.L. Henning, “SPEC CPU2006 Benchmark Descriptions,” 2006.

[10] Inphi. LRDIMM Isolation Memory Buffer (iMB™) Component. http://www.inphi.com/products-technology/computing-storage-products/products/imbtrade02-gs02.php

[11] B. Jacob, D. Wang, S. Ng, “Memory Systems: Cache, DRAM, Disk”, Morgan Kaufmann Publishes, San Diego, 2008. Pg 391.

[12] J. Kennedy et al., J. Hoffmann, W. Hokenmaier, R. Houghton and T. Vogelsang, “A 3.6-Gb/s Point-to-Point Heterogeneous-Voltage-Capable DRAM Interface for Capacity-Scalable Memory Subsystems,” in IEEE Journal of Solid-State Circuits, vol. 40, no. 1, Jan. 2005.

[13] J. Ko et al., “An RF/Baseband FDMA-Interconnect Transceiver for Reconfigurable Multiple Access Chip-to-Chip Communication,” in 2005 IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers, February 2005.

[14] K. Kundert, "Introduction to RF Simulation and Its Application," IEEE Journal of Solid-State Circuits, Vol. 34, No. 9, Sep. 1999.

[15] J. McCalpin, “Memory bandwidth and machine balance in current high performance computers,” IEEE Computer Society Technical Committee on Computer Architecture (TCCA) Newsletter, Dec. 1995.

[16] Micron. 1Gb: x4,x8,x16 DDR3 SDRAM Features. 2006.

[17] K. Oh et al., “A 5-Gb/s/pin Transceiver for DDR Memory Interface With a Crosstalk Suppression Scheme,” in IEEE Journal of Solid-State Circuits, vol. 44, no. 8, Aug. 2009.

[18] N. Rafique, W.-T. Lim and M. Thottethodi, “Effective Management of DRAM Bandwidth in Multicore Processors,” in PACT, 2007.

[19] V. Reddi, A. Settle, D.A. Connors and R.S. Cohn, “Pin: A Binary Instrumentation Tool for Computer Architecture Research and Education,” 2004.

[20] E. Tadmor, S. Nezzar and L. Vese, "Multiscale hierarchical decomposition of images with applications to deblurring, denoising and segmentation", Commun. Math. Sci., vol. 6, p.281 , 2008.

[21] M. Tan et al., “A High-Speed Optical Multi-Drop Bus for Computer Interconnections,” in IEEE Micro, 95:945-953, Jun. 2009.

[22] D. Wang, B. Ganesh, N. Tuaycharoen, K. Baynes, A. Jaleel, Bruce Jacob, “Dramsim: A Memory-System Simulator,” 2006.

[23] D. Vantrease et al., “Corona: System Implications of Emerging Nanophotonic Technology,” in Proceedings of ISCA-35, 2008.

[24] L. Vese and S. Osher, “Image Denoising and Decomposition with Total Variation and Oscillatory Functions,” Journal of Mathematical Imaging and Vision, Vol. 20, No. 1–2, pp. 7–18, 2004.

[25] W.A. Wulf and S.A. McKee, “Hitting the Memory Wall: Implications of the Obvious,” Computer Architecture News, vol. 23, no. 1, Mar. 1995, pp. 20–24.

[26] Yanovsky, I., C.L. Guyader, A. Leow, P. Thompson, L. Vese, “Nonlinear elastic registration with unbiased regularization in three dimensions,” Computational Biomechanics for Medicine III, MICCAI 2008 Workshop (2008)