Citation:
Marino, MD (2016) ABaT-FS: Towards adjustable bandwidth and temperature via frequency scaling in scalable memory systems. MICROPROCESSORS AND MICROSYSTEMS, 45. pp. 339-354. ISSN 0141-9331 DOI: https://doi.org/10.1016/j.micpro.2016.06.013

Link to Leeds Beckett Repository record:
https://eprints.leedsbeckett.ac.uk/id/eprint/8639/

Document Version:
Article (Accepted Version)

$ABaT-FS$: Towards Adjustable Bandwidth and Temperature via Frequency Scaling in Scalable Memory Systems

Abstract—**In the context of the traditional memory design trade-off between memory width and frequency scaling (FS) which is employed to improve bandwidth, we propose $ABaT-FS$, a hardware scheduling mechanism that performs FS on ranks in scalable memory systems aiming to control rank operating temperature and reduce power: by defining FS intensity as the ratio between the amount of time FS is applied and the total selected scheduled cycle, $ABaT-FS$ on average is able to employ different rank frequencies. To understand the impact on memory bandwidth and temperature, we propose a design space exploration of $ABaT-FS$ with different FS intensities - FS applied to different percentages (0%, 25%, 50%, 75% and 100%) of the cycle time interval. Our findings show that for the 100% of FS intensity, bandwidth increases proportionally while rank temperature is increased of about + 23.7 degree Celsius%, and energy-per-bit magnitude is decreased in up to 67%.**

I. Introduction

The increase on the number of cores in the multicore era has been pushing the pressure on the memory system, with noticeably higher levels of contention and latency. Combined to this core growth, high demanding memory bandwidth applications further leverage these levels of memory pressure.

Double data rate (DDR) traditional rank - set of memory dynamic random-access memory (DRAM) chips which are simultaneously enabled by a common chip selection signal [24][25] - design towards bandwidth improvement is based on the following pair of parameters: (i) rank width and (ii) clock frequency - or simply frequency. (i) Rank width factor is determined by its total number of bits. Assuming 1:1 ratio between the number memory controllers (MCs, which translate cache requests into rank commands and data requests and responses) and ranks and by scaling memory width via scaling the number of MCs and ranks, memory width is scaled, thus allowing a larger parallelism or bandwidth. Despite that, this increase in memory parallelism through MC scalability is restricted by I/O-pin counts, area, density, and costs.

In traditional DDR-systems, due to these I/O pin restrictions, (ii) rank clock frequency - simply rank frequency - has been more intensively used than rank width as a mechanism to improve bandwidth. The results reported in [16] indicate a rank frequency scaling (FS) factor of about 8x along the design of DDR-family generations. Nonetheless, FS scaling significantly affects power/energy and temperature [7].

We see an opportunity to address I/O-pin count restrictions and combine it to FS in order to improve bandwidth and reduce power in scalable memory systems which employ DDR-like memories [3][19][20][30]. For example, high scalable optical- [3] and radio-frequency-based [20] (RF-based) interfaces present low pin counts, which allows (i) large memory widths via large MC-scalability, thus allowing bandwidth enhancement: in this study, memory width is represented by the total rank-count width, and it scales as rank counts width (and MCs) are scaled. Furthermore, in these systems (ii) FS can be applied at the rank level aiming to improve bandwidth. Moreover, the utilization of DDR ranks allows the utilization of FS at the rank level.

In this context of scalable memory systems, aiming to trade-off FS and its effects such as average rank temperature operation - defined as the average temperature of each individual chip that belongs to that rank - and energy-per-bit with bandwidth, we propose $ABaT-FS$ hardware scheduling algorithm. $ABaT-FS$ is able to select a range of different rank frequencies - which are set according to different levels of FS intensity. $ABaT-FS$ leverages the area of memory systems with the following contributions:

- We perform a design space exploration of different FS intensities during $ABaT-FS$ cycle. $ABaT-FS$ is able to change FS parameter in order to improve bandwidth. $ABaT-FS$ control unit (CU) controls rank temperature increase via adjusting FS intensity, i.e., the amount of time FS is applied to the ranks. To the best of our knowledge, it is the first time a FS mechanism is applied to scalable-memory systems which employ DDR memories.
- $ABaT-FS$ scheduling uses bandwidth - which can be observed as a function of rank frequency - in order to control rank temperature increase when performing FS.
- Based on real rank temperatures [7], $ABaT-FS$ employs an approach to estimate the behavior of rank temperature as a function of the frequency. $ABaT-FS$ can determine the temperature by indirectly controlling rank bandwidth and it can also provide different rank bandwidths based on FS intensity. To the best of our knowledge, it is the first time bandwidth and temperature are used as part of an FS mechanism on scalable memory systems.
- We perform a design space exploration to investigate memory frequency versus temperature trade-offs in $ABaT-FS$ by performing an architectural exploration of different time FS intensities. It is the first time that an architectural exploration of frequency versus temperature is performed in scalable-memory systems that employ DDR-memories.

In particular, considering optical and radio-frequency (RF) memory solutions which employ DDR ranks as representative of scalable memory systems, without any loss in generality, we select the category of RF scalable memory systems versus optical-ones as representative of scalable systems, given [29] their better integration with complementary metal-oxide-semiconductor (CMOS) and lower temperature sensitivity. The exploration of $ABaT-FS$ is performed at the level of

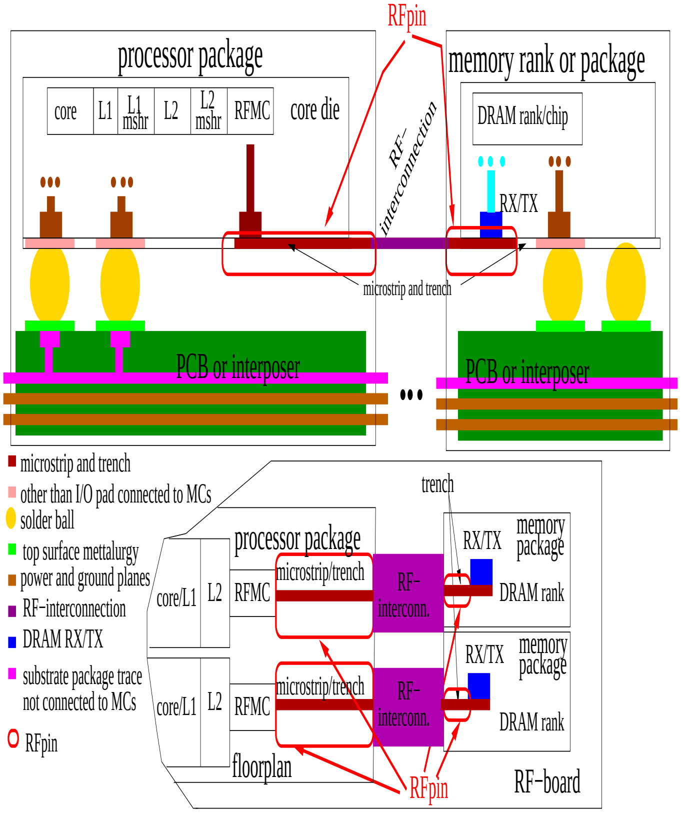

Fig. 1: RF-based scalable memory systems: memory path and reduced floorplan, modified from [20]; interconn.: interconnection

designing, modeling, simulation, and evaluation with several bandwidth-bound benchmarks. Due to a larger variety of voltage magnitudes available at the rank level, RF-based MCs (RFMCs) and memory channels, as well as power/temperature effects on the refresh mechanism, we leave a complete evaluation of dynamic voltage/frequency scaling (DVFS) and refreshing as a future effort.

This paper is organized as follows: Section II presents the background and motivations of $ABaT-FS$. and the trade-off between frequency and width, the likely effects on performance, power, and temperature. Section III describes $ABaT-FS$ strategies. Section IV presents the experiments and results, while Section V performs a sensitivity analysis of the bandwidth/temperature aspects, and Section VI the related work. Section VII concludes it.

II. Background and Motivation

We start this section having an overview of scalable memory systems, where we also show the bandwidth and power motivations applied to memory design, represented by its scalable width and rank frequency. In the sequence, we describe the effects of rank frequency over its default operating temperature.

To have a general overview of RF-based scalable systems, we exemplify in Figure 1 the typical structures utilized in DIMM Tree [30] (rank or DIMM - dual inline memory module) and RFiof [20]. The memory interfaces present RF-transmitter/receiver (TX/RX) elements placed at the MCs - to form a RFMC, i.e., a MC coupled with RF TX/RX, and at the ranks in order to respectively perform modulation and demodulation of digital cache requests/memory responses into RF waves. These systems present a proper low error-bit-rate (BER) over RF interconnection with structures to avoid crosstalk. To connect RFMCs to the ranks proper RFpins are utilized, which with RF-interconnection and RFMCs allow to achieve larger data rates with higher bandwidth-per-pin as reported in [20]. As a result, RFpins allow RFMC scalability, which enables to scale the memory width parameter.

Moreover, as a consequence of using scalable pins, in these scalable memory systems, bandwidth achieved by scaling MCs is significantly higher than a typical 2-MC or 4-MC system respectively employed in typical personal computers' microprocessors [19][20][30]. For example, in RFiof [20] bandwidth achieved with 32 RFMCs is up to 7.2x higher than a 5-MC DDR-based system assumed as a baseline in a 32-core processor configuration.

Another important advantage of these scalable systems is in terms of power. For example, in DIMM Tree [30], rank power consumption of the RF-interface elements (TX/RX and RF-memory channel) corresponds to 3-4% of a traditional DDR3 (double data rate type three synchronous DRAM) rank, therefore, of the same order of magnitude of a traditional memory system. Furthermore, given the larger bandwidths achieved, since there are a larger number of memory channels, replacing MCs with RFMCs is architecturally interesting in terms of energy-per-bit interconnection magnitudes as reported in [20].

Besides bandwidth and energy-per-bit advantages, RFMCs have significantly reduced areas when compared to traditional MCs [19], and are therefore similarly regarded as architectural replacements of MCs.

A. Temperature and FS Trade-offs

Rank temperature parameter is subject to rank operating frequency. For example, Micron report [7] indicates that rank temperature increases as memory clock frequencies are scaled. To illustrate these effects, we repeat the results obtained in Micron reports [7] in Figure 2, where we can observe that temperature increases of about 11 degree Celsius for a frequency augment of 333 MHz.

According to Micron [7], in order to avoid permanent damages at the rank banks, it is recommended to utilize temperatures lower than the maximum rank temperatures - range from 85 to 95 degree Celsius.

In the case of frequency increase, it is fundamental to determine the percentage of frequency increase and the duration of the interval along which the augment of the frequency is applied in order to guarantee its appropriate regular operation. In this context, we are assuming that rank frequency increase does not cause any type of circuit damage such as current microprocessors, where most of the cores are kept at lower clock frequencies and one or more cores' clock frequency is boosted [2][11] aiming to enhance performance of sequential applications and/or to reduce power.

We turn to the discussion of $ABaT-FS$.

III. ABaT-FS

$ABaT-FS$ is a hardware scheduling technique which applies FS to each rank in order to improve bandwidth. In $ABaT-FS$, FS is applied at the scheduled time intervals.

During the scheduled interval, ranks' operation temperature raises to higher levels than regular standard ones. To avoid rank exposition at higher temperatures for longer times, which are likely to damage its proper operation, all ranks are kept under higher frequency settings for a percentage of the time interval, and after this has elapsed, FS is disabled and the operating rank frequency returns to its default standard magnitude.

In $ABaT - FS$, there is a controller unit (CU) responsible for managing the operations of time scheduling, activation/de-activation of FS, and indirect temperature control. Each time CU unit performs a scheduling round, all the ranks available are submitted to the higher frequency, while during the de-scheduled intervals ranks start to return to their default operating frequency, in order to get cooler and avoid temperature stressing as further described.

A. Scheduling

Before describing the scheduling mechanism in $ABaT - FS$, we define the $ABaT - FS$ total schedule cycle time. Assuming the most general memory access method performed as interleaving each cache lines along the respective rank, we define the $ABaT - FS$ total schedule time as the time to perform memory accesses corresponding to 20x the total number of ranks.

Therefore, the total scheduled cycle time is measured in number of memory accesses. For example, considering the total number of ranks as 32 - such as in current scalable memory systems [20], the total scheduled cycle time corresponds to 640 memory accesses. The scheduled magnitude (640 memory accesses) is selected to guarantee that enough memory accesses are performed in order to have FS impacting bandwidth and temperature.

In the scheduling mechanism, the key parameter is $ABaT - FS$ FS intensity. FS intensity is defined as the ratio between the amount of time while FS is kept active in respect to the total scheduled cycle time. We express this ratio in terms of percentage of the total scheduled cycle time. In the CU there is a dedicated register which contains FS intensity magnitude (FS_intensity_register). Different settings of this parameter result different average clock frequencies as explained in the sequence. In $ABaT - FS$:

- the minimum FS intensity - which is 0% - corresponds to not applying FS.
- the maximum FS intensity - which is 100% - corresponds to apply maximum FS level (i.e., an increase of 100% clock frequency). In order to evaluate its bandwidth and energy-per-bit benefits, We assume that this maximum increase does not cause a temperature increase that damages any of the rank device elements.

By regulating FS intensity, larger bandwidths and operating temperatures can be achieved. We further detail how FS intensity and its effects on bandwidth and temperature in the following subsections.

We now address how FS intensity is utilized in $ABaT - FS$ by defining:

- The interval along which FS is applied is defined as $Hint$. We generically define higher frequencies along $Hint$ as $Hfreq$.
- The de-scheduled interval, where ranks are set with the same default baseline operating frequency, defined as $Dfreq$. An interval with these features is defined as $Dint$.
- $Hint + Dint$ performs the total scheduled time, which as previously defined corresponds to 20x the total number of ranks.

The concept of FS intensity is simply defined as:

$$FS_intensity \ = \ Hint \ / \ (Hint \ + \ Dint) \tag{1}$$

The larger the FS intensity, the larger $Hint$ comparatively to $Dint$. The smaller $Hint$, the smaller FS intensity, the smaller the bandwidth. We further mathematically describe how FS intensity affects bandwidth.

After $Hint$ time, rank temperature is likely to be higher, which we define as Th. Along the interval which ranks are operating at their default baseline frequency $Dfreq$, their temperature is defined as Td. Based on these definitions, we can establish the following straightforward formulations:

$$Hfreq > Dfreq \tag{2}$$
$$Th > Td \tag{3}$$

We discuss about these parameters and their implications in the following sections.

B. Bandwidth and FS

To understand the approach adopted in $ABaT - FS$, we assume that all ranks have the same operating frequency and define the dependency between bandwidth and rank frequency - in scalable memory systems as:

$$bandwidth = totalrankwidth * rankfrequency \tag{4}$$

Giving we have multiple ranks, total rank width is obtained via the product of each individual rank width - defined as rank width, and assumed identical rank width over all ranks - with rank frequency:

$$bandwidth = number\ of\ ranks * rank\_width * rank\ frequency \tag{5}$$

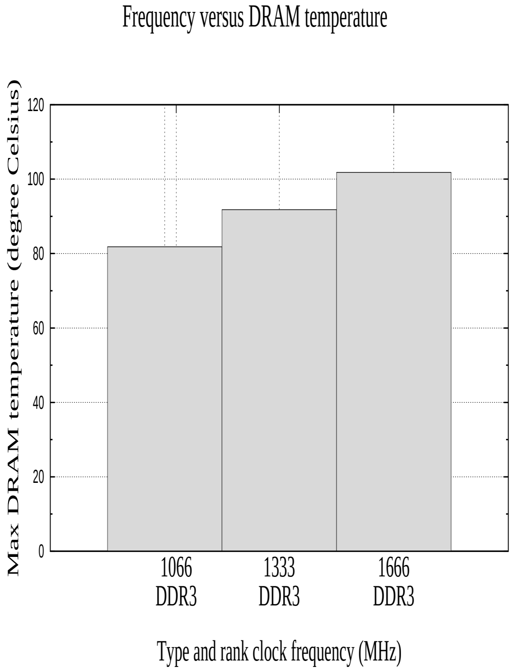

Fig. 2: temperature versus frequency: temperature for 1066MHz and 1333MHz are repeated from [7], while the one for 1666MHz is extrapolated.

Therefore, as we increase rank frequency, bandwidth increases proportionally.

In the sequence, we calculate equation 5 for $Hfreq$ and $Dfreq$ frequencies:

$$bandwidthH = number\ of\ ranks * rank\_width * Hfreq \tag{6}$$

$$bandwidthD = number\ of\ ranks * rank\_width * Dfreq \tag{7}$$

Since $bandwidthH$ is achieved when CU applies $Hfreq$, i.e., during $Hint$ and $bandwidthD$ is achieved when CU applies the baseline $Dfreq$ - during $Dint$, the average geometric mean bandwidth during the entire interval, which we define as $bandwidthA$ is:

$$bandwidthA = (bandwidthH\ * Hint + bandwidthD\ * Dint)/ \\ (Hint\ + Dint) \tag{8}$$

which we can re-write as:

$$\begin{aligned} bandwidthA = (number\ of\ ranks * rank\_width * \\ (Hfreq * Hint + Dfreq * Dint))/ \\ (Hint\ + Dint) \end{aligned} \tag{9}$$

or

$$\begin{aligned} bandwidthA = (number\ of\ ranks * rank\_width * \\ (Hfreq * Hint/(Hint\ + Dint) + \\ Dfreq * Dint\ /(Hint\ + Dint))) \end{aligned} \tag{10}$$

Combining the definition of FS (equation 1), $Dint\ = Dint\ +\ Hint\ - Hint$, and the latter equation, we have:

$$\begin{aligned} bandwidthA = (number\ of\ ranks * rank\_width * \\ (Hfreq * FS\ + Dfreq(1\ - FS)))\ , \end{aligned} \tag{11}$$

which represents the core of $ABaT - FS$ mechanism to improve bandwidth. Therefore, the larger the FS intensity, the larger the bandwidth. The smaller FS intensity, the smaller the bandwidth.

C. $ABaT - FS$ Scheduling Algorithm

In Algorithm 1, $ABaT - FS$ scheduling algorithm is described in a pseudo-language. $ABaT - FS$ algorithm can be summarized as follows:

1) To control FS intensity and its effects on bandwidth and temperature, CU utilizes FS_intensity_register. This register indicates the number of memory accesses of the $Hint$ interval, i.e., while ranks are subject to higher frequency $Hfreq$. While along $Dint$ interval, $ABaT - FS$ configures ranks with the default frequency $Dfreq$).
2) CU unit estimates rank temperature limits by using equation 14, which we further discuss.
3) For each address access, as previously assumed accessed on an interleaving fashion, at the CU there is a memory access counter which is incremented and reset after the total scheduled time is reached.
4) For all ranks $Hfreq$ is set when the memory access counter is smaller than the FS_intensity_register, while $Dfreq$ is set when this counter is larger or equal than the FS_intensity_register.

```
while (there are memory addresses to be interleaved) do
    if (memory_access_counter $<=$ FS_intensity_register)
    then
        set_rank_frequency(Hfreq);
    end
    else
        set_rank_frequency(Dfreq);
    end
    memory_access_counter++;
    if (memory_access_counter $>$ total scheduled time)
    then
        memory_access_counter = 1;
    end
end
```

Algorithm 1: ABaT-FS bandwidth algorithm; memory_access_counter initialized with 1.

D. Temperature control in ABaT-FS

The core of $ABaT - FS$ mechanism is based on the trade-off between $Hfreq$, $Dfreq$, and the FS intensity. In order to approach temperature control it is fundamental to understand the temperature boundaries which can restrict rank operation and bandwidth. According to this strategy, we start this subsection by investigating the typical temperature DRAM devices are designed to operate and the impact of operating at temperatures larger than supported. In the sequence, within these restrictions we propose the strategy $ABaT - FS$ performs temperature control.

According to Micron thermal report [7], typical DRAM device temperature operations are in the range +0 to +95 degree Celsius. Furthermore, according to the same manual, these temperature boundaries should be followed in order to guarantee proper functionality of the ranks devices.

Considering rank maximum temperatures and bandwidth/power trade-offs previously discussed, we can generically define the $Dtemp$ as the temperature ranks achieve when subject to $Dfreq$, which we can generically express as:

$$Dtemp = f(Dfreq) \tag{12}$$

Similarly, we define $Htemp$ for the $Hfreq$ case:

$$Htemp = f(Hfreq) \tag{13}$$

Since $Htemp$ and $Dtemp$ are functions of the frequency ($Hfreq$ and $Dfreq$ respectively), and $ABaT - FS$ is responsible for rank frequency, $ABaT - FS$ indirectly controls the rank temperature.

To estimate $ABaT - FS$ temperature control, we utilize DDR3 ranks and Micron thermal reports [7]. By performing a significant research in these thermal reports we develop a polynomial empirical interpolation of temperature as a function of rank frequency - such as illustrated in Figure 2. As a result, rank temperature can be expressed as a function of the rank frequency (freq) as follows:

$$temperature = -5.21848 * 10^{-6}\ * f^2 + 0.0456803\ * f + 40.1809\ , \tag{14}$$

For instance, according to this equation, for ranks configured with $Hfreq$ of 1333 MHz, temperature achieves $Dtemp = 91.8$ degree Celsius as observed in Figure 14. The proposed formulation is applicable to the specific DDR3 ranks utilized in [7] experimentation. However, it can be extended to a large variety of DDR2 (double data rate type two synchronous DRAM), DDR3, and DDR4 (double data rate fourth generation DRAM) families when a larger range of DDR rank features are considered.

Given that this formulation respects the typical upper boundaries of temperature operation (+95 degree Celsius [7]), FS magnitudes indirectly should also follow it. We further exemplify how in $ABaT - FS$ FS intensities follow temperature behaviors represented by equation 14 in Section IV.

E. Rank Power and Frequency

To understand frequency effects on the rank power behavior, we use the formulation developed by Micron [26], which derives the power by de-rating - decrease the rank frequency rate - the power spent at the default frequency $Dfreq$ to determine the power on a generic frequency Dfreq ($rank\_pw(freq)$):

$$rank\_pw(freq) = rank\_pw(Dfreq)\ *\ freq/Dfreq \qquad (15)$$

with $rank\_pw(freq)$ meaning the rank power configured at a frequency freq. From this equation, assuming $freq$ set with $Hfreq$, $rank\_power(freq)$ also increases. Now that we have determined the effects on power, we can determine the effects on the energy-per-bit levels.

F. Rank Energy-Per-Bit

The straightforward relationship we can stablish among energy, power, and bandwidth can be expressed as:

$$energy - per - bit = rank\_pw\ /\ rank\_bandwidth \qquad (16)$$

where we observe that $memory\ energy - per - bit$ depends on the behavior of the $total\ power$ spent and on the bandwidth demanded by the application which, given this specific behavior, we approach in Section IV.

G. Combining Temperature, Power, and Bandwidth

(i) Application bandwidth demands, (ii) power/energy savings, and (iii) rank temperature stress are some of the parameters which can trigger $ABaT - FS$ scheduling. For example, equation 11 can be used individually - such as in the presented $ABaT - FS$ mechanism - or combined to the application needs or OS.

In case (i), applications which have a bandwidth-bound phase such as the parallel initialization of larger matrices, or even bandwidth-bound applications such as NPB [1], can benefit from the increase on bandwidth. In this case, with the correct settings provided by the application, $ABaT - FS$ can be tunned to increase bandwidth performance (Equation 13), if correctly tunned/matched to phases along the execution of the application.

Case (ii) should be combined to case (i) in order to have the application itself optimizing its bandwidth (performance) and power. By speeding it up via rank bandwidth increase or by reducing memory bandwidth via reducing frequency increase in application sections - such as cache-bound phases, power/energy could be approached.

In case (iii), equation 13 can be used to control upper rank temperatures and, in case of inefficient, deficient, or cooling failure, it can be used to decrease rank frequency, as a consequence, its temperature. An example would be combine $ABaT - FS$ with OS scheduling, so that failures or rank device damages due to high rank temperatures could be avoided.

tool	description
Cacti [5]	cache latencies configured with
DRAMsim [6]	Capture memory transactions from DRAMsim and simulate them with 4 to 32 RFMCs, as well as 5 MCs to emulate the baseline. Respond to M5 with the result of the memory. transaction. Determine power spent in each rank. Determine the number of memory accesses.
M5 [27]	configured as 32-core, OOO processor generates memory transactions, which are passed to DRAMsim [6].
MSHR counts from [15]	configured as 32-core, OOO processor since the 3Dstacking and the RF-systems simulated both have multiple MC-system
RF-crossbar	implemented in M5 [27] with RF settings from [17][23]
RF-communication delays	RF-circuitry modeling and scaling [17][22]
(TX/RX) power	RF-modeling predictions [30]
Temperature Determination	Utilize bandwidth obtained when performing bandwidth versus FS intensity behavior and equation 14

TABLE I: methodology: tools and description

Given that the previously mentioned parameters also depend on the application behavior and/or OS, we leave a deep investigation of these aspects as a future research.

H. Implementation and Overhead of ABaT-FS

We assume that the CU is implemented as a hardware unit together with the decoder unit of the pipeline stage.

To implement $ABaT - FS$, several hardware elements are required:

1) a counter to measure the total cycle time in terms of memory accesses.
2) FS_intensity_register to adjust $Hint$ interval.
3) A small-magnitude memory to store the temperatures and frequencies in order to implement equation 14, i.e., perform the upper temperature limit detection.

We briefly describe the overhead of this unit in Section V.

IV. Experimental Section

In this section, we perform a series of experiments to evaluate $ABaT - FS$ mechanism in terms of bandwidth/latency versus FS intensity, temperature, and energy-per-bit magnitudes.

A. Methodology

Before describing the methodology adopted, we describe the baseline. We define our baseline with 32 cores and 32 MCs in order to have scalable MC counts and to maintain the ratio core:MC the same (32:32). In addition, in this baseline configuration we set the rank frequency to represent the FS intensity of 0%, which correspond to the lower rank frequency of 666 MHz. We further justify, describe, and discuss about rank frequency settings.

To have a global picture of the methodology employed in this study, we have listed all the simulators employed and the description of their purpose in Table I. The general methodology employed to obtain bandwidth is adopted from [18]: by using bandwidth-bound benchmarks to stress the memory system, we combine M5 [27] and DRAMsim [6] simulators as follows. A 32-core processor model is created in M5 [27], and as memory transactions are generated in M5 upon benchmark execution, these are captured in DRAMsim [6] which is properly configured with core:MC ratio of 32:32 (previously explained). In the sequence DRAMsim responds to M5 with the result of each transaction. In this environment, we confirm the appropriate calibration of the bandwidth of one rank: about 2.0 GBytes/s as indicated in the manuals [25].

We employ a 4.0GHz (Alpha ISA) and 4-wide out-of-order (OOO) core (as current microprocessors) to guarantee significant memory pressure, while having RFMCs at 2.0GHz. We used Cacti [5] to obtain cache latencies and adopted miss-status handling register (MSHR) counts similarly to current microprocessors [15]. Processors are connected as a clustered architecture, and we utilize scalable L2-MSHR structures (based on [13]), while the 1 MB/core L2 caches are interconnected via an RF-crossbar with 1-cycle latency (adopting same timing settings of [17][23]: 200ps of TX-RX delays, plus the rest of the cycle to transfer 64 Bytes using high speed and modulation). The RF-crossbar upper bandwidth was (i) designed so that when ranks are scaled it does not restrict total bandwidth; and (ii) to resemble real delays [17].

As each RFMC is connected to a different DDR rank, cache lines interleaved along the RFMCs are interleaved along different ranks. Finally, we employ closed page mode (server environment) in all experiments. To finalize, all architectural parameters are summarized in Table II.

We turn to the methodology employed to obtain power and energy-per-bit parameters. To obtain total power, we consider the DRAMsim power models, which follow Micron formulation [25]: total power includes the power spent on all ranks and interconnection. To determine the total energy-per-bit spent, we employ the power magnitudes obtained in DRAMsim power infrastructure and which are also based on Micron formulations [25] and combine them to the memory bandwidth extracted from the benchmark - if it is designed to measure bandwidth, otherwise from the number of memory transactions (DRAMsim) and execution time. We compare these energy-per-bit results to the ones predicted by the equation 16.

To model delays involved in the RF communication, we have considered RF-circuitry modeling and scaling proposed by Frank Chang et al. [17][22]. In these models, modulation and line separation are taken into account targeting to keep a low bit error rate (BER). Finally, these models are validated with prototypes for different transmission lines [9][22], while follow the International Technology Roadmap for Semiconductors (ITRS) [12].

Similar to the methodology employed in [15] which was designed to stress the memory system, we have selected bandwidth-bound benchmarks with a significant number of misses per kiloinstructions (MPKI): (i) STREAM [21] suite, which we decompose in its four sub-benchmarks (Copy, Add, Scale, and Triad); (ii) pChase [28] with pointer chase sequences randomly accessed in order to estimate bandwidth and latency with a random behavior; (iii) SP and MG from NPB [1] as representative of scientific bandwidth-bound applications. STREAM and pChase are designed to evaluate bandwidth, while the latter is also designed to evaluate latency.

Core	4.0 GHz, OOO-Core, 4-wide issue, turnament branch predictor
technology	22 nm
L1 cache	32kB dcache + 32 kB icache; associativity = 2 MSHR = 8, latency = 0.25 ns
L2 cache	1MB/per core ; associativity = 8 MSHR = 16; latency = 2.0 ns
RF-crossbar	latency = 1 cycle, 64GB/s
RFMC trans. queue	1 to 32 RFMC; 1 MC/core, 2.0GHz, on-chip buffer size = 32/MC, close page mode
Memory rank	DDR3 666MT/s to 1333MT/s via FS 1 rank/MC, 1GB, 8 banks, 16384 rows, 1024 columns, 64 bits, Micron MT41K128M8 [25] tras=26.7cycles, tcas=trcd=8cycles
RF interconnection size, delay	length of 2.5 cm, 0.185ns 2.5 cm, 0.185ns

TABLE II: modeled architecture parameters

Benchmark	Input Size	read : write	MPKI
Copy, Add, Scale, Triad (STREAM)	8.5Mdoubles per core, 2 interations	2.54:1	54.3
pChase	30.1MB/thread, 3 iterations, random	158:1	116.7
Scalar Pentadiagonal:SP (NPB) Pentadiagonal (NPB)	Class A, 2 iterations 2 iterations	1.9:1	11.1
Multigrid:MG (NPB) (NPB)	Class A, 3 iterations 3 iterations	76:1	16.9

TABLE III: benchmarks and input sizes

Table III lists the benchmarks experimented, input sizes, read-to-write rate, and L2 MPKI obtained in the experiments. In all benchmarks, parallel regions of interest are executed until completion, and input sizes guarantee that all memory space used is stressed. The adopted input sizes are selected based on simulation times and stressing memory capability. Average results are calculated based on harmonic average.

We perform the evaluation of $ABaT - FS$ for different FS intensities to understand the effects on bandwidth and temperature. The experimented FS intensities cover the range 0%, 25%, 50%, 75%, and 100%.

As previously mentioned, we have selected low magnitude frequency DDR3-ranks (666 MHz/Mtransactions/s/ or MT/s based on the DDR3 model Micron MT41K128M8 of 1GB [25], and described in Table II) according to the methodology developed in [19][20], due to two different reasons: (i) maximum crossbar bandwidth of 64GB/s, which limits the total data rate; (ii) since the rank scalability is much higher in RF-scalable memory systems [20][30], the utilization of a lower data-rate rank - when compared to the employed in typical microprocessors - is advisable in order to save power and while bandwidth improving is obtained by scaling MCs/ranks. Therefore, we adopted $Dfreq = 666MHz$, while $Hfreq = 1333MHz$. We assume that this rank is able to work properly when FS of 100% is applied, i.e., $Hfreq = 1333MHz$ for the scheduled time. Therefore FS intensity range covers $Hfreq - Dfreq = 666MHz$.

In order to determine rank temperatures, we first determine the behavior of bandwidth versus FS intensity. Assuming that STREAM benchmarks have a read-to-write ratio behavior similar to the benchmark utilized in Micron thermal experiments [7], we utilize the bandwidth magnitudes obtained in these benchmarks as inputs for equation 14 in order to obtain rank temperatures.

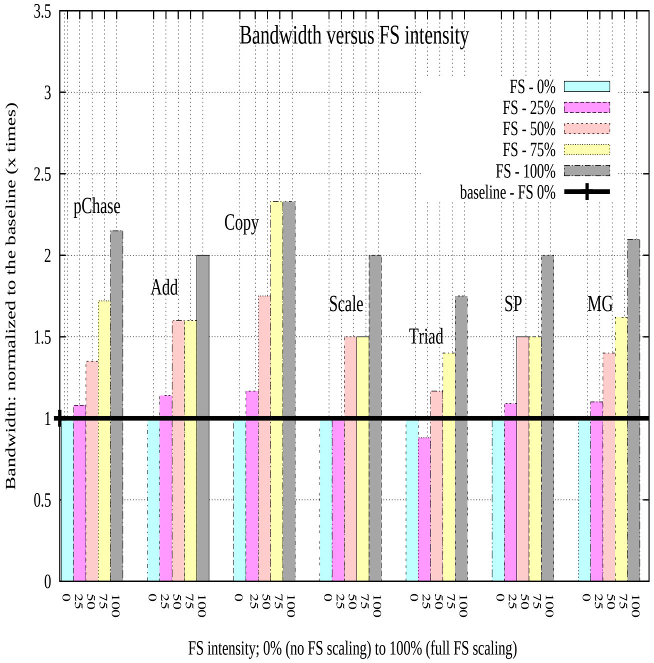

Fig. 3: FS intensity versus bandwidth

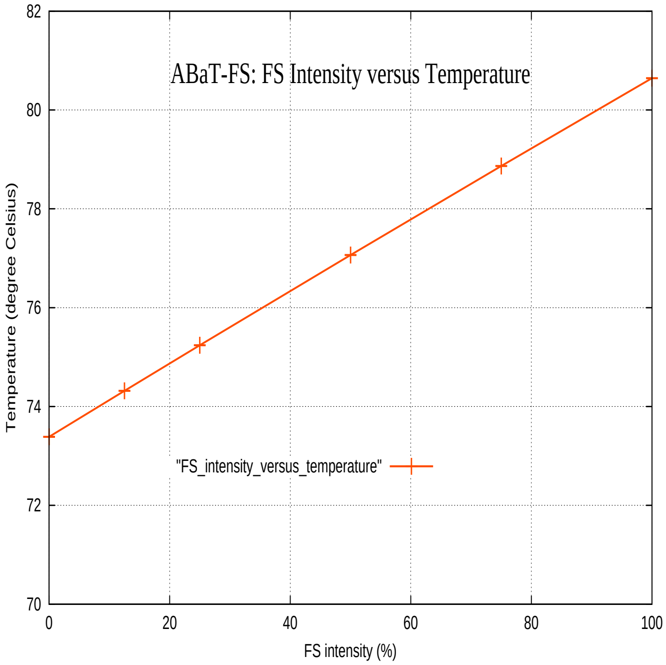

Fig. 4: FS intensity versus temperature

B. Bandwidth versus FS intensity

Since in this paper we are focusing on $ABaT - FS$ FS mechanism, we are particularly evaluating $ABaT - FS$ bandwidth effects individually (case (i), Subsection III-G). Bandwidth results of the different benchmarks are plotted in Figure 3. In this figure, we can observe the core of $ABaT - FS$ algorithm, represented by equation 11, along each benchmark: as we increase FS intensity, bandwidth increases proportionally. Similarly, the smaller FS intensity, the smaller the bandwidth.

The expected bandwidth behaviors shown are according to the predicted equation 11 in regards to bandwidth versus FS scaling. However we have also noticed a not expected bandwidth behavior for pChase, Copy and MG: for the largest FS intensity, bandwidth has increased more than expected - i.e., more than 2x. By deeply investigating the statistics of our simulation environment, we have concluded that this behavior happens due to small benchmark input sizes, given that similar rank frequency ranges and larger input sizes have not produced these behaviors in other reports [19][20]. By adopting larger input sizes, bandwidth corresponding to the lowest FS intensities would present a larger magnitude, which would make the largest ones lower when performing relative bandwidth calculation. Finally, the best bandwidth results are obtained for pChase, Copy, and MG, and the worst ones for Triad, and Add.

C. Bandwidth versus Temperature

Having in mind that $Dfreq = 666MHz$ and $Hfreq = 1333MHz$ - therefore $Hfreq = 2 * Dfreq$, we proceed temperature control determination by plotting equation 14 with different FS intensities. The result of this investigation is shown in Figure 4.

To proceed rank temperature determination, given that STREAM benchmarks [21] are developed to measure bandwidth, we employ the bandwidth obtained in these benchmarks as x-axis inputs in Figure 4. As a result we obtain the correspondent temperature range (Y axis), in the X-Y-degree-Celsius range. Therefore, 100% FS intensity range corresponds to an increase of 23.7 degree Celsius. As previously explained, these temperatures are within predicted rank temperature ranges [7].

The temperature observed in Figure 4 shows that temperature increases proportionally, via the increase of FS rank intensities. The maximum temperature achieved is about above +80.7 degree Celsius for a maximum of 100% of FS intensity range (666MHz to 1333MHz). In this case the temperature achieved when having FS 100% is not a restriction to the operation of the ranks, i.e., the memory system.

However, if the obtained temperature constitutes a restriction, it is advisable to employ lower FS intensities to guarantee its proper operation. For example, according to this figure, if rank temperature is restricted to +79 degree Celsius, a maximum FS intensity of 75% could be employed. In this case, CU should set FS intensity to 75%. Using this strategy in $ABaT - FS$, temperature is limited by bandwidth correspondent to 75% of FS intensity.

The prediction of these temperatures here developed is based on the Micron thermal experiments [7] previously described, which take into consideration the execution of a specific benchmark, with a specific read/write memory hit/miss ratio. Therefore, it is necessary to perform a similar investigation of read/write memory hit/miss ratios along the benchmark utilized in Micron report [7] in order to determine a more general formulation where we can express temperature as a function of the bandwidth and FS intensity. In addition, if higher FS intensities are required, more appropriated cooling mechanisms should be utilized, which we leave as a further investigation.

D. Latency versus FS intensity

In order to estimate latencies, we plot pChase [28] latencies in Figure 5. The results confirm a significant latency reduction, which is about 54% for a frequency increase of 100%. It is

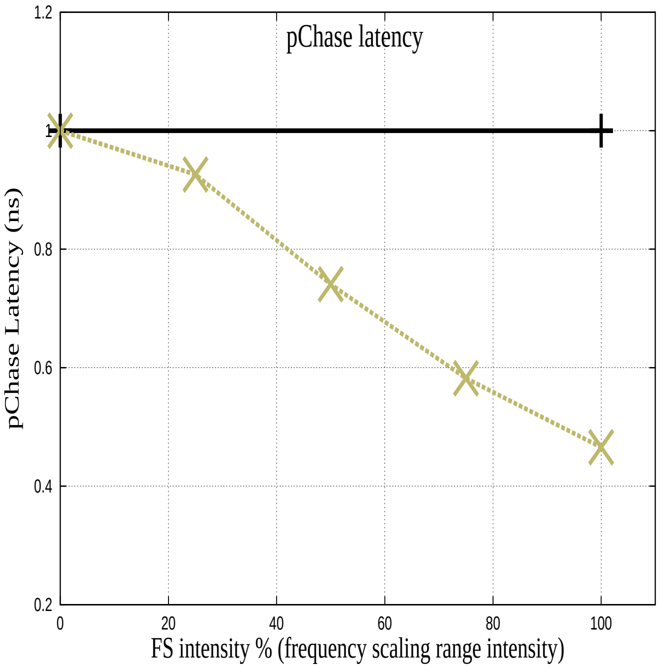

Fig. 5: FS intensity versus latency

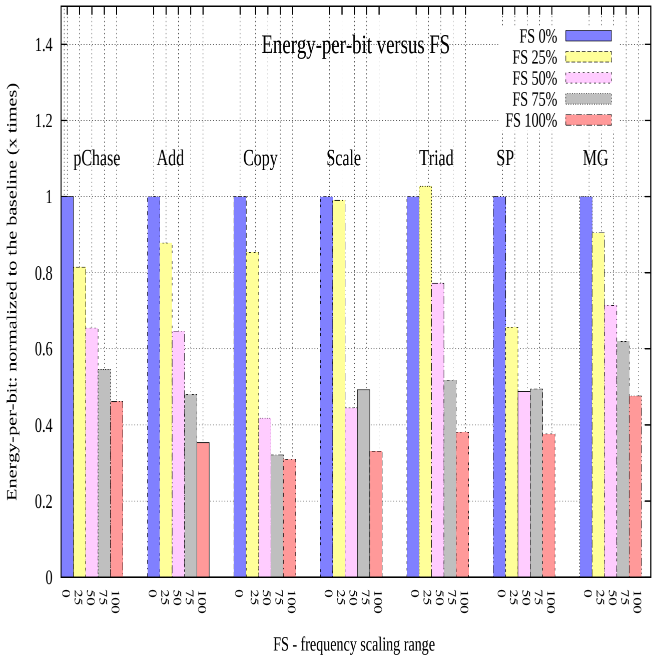

Fig. 6: FS versus energy

important to mention the generality of this result since it is valid for for pChase with chases set with random behavior.

As a more general conclusion, based on bandwidth and latency experiments, we observe the expected behavior of Litte's law [14], i.e., as memory bandwidth increase due to FS increase is followed by memory latency reduction.

E. Memory energy versus FS intensity

To understand the behavior of memory energy-per-bit we consider bandwidth results and power magnitudes. Regarding bandwidth, we have evaluated in Subsection IV-B and noticed that increase of FS intensities impacts rank bandwidth proportionally. Regarding power magnitudes, as indicated in equation 15, we expect a unproportional increase of the power spent at each rank when larger FS intensities are employed since it includes not only the power spent at the rank, but also those of each channel and I/O pin power.

Finally, observing the memory energy-per-bit experiment plotted in Figure 6, we confirm the predicted memory energy-per-bit behavior developed in Subsection III-F: by combining the bandwidth obtained and unproportional power magnitudes expected, we observe that as we increase FS intensities memory energy-per-bit decrease unproportionally. The maximum energy-per-bit magnitude reduction (about 67%) observed is obtained for Copy benchmark.

V. Sensitivity Analysis

We also perform a sensitivity analysis to assess the impact of the key aspects in $ABaT - FS$ design: rank temperature, number of memory ranks/channels, proportionality of MCs and ranks, maximum acceptable performance degradation, and benchmarks/number of cores.

1) Rank Temperature: The estimation we developed (which turned to equation 14) shows a baseline temperature $Dtemp$ as +67 degree Celsius for the baseline frequency $Dfreq = 666MHz$. According to the estimated temperatures in Figure 4, the maximum temperature achieved for FS intensity of 100% is about $Htemp = +80.7$ degree Celsius which is smaller than the maximum levels of accepted operating temperatures (+95 degree Celsius [7]). As a result, baseline temperature has increased of 23.7 degree Celsius (+80.7 - 67) as a consequence of FS application in $ABaT - FS$.

As previously stated, these temperature estimations are based on the benchmarks run in the experiments developed by Micron [7], which stress the memory system. In this report, to consider appropriate stress conditions, we have assumed that STREAM benchmarks have similar stress conditions in order to obtain proper temperature estimation. Therefore, for less memory intensive benchmarks with smaller MPKI ratios - such as MG and SP benchmarks - temperature levels are likely to be smaller than the predicted ones, thus less restrictive in terms of bandwidth and allowing larger levels of FS intensities.

2) Number of MCs and Ranks: We have performed the evaluation of FS impact utilizing 32 MCs and 32 ranks. Since FS directly affects the behavior of each rank, it does not depend on the number of MCs.

By regulating FS intensity in $ABaT - FS$ via controlling the time while ranks are submitted to FS, different frequencies can be set at the rank level, which allows to achieve different levels of bandwidth. By accounting these different levels of bandwidth for each rank, the entire memory system is able to have different bandwidth levels as demonstrated in Subsection IV-B.

3) Proportionality of MCs and ranks: In this study we have assumed the same number of MCs and ranks (ratio MCs:ranks is 32:32). By applying $ABaT - FS$ on memory systems where the proportion of ranks is larger than MCs, we are likely to observe similar effects to the presented here, since similarly to the previous analysis regarding the number of MCs/ranks, FS is applied at the rank level.

4) Maximum Acceptable Performance Degradation: We consider $ABaT - FS$ of low complexity in terms of overheads, given that very simple elements such as registers (FS_intensity_register) and a table to implement equation 14 are employed in order to control temperature as a function of FS intensity. These elements are of simple implementation complexity, therefore thus not requiring a significant amount of chip area and cost to be implemented.

5) High-Bandwidth Benchmarks and Number of Cores: In this study, as related reports [20][19] which evaluate scalable memory systems, we have employed benchmarks with a 32-core-OOO-microprocessor dimensioned to evaluate 32 MCs/ranks. By observing and comparing the related studies [19] and [20], which respectively employ 16 and 32-core-OOO-microprocessors, we expect similar effects by employing a larger number of MCs/ranks - such as 64 or 128 MCs/ranks - and re-dimensioning benchmark input sizes.

VI. Related Work

HMC [10] is a recent memory solution designed to target 3Dstacking and off-chip memory systems. In the case of off-chip memories, either HMC or this study use an external memory package as memory ranks. HMC organizes its memory package by employing sets of banks on the memory dies, and processor/memory communication is done via serial/deserial operations, with 10-Gbit/s-I/O-links. Although MC-scalability is smaller in HMC when compared to the scalable memory systems here studied, FS scaling combined to MC-scalability can be generically applied in HMC.

Udipi [4] proposes a series of hardware and software mechanisms to approach the bandwidth problem via the utilization of optical-based interfaces, appropriated memory organization for taking advantage of optical transmission, and MC optimizations to improve power and performance. This study targets to study FS effects on the performance aspects represented by latency and bandwidth not only using RF transmission, but also in optical ones which utilize DDR-like memories.

Memscale [8] is a set of software and hardware mechanisms which include OS policies, and specific hardware power techniques to trade-off memory energy and performance in typical memory systems. Differently, in this study $ABaT-FS$ employs FS to improve bandwidth in scalable memory systems, which is also used to control rank temperatures.

While DIMM Tree [30], RFiop [19], and RFiof [20] are RF-memory organizations which aim to improve power and performance, the two latter ones also focus on pad/pin reduction and RFMC scalability. Bandwidth versus FS trade-offs here investigated are applicable to any of these systems, as well as to optical-based ones.

VII. Conclusions and Future Work

In this study we proposed $ABaT-FS$ mechanism and evaluate its impact on bandwidth, latency, and energy-per-bit. Our analysis demonstrates that memory energy-per-bit magnitudes decrease as bandwidth is increased via FS intensity increase. Moreover, in $ABaT-FS$, we developed an empirical method to predict the behavior of temperature versus FS, and how it can be indirectly controlled via bandwidth/FS intensity.

We plan an integration of $ABaT-FS$ to deal with different benchmarks simultaneously running and to the operating system (OS) to understand how bandwidth and temperature can be specifically tuned to match the demands of these benchmarks. We also plan an extension of this work by improving temperature estimation to a broad variety of families of ranks, as well as incorporating FS effects on the refresh mechanism.

References

[1] NAS Parallel Benchmarks. Accessed date: 03/11/2013; http://www.nas.nasa.gov/Resources/Software/npb.html/.

[2] AMD Reveals Details About Bulldozer Microprocessors, 2011. accessed date: 14/08/2013 - http://www.xbitlabs.com/news/cpu/display/20100824154814 _AMD_Unveils_Details_About_Bulldozer_Microprocessors.html.

[3] Amit Hadke. Design and Evaluation of an Optical CPU-DRAM Interconnect. In *Master of Science Thesis*, pages 1–90, University of California at Davis, USA, 2009. University of California, Department of Computer Science.

[4] Aniruddha N. Udip. Designing Efficient Memory for Future Computing Systems . In *PhD Thesis*, pages 1–126, Utah, USA, 2012. University of Utah, School of Computing.

[5] CACTI 5.1. Accessed Date: 04/16/2013; http://www.hpl.hp.com/techreports/2008/HPL200820.html.

[6] David Wang et al. DRAMsim: a memory system simulator. *SIGARCH Comput. Archit. News*, 33(4):100–107, 2005.

[7] DDR3 Thermals. Accessed date: 04/08/2014 ; http://www.micron.com/~/media/Documents/.../ddr3_thermals_nonNDA.pdf.

[8] Deng, Q. et al. Memscale: active low-power modes for main memory. In *Proceedings of the Sixteenth ASPLOS*, pages 225–238, New York, NY, USA, 2011. ACM.

[9] G. Byun et al. An 8.4Gb/s 2.5pJ/b Mobile Memory I/O Interface Using Bi-directional and Simultaneous Dual (Base+RF)-Band Signaling. IEEE, 2011.

[10] Hybrid Memory Cube Specification 1.0. Accessed date: 03/03/2014 ; http://www.hybridmemorycube.org/.

[11] The Intel Xeon Processor E7 v2 Family. Accessed date: 03/10/2014; http://www.intel.com/content/www/us/en/processors/xeon/xeon-processor-e7-family.html?wapkw=intel+xeon+e7.

[12] ITRS HOME. Accessed date: 09/12/2012 ; http://www.itrs.net/.

[13] J. Tuck et al. Scalable Cache Miss Handling for High Memory-Level Parallelism. In *MICRO*, pages 409–422, DC, USA, 2006. IEEE.

[14] Little, J. D. C. (1961). ”A Proof for the Queuing Formula: L = W”. Operations Research 9 (3): 383387. Accessed date: 03/20/2014 ; http://dx.doi.org/10.1287/opre.9.3.383.

[15] Loh, Gabriel H. 3D-Stacked Memory Architectures for Multi-core Processors. In *ISCA*, pages 453–464, DC, USA, 2008. IEEE.

[16] LPDDR4 Moves Mobile. Mobile Forum 2013, presented by Daniel Skinner, Accessed date: 02/03/2014; http://www.jedec.org/sites/.../D_Skinner_Mobile_Forum_May_2013_0.pdf.

[17] M. Frank Chang et al. CMP Network-on-Chip Overlaid With Multi-Band RF-interconnect. In *HPCA*, pages 191–202, 2008.

[18] Marino, M. D. On-Package Scalability of RF and Inductive Memory Controllers. In *Euromicro DSD*. IEEE, 2012.

[19] Marino, M. D. RFiop: RF-Memory Path To Address On-package I/O Pad And Memory Controller Scalability. In *ICCD, 2012, Montreal, Quebec, Canada*. IEEE, 2012.

[20] Marino, M. D. RFiof: An RF approach to the I/O-pin and Memory Controller Scalability for Off-chip Memories. In *CF, May 14-16 , Ischia, Italy*. ACM, 2013.

[21] McCalpin, J. D. Memory Bandwidth and Machine Balance in Current High Performance Computers. *IEEE TCCA Newsletter*, pages 19–25, Dec. 1995.

[22] M.C.F. Chang et al. Advanced RF/Baseband Interconnect Schemes for Inter- and Intra-ULSI Communications. 52:1271–1285, Jul 2005.

[23] M.C.F. Chang et al. Power reduction of CMP communication networks via RF-interconnects. In *MICRO*, pages 376–387, Washington, USA, 2008. IEEE.

[24] Memory rank. Accessed date: 08/28/2014 ; http://en.wikipedia.org/wiki/Memory_rank.

[25] Micron manufactures DRAM components and modules and NAND Flash. Accessed date: 02/12/2014 ; http://www.micron.com/.

[26] Calculating Memory System Power for DDR3 Introduction. Accessed date: 03/28/2014 ; http://www.micron.com/.

[27] Nathan L. Binkert et al. The M5 Simulator: Modeling Networked Systems. *IEEE Micro*, 26(4):52–60, 2006.

[28] The pChase Memory Benchmark Page. Accessed date: 04/12/2014 ; http://pchase.org/.

[29] Sai-Wang Tam et al. RF-Interconnect for Future Network-on-Chip. pages 255–280, 2011.

[30] K. e. a. Therdsteerasukdi. The dimm tree architecture: A high bandwidth and scalable memory system. In *ICCD*, pages 388–395. IEEE, 2011.

VIII. Vitae

Mario Donato Marino is currently an independent researcher in Italy. He has received a best paper award on an international top conference and has co-authored 38 international articles in journals, conferences, and workshops which include computer architecture, microprocessor evaluation, systems, high-performance computing, distributed computing, parallel computing, and performance evaluation. He has worked in several institutions such as University of Sao Paulo as an assistant professor and University of Texas at Austin and of Virginia as researcher. He has served as organization chairman in one international conference and has been serving a number of committees in conferences, workshops, and one journal editorial board. He has received an award on teaching and is a member of the Institute of Electrical and Electronics Engineers (IEEE) and a member Association of Computer Machinery (ACM).