DimNoC: A Dim Silicon Approach towards Power-Efficient On-Chip Network

Jia Zhan[†], Jin Ouyang[§], Fen Ge[*], Jishen Zhao[‡], Yuan Xie[†]

[†]University of California Santa Barbara, [§]NVIDIA Corporation,
[*]Nanjing University of Aeronautics and Astronautics, [‡]University of California Santa Cruz

[†]{jzhan, yuanxie}@ece.ucsb.edu, [§]jouyang@nvidia.com,
[*]gefen@nuaa.edu.cn, [‡]jishen.zhao@ucsc.edu

ABSTRACT

The diminishing momentum of Dennard scaling leads to the ever increasing power density of integrated circuits, and a decreasing portion of transistors on a chip that can be switched on simultaneously—a problem recently discovered and known as *dark silicon*. There has been innovative work to address the "dark silicon" problem in the fields of power-efficient core and cache system. However, dark silicon challenges with Network-on-Chip (NoC) are largely unexplored. To address this issue, we propose *DimNoC*, a "dim silicon" approach, which leverages *drowsy SRAM* and *STT-RAM* technologies to replace pure SRAM-based NoC buffers. Specifically, we propose two novel hybrid buffer architectures: 1) a Hierarchical Buffer (HB) architecture, which divides the input buffers into a hierarchy of levels with different memory technologies operating at various power states; 2) a Banked Buffer (BB) architecture, which organizes drowsy SRAM and STT-RAM into separate banks in order to hide the long write-latency of STT-RAM. Our experiments show that the proposed *DimNoC* can achieve 30.9% network energy saving, 20.3% energy-delay product (EDP) reduction, and 7.6% router area decrease compared with pure SRAM-based NoC design.

Categories and Subject Descriptors: C.2 [Computer-Communication Networks]: Network Architecture and Design

General Terms: Performance, Design

Keywords: Network-on-Chip, Dark Silicon, STT-RAM

1 Introduction

Dennard scaling, which has offered us near-constant chip power with doubling number of transistors, has come to an end [8]. Computer designers are seeking ways to stay on the performance curve without exceeding the thermal design power (TDP) by using emerging many-core processors. Dynamic voltage/frequency scaling (DVFS) of cores can mitigate the issue. However, many-core processors integrate increasingly more transistors than those can remain powered-on simultaneously. Recent studies refer to the fraction of chip which is entirely powered-off as "dark silicon" [28].

To address the challenges of many-core scaling in the dark silicon era, most prior studies focus on core or memory sub-system optimization, while dark-silicon-aware on-chip interconnect design has not drawn much attention. Network-on-Chip (NoC) has a significant impact on the overall performance of many-core processors and consumes a large portion of total power budget. Recent research and industrial prototypes have validated that about 10%–36% [3, 10, 26] of total chip power is consumed by NoC. Moreover, in the dark silicon era, the majority of on-chip core/cache components may be put in the dormant mode due to the TDP limit. Without further optimization, the NoC components must remain active, because a gated-off router will block packet-forwarding and the access to shared caches/directories. As a result, the ratio of NoC power rises substantially among the remaining on-chip active resources, e.g. 42% in a 32 core system [30].

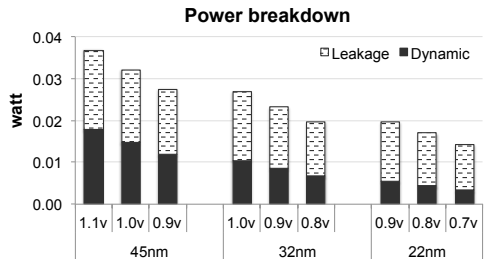
Driven by these observations, there have been some research efforts to aggressively shut down parts of the NoC, leveraging conventional power-gating [4, 6, 17, 23, 24] techniques to reduce idle power. However, NoC power-gating is heavily dependent on the traffic pattern. In order to benefit from power-gating, an adequate idle period (break-even time) of a router is required to secure any appreciable power saving. However, the major performance penalties for waking up the power-gated blocks (in the range of 10∼20 cycles [4, 6, 24] depending on the frequency) are undesirable. To alleviate this drawback and achieve better power saving, we explore the replacement of conventional NoC buffers by **drowsy SRAM**, which uses similar circuitry in drowsy cache [9] to introduce an intermediate sleep mode between power-on and power-gating state. In addition to power management techniques, emerging non-volatile memory (NVM) technologies enable new opportunities to enhance the power-efficiency of NoC. Since buffer is the dominant leakage consumer in NoC, it may be constituted by NVM whose leakage power is considerably lower than conventional SRAM. Among alternative NVM technologies (PCM, STT-RAM, ReRAM), **STT-RAM** is particularly promising to replace NoC buffers because of its high cell density, non-volatility feature, low leakage power consumption, and high endurance [11].

In summary, we propose DimNoC which uses much finer-granularity power saving techniques rather than power-gating in prior dark-silicon researches. To that end, our work can be viewed as a dim-silicon approach dimming the power consumption with novel buffer-management schemes. It introduces hybrid buffer architectures which combine the benefits of both *drowsy* SRAM and *non-volatile* STT-RAM. The rational behind the hybrid organization is that, we can use drowsy SRAM to reduce the wake-up penalty of every sleeping block, and leverage the low-leakage property of STT-RAM to further eliminate unnecessary power-gating operations, and thus, reduce the number of costly wake-up operations. Moreover, the combination of these two technologies mitigates the disadvantages of each other: the long-latency write operations of STT-RAM can be reduced by steering frequent write accesses to drowsy SRAM, and the hardware overhead of drowsy circuit can be compensated by the area saving from the high-density STT-RAM cells. Specifically, two novel hybrid buffer architectures are proposed in this paper: 1) *Hierarchical Buffer (HB)*, in which each input buffer of the NoC router is organized as multiple levels of storage, where each level is composed of a different memory technology. Depending on the traffic load, different levels are activated successively. STT-RAM is designed to be lastly activated to

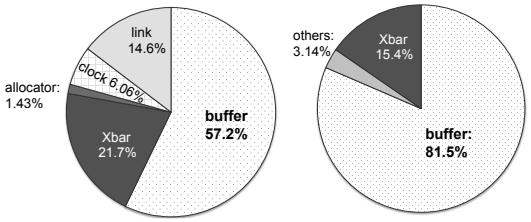
This work is supported in part by NSF grant 1500848, 1461698, and 1213052. This work is part of the ASKS project (`http://www.ece.ucsb.edu/~yuanxie/projects/ASKS/`).

DAC '15, June 07 - 11, 2015, San Francisco, CA, USA

http://dx.doi.org/10.1145/2744769.2744824.



(a) Router power breakdown (dynamic power vs leakage power) when varying the operating voltage and frequency.



(b) Dynamic power breakdown (c) Leakage power breakdown

Figure 1: Dynamic and leakage power breakdown of a virtual-channel based router ((b) and (c) are for 32nm and 0.9v case in (a)).

avoid frequent slow write operations. 2) *Banked Buffer (BB)*, in which drowsy SRAM buffers and STT-RAM buffers are logically interleaved within each VC, but physically separated into multiple banks to *hide the long write latency of STT-RAM*.

2 Challenges and Opportunities

Router Power Analysis. To better understand the power distribution of a router, we simulate a classic wormhole router with DSENT [25] and plot the power breakdown of both dynamic and leakage power for different process technologies: 45nm, 32nm, and 22nm. We use a frequency of 1GHz, and the supply voltages for different technologies are, by default, 1.0V, 0.9V, and 0.8V, respectively. For sensitivity studies, we vary the supply voltage around the default values. The flit width is set to be 128 bits. Each input port of a router comprises 2 virtual channels (VC) and each VC is 4-flit deep. The power numbers are estimated with an average injection rate of 0.3 flits/cycle.

As shown in Figure 1a, for a certain process technology, the percentage of leakage power increases as the supply voltage reduces; As technology scales from 45nm to 22nm, the ratio of leakage power increases and substantially outweighs that of dynamic power. For example, leakage power is around 63.4% of total power with a supply voltage of 0.9V at 32nm technology. These results also reveal the power crisis in the dark silicon era due to the increasing leakage power. The breakdown of dynamic and leakage power consumption for different router components are shown in Figure 1b and Figure 1c (32nm, 0.9V), respectively. In summary, Figure 1 reveals that: (1) *Leakage power dominates the power budget in the dark silicon era.* (2) *Buffers become the primary power consumer for NoC*.

The Limitation of Conventional Power Gating. In the dark silicon background, as a significant percent of transistors cannot be switched on due to TDP limit, researchers started to explore power-gating on NoC: completely shut down idle routers or links and then wake them up when new packets arrive. Power-gating is implemented by inserting appropriately sized transistor(s) with high threshold voltage between V_{dd} and the logic block. Therefore, the entire router block can be switched between on and off states by asserting and de-asserting the sleep signal. However, applying power-gating on NoC has been elusive, because frequently switching routers/links off and on will incur not only significant energy overhead, but also considerable wake-up delay. The inter-arrival period of packets has to be long enough in order to compensate for the overheads.

Therefore, the benefit of applying power-gating on NoC is largely traffic-dependent. As a motivational example, we run the `disparity` workload on a NoC-based sixteen-core system and conduct statistical analysis on the intervals of consecutive flit arrivals. As the histogram

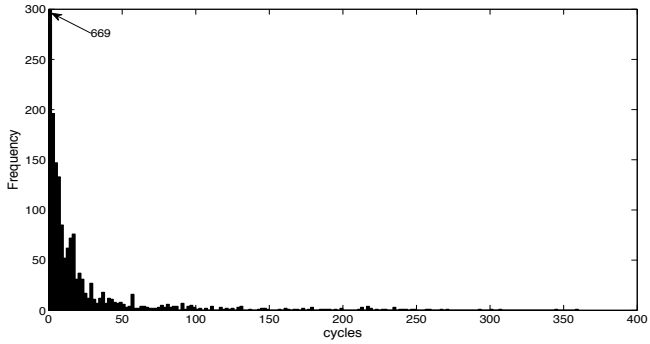


Figure 2: Histogram of inter-arrival time for a router

in Figure 2 shows, some long inter-arrival periods exist in the network which offers the opportunity of power-gating. Nevertheless, most intervals are less than 10 cycles (the break-even time of routers is approximately 10 cycles [4, 6, 24]) and among them a significant portion are less than 2 cycles. Consequently, in this example, conventional power-gating techniques will be detrimental to the network performance because of frequent on-off state transitions of network routers, and even severely offset the power saving gained from long idle intervals.

Fine-Grained Power Management. Conventional power-gating techniques shut down a router only when it is completely idle, which limit their application in skewed or heavily-loaded network traffic. Correspondingly, the whole router needs to be woken up for a new arrival, which incurs substantial wake-up overhead. However, depending on the traffic, buffer utilization may vary over time (temporal difference) and different VCs have different occupancies (spatial difference). To test this hypothesis, using the same `disparity` example, we profile the buffer occupancies for two VCs in a physical port, the result of which is presented in Figure 3. Within each VC, we found that their utilization differ most of the time. Furthermore, when comparing VC 0 with VC 1, their buffer utilizations differ at most time. There are periods when VC 0 is busy while VC 1 is idle, and vice versa.

The variation of buffer utilization over time and space indicates indicates that only a subset of buffers are required at most time. A finer-granularity power management of buffers can potentially save significant power. In contrast, prior proposals of bufferless NoC [19] or elastic-buffered NoC [18] aggressively carve away nearly all buffers from NoC to save leakage power (and area). Consequently, these designs suffer from performance penalties under high load and do not support VCs for multiple traffic classes.

3 Our Method: DimNoC

Instead of completely shutting down the on-chip components in response to power shortage, another option is to operate them under-clocked, namely "dim silicon". Conventional approaches towards this direction are employing dynamic voltage and frequency scaling (DVFS) on individual NoC routers/links [20, 29]. While these approaches have shown some promising power saving, their feasibility is unclear as the per-node voltage regulators incur non-negligible area overhead and the re-timing latency leads to significant performance degradation.

In order to achieve a practical dim-silicon NoC (DimNoC) design, our study focuses on optimizing buffer architecture—the primary NoC power consumer—rather than adopting conventional DVFS in individual routers. We leverage both drowsy SRAM and STT-RAM techniques to design power-efficient NoC buffers. In particular, the drowsy SRAM buffers introduce an intermediate power state between power-on and power-gating to achieve better performance-power trade-off; the STT-RAM buffers dissipate low leakage power, endure frequent packet accesses, and consist of high-density cells to save area.

Furthermore, we integrate drowsy SRAM with STT-RAM to simultaneously leverage the advantages of both memory technologies to 1) reduce wake-up penalty of gated buffers using drowsy circuit, 2) avoid unnecessary power-gating and wake-up by utilizing low-leakage STT-RAM, and 3) reduce the area budget with the dense STT-RAM

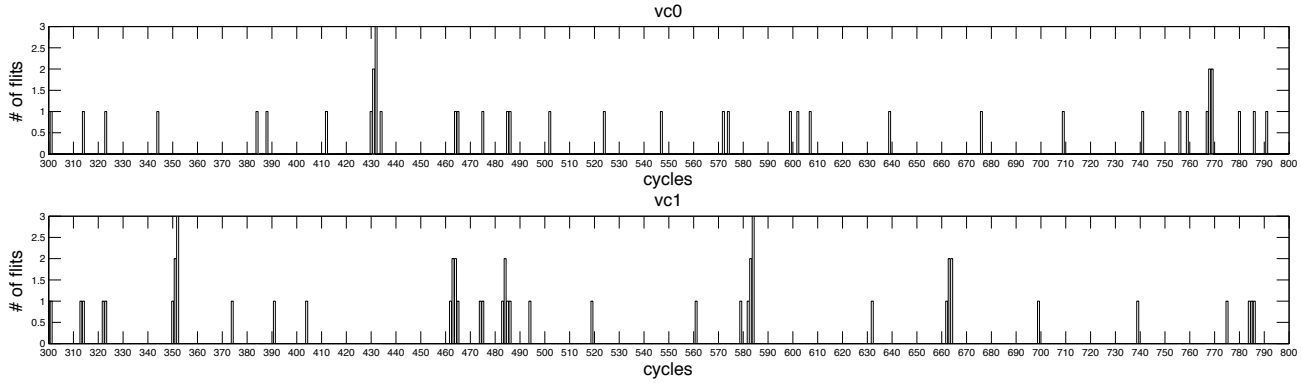


Figure 3: Buffer utilization for different VCs

cells. As a result, our design can substantially reduce network power and area with negligible performance overhead.

3.1 Drowsy SRAM Buffers

Our first technique dims the network by introducing an intermediate "drowsy" state between "on" and "off", in which a router operates at a low-power state which retains the data but does not allow for write accesses. Since the voltage in the drowsy state is lower than the "on" state but higher than the "off" state, entry to and exit from the "drowsy" state are both more efficient than the power-gated state.

Figure 4 shows our drowsy buffer, which is motivated by *drowsy cache* [9]. The drowsy circuit includes a drowsy bit, a voltage controller, and a word-line gating circuit. Depending on the state of the drowsy bit, the voltage controller switches the operating voltage between high (active) or low (drowsy) states. Additionally, the word-line gating circuit prevents access to the drowsy cells and avoids data corruption. .

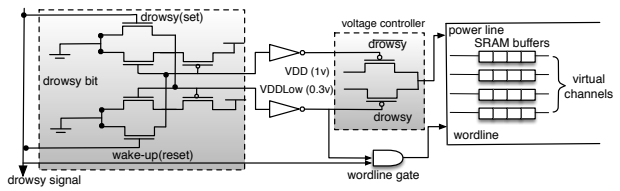


Figure 4: Circuit design for drowsy buffer. In the drowsy bit, the word line, bit lines, and two pass transistors are not shown for simplicity.

Drowsy SRAM has shorter transition delay than power-gated SRAM, despite consuming higher leakage power. As shown in Figure 2, there are a lot of short intervals that are less than the break-even time ($\sim$10 cycles) which cannot be leveraged by power-gating, but can be used to enter the drowsy state whose wake-up latency is only 1 to 2 cycles [9].

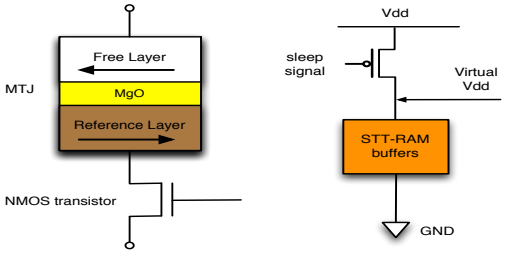
3.2 STT-RAM Buffers

In addition to modifying SRAM buffer circuitry, we also propose to adopt emerging non-volatile memory (NVM) technologies to replace SRAM buffers. STT-RAM has the following advantages which make it a good candidate to "dim" NoC buffers in the dark silicon era.

3.2.1 Characteristics

STT-RAM relies on non-volatile, resistive information storage in a cell, and thus exhibits *near-zero leakage in the data array*. Figure 5a shows the structure of a 1T1J STT-RAM cell, which comprises of an access transistor and a Magnetic Tunnel Junction (MTJ) for binary storage. An MTJ contains two ferromagnetic layers (the reference layer and the free layer) and one tunnel barrier layer (MgO). The directions of these two layers determine the high/low resistance of the MTJ, which indicate the "1"/"0" state. The 1T1J cell size (about 6 - 50 F^2) is smaller compared to typical 6T SRAM cells (about 120 - 200 F^2). As a result, for the same capacity as SRAM, *high-density* STT-RAM cuts the buffer area budget.

The *endurance* of STT-RAM (10^{15} writes [11]) is significantly higher than other NVM (e.g. 10^9 writes for PCM). This makes



(a) 1T1J STT-RAM cell (b) Power-gating STT-RAM

Figure 5: (a) shows the 1T1J cell structure of the STT-RAM data array. (b) applies power-gating on STT-RAM buffers to further cut the leakage power of peripheral circuit.

STT-RAM superior to other NVM in enduring frequent packet accesses in the on-chip network.

3.2.2 Power-Gating STT-RAM

As the data array of STT-RAM has negligible leakage, the peripheral circuit becomes the dominant leakage consumer in STT-RAM buffers. Simulation results from NVsim [7] show that nearly half of STT-RAM die area is occupied by peripheral circuitry, i.e., half of the chip is leaky. Therefore, we propose to shut off the peripheral leakage power through power-gating. Figure 5b shows the gating circuit, which uses a sleep transistor inserted between the V_{dd} and the STT-RAM buffers to control the on/off states.

3.2.3 Relaxing Non-volatility

STT-RAM achieves comparable read latency and energy as SRAM. However, the write access time and the write energy of STT-RAM are relatively higher than an SRAM write operation. With the advance of technology, recent designs have demonstrated shorter write latency of 2 - 4 *ns* [5, 16, 21], which corresponds to 2 - 4 cycles for 1 GHz clock frequency.

Even with more conservative STT-RAM design, the write overhead can be mitigated by relaxing its non-volatility [13]. STT-RAM generally has more than 10 years of retention time. This long retention time is unnecessary for on-chip buffers because they only serve as temporary storage for packets. Even under high network load, the worst-case queuing time of packets is at the magnitude of microseconds (μs). Therefore, by relaxing the non-volatility of STT-RAM from years to a few *ms* or even μs, faster write speed and smaller write energy can be obtained. It has been demonstrated by Jog et al. [13] that a 10*ms* retention time of MTJ can be achieved at a switching time of 2*ns* with 61μA write current.

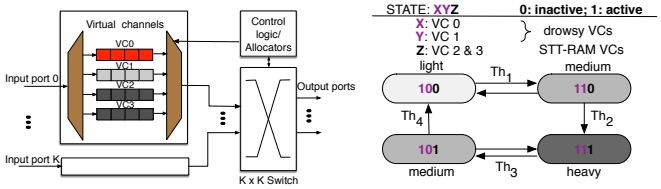
Although the significant reduction of STT-RAM write latency enables integration of STT-RAM as NoC buffers, we do not aggressively assume the write latency gap between STT-RAM and SRAM can be completely eliminated. Instead, as illustrated later, we propose hybrid buffer designs to further reduce STT-RAM write overhead and even seamlessly hide the long write latency through buffer *banking*. A recent work [12] leverages STT-RAM in NoC buffers, but only for performance purposes by designing larger buffers with dense STT-RAM cells. Their design suffers from power overhead in moderate or high network load.

3.3 Hybrid Buffer Architectures

In this subsection, we propose two novel hybrid buffer architectures to take the advantages of both drowsy SRAM and non-volatile STT-RAM for power-efficient DimNoC design in the dark silicon era.

3.3.1 Hierarchical Buffer

To allow fine-grained activation/power-gating of different VCs, we divide the VCs into a hierarchy of levels and activate different levels successively based on the network load. We employ drowsy SRAM in lower-level VCs, which switch between power-on and drowsy state. Being aware of the frequent short inter-arrival periods of packets, we do not shut down the drowsy buffers in order to reserve resources for instant wake-up. Correspondingly, higher-level VCs are made of STT-RAM and will be activated lastly to accommodate heavy traffic, thus avoiding frequent costly write accesses. Moreover, power-gating techniques will be applied on STT-RAM VCs to take advantage of the long idle periods under very light traffic.



(a) The Hierarchical Buffer (HB) architecture (b) State transition

Figure 6: The Hierarchical Buffer (HB) architecture with four VCs as an example. VC 0 & 1 are drowsy VCs, and the rest are STT-RAM VCs. In (a), VC 0 is active, VC 1 is in drowsy state, whereas VC 2 and 3 are power-gated. Diagram (b) depicts the state transitions for different VCs.

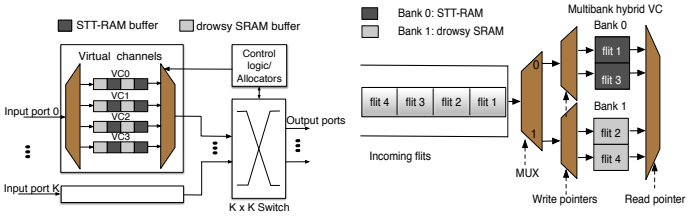
Figure 6a depicts a Hierarchical Buffer (HB) design, where VC 0 and 1 are drowsy VCs, while VC 2 and 3 are STT-RAM VCs. Note that we use four VCs here for illustration purposes, less or more number of VCs are also applicable. We separate drowsy VCs and STT-RAM VCs into multiple levels, and allow fine-grained activation of different levels in a hierarchical manner. For example, in Figure 6a, VC 0 (level 1) is active, VC 1 (level 2) is in drowsy state, whereas the rest of STT-RAM VCs (level 3) are power-gated.

Figure 6b shows the state transition diagram of different VCs in response to variation of network traffic. X, Y, and Z are used to represent the status of VC 0, 1, and {2,3}, respectively. "1" means *active* whereas "0" stands for *inactive*. Note that *inactive* means *drowsy* for VC 0 and 1, and *power-gated* for VC 2 and 3. Initially, assuming the traffic load is light, only VC 0 is activated and the rest of VCs are inactive (100). When the incoming traffic exceeds a certain threshold (Th_1) of buffer occupancy, the remaining drowsy VC 1 will be promptly activated (110). In case of further increase of network load or abrupt arrival of burst packets that exceeds another threshold (Th_2), the STT-RAM VCs will be switched on (111). Reversely, when the network traffic decreases (Th_3), we do not power-gate STT-RAM VCs immediately to avoid unnecessary wake-up due to network fluctuation, but instead put VC 1 into drowsy state (101). Finally, if the traffic further decreases (Th_4), the STT-RAM VCs will be shut down (100). Also, 110⇒100 and 101⇒111 happen under network fluctuation.

3.3.2 Banked Buffer

Apart from the VC-based partitioning which separates the drowsy SRAM VCs from STT-RAM VCs, we propose a more fine-grained hybrid buffer architecture, in which drowsy SRAM and STT-RAM buffers are logically interleaved within each VC, but physically organized as separate banks. This design allows for simultaneous access to multiple banks so as to hide the longer write latency of STT-RAM.

Figure 7a shows the logical view of the Banked Buffer (BB), where STT-RAM buffers and drowsy SRAM buffers are organized in an interleaved fashion within each VC. Figure 7b shows the physical architecture inside a single VC. Specifically, each VC is separated into two banks. One is the STT-RAM bank, and the other is the drowsy



(a) Logical view of the Banked Buffer (BB) (b) Physical design of a banked VC

Figure 7: A Banked Buffer (BB) architecture where each VC consists of both STT-RAM and drowsy SRAM in an interleaved fashion. Within each VC, the STT-RAM buffers and drowsy SRAM buffers are separated into two banks.

SRAM bank. Then, the incoming flits will go through a bank-selection multiplexer and write into the appropriate bank. For example, assuming a two-cycle write latency for STT-RAM, during cycle 0, flit 1 will be written into the STT-RAM bank. Subsequently at cycle 1, flit 2 will be directed into the drowsy SRAM bank and complete the write operation. Meanwhile, flit 1 will also complete its two-cycle write operation in the STT-RAM bank at this cycle, and thus both flits are readable at the next clock cycle. Similarly, the following flit 3 and flit 4 will be transferred into the STT-RAM bank and drowsy SRAM bank, respectively. In this way, the long write latency of STT-RAM can be hidden. In general, for a n-cycle STT-RAM write latency, we can divide the STT-RAM buffers into n-1 banks to hide the long write latency.

4 Experiments

We use a Pin [22] based functional simulator to collect instruction traces from applications. The traces are then fed into a cycle-level trace-driven multicore simulator integrated with Garnet [1] and DSENT [25] to evaluate the network performance and power with 32nm technology. Detailed system configurations are listed in Table 1.

Table 1: System and Interconnect configuration

core count	16	topology	4 × 4 2D Mesh
L1 I & D cache	private, 32KB	router pipeline	four-stage
L2 cache	shared, 512KB/bank	VC count	4 VCs per port
cacheline size	64B	buffer depth	4 buffers per VC
frequency	1GHz	packet length	5 flits, 16B/flit

We use NVsim [7], combined with statistics collected from recent STT-RAM prototypes [15] to estimate the latency and energy consumption of each SRAM and STT-RAM access. Furthermore, we obtain the scaled latency and energy of accesses to drowsy SRAM [9] and STT-RAM with relaxed non-volatility [13]. Table 2 shows the parameters of various designs.

We evaluate our hybrid buffer designs with the San Diego Vision Benchmark Suite [27]. Table 3 summarizes and compares various buffer design techniques which are evaluated in our experiments. We compare our design with a prior work that employs look-ahead routing based scheme [17], and a recent NoC power-gating design called Node-Router Decoupling (NoRD) [4].

4.1 Energy and Performance Evaluation

4.1.1 Energy Savings

Because our primary goal is to conserve network energy, we run different benchmarks with all buffer design techniques listed in Table 3. Figure 8 shows the results of total network energy. Note that we assume the wake-up latency of conventional power-gating, look-ahead routing based power-gating, and drowsy buffers are ten cycles [4], five cycles [17], and two cycles [9], respectively. In addition, the write latency of STT-RAM is two cycles. We also perform sensitivity study of the impact of longer STT-RAM write-latencies and present the results in the later part of this section. We make four important observations from the evaluation results:

• Compared with the baseline *All_SRAM* design, *All_SRAM_PG*, which employs conventional power-gating techniques to turn off an

Table 2: 1KB SRAM and STT-RAM buffer configurations

	cell size (F^2)	write latency (cycle)	read latency (cycle)	write energy/bit (pJ)	read energy/bit (pJ)	leakage power (mW)
SRAM	146	1	1	0.049	0.063	1.797
DrowsySRAM	146	1	1	0.049	0.063	0.165
STT-RAM (10yr)	50.67	10	1	0.534	0.153	0.042
STT-RAM (10ms)	45.60	2	1	0.286	0.082	0.044

Table 3: Comparisons of different buffer design techniques

All_SRAM	Baseline, which employs pure conventional SRAM based buffers
All_SRAM_PG	Buffers are designed with pure SRAM, and conventional power-gating technique is applied
All_SRAM_PG_LA	Buffers are designed with pure SRAM, and look-ahead routing based power-gating technique is applied [17]
NoRD	Buffers are designed with pure SRAM, and bypass paths are introduced to deliver packets without waking up gated routers [4]
All_DrowsySRAM	Buffers are designed with pure *drowsy* SRAM
All_STTRAM	Buffers are designed with pure STT-RAM
HB	Hierarchical Buffer: Lower-level VCs are pure SRAM based, and higher-level VCs are pure STT-RAM based
BB	Banked Buffer: Drowsy SRAM and STT-RAMs are constructed into separate banks, and are accessed in an interleaved fashion in each VC

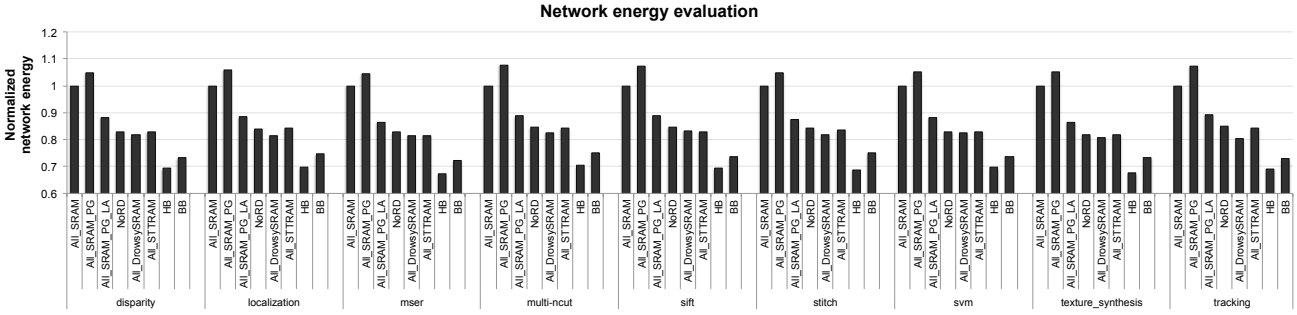


Figure 8: Energy comparisons for different buffer designs and power management schemes

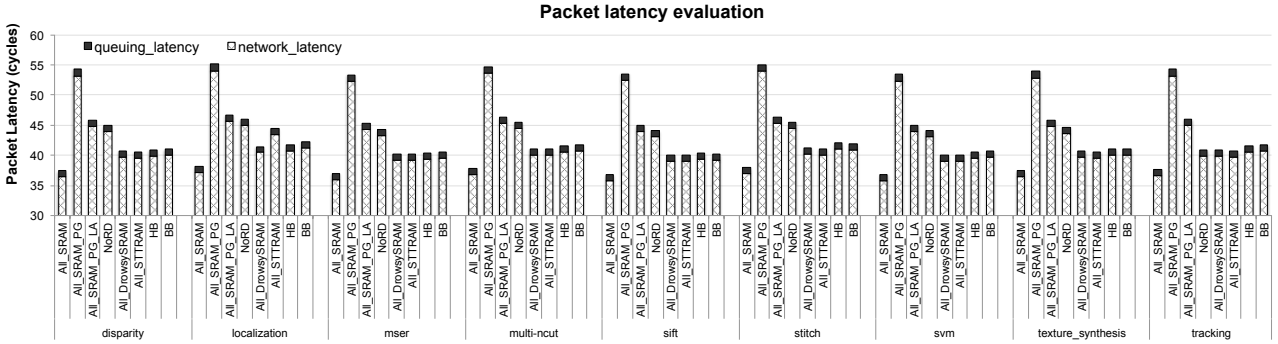


Figure 9: Packet latency comparisons for different buffer designs and power management schemes

entire router whenever it is idle, indeed incurs 5.82% energy overhead due to frequent router wake-up, averaged over all the benchmarks.

• *All_SRAM_PG_LA* [17] employs look-ahead routing to hide some wake-up latency and achieves 11.9% energy saving over the baseline. *NoRD* [4] provides bypass to transmit packets through gated routers and thus increases the power-gating duration. It reduces the network energy by 16.4%.

• *All_DrowsySRAM* put routers into low-power drowsy state instead of completely shutting them down to further accelerate the wake-up process. On average, it cuts down the energy by 18.2%.

• *All_STTRAM* replaces SRAM by STT-RAM, and still achieves energy savings compared with the baseline, indicating the benefit of low leakage outweighs the overhead of high write energy of STT-RAM. However, the energy saving (17.1%) is not very significant.

• To avoid unnecessary write operations to STT-RAM, *HB* uses hybrid buffers and only activates the STT-RAM VCs at high load. As a result, it achieves 30.9% energy savings on average. Alternatively, *BB* accesses the drowsy SRAM bank and STT-RAM bank in an interleaved fashion within each VC, which successfully hides the long write latency of STT-RAM and achieves 26.3% energy saving on average.

4.1.2 Performance Overhead

The proposed buffer designs achieve energy savings by deactivating some of the network resources, which will unavoidably sacrifice the network performance to some degree. Figure 9 shows the average packet latencies when running different benchmarks, including *queuing latency* in the network interface and *network latency* from source nodes to destination nodes.

As expected, *All_SRAM_PG* leads to significant network delay, adding 44.9% on average across all the benchmarks. By mitigating the wake-up delay, *All_SRAM_PG_LA* and *All_DrowsySRAM* amortize the latency overhead to 22.3% and 8.8% on average, respectively. *NoRD* reduces the number of wake-up operations through bypass but still suffers from 18.6% performance overhead due to packet detours.

For *All_STTRAM*, the average network latency overhead is only 9.4% without applying power-gating. *HB* and *BB* slightly increase the overhead to 10.4% and 10.8%, respectively. This is because they further apply fine-grained power-gating on individual VCs.

4.2 Sensitivity Study on STT-RAM writes

The aforementioned STT-RAM buffer designs assume a two-cycle write latency by sacrificing the retention time of STT-RAM cells. Here, we conduct sensitivity studies with different write latencies of STT-RAM. Intuitively, when increasing the retention time of STT-RAM, the write latency and the write energy of STT-RAM cells increase. As a result, the performance overhead of *All_STTRAM* and *HB* will increase, and the corresponding energy savings will decrease. For *BB*, it hides the long write latency of STT-RAM through banking.

However, the increase of hardware overhead and write energy overhead of STT-RAM will still decrease its energy saving.

Therefore, we use energy-delay product (EDP) as a metric to evaluate different buffer designs, where E and D refer to the network energy and latency, respectively. Figure 10 shows the EDP results when varying the write latency of STT-RAM, using `disparity` as an example. As illustrated in Section 3.2.3, the write latency of STT-RAM is between 2 to 4 cycles through relaxing of non-volatility.

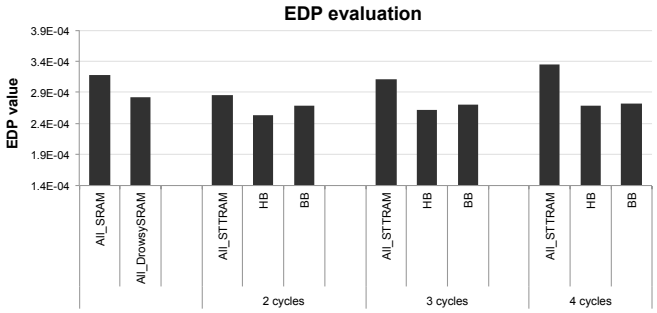


Figure 10: Comparisons of EDP with different designs when varying the write latency of STT-RAM.

We can see that, *All_STTRAM*, *HB*, and *BB* achieve lower EDP values than the *All_SRAM* baseline when the write latency of STT-RAM is two cycles. However, as the write latency increases, the EDP values for *All_STTRAM* increases dramatically and exceeds that of the baseline. *HB* achieves lower EDP than *BB*, but the gap between these two designs is shrinking with the increase of STT-RAM write latency. Specifically, compared to the *All_SRAM* baseline, the *HB* design can achieve 20.3%, 17.8%, and 15.2% EDP reduction for STT-RAM write latency of 2, 3, and 4 cycles, respectively.

4.3 Hardware Implementation

For *HB* as shown in Figure 6a, the drowsy circuit is introduced to low-level VCs, and the higher-level VCs are replaced by STT-RAM. Sleep transistors are also needed for power-gating purposes. Moreover, for *BB* as shown in Figure 7, additional multiplexers are required to access different buffer banks. To evaluate the area and power overhead of our proposed hybrid buffer designs, we synthesize a parametrized RTL router implementation [2] using Synopsis Design Compiler, and substitute the power/area numbers of SRAM and STT-RAM buffers through NVsim [7] simulation. For fair comparison, we keep the buffer capacity consistent throughout all buffer designs.

Figure 11 shows the area breakdown of different router architectures. *All_STTRAM* significantly reduces the area because of the high cell density of STT-RAM. It achieves 17.8% area saving compared with the baseline *All_SRAM*. The Hierarchical Buffer (*HB*) design also achieves 7.6% area saving. More complex logic units are required for the Banked Buffer (*BB*), but BB only incurs a diminutive 4.1% area overhead compared to the pure SRAM design. Note that the power consumption of these extra components is incorporated in our energy evaluation (Figure 8).

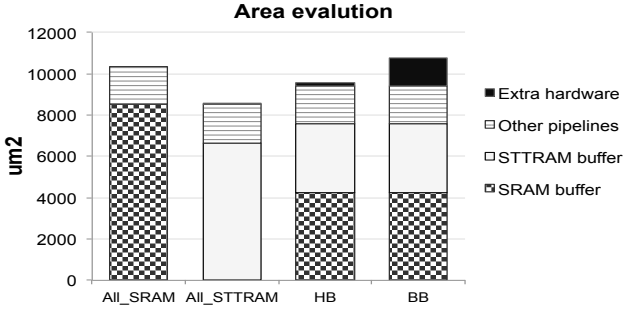


Figure 11: Router area breakdown for different buffer designs

5 Conclusion

In this work, we propose *DimNoC* to tackle the dark silicon problem from the NoC's perspective, which integrates drowsy SRAM and STT-RAM to design buffers in a router. Two hybrid buffer architectures, Hierarchical Buffer (HB) and Banked Buffer (BB), are proposed with efficient power-management strategies. Experimental results on the San Diego Vision Benchmark Suite show that *DimNoC* can achieve 30.9% energy savings on average, 20.3% network energy-delay product (EDP) reduction, and 7.6% router area reduction.

6 References

[1] N. Agarwal, T. Krishna, L.-S. Peh, and N. K. Jha. GARNET: A detailed on-chip network model inside a full-system simulator. In *ISPASS*, pages 33–42, 2009.

[2] D. U. Becker. *Efficient microarchitecture for network-on-chip routers*. PhD thesis, Stanford University, 2012.

[3] S. Bell et al. Tile64-processor: A 64-core soc with mesh interconnect. In *ISSCC*, pages 88–598, 2008.

[4] L. Chen and T. M. Pinkston. Nord: Node-router decoupling for effective power-gating of on-chip routers. In *MICRO-45*, pages 270–281, 2012.

[5] K. C. Chun et al. A scaling roadmap and performance evaluation of in-plane and perpendicular MTJ based STT-MRAMs for high-density cache memory. *IEEE Journal of Solid-State Circuits*, 48(2):598–610, 2013.

[6] R. Das, S. Narayanasamy, S. Satpathy, and R. Dreslinski. Catnap: Energy proportional multiple network-on-chip. In *ISCA*, pages 320–331, 2013.

[7] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi. Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory. *TCAD*, 31(7):994–1007, 2012.

[8] H. Esmaeilzadeh et al. Dark silicon and the end of multicore scaling. In *ISCA*, pages 365–376, 2011.

[9] K. Flautner, N. S. Kim, S. Martin, D. Blaauw, and T. Mudge. Drowsy caches: simple techniques for reducing leakage power. In *ISCA*, pages 148–157, 2002.

[10] Y. Hoskote, S. Vangal, A. Singh, N. Borkar, and S. Borkar. A 5-GHz mesh interconnect for a teraflops processor. *Micro, IEEE*, 27(5):51–61, 2007.

[11] ITRS. Process Integration, Devices, and Structures (PIDS). `http://www.itrs.net/Links/2013ITRS/2013Tables/PIDS_2013Tables.xlsx`, 2013.

[12] H. Jang et al. A hybrid buffer design with STT-MRAM for on-chip interconnects. In *NoCS*, pages 193–200. IEEE, 2012.

[13] A. Jog et al. Cache revive: architecting volatile STT-RAM caches for enhanced performance in CMPs. In *DAC*, pages 243–252. ACM, 2012.

[14] J. Kim, W. J. Dally, B. Towles, and A. K. Gupta. Microarchitecture of a high-radix router. In *ISCA*, pages 420–431, 2005.

[15] J. P. Kim et al. A 45nm 1Mb embedded STT-MRAM with design techniques to minimize read-disturbance. In *VLSIC*, pages 296–297. IEEE, 2011.

[16] E. Kitagawa et al. Impact of ultra low power and fast write operation of advanced perpendicular MTJ on power reduction for high-performance mobile CPU. In *IEDM*, pages 29–4. IEEE, 2012.

[17] H. Matsutani, M. Koibuchi, D. Wang, and H. Amano. Run-time power gating of on-chip routers using look-ahead routing. In *ASPDAC*, pages 55–60, 2008.

[18] G. Michelogiannakis and W. J. Dally. Elastic buffer flow control for on-chip networks. *Computers, IEEE Transactions on*, 62(2):295–309, 2013.

[19] T. Moscibroda and O. Mutlu. A case for bufferless routing in on-chip networks. In *ACM SIGARCH Computer Architecture News*, volume 37, pages 196–207, 2009.

[20] U. Y. Ogras, R. Marculescu, P. Choudhary, and D. Marculescu. Voltage-frequency island partitioning for GALS-based networks-on-chip. In *DAC*, pages 110–115, 2007.

[21] T. Ohsawa et al. A 1.5 nsec/2.1 nsec random read/write cycle 1Mb STT-RAM using 6T2MTJ cell with background write for nonvolatile e-memories. In *VLSIT*, pages C110–C111. IEEE, 2013.

[22] H. Patil, R. Cohn, M. Charney, R. Kapoor, A. Sun, and A. Karunanidhi. Pinpointing representative portions of large Intel® Itanium® programs with dynamic instrumentation. In *MICRO*, pages 81–92, 2004.

[23] M. Powell, S.-H. Yang, B. Falsafi, K. Roy, and T. Vijaykumar. Gated-Vdd: a circuit technique to reduce leakage in deep-submicron cache memories. In *ISLPED*, pages 90–95, 2000.

[24] A. Samih et al. Energy-efficient interconnect via router parking. In *HPCA*, pages 508–519. IEEE, 2013.

[25] C. Sun et al. DSENT-a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling. In *NOCS*, pages 201–210, 2012.

[26] M. B. Taylor et al. The Raw microprocessor: A computational fabric for software circuits and general-purpose programs. *Micro, IEEE*, 22(2):25–35, 2002.

[27] S. K. Venkata, I. Ahn, D. Jeon, A. Gupta, C. Louie, S. Garcia, S. Belongie, and M. B. Taylor. SD-VBS: The San Diego vision benchmark suite. In *IISWC*, pages 55–64, 2009.

[28] G. Venkatesh et al. Conservation cores: reducing the energy of mature computations. In *ACM SIGARCH Computer Architecture News*, volume 38, pages 205–218, 2010.

[29] J. Zhan et al. Optimizing the NoC Slack Through Voltage and Frequency Scaling in Hard Real-Time Embedded Systems. *TCAD*, 33(11):1632–1643, 2014.

[30] J. Zhan, Y. Xie, and G. Sun. NoC-Sprinting: Interconnect for fine-grained sprinting in the dark silicon era. In *DAC*, pages 1–6. ACM, 2014.