Power-gated MRAMs for Memory-Based Computing with improved broadcast capabilities

Jean-Philippe Diguet

Power-gated MRAMs for Memory-Based Computing with improved broadcast capabilities

Jean-Philippe Diguet
CNRS, Lab-STICC, Lorient, France

***Abstract*—Emerging non-volatile memory technologies open new perspectives for original computing architectures. Recent work have demonstrated the potential of flexible architectures that rely on power-gated distributed Magnetoresistive Random-Access Memory (MRAM). The proposed architecture uses a Network-on-Chip (NoC) to interconnect MRAM-based clusters and distribute application-specific commands to MRAM devices by means of packets. Configurable Network Interfaces allow to transform MRAM devices into smart units able to respond to incoming commands. With such an approach, the NoC becomes the energy bottleneck. In this paper we introduce a possible next step, which consists in optimising the NoC architecture by means of Wireless Links that provide Broadcast capabilities and reduce latency of multi-hop communications.**

I. Memory-Based Computing context

The rise of large manycore architectures supports the increase of computing capabilities, but this evolution is facing multiple challenges. The limitation of power consumption and heat dissipation is a well known issue also identified as dark silicon [1]. The interconnect scalability is another important issue and if flexibility and bandwidth are partially solved with wired Network On Chip (NoC), the resulting latency increase remains a serious challenge [2]. As a consequence available computing resources cannot be fully exploited. Specialisation of processor architectures is a solution to improve the energy efficiency demands such as Multimedia, Graphics, Networks as well as Computer Vision and Machine Learning. Today the evolution in memory design and on-chip memory integration allows for architectures dedicated to application domains characterized by extensive data reuse and where data access dominate computations. This is particularly true for power-gated non-volatile memories (NVM) that open new design perspectives in application domains such as Databases, Search Engines and more generally Associative Memories. In these application domains, writings are negligible compared to readings and the computing model is mostly based on broadcast of requests and asynchronous responses.

Increasing the on-chip memory resources reduces the access time and the power consumption if a memory block can be switched to sleep modes with low leakage power and being activated only when it is required. The benefit of such a solution is demonstrated in [3] that introduces a type of memory-based computing (MBC), where a NoC is used to broadcast requests to distributed NVM blocks enhanced with full power-gating (PG) to switch-off unused memories. Fig.1 presents the proposed NoC-MBC (NMBC) architecture, where most nodes are memories with power-gated blocks and few are processors. The approach also relies on the implementation of functions

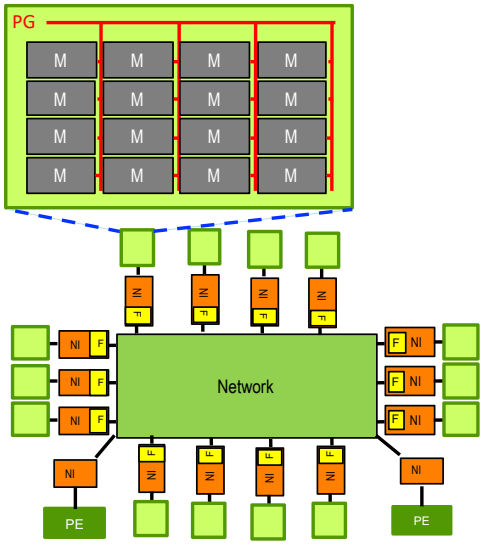

Figure 1: Power-gated NMBC architecture

(F) in Network Interfaces (NI), these functions are application specific and so can so be implemented with configurable logic.

However, the scalability of such a solution comes with a penalty on the NoC latency that increases with the number of routers so the number of hops [2]. New interconnect technologies are emerging to address the long range multi-hop communication bottleneck such as three-dimensional (3D), photonics, and RF/wireless NoCs (WiNOCs). The photonic NoC achieves a higher bandwidth than WiNoC, but a high bandwidth is not necessarily the first constraint. For instance parallel computing requires broadcast operations for synchronization based on short messages [4]. Photon+IC NoC does not immediately provide broadcast capability which is naturally available with RF interconnects. Beyond broadcast capabilities, WiNoCs provide an end-to-end single-hop communication and so can be an efficient solution for improving the performance of parallel applications significantly. In this study we discuss the combination of NMBC and WiNOC.

II. Previous work and results

A. First architecture principle

The proposed architecture model is based on a NoC that interconnects three types of IP blocks: memory clusters, processing elements and managers. In this approach we consider NoC packets as commands initiated by managers as depicted in Fig.2. Memories are distributed in M_C clusters, each cluster

has a unique NI and is composed of M_B memory blocks. The managers are in charge of a set of requests to process and send packets to memory clusters mainly in a broadcast way. They also apply a processor selection for post-processing and can so perform load balancing. All the managers can work in parallel. The decoding of requests is processed by the NI that implement additional logic elements. NIs are, in particular, aware of a data mapping and manage the communication of results to processors, the NIs can also provide managers with monitoring data. The memory blocks are technologically independent from the NoC and the processing units. We used different original implementations of STT-MRAM from Tohoku University with cell and peripheral power-gating capabilities.

B. Database case study

Finally, the effectiveness of the proposed approach is demonstrated through a relevant case study of a database search application implemented with a neuromorphic architecture based on Sparse-Neural-Network (SNN) [5]. A record is a binary sequence of bits and is associated to a clique of neurons, each clique materializes a codeword. The information is coded with Matrices M_{ij} that represent connexions between neurons from two clusters. These matrices are stored in the distributed memory blocks and can so be consulted in parallel. The information retrieval or query consists in the selection of the best candidate neuron for each cluster. The selection of the neuron is based on its connexions with neurons in other clusters corresponding to known fields. The information about existing connection is provided by concerned memories that respond to requests. In practice, all memories are not addressed at the same time leading to inherently partial resource usage in time. Therefore, applying power-gating techniques to unused resources can offer important power savings.

C. Results

The results show that full power gating option gives the best results in terms of power efficiency thanks to small wake up times. Secondly we observe better results with smaller memory granularities (256×256 vs $4 \times 128 \times 128$ bits) and it worth taking advantage of the asymmetry between readings and writings occurrences. The best result are actually obtained with a third memory model with a maximum write bit width of 32 bits, and a read bit width of 32, 64, 128 of 256 bits which are activated selectively according to the reading requirements. Finally, the Energy reduction of the best configuration reaches 87% compared to the SRAM case. Another important point is the distribution of the power consumption. With SRAM memories, the NoC and the memories account for 35% and 57% respectively. But with the best MRAM type, the NoC represents 71% and the memory only drops down to 13% of the total power consumption. The conclusion is that the next optimisation effort must focus on the NoC implementation.

III. WiNoC

A. Architecture principle

The efficiency of the NMBC architecture means splitting an important memory space into a large set of memory blocks connected to a big-scale NoC. This is an issue since latency

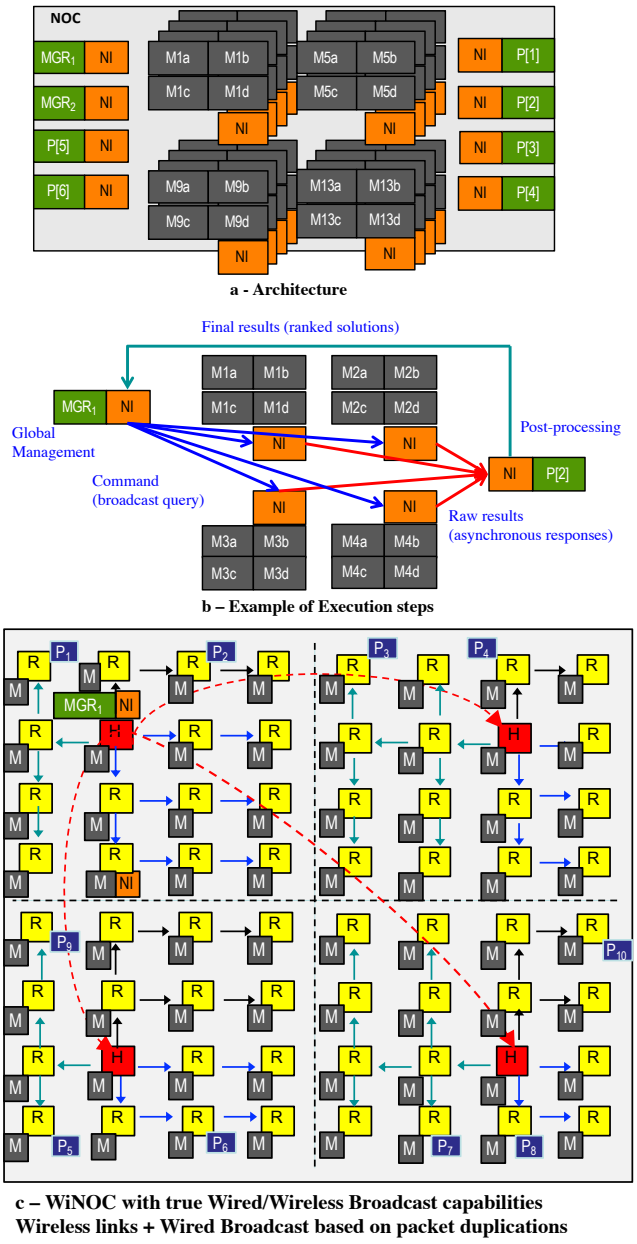

Figure 2: NBMC principle, 64 WiNOC-based implementation

will increase and the energy efficiency decrease with the NoC sizes. So, the imbalance between the PG-MRAM and the NoC will be amplified. The NMBC model is also characterised by a lot of broadcast messages. These two conditions correspond to the conditions that make and hybrid Wireless / Wired NoC a promising solution. In a previous study [4], we have estimated that a WiNoC (60GHz, 65nm) can improve the latency (40%) of parallel applications that introduce about 5% of broadcast communications for barrier synchronization in the case of PARSEC benchmarks using barrier synchronisation. The power consumption of a wireless router is important, however when power-gating techniques are used along with a token passing protocol [6] or a collision detection algorithm [7] then the Energy efficiency can also be significantly improved. In the

NMBC case broadcast plays an important role as shown on Fig. 2-b. Since in the first step the same message is sent to all memories and then only some of them will send a response to the processors in charge of the post-processing. Moreover the more small memories and the the more the NMBC approach is efficient, so large networks of memories are expected. It menas that long path will be numerous so we can expect also to save power by replacing long distance wired multi-hops by wireless single-hops. Fig2-c), shows how we can apply the concept of the WiNoC with four hybrid routers (H) that implement wireless communications. In this example, the NoC is divided in 4 groups, and a wireless is used for broadcast communications between groups, whereas locally a conventionnel mesh NoC is used. Note that we can avoid duplication of packets with a source routing approach using for instance a tree-based routing method like Whirl [8]. The best use of the WiNOC must consider a dynamic choice of paths using Wireless and Wired routers to globally reduce the latency and improve the Energy efficiency. The asynchronous nature of the MBC architecture provide a flexibility that favors this type of optimization.

B. Preliminary case study

We consider the architecture described in Fig. 2-c and simulate random communications based on a pattern, which is representative of the expected traffic The Manager sends a search query where 7 out of 11 fields are unknown. The request is sent to all 63 memory nodes, then 28 (randomly chosen) memories, corresponding the 7x4 involved interconnection memories, send a response, namely one paquet to one out of seven processors randomly chosen to process the 7 missing fields. Finally each of these 7 processors sends a unique paquet to the Manager. This pattern is then repeated for each request.

We simulate this case study with our modified version of Noxim [9] that implements wireless links and wired broadcast based on packet replication using the WHIRL algorithm. In this experiment, the WiNoC is not used for long distance but only for broadcast communications. The power consumption is derived from the NoC activities and power models. We consider the 60GHz transceiver from [2] for wireless links. This model is based on a 65nm technology and implements a OOK modulation, it consumes 32.2 mw and 6.3mW in active and sleep modes respectively. We used ORION 3.0 for the estimation the wired NoC power consumption also with a 65nm technology. The results are given in Table I. We can observe that broadcast mechanism allows to significantly decrease the execution time (0.56) and and the Energy per request (0.79). Moreover the average Latency (source / destination) is drastically reduced (0.18) so the congestion risks. These first results are promising and show an opportunity to improve the Energy efficiency and the performances (request rate) of the NoC so of the NMBC architecture. Considering the reduction of the latency we can also expect significant improvements with multiple managers processing numerous requests in parallel. Note also that the use of wireless links for long distance unicast communication also offer perspectives for very large networks with hundreds of memories where multi-hop paths would be prohibitive.

For 1 Manager query	**Wired NoC**	**Hybride NoC**	**Ratio**
Broadcast implementation	Multiple Unicast	Wireless: single hop Wired: replication	-
Routing	XY	WHIRL	-
Energy	2.636 μJ	2.085 μJ	0.79
Execution Time	57600 cycles	32500 cycles	0.56
Latency (all packets)	97.3 cycles	17.6 cycles	0.18
Latency (only broadcast packets)	75.2 cycles	10.3 cycles	0.14

Table I: Wired NoC vs hybrid wired/wireless NOC for a typical search engine request with the NMBC architecture. routers: 5 32b ports, 2 stage pipeline, 4 flits buffer, 2 virtual channels, 2 GHz

IV. Conclusion

A WiNOC architecture is a promising solution to improve the previously NMBC proposed architecture. An in-depth study is required to size the NoC and to implement the best use of Wireless and Wired routers for performance and energy optimization. But the first experiments confirm the idea that benefits can be expected from efficient broadcast implementation including Wireless links and avoiding multiple and redondant unicast packets. After memories and NoC optimisations, the next step should focus on the PE that were designed in our first study to implement the winner-take-all post-processing. In a more general-purpose MBC architecture, the PE should be configurable to implement different types of post-processing operations. Finally we can also expect significant WiNOC power reduction with technology node, digital communication techniques (e.g. ECC), radio channel accurate modelling and correction including absorption layer for instance [10].

Acknowledgments

Thanks to Dr. Navonil Chatterjee (Lab-STICC / UBS University) for simulations and fruitful discussions.

References

[1] H. Esmaeilzadeh, E. Blem, R. S. Amant, K. Sankaralingam, and D. Burger, "Dark silicon and the end of multicore scaling." in *38th Annual Int. Symp. on Computer Architecture*, June 2011.

[2] S. Deb, K. Chang, A. Ganguly, X. Yu, P. Pande, D. Heo, and B. Belzer, "Design of an energy efficient cmos compatible noc architecture with mm-wave wireless interconnects," *IEEE Transactions on Computers (TC)*, vol. 62, no. 12, Dec. 2013.

[3] J.-P. Diguet, N. Onizawa, M. Rizk, J. Sepulveda, A. Baghdadi, and T. Hanyu, "Networked power-gated mrams for memory-based computing," *IEEE Trans. on VLSI*, vol. 26, no. 12, Dec. 2018.

[4] H. K. Mondal, R. C. Cataldo, C. Marcon, K. Martin, S. Deb, and J.-P. Diguet, "Broadcast and power-aware wireless noc for barrier synchronization in parallel computing," in *31st IEEE Int. System-on-Chip Conference (SOCC)*, Washington DC, USA, Sep. 2018.

[5] V. Gripon and C. Berrou, "Sparse neural networks with large learning diversity," *IEEE Trans. on Neural Networks*, vol. 22, no. 7, Jul. 2011.

[6] N. Mansoor, P. J. S. Iruthayaraj, and A. Ganguly, "Design methodology for a robust and energy-efficient millimeter-wave wireless network-on-chip," *IEEE Transactions on Multi-Scale Computing Systems*, vol. 1, no. 1, pp. 33–45, Jan 2015.

[7] S. Abadal, J. Torrellas, E. Alarcón, and A. Cabellos-Aparicio, "Orthonoc: A broadcast-oriented dual-plane wireless network-on-chip architecture," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 3, pp. 628–641, March 2018.

[8] T. Krishna *et al.*, "Towards the ideal on-chip fabric for 1-to-many and many-to-1 communication," in *44th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2011.

[9] V. Catania *et al.*, "Noxim: An open extensible and cycle-accurate network on chip simulator," in *IEEE 26th Int. Conf. onApplication-specific Systems Architectures and Processors (ASAP)*, 2015.

[10] I. E. Masri, H. K. Mondal, T. L. Gouguec, C. Roland, P.-M. Martin, R. Allanic, C. Quendo, and J.-P. Diguet, "Accurate channel models for realistic design space exploration of future wireless nocs," in *12th IEEE/ACM Int. Symp. on Networks-on-Chip (NOCS)*, 2018.