Broadcast-Enabled Massive Multicore Architectures: A Wireless RF Approach

Sergi Abadal, Benny Sheinman, Oded Katz, Ofer Markish, Danny Elad, Yvan Fournier, Damian Roca, Mauricio Hanzich, Guillaume Houzeaux, Mario Nemirovsky, Eduard Alarcón, and Albert Cabellos-Aparicio

Abstract—Broadcast has been traditionally regarded as a prohibitive communication transaction in multiprocessor environments. Nowadays, such constraint largely drives the design of architectures and algorithms all-pervasive in diverse computing domains, directly and indirectly leading to diminishing performance returns as we reach the manycore era. Novel interconnect technologies could allow to revert this trend by offering, among others, improved broadcast support even in large-scale chip multiprocessors. In this position paper, the prospects of wireless on-chip communication technologies pointing towards low-latency (a few cycles) and energy-eff cient (a few pJ/bit) broadcast are outlined. This work also discusses the challenges and potential impact of adopting these technologies as key enablers of unconventional hardware architectures and algorithmic approaches, in the pathway of signif cantly improving the performance, energy eff ciency, scalability and programmability of manycore chips.

Index Terms—Broadcast; Wireless Network-on-Chip; Manycore; Hardware Architecture; Programming Models; Parallel Algorithms

1 Introduction

In the ever-changing world of microprocessor design, multicore architectures are currently the dominant trend for both conventional and high-performance computing. These were conceived to overcome the complexity and power scalability hurdles of processors with a single CPU; however, the scalability concerns have now migrated to facets such as memory management, programmability or the limits of parallelism speedup.

Inherent parallelism limits aside, these scalability concerns are generally dependent on the architecture or programming model of choice. A long-running debate has brought up strong arguments for the adoption of two widely-known models, namely, shared memory and message passing [1], as summarized in Fig. 1. Shared memory provides remarkable programmability and compatibility with legacy code, while incurring a low overhead due to data replication for the small scale. However, it is argued that its scalability is limited by performance and architectural complexity issues. On the contrary, message passing offers unique validation and performance benefits, which come at the cost of placing an increasingly heavy burden upon the programmer. The differences between these two extremes contrast with one common point: most of the scalability issues are tightly coupled to on-chip communication limitations. Unsurprisingly, the capabilities of the on-chip interconnect have steered the design of cache coherence and parallel algorithms through the decades.

The ability of the interconnect to handle multicast and broadcast communication has been one of the main drivers of this trend. For a few cores, buses with ordered broadcast capabilities are feasible and broadcast-based architectures deliver remarkable performance. As the number of cores grows, though, interconnect design shifts to the Network-on-Chip (NoC) paradigm. NoCs are more suitable for concurrent point-to-point communications, whereas broadcast becomes a costly feature: the latency increases proportionally to the average distance to the furthest node (e.g. $O(3k)$ *hops* in a planar k-ary mesh) and the available bandwidth drops due to contention: Moreover, ordering consistency of delivered messages is rarely guaranteed. These issues encourage the design of architectures and algorithms that use concurrent unicast messages, negatively impacting upon complexity and performance.

In shared memory systems, memory coherence and consistency are the main implicit sources of on-chip communication. With the advent of NoCs, broadcast-based snooping mechanisms have progressively given way to directory-based protocols, which generally limit the use of multicast to the invalidation of cache blocks on a shared write. However, this comes at well-known costs: area and energy overheads are required to track the sharers of the cache blocks, whereas the use of the directory as ordering point introduces indirection in the critical path of misses and, therefore, penalizes performance. Although these aspects may not completely preclude the use of multicast-avoiding architectures in future manycore chips [2], diminishing performance returns can be expected.

In message passing systems, communication is explicit and needs to be carefully orchestrated in order to maximize performance. A set of routines are available to the programmer to implement simple and collective communication. In MPI, `MPI_Bcast` performs a broadcast to all cores, whereas `MPI_Allreduce` and `MPI_Allgather` are all-to-

- *Sergi Abadal, Eduard Alarcón and Albert Cabellos-Aparicio are with the NaNoNetworking Center in Catalonia (N3Cat), Universitat Politècnica de Catalunya, Barcelona, Spain.*
  Corresponding E-mail: abadal@ac.upc.edu
- *Benny Sheinman, Oded Katz, Ofer Markish, Danny Elad are with the mmWave Technologies Group, IBM Research - Haifa, Israel.*
- *Yvan Fournier is with EDF R&D, Chatou, France.*
- *Mario Nemirovsky is an ICREA Senior Research Professor at the Barcelona Supercomputing Center (BSC), Barcelona, Spain.*
- *Damian Roca, Mauricio Hanzich and Guillaume Houzeaux are with the BSC, Barcelona, Spain.*

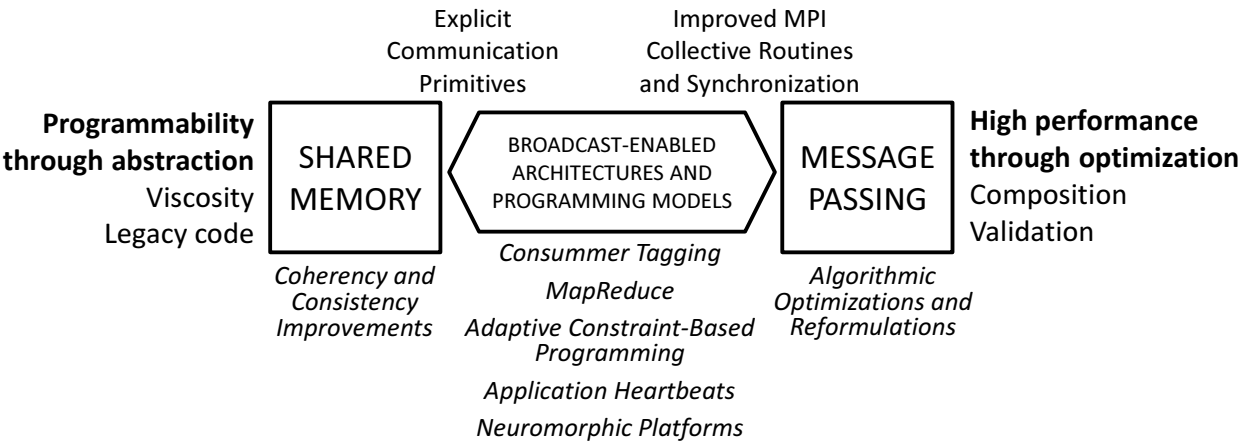


Fig. 1. Representation of the manycore design space opened by broadcast between the shared memory and the message passing paradigms.

all routines that directly use broadcast at some point. The main issue is that, as the number of possible destinations increases, the cost of these operations becomes prohibitive and strongly limits their use. Avoiding broadcast is therefore a widespread practice, even though it lowers the maximum achievable performance and increases the complexity of parallel programming even further.

The contradicting tendencies found in both programming models suggest that an effective broadcast plane may be not only valuable, but even necessary to avoid slowing down progress. The potential performance and programmability benefits would be substantial since:

- The cost of operations based upon broadcast, e.g. synchronization, would be significantly reduced.
- The performance of a wide set of architectures and programming models would be greatly improved.
- The constraints relative to the use of broadcast would be relaxed, enabling the adoption of techniques overlooked at the manycore scale and opening a set of unexplored possibilities.

In this position paper, we aim to defy conventional wisdom by turning broadcast communication into the central pillar of future manycore architectures and algorithms. To support this vision, we propose a hybrid network composed of two dedicated planes: a wireless technology-enabled globally shared medium that will provide native hardware support for broadcast, and a conventional NoC that will serve the rest of the communication flows. While the return of shared media to the multicore scenario has been already discussed in the literature [3], most proposals use the shared medium *in combination with* the conventional NoC to better adapt to the demands of current multicore architectures. Instead, by considering both network planes to be independent and using the shared medium exclusively for broadcast, our approach has a strong broadcast-oriented edge that may impact on the design of new manycore architectures and algorithms in many ways.

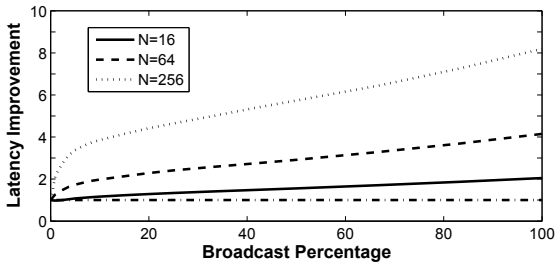
Here, we first quantify the potential of the proposal, to then analyze the underlying implementation requirements and discuss how close is wireless on-chip technologies to fulfilling them. After that, we delve into the possible architectural and algorithmic breakthroughs that may stem from a more relaxed use of broadcast communication. As a roadmap for future investigations, we conclude the paper by outlining the main challenges that need to be addressed from the implementation, communication, and architecture perspectives to realize this vision.

2 Future Enablers of an Effective Broadcast Plane

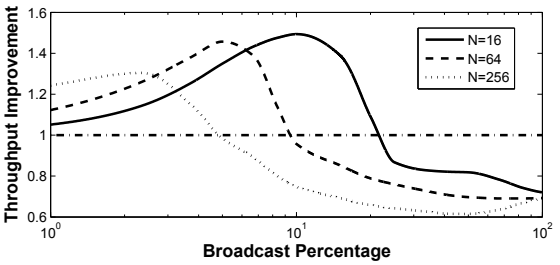
Considerable efforts have been recently devoted towards slowing down the broadening of the gap between multicast requirements and conventional NoC performance. MIT leaded several of the hardware research initiatives, which sought to improve broadcast performance and provide ordered delivery over conventional NoC fabrics. First designs attained a latency reduction of between 14% and a 50% in broadcast-intensive scenarios, revealing execution speedups of up to 40% in PARSEC benchmarks. Their research later culminated into SCORPIO [4], a 36-core prototype that implements snooping coherence and consistently achieves a 25% execution speedups with respect to directory-based schemes. Even though these works are a huge step towards demonstrating the feasibility and benefits of broadcast-based manycore architectures, it also recognizes important scalability concerns beyond 100 cores.

Our proposal breaks away from the SCORPIO approach and, instead, advocates for the use of a hybrid NoC composed of two *independent* network planes: a conventional NoC that would serve unicast flows, and a globally shared medium that would transport broadcast messages. Communication within the globally shared medium will be implemented by means of a Wireless Network-on-Chip (WNoC) where antennas are integrated on a per-core basis, and with a transmission bandwidth of approximately one message per clock cycle. This scheme not only provides scalable and low-latency support for broadcast, but also removes a potentially heavy burden from the conventional NoC.

To quantify the potential improvements from a network perspective, we simulated SCORPIO and then added an independent shared medium to be used as complementary broadcast plane. Traffic is synthetic, composed of one-flit



(a) Latency



(b) Throughput

Fig. 2. Performance improvement of a hybrid NoC with respect to that of SCORPIO as a function of the broadcast percentage and for different system sizes. Dash-dot line represents the break-even boundary.

messages with a variable percentage of broadcast transmissions. As shown in Figure 2a, the low-load latency is cut by a factor proportional to both the system size and the percentage of broadcast transmissions. As shown in Figure 2b, the throughput improves for a wide range of broadcast percentages. Beyond a certain point, however, the superior bisection bandwidth of the conventional NoC outweighs the benefits of having a hybrid network, suggesting the need of load balancing mechanisms in the presence of extreme broadcast demands.

Besides wireless RF, two other emerging technologies (RF through transmission lines and nanophotonics) have recently been in the spotlight due to their unique properties [3] and could be also considered for the implementation of the broadcast plane. However, design complexity issues in the former and laser power problems in the latter hinder their use as globally scalable shared media. Instead, the wireless RF paradigm built upon the pioneering work by Chang et al [5], shows outstanding promise given its inherent broadcast nature and its:

- *Simplicity:* since no path infrastructure is required, non-intrusive solutions can be devised and chip floorplanning costs reduced. The concept of *wireless core* module, wherein one wireless unit is integrated per core or group of cores, could simplify the design of large-scale architectures.
- *Flexibility:* the logical topology or other transmission parameters can be modified to provide adaptability without the need of any physical modification.

The realization of the proposed vision is enabled by the huge advancements made in different key technologies, which impact upon the area occupation, energy efficiency and bandwidth offered by the broadcast plane.

2.1 Scaling CMOS Technology

To fully take advantage of the proposed approach, wireless communication must be implemented at the core level through a single broadband channel with a capacity of around one flit per cycle. This implies that the wireless unit should [i] have a size between 0.01 and 0.1mm^2, [ii] offer a data rate of several tens of Gbps with an extremely low Bit Error Rate (BER), and [iii] consume less than 1 pJ/bit/core to provide an energy efficiency commensurate to that of related work [4]. To reach these goals, antennas and transceivers need to operate at high frequencies since this reduces both the area occupation and energy per bit and increases the available bandwidth [6]. Specifically, a frequency of several hundreds of GHz seems an appropriate target.

A growing number of publications report advancements in the design of on-chip antennae [7], [8] and transceivers [6] working at the targeted frequency bands. Actual implementations currently modulate data on a high frequency carrier in the V band range (40-75 GHz) using simple schemes that reduce the area and power footprint. For instance, Yu *et al* propose a 60-GHz transceiver that performs close to the aforementioned objectives by providing 16Gbps with a BER of 10^{-15}, while occupying 0.3mm^2 and consuming 30mW ($\sim$2pJ/bit/core) [9]. As CMOS technology evolves and advanced CMOS devices such as FinFETs and III-V on silicon are implemented, it becomes possible to further raise the carrier frequency into the mmWave region (up to 300 GHz), thereby significantly increasing the available signal bandwidth and decreasing the energy required per bit of data. Promising results were achieved in recent years also with fast plasma-wave detectors implemented in scaled CMOS technologies [10]. Following these advancements, first transmitter/receiver circuits have been already proposed for multigigabit communication at frequencies up to 400 GHz and imaging up to 800 GHz [6].

2.2 Graphene Technology

Novel technologies such as graphene are being carefully inspected as they could introduce further improvements at the chip scale. On the one hand, graphene-based planar antennas are expected to radiate in the lower part of the terahertz band (300 GHz to 3 THz) while being barely a few micrometers in size, this is, between one and two orders of magnitude smaller than their metallic counterparts [11]. With this, the footprint of the WNoC in terms of area would be significantly reduced. On the other hand, preliminary results have shown that graphene-based components are excellent candidates for ultra-high-frequency applications: impressive cut-off frequencies of 350 GHz have been obtained in Graphene Field-Effect Transistors (GFETs) due to the high carrier mobility in the nanomaterial [12]. Graphene is also uniquely suited for Low Noise Amplifiers (LNAs) as it theoretically offers high frequency and low noise.

2.3 Surface Wave Technology

A recent work has proposed the use of engineered surfaces that support the propagation of Zenneck surface waves [13]. Early-stage results confirm that, instead of propagating in

all directions, radiated signals are bound to the surface and propagate along it. This shrinks the spreading loss from $O(1/d^2)$ to approximately $O(1/d)$, implying a strong reduction on the path loss and, therefore, on the energy required to obtain the target BER at the receivers. Moreover, surfaces can be designed to support the bounding effect for a broad range of frequencies. This paradigm also practically eliminates the propagation of signals at the vertical direction, opening the door to future stacked designs with different orthogonal ultra-broadband channels enabled by a sort of *surface multiplexing*. If feasible, this technology could represent an important breakthrough towards the implementation of energy-efficient and high-throughput shared media.

3 Broadcast-Enabled Architectures and Algorithms

Having an effective broadcast plane with low latency and total ordering potentially relaxes a large number of constraints cast upon architects and parallel programmers. In shared memory, the introduction of such network plane is expected to bring the manycore coherence solution space much closer to the area characterized by low architectural costs, low complexity, and high performance. The cost of synchronization will also be strongly reduced. Due to *atomic* nature of broadcasts, locks will be more effective and easier to implement; whereas barriers may take advantage of the peculiarities of RF signaling over a shared medium to virtually eliminate serialization in counting arrivals [14]. This could lead to performance and programmability improvements stemming from both the exploitation of parallelism on a finer granularity and the reduction of the penalty of maintaining sequential consistency.

In message passing, all-to-all routines could be redefined seeking greater performance and lower cost. These improvements may enable the development of alternative approaches for widely used kernels and applications. The achievable speedup could be dramatic in applications where global sharing and coordination are the main bottlenecks. As an example, Fig. 3 shows the trace of communication actions within the so-called Deflated Conjugate Gradient iterative solver for a specific example described in [15]. This solver performs much better than the classical Conjugate Gradient solver in terms of iterations, but exhibits a bottleneck due to an *allreduce* communication. For this particular example, the *allreduce* in question is of 4000b and its duration is that of a matrix-vector product. Although the iterative solver approach can achieve a speedup of ten in terms of CPU time in this case, the duration of the *allreduce* operation reduces the speedup to four. This demonstrates the great importance of such global communications in iterative solvers.

While the potential advantages are manyfold, thus far there is little manycore architecture research revolving around advancements on efficient on-chip broadcast. A new breed of race-free coherence protocols based on the use of broadcast is proposed in [16]. Broadcasts are used to acquire fine-grained mutexes that serialize requests to conflicting addresses and, thus, eliminate race conditions. Other works have analyzed the impact of improved broadcast

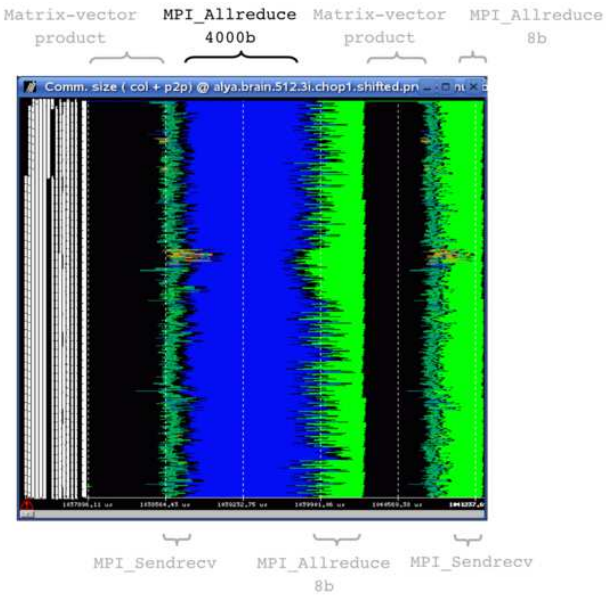


Fig. 3. Graphical representation of the inter-process communication over time. Each row represents a thread.

over the performance of traditional architectures rather than proposing new schemes. For instance, snooping and limited directory protocols have been evaluated using the SCORPIO prototype [4], revealing an speedup of 20% when using snooping coherence.

We believe that this scarce related work is just the tip of an iceberg of novel methods for manycore computing, which will be revealed once the feasibility of a globally shared medium is demonstrated. We anticipate the following advancements towards greater performance and programmability:

Coherence Mechanisms: traditional broadcast-based coherence mechanisms will be revisited. For instance, the protocol implemented by the Sun's Fireplane system would map especially well to the proposed hybrid network, as it originally snooped addresses over a bus and transferred data through a switched network. Additionally, new innovations such as race-free protocols [16] will be embraced. Also, the use of explicit communication primitives in shared memory environments could be considered, even if that means trading off programmability for performance. For instance, critical variables could be directly managed through the wireless medium to, for instance, eliminate potentially long delays in the critical path of L2 misses.

Programming Models: the differentiated management of critical variables has been already proposed in order to eliminate the bottlenecks caused by producer-consumer access patterns. In this approach, referred to as *Consumer Tagging*, the programmer tags the potentially conflicting addresses and the consumers associated to them. This represents a particular case of the *"always update through broadcast"* approach explained above, where the programmer does not need to know the consumer set. In related work, Psota *et al* discuss this case along with other new programming models that aim to combine the flexibility of shared medium and the hand-tuned performance of message passing, and that

would benefit from an improved broadcast mechanism [17]. They analyze *Adaptive Constraint-Based Programming* and *Application Heartbeats*, both of which periodically broadcast information. The former uses broadcast to communicate or update a set of known constraints to worker cores; whereas, in the latter, worker cores broadcast their performance to scheduler cores, which later send instructions and performance goals to worker cores also through broadcast. Finally, we speculate that an efficient broadcast plane could be also used to improve the performance and redundancy properties of the well-known MapReduce model.

Software Architecture and Algorithms: a first objective will be to revisit the MPI routines that implement all-to-all communication, to then re-evaluate a set of representative algorithms. Native broadcast support will clearly benefit routines that use broadcast primitives, i.e. `MPI_Allgather` and `MPI_Allreduce`; whereas improving other all-to-all routines may rely on the careful orchestration of communication through the wired and wireless networks. These collective communication patterns are of great relevance in pervasive applications such as computational mechanics solvers in general, e.g. computational fluid dynamics (CFD) or computational structure dynamics (CSD); and in the specific case of iterative solvers, where selected parts of the code generally show degraded scaling behavior at high core counts.

If a natural broadcast communication can be provided, a vast area for re-definition of algorithms will be open beyond simply improving all-to-all communication support. The designer will need to take into consideration optimizations that may involve the intense sharing of data among a large set of threads, or the use of oftentimes avoided synchronization primitives and all-to-all communication patterns. For example, after many OpenMP loops there is an implicit barrier that may be enhanced using broadcast-based synchronization mechanisms like the proposed in [14].

Even though substantial performance improvements could be obtained through optimization, maximum gains will be only achieved by going back to the original problem and re-implementing a given algorithm without the constraints imposed by conventional on-chip communication. This is indeed a daunting task, yet cannot be totally discarded as the potential scalability, performance, or efficiency improvements may be huge.

Novel Computing Systems: On a separate note, it is worth remarking that the introduction of an effective broadcast plane may have a profound impact on the design of neuromorphic computing systems. Recent years have seen the rise of such brain-emulating systems, which model a set of neurons within each core and communicate cores with multicast messages that simulate the behavior of neural spikes. Given the multicast-driven nature of such communication, important improvements can be expected.

4 Associated Challenges

The adoption of broadcast communication as the basis of a new generation of manycore processors brings up a wide variety of research challenges spanning four well-defined areas:

4.1 Wireless Transceiver Implementation

In order to incorporate the envisioned wireless NoC transceivers in advanced CMOS technologies, many challenges need to be overcome. The large bandwidth requirement compels the use a carrier frequency in the upper mmWave range. Although technology scaling allows using CMOS devices at mmWave or even THz frequencies, aspects such as the device parasitics, low supply voltage, device dimension limitations, and the complex metal stack (which adds parasitic intrinsic inductance) limit the transceiver operation frequency and performance.

Silicon area constraints suggest the use of simple circuit topologies, employing direct conversion architectures and injection-locked Voltage Controlled Oscillators (VCOs) instead of full Phase Locked Loops (PLLs). The design is further complicated by the small available area, power budgets, substrate noise, and the use of a technology optimized for digital operation rather than RF-oriented. These challenges can be addressed by the use of forward body biasing to lower the threshold voltage, the inclusion of dedicated CMOS devices with improved RF performance, or the adoption of improved technologies such as III-V semiconductors and 3D integration.

Designing an on-chip mmWave antenna close to the transceiver requires a high level of integration and area reuse. Using on-chip antennae has the benefit of enabling higher pattern accuracy and the dismissal of waveguide transitions to the package, which are a cause of severe losses. Embedded antennae, however, may have lower efficiency due to power absorption in the doped silicon chip, and the high interconnect metal density can interfere with their operation. Using artificial screening surfaces such as electromagnetic band-gaps, or applying backside etching might be necessary to minimize the on-chip losses. Since the antennae must operate inside a multicore environment, reflections between cores could cause channel degradation due to destructive interference and dispersion. In addition, the antenna is required to be omnidirectional and communicate with several nearby cores. Although an omnidirectional antenna consumes smaller area, the structure is more sensitive to its environment and this may enhance its vulnerability. On the other hand, one may also try to harness the reflections in order to actually enhance the antenna performance.

These challenges will require simulating the antennae together with the surrounding cores, which is computationally heavy due to the large physical volume of the model. One might minimize the effort by enforcing periodicity and simulating only a unit cell, or by using fast two dimensional methods taking advantage of the planar nature of the system.

4.2 Communications and Networking

The peculiarities of the on-chip wireless scenario require us to rethink the entire protocol stack with respect to classical wireless networks. Finding lightweight modulation schemes able to maintain a graceful compromise between data rate and transceiver complexity is a grand challenge by itself. However, given a shared medium in such a communication-intensive context, the way nodes gain access to it plays an

even more decisive role in determining the performance of WNoC. Therefore, the main challenge is to develop Medium Access Protocol (MAC) capable of guaranteeing error-free broadcast delivery in a cost-effective and fair manner. Unfortunately, the optimal solution generally depends on the level of contention. Schemes where nodes contend for the channel perform remarkably well as long as channel accesses are infrequent, which limits its throughput; whereas the use of arbitration mechanisms that avoid collisions deliver better performance in high-contention situations. A possible solution to this issue would be the development of protocols that dynamically switch between both options as a function of the level of contention. The decision process can be reactive based on an estimation of the level of contention, or proactive based on hints that may be intelligently placed by the compiler.

4.3 Architectures and Algorithms

The development of disruptive manycore methods and applications may find the first challenge in the lack of multi-scale simulation tools that encompass all the novelties introduced by our proposal, as well as its expected computational cost. To reduce it, the behavior of irrelevant parts of application code can be represented by performance modeling or tested on separate simulators with reasonable effort; whereas the performance of the region of interest should be accurately simulated.

From an architectural perspective, another important aspect to investigate is the impact of the RF transceiver-induced interference upon the architecture of a processing tile. The challenge here is to find the optimal component distribution which minimizes crosstalk effects without compromising wire routability. To solve this, RF interference needs to be modeled and integrated within architectural exploration tools.

From a programming perspective, the main challenge resides in the identification of the optimization opportunities stemming from the improved broadcast support. While several cases are widely known in the research community, others may require further investigation. Such exploratory work would benefit from the refining of profiling techniques, which could allow researchers to define new situations of interest and capture them in parallel applications on a phased basis.

5 Conclusions

The cost of broadcast has been constraining the design of manycore processors and of the algorithms that run upon them. However, advancements in CMOS RF, graphene RF and surface wave technologies could lead a paradigm shift, as native hardware support for low-latency and low-power broadcast could be implemented via wireless communication even in manycore chips. In shared memory, this approach would allow overcoming the so-called *coherence wall* by reducing complexity and increasing performance. It would also improve support for new programming models and computing systems. In message passing environments, the unprecedented availability of inexpensive broadcast could open the door to a wealth of optimization and reformulation opportunities. Even though this vision requires that a broad set of research challenges be addressed, the early-stage demonstration of its potential benefits is expected to create a technology pull that will pave the way for the realization of next-generation manycore processors.

Acknowledgments

The authors gratefully acknowledge support from the INTEL Student Honor Program and the Samsung GRO program.

References

[1] R. Kumar *et al.*, "The case for message passing on many-core chips," in *Multiprocessor System-on-Chip*, Springer, 2011, pp. 115–123.
[2] M. M. K. Martin *et al.*, "Why on-chip cache coherence is here to stay," *Commun. ACM*, vol. 55, no. 7, p. 78, 2012.
[3] J. Kim and K. Choi, "Exploiting New Interconnect Technologies in On-Chip Communication," *IEEE Trans. Emerg. Sel. Topics Circuits Syst.*, vol. 2, no. 2, pp. 124–136, 2012.
[4] B. Daya *et al.*, "SCORPIO: a 36-core research chip demonstrating snoopy coherence on a scalable mesh NoC with in-network ordering," in *Proceedings of ISCA-41*, 2014, pp. 25–36.
[5] M.-C. F. Chang *et al.*, "RF/wireless interconnect for inter- and intra-chip communications," *Proc. IEEE*, vol. 89, no. 4, pp. 456–466, 2001.
[6] S. Abadal *et al.*, "On the Area and Energy Scalability of Wireless Network-on-Chip: A Model-based Benchmarked Design Space Exploration," *IEEE/ACM Trans. Netw.*, vol. 23, no. 5, 2015.
[7] D. Hou *et al.*, "Silicon-based On-chip Antenna Design for Millimeter-wave / THz Applications," in *Proceedings of the EDAPS '11*, 2011, pp. 130–133.
[8] B. Klein *et al.*, "Design of a Cloverleaf Antenna for an Antenna Coupled Bolometer for Room Temperature THz Imaging," in *Proceedings of the ISCDG '13*, 2013, pp. 1–4.
[9] X. Yu *et al.*, "Architecture and Design of Multi-Channel Millimeter-Wave Wireless Network-on-Chip," *IEEE Des. Test.*, vol. 31, no. 6, pp. 19–28, 2014.
[10] D. Glaab *et al.*, "Terahertz heterodyne detection with silicon field effect transistors," *Appl. Phys. Lett.*, vol. 96, no. 52, pp. 042106, 2010.
[11] S. Abadal *et al.*, "Graphene-enabled Wireless Communication for Massive Multicore Architectures," *IEEE Commun. Mag.*, vol. 51, no. 11, pp. 137–143, 2013.
[12] Y. Wu *et al.*, "Graphene Electronics: Materials, Devices, and Circuits," *Proc. IEEE*, vol. 101, no. 7, pp. 1620–1637, 2013.
[13] A. J. Karkar *et al.*, "Hybrid wire-surface wave interconnects for next-generation networks-on-chip," *IET Comput. Digit. Tec.*, vol. 7, no. 6, pp. 294–303, 2013.
[14] J. Oh *et al.*, "TLSync: support for multiple fast barriers using on-chip transmission lines," in *Proceedings of the ISCA-38*, 2011, pp. 105–115.
[15] R. Löhner *et al.*, "Deflated Preconditioned Conjugate Gradient Solvers for the Pressure-Poisson Equation: Extensions and Improvements," *Int. J. Numer. Meth. Eng.*, vol. 87, no. 1-5, pp. 2–14, 2011.
[16] D. Vantrease *et al.*, "Atomic Coherence: Leveraging nanophotonics to build race-free cache coherence protocols," in *Proceedings of the HPCA '11*, 2011, pp. 132–143.
[17] J. Psota *et al.*, "ATAC : Improving Performance and Programmability with On-Chip Optical Networks," in *Proceedings of the ISCAS '10*, 2010, pp. 3325–3328.