Rotary Router: An Efficient Architecture for CMP Interconnection Networks

Pablo Abad, Valentín Puente, Pablo Prieto, and Jose Angel Gregorio
University of Cantabria
Av. Los Castros S/N
39005 Santander, Spain

{pablo,vpuente,prietop,jagm}@atc.unican.es

ABSTRACT

The trend towards increasing the number of processor cores and cache capacity in future Chip-Multiprocessors (CMPs), will require scalable packet-switched interconnection networks adapted to the restrictions imposed by the CMP environment. This paper presents an innovative router design, which successfully addresses CMP cost/performance constraints. The router structure is based on two independent rings, which force packets to circulate either clockwise or anti-clockwise, traveling through every port of the router. It uses a completely decentralized scheduling scheme, which allows the design to: (1) take advantage of wide links, (2) reduce Head of Line blocking, (3) use adaptive routing, (4) be topology agnostic, (5) scale with network degree, and (6) have reasonable power consumption and implementation cost. A thorough comparative performance analysis against competitive conventional routers shows an advantage for our proposal of up to 50 % in terms of raw performance and nearly 60 % in terms of energy-delay product.

Categories and Subject Descriptors

C.2.1 [**Parallel Architectures**]: *Distributed architectures.*

General Terms

Algorithms, Performance, Design, Experimentation.

Keywords

Router architecture, Interconnection networks, Chip Multiprocessors, NUCA.

1. INTRODUCTION

Chip Multiprocessors (CMP) seem to be the most effective way to deal with the increasing design complexity for actual and future microarchitectures. The organization of their memory hierarchy is a first-order design issue at chip level, including its supporting communication subsystem. Like the bonding substrate of other components in the chip, the interconnection network must scale with future transistor budget increments. Nowadays, centralized structures are common in commercial CMPs [15][25], but a larger number of functional blocks [3][4] in the system will eventually require decentralized structures. Packet switched point-to-point networks are postulated as the best candidate to accomplish this challenge [5].

Point-to-point networks are a well-known topic in classical multiprocessor environments and this knowledge can only be applied to the CMP scenario if technological constraints are kept in mind. It is mandatory to consider the boundary conditions prior to proposing a network architecture for interconnecting the building blocks of a CMP architecture. Inside the chip, network and upper levels of the system are significantly closer than in off-chip networks. This means lower latency and higher raw bandwidth availability but at the cost of increasing implementation costs and power restrictions.

With CMP architectures, both power consumption and area requirements are decisive factors in the network design [2]. The growing number of processing elements and restrictions in refrigeration systems limit the feasibility of using complex interconnection network [23]. Moreover, a complex network will decrease the total budget of transistors devoted to other crucial components of the system. The work in [17] details how a high performance but complex network will reduce cache sizes and/or other crucial elements lowering the overall system performance.

Low cost router architecture imposes the usage of input buffers with simple FIFO policy and, as in off-chip networks, this technique will produce Head of Line (HOL) blocking [9]. However, complex input buffers, output buffers (physical or virtual) or any other kind of centralized internal storage to mitigate HOL blocking are not feasible in a CMP framework. In the same way, real traffic patterns are distant from the balanced usage of network resources when deterministic routing algorithms are employed. Network-state dependent routing algorithms should be used in order to achieve best utilization and maximum performance. However, adaptivity also increases the router complexity because of the scheduling and deadlock avoidance mechanisms.

For on-chip networks, link wire availability is substantially higher than for off-chip networks. Cache lines and/or coherence protocol commands compose the traffic interchanged among on-chip building blocks. Because of both these characteristics, packet length in CMP systems will be noticeably smaller than in an off-chip network [8]. For example in [5] and [30], a 128-bit wide link is considered a suitable choice for on-chip networks. If we

ISCA'07, June 9–13, 2007, San Diego, California, USA.

consider a system with 64-byte cache lines and 16-byte protocol commands, the packet size ranges from five phits (command plus cache line) to one phit (command). If we compare this with some off-chip multiprocessors, such as the one based on Alpha 21364 [21], the maximum packet size is up to 5 times longer. At first glance, having small packets is not a problem until we analyze its impact on conventional routers performance. Figure 1 shows the performance of an 8x8 Torus network with Bubble routers (either adaptive or deterministic as in [28]) under synthetic uniform traffic for different packet lengths. Both routers use FIFO input buffers, and were tested under constant buffer space. Reducing the packet length from 20 phits down to 2 reduce the potential performance of the adaptive router by almost 45%. In other words, if router links are made five times wider, only 55% of the bandwidth improvement will be effective. This additional contention is due to more frequent packet arbitration. Although the adaptive router employed uses a feasible but aggressive arbiter, similar to the one employed in the Alpha 21364 router [21], its behavior when packets are extremely short degrades performance. As we can see, deterministic routers present much lower sensitivity to packet length, but with up to 30% performance loss compared to adaptive routers with large packets. Consequently, if we want to empower network performance using adaptive routing we need arbitration mechanisms immune to packet length.

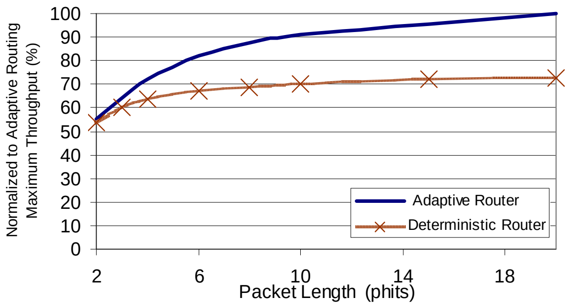

Figure 1. Packet length impact in adaptive and deterministic input buffered routers.

All in all, the challenge faced in the design of a router for a CMP interconnection network is a hard task. Reaching the necessary trade-offs slightly modifying conventional router architectures seems very difficult. For this reason, we try to address this problem from a radically different point of view. The present work copes with this situation by proposing a new architecture that fulfils the main requirements successfully with a sustainable cost. The architecture is based on a router, denoted as Rotary Router, which not only minimizes effects of small packets but also takes advantage of them, has no appreciable HOL blocking, and allows the use of topology agnostic adaptive routing.

The rest of the paper is organized as follows: Section 2 introduces the Rotary Router architecture. Section 3 explains how network anomalies are avoided. Section 4 shows some performance results. Section 5 addresses the implementation cost of the router. Section 6 discusses related research and, finally, Section 7 states the main conclusions of the paper.

2. THE ROTARY ROUTER

In this section, we will provide a detailed router architecture and describe its operation. We will focus on the main differences of the Rotary Router compared to more classic architectures and on the advantages it presents when working with CMPs. Aspects such as flow control mechanisms and routing algorithm are also described.

2.1 General Router Structure

Trying to avoid the appearance of negative effects present in input buffered structures, the introduction of radical changes in the router design seems essential. On the one hand, in order to minimize contention effects on performance, the Rotary Router should not make use of centralized arbitration mechanisms nor centralized crossbar. For this reason, arbitration should be done independently at each router output port and independent of the number of input ports. On the other hand, non-FIFO buffers involve a high cost [32], so in order to deal with the HOL blocking problem while maintaining buffer FIFO policy, we need some mechanism that allows the packets at the head of the queue to leave the buffer, even when they have not obtained their profitable output port. This would enable the advance of the packets waiting behind the one blocked in the head of the buffer. Finally, it would be preferable that the number of router ports or the routing algorithm do not increase router complexity. In order to address all the aforementioned requirements, the way of connecting the components inside the router has to be completely new, while some common elements present in conventional architectures should disappear.

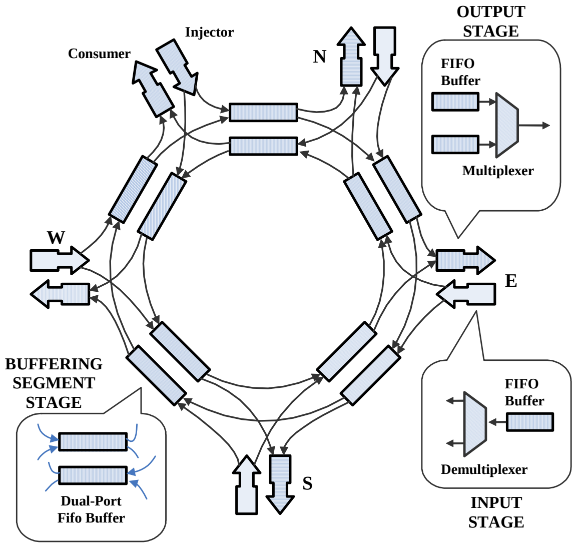

Figure 2. Rotary Router sketch.

Figure 2 shows a sketch of the router for a 2-degree network with one host attached. The structure of the Rotary Router is based on two independent rings, which force packets to circulate either clockwise or anti-clockwise, traveling from port to port of the router. Each ring is built with a group of Dual-port FIFO Buffers (DFB). The operation of the Rotary Router is simple, when a packet arrives at a router input port it is sent to one of the rings which forms the router. The packet starts moving towards its output port using the DFBs of the ring. Once the packet reaches a profitable output port, there are two possible cases; if the suitable output port is available, the packet will leave the ring and advance to the next router. Otherwise, i.e. another packet is in transit through the same output port or the remote node has not enough room for a packet, the packet will keep on circulating the ring until reaching another profitable output port. When necessary, the packet will be forced to complete a full turn in the ring and start a second turn. The packet is able to do as many turns as needed to leave the router.

The circulation of packets inside the router presents some potential drawbacks but numerous advantages. Each packet is able to go through any output port avoiding the usage of a centralized arbitration. Adaptive routing does not add complexity to the router. It is enough to allow a packet to leave the router by any available output that approximate it to its destination. The reduction of the HOL blocking is obvious; when a packet at the head of a DFB is not able to obtain a grant for the output port, it moves toward the next DFB in the ring, allowing the advance of the rest of the packets behind it.

2.2 Router Building Blocks

The Rotary Router is made up of a number of building blocks proportional to the number of router ports (see Figure 2). The three types of blocks are INPUT, OUTPUT and BUFFERING SEGMENT. The first and second constitute the input and output ports of the router and the third one makes up the rings. The structure and complexity of each block are both independent of the node degree making the router easily scalable. Next, we will describe the basic function of each block.

2.2.1 Input Stage

The datapath of this block is made up of a FIFO buffer and a demultiplexer. Depending on network topology, current node and destination node, this stage is responsible for computing the profitable output ports for each packet entering the router. These can be obtained by using any valid method (table-based routing, arithmetic routing, etc.) and must be added to the packet header. The overhead introduced by this information is small because it is only used inside the router and then discarded. The input stage also selects the ring direction in which the packet will move inside the router. This choice is made in order to minimize the delay of the packet in the router. The delay depends mainly on two factors, the number of DFBs traversed by the packet, and the time spent going through each DFB. At low load, ring selection minimizes the number of DFBs needed to reach the nearest profitable output. At medium-high load, the time spent traversing a DFB becomes the dominating factor, and under these conditions, the selection mechanism changes and the ring with lower occupation will be chosen. In order to take the right decision, the input stage only needs to check the occupation of the DFBs connected to it, because the occupation of all the buffers in the same ring is similar, as we will see in Section 2.3.

2.2.2 Output Stage

This stage is responsible for getting the packets out of the rings and sending them to a neighbor router. Given that packets belonging to both rings can try to access the same output port at the same time, this stage has two buffers and one multiplexer for sharing the unique physical channel. The multiplexer employs a fair policy to guarantee uniform usage of both rings. Note that Virtual Channels are not required. The presence of output buffers is essential to achieve optimal network performance. This output buffering space helps to improve the links utilization because it stores packets capable of using the link as soon as it becomes available. Along with the input stage, this stage is responsible for applying Flow Control mechanism between contiguous routers.

2.2.3 Buffering Segment Stage

This can be considered as the most important block of the router. This stage provides the router with its functionality and is the source of the multiple advantages of the Rotary Router. This part is made up of two DFBs connecting every two router ports. Each DFB has two pairs of Read/Write (R/W) ports. One pair is used to build a ring (connecting with previous and next DFBs) in which packets turn, while the other one connects the buffer to Input and Output Stages. On each router two independent rings are found, each made up of a number of DFBs equal to the number of input ports of the router. The two rings have opposite directions in order to minimize the number of DFB traversals. This stage must decode the routing information generated by the input stage and included in every packet header. If a profitable output port of the router is available, the Buffering Segment Stage must use the read port connected to the corresponding Output Stage. Otherwise, the read port connected to the write port of the next DFB should be used.

2.3 Flow Control and Routing Algorithm

In the Rotary Router, flow control and routing mechanisms are strongly bounded to the deadlock avoidance method. In this section we will introduce these mechanisms briefly. In section 3, we will complete the missing details. As shown in Figure 3, three flow control mechanisms coexist inside the router; one of them controls the advance between routers, another one manages packet movement in the rings inside the router and the last one controls the access to these rings.

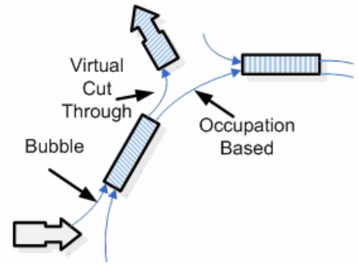

Figure 3. Flow control mechanisms in the Rotary Router

Virtual Cut Through is used to control the advance of packets among routers [11]. The packet injection to the rings is regulated by the Bubble flow control mechanism [28]. In order to allow the access of a packet to any ring, the buffer requested must have room for at least two packets. If the input port is connected to a processing element, the number of required holes increases to three. Finally, in order to manage the advance of packets in the rings inside the router an occupation-based flow control is used. Each DFB keeps information about its own occupation level, and this information is shared between contiguous DFBs. In an on-chip environment this is not a problem, due to the router blocks' proximity [6]. In this way we can allow a packet to advance to the next DFB only if the occupation of the destination buffer is less than or equal to the occupation of the current buffer. This flow control helps to balance the occupation of all the DFBs in the ring and to equalize the ring injection probability at each input port.

With respect to the routing algorithm, the packet always tries to leave each router on its path through a profitable output port. Therefore, the routing algorithm can be qualified as minimally adaptive. However, under certain conditions, a packet has the possibility of being misrouted. If a packet makes a predetermined (and large enough) number of complete turns in one of the rings (without success in any profitable output port), it will be marked for misrouting. From that moment, this packet must use the first available output port (except consumer). Once the packet leaves the router, the mark is cleared and the packet will advance trying to follow minimal path again.

3. AVOIDING ANOMALIES

Next, we will show that the routing and flow control algorithms employed in our proposal avoid the appearance of any anomaly typical of interconnection networks: deadlock, livelock or starvation. In this section, due to space restrictions, we will provide an informal explanation about how Rotary Router successfully avoids all of these anomalies. A formal proof can be found in [35].

3.1 Deadlock and Livelock

The deadlock freedom can be shown in two related parts. First, we can easily guarantee packet movement inside any router. In effect, Bubble flow control [28] assures that input ports cannot exhaust all the buffering space in the internal rings of a router because to inject a new packet there must be room for at least another one. Thus, there will always be a hole inside the ring where a packet can advance.

Secondly, we can see that any packet can always move between routers. The Bubble flow control guarantees the existence of at least one hole in any ring, i.e. 2*N holes for an N-router network. Nevertheless, this level of occupation can never be reached because of the restricted injection control that is applied. In effect, the growth in network occupation comes only from the injection of new packets from processing elements, so in order to get a full network occupation the last packet must arrive from a host injection port. However, as was previously mentioned in Subsection 2.3, in order to inject a new packet into the network there must be room for at least three packets and therefore in one ring there will be two holes or 2*N+1 holes in the entire network. Even after reaching this extreme network configuration, the extra hole never stays in the same router because after a finite amount of time every packet in a neighbor router will be marked for misrouting, making the movement of a packet compulsory. Additionally, the hole never performs cyclic paths because the probability of injection of a packet from the counter-clockwise or clockwise neighbor ring is the same (remember that the output stage selects randomly the packet to be ejected), therefore traveling through every router. In a heavily loaded situation, the misrouting possibility is higher and consequently the packet movement is agitated, which helps to avoid cyclic movements for any traffic pattern. For this reason, in a finite amount of time the extra hole will circulate reaching all the network routers, which makes the network deadlock-free. Eventually, the misrouting means that neither packets repeat cyclic paths, which implies, statistically, that every packet will be able to reach its destination in a finite amount of time, avoiding network livelock. Note that the approach employed is similar to [16], although with a much higher feasibility.

It should be noted that the Rotary Router architecture delays the appearance of extremely congested situations and alleviates the HOL blocking effect. This has a positive effect on deadlock, because congestion usually accelerates buffer space exhaustion. For this reason, the probability of reaching a configuration in the network with the theoretical minimal number of holes (2*N+1) is negligible.

The deadlock avoidance method guarantees that the Rotary Router is deadlock-free for every network topology without requiring additional hardware resources. This characteristic could present multiple advantages for the router implementation, simplifying implementation cost (no virtual channels), applying it to any topology, or easing the inclusion of fault-tolerance mechanisms. Contrary to our proposal, deadlock avoidance is usually connected to the relation between network topology and routing algorithm. We believe that this fact is an enormous advantage over other proposals not only in the CMP context but in general.

3.2 Starvation

Two different traffic flows can suffer starvation: injection and/or in-transit. For the injection traffic, the resource access control mechanism is unfair, because packets at network injection ports need three holes to enter a router, while at transit ports two holes are enough. This situation can cause starvation. However, the probability of persistence in time of this situation is extremely low. In practical cases, such as the router employed in BlueGene/L [1], a tunable mechanism called in-transit-priority restriction (IPR) is used. But this mechanism is never deactivated, because it is shown that this is not a relevant issue from a practical point of view.

Nevertheless, starvation of in-transit traffic needs careful attention. Traffic with uniform destination patterns is unusual when working with real applications. Mostly we will find traffic patterns with restricted source-destination pairs. This kind of patterns causes an unbalanced use of router input ports, and can generate an unbalanced occupation of the buffering space, making the access to the rings more difficult for low activity input ports. This would have a negative impact on performance, due to the inefficient use of communication links (most of the packets belong to the same traffic flow). In order to avoid this situation the optimal solution would be to have a combination of packets, which makes a balanced usage of every router output port. However, this ideal solution is not viable from a practical point of view, because it is impossible to know in advance which output ports a packet will use when it is in the ring. We can approach the optimal solution by attempting to balance buffer occupation among every router input port; this means, the least used router input ports will be given ring injection preference over the most used. This mechanism works as follows; when the number of packets in a router from the same input port (excluding the injection one) grows over a certain limit, this input port modifies its flow control, increasing the required number of holes to inject a packet into the ring. This gives the rest of the input ports more chance to inject. Once the number of packets from the restricted port falls below the limit, the flow control modification is suppressed, i.e. the number of holes required to inject returns to its original value. It must be noted that the flow control modification never applies to all the transit ports simultaneously and therefore it does not affect the deadlock-avoidance mechanism.

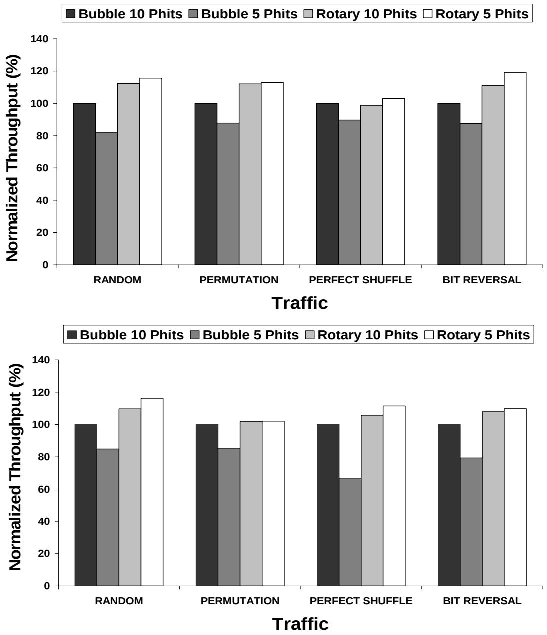

Figure 4. Maximum normalized throughput (up) 4x4 torus (down) 8x8 torus.

The added hardware complexity of this method could be negligible. For example, it could be implemented with five counters and four comparators. Four counters store the number of packets from each input port, and the fifth stores the sum of all the inputs. Each input port compares the value of its own counter with the value of the sum counter. Depending on the result of this comparison, the restrictive injection will be applied or not.

4. PERFORMANCE EVALUATION

The Rotary Router has been presented as a router architecture which takes advantage of CMP characteristics better than classic alternatives. Up to now, few proposals for CMP interconnection network exist, and those that do continue to use slightly modified off-chip router architectures [2][17][22]. Instead of using some previous CMP proposals we will compare our proposal with the Adaptive Bubble Router (ABR) [1][28][27] for two reasons: we will be able to see the potential advantage of the Rotary Router in the CMP context and explain why good off-chip proposals may not be a suitable choice for on-chip networks.

4.1 Synthetic Workloads

In a first phase of the evaluation, we will show the effect of link widening on performance for each router. To do so we use the interconnection networks simulator SICOSYS [29], which allows us to take into account most of the VLSI implementation details with high precision but with much lower computational effort than hardware-level simulators. For the Bubble Router a four-cycle pipeline has been employed (FIFO buffer, routing unit, arbitration, crossbar), while for the Rotary Router the pipeline is divided as follows; one cycle for the INPUT stage, one cycle for the OUTPUT stage and one cycle for each DFB. On an empty network the average number of DFBs traversed by each packet is two, which makes a total pipeline of four cycles. In Section 5.1, we will study if DFB can operate under such an assumption. Despite being topology agnostic, due to space restrictions, the networks considered in the simulations are two-dimensional torus with 16 (4x4) and 64 (8x8) nodes. Note that Adaptive Bubble router applicability is restricted to *k-ary n-cubes*. Synthetic traffic patterns with uniform and modal distribution have been used. Results have been obtained simulating 200,000 cycles for each traffic pattern analyzed after a warm-up phase of 20,000. These patterns are Random, Transpose Matrix, Perfect Shuffle and Bit Reversal.

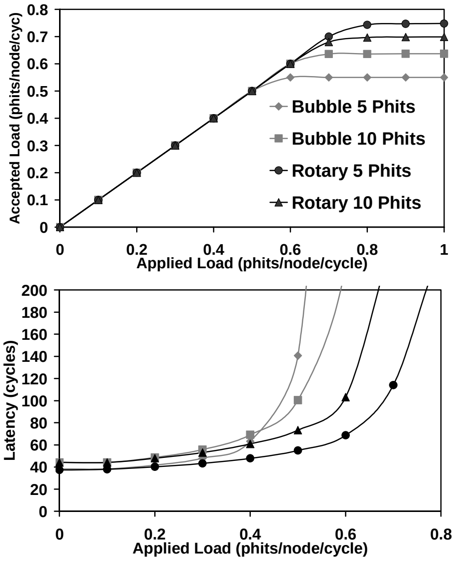

Figure 5. Throughput and latency for a 64-node 2D torus under random traffic.

4.1.1 Maximum Sustained Throughput

We analyzed network performance evolution for two different packet sizes, 10 and 5 phits (this can be interpreted as doubling link width, from 64 to 128 bits with packets of 80 bytes). Total buffering storage space is kept constant in both routers, making comparison as fair as possible. The total buffer capacity in both routers is 5.6KB, assuming the aforementioned phit sizes. By doubling link width (5 phits packet), phit size also doubles, so in order to keep router area constant, the number of phits a router can hold must be divided by two. Values for every traffic pattern and packet size are presented in Figure 4. The results have been normalized taking as reference the results obtained for the ABR router with 10-phit packets. It can be seen that in every case the Rotary Router obtains better results. Working with 10 phit packets, results show that HOL blocking avoidance and buffer space utilization helps the Rotary Router to accept higher load levels before saturation. When packet size is reduced from ten to five phits, we observe a completely opposite effect on performance for both routers. In a structure with a centralized arbiter, such as ABR, a higher number of arbitrations produces more collisions, causing resource underutilization and the early appearance of network congestion. On the other hand, in the Rotary Router a smaller packet size implies higher frequency of output requests, increasing link utilization and therefore performance. As can be seen in Figure 4, the Rotary Router outperforms the classic router by up to 42% in a 4x4 torus (in Bit-reversal traffic) and 58% in an 8x8 torus (in Perfect-shuffle traffic).

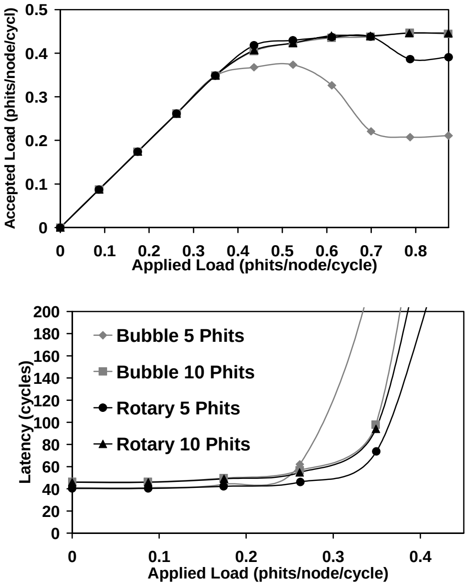

Figure 6. Throughput and latency for a 64-node 2D torus under transpose matrix traffic.

4.1.2 Variable Load Results

To complete this quantitative comparison, we need to consider network behavior under variable network loads. Throughput and latency measurements were made for every traffic pattern analyzed, with load levels up to one phit/node/cycle.

In Figure 5 and Figure 6, we can see throughput and latency values obtained for an 8x8 torus under Random and Transpose Matrix traffic. Here we can observe in more detail the double effect of packet size reduction on input buffered router performance. First, it causes the earlier appearance of the saturation point, and second, after saturation the ABR suffers an important degradation of its throughput, as shown in Figure 6. With non-uniform traffic patterns, the effect of contention and HOL blocking is much greater. These types of traffic highly stress some communication links, routing most of the packets through them. The performance of routers connected to these links degrades extremely fast, extending the congestion to the rest of the network even at medium traffic load. The HOL blocking effect makes persistent traffic flows obstruct the remaining traffic flows. Once again, architectural structure and injection control methods present in the Rotary Router help to alleviate all these problems, not only increasing maximum achievable throughput but stabilizing network performance beyond saturation point. For the rest of the traffic patterns analyzed we observe similar behavior.

4.2 Real Workloads

In this section, we will show the benefits of the Rotary Router on the whole system performance. For this purpose we will use the complete system simulator Simics [19], extended with the GEMS timing infrastructure [20]. GEMS provides a detailed model of the memory system and a state-of-the-art detailed processor model. SICOSYS has been integrated into the simulator GEMS, replacing its original network simulator. The simulated system is a 16-processor CMP with shared SNUCA L2 based on [12]. The main simulation parameters are shown in Table 1.

Table 1. Main simulation parameters.

Number of Cores	16
Window Size / outstanding memory requests per CPU	256 / 16
Issue Width	4
L1 I/D cache	Private, 32KB, 2-way, 64Bytes block, 1-cycle
Direct branch predictor	4KB YAGS
Indirect Brach Predictor	256 entries (cascaded)
L2 cache	SNUCA, 16x16 banks, 4 per router
L2 cache bank	128KB, 16-way, 3-cycles, Pseudo LRU, 64 Bytes block
Main Memory	4GB, 260 cycles, 320 GB/s
Command size	16 bytes
Network Topology	8x8 torus
Network Link	128 bits / 1 cycle latency

The applications considered in this study are two transactional and two scientific workloads. The server workload we will use is a Static Serving Web server benchmark based on SURGE running on top of Apache web server (Surge), and SPECjbb2000 (Java). The numerical workloads used are LU and FT from NAS Parallel Benchmark, using the OpenMP implementation, Version 3.1. In all cases, a variable number of runs are performed with pseudo-random perturbation in order to estimate workload variability.

In Figure 7, results with expected average normalized execution time are provided. The confidence interval is 95%. As we can see, the conclusions obtained with synthetic traffic remain largely unchanged. The Rotary Router outperforms classic input buffered router by up to 20%.

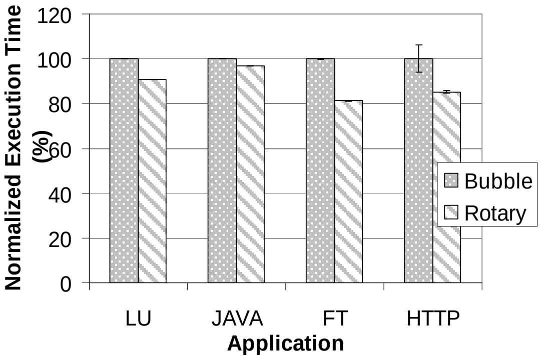

Figure 7. Normalized execution time.

5. IMPLEMENTATION VIABILITY

Ideally, any computer architecture proposal must be wrapped-up with an implementation feasibility study. Moreover, CMP requirements such as power consumption or hardware complexity become first order design issues. In the Rotary Router there is an additional reason to carry out such a study, related to router structure; packets have a special way of moving, they do not stop turning until they are able to find an available output port. For this cause, a packet makes multiple buffer reads and writes at every router. At first glance, this seems to have a very negative impact on router power consumption, but we will prove through simulation results that the real behavior is different. In addition, implementation cost of the building block should be made clear.

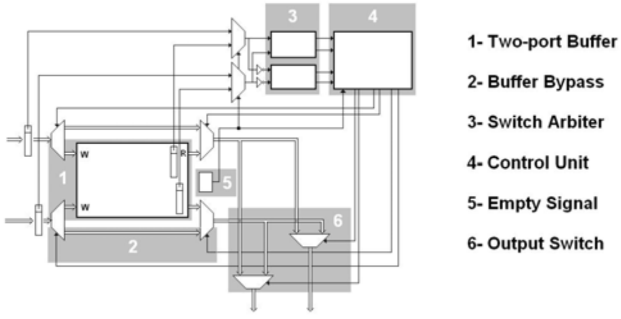

Figure 8. Structure of the DFB.

5.1 Delay and Area

Multiport buffer design is a hard issue in terms of hardware complexity and delay. As the number of ports increases, the implementation cost grows drastically. For this reason, we must prove that the router complexity and the hardware delay present suitable values. INPUT and OUTPUT stages are simple; the INPUT stage only performs link synchronization and creates packet routing information, and the OUTPUT stage consists of two FIFO buffers and a multiplexer which controls link access. These two stages represent a small portion of router area and are significantly less critical than the DFB. In the Rotary Router the DFB requires only two ports, regardless of network topology, so the complexity associated with the number of ports is still affordable. Next, we provide an approach to the DFB physical design, focusing on the delay of a packet going through this stage. Figure 8 shows a possible implementation for this component. As can be seen, this element is made up of a FIFO buffer with two read ports and two write ports, a control unit and two multiplexers which act as a small crossbar. The hardware implementation of the buffer can be based on the structure presented in [10]. The original implementation was designed to work as a whole router (holding a large number of packets), and presents a three-stage pipeline. In the Rotary Router the buffering space of each DFB is considerably lower, which highly reduces memory access time. For this reason, the writes or reads to/from that buffer do not limit clock period. In order to minimize DFB delay, two demultiplexers and two multiplexers have been added to the 2-port buffer. This allows a packet to bypass the buffer and access the output multiplexer directly, avoiding buffer stages when it is empty. The DFB operates as follows; when a new packet arrives at an input port, it activates the control unit. If the control unit detects that the buffer is empty, the packet is chosen for arbitration before being written to the buffer. If the packet receives a grant for an output, it bypasses the buffer through a demultiplexer and a multiplexer moving directly to the output. If the packet does not gain access to an output, it is written in the buffer. When the buffer is not empty, the packets chosen for arbitration are those in the head of the buffer, and any packet arriving at an input port is written in the buffer. A round-robin policy is employed to arbitrate each read port access, in order to ensure fairness.

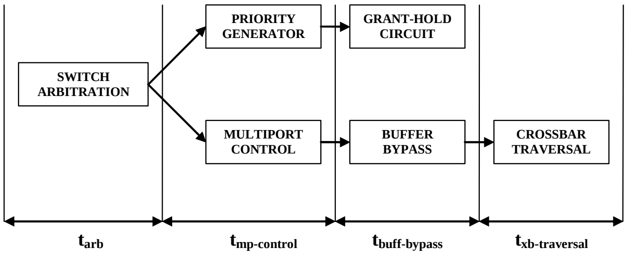

Figure 9. Atomic modules of the DFB.

The objective of the DFB design is to obtain a delay of only one clock cycle when the buffer is empty. This will achieve under low load conditions a router pipeline of 4 cycles on average, equaling it to the Bubble Router pipeline. In order to verify that the DFB fulfils delay requirements, a delay study based on the model presented in [24] was carried out. This model employs the Logical Effort theory [31], in which the delay of a circuit is calculated as the sum of two components known as effort delay (Teff) and parasitic delay (Tpar). The former is the effort required to perform a logic function plus the effort required to drive an electrical load. The parasitic delay is the intrinsic delay of a gate due to its own internal capacitance. Both terms allow the delay to be expressed in a technology independent unit called FO4. In Figure 9, we can see the atomic modules of the DFB and their dependencies. In an effort to obtain accurate delay values, each module has been designed at gate level.

The round-robin arbiter used in the DFB is similar to the one explained in [6]. As the buffer has only two outputs, we will use a two-bit arbiter. The delay value for switch arbitration can be calculated in terms of the number of requests needed for arbitration, following the procedure explained in [24]:

$$t_{arb} = T_{par_{arb}}(n) + T_{eff_{arb}}(n) = 0.6 \cdot n + 1.6 = 2.8FO4_{(n=2)}$$

The priority generation (for the next round of arbitration) and the grant-hold circuit can be computed in parallel with the next DFB modules, being outside the critical path. The DFB Control module controls the path used by a packet to traverse the dual-port buffer and generates control signals addressed to neighboring blocks. It takes as inputs the two grant signals generated by the previous module and control signals from the next blocks. Its functionality is implemented with a two-gate level circuit that generates the signals to control every multiplexer and demultiplexer. Following the aforementioned procedure, the delay for this module in terms of FO4 is:

$$t_{mp-control} = T_{eff_{mp-control}} + T_{par_{mp\_control}} = 1.46FO4$$

The last two modules in the critical path take into account the delay produced by packet movement from an input to an output of the DFB. First, the phit must go through the multiplexer and the demultiplexer which form a buffer bypass. Finally, the last multiplexers guide the phit to an output of the DFB. Their corresponding delays are:

$$t_{buff-bypass} = 9.4FO4 \qquad t_{xb-traversal} = 5.7FO4$$

The model returns a delay of 19.36FO4 for the Rotary Router DFB. In order to compare this result, we have repeated the delay study for the arbitration stage (the stage with the longest delay) of the Bubble Router. Using the hardware description presented in [27] we can infer a delay value of 23.56FO4 for arbitration and crossbar traversal. Both delay values confirm that Rotary Router DFB traversal can be implemented in one cycle and consequently corroborate that the Rotary pipeline used for the performance evaluation is realistic.

Although estimation of area for this possible implementation is not provided, it would seem straightforward to conclude that the area will be dominated by buffering space. In order to manage only two inputs and outputs the control implementation cost will be negligible. Considering this along with the fact that the conventional router has the same total storage capacity as the Rotary Router, the whole area required should be advantageous to our proposal because Adaptive Router will require larger control logic. According to [2] in a conventional router, the area devoted to control could be significant.

5.2 Power

Finally, an evaluation of network power consumption was carried out in order to assess the viability of the proposal. We use a simulation tool called Orion [33]. Orion is a power simulator based on component capacity models [34]. These models are architectural-level parameterized, which allows us to explore the effect of different network architectures on power consumption. Orion is built to be connected to other network simulators, so in our experiments it has been used together with SICOSYS. Leakage current or clock signal distribution power consumption have not been taken into account; these values can be considered similar for both network architectures. Power consumption simulations have been carried out for 0.10μm technology and 1GHz frequency. Models for the Rotary Router components have been taken from the models present in the simulator.

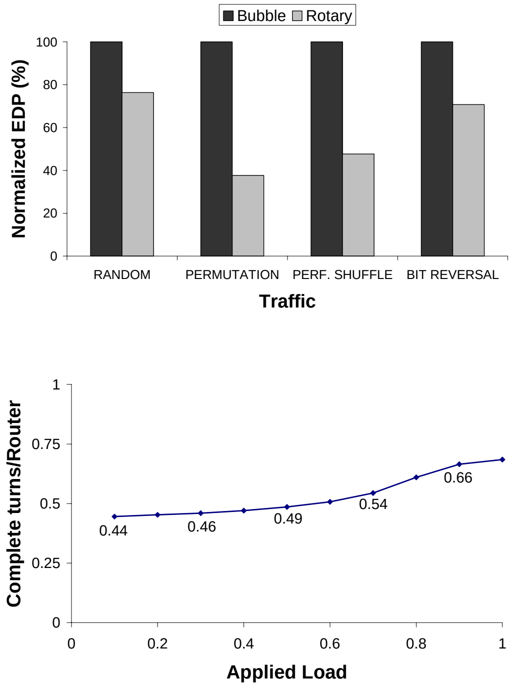

Figure 10. Power consumption for an 8x8 torus network and Mobility of the packets.

As described before, packet movement has a negative effect on network power consumption, but the Rotary Router benefits lead to improvements in a program's overall Energy-Delay product (EDP) since they reduce the program execution time [7]. The energy-delay product is calculated as the product of a workload execution time and the energy expended by the workload. We approach the workload simulation as the reception of a fixed number of packets, high enough to ensure that those packets injected into the empty network at start-up do not influence the reported results. In the simulations performed the number of packets chosen as a workload was 40,000 (200,000 phits). Figure 10 shows that for every traffic pattern the Rotary Router presents noticeably better Energy-Delay results than ABR.

The reasons behind these results are that the performance improvements are able to hide power consumption increases and the Rotary Router power consumption is not as high as might be thought. As shown in Figure 10, the average number of complete turns by a packet in a low loaded router is 0.44. This means that most of the packets find an available output on their first turn. Increasing load level returns higher mobility values, but never as high as expected. In fact, at the maximum load level the average mobility is still below one, which means that on average packets never start a second turn in the router rings. The multiple flow controls applied in different network places avoid network congestion, easing packet advance between routers and reducing the number of complete turns a packet makes. Moreover, in classic routers a packet stored in a transit buffer never moves, but centralized arbitration causes a power consumption which is far from negligible.

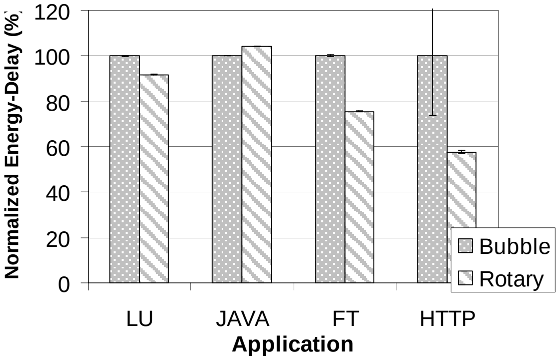

Figure 11. EDP of the networks for real applications.

Finally, Figure 11 shows the EDP for the real workloads. As GEMS does not take into account power consumption for the simulated system, the measure provided only includes the network energy. Although the power consumption of the network is significantly lower than the rest of the system, the reduction achieved by the Rotary Routers is clear in most cases. The advantage is smaller than in synthetic traffic because the average applied loads over the network by real applications are generally lower. In fact, in the JAVA application the results are unfavorable for the Rotary Router because the reduction in execution time does not compensate the higher power requirements for our proposal. However, it must be noted that even this application reduces its execution time (see Figure 7) which implies a potential EDP reduction in the rest of the system that GEMS does not report.

6. RELATED WORK

Although the existing bibliography related to CMP is profuse, studies focusing on the interconnection networks are scarce. Among the works where this subsystem was considered, we can cite [17]. In this work, a thorough study of the impact of the interconnection system cost and performance was carried out. However, it only analyzed centralized structures, but mentioned point-to-point interconnects study as a future line.

One of the seminal papers where point-to-point packet switched networks are suggested for systems on-chip was [5]. The authors argue for this type of interconnection and show its growing advantages over ad-hoc global wiring structures when the total budget of transistors is high. First approximations about the requirements for on-chip systems were provided and interconnection network design issues were discussed. However, only traditional router architectures were considered. In recent works, such as [2], the same router architecture has been employed.

Recently, new proposals for the architecture for network interconnect in on-chip systems have been presented. In [22] the authors introduced a router with 1-cycle pass-through delay in non-contended conditions. The router uses a conventional input-buffered and centralized arbitration scheme. Therefore, no special mechanism was provided to optimize resource usage and delay congestion. In [14][13] a partitioned organization for the crossbar was proposed for on-chip networks. That architecture requires look-ahead routing. In [18] the authors tried to overcome the performance degradation of the route decision time proposing a mechanism to increase the throughput of an adaptive router in network-on-chip, through the use of a faster clock during the service of the body phit.

Finally, other works focus on complementary areas to router architecture. For example, in [26][2] the authors analyze how different topologies behave in the network-on-chip context. In the later, the authors demonstrate that replicating networks in an on-chip environment increases power dissipation but can improve performance and energy efficiency.

7. CONCLUSIONS

We have presented a novel router architecture, especially targeting CMP systems. The router utilizes a decentralized and scalable structure based on rings. Making packets circulate in the rings has important benefits: it eliminates HOL blocking; it improves the performance when the size of packets decreases and it provides a deadlock avoidance mechanism that is topology agnostic without using virtual channels. Raw performance results from a wide range of loads demonstrate its noticeable advantage over very competitive conventional router architectures.

The feasibility analysis shows how the proposal has reasonable cost in terms of area and power consumption for a possible implementation. Although the analysis is based in approximate models, in the achieved results our proposal has a considerable advantage over off-chip router alternatives.

The work presented has taken into account only first order details of CMP systems. It is possible to adjust some subtle details in the network in order to maximize coherence protocol requirements. In addition, it is possible to extend the applicability of the idea to other fields. The topology independence of the deadlock avoidance mechanism could be useful in some off-chip networks.

8. ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers as well as Cruz Izu for their constructive comments. This work has been supported by the Spanish Science and Education Ministry under grant TIN2004-07440-C01-01.

9. REFERENCES

[1] N.R. Adiga, et al., “An Overview of the BlueGene/L Supercomputer”, Supercomputing 2002.

[2] J. Balfour, W.Dally, “Design Tradeoffs for Tiled CMP On-Chip Networks”, International Conference on Supercomputing (ICS) 2006.

[3] S.Borkar, et al. “Platform 2015: Intel Platform and Evolution for the Next Decade”, Technology@Intel Magazine, March 2005.

[4] D. Burger, S. Keckler, K. McKinley, M. Dahlin, L. John, C. Lin, C. Moore, J. Burrill, R. McDonald, W. Yoder “Scaling to the end of Silicon with EDGE Architectures” IEEE Computer. Volume 37, No 7, pp.44-55, July 2004.

[5] W. Dally, B. Towles, “Route Packets, Not Wires: On-Chip Interconnection Networks”, Design Automation Conference (DAC) 2001.

[6] W. Dally, B. Towles, “Principles and Practices of Interconnection Networks”. Morgan Kaufmann, 2004.

[7] R. Gonzalez, M. Horowitz, "Energy Dissipation In General Purpose Microprocessors", IEEE Journal of Solid-State Circuits, Vol. 31, No. 9, pp. 1277-1284, September 1996.

[8] P. Gratz, C. Kim, R. McDonald, S. W. Keckler, D. Burger, "Implementation and Evaluation of On-Chip Network Architectures", International Conference on Computer Design (ICCD), 2006.

[9] M. J. Karol, M. G. Hluchyj, S. P. Morgan, "Input versus Output queuing on a space-division packet switch", IEEE Trans. Communication., Vol. 35, no. 12, pp. 1347-1356, December 1987.

[10] M. Katevenis, P. Vatsolaki, A. Efthymiou. "Pipelined Memory Shared Buffer for VLSI Switches", ACM SIGCOMM 1995.

[11] P. Kermani, L. Kleinrock, "Virtual Cut-Through: A New Computer Communication Switching Technique". Computer Networks, Vol. 3, pp. 267-286, September 1979.

[12] C. Kim, D. Burger, S. W. Keckler. "An Adaptive, Non-Uniform Cache Structure for Wire-Dominated On-Chip Caches", International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2002.

[13] J. Kim, C. Nicopoulos, D. Park, V. Narayanan, M. S. Yousif, C. R. Das, "A Gracefully Degrading and Energy-Efficient Modular Router Architecture for On-Chip Networks", International Symposium on Computer Architecture (ISCA), 2006.

[14] J. Kim, D. Park, T. Theocharides, N. Vijaykrishnan, C. R. Das. "A low latency router supporting adaptivity for on-chip interconnects". Design Automation Conference (DAC) 2005.

[15] P. Kongetira, K. Aingaran, K. Olukotun, "Niagara: A 32-way Multithreaded SPARC Processor", IEEE Micro. Vol. 25, No. 2, pp. 21-29, March/April 2005 .

[16] S. Konstantinidou, L. Snyder, "The Chaos Router", IEEE Trans. Computers, Vol. 43, No. 12, pp. 1386-1397, December 1994.

[17] R. Kumar, V. Zyuban, D. Tullsen, "Interconnections in Multi-Core Architectures: Understanding Mechanisms, Overheads and Scaling", International Symposium on Computer Architecture (ISCA), 2005.

[18] S. E. Lee, N. Bagherzadeh, "Increasing the Throughput of an Adaptive Router in Network-on-Chip (NoC)" CODES+ISSS'06, 2006.

[19] P. S. Magnusson, M. Christensson, J. Eskilson, D. Forsgren, F. Larsson, A. Moestedt, B. Werner, "Simics: A Full System Simulation Platform". Computer, Vol. 35, No.2, pp. 50-58, February 2002.

[20] M. Martin, D. Sorin, B. Beckmann, M. Marty, M. Xu, A. Alameldeen, K. Moore, M. Hill, D. Wood, "Multifacet's General Execution-driven Multiprocessor Simulator (GEMS) Toolset", SIGARCH Comput. Archit. News, Vol.33, No.4, pp.92–99, November 2005.

[21] S. Mukherjee, P. Bannon, S. Lang, A. Spink, D. Webb, "The Alpha 21364 Network Architecture", IEEE Micro, vol. 22, no. 1, pp 26-35, Jan-Feb 2002.

[22] R. Mullins, A. West, S. Moore "Low-Latency Virtual-Channel Routers for On-Chip Networks", International Symposium on Computer Architecture (ISCA), 2004.

[23] K. Olukotun, L. Hammond, "The future of Microprocessors" ACM Queue, Vol. 3, No. 7, September 2005.

[24] L. Peh, W. Dally, "A Delay Model and Speculative Architecture for Pipelined Routers", International Symposium on High-Performance Computer Architecture (HPCA) 2001.

[25] H.Hofstee, "Power Efficient Processor Architecture and The Cell Processor", International Symposium on High-Performance Computer Architecture (HPCA), 2005

[26] P.Pande, C.Grecu, M. Jones, A.Ivanov, R.A. Saleh, "Performance Evaluation and Design Trade-Offs for Network-on-Chip Interconnect Architectures", IEEE Trans. Computers, Vol. 54, No. 8, pp.1025-1040, February 2005.

[27] V. Puente, C. Izu, R. Beivide, J.A. Gregorio, F. Vallejo, J.M. Prellezo, "The Adaptive Bubble Router", Journal of Parallel and Distributed Computing, Vol. 61, No. 9, September 2001.

[28] V. Puente, J.A. Gregorio, J. M. Prellezo, R.Beivide, J. Duato, C. Izu, "Adaptive Bubble Router: a Design to Improve Performance in Torus Networks", International Conference of Parallel Processing (ICPP) 1999.

[29] V.Puente, J.A. Gregorio, R.Beivide, "SICOSYS: An Integrated Framework for studying Interconnection Network in Multiprocessor Systems", Euromicro Workshop on Parallel and Distributed Processing, 2002.

[30] K. Sankaralingam et. al. "Distributed Microarchitectural Protocols in the TRIPS Prototype Processor", International Symposium on Microarchitecture (MICRO), 2006.

[31] I. Sutherland, R. F. Sproull, D. Harris, "Logical Effort: Designing Fast CMOS Circuits", Morgan Kaufmann, 1999.

[32] Y. Tamir and G.L. Frazier. "Dynamically-Allocated Multi-Queue Buffers for VLSI Communication Switches" IEEE Trans. on Computers, Vol.41, No. 6, pp 725-737, June 1992.

[33] H. Wang, X. Zhu, L. Peh, S. Malik, "Orion: A Power-Performance Simulator for Interconnection Networks.", International Symposium on Microarchitecture (MICRO), 2002.

[34] S. J. E. Wilton and N. P. Jouppi. "CACTI: An Enhanced Cache Access and Cycle Time Model", IEEE Journal of Solid-State Circuits, May 1996, pp 677-688.

[35] http://www.atc.unican.es/investigacion/publicaciones/publicaciones_files/publ_109.pdf