Introducing the superGT network-on-chip: SuperGT QoS: more than just GT

Citation for published version (APA):

Document status and date:
Published: 01/01/2007

Publisher Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
Introducing the SuperGT Network-on-Chip

SuperGT QoS: more than just GT

Théodore Marescaux
IMEC vzw,
Kapeldreef 75,
3001 Leuven, Belgium
Theodore.Marescaux@imec.be

Henk Corporaal
TU Eindhoven,
Den Dolech 2,
5612 AZ Eindhoven, The Netherlands
H.Corporaal@tue.nl

ABSTRACT
Predictability of multi-processor systems-on-chip communication is critical and needs to be addressed by providing the right mix of soft and hard real-time guarantees. To this end, state-of-the-art packet-switched networks-on-chip (NoC) provide different levels of quality-of-service (QoS) such as best effort (BE) and guaranteed throughput (GT). Unfortunately, GT resources have to be reserved for the worst-case, resulting in over-allocated resources.

We introduce the SuperGT NoC, a packet-switched NoC that, besides BE and GT, supports a new SuperGT QoS. A SuperGT connection combines guaranteed and non guaranteed traffic while maintaining in-order packet delivery. Time-slots are allocated to provide guarantees and extra BE resources are claimed by injecting data during free slots.

Simulation results demonstrate the advantages of SuperGT over GT. Synthesis results of the SuperGT virtual channel manager show that the SuperGT router is an inexpensive enhancement to state-of-the-art packet-switched NoCs.

Categories and Subject Descriptors
C.1.2 [Processor Architectures]: Multiprocessors

General Terms
Design, Performance

Keywords
MP-SoC, NoC, QoS

1. INTRODUCTION

Many of the applications targeted at multi-processor systems-on-chip (MP-SoC), such as multi-media applications, display a mixture of predictable streaming data and of less predictable traffic such as cache misses or operating-system (OS) communication. State-of-the-art NoCs, such as Æthereal, provide connection oriented guaranteed services, such as guaranteed throughput (GT), with TDMA sharing of the bandwidth and use BE services to fill in the bandwidth reserved but not claimed by the GT QoS [3, 4].

A connection is defined as the ensemble of a request path and a response path connecting the network interfaces of two communicating IP-blocks over the NoC. A connection is assigned a particular QoS class and is typically established for a duration much longer than a packet round-trip time.

Consider a system with a GT connection between an L1 cache and an L2 memory. As cache misses are not completely predictable it is necessary to largely over-allocate bandwidth, possibly preventing other GT connections from being allocated. Instead, it would be desirable to reserve enough bandwidth for the predictable traffic and absorb peaks of less predictable traffic with additional non guaranteed bandwidth. In Figure 1, the offered traffic load has peaks higher than the amount of bandwidth reserved. With a regular GT connection, throughput saturates hence increasing overall application latency.

The main contribution of this paper is SuperGT, a new QoS class that allows the NoC to follow much closer the offered traffic load. The traffic source is still offered guaranteed throughput for the amount reserved. In addition, traffic peaks over the reservations are transmitted in non-guaranteed (best-effort) manner. Advanced virtual-channel management is performed to guarantee in-order-packet delivery over a SuperGT connection.

This paper presents the SuperGT QoS and introduces the routers and network interfaces of the SuperGT NoC. This NoC supports three QoS classes: BE, GT and SuperGT.

Figure 1: SuperGT follows offered load above reservations, GT saturates at reserved amount.

The rest of the paper is organized as follows. Section 2 introduces background information and related work. Section 3 details the concept and architecture of the SuperGT NoC. Sections 4 and 5 present respectively performance results of the SuperGT NoC and synthesis results of its virtual channel manager in 90nm standard cell technology. Finally, Section 6 concludes.

2. BACKGROUND

This section introduces the related work on QoS for NoC and discusses packet ordering.
2.1 Related work: QoS on NoCs

Bolotin and al. propose QNoC [2], an NoC that supports four levels of QoS (signalization, real-time, read/write and block-transfer) by employing pure packet-switching with four prioritized virtual channels. All communication properties have to be estimated at design-time in simulation to provide the expected QoS levels, limiting the scope to application-specific NoCs with static configurations.

The MANGO NoC [1] is another asynchronous NoC that mixes a BE router and a guaranteed services (GS) router to provide both best-effort and real-time guarantees. Connection oriented GS are provided by reserving virtual circuits through the NoC. End-to-end flow control is inherent. BE services are connection-less and can be used to setup GS connections at run-time. MANGO has a work-conserving [9] scheduling discipline: the offered application load on a GS connection is immediately served.

The techniques to provide BE and GT that are based on Time Division Multiple Access (TDMA) techniques for synchronous NoCs, such as in the Nostrum NoC [5] and the Æthereal NoC [3, 4, 7] are the closest to the one applied here. They all have non-work-conserving scheduling disciplines.

The Sonics MXNoC performs (at the edge of the network) priority promotion and demotion of packets, based on a token-bucket scheme. Arbitration inside the network fabric, is based on epochs (a form of TDMA compatible with asynchronous networks) [8]. Because of the proposed priority switching mechanism, in-order packet delivery for streaming connections cannot be guaranteed.

The SuperGT NoC contribution can be best understood as a potential evolution of the GS-BE central programming architecture of Æthereal (ÆGS−BE). ÆGS−BE is a pure packet-switched NoC, where both BE and GT1 packets have headers and where time-slot tables (Figure 6) are only situated in the network interfaces [4].

2.2 Packet Delivery Order

Packet re-ordering at the level of the network interface is impractical as it requires expensive re-ordering buffers and greatly increases latency. Many NoCs try thus to avoid out-of-order packet delivery. For instance, the Æthereal ÆGS−BE guarantees in-order delivery for both GT and BE services. It is then possible to construct a service similar to the proposed SuperGT by using a mix of GT and BE packets on the same connection2 between two IP-blocks, such as an L1 cache and the L2? GT packets would be used to provide the reserved bandwidth and BE packets to absorb the offered load that is superior to reservations. However, as this section demonstrates, such a scheme would lead to out-of-order packet delivery.

Let us assume a packet-switched router with two virtual channels, similar to ÆGS−BE. The first virtual channel (VC\textsubscript{GT}) is assigned the highest priority and is used for hard real-time GT traffic, whereas the second one (VC\textsubscript{BE}) of lower priority is used for BE traffic. Assuming that the injection-time of flits (flits, or flow-control units, are sub-units of packets) on the VC\textsubscript{GT} respects a congestion-free allocation schedule [3, 4], we can deduce two properties:

1. Any two flits injected on VC\textsubscript{GT} never collide, as a consequence of the congestion-free allocation.

2. All flits on VC\textsubscript{GT} have higher priority than flits injected on VC\textsubscript{BE}. In case of a conflict between a GT and a BE flit, the GT flit is guaranteed to have access to the resources, the BE flit is buffered locally.

The direct consequence of these properties is that GT flits are guaranteed to never be delayed on their path on the NoC, thus ensuring hard real-time guarantees on their delivery time. Note that in this configuration of packet-switched NoCs no time-slot tables are required at the routers, but only at the GT network interface (NI) [6, 3].

Also, GT flits may not be buffered on the routers. The end consumer of the GT flits should therefore be fast enough to guarantee that all GT data is taken out of the NoC timely. This condition is very difficult to guarantee in realistic systems, hence techniques such as end-to-end credit-based flow-control techniques are used to inform the producer NI of the level of occupation of the FIFOs inside the consumer NI. This allows the producer NI to control the amount of data it may inject in the NoC and thus guarantee that the NoC never gets congested on VC\textsubscript{GT}.

On a router, in the absence of GT packets on VC\textsubscript{GT}, the bandwidth can be reused for BE traffic (of other connections) on VC\textsubscript{BE}. This happens automatically when inter-leaving virtual channels on classic packet-switched networks. Given deterministic routing, packet-order delivery is guaranteed.

Assume now the network interface of the aforementioned cache can inject BE packets as well as GT packets (on the same connection) destined to the same L2 remote memory. For instance, the NI sends flits 2, 3, 4, 5, 6 on this hypothetical GT+BE connection (Figure 2(a)). The allocation of time-slots ensures that flits 3, 5 have GT priority and go on VC\textsubscript{GT} on the underlying router, whereas flits 2, 4, 6 are injected on VC\textsubscript{BE}. Assume, the low priority 1-flit packet 2 has lost contention to a low-priority packet on another connection. Packet 2 is buffered locally long enough for guaranteed 1-flit packet 3 to take over, disrupting packet order. Special attention is thus required to provide in-order packet-delivery on such a connection with reserved and non-reserved resources.

![Figure 2: (a) packet order inversion on $\mathcal{E}_{GS-BE}$ for a mixed BE/GT QoS connection, (b) QoS levels in SuperGT NoC.](image)
3. SuperGT NoC

The SuperGT NoC is a packet-switched NoC that provides 3 QoS classes: BE, GT and SuperGT. Packets of all three QoS classes have headers and are guaranteed to be delivered in-order. Priority and flit-type information are transmitted out-of-band.

The SuperGT router is a packet-switched router with 2 virtual channels with priorities, featuring a SuperGT virtual channel, VC_{sGT}, next to a high-priority-escape virtual channel, VC_{escHi}. Similarly to the \(E_{GS-BE}\) architecture, GT packets are assigned a high-priority (Hi) and are sent on VC_{escHi}, whereas BE packets have a low priority (Lo) and are sent on VC_{sGT}.

SuperGT packets are special. Not only can the network interface assign them either of the priority levels (Hi and Lo), but packet priority can be promoted or demoted by routers along the path. Moreover, whereas VC_{escHi} is only used by SuperGT packets with Hi priority, VC_{sGT} accommodates for both low and high priority traffic (Figure 2(b)). Section 3.1 details the SuperGT virtual channel management in a router.

On a given SuperGT connection, flits are injected with a high priority (Hi) during the allocated time-slots and with a low priority (Lo) during free time-slots. A contention-free central time-slot allocation is performed for GT traffic together with the Hi part of the SuperGT traffic.

3.1 SuperGT Virtual Channel Management

The SuperGT mechanism relies on a particular management of the VC_{escHi} and VC_{sGT} virtual channels and does not depend on the type of queuing (input queuing, virtual output queuing, etc.) nor on the switching method. For simplicity, this paper describes a wormhole-switched, 2x2 input-queued SuperGT router (Figure 3). Every input port on the SuperGT router has a SuperGT virtual channel manager (sgtVCM). Its role is to buffer flits and keep in-order packet delivery for all three QoS classes.

![Figure 3: 2x2 input-buffered SuperGT Router and its two sgtVCM.](image)

To provide in-order delivery on a SuperGT connection the sgtVCM may have to demote guaranteed Hi packets to Lo and simultaneously promote best-effort Lo packets to real-time Hi priority. Priority switching is only performed on VC_{sGT} when flits from a Lo priority packet are buffered on the VC_{sGT} FIFO and a Hi priority packet belonging to the same connection enters the sgtVCM. The incoming flit is pushed at the tail of the VC_{sGT} FIFO and its priority token (1-bit out-of-band field) is transferred to the flit at the head. This mechanism conserves both in-order packet delivery and the number of priority tokens on a given connection, thus its real-time properties.

Figure 4 details the priority switching process in time. Assume a Lo packet \{1, 2, 3\} has lost contention on router \(R + 2\) (1 is a Head flit, 2 is a Body flit and 3 is a Tail flit). The routers in the example implement wormhole switching and the VC_{sGT} FIFO can buffer two Lo flits. Flits \{1, 2\} are therefore blocked at router \(R + 2\) and flit 3 is blocked at router \(R + 1\). A Hi packet \{A, B, C\} enters router \(R\) at time \(T\) on the same connection as the previous packet \{1, 2, 3\}. At time 2\(T\) on router \(R + 1\) the first priority switching occurs between the Head flit A and the Tail flit 3. At time-slot 3\(T\) two priority switchings occur, the first one at router \(R + 1\) between the Body flit B and the Head flit A and the second one at router \(R + 2\) between the Tail flit 3 and the Head flit B. The priority switching is pipelined, so that at router \(R + 1\) the Head flit 1 is leaving during time-slot 3\(T\) and the Tail flit 3 is leaving at time-slot 5\(T\). The packet \{A, B, C\} has been completely demoted to Lo and is buffered on router \(R + 1\) and \(R + 2\) (in place of the packet \{1, 2, 3\}) until the contention on Lo on router \(R + 2\) is won (or until a subsequent packet with Hi priority on the same connection transfers its priority). A minimum amount of bandwidth corresponding to the three allocated slots is guaranteed.

![Figure 4: Priority switching on VC_{sGT} for packets on the same connection.](image)
packet \{1,2,3\}. The reason to this constraint is simple: if the Lo packet \{Z\} from another connection was allowed to be buffered on router \(R+1\) its VC_{sGT} FIFO would contain the following flits: \{Z,3\}. When the Hi packet \{A,B,C\} would enter router \(R+1\) on the same connection as flit 3, the priority switching mechanism would not make sense for flit \{Z\} as it is on a different connection\(^3\).

Since we have seen that priority switching only makes use of VC_{sGT} one may wonder what the purpose of VC_{e.si}Hi is. Consider the situation of Figure 5 where Lo flits of a given SuperGT (or BE) connection \{(1,2,3)\} are being buffered on VC_{sGT} and Hi flits from a packet on a different SuperGT connection \{(A,B,C)\} enter the sgtVCM. As the connections are different, re-ordering does not make sense, so the Hi packet must take over the Lo flits buffered on VC_{sGT}. In this case only the Hi flits are pushed on the other virtual channel of the sgtVCM, the escape channel VC_{e.si}Hi, and keep their priority tokens. Packet \{1,2,3\} stays in place on VC_{sGT}, buffered across routers \(R+1\) and \(R+2\) and packet \{A,B,C\} uses the escape virtual channel VC_{e.si}Hi to exit router \(R+2\) in the reserved time. Note that packet \{A,B,C\} has totally left router \(R+2\) at time-slot \(6T\), exactly at the same time as packet \{1,2,3\} in the previous example, where priority switching was needed.

The same sgtVCM supports also BE and GT connections. Only SuperGT connections mix packets of Lo and Hi priorities and thus may require priority switching. As BE and GT connections always keep the same priority level, the priority switching mechanism of the sgtVCM is simply never triggered for these type of connections.

The sgtVCM contains a simple connection tracker module to indicate whether the current packet belongs to the same connection as the packet that occupies (or last occupied) the VC_{sGT} FIFO. The connection tracker checks the connection ID field (Section 3.2) in the header flit of the entering packet against the stored value of the previous one.

### 3.2 SuperGT Network Interface
The SuperGT network interface is responsible for driving the priority signal, besides creating and injecting flits into the NoC. The priority signal is set to Hi when a flit is injected during an allocated time-slot, giving this flit a priority token. The connection ID, required for connection tracking, is simply composed by the NI by adding a few bits to identify the traffic source to the routing information in the header flit (whether using source or destination routing).

To guarantee contention-free allocation in the system all (GT and SuperGT) packets that potentially use the high-priority escape channel VC_{e.si}Hi need to free the switch configuration at any router \(R\) during their allocated slots at \(R\). This is ensured by having the network interface to inject a Tail flit (with payload) before or during its last consecutively allocated time-slot to close any packet it had opened. Figure 6 shows a NI managing a SuperGT connection with three consecutively allocated time-slots in the time-slot wheel. A possible packetization policy for Hi packets is to send a Head flit (H) in the first allocated slot (1), a Body flit (B) in the second one (2) and a Tail flit (T) in the last consecutively allocated slot (3). We note this packetization scheme \{H_{Hi,1},B_{Hi,2},T_{Hi,3}\}. All other legal packetization for three consecutively allocated time-slots are: \{-,H_{Hi,2},T_{Hi,3}\}, \{H_{Hi,1},T_{Hi,2},-\}, \{-,-,-\} where – is either an Atomic (combined Head + Tail) flit or no flit.

![Figure 5: Channel escape on VC_{e.si}Hi for packet \{A,B,C\} from a different connection.](image)

![Figure 6: Packetizations of size three in a SuperGT network interface with three allocated slots.](image)
number of allocated time-slots), a SuperGT NI may inject Lo priority packets outside of the allocated time-slots, provided two constraints are fulfilled.

First, to effectively guarantee throughput, Lo packets are only injected when there is sufficient data to create one (or several) Lo packet(s) and one Hi packet (in the example of Figure 6 six flits are required). Even though packets are served more frequently than for a GT connection, this packetization policy is clearly non-work-conserving. This constraint is required to push Lo packets (by transferring them the Hi priority) from the same connection that are eventually buffered in a router along the path. It implies that SuperGT has throughput guarantees at the granularity of a time-slot wheel, whereas for GT, throughput is guaranteed for an individual packet.

The second constraint requires a Hi packet to have the same length as all Lo packets it is likely to push. This ensures the transfer of the correct number of priority tokens during a possible priority switching. In its simplest form, all packets for a given SuperGT connection need to have the same size. More complex techniques need to ensure that all Lo packets on a given SuperGT connection have been consumed. A possibility would be to use the end-to-end credit-based flow-control scheme to track packet consumption.

4. SYSTEM SIMULATION

This section demonstrates, at a cycle-accurate level, the SuperGT idea on a realistic example. We compare the performance advantages of the SuperGT QoS over a strict GT using SystemC simulation transaction-level models of the SuperGT NoC.

4.1 Simulation Model

To evaluate the performance of the SuperGT system, an instance composed of three processor nodes and one memory node has been created within NoCTurn, a custom SystemC MP-SoC platform simulator. NoCTurn is a cycle-annotated simulator where the NoC-level transaction is a flit (the experiments use 3 phits per flit). Each processor, a TIC62 instruction set simulator, is connected to a local bus that gives direct access to a private L1 memory and to a bus-shell adapter that allows access to remote memories through the NoC (Figure 7). The bus-shell behaves as a 4-way set-associative write-back cache that transmits cache misses and line evictions as bursts of data onto the NoC. The memory node is a slave on the NoC, to which it connects through a similar bus-shell unit.

As Figure 7 shows, the four bus shells in the system are attached to four network interfaces that are connected to a single router, resulting in a star topology. The contention on the L2 link is sufficient to illustrate the behavior of the various QoSs. This system has been used to create two experimental setups. One uses GT and BE traffic (similar to the $E_{GS-BE}$) whereas the other uses SuperGT and BE traffic. The connections P0-L2 and P1-L2 are GT (SuperGT) and have the same time-slot allocation for both experiments. P1-L2 is allocated half the bandwidth of P0-L2. Connection P2-L2 only supports BE traffic.

4.2 Simulation Results

All processors access data stored in L2 through caches. Cache-misses and line evictions are converted to NoC transactions by the NI bus-shell adapter.

Processors P0 and P1 run a data parallel version of cavity detection, a simple, but computationally intensive, image-filtering application from the medical domain. Input and output images (640x400 pixels) are stored in L2 (Figure 7).

Processor P2 performs FFTs on data also stored in L2 over a BE connection. In the SuperGT/BE setup, its traffic competes with the Lo priority traffic on the SuperGT connections P0-L2 and P1-L2. To compare the effect of BE load over the Lo traffic of the SuperGT we ran two experiments, one with a low BE load (FFT over 1024 points) and another with a high load (FFT over 65535 points).

![Figure 7: System simulation setup.](image)

![Figure 8: Comparing SuperGT and GT QoS.](image)

Figure 8(a) compares the average throughput on the request paths ($P0 \rightarrow L2$ and $P1 \rightarrow L2$) for both setups (the network is loaded with the high BE load of P2)\(^4\). The family of 4 curves presents the same shape, an initial peak of traffic (corresponding to L2 initialization performed by the TI processors), followed by a slow ramping-up corresponding to data transfers of the cavity application.

As expected for the GT experiment, the throughput of $P0 \rightarrow L2$ is higher than $P1 \rightarrow L2$ as it has more bandwidth allocated. For the SuperGT experiment, as Lo packets can be injected additionally to the reserved bandwidth, not only the throughput on both connections is higher than on their

\(^4\)As the applications have a very high data reuse locality, the traffic load they require from the NoC is low (though its shape is realistic). To create substantial NoC load the flit clock (96 bits per flit) has been slowed down to 120ns.
GT counterparts, but also the difference in time-slots allocated is absorbed. Indeed the throughput curves $P0 \rightarrow L2$ and $P1 \rightarrow L2$ are almost superimposed. Thanks to its guaranteed $Hi$ packets and additional non-guaranteed $Lo$ packets, the SuperGT QoS improves application throughput by 14.4% (resp. 35.6%) on the connection $P0 \rightarrow L2$ (resp. $P1 \rightarrow L2$) with respect to the GT connections.

Figure 8(b) compares the average throughput on the best effort request path ($P2 \rightarrow L2$) for both setups and under both BE traffic loads. The two top curves correspond to the high BE traffic load and the two bottom ones to the low BE traffic load. For both setups, the BE traffic competes with high priority traffic from the reserved resources. In the SuperGT versus BE setup, the BE traffic additionally competes with $Lo$ packets from the SuperGT connections. This effect is visible as the BE curves in the SuperGT experiment present a lower throughput than their GT counterparts.

Interestingly, one can see the effect of a drop in application throughput in the SuperGT connections. Indeed, an inflection point at $t = 2.10^9 \text{ns}$ only exists for the SuperGT/BE setup (at both BE loads). This corresponds exactly to the end of the initialization traffic of processors $P0$ and $P1$ (Figure 8(a)). As throughput drops on SuperGT paths $P0 \rightarrow L2$ and $P1 \rightarrow L2$, it increases on the $P2 \rightarrow L2$ path.

5. SYNTHESIS RESULTS

We have implemented the differentiating block of the SuperGT router, its virtual channel manager. The $V_{c,\text{GT}}$ and $V_{c,\text{shift}}$ FIFOs are 34 bits wide (32-bit data and 2-bit for flit information) and there are 4 flits per flit (due to restrictions of the FIFO model). The sgtVCM has been synthesized at gate level with Synopsys Physical Compiler, in 90$\mu$m TSMC standard cell technology, worst case conditions. This example uses 21 bits for the connection ID register (15 bits for source routing and 6 bits for NI ID).

$$
\begin{array}{|c|c|c|c|}
\hline
\text{Component} & \text{Comb.} & \text{Non-comb.} & \text{Total (um$^2$)} \\
\hline
V_{c,\text{GT}}(4\text{-deep}) & 6492 & 9894 & 16386 \\
V_{c,\text{shift}}(1\text{-deep}) & 1846 & 2629 & 4475 \\
\text{Control logic} & 378 & 480 & 858 \\
\hline
\text{Total} & 8716 & 13003 & 21719 \\
\hline
\end{array}
$$

Table 1: sgtVCM area (TSMC STD Cell 90nm).

The total area of the sgtVCM is under 0.0217mm$^2$, out of which 0.0209mm$^2$ are for the standard cell implementation of the FIFOs. In a commercial design these could be replaced by custom or semi-custom implementations, that can be a factor 3 to 5 less expensive. In essence, the overhead of 4% due to the logic of the sgtVCM is negligible compared to the area of the FIFOs being managed, making the sgtVCM an inexpensive addition to a regular packet-switched router with 2 virtual channels.

For the sake of comparison, if the 6x6 $E_{GS-BE}$ router with 2 virtual channels of [4] was to be extended with the logic of 6 sgtVCMS its area would be increased by only 6.6% (in 0.13$\mu$m technology the $E_{GS-BE}$ area is 0.175mm$^2$ and the logic of one sgtVCM 1917$\mu$m$^2$). For a 6x6 mesh NoC the total area increase is only of 7.2% (4.15 vs 4.45 mm$^2$). This unoptimized implementation of the sgtVCM can be clocked at over 730MHz allowing the implementation of high-performance SuperGT routers (the $E_{GS-BE}$ router implemented in 0.13$\mu$m technology runs at 500MHz).

6. CONCLUSION

This paper presents the SuperGT packet-switched router and network interface. Thanks to advanced virtual channel management the SuperGT NoC supports three QoS classes: BE, GT and SuperGT. As for GT, the SuperGT QoS class allows to reserve resources on a connection to guarantee bandwidth. Additionally, on the same connection, non-guaranteed bandwidth can be granted while guaranteeing in-order packet delivery.

Simulation results show throughput improvements up to 35.6% for the SuperGT QoS over the GT QoS. Synthesis results of the SuperGT virtual channel manager, the critical block in the SuperGT router, show an area increase of only 6.6% in a state of the art router of large arity. The SuperGT VCM is an inexpensive enhancement to state-of-the-art packet-switched networks-on-chip with differentiated services. It adds the SuperGT QoS, beneficial to many MP-SoC applications.

7. REFERENCES