

(19) World Intellectual Property Organization  
International Bureau



(43) International Publication Date  
31 August 2006 (31.08.2006)

PCT

(10) International Publication Number  
**WO 2006/089559 A1**

(51) International Patent Classification:  
**H04L 12/56** (2006.01) **G06F 9/38** (2006.01)  
**H01L 21/00** (2006.01)

(74) Agent: **INSPICOS A/S**; Bøge Allé 3, P.O. Box 45,  
DK-2970 Hørsholm (DK).

(21) International Application Number:  
PCT/DK2006/000118

(22) International Filing Date:  
28 February 2006 (28.02.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:  
PA 2005 00306 28 February 2005 (28.02.2005) DK  
60/656,376 28 February 2005 (28.02.2005) US

(71) Applicant (for all designated States except US): **DAN-**  
**MARKS TEKNISKE UNIVERSITET** [DK/DK]; Afdel-  
ing for Forskning og Innovation, Anker Engelundsvej 1,  
Bygning 101A, DK-2800 Lyngby (DK).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **BJERREGAARD,**  
**Tobias** [DK/DK]; Dalgas Have 34, 2. tv., DK-2000 Fred-  
eriksberg C (DK). **SPARSØ, Jens** [DK/DK]; Skovdraget  
4, DK-2880 Bagsværd (DK).

(81) Designated States (unless otherwise indicated, for every  
kind of national protection available): AE, AG, AL, AM,  
AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN,  
CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI,  
GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE,  
KG, KM, KN, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV,  
LY, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI,  
NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG,  
SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US,  
UZ, VC, VN, YU, ZA, ZM, ZW.

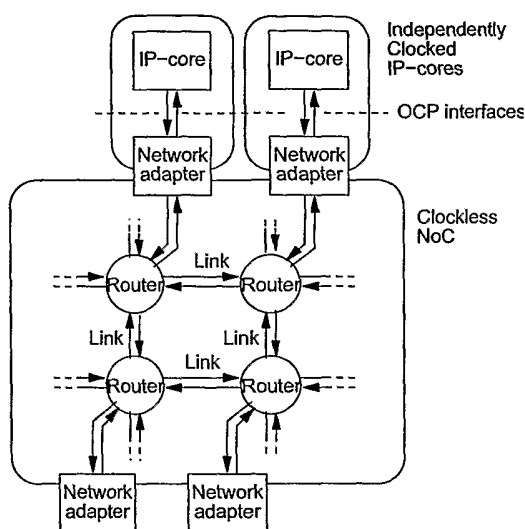
(84) Designated States (unless otherwise indicated, for every  
kind of regional protection available): ARIPO (BW, GH,  
GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM,  
ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),  
European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI,  
FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT,  
RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA,  
GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— with international search report

For two-letter codes and other abbreviations, refer to the "Guid-  
ance Notes on Codes and Abbreviations" appearing at the begin-  
ning of each regular issue of the PCT Gazette.

(54) Title: A NETWORK, A SYSTEM AND A NODE FOR USE IN THE NETWORK OR SYSTEM



(57) Abstract: A network of interconnected nodes, a system comprising the network and a node for the network/system. Each output of the nodes comprises an output buffer, and each node is adapted to inform a node transmitting data to an output buffer, whether the buffer is not full. The transmitting node is adapted to only transmit data when the buffer is not full. Arbitration is only required at the outputs of the node, and guaranteed services may be obtained for data through the network with suitable arbitration. The network and system is especially adapted as a Network on a Chip.

## A NETWORK, A SYSTEM AND A NODE FOR USE IN THE NETWORK OR SYSTEM

The present invention relates to a network comprising a number of nodes interconnected by links, and in particular to a network which may be provided on a single chip to interconnect individual applications or circuits on that chip.

- 5 Technologies of this type may be seen in: "Automatic hardware-efficient SoC integration by QoS network on chip", Electronics, Circuits and systems, 2004, ICECS 2004, Proceedings on the 2004 11<sup>th</sup> IEEE international conference on Tel Aviv, Israel, Dec. 13-15, 2004, Piscataway, NJ, USA, IEEE, pp 479-482, "Virtual channel designs for guaranteeing bandwidth in asynchronous network-on-chip", Bjerregaard T et al, Norchip conference, 2004, 10 proceedings, Oslo, Norway, 8-9 Nov. 2004, Piscataway, NJ, USA, IEEE, pp 269-272, "Trade-offs in the design of a router with both guaranteed and best-effort services for networks on chip – DATE03", IEE proceedings E. Computers & Digital Techniques, Institution of electrical engineers, Stevenage, GB, vol. 150, No. 5, 22 September 2003, pp 294-302, as well as in "A channel library for asynchronous circuit design supporting mixed-mode modelling", 15 Bjerregaard T et al, proceedings on the 14<sup>th</sup> international workshop on power and timing modelling optimization and simulation (PATMOS 2004), September 2004, pp 301-310.

- Due to the effects of scaling micro-chip technologies, it has become unavoidably necessary to differentiate between local and global on-chip communication. The concept of communication-centric design derives from the fact that global communication is becoming 20 more and more costly, in relation to processing power. Physical issues of deep submicron technologies as well as design complexity issues make current, poorly scalable solutions for global on-chip communication, such as busses, unsuited for future large scale system-on-chip (SoC) designs. There is a general consensus that shared, segmented interconnection networks, so called networks-on-chip (NoC), constitute a viable solution space to the 25 problems faced today.

- NoC facilitates a truly modular and scalable design approach in which IP reuse is an integral part. This is called upon in order to make efficient use of the exploding amount of resources available to chip designers, and so within the constraints of ever shortening design cycles. Chip-wide synchronous operation is becoming prohibitively difficult to achieve in large, high 30 performance chips. Possible solutions are in the direction of the *globally asynchronous locally synchronous* (GALS) concept, by which a chip consists of synchronous islands which communicate asynchronously. Since a NoC spans the entire chip, it seems natural not to attempt a globally synchronous implementation but instead use a GALS approach or even a fully asynchronous, i.e. clockless, realization. This is illustrated in Figure 1. The very nature 35 of a clockless NoC also makes the integration of cores with different timing characteristics an

integral part of the design flow, promoting the concept of modularity. Clockless circuits have a number of advantages over clocked ones. Since they make use of self-timed, data-driven control they operate at the maximum speed possible, and they have zero dynamic power consumption when idle. These advantages make them ideal for implementing NoC.

- 5 The present network, however, is useful independently on the choice of clocking of the chip or parts thereof, as well as the network. In addition, the present network may additionally be used not as a NoC, but as a more macroscopic network between individual, physically separate elements.

- 10 In a first aspect, the invention relates to a network comprising a number of nodes interconnected by links, wherein:

each node:

- 15 has a number of inputs and a number of outputs, each output comprising a number of output buffers, each input and each output being adapted to receive and transmit data with a predetermined data rate, and

has a forwarding element adapted to forward data, from the inputs thereof to the output buffers, at a rate being at least the combined rate with which the inputs are adapted to receive data,

- 20 a number of "data channels" are each defined as a plurality of serially interconnected nodes, for each node in a channel, an output buffer is defined of an output interconnected to an input of an other node of the channel,

each node in the channel having an input interconnected to another node of the channel further being adapted to inform the other node when the output buffer of the node and of the channel is not full, and

- 25 each node in the channel having an output interconnected to another node of the channel further being adapted to transmit data from the output buffer of the node and of the channel to the other node when the other node informs the node that the output buffer of the other node and of the channel is not full.

The overall functionality of the network is that one or more channels are defined along which data is transmitted. A channel normally is a series – or string – of nodes transmitting the data from a first node, via a number of intermediate nodes, to a last node. A channel may be defined as a series of output buffers, one output buffer in each of the nodes of the channel.

5 Thus, a node may be a part of a number of channels, but preferably an output buffer is only a part of one channel.

According to the invention, a receiving node is adapted to inform a transmitting node in the same channel that the buffer is not full. Thus, a hop-by-hop backpressure may be provided in which data is only transmitted from a transmitting node to a next, receiving node, along the  
10 channel, when the receiving node has space for the data on the pertaining output buffer of the receiving node. As a consequence thereof, lossless data transport may be obtained.

In the present connection, "interconnection" means that the nodes are adapted to transmit data to each other. Thus, this interconnection may be a physical interconnection, such as by wires or cables (electrically or radiation conducting), or a wireless interconnection, such as  
15 using radiation or radio waves. Therefore, a link may be formed by a cable or other physical element or may be a wireless connection between the input and the output in question.

Naturally, the interconnection of the nodes may be adapted to the application and functionality of the network. This will depend on the number of nodes in the network and the number of inputs and outputs of the individual nodes. A fully interconnected network, where  
20 all nodes are connected to each other, may be provided, and networks may alternatively be provided where data may be transmitted between a first and a second node only via one or more other nodes. In the last situation, a data channel then incorporating buffers in the first and the second nodes will then also incorporate buffers in the intermediate nodes.

In fact, an input or an output may be adapted to operate at a number of different data rates,  
25 but the overall functionality of the forwarding element is to ensure that the data received or receivable on the inputs of the node may also be forwarded to the output buffers thereof without stalling or congestion in the forwarding element, such that a predictable delay can be determined in the forwarding element. Again, this facilitates data transport without dropping data due to congestion or collision of data in elements.

30 Preferably, the forwarding element is adapted to forward data from each input to an output buffer with at least the data rate with which the individual input is adapted to receive data.

It is noted that the nodes need only have buffers at the outputs thereof when the transmitting nodes only transmit, when the buffer to receive the data is not empty. Thus, data may be forwarded from the output buffer of the transmitting node, via the output of that node, to the input of the receiving node, and to the relevant output buffer thereof via the forwarding element without need of buffering.

Returning to the data channels, these may be defined on beforehand in any suitable manner. Knowing, e.g., the data transfer pattern and requirements of the individual parts interconnected by the network, the channels may be fully defined before even providing the network. In this manner, the interconnection between the nodes may be adapted to the actual requirements only, and interconnections may not be required or provided between all pairs of nodes.

In general, a data channel may be defined only by the identity of the output buffers of the nodes of the channel, when each output of the nodes is interconnected to only one (if any) input of another node. The identity of two output buffers automatically identifies the output of the transmitting node and the input (even though this need not be required) of the receiving node having the other buffer.

In one embodiment, all inputs of a node are treated the same, and data received on each input is treated in the same manner independently on which input the data was received on. This may be the situation where the data channel is defined by having the nodes determine the identity of the pertaining output buffer from the data received. Thus, the data received may comprise a part, such as a header, identifying either the channel or the particular output buffer for which it is destined in this node. When transmitting the data, this node may then replace the identifying part with a part for use in the next node for again selecting the correct buffer.

If the identifying part of the data was to identify the channel (such as an identity number thereof), the node will hold information identifying the correct buffer on the basis of the channel identity.

Thus, the node may comprise a memory for holding information for use in identifying the correct output buffer on the basis of information derived from the data received, and/or the memory may hold information for adding to the received data (or a part thereof, if a part of the data has been stripped therefrom) before transmission to the next node in the channel.

In an alternative embodiment, the mere identity of the input of the node receiving the data may identify the correct output buffer. This, however, reduces the number of channels using

the node to the number of inputs thereof. Additionally, this puts restraints on the interconnection or routing of the network.

The information relating to the output buffer being not full may be transmitted continuously or only when changes occur to this status.

- 5 This information may be transmitted to all nodes having an output interconnected to an input of the present node. In that situation, these nodes preferably are adapted to, from that information, determine from which of their respective output buffers, data is to be or not to be transmitted (the identification of the channel to which the full or not full output buffer relates). Alternatively, the information may be transmitted only to a node from which data is
- 10 transmitted to the pertaining output buffer. In that situation, the receiving node (transmitting the full/not full information) will determine (such as maybe as a part of an initial setting up of the data channels) to which transmitting node, the information is to be transmitted.

- In fact, in this respect, those nodes in a channel that each has an input interconnected to another node of the channel, further is adapted to inform the other node of an amount of
- 15 data receivable in the output buffer of the node. Thus, the transmitting node may be informed of how much data may be transmitted while being assured that the pertaining output buffer is able to store it. Naturally, the pertaining buffer may be able to store any amount of data.

- Having now received data into one or more output buffers of an output of a node, this data is
- 20 to be forwarded on the link to an input of another node. Consequently, each node preferably further comprises a scheduler for controlling transmission of data from the output queues of an output to the link connected to the output. Otherwise, collisions and a resulting loss of data may occur.

- A large number of schedulers exist, such as "fair share", "fixed bandwidth", bucket-
- 25 controlled, etc.

Thus, this use of priorities and scheduling will not only provide the hop-by-hop backpressure but also guaranteed services for the individual channels. Consequently, a network may be provided which guarantees both delivery of the data and a maximum latency of the data or a minimum bandwidth over the individual links and thus over the full path or channel.

- 30 Even though the present network may be set up of independent, physically separated elements, it is especially suited as a network being embedded in a single integrated circuit. In fact, this network may then be used for transporting data between other elements also

embedded in this circuit. Naturally, this circuit may be any type of circuit, such as a CMOS integrated circuit, a GaAs integrated circuit and a SiGe integrated circuit.

In one embodiment, the inputs and the outputs of the nodes are adapted to transmit and receive data in accordance with an asynchronous handshake protocol. In this manner, no  
5 global – or even local – clocking is required, and the network may actually provide a faster data transport than if it was globally clocked.

Thus, the forwarding elements of the nodes may be adapted to operate in the absence of a common clocking signal. Also, each forwarding element could then comprise a number of sub-elements, and wherein the sub-elements of each forwarding element are adapted to  
10 operate in the absence of a clocking signal common to all sub-elements of a forwarding element. Thus, even the individual parts of the node may not require a common clocking signal. One manner of obtaining this is to have the parts act only in the presence of data. If no data are present, the sub-elements will be idle.

In another aspect, the invention relates to a system comprising a network as described  
15 previously as well as a data provider interconnected to an input of a first node of a data channel, and a data receiver, comprising a buffer, interconnected to an output of a second node of the data channel, wherein:

- the data receiver is adapted to inform the second node when the buffer of the data receiver is not full, and

- 20 - the data provider is adapted to transmit data to the first node when the first node informs the data provider that its output buffer is not full.

In this situation, the data channel is defined as a number of consecutive nodes starting with (in the direction of data transport from the provider to the receiver) the first node and ending with (with any number of intermediary nodes) the second node, which transmits the data to  
25 the receiver.

Naturally, the network may have any number of receivers and transmitters, and any receiver may receive data from any number of the transmitters, and any transmitter may transmit data to any number of the receivers.

In this aspect, the transmitter is adapted to respond to the "buffer not full" information in the  
30 same manner as a normal node, and the receiver is adapted to transmit a "buffer not full" information to the node of the network in order to again ensure no loss of data.

A third aspect of the invention relates to a node for use in the above-described network or system, the node:

- having a number of inputs and a number of outputs, each output comprising a number of output buffers, each input and each output being adapted to receive and transmit data with a predetermined data rate,
- having a forwarding element adapted to forward data, from the inputs thereof to the output buffers, at a rate being at least the combined rate with which the inputs are adapted to receive data,
- being adapted to inform one or more other nodes interconnected to an input of the node when one or more output buffers of the node is/are not full, and
- being adapted to transmit data from an output buffer thereof, when a node interconnected to an output of the node informs the node that an output buffer of the other node is not full.

All the above considerations relating to the nodes are also relevant in this situation.

- Thus, the node may further comprise means for obtaining information, such as predetermined or predefined information, relating to pairs of an output buffer of the node and an input of the node from which data is to be forwarded to the output buffer; the node being adapted to, when an output buffer of a pair is not full, inform a node interconnected to the input accordingly. Thus, the information is provided specifically to the node providing data for that particular output buffer.

Alternatively, all nodes interconnected to an input of a certain node may be informed of the full or not-full status of all output buffers of the certain node. Then, the interconnected nodes themselves will determine whether data can be forwarded from an output buffer thereof to the certain node.

- Also, the information provided need not merely be "not-full" but may actually relate to an amount of data which may be received in the buffer.

As mentioned above, the node preferably further comprises a scheduler for controlling transmission of data from the output queues of an output to the link connected to the output.

One particular manner of providing the forwarding element is one comprising:

- in or at each output of the node, one or more switching elements being adapted to switch data from any one of X inputs to any one of X outputs, X being an integer, each output of the switching elements being adapted to provide data into a different one of the output buffers of the node, and

- in or at each input of the node, an element adapted to provide received data into one of a plurality of data paths, each data path being adapted to provide data to an input of a different one of the switching elements of the node.

In this respect, "symmetrical" switching means are used, symmetrical meaning that the number of inputs and outputs being the same. These switching means have the advantage that they, lay-out-wise on a chip, are very compact. In this manner, an optimum functionality is obtained for a given area.

Thus, the internal data transport in the forwarding element is split into two decisions, one being made in the providing element deciding to which switching element, the data is provided, and a second decision is made in the switching element identifying the final output buffer. As mentioned above, these decisions may be made on information provided or received in a number of manners, such as fixed routing based on where the data is received from, data in the actual

Preferably, X equals Y-1, where Y is the number of inputs of the node, and wherein the number of output buffers for each output equals  $2*Y$ . Then, all inputs and outputs of all switching means are used, optimizing the functionality per area of the forwarding element.

Naturally, fewer output buffers may be used, where all outputs of the switching elements are then not used. Alternatively, more output buffers may be used, where the additional buffers may then be used for other traffic or data which is not part of the data channel structure and addressing structure. Such data may be any type of data, such as connection-oriented data, source-routed data and/or so-called best effort data traffic.

In a preferred embodiment, the forwarding element comprises:

- in or at each output of the node, one or more switching elements being adapted to switch data from any one of X inputs to any one of V outputs, X and V being integers, each output of the switching elements being adapted to provide data into a different one of the output buffers of the node, and

- in or at each input of the node, an element adapted to provide received data into one of a plurality of data paths, each data path being adapted to provide data to an input of a different one of the switching elements of the node.

Thus, the data from each input is directed toward an identified one of the switching elements.

5 Each output has one or more switching elements, and the output buffers thereof are each fed by an output of a switching element.

Depending on the number of switching elements as well as the number of inputs, outputs and output buffers of the node, the switching elements may have any suitable number of inputs, X, and outputs, V.

10 Preferably, X and V are identical and equal  $Y-1$ , where Y is the number of inputs of the node, and wherein the number of output buffers for each output equals  $Q*(Y-1)$ , Q being an integer. In this manner, the switching elements are symmetrical (have the same number of inputs and outputs) which optimizes the area/performance of the circuit. In addition, each output may have Q of such symmetrical switching elements.

15 In one embodiment, the inputs and the outputs of the node are adapted to transmit and receive data in accordance with an asynchronous handshake protocol, so that the data transport between nodes may be asynchronous. In addition, each forwarding element may comprise a number of sub-elements, and wherein the sub-elements are adapted to operate in the absence of a clocking signal common to all sub-elements of a forwarding element. Thus,  
20 the data forwarding inside the node may also be asynchronous, providing additional advantages.

In the following, a preferred embodiment of the invention is illustrated with reference to the drawing, wherein:

- Figure 1 illustrates a shared clockless network connecting independently  
25 clocked IP cores facilitating modularity in large scale SoC designs,
- Figure 2 illustrates the overall concept of a router or node comprising a preferred node according to the invention,
- Figure 3 illustrates a generic, output buffered virtual channel router architecture,
- 30 - Figure 4 illustrates that the present router or node implements a non-blocking input to output switching,
- Figure 5 illustrates the input/output port pair of the switching fabric of a 5x5 GS router with 8 VCs/port,

- Figure 6 illustrates share-based VC control,
- Figure 7 illustrates a BE router/node,
- Figure 8 illustrates the BE router/node integrated into a GS router, using a subset of the VCs,
- 5 - Figure 9 illustrates the 4-phase bundled data handshake protocol, on which the preferred router is based and
- Figure 10 illustrates how the links of the preferred system are implemented using a delay insensitive handshake protocol, while the routers themselves are implemented using a bundled data protocol.

10

The present implementation is based on clockless circuit techniques. As mentioned in the introduction, it is getting increasingly difficult to implement a globally spanning, synchronous clock signal. Though a clockless implementation of the network links and routers effectively solves many problems, simply by not needing a global clock signal, the basic ideas of the presented network, relating to the establishment of logically independent data channels on which guaranteed services can be provided, are not limited to such an implementation. Other styles of implementation are, without being limited to: a fully synchronous system in which all routers and links are clocked by a common synchronous clock, mesochronous in which routers and links are clocked independently by a clock of the same frequency but possibly at a different phase, asynchronously in which each router is clocked by completely independent clocks, and handshaked asynchronous or clockless in which all data movement in the system is handled without the need of a clock common to any sub-element in any part of the system. This is commonly known as asynchronous handshake flow control. This is the style of implementation on which the preferred network is based. A data path is separated into combinatorial regions by pipeline stages. As illustrated in Figure 9, data integrity during transmission from one pipeline stage to the proceeding one, is ensured by an asynchronous handshake protocol. In the figure the commonly known 4-phase bundled data protocol is illustrated, in which the sending stage issues a request signal indicating that data is valid. The receiving stage issues an acknowledge signal upon latching the data. After this a return-to-zero phase is commenced during which first the sender withdraws its request, then the receiver withdraws its acknowledge signal. Other well known handshake protocols may also be used, such as delay-insensitive ones, or a mix of protocols.

20

25

30

In the following, we present a worm-hole NoC router to be used in MANGO (*Message-passing Asynchronous Network-on-Chip providing Guaranteed services through OCP interfaces*). The present implementation is based on clockless circuit techniques. It has features, area and performance comparable to a similar class of clocked routers. Additional benefits are an inherent support for GALS systems and zero dynamic idle power. An interesting feature lies in the concepts of establishing the GS paths and the application of methods developed into

implementing non-blocking behaviour amongst several independent streams in the shared part of the link/router. The network is demonstrated with a 0.12  $\mu\text{m}$  CMOS standard cell implementation. Presently, the focus will be more on the router architecture and less on circuit realizations. In Section 1, argumentation for the implementation of GS in NoC is provided, and in Section 2, an overview of MANGO is given. In Sections 3 and 4, the details of the router architecture are explained, and it is explained how it is possible to provide GS via connections. Finally, in Section 5, the present implementation is explained.

## 1 Background

When designing NoCs, much research done in the area of multi computer networks can be readily built upon. The trade-offs differ however, and while pin count and global wiring have been the main limiting factors in multi computer networks, NoC designers face constraints on area and power. Also latency requirements are more stringent in tightly coupled single-chip systems. As a result, NoC demands simple solutions to ensure high speed and low area. Traditionally, multi computer networks support best effort (BE) routing. Much NoC research builds on this by improving BE routing efficiency under the constraints of single-chip system design. In particular the use of virtual channels (VCs) – logically independent channels sharing a physical link – is employed. In this implementation, it is proposed to use the VC routing resources somewhat differently, i.e. for establishing GS connections. While an improvement of BE routing simply makes better use of existing wires, connection-oriented GS routing adds explicit new features: packets travel along virtual circuits, avoiding the mutual influence that BE packets routed on the same logical network may experience. From an application-level perspective, the predictability that GS incurs promotes system integrity, and simplifies functional verification since there is less call for taking into account the complex dynamics of a shared communication structure. At the circuit level, the circuits needed to implement GS also turn out to be simpler than those needed for BE, since less arbitration is called for during routing control.

## 2 The preferred (MANGO) Router

As shown in Figure 1, a MANGO network, as other NoCs has network adapters (NA), routers/nodes and links. Each IP core is connected to the network through an NA, providing high level communication services, i.e. OCP transactions, on the basis of primitive services implemented by the network. Each NA, which also performs the synchronization between the clocked IP core (if clocked at all) and the clockless network, is connected to a router/node. In the presently preferred embodiment, the routers are connected by links in a grid type structure, either homogeneous or heterogeneous. To keep the speed up, long links can be

implemented as pipelines. Naturally, any type of link structure and interconnection may be chosen.

Figure 2 shows a conceptual illustration of the preferred MANGO router. The router implements a number of uni-directional ports. Two of these are local ports which connect to the local NA. The local ports have a number of physical interfaces. The remaining ports are network ports which, via point-to-point links, connect the router to neighbouring routers. Each of these ports implements a number of independently buffered VCs. Network ports and local ports implement the same type of interface. It is the function of the NA to translate communication to and from network packet format. Internally, the preferred router has a BE router, a GS router, output buffers, and link arbiters. The BE and the GS routers are separately implemented. This is done in order to make the router modular. The BE router dynamically source-routes connection-less data packets, according to the routing path defined in the packet header. A subset of the VCs are allocated for BE routing. The GS router uses the remaining VCs to route header-less data streams on statically programmable connections. In order to make hard service guarantees, GS connections must be logically independent of other network traffic. In MANGO, a connection implements a logical point-to-point circuit between two different local ports in the network, by reserving a sequence of independently buffered VCs. The GS router provides non-blocking switching between the input ports and the output buffers, therefore *GS can be realized purely on the basis of link access arbitration*. This will become clear in the following sections. The GS connections are set up by programming these into the GS router via the BE router.

In the present implementation, a connection is predefined, and a node may ascertain the correct VC from information added to the data. However, as is also clear from further above, a large number of manners exist of providing the information defining the data channels.

In addition, a BE router is implemented in the present implementation in order to provide the backpressure signals and in order to provide additional, which may be forwarded on the links, when no GS information (information presently carried on the data channels) is waiting. This BE router may be used for a large number of applications but is not required according to the invention.

### 3 The GS Router

#### 3.1 Overview

In the following description, we address data items in the network as flits (flow control units). This represents any data item quantification of a number of possible quantifications, i.e. flits, packets, or other terminology. Figure 3 shows a generic prior art output buffered VC router with  $P$  input and  $P$  output ports; a PXP switch is followed by a split module, which splits the incoming traffic to individual VC buffers at the output ports. Since several input ports may attempt to access the same output port simultaneously, congestion may occur. This makes the architecture unsuitable for providing service guarantees.

- Figure 4 illustrates the connection-oriented GS router architecture according to the preferred embodiment. Again,  $P$  is the number of ports on each routing node, while  $V$  is the number of VCs on each port. As touched upon in Section 2, a connection is a reserved sequence of VCs through the network. In the router, incoming flits are guided through the switching module (Section 3.2) – by way of *steering bits* which are appended to the flits in the previous router – to the VC buffer that has been reserved for the given connection. To prevent flits of different VCs from stalling and blocking each other on the links, VC control is employed. This ensures that flits will only be transmitted on a link, if there is free buffer space in the target VC buffer. The VC control module (Section 3.3) establishes a VC control channel from the VC buffers back to the VC buffers in the previous router; a step back on a given connection. For each connection, a router thus stores the steering bits needed to guide flits to the VC buffer reserved for the connection in the *next* router, as well as *control channel bits* used to establish a VC control channel back to the VC buffer in the *previous* router. This information is programmed into the router using BE packets (via the programming interface, Figure 2). As seen, the setup information for each hop on a connection is stored in two places: one for the flit forward path, and one for the VC control reverse path. This overhead was accepted because it facilitates very simple circuits. The switching module is non-blocking. This means that the latency from the time that a flit is granted access to a link, until it arrives at the designated VC buffer in the following router, is constant. In order to provide hard GS for a connection there is thus only call for link access arbitration. The link arbiter (Section 4.4) is the key element in providing GS. It arbitrates amongst the VCs contending for access to the link, implementing the type of GS that is provided.

### 3.2 The Switching Module

The switching module makes it possible for all input ports to route flits to any combination of VC buffers. It is made entirely without any form of arbitration necessary, making its performance highly predictable, and simplifying its design, saving both area and routing latency. Since a VC buffer is part of only one connection, no congestion will occur as only one input will attempt to route to one VC buffer at any given time. Figure 5 shows an input/output port pair in a 5X5 port router with eight VCs on each port. This configuration will form the basis for the implementation being used as an example of the architecture. The diagram indicates how the remaining ports are connected. An input port needs only connect to four output ports as it is not useful to route flits back in the direction they came from. The diagram also indicates how a total of five steering bits are used in this embodiment. The first three bits are used by a split module, which directs the flit to one of two 4X4 switches at each output port. At each switch, two more steering bits are used to direct the flit to one of four VC buffers. At each stage, the steering bits used are stripped, as they are no longer needed.

The switching module, which constitutes a considerable part of the total router area, scales linearly with the number of VCs, and thus with the number of connections supported.

### 3.3 The VC Control Module

In the present embodiment, *share-based* VC control is employed, which we have described in [5]. The scheme, illustrated in Figure 6, uses a single wire per VC to implement non-blocking access to a shared media, e.g. a link. When admitting a flit to the media, the *sharebox* locks, not allowing further flits to pass. The flit passes across the media, to the *unsharebox* at the far side. The *unsharebox* implements a latch, into which the flit is accepted. When the flit in turn leaves the *unsharebox*, the *unlock* control wire toggles. This unlocks the *sharebox*, admitting another flit to the media. As long as the media is deadlock free, no flit will stall within it. In the MANGO router, the link and the switching module constitute the shared media. Even if the link cycle for each flit transmitted on a VC is long, the full link bandwidth is exploited by the unlock handshake of different VCs overlapping.

A single VC can normally not utilize the full link bandwidth. However with an appropriate link access scheme, the goal to provide well defined worst case guarantees, can be met. The cycle time of the VC link is sensitive to the forward latency of the flits. In this regard, clockless circuits are advantageous, since they make possible a very short forward latency per pipeline stage. An advantage of the share-based VC control scheme is that it is much cheaper, both area and power wise, than the commonly used credit-based VC control

scheme. As shown in [5], the two schemes can be used together, controlling access to the same link. This is useful when the same link is used for both GS and BE routing.

During BE routing it is clearly an advantage to use credit based VC control to improve the average case performance. As indicated in Figure 4, each input port of the router implements an output unlock wire for each VC, likewise each output port implements an input unlock wire per VC. Under the share-based VC control scheme, the VC control module establishes control channels between arbitrary VC buffers and the inputs, simply by circuit switching the unlock wires, originating in the VC buffers, unto the VC control outputs at the input ports. Thus the VC control module implements a non-blocking  $(P \times V) \times (P \times V)$  switch. The control channel bits stored in the router map the VC buffers to the appropriate input VC wires, according to the connection path. In our 5X5, 8 VC router we simply use 5\*8 instantiations of a (5-1)\*8-input multiplexer. For larger number of VCs, it might prove worthwhile to implement a more complex switch structure, e.g. a Clos network, which scales better. The mapping between input and output VCs can be considered static during connection usage.

### 3.4 Link Access

The MANGO router implements output buffers. Since a connection is a reserved sequence of VCs, the target VC buffer of a flit entering the router is deterministic. There is no need for arbitrating amongst VCs on different output ports. Placing the VC buffers at the output simplifies the link access arbitration scheme. Thus simple, fast, and low area circuits can be used to implement access schemes. Since the path across the link, through the switching module, to the target VC buffer is fully deterministic and congestion free, by way of the non-blocking switching module described in section 3.2, the link access is the only critical point on a connection, with regard to congestion. If guarantees can be made for link access, guarantees can be made for the entire connection. Thus link access arbitration is what facilitates GS in MANGO.

The *link arbiter* shown in Figures 2, 4, 5 and 8 implements the arbitration scheme. According to the arbitration amongst contending VCs, the flits are merged onto the shared link, which connects to the neighbouring router. The steering bits, which indicate the destination of the flit, are appended before the flit is transmitted. In the router implementation presented herein, a demonstration is given of the general architecture implemented a fair-share access scheme, by which each of 8 VCs are guaranteed a minimum of one 8th of the total link bandwidth. To keep the area down, the present output buffers are a single flit deep plus one flit in the unsharebox. This is enough to ensure the fair-share scheme to function over a sequence of links, providing a hard lower bound on the throughput of a connection. If a VC does not use its allocated bandwidth, the link is automatically used by another contending

VC. It is seen how the GS routing functionality is decoupled from the switching functionality in the router. This makes it an easy and modular task to instantiate new GS schemes.

Naturally, other types of scheduling may be used. Different schedulers have different advantages and disadvantages. The present network/router is basically indifferent as to the type of scheduling used.

A particular advantage of the present network/router is that it is indifferent to the type of scheduling used, and because of the non-blocking forwarding element (the switching module) within the router, this scheduling determines the type of guaranteed services provided on the data channels.)

Global communication will incur an increasing part of the total power budget in future microchip technologies. Also the delay uncertainty of globally propagated signals forces circuit designers to allocate progressively safe timing margins for these, at the expense of performance. For these reasons we choose to implement our inter-router links using a 1-of-4 coded delay insensitive (DI) 4-phase asynchronous signaling protocol. Figure 10 illustrates how the routers themselves are implemented using area efficient bundled data control. Within the relatively limited physical extent of a router, wire delays can be considered small compared with gate delays, and it is thus possible to make use of speed independent control circuits. On the long global inter-router links, DI signaling is employed. In particular, 1-of-4 DI encoding ensures a balanced trade-off between power consumption and area usage. Four wires are needed for each 2 bits communicated, plus one wire for the acknowledge signal. Because data validity is encoded into the data itself, DI signaling eliminates timing problems caused by delay differences between data and control signal propagation. This is a highly desirable property on long global links such as inter-router pipelines, and helps in obtaining a modularized system design flow.

In the current version of MANGO, we implement 4-phase signaling. For each transaction we thus need 2 global wire toggles for each 2 bits communicated, plus 2 toggles of the acknowledge wire. As the power cost of toggling global wires increase in future technologies, relative to that of processing logic, it will prove worthwhile to implement 2-phase signaling on the inter-router links. This would reduce the amount of toggling of global wires to 1 toggle for each 2 bits plus 1 toggle for the acknowledge. The overhead in implementing 2-phase pipelines lies in the encoding and decoding front-end circuits at the router ports, and in the completion detection in each pipeline stage.

#### 4 The BE Router

As mentioned above, the BE router is introduced as a manner of providing the back pressure signals between the nodes in order to take up any excess bandwidth on the links. In addition, this router may be used for traffic which does not progress along the data channels defined. The BE router may be any type of router in that it merely uses the links and interconnections of the nodes. The determination, in a node, of the pertinent output of received data, and even the type of data, may be selected independently of the GS traffic.

The BE router, shown in Figure 7, implements a simple source routing scheme. The first flit of a packet is the header flit. The two MSBs of the header indicate one of four output ports.

Choosing a direction back to where it came from, the packet is routed to the local port. The header is then rotated two bits, positioning the header bits for the next hop. With 32-bit flits, a packet can make a total of 15 hops. The packets are of variable length; a control bit is used to indicate the last flit. The outputs arbitrate fairly between input ports contending for access. Once an input port has gained access, it will retain this until the last flit has been routed, thus keeping packet coherency. The interface used to program GS connections into the GS router, is implemented as an extension on port 0, the local port. Figure 8 is an illustration of the router, indicating how the BE router is integrated into the GS router architecture. To avoid deadlocks, XY-routing is employed. When a flit enters the BE router, three steering bits have been stripped, and a total of 34 bits remain (see Figure 5), 32 of which are flit data. Of the other two, one is the control bit used to indicate the last flit. The remaining bit can be used to indicate one of two BE VCs. This is not used in the present implementation, but can be used to extend the BE router to provide more complex deadlock free routing, adaptive VC allocation, etc. The VC control of the BE channels is handled separately from the VC control module and can be implemented using a credit-based scheme. The focus of the present embodiment is means of providing GS by way of the non-blocking GS router. The BE router has been included for the sake of completeness and holds lots of potential for improvement. First of all, it could implement credit-based VC control, in order to facilitate improved average case performance. The modular architecture of the MANGO router makes it easy to plug in a new BE router.

#### 5 Implementation and Results

A 32-bit MANGO router has been implemented using 0.12  $\mu\text{m}$  CMOS standard cells. The router is designed with 4-phase bundled data clockless control circuits. It is a relatively small entity, thus internal timing is quite predictable. The links between neighbouring routers are much longer, and thus more sensitive to timing variations. In order to make assembling a NoC-based SoC a modular and timing safe exercise, and in order to save power, delay

insensitive signalling may be used between routers, e.g. 1-of-4 signalling. The router has local input and output ports providing 5 interfaces each (4 for GS and 1 for BE) and 4 input and 4 output network ports (each implementing 8 VCs). The router simultaneously supports connection-less BE routing plus a total of 32 independently buffered GS connections. Each VC

5 provides one 8th of the bandwidth on a link.

The performance in net list simulations using worst-case timing parameters (1.08 V/125 C) was 515 MHz per port (795 MHz under typical timing conditions). The pre-layout area was 0.188 mm<sup>2</sup>. The area usage is detailed in Table 1. The switching module and the VC buffers together account for more than half of the total area. Much area could be saved by using

10 custom-designed buffers.

| Module           | Area                        |
|------------------|-----------------------------|
| Connection table | 0.005 mm <sup>2</sup>       |
| Switching module | 0.065 mm <sup>2</sup>       |
| VC buffers       | 0.047 mm <sup>2</sup>       |
| Link access      | 0.022 mm <sup>2</sup>       |
| VC control       | 0.016 mm <sup>2</sup>       |
| BE router        | 0.033 mm <sup>2</sup>       |
| <b>Total</b>     | <b>0.188 mm<sup>2</sup></b> |

Table 1. Area usage in the MANGO router.

## CLAIMS

1. A network comprising a number of nodes interconnected by links, wherein:

each node:

- has a number of inputs and a number of outputs, each output comprising a number of output buffers, each input and each output being adapted to receive and transmit data with a predetermined data rate, and
- has a forwarding element adapted to forward data, from the inputs thereof to the output buffers, at a rate being at least the combined rate with which the inputs are adapted to receive data,

a number of "data channels" are each defined as a plurality of serially interconnected nodes, for each node in a channel, an output buffer is defined of an output interconnected to an input of an other node of the channel,

each node in the channel having an input interconnected to another node of the channel further being adapted to inform the other node when the output buffer of the node and of the channel is not full, and

each node in the channel having an output interconnected to another node of the channel further being adapted to transmit data from the output buffer of the node and of the channel to the other node when the other node informs the node that the output buffer of the other node and of the channel is not full.

2. A network according to claim 1, wherein each node further comprises a scheduler for controlling transmission of data from the output buffers of an output to the link connected to the output.

3. A network according to any of the preceding claims, wherein those nodes in a channel that each has an input interconnected to another node of the channel, further is adapted to inform the other node of an amount of data receivable in the output buffer of the node.

4. A network according to any of the preceding claims, the network being embedded on a single integrated circuit.

5. A network according to any of the preceding claims, wherein the inputs and the outputs of the nodes are adapted to transmit and receive data in accordance with an asynchronous handshake protocol.
6. A network according to any of the preceding claims, wherein the forwarding elements of the nodes are adapted to operate in the absence of a common clocking signal.
7. A network according to claim 6, wherein each forwarding element comprises a number of sub-elements, and wherein the sub-elements of each forwarding element are adapted to operate in the absence of a clocking signal common to all sub-elements of a forwarding element.
8. A system comprising a network according to any of the preceding claims as well as a data provider interconnected to an input of a first node of a data channel, and a data receiver, comprising a buffer, interconnected to an output of a second node of the data channel, wherein:
- the data receiver is adapted to inform the second node when the buffer of the data receiver is not full, and
  - the data provider is adapted to transmit data to the first node when the first node informs the data provider that its output buffer is not full.
9. A node for use in the network according to any of claims 1-7 or the system according to claim 8, the node:
- having a number of inputs and a number of outputs, each output comprising a number of output buffers, each input and each output being adapted to receive and transmit data with a predetermined data rate,
  - having a forwarding element adapted to forward data, from the inputs thereof to the output buffers, at a rate being at least the combined rate with which the inputs are adapted to receive data,
  - being adapted to inform one or more other nodes interconnected to an input of the node when one or more output buffers of the node is/are not full, and

- being adapted to transmit data from an output buffer thereof, when a node interconnected to an output of the node informs the node that an output buffer of the other node is not full.

10. A node according to claim 10, further comprising means for obtaining information relating to pairs of an output buffer of the node and an input of the node from which data is to be forwarded to the output buffer; the node being adapted to, when an output buffer of a pair is not full, inform a node interconnected to the input accordingly.

11. A node according to claim 9 or 10, further comprising a scheduler for controlling transmission of data from the output buffers of an output to the link connected to the output.

12. A node according to any of claims 9-11, further being adapted to inform the one or more nodes interconnected to an input of the node of an amount of data receivable in the output buffer(s) of the node.

13. A node according to any of claims 9-12, wherein the forwarding element comprises:

- in or at each output of the node, one or more switching elements being adapted to switch data from any one of X inputs to any one of V outputs, X and V being integers, each output of the switching elements being adapted to provide data into a different one of the output buffers of the node, and

- in or at each input of the node, an element adapted to provide received data into one of a plurality of data paths, each data path being adapted to provide data to an input of a different one of the switching elements of the node.

14. A node according to claim 13, wherein X and V are identical and equal Y-1, where Y is the number of inputs of the node, and wherein the number of output buffers for each output equals  $Q \cdot (Y-1)$ , Q being an integer.

15. A node according to any of claims 9-14, wherein the inputs and the outputs of the node are adapted to transmit and receive data in accordance with an asynchronous handshake protocol.

16. A node according to claim 15, wherein each forwarding element comprises a number of sub-elements, and wherein the sub-elements are adapted to operate in the absence of a clocking signal common to all sub-elements of a forwarding element.

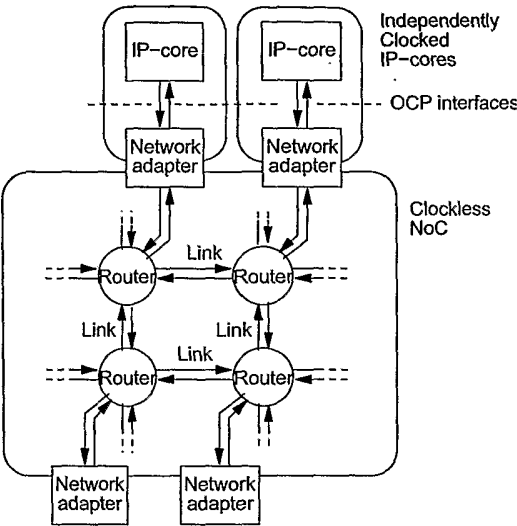


Fig. 1

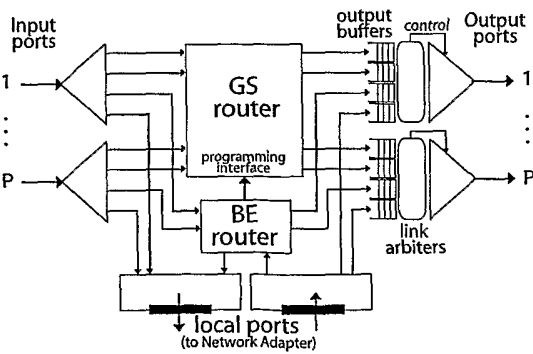


Fig. 2

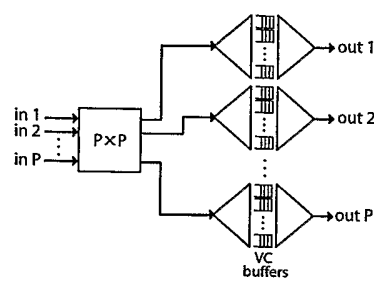


Fig. 3

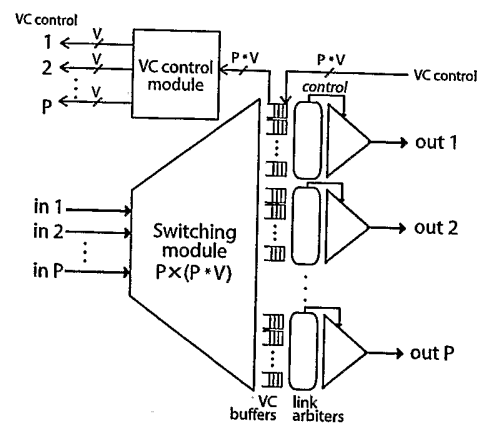


Fig. 4

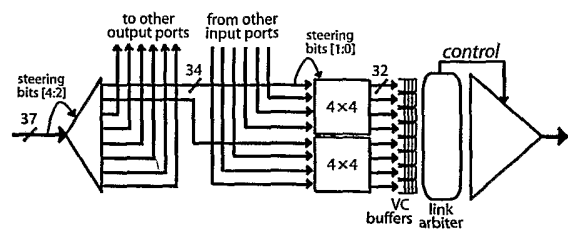


Fig. 5

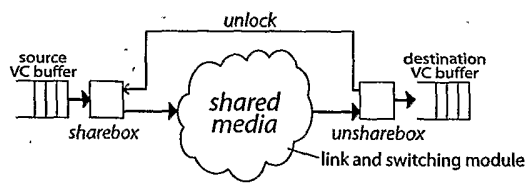


Fig. 6

4/6

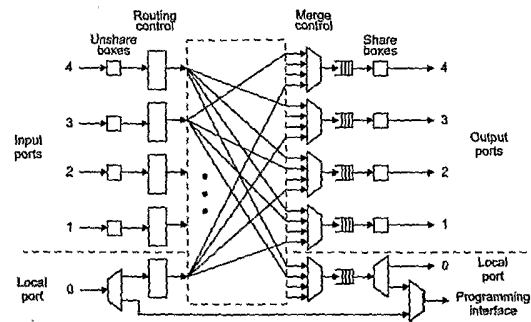


Fig. 7

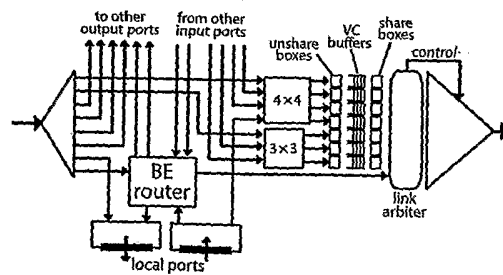


Fig. 8

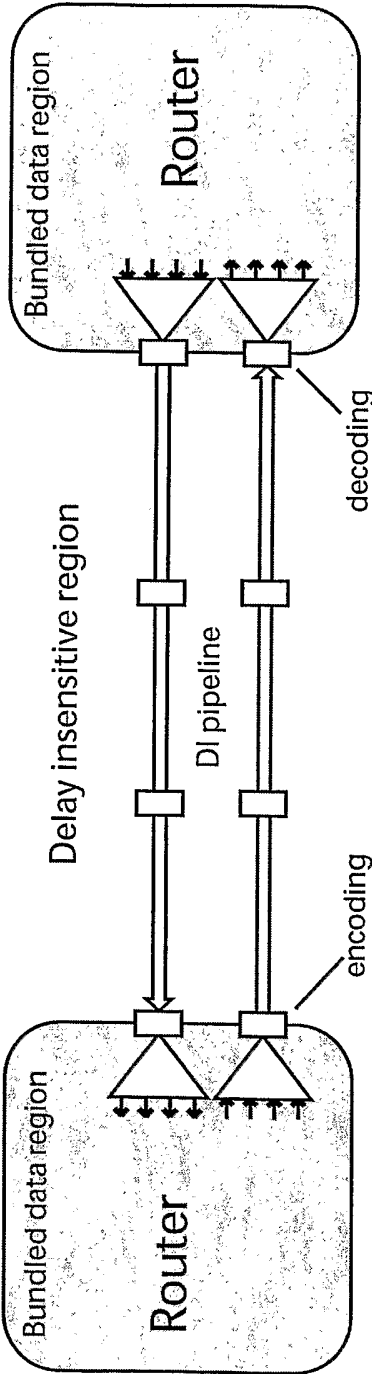


Fig. 9

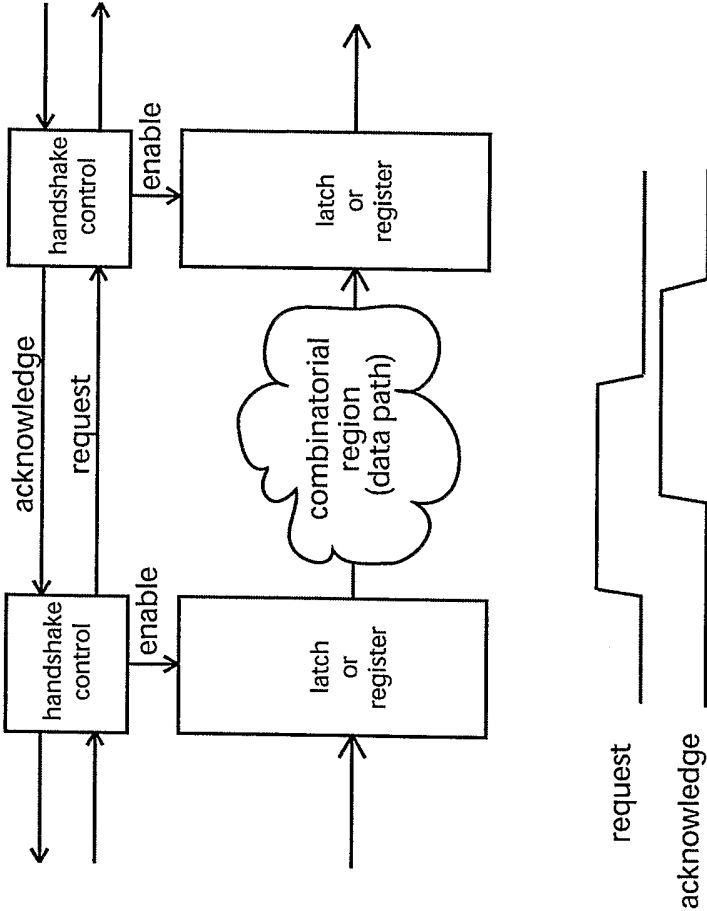


Fig. 10

# INTERNATIONAL SEARCH REPORT

International application No

PCT/DK2006/000118

## A. CLASSIFICATION OF SUBJECT MATTER

INV. H04L12/56 H01L21/00 G06F9/38

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

H04L H01L G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                                                                                                                                                                                                                                                                             | Relevant to claim No. |
|-----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| X         | BOLOTIN E ET AL: "Automatic hardware-efficient SoC integration by QoS network on chip"<br>ELECTRONICS, CIRCUITS AND SYSTEMS, 2004.<br>ICECS 2004. PROCEEDINGS OF THE 2004 11TH IEEE INTERNATIONAL CONFERENCE ON TEL AVIV, ISRAEL DEC. 13-15, 2004, PISCATAWAY, NJ, USA, IEEE, 13 December 2004 (2004-12-13), pages 479-482, XP010774300<br>ISBN: 0-7803-8715-5 | 1-4, 8-12             |
| Y         | paragraph [2.1.QNOC.TOPOLOGY]<br>paragraph [2.4.QNOC.ROUTERS]; figure 2<br>-----<br>-/-                                                                                                                                                                                                                                                                        | 5-7,<br>13-16         |

☒ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

### \* Special categories of cited documents :

- \*A\* document defining the general state of the art which is not considered to be of particular relevance
- \*E\* earlier document but published on or after the international filing date
- \*L\* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- \*O\* document referring to an oral disclosure, use, exhibition or other means
- \*P\* document published prior to the international filing date but later than the priority date claimed

- \*T\* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
- \*X\* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
- \*Y\* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
- \*G\* document member of the same patent family

Date of the actual completion of the international search

27 April 2006

Date of mailing of the international search report

08/05/2006

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2  
NL - 2280 HV Rijswijk  
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,  
Fax: (+31-70) 340-3016

Authorized officer

Lefebvre, L

## INTERNATIONAL SEARCH REPORT

International application No

PCT/DK2006/000118

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                                                                                                                                                                                                                                                                                                                                                                  | Relevant to claim No. |
|-----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| Y         | <p>BJERREGAARD T ET AL: "Virtual channel designs for guaranteeing bandwidth in asynchronous network-on-chip"<br/> NORCHIP CONFERENCE, 2004. PROCEEDINGS OSLO, NORWAY 8-9 NOV. 2004, PISCATAWAY, NJ, USA, IEEE,<br/> 8 November 2004 (2004-11-08), pages 269-272, XP010791779<br/> ISBN: 0-7803-8510-1<br/> paragraph [III.ACCESS.CONTROL]; figures 2-5<br/> paragraph [IV.BANDWIDTH.GUARANTEES]</p>                                                 | 5-7,<br>13-16         |
| A         | <p>RIJPKEMA E ET AL: "Trade-offs in the design of a router with both guaranteed and best-effort services for networks on chip - DATE03"<br/> IEE PROCEEDINGS E. COMPUTERS &amp; DIGITAL TECHNIQUES, INSTITUTION OF ELECTRICAL ENGINEERS. STEVENAGE, GB,<br/> vol. 150, no. 5,<br/> 22 September 2003 (2003-09-22), pages 294-302, XP006021112<br/> ISSN: 0143-7062<br/> paragraph [3.NETWORKS.ON.CHIP]<br/> paragraph [4.COMBINED.GT.BE.ROUTER]</p> | 1-16                  |
| A         | <p>TOBIAS BJERREGAARD, SHANKAR MAHADEVAN AND JENS SPARSO: "A Channel Library for Asynchronous Circuit Design Supporting Mixed-Mode Modeling"<br/> PROCEEDINGS OF THE FOURTEENTH INTERNATIONAL WORKSHOP ON POWER AND TIMING MODELING OPTIMIZATION AND SIMULATION (PATMOS 2004), September 2004 (2004-09), pages 301-310, XP002339818<br/> paragraph [2.2.MIXED.MODE.COMMUNICATION];<br/> figure 8</p>                                                | 14                    |