US 20060227774A1

(19) **United States**

(12) **Patent Application Publication**

(10) **Pub. No.: US 2006/0227774 A1**

(43) **Pub. Date: Oct. 12, 2006**

Hoenicke

(54) **COLLECTIVE NETWORK ROUTING**

(75) Inventor: **Dirk Hoenicke**, Ossining, NY (US)

Correspondence Address:
SCULLY SCOTT MURPHY & PRESSER, PC
400 GARDEN CITY PLAZA
SUITE 300
GARDEN CITY, NY 11530 (US)

(73) Assignee: **International Business Machines Corporation**, Armonk, NY

(21) Appl. No.: **11/100,207**

(22) Filed: **Apr. 6, 2005**

Publication Classification

(51) **Int. Cl.**
H04L 12/28 (2006.01)

(52) **U.S. Cl.** **370/389**

(57) **ABSTRACT**

Disclosed are a unified method and apparatus to classify, route, and process injected data packets into a network so as to belong to a plurality of logical networks, each implementing a specific flow of data on top of a common physical network. The method allows to locally identify collectives of packets for local processing, such as the computation of the sum, difference, maximum, minimum, or other logical operations among the identified packet collective. Packets are injected together with a class-attribute and an opcode attribute. Network routers, employing the described method, use the packet attributes to look-up the class-specific route information from a local route table, which contains the local incoming and outgoing directions as part of the specifically implemented global data flow of the particular virtual network.

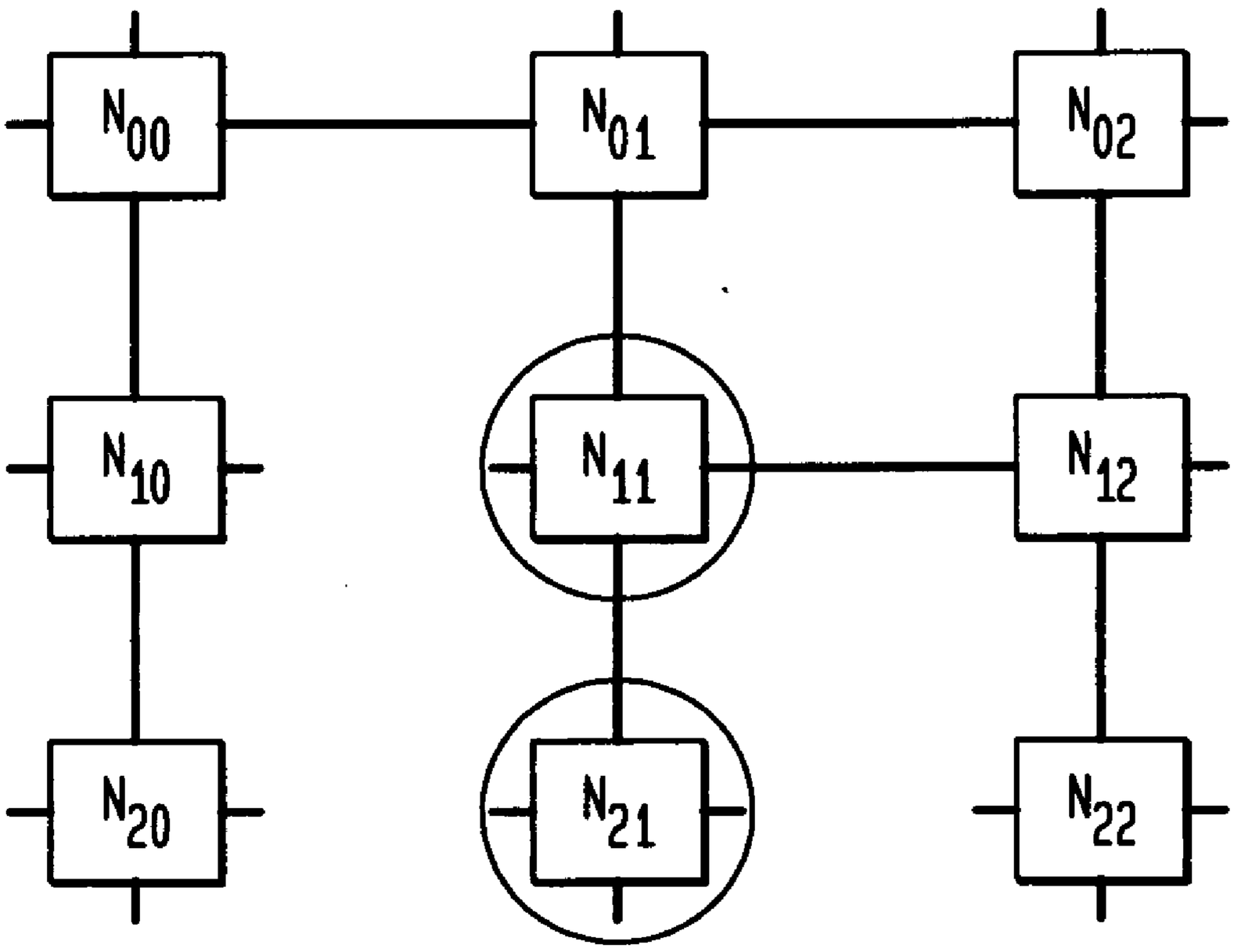


FIG. 1

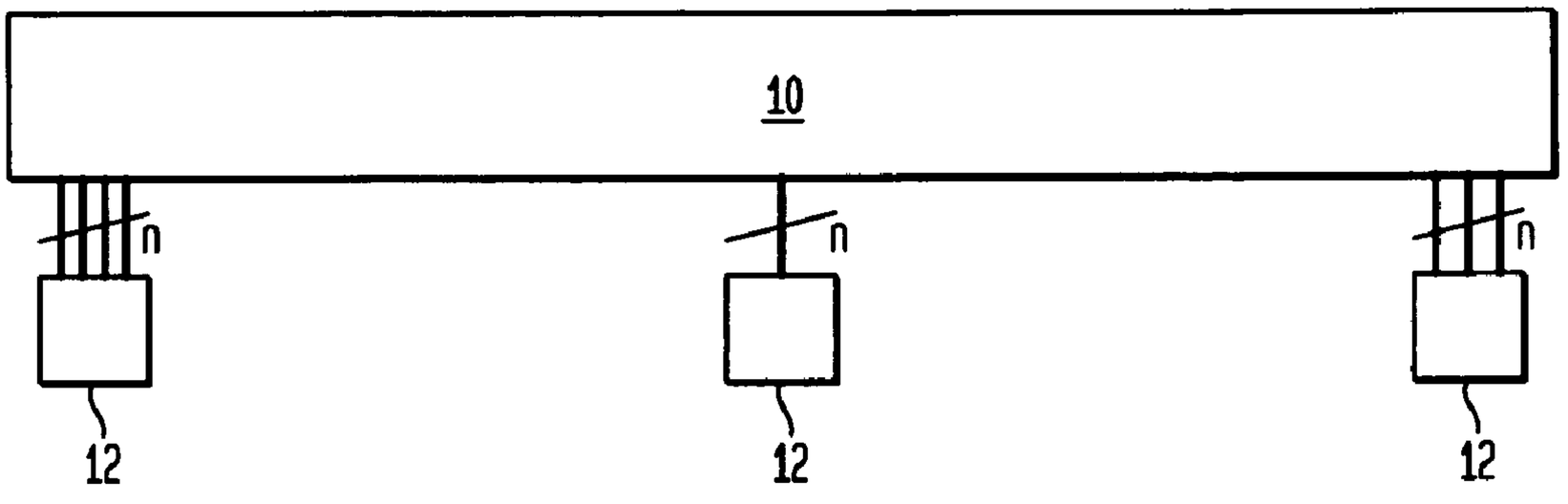


FIG. 2

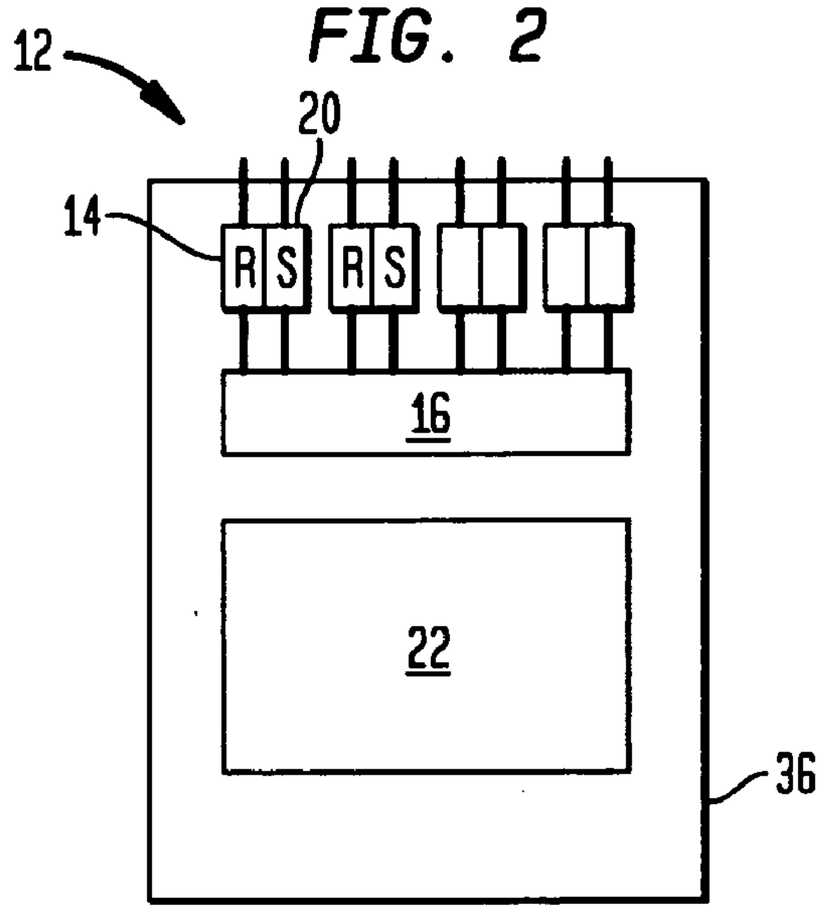


FIG. 3

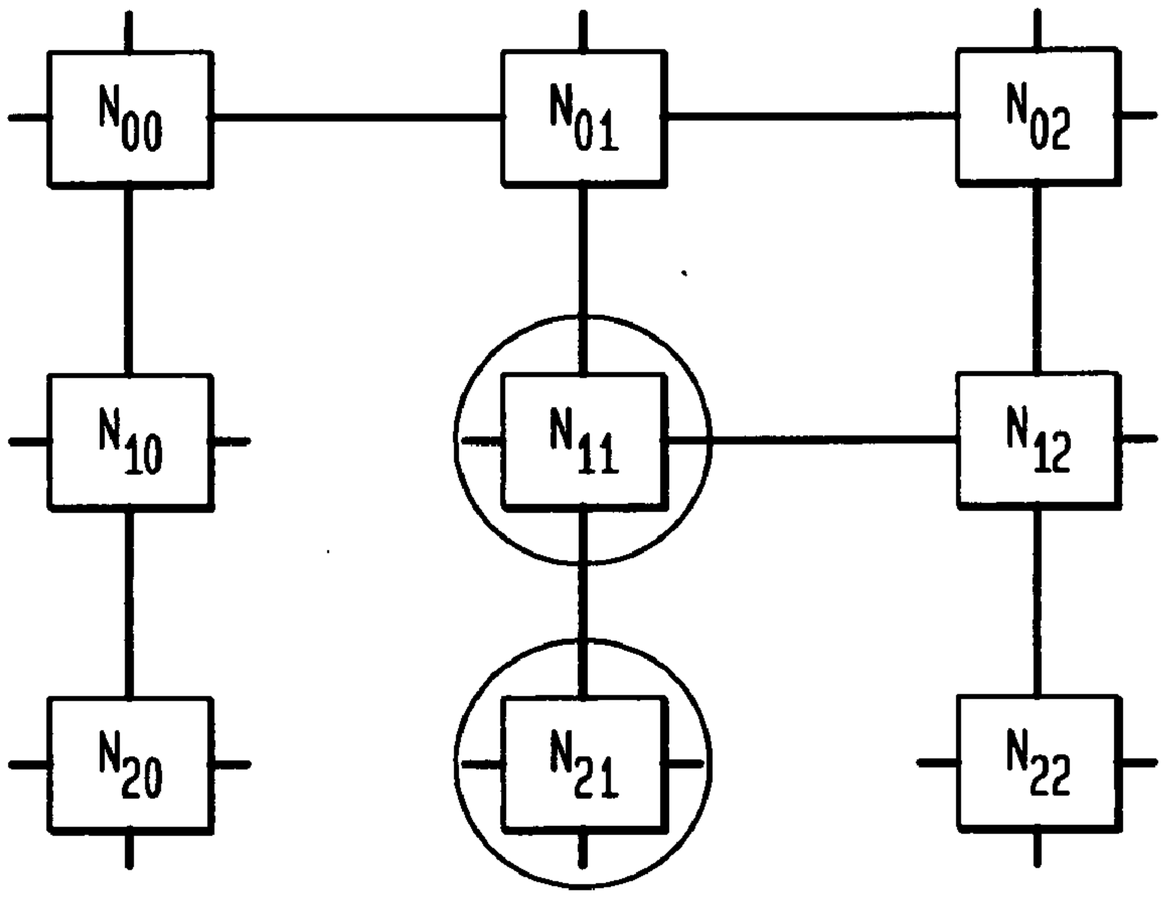


FIG. 4

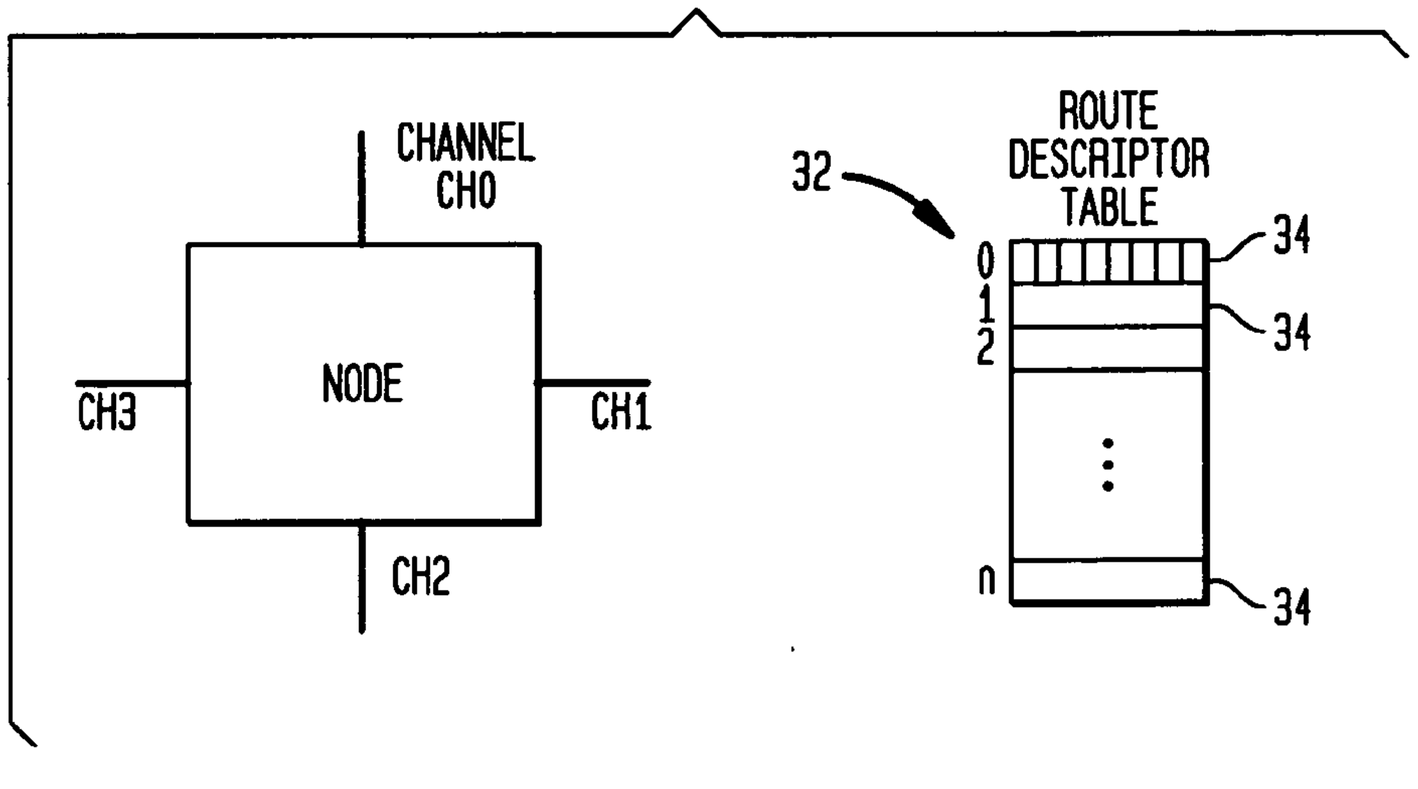
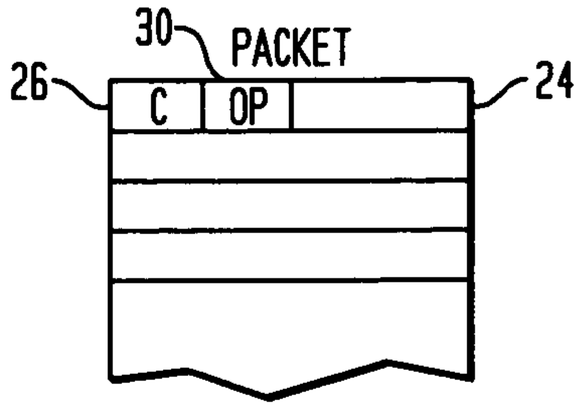


FIG. 5

		CH3	CH2	CH1	CH0	LCL	CH3	CH2	CH1	CH0	LCL
NODE A	0:	0	1	1	0	1	0	0	0	1	0
	1:										
NODE B	0:	0	0	0	0	1	0	0	0	1	0
	1:										

FIG. 6



COLLECTIVE NETWORK ROUTING

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of Provisional Application No. 60/625,026, for "Collective Network Routing," filed Nov. 4, 2004.

GOVERNMENT CONTRACT

[0002] This invention was made with Government support under Subcontract B517552 under prime contract W-7405-ENG-48 awarded by The Department of Energy. The Government has certain rights in this invention.

BACKGROUND OF THE INVENTION

[0003] 1. Field of the Invention

[0004] This invention generally relates to the field of high-speed digital data processing systems; and more specifically, the invention relates to methods and systems for routing messages in computer systems.

[0005] 2. Background Art

[0006] Massively parallel computer systems comprise a large number of data processing elements, which are typically connected using a network. Each node connected to the said network typically is comprised of a network interface and the local data processing elements. The network interface receives data from the network, which is addressed to this particular node, and the network interface also injects the local results into the network. Data is typically routed through the network in packets; and the packets are routed by a plurality of routers, typically one router per node. The network, specifically the plurality of the network routers, ensures the movement of the injected packets between the connected nodes towards the desired packet destinations.

[0007] Typically, the node, which produces a data packet, specifies the desire destination of that packet by specifically providing a unique address of the said packet destination. Upon injection of such an attributed packet, the plurality of network routers make local routing decisions to incrementally reduce the distance of the packet to its destination by forwarding the packet to a connected node closer to the specified destination. This universal point-to-point style of communication is state of the art and used by most of today's implemented computer networks. The drawback of using addresses as part of the packet attributes is the limitation of the network scalability to the maximal number of addresses presentable with the bits dedicated to the packet address.

[0008] Furthermore, additional auxiliary networks have been used to implement special support for collective communication such as global broadcasts to all connected nodes (CM-5). These networks have typically the topology of a tree or a fat tree, since the tree topology provides the minimal distance between any two connected nodes and, thus, minimal communication latency.

[0009] There are several constraints imposed to particular nodes by the tree topology. For example, the dedicated root node splits the network into two domains, left and right. Traffic from one domain targeted to the other domain must go through the root node under any circumstances. A broken root-node, router and/or links, will render the entire network useless since no packets can be routed from the left to the right partition. In addition, leaf nodes in a tree network have only one connection to the network. If this link is broken, the entire network is also not functional anymore.

SUMMARY OF THE INVENTION

[0010] An object of this invention is to provide an improved method and system for routing data packets through multi-node computer networks.

[0011] Another object of the invention is to avoid using addresses to route data packets through computer networks by classifying the packets into a limited set of classes for which the packet behavior can be specified in detail on a per-node basis.

[0012] A further object of the present invention is to allow the nodes of a multi-node computer network to utilize an arbitrary number of links between nodes while still assuring deterministic packet routes through the network and thus allow for well-defined, well-behaving collective operations.

[0013] Another object of this invention is to route data packets through a multi-node computer network by employing a general address-less static routing method applicable to arbitrary network topologies, which allows to embed a plurality of virtual logical networks in one physical network.

[0014] A further object of the invention is to enable a multi-node computer system to define and process collective packet operations such as global packet reductions (global maximum or similar) on networks of arbitrary size and shape.

[0015] These and other objectives are attained with a method of and a system for routing data packets in a computer network having a multitude of nodes and a multitude of links connecting the nodes together, and wherein each data packet includes a class identifier. The method comprises the steps of, each node, for each of a defined set of data packets, checking or looking at the data packet to identify the class of the data packet, and routing the data packet from the node based on the identified class of the data packet.

[0016] The preferred embodiment of the invention, described in detail below, provides a method and apparatus for identifying collectives of packets among a plurality of packets on a network in a system of connected data processing elements which are connected using an arbitrary network topology. The preferred method yields local routing decisions as well as decisions whether collective packet reduction operations or other operations should be applied to the identified packet collective. The local result of the said collective packet operation is routed to the collective of connected nodes and/or locally received. The preferred embodiment of the invention allows the specification of packet data reductions among an arbitrary set of nodes, connected using an arbitrary interconnection topology. In addition to packet reductions, the invention also can be utilized to multi-/broadcast packets among one or more configurable sets of nodes in a network.

[0017] The preferred embodiment of the invention provides a number of important advantages. For instance, the invention avoids using addresses—and thus avoids their

associated limitations—by classifying packets into a limited set of classes for which the packet behavior can be specified in detail on a per-node basis using, for example, local class descriptor tables. Each class may have a virtually unlimited set of nodes participating.

[0018] In addition, this preferred embodiment allows nodes to utilize an arbitrary number of links between nodes while still assuring deterministic packet routes through the network and, thus, allows for well-defined, well-behaving collective operations. Since the effective topology is defined per packet class, the topology may be modified dynamically, for example to compensate for broken links or to extend or shrink the affected network partition. Changes of the logical network topology may also be transparent to the application.

[0019] Further, the preferred embodiment of the invention disclosed herein solves several problems of auxiliary networks by employing a general address-less static routing method applicable to arbitrary network topologies, which allows to embed a plurality of virtual logical networks in one physical network; for example tree networks with redundant links or irregular networks. The absence of source or target addresses allows the application of this invention to networks of any size and shape.

[0020] In addition, the invention, in its preferred embodiment, allows to define and process collective packet operations such as global packet reductions (global sum or global maximum or similar) on networks of arbitrary size and shape.

[0021] Further benefits and advantages of the invention will become apparent from a consideration of the following detailed description, given with reference to the accompanying drawings, which specify and show preferred embodiments of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

[0022] **FIG. 1** shows the general architecture of a network comprising a plurality of nodes connected by an interconnection network of the degree n.

[0023] **FIG. 2** describes the general structure of a single node of the network of **FIG. 1**.

[0024] **FIG. 3** depicts a sparsely connected network topology, suitable for the invention described herein.

[0025] **FIG. 4** shows a route descriptor table comprising n route descriptors describing the packet behavior for a node with four network links plus one local client.

[0026] **FIG. 5** shows one exemplary route descriptor configuration for class **0** of the two nodes A and B of the network shown in **FIG. 3**.

[0027] **FIG. 6** illustrates a data packet that may be used in the practice of this invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0028] The herein described invention solves the problem to describe packet routes of single packets and to define packet collectives for collective packet operations among a plurality of nodes connected by a network with arbitrary topology of degree n as shown at **10** in **FIG. 1**. Each node **12** itself comprises the network interface and the local client, which contains the processing elements for data processing of the received data and for injecting results of the local computation into the network.

[0029] The general structure of the network interface with four links, for a network of degree four, is also shown in **FIG. 2**. Each network link comprises a network receiver **14**, which receives packets from the network link and presents the packets to the arbiter **16**, which routes the packets, via sender **20**, towards the targets specified using the collective class routing method herein described. The network interface also includes a local client CPU and memory, represented at **22**.

[0030] In particular, the arbiter **16** first evaluates the packet header information, shown at **24** in **FIG. 6**, such as the class **26** of the packet and the specified packet opcode **30**. Using the class information, the arbiter **16** retrieves the appropriate route descriptor from the route descriptor table, shown at **32** in **FIG. 4**. The route descriptor table can be read and written and contains the specific description **34** of the packet behavior for all available packet classes for this particular node. The route descriptor table may be different for each node in the network, depending on the position of the node in the logical network structure and on the availability of physical connections to neighboring nodes.

[0031] The route descriptor table **32** may be initialized immediately after booting the local client. During runtime, packet classes may be allocated and initialized to implement a specific communication pattern such as packet broadcast or a packet reduction. From that time on, packets, injected with the appropriate class-tag, follow the configured packet routes according to the route descriptors deposited in the nodes along the packet path through the network. For collective packet operations such as reductions, the route descriptor **34** also indicates which packets are members of the collective. The router **36** will wait until the packet collective is complete and it will then forward the packet collective while applying the specified packet operation.

[0032] The collective is considered to be complete if all channels which are identified as source-channels have collective packets available. Whether a packet is considered a collective packet or not is specified on a per-packet basis using the packet opcode field **30** of the packet header **22**, shown in **FIG. 6**.

[0033] For the configuration shown in **FIG. 5**, for example, Node A considers a packet collective as complete if the receivers of channel **2**, channel **1**, and the local client signal the availability of a packet, which is marked with an opcode such as ADD or MAX, which in turn indicates a collective operation. In that case, the router applies the specified operation to the packet collective and routes the result of that operation to the sender of channel **0**.

[0034] Packets, not marked as operands of collective packet operations, are simply routed to the target channels without waiting for the other sources. In general the rule for forwarding packets is: if the packet enters the node from a channel which is marked as a source channel for the given packet class, then the packet or the packet collective is routed to the specified target channels of that class. If the channel is not an explicit source channel, then the packet is routed to all the source channels and no collective operation is applied even though the packet opcode may request a collective operation.

[0035] The apparatus enabling the method of collective class routing, the route descriptor table **32**, is an array of registers, which can be read and written, comprising at least two bits per potential packet route (channel). One bit indicates whether the particular channel is a dedicated source channel for that packet class. The other bit is set for all channels which are dedicated targets for packets of the given class. Thus the descriptor describes the packet routes and the packet collective for collective operations.

[0036] With reference to **FIG. 6**, each packet header comprises the packet class **26** and the packet opcode **30**. The class is used to identify the appropriate packet routes, and the opcode is used to decide whether collective operations should be applied. **FIG. 5** shows an example of a particular configuration for class **0** of the two nodes A and B of the network shown in **FIG. 3**. It should be noted that a number of configurations may coexist in parallel, each using a different class with different sources and targets specified.

[0037] While it is apparent that the invention herein disclosed is well calculated to fulfill the objects stated above, it will be appreciated that numerous modifications and embodiments may be devised by those skilled in the art, and it is intended that the appended claims cover all such modifications and embodiments as fall within the true spirit and scope of the present invention.

What is claimed is:

1. A method of routing data packets in a computer network having a multitude of nodes and a multitude of links connecting the nodes together, each data packet including a class identifier, the method comprising:

each node, for each of a defined set of data packets,

looking at the data packet to identify the class of the data packet; and

routing the data packet from the node based on the identified class of the data packet.

2. A method according to claim 1, wherein each node includes a route descriptor table including one or more route descriptors, each route descriptor (i) specifying a route from the node, and (ii) being associated with one of a plurality of packet classes, and wherein the routing step includes the steps of,

each node, for each of the defined set of data packets,

identifying the route descriptor, in the route descriptor table of the node, associated with the class of the data packet; and

routing the data packet on the route specified by the identified route descriptor.

3. A method according to claim 2, wherein each node includes one or more channels, and each route descriptor table includes an array of registers comprising at least two bits for each channel of the node in which the descriptor table is located.

4. A method according to claim 1, wherein at each node, the defined set of data packets includes data packets received at the node and data packets originated at the node.

5. A method as specified in claim 1, wherein an opcode associated with each packet or packet class indicates that the plurality of selected, specified input directions defines a local collective of packets which is subject to the said collective packet operation.

6. A method as specified in claim 5, wherein the identified collective of packets is subject to arithmetical/logical data processing operations to combine the members of the packet collective thereby reducing the number of packets to be routed through the network.

7. A method as specified in claim 5, wherein the identified collective of packets is subject to arithmetical/logical data processing operations to compute a number of resulting packets based on the data of the identified packet collective without reducing the number of packets to be routed through the network.

8. An apparatus for routing data packets in a computer network having a multitude of nodes and a multitude of links connecting the nodes together, each data packet including a class identifier, the apparatus comprising:

a plurality of checking means, each of the checking means being located at a respective one of the nodes for checking each of a defined set of data packets, to identify the class of the data packet; and

a plurality of routing means, each of the routing means being located at a respective one of the nodes to route data packets from the node based on the class of the data packets.

9. Apparatus according to claim 8, further comprising a plurality of route descriptor tables, each of the route descriptor tables being located at a respective one of the nodes, each route descriptor table including one or more route descriptors, each route descriptor (i) specifying a route in the network and (ii) being associated with one of a plurality of packet classes, and wherein each routing means includes:

means for identifying the route descriptor, in the route descriptor table at the node at which the routing means is located, associated with the identified class of the data packet; and

means for directing the data packet onto the route specified by the identified route descriptor.

10. Apparatus according to claim 9, wherein each node includes one or more channels, and each route descriptor table includes an array of registers comprising at least two bits for each channel of the node in which the descriptor table is located.

11. Apparatus according to claim 8, wherein, at each node, the defined set of data packets includes data packets received at the node and data packets originated at the node.

12. A method as specified in claim 8, wherein an opcode associated with each packet or packet class indicates that the plurality of selected, specified input directions defines a local collective of packets which is subject to the said collective packet operation.

13. A method as specified in claim 8, wherein in addition to packet routes also specifies a plurality of nodes on the network which participate in a particular collective operation, comprising two additional bits per local route descriptor; one bit corresponding to the local contribution of the actual node to the said collective operation and another bit correlating to the local reception of the results of the said operation.

14. The method specified in claim 13, wherein the packet class and opcode information is also subject to a collective operation such as but not limited to substituting the class or opcode with certain predefined values or increment/decrement operations.

15. The method specified in claim 14, wherein the packet class and/or opcode information is being utilized to apply pre-processing steps, such as reverting the word order, to the data locally injected and/or to apply post-processing steps to the said data.

16. A program storage device readable by machine, tangibly embodying a program of instructions executable by the machine to perform method steps for routing data packets in a computer network having a multitude of nodes and a multitude of links connecting the nodes together, each data packet including a class identifier, said method steps comprising:

each node, for each of a defined set of data packets,

looking at the data packet to identify the class of the data packet; and

routing the data packet from the node based on the identified class of the data packet.

17. A program storage device according to claim 16, wherein each node includes a route descriptor table including one or more route descriptors, each route descriptor (i) specifying a route from the node, and (ii) being associated with one of a plurality of packet classes, and wherein the routing step includes the steps of,

each node, for each of the defined set of data packets,

identifying the route descriptor, in the route descriptor table of the node, associated with the class of the data packet; and

routing the data packet on the route specified by the identified route descriptor.

18. A program storage device according to claim 17, wherein each node includes one or more channels, and each route descriptor table includes an array of registers comprising at least two bits for each channel of the node in which the descriptor table is located.

19. A program storage device according to claim 16, wherein at each node, the defined set of data packets includes data packets received at the node and data packets originated at the node.

20. A program storage device according to claim 16, wherein an opcode associated with each packet or packet class indicates that the plurality of selected, specified input directions defines a local collective of packets which is subject to the said collective packet operation.

21. A program storage device according to claim 20, wherein the identified collective of packets is subject to arithmetical/logical data processing operations to combine the members of the packet collective thereby reducing the number of packets to be routed through the network.

22. A method of identifying a collective of data packets on a computer system, the computer system including a multitude of interconnected processing nodes, and wherein a multitude of data packets are routed in the computer system, the method comprising the steps of:

allocating a class identifier to identify a given class of data packets;

providing each data packet in said given class with said class identifier;

providing each of the nodes with a set of channels for receiving and holding data packets; and

each of at least some of the nodes,

i) identifying a subset of the set of channels of the node,

ii) evaluating the data packets at the node to identify the class of the data packet, and

iii) identifying said collective as complete when all of the channels of said subset have a data packet of the given class.

23. A method according to claim 22, wherein each of said at least some of the nodes includes a route descriptor table identifying routes for data packets from said node, and further comprising the step of using the route descriptor table to identify a route for the collective of data packets from the node.

24. A method according to claim 22, further comprising the step of initializing the plurality of local route descriptors to reflect the local flow of data for the desired global operation associated with that class by selectively enabling the desired input and output directions.

25. A method according to claim 24, further comprising the steps of:

using the class identifier of the incoming packets to select one of the local route descriptors; and

using the selected route descriptor to determine the plurality of valid input and output directions.

26. A method according to claim 25, further comprising the step of comparing the incoming direction of the packet with the specified incoming directions of the packet class to determine whether the packet is to be routed to the specified output directions or to the specified input directions.

27. A method as specified in claim 26, wherein an opcode associated with each packet or packet class indicates that the plurality of selected, specified input directions defines a local collective of packets which is subject to the said collective packet operation.

28. A method a specified in claim 27, wherein the identified collective of packets is subject to arithmetical/logical data processing operations to combine the members of the packet collective thereby reducing the number of packets to be routed through the network.

* * * * *