(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 October 2007 (18.10.2007)

PCT

(10) International Publication Number
WO 2007/116391 A2

(51) International Patent Classification:
H04J 3/22 (2006.01)

(21) International Application Number:
PCT/IL2007/000394

(22) International Filing Date: 27 March 2007 (27.03.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
11/279,045 7 April 2006 (07.04.2006) US

(71) Applicant *(for all designated States except US)*: **CORRIGENT SYSTEMS LTD.** [IL/IL]; 126 Yigal Alon Street, 67443 Tel Aviv (IL).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **ZELIG, David** [IL/IL]; 1 Hadganim Street, 53482 Givatayim (IL). **SOLOMON, Ronen** [IL/IL]; 18/2 Bloch Street, 53229 Givatayim (IL). **KHILL, Uzi** [IL/IL]; 7a Tormos Street, Netanya (IL).

(74) Agents: **SANFORD T. COLB & CO.** et al.; P.O. Box 2273, 76122 Rehovot (IL).

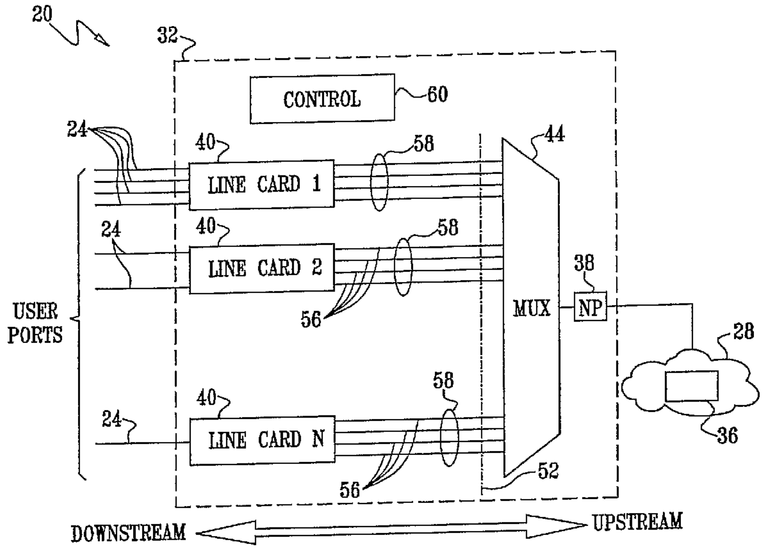
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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, MT, 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:
— *without international search report and to be republished upon receipt of that report*

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

(54) Title: TWO-WAY LINK AGGREGATION



(57) Abstract: A method for communication includes coupling a network node (36) to one or more interface modules (40) using a first group of first physical links (56) arranged in parallel. Each of the one or more interface modules is coupled to a communication network using a second group of second physical links (24) arranged in parallel. A data frame having frame attributes sent between the communication network and the network node is received. A first physical link out of the first group and a second physical link out of the second group are selected in a single computation based on at least one of the frame attributes. The data frame is sent over the selected first and second physical links. This method allows two or more link aggregation groups (58, 68) to be concatenated, using a single processing stage to determine port assignment for each frame in each of the link aggregation groups.

TWO-WAY LINK AGGREGATION

FIELD OF THE INVENTION

The present invention relates generally to communication networks, and particularly to methods and systems for link aggregation in network elements.

BACKGROUND OF THE INVENTION

Link aggregation (LAG) is a technique by which a group of parallel physical links between two endpoints in a data network can be joined together into a single logical link (referred to as the "LAG group"). Traffic transmitted between the endpoints is distributed among the physical links in a manner that is transparent to the clients that send and receive the traffic. For Ethernet™ networks, link aggregation is defined by Clause 43 of IEEE Standard 802.3ad, Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications (2002 Edition), which is incorporated herein by reference. Clause 43 defines a link aggregation protocol sub-layer, which interfaces between the standard Media Access Control (MAC) layer functions of the physical links in a link aggregation group and the MAC clients that transmit and receive traffic over the aggregated links. The link aggregation sub-layer comprises a distributor function, which distributes data frames submitted by MAC clients among the physical links in the group, and a collector function, which receives frames over the aggregated links and passes them to the appropriate MAC clients.

SUMMARY OF THE INVENTION

In various communication applications, users connect to a communication network through a network element, such as an access concentrator or aggregator for obtaining different data services.

Embodiments of the present invention that are described hereinbelow provide improved methods and systems for connecting users to a communication network with increased capacity and quality of service. The network element comprises one or more user interface modules (UIMs), each serving one or more user ports. In some embodiments, each UIM is connected to the communication network using two or more physical links arranged in parallel, in order to provide sufficient bandwidth at a given quality-of-service (QoS) level.

Upstream data frames sent from the user ports to the communication network and downstream data frames sent from the communication network to the user ports are distributed among the parallel physical links, so as to balance the traffic load among the links. The load balancing enables each UIM, and the network element as a whole, to deliver a higher bandwidth at a given QoS or to improve the QoS at a given bandwidth.

In some embodiments, the UIMs are coupled to a backplane of the network element, and the parallel physical links comprise backplane traces. In some embodiments, the physical links are configured as an Ethernet link aggregation (LAG) group. Distribution of frames to individual physical links typically comprises applying a suitable mapping function, such as a hashing function. The mapping function typically uses frame attributes, such as various header fields of the frame, to determine a physical link over which to send each frame.

Unlike some known network element configurations, in which each user port is fixedly assigned to a specific backplane trace, the load balancing operation in embodiments of the present invention enables statistical multiplexing of the frames, in which there is no direct relationship or connection between user ports and backplane traces.

In some embodiments, two or more physical user ports are aggregated into a LAG group external to the network element, so as to form an aggregated user port having a higher bandwidth. The frames to and from the aggregated user port are distributed among the physical user ports of the external LAG group to balance the load and enable higher bandwidth and higher QoS. In the downstream direction, a combined mapping operation for frames addressed to the aggregated user port determines an individual backplane trace and a physical user port over which to send each frame in a single mapping computation.

Several system configurations that implement the disclosed methods are described hereinbelow. Bandwidth allocation considerations for allocating sufficient bandwidth in the various resources of the network element are also described and demonstrated.

Although some of the embodiments described herein relate specifically to access concentrators, aspects of the present invention are also applicable to link aggregation performed by network elements of other sorts.

There is therefore provided, in accordance with an embodiment of the present invention, a method for communication, including:

coupling a network node to one or more interface modules using a first group of first physical links arranged in parallel;

coupling each of the one or more interface modules to a communication network using a second group of second physical links arranged in parallel;

receiving a data frame having frame attributes sent between the communication network and the network node;

selecting, in a single computation based on at least one of the frame attributes, a first physical link out of the first group and a second physical link out of the second group; and

sending the data frame over the selected first and second physical links.

In an embodiment, the network node includes a user node, and sending the data frame includes establishing a communication service between the user node and the communication network.

In another embodiment, the second physical links include backplane traces formed on a backplane to which the one or more interface modules are coupled.

In yet another embodiment, at least one of the first and second groups of physical links includes an Ethernet link aggregation (LAG) group.

In still another embodiment, coupling the network node to the one or more interface modules includes aggregating two or more of the first physical links into an external Ethernet LAG group so as to increase a data bandwidth provided to the network node.

In an embodiment, coupling each of the one or more interface modules to the communication network includes at least one of multiplexing upstream data frames sent from the network node to the communication network, and demultiplexing downstream data frames sent from the communication network to the network node.

In another embodiment, selecting the first and second physical links includes balancing a frame data rate among at least some of the first and second physical links.

In an embodiment, selecting the first and second physical links includes applying a mapping function to the at least one of the frame attributes. In another embodiment, applying the mapping function includes applying a hashing function. In still another embodiment, applying the hashing function includes determining a hashing size responsively to a number of at least some of the first and second physical links, applying the hashing function to the at least one of the frame attributes to produce a hashing key, calculating a modulo of a division operation of the hashing key by the hashing size, and selecting the first and second physical links responsively to the modulo. In an embodiment, selecting the first and second physical links responsively to the modulo includes selecting the first and second physical links responsively to respective first and second subsets of bits in a binary representation of the modulo.

In an embodiment, the at least one of the frame attributes includes at least one of a layer 2 header field, a layer 3 header field, a layer 4 header field, a source Internet Protocol (IP) address, a destination IP address, a source medium access control (MAC) address, a destination MAC address, a source Transmission Control Protocol (TCP) port and a destination TCP port.

In another embodiment, coupling the network node to the one or more interface modules and coupling each of the one or more interface modules to the communication network include specifying bandwidth requirements including at least one of a committed information rate (CIR), a peak information rate (PIR) and an excess information rate (EIR) of a communication service provided by the communication network to the network node, and allocating a bandwidth for the communication service over the first and second physical links responsively to the bandwidth requirements.

There is additionally provided, in accordance with an embodiment of the present invention, a method for connecting user ports to a communication network, including:

coupling the user ports to one or more user interface modules;

coupling each user interface module to the communication network via a backplane using two or more backplane traces arranged in parallel;

receiving data frames sent between the user ports and the communication network, the data frames having respective frame attributes;

for each data frame, selecting responsively to at least one of the respective frame attributes a backplane trace from the two or more backplane traces; and

sending the data frame over the selected backplane trace.

Apparatus for connecting a network node with a communication network and for connecting user ports to a communication network are also provided.

The present invention will be more fully understood from the following detailed description of the embodiments thereof, taken together with the drawings in which:

BRIEF DESCRIPTION OF THE DRAWINGS

Figs. 1 and 2 are block diagrams that schematically illustrate communication systems, in accordance with embodiments of the present invention;

Fig. 3 is a block diagram that schematically illustrates elements of a communication system, in accordance with an embodiment of the present invention;

Fig. 4 is a flow chart that schematically illustrates a method for single-stage hashing, in accordance with an embodiment of the present invention; and

Fig. 5 is a block diagram that schematically illustrates bandwidth allocation aspects of a communication system, in accordance with an embodiment of the present invention.

DETAILED DESCRIPTION OF EMBODIMENTS

SYSTEM DESCRIPTION

Fig. 1 is a block diagram that schematically illustrates a communication system 20, in accordance with an embodiment of the present invention. System 20 interconnects a plurality of user ports 24 to a communication network 28. Network 28 may comprise a wide-area network (WAN), such as the Internet, a network internal to a particular organization (Intranet), or any other suitable communication network.

A network element 32, such as an access concentrator, connects user ports 24 to a node 36 in network 28, typically via a network processor (NP) 38. Node 36 may comprise any suitable network element, such as a switch. The network element enables bi-directional communication in both upstream (i.e., user ports to network 28) and downstream (network 28 to user ports) directions.

In some embodiments, system 20 provides data services to users via network element 32. Typically, the system uses a Layer 2 communication protocol, such as an Ethernet™ communication protocol, in which data is transferred among the different system components using Ethernet frames. In some cases, services use higher level protocols, such as Multi-protocol label switching (MPLS). Alternatively, the Internet Protocol (IP) or other layer 2 or layer 3 protocols may also be used.

Network element 32 comprises one or more user interface modules (UIMs), such as line cards 40. Each line card is assigned to process data frames of one or more user ports. In some embodiments, the line cards are plugged into, mounted on, or otherwise coupled to a backplane 52, which distributes digital signals carrying the frames to and from line cards 40. Backplane 52 comprises physical links, such as backplane traces 56, typically in the form of printed circuit board (PCB) conductors. Each backplane trace 56 has a finite bandwidth and can support a certain maximum frame throughput.

A multiplexer (MUX) 44 is coupled to backplane traces 56. In the upstream direction, MUX 44 multiplexes upstream frames coming out of line cards 40 to produce an upstream output. The upstream output is provided to network processor 38. The upstream output is then sent by network processor 38 via a network connection 48 to network 28. In the downstream direction, network processor 38 of network element 32 accepts a downstream input comprising downstream frames addressed to user ports 24 from node 36 through connection 48. MUX 44 sends each frame to the appropriate user port via the appropriate line-card, using methods which will be explained below.

In many cases, a particular line card is connected to MUX 44 using two or more parallel backplane traces, in order to support the total bandwidth of the user ports assigned to this line card. In general, the statistical distribution of frames sent over different backplane traces may differ significantly from trace to trace, even for traces that belong to the same line card. In order to improve the capacity of the line cards, and consequently of network element 32 as a whole, it is desirable to balance the load of frames sent over the different traces, so as to avoid situations in which a certain backplane trace is overloaded, while excess capacity is available on a neighboring trace.

In many cases, the bandwidth offered by network element 32 to its users is specified in terms of quality of service (QoS) figures of merit, such as a guaranteed bandwidth (sometimes denoted CIR – Committed Information Rate) and a peak bandwidth (sometimes denoted PIR – Peak Information Rate). An alternative definition sometimes specifies the excess information rate (EIR), wherein CIR+EIR=PIR. Balancing the frame load between backplane traces 56 reduces the probability of lost frames due to overloading of backplane traces. The load balancing thus improves the QoS figures of merit, increases the quality of service offered by the concentrator and/or enables a higher bandwidth for a given QoS.

One way of balancing the load is to distribute the data rate of the Ethernet frames as uniformly as possible among the backplane traces. Alternatively, any other load balancing criterion can be used. In order to achieve load balancing, each group of backplane traces belonging to a particular line card is configured as an Ethernet link aggregation (LAG) group 58. Each LAG group 58 is considered by the relevant line card and by multiplexer 44 to be a single logical link having an aggregated bandwidth (i.e., capacity) equal to the sum of the bandwidths of the individual backplane traces in the group. As a result, there is no pre-assigned relationship or connection between any given user port 24 and a specific backplane trace 56, as in some conventional network elements. Ethernet frames are statistically multiplexed so as to balance the load among the backplane traces.

In some embodiments, Ethernet frames are mapped to individual backplane traces in the LAG group in accordance with a suitable mapping function, such as a hashing function. The mapping function distributes the frames among the different backplane traces so as to balance the load between the traces. In some embodiments, the mapping function hashes one or more frame attributes, such as header fields of the Ethernet frame, to produce a hashing key. The hashing key corresponds to an index of the backplane trace over which the frame is to be sent. Header fields may comprise any suitable layer 2, 3 or 4 headers of the Ethernet frame, such as source Internet Protocol (IP) address, destination IP address, source medium access control (MAC) address, destination MAC address, source Transmission Control Protocol (TCP) port and destination TCP port, MPLS label, virtual circuit (VC) label, virtual local area network identification (VLAN-ID) tag, as well as any other suitable header field.

In some embodiments, in the upstream direction, each line card 40 performs an upstream mapping of upstream frames to the individual backplane traces in its LAG group using a suitable mapping function. The mapping function typically uses attributes of the upstream frames, as described above.

In some embodiments, in the downstream direction, a control module 60 in network element 32 determines a downstream mapping of each downstream frame received through connection 48 to the appropriate backplane trace, using a suitable mapping function. The mapping performed by module 60 should naturally consider the user port to which the frame is addressed, in order to send the frame to the line card serving this user port. Module 60 controls multiplexer 44 in order to send each frame over the appropriate backplane trace. In addition to determining the appropriate line card for each frame (i.e., determining over which LAG group 58 to send the frame), the downstream mapping also balances the load of frames within each LAG group, responsively to attributes of the downstream frames.

For example, the downstream mapping can be implemented in two sequential stages. First, for each Ethernet frame, module 60 determines the appropriate line card to which the frame should be sent, depending on the destination user port. Then, within the backplane traces belonging to the LAG group of the appropriate line card, module 60 selects a particular trace and controls MUX 44 to send the frame over the selected trace. Alternatively, any other suitable mapping method can be used by module 60.

Control module 60 may be implemented in hardware, as software code running on a suitable processor, or as a combination of hardware and software elements. Module 60 may comprise an independent module, or be integrated with other components of network element 32. Different embodiments of network element 32 may comprise any number of line cards 40. Each line card may serve any number of user ports 24 and any number of backplane traces 56. In general, there need not be a direct relation between the number of user ports and the number of backplane traces. Each user port and each backplane trace may have any suitable bandwidth. User ports can have equal or different bandwidths. Similarly, the backplane traces within each LAG group 58 can have equal or different bandwidths.

Fig. 2 is a block diagram that schematically illustrates communication system 20, in accordance with another embodiment of the present invention. In some cases, a particular user requires a bandwidth higher than the bandwidth of a single user port 24. For this purpose, in the exemplary embodiment of Fig. 2, a number of user ports 24 are configured to form an aggregated user port 64. User ports 24 forming port 64 are configured as an Ethernet LAG group, referred to as an external LAG group 68. Thus, the bandwidth of port 64 is generally equal to the sum of bandwidths of the individual user ports 24 in external LAG group 68, in both upstream and downstream directions.

External LAG group 68 may comprise any number of user ports 24, belonging to any number of line cards 40. In general, a particular line card 40 may have some of its user ports 24 assigned to an external LAG group and other user ports 24 used individually or assigned to another external LAG group.

An external multiplexer (MUX) 72 performs the multiplexing and de-multiplexing (mapping) functions of the external LAG group. MUX 72 is external to network element 32 and is often located in user equipment 76, separate and distinct from network element 32. Aggregated port 64 is typically connected on the downstream side to a user node, such as a layer 2 or layer 3 switch (not shown).

In the downstream direction, MUX 72 multiplexes the downstream frames arriving over the user ports of external LAG group 68 to port 64. In the upstream direction, MUX 72 applies a suitable mapping function, such as a hashing function, to balance the load of upstream frames sent from port 64 over the different user ports of group 68. The mapping function uses attributes of the upstream frames, as explained above. Such load balancing helps to increase the upstream bandwidth of port 64 and/or its quality of service.

SINGLE-STAGE DOWNSTREAM MAPPING METHOD

In the system configuration of Fig. 2, consider the downstream frames arriving from network 28, via network processor 38, to multiplexer 44 and addressed to aggregated user port 64. In general, the processing of these frames in network element 32 comprises two consecutive mapping operations. First, each frame is mapped to one of user ports 24 in external LAG group 68. Determining the user port implicitly determines through which line card 40 the frame will pass. Then, the same frame is mapped to one of backplane traces 56 in the appropriate LAG group 58 that serves the selected line card. In some embodiments, in order to reduce the hardware complexity and/or computational complexity of network element 32, control module 60 performs a single combined mapping operation that combines the two mapping operations described above.

Thus, the combined mapping comprises a single hashing operation that determines, for each such downstream frame, both the backplane trace 56 over which the frame is to be sent to one of line cards 40, and the user port 24 to be used within external LAG group 68.

Fig. 3 is a block diagram that schematically illustrates elements of system 20, in accordance with an embodiment of the present invention. Fig. 3 is a simplified diagram, shown for the purpose of explaining the single-stage hashing method. As such, system elements unnecessary for explaining the method are omitted from the figure. The exemplary configuration of network element 32 in Fig. 3 comprises four line cards 40. Each line card is connected to MUX 44 using four backplane traces 56. The four backplane traces of each line card are configured as a LAG group. Each line card 40 serves a single user port 24. The four user ports are configured as an external LAG group to form aggregated user port 64, as explained above.

Fig. 4 is a flow chart that schematically illustrates a method for single-stage downstream hashing, in accordance with an embodiment of the present invention. Although the following description demonstrates the method using the simplified system configuration of Fig. 3 above, the method can also be applied in any other system configuration comprising two stages of link aggregation, such as the configurations discussed in the descriptions of Figs. 1 and 2 above.

The method begins with control module 60 determining a hashing size parameter denoted N_{bpow}, at a hash size definition step 80. N_{bpow} is defined as $N_{bpow}= N_{extp} \cdot N_{bpt}$, wherein N_{extp} denotes the number of user ports in external LAG group 68, and N_{bpt} denotes the number of backplane traces 56 in each LAG group 58. For simplicity, it is assumed that all line cards 40 having a user port in external LAG group 68 have the same number of backplane traces 56 connecting them to MUX 44. In the present example, $N_{extp}=4$, $N_{bpt}=4$, therefore $N_{bpow}=16$.

In different embodiments of network element 32, the value of hashing size N_{bpow} can be hard-wired in the concentrator design. Alternatively, the hashing size can be provided to the network element as part of its configuration, or determined by control module 60 responsively to the configuration setting of network element 32, as detected by module 60. In general, different line cards 40 may comprise different numbers of user ports belonging to external LAG group 68.

During normal operation, network element 32 receives downstream frames via network connection 36, at a frame reception step 82. For each downstream frame, control module 60 calculates a hashing key of the frame, at a hash key calculation step 84. As explained above, the hashing key is typically calculated by applying a suitable hashing function to the frame attributes of the downstream frame.

Module 60 performs an integer modulo-N_{bpow} division of the hashing key, at a mapping calculation step 86. Module 60 divides the hashing key of the downstream frame by N_{bpow} and retains the modulo, or the remainder of the division operation, as a mapping index. The mapping index can be represented as a binary number having a fixed number of bits, depending on the value of N_{bpow}. In the present example in which N_{bpow}=16, the mapping index, being a remainder after division by 16, is an integer number in the range 0...15, which can be represented using 4 bits.

Module 60 partitions the binary representation of the mapping index into two parts having N1 and N2 bits. Module 60 uses N1 bits as a user slot/port index, indicating over which user port 24 in external LAG group 68 the frame should be sent. The remaining N2 bits are used as a backplane trace index, indicating over which of the backplane traces of the relevant line card the frame should be sent. In the present example, the four bit mapping index is partitioned so that two bits encode the user port and two bits encode the backplane trace. In other words, N1=N2=2. The user port index and the backplane trace index jointly define the combined mapping operation for determining how the particular downstream frame is to be handled. Alternatively, module 60 can apply any other suitable method for determining the combined mapping responsively to the frame attributes.

Having determined the combined mapping, module 60 controls MUX 44 so as to send the downstream frame, at a frame sending step 88. MUX 44 sends the frame to the appropriate line card over the appropriate backplane trace, responsively to the user port index and the backplane trace index, respectively. In some embodiments, MUX 44 sends the user port index to the line card along with the frame. The line card then selects the user port over which to send the frame to aggregated port 64 responsively to the user port index.

After sending the frame over the appropriate backplane trace and user port, the method returns to reception step 82 to process the next downstream frame.

BANDWIDTH ALLOCATION CONSIDERATIONS

Fig. 5 is a block diagram that schematically illustrates bandwidth allocation aspects in communication system 20, in accordance with an embodiment of the present invention. In the exemplary configuration of Fig. 5, network element 32 comprises two line cards 100 and 101, each having a similar structure and functionality to line cards 40 of Figs. 1-3 above. Network element 32 connects network 28 with a user node, in the present example comprising a layer-2 switch 102. Port 64 is served by three user ports 24 configured as an external LAG group 68. As explained above, frames sent between switch 102 and network 28 undergo two stages of link aggregation. In addition to port 64, network element 32 also supports an independent user port 104 on line card 101. Frames sent between port 104 and network 28 undergo link aggregation only once.

Each of user ports 24 and 104 is coupled to a respective queue 106, which queues the frames of this port. In some embodiments, each port may have separate queues for upstream and downstream frames. Each line card comprises a link aggregator 108, which performs the aggregation of backplane traces 56 of this line card into the respective LAG group 58 in both upstream and downstream directions.

The configuration shown in Fig. 5 is an exemplary configuration chosen for the sake of simplicity. The bandwidth allocation calculations given below can be used in conjunction with any other suitable configuration of system 20, such as configurations having different numbers and arrangements of line cards, user ports, aggregated user ports and backplane traces.

Consider a particular communication service provided by network 28 to aggregated port 64. The bandwidth allocation of this communication service is commonly specified in terms of CIR and PIR or CIR and EIR figures-of-merit, as explained above. In order to provide these bandwidths at the specified quality-of-service, it is desirable to allocate sufficient bandwidth in the different physical resources used for transferring the frames. These resources comprise, for example, user ports 24, backplane traces 56 and different queues 106 in line cards 100 and 101. In some embodiments, a high quality of service is achieved by defining suitable bandwidth margins for the different physical resources. The following description details an exemplary calculation for allocating sufficient bandwidth to the different resources of system 20. Alternatively, any other suitable bandwidth allocation method can be used.

In the upstream direction, let $BW_{SERVICE}$ denote the total bandwidth of the communication service in question. ($BW_{SERVICE}$ may refer to the CIR, EIR or PIR of the service, as applicable. Similarly, the different bandwidth calculations given throughout the description below may be used to allocate CIR, EIR or PIR to the different system resources.) Upstream frames belonging to this service originating from switch 102 are mapped and distributed by MUX 72 among the three user ports 24 of external LAG group 68. In theory, the mapping operation should distribute the frames evenly among the user ports, so that each port receives a bandwidth of $BW_{SERVICE}/3$. In practice, however, the actual bandwidth distribution may deviate significantly from these values. In particular, deviations are likely to occur in embodiments in which the mapping operation is a data-dependent operation, such as hashing.

Therefore, in order to account for these deviations, each of the three user ports of external LAG group 68 is allocated a higher bandwidth given by:

[1] $BW_{ELAGPORT} = (BW_{SERVICE}/3) \cdot MARGIN_{UPELAG}$

wherein $MARGIN_{UPELAG}$ denotes an upstream bandwidth margin of each user port 24 in the external LAG group. (In other words, $MARGIN_{UPELAG}$=1.5 corresponds to a 50% bandwidth margin.) In general, the bandwidth allocated to a particular user port 24 should also be allocated to the respective queue 106 that queues the frames of this port.

After being mapped and sent over one of user ports 24, the upstream frames are processed by one of line cards 100 and 101. As part of this processing, the frames are mapped again by link aggregator 108 in the line card, so as to distribute them among the four backplane traces 56 of LAG group 56. It is thus desirable to allocate sufficient bandwidth on each of backplane traces 56. Assuming an optimal (uniform) distribution among the backplane traces, the bandwidth received by each backplane trace can be written as $BW_{ELAGPORT}$*#$PORTS_{ELAG}$/#BPT, wherein #$PORTS_{ELAG}$ denotes the number of user ports of external LAG group 68 that are processed by the particular line card, and #BPT denotes the number of backplane traces 56 in the LAG group 58 of this line card.

Since, as explained above, the actual distribution achieved by the mapping operation (in this case, the mapping between backplane traces performed by aggregator 108) often deviates from uniform distribution, a suitable margin denoted $MARGIN_{UPBPLAG}$ is added. Thus, the bandwidth allocation of each backplane trace can be written as $BW_{ELAGPORT}$*#$PORTS_{ELAG}$*$MARGIN_{UPBPLAG}$/#BPT, wherein #$PORTS_{ELAG}$ denotes the number of user ports in the particular line card. In some cases, however, this bandwidth allocation is greater than $BW_{SERVICE}$, the total bandwidth of the communication service.

Clearly there is no need to assign to any single backplane trace a bandwidth that is greater than the total service bandwidth. Therefore, the bandwidth allocated to each backplane trace can be written as:

$$[2]\quad BW_{BPT} = \mathrm{Min}\begin{Bmatrix} BW_{ELAGPORT} \cdot \#PORTS_{ELAG} \cdot \\ MARGIN_{UPBPLAG} /\#BPT, BW_{SERVICE} \end{Bmatrix}$$

Note that according to equation [1], $BW_{ELAGPORT}$ already contains $MARGIN_{UPELAG}$, the bandwidth margin of the external LAG. $\#PORTS_{ELAG}$ refers to the specific line card in question.

The total bandwidth allocated on LAG group 56 is thus given by

$$[3]\quad BW_{BPLAG} = \mathrm{Min}\begin{Bmatrix} BW_{ELAGPORT} \cdot \#PORTS_{ELAG} \cdot \\ MARGIN_{UPBPLAG} /\#BPT, BW_{SERVICE} \end{Bmatrix} \cdot \#BPT$$

In cases where external LAG is not used, such as for upstream frames originating from independent port 104, the total bandwidth allocated on LAG group 56 is given by the simpler expression:

$$[4]\quad BW_{BPLAG} = \mathrm{Min}\begin{Bmatrix} BW_{SERVICE} \cdot MARGIN_{UPBPLAG} /\#BPT \\ , BW_{SERVICE} \end{Bmatrix} \cdot \#BPT$$

In the downstream direction, as explained above, downstream frames undergo only a single mapping operation. When external LAG is used, using a similar calculation, the total bandwidth allocated to LAG group 56 for downstream frames of a particular service can be written as:

$$[5]\quad BW_{BPLAG} = \mathrm{Min}\begin{Bmatrix} BW_{ELAGPORT} \cdot \#PORTS_{ELAG} \cdot \\ /\#BPT, BW_{SERVICE} \end{Bmatrix} \cdot \#BPT$$

wherein a suitable downstream bandwidth margin is assumed to be already included in $BW_{ELAGPORT}$. When external LAG is not used, such as for downstream frames addressed to independent port 104, the total bandwidth allocation can be written as:

$$[6]\quad BW_{BPLAG} = \mathrm{Min}\begin{Bmatrix} BW_{SERVICE} \cdot MARGIN_{DNBPLAG} /\#BPT \\ , BW_{SERVICE} \end{Bmatrix} \cdot \#BPT$$

wherein $MARGIN_{DNBPLAG}$ denotes a downstream bandwidth margin for the backplane traces, which may be the same as or different from $MARGIN_{UPBPLAG}$.

Although the methods and systems described herein mainly address link aggregation of backplane traces and external link aggregation in Ethernet communication systems, the principles of the present invention can also be used in any system configuration in which a user interface module connects one or more user ports to a communication network via two or more parallel physical links. In particular, the principles of the present invention can also be used in other applications involving the selection of physical links in two successive link aggregation stages using a single computation.

It will thus be appreciated that the embodiments described above are cited by way of example, and that the present invention is not limited to what has been particularly shown and described hereinabove. Rather, the scope of the present invention includes both combinations and sub-combinations of the various features described hereinabove, as well as variations and modifications thereof which would occur to persons skilled in the art upon reading the foregoing description and which are not disclosed in the prior art.

CLAIMS

1. A method for communication, comprising:

coupling a network node to one or more interface modules using a first group of first physical links arranged in parallel;

coupling each of the one or more interface modules to a communication network using a second group of second physical links arranged in parallel;

receiving a data frame having frame attributes sent between the communication network and the network node;

selecting, in a single computation based on at least one of the frame attributes, a first physical link out of the first group and a second physical link out of the second group; and

sending the data frame over the selected first and second physical links.

2. The method according to claim 1, wherein the network node comprises a user node, and wherein sending the data frame comprises establishing a communication service between the user node and the communication network.

3. The method according to claim 1, wherein the second physical links comprise backplane traces formed on a backplane to which the one or more interface modules are coupled.

4. The method according to claim 1, wherein at least one of the first and second groups of physical links comprises an Ethernet link aggregation (LAG) group.

5. The method according to claim 1, wherein coupling the network node to the one or more interface modules comprises aggregating two or more of the first physical links into an external Ethernet LAG group so as to increase a data bandwidth provided to the network node.

6. The method according to claim 1, wherein coupling each of the one or more interface modules to the communication network comprises at least one of multiplexing upstream data frames sent from the network node to the communication network, and demultiplexing downstream data frames sent from the communication network to the network node.

7. The method according to claim 1, wherein selecting the first and second physical links comprises balancing a frame data rate among at least some of the first and second physical links.

8. The method according to any of claims 1-7, wherein selecting the first and second physical links comprises applying a mapping function to the at least one of the frame attributes.

9. The method according to claim 8, wherein applying the mapping function comprises applying a hashing function.

10. The method according to claim 9, wherein applying the hashing function comprises determining a hashing size responsively to a number of at least some of the first and second physical links, applying the hashing function to the at least one of the frame attributes to produce a hashing key, calculating a modulo of a division operation of the hashing key by the hashing size, and selecting the first and second physical links responsively to the modulo.

11. The method according to claim 10, wherein selecting the first and second physical links responsively to the modulo comprises selecting the first and second physical links responsively to respective first and second subsets of bits in a binary representation of the modulo.

12. The method according to any of claims 1-7, wherein the at least one of the frame attributes comprises at least one of a layer 2 header field, a layer 3 header field, a layer 4 header field, a source Internet Protocol (IP) address, a destination IP address, a source medium access control (MAC) address, a destination MAC address, a source Transmission Control Protocol (TCP) port and a destination TCP port.

13. The method according to any of claims 1-7, wherein coupling the network node to the one or more interface modules and coupling each of the one or more interface modules to the communication network comprise specifying bandwidth requirements comprising at least one of a committed information rate (CIR), a peak information rate (PIR) and an excess information rate (EIR) of a communication service provided by the communication network to the network node, and allocating a bandwidth for the communication service over the first and second physical links responsively to the bandwidth requirements.

14. A method for connecting user ports to a communication network, comprising:

coupling the user ports to one or more user interface modules;

coupling each user interface module to the communication network via a backplane using two or more backplane traces arranged in parallel;

receiving data frames sent between the user ports and the communication network, the data frames having respective frame attributes;

for each data frame, selecting responsively to at least one of the respective frame attributes a backplane trace from the two or more backplane traces; and

sending the data frame over the selected backplane trace.

15. The method according to claim 14, wherein at least some of the backplane traces are aggregated into an Ethernet link aggregation (LAG) group.

16. The method according to claim 14 or 15, wherein selecting the backplane trace comprises applying a hashing function to the at least one of the frame attributes.

17. Apparatus for connecting a network node with a communication network, comprising:

one or more interface modules, which are arranged to process data frames having frame attributes sent between the network node and the communication network;

a first group of first physical links arranged in parallel so as to couple the network node to the one or more interface modules;

a second group of second physical links arranged in parallel so as to couple the one or more interface modules to the communication network; and

a control module, which is arranged to select for each data frame sent between the communication network and the network node, in a single computation based on at least one of the frame attributes, a first physical link out of the first group and a second physical link out of the second group over which to send the data frame.

18. The apparatus according to claim 17, and comprising a backplane to which the one or more interface modules are coupled, wherein the second physical links comprise backplane traces formed on the backplane.

19. The apparatus according to claim 17, wherein at least one of the first and second groups of physical links comprises an Ethernet link aggregation (LAG) group.

20. The apparatus according to claim 17, wherein two or more of the first physical links are aggregated into an external Ethernet LAG group so as to increase a data bandwidth provided to the network node.

21. The apparatus according to claim 17, and comprising a multiplexer, which is arranged to perform at least one of multiplexing upstream data frames sent from the network node to the communication network, and demultiplexing downstream data frames sent from the communication network to the network node.

22. The apparatus according to claim 17, wherein the control module is arranged to balance a frame data rate among at least some of the first and second physical links.

23. The apparatus according to any of claims 17-22, wherein the control module is arranged to apply a mapping function to the at least one of the frame attributes so as to select the first and second physical links.

24. The apparatus according to claim 23, wherein the mapping function comprises a hashing function.

25. The apparatus according to claim 24, wherein the control module is arranged to determine a hashing size responsively to a number of at least some of the first and second physical links, to apply the hashing function to the at least one of the frame attributes to produce a hashing key, to calculate a modulo of a division operation of the hashing key by the hashing size, and to select the first and second physical links responsively to the modulo.

26. The apparatus according to claim 25, wherein the control module is arranged to select the first and second physical links responsively to respective first and second subsets of bits in a binary representation of the modulo.

27. The apparatus according to any of claims 17-22, wherein the at least one of the frame attributes comprises at least one of a layer 2 header field, a layer 3 header field, a layer 4 header field, a source Internet Protocol (IP) address, a destination IP address, a source medium access control (MAC) address, a destination MAC address, a source Transmission Control Protocol (TCP) port and a destination TCP port.

28. The apparatus according to any of claims 17-22, wherein the communication network is arranged to provide a communication service to the network node, the service having specified bandwidth requirements comprising at least one of a committed information rate (CIR), a peak information rate (PIR) and an excess information rate (EIR), and wherein the first and second groups of physical links are dimensioned to provide an allocated bandwidth for the communication service responsively to the bandwidth requirements.

29. Apparatus for connecting user ports to a communication network, comprising:

one or more user interface modules coupled to the user ports, which are arranged to process data frames having frame attributes sent between the user ports and the communication network;

a backplane having the one or more user interface modules coupled thereto and comprising a plurality of backplane traces arranged in parallel so as to transfer the data frames between the one or more user interface modules and the communication network;

a control module, which is arranged to select, for each data frame, responsively to at least one of the frame attributes, a backplane trace from the plurality of backplane traces over which to send the data frame.

30. The apparatus according to claim 29, wherein at least some of the backplane traces are aggregated into an Ethernet link aggregation (LAG) group.

31. The apparatus according to claim 29 or 30, wherein the control module is arranged to apply a hashing function to the at least one of the frame attributes so as to select the backplane trace.

FIG. 1

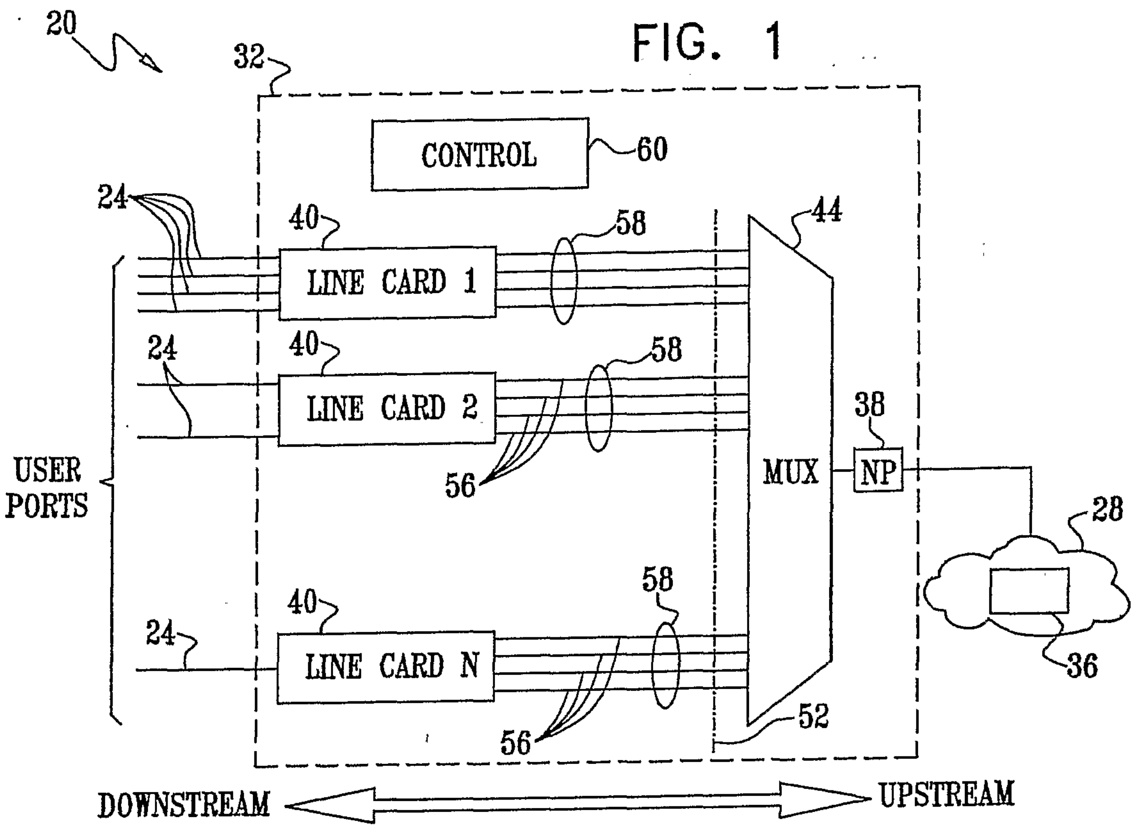

20
32
CONTROL
60
24
40
LINE CARD 1
58
44
40
24
LINE CARD 2
58
56
USER PORTS
MUX
38
NP
28
36
40
24
LINE CARD N
58
56
52
DOWNSTREAM
UPSTREAM

FIG. 2

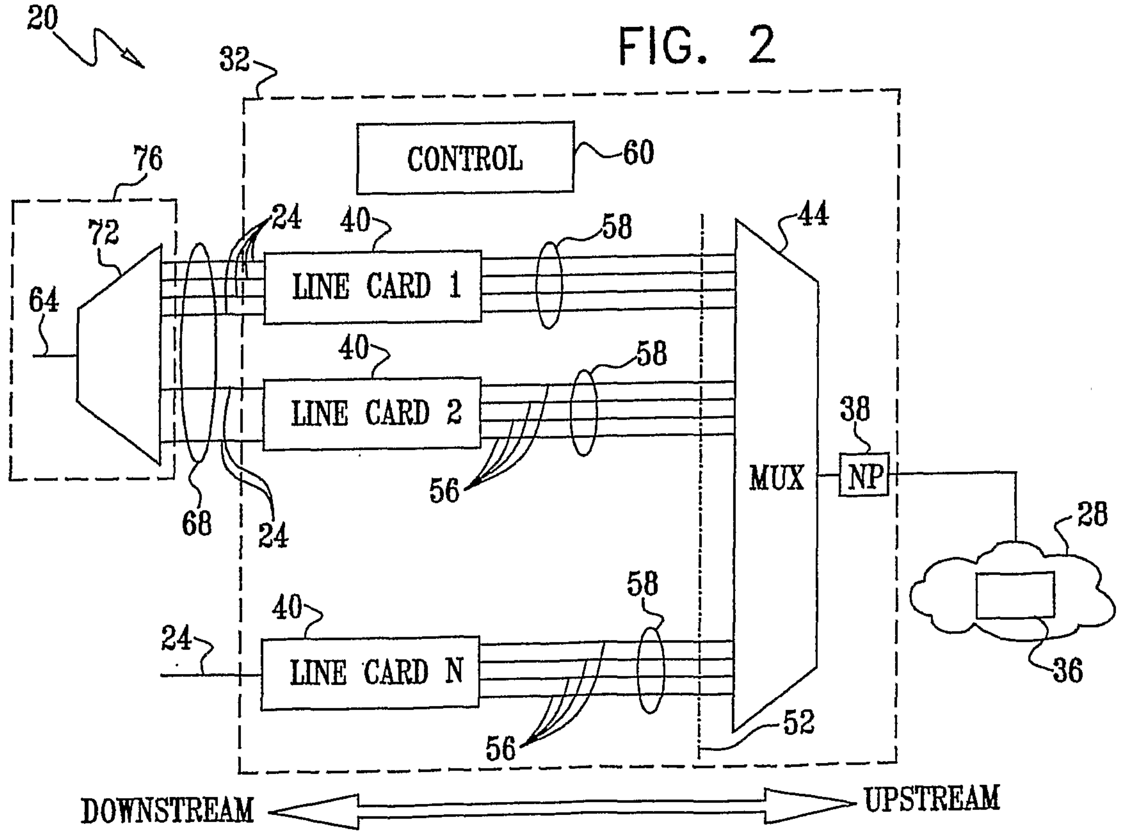

20
32
CONTROL
60
76
72
64
24
40
LINE CARD 1
58
44
40
LINE CARD 2
58
56
68
24
MUX
38
NP
28
36
40
24
LINE CARD N
58
56
52
DOWNSTREAM
UPSTREAM

FIG. 3

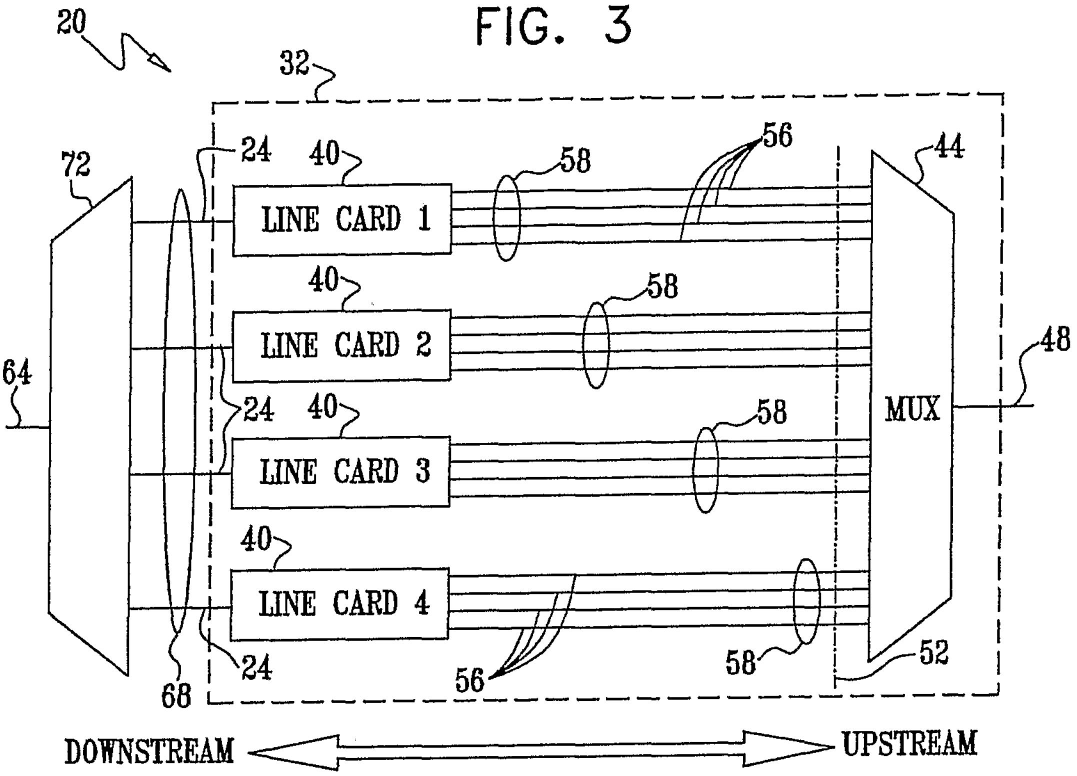

20
32
24
40
56
44
72
58
LINE CARD 1
40
58
LINE CARD 2
48
64
24
40
58
MUX
LINE CARD 3
40
LINE CARD 4
24
56
58
52
68
DOWNSTREAM
UPSTREAM

FIG. 4

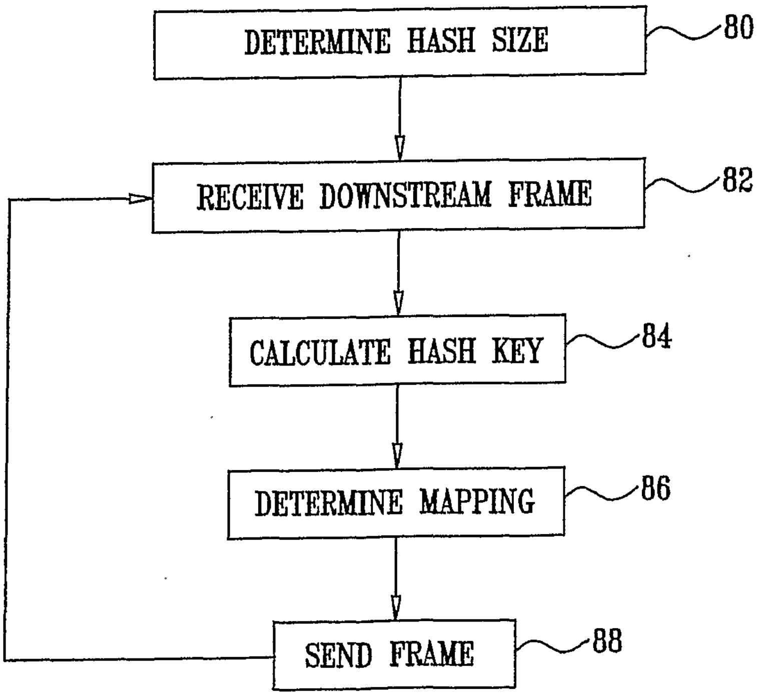
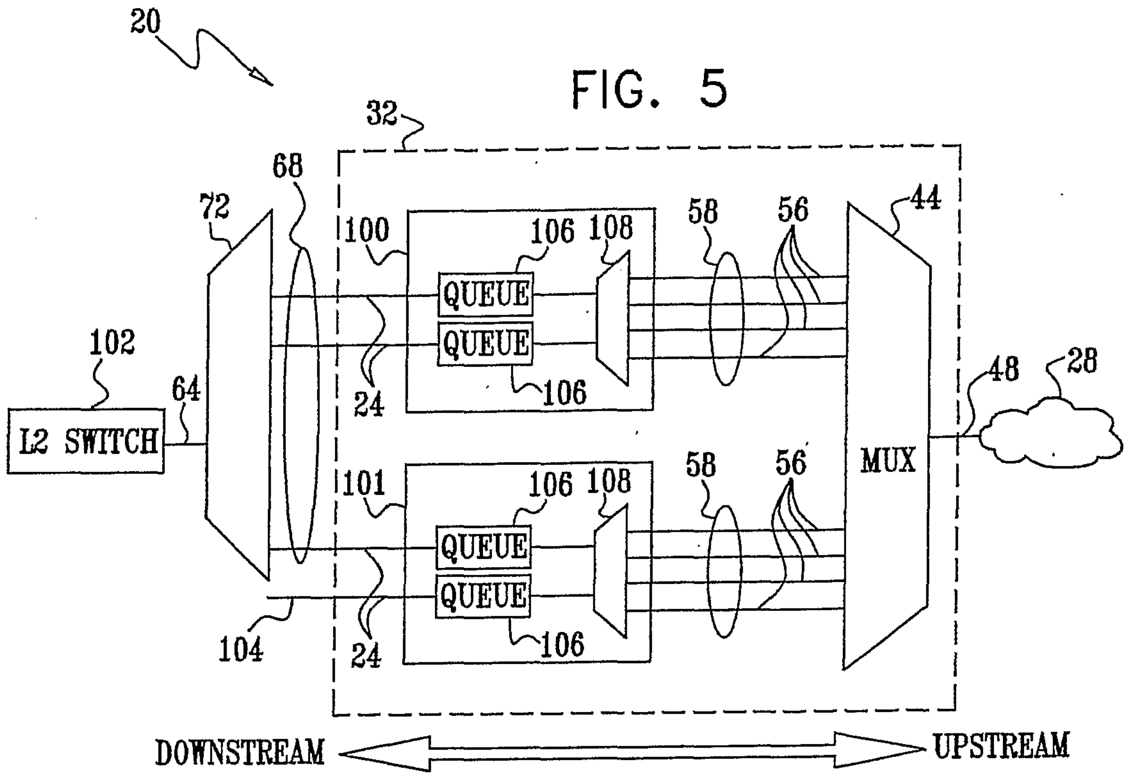

DETERMINE HASH SIZE
80
RECEIVE DOWNSTREAM FRAME
82
CALCULATE HASH KEY
84
DETERMINE MAPPING
86
SEND FRAME
88

FIG. 5


20
32
68
72
100
106
108
58
56
44
QUEUE
QUEUE
24
106
102
64
L2 SWITCH
48
28
MUX
101
106
108
58
56
QUEUE
QUEUE
104
24
106
DOWNSTREAM
UPSTREAM