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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
12 February 2009 (12.02.2009)

PCT

(10) International Publication Number
WO 2009/019258 A1

(51) International Patent Classification:
H04L 12/40 (2006.01) *H04L 12/56* (2006.01)

(21) International Application Number:
PCT/EP2008/060246

(22) International Filing Date: 4 August 2008 (04.08.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
07015525.4 7 August 2007 (07.08.2007) EP

(71) Applicant *(for all designated States except US)*: **NOKIA SIEMENS NETWORKS OY** [FI/FI]; Karaportti 3, FI-02610 Espoo (FI).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **RAMALHO RIBEIRO DOS SANTOS, José Miguel** [PT/PT]; Rua José Cardosa Pires n°1, 8°B., P-1750-430 Lisboa (PT). **DE FRIAS REBELO NUNES, Pedro Ricardo** [PT/PT]; Rua Antero de Figueredo, 7-5°A, P-2795016 Linda-A-Velha (PT).

(74) Common Representative: **NOKIA SIEMENS NETWORKS OY**; COO RTP IPR / Patent Administration, 80240 Munich (DE).

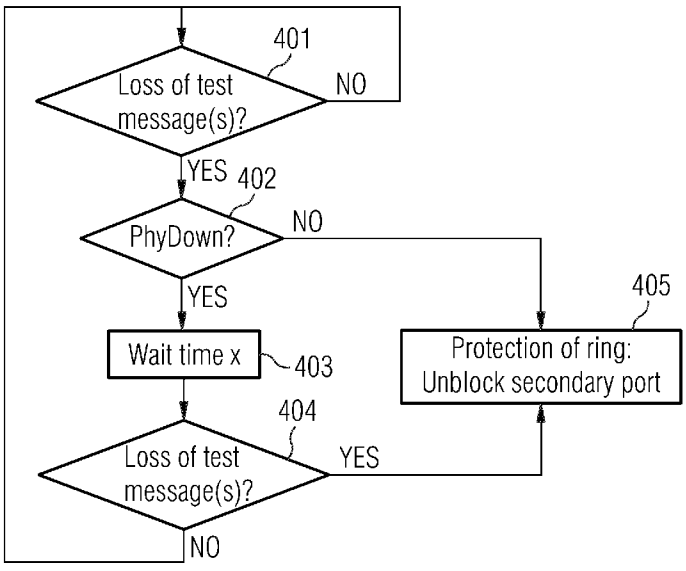
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, 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, ST, 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, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, 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*

(54) **Title:** METHOD, DEVICE AND COMMUNICATION SYSTEM TO AVOID LOOPS IN AN ETHERNET RING SYSTEM WITH AN UNDERLAYING 802.3ad NETWORK

FIG 4



(57) **Abstract:** A method is provided to be run in a network, the network comprising several network elements that are connected via a ring, wherein one of the network element is a ring master comprising a primary port and a secondary port. The method comprises the steps of (i) a failure is detected by the ring master; and (ii) the ring master checks for a second message and based on the content of the second message unblocks the secondary port. Also an associated device as well as a communication system comprising such device are provided.

WO 2009/019258 A1

Description

METHOD, DEVICE AND COMMUNICATION SYSTEM TO AVOID LOOPS IN AN ETHERNET RING SYSTEM WITH AN UNDERLAYING 802.3AD NETWORK

The invention relates to a method to be run in a network and to an associated device as well as to a communication system comprising said device.

An Ethernet Ring Protection (ERP) mechanism and protocol are disclosed in, e.g., EP 1 062 787 B1. In addition, there exists another ring protection mechanism called Ethernet Automatic Protection Switching (EAPS) as described in, e.g., IETF RRC3619.

Such ring protection mechanisms comprise a ring master RM (also referred to as a redundancy manager) to coordinate ring protection activities.

Protection in this sense means in particular that a link-layer loop in a physical Ethernet is avoided. The ring master is equipped to prevent the ring from forming such Ethernet loops.

When the ring master is notified that the ring is healthy (e.g., via test packets that are sent by the ring master via both of its ports), i.e. all ring nodes (network elements) and links (segments or arcs) are operational, the ring master breaks the link-layer loop by blocking traffic reception and transmission at one of its ring ports (the ring master's secondary port).

All traffic is blocked at that secondary port except for Ethernet ring protection control traffic, e.g., test packets. Preferably, such control traffic is sent via a control virtual LAN (VLAN).

From a link-layer's perspective, blocking traffic at the ring master's secondary port transforms the ring's topology into a chain of nodes (network elements). This is necessary in typical layer 2 (L2) networks (see also document IEEE 802.1 for further explanation). The ring master blocking its secondary port resulting in a topology of a chain of network elements is considered a normal operational state of the Ethernet Ring Protection mechanism.

Fig.1 shows such an ERP structure. The ring comprises network elements or nodes 101 to 106, wherein the node 101 is a Ring Master RM (also referred to as redundancy manager) with a primary port P and a secondary port S. As stated before, in normal operation, the Ring Master blocks its secondary port S resulting in the nodes 101 to 106 building a chain topology for the user traffic.

Link or Port Failure:

When a failure emerges in the ring, e.g., a link failure of a ring segment, the Ring Master unblocks its secondary port S thereby reestablishing communication between all ring nodes.

The failure can be directly detected by the Ring Master itself if the failure occurs at one of its ports. Alternatively, the Ring Master can be notified by another network element of the ring about a failure detected at one of the network element's ports. In such case, the affected network element sends a Link Down message to the Ring Master. The Ring Master subsequently unblocks its secondary port S (see **Fig.2**).

Failure Recovery:

When a network element of the ring detects that a failure recovered, it sends a notification to the Ring Master indicating that the link or port is operative again. This can be achieved by the network element sending a Link Up message

to the Ring Master. The network element will switch over to a pre-forwarding state blocking all traffic except test packets (health-check messages conveyed via the VLAN). In this pre-forwarding state the network element waits for a message from the Ring Master to switch over to normal operation (or forwarding state) again.

The Ring Master blocks the secondary port S again and sends the message to the network element to get back to normal operation. The Ring Master allows the network element to migrate from its pre-forwarding state to normal operation (forwarding state) only after the Ring Master blocked its secondary port S. This avoids configuration of a link-layer loop.

Preferably, the Ring Master assesses the operational state of the whole ring by frequently sending heath-check packets via both of its ring interfaces, i.e. via its primary port P and its secondary port S. These health-check packets (also referred to as test packets) may be conveyed via a control VLAN. If the ring is operational, the Ring Master receives its test packets sent via the respective other interface. If the test packets are not received, the ring may be broken and protection recovery actions should be initiated.

Link Aggregation:

One basic element in an Ethernet network is a Link Aggregation (LAG). LAG is defined in, e.g., document IEEE 802.3ad.

An optional Link Aggregation sublayer for use with CSMA/CD MACs is defined in IEEE 802.3ad. Link Aggregation allows one or more links to be aggregated together to form a Link Aggregation Group, such that a MAC Client can treat the Link Aggregation Group as if it were a single link. To this end, it specifies the establishment of DTE to DTE logical links,

consisting of N parallel instances of full duplex point-to-point links operating at the same data rate.

Link Aggregation comprises an optional sublayer between a MAC Client and the MAC (or optional MAC Control sublayer).

It is possible to implement the optional Link Aggregation sublayer for some ports within a system while not implementing it for other ports; i.e., it is not necessary for all ports in a system to be subject to Link Aggregation.

Since Ethernet link bandwidth usually increases by multiplication of 10 (e.g., 10Mbps, 100Mbps, 1Gbps, etc.), LAG defines how to aggregate several (n) Ethernet links, all of the same rate, to a larger link with a bandwidth amounting to

n*{single link rate}

(e.g. 8 links of 100Mbps create a LAG with 800Mbps).

An important aspect of LAG is its protection: If one of the physical links composing a LAG fails, the traffic can still be conveyed by the remaining links of the LAG and hence the traffic through the LAG is kept up.

In Ethernet Link Aggregation (LAG) according to IEEE 802.3, several Ethernet physical interfaces are combined into one single logical interface. To the Ethernet client layer, only one "logical" interface is presented. This mechanism is used to, e.g., increase bandwidth between two nodes or to allow load-sharing between several physical links. The link aggregation only fails when all physical links fail. As long as there exists at least one link that is operative, the traffic that was transmitted or received can be redirected to be transmitted or received over still operational physical link(s) that belong to the same link aggregation. This redirection operation may take one second to be completed.

When Ethernet Ring Protection is applied over Link Aggregation and one physical link fails, as stated before, the Link Aggregation will still be operational, also from the Ethernet Ring Protection point of view. However, if the failed physical link was chosen to convey ERP test packets, these packets will be lost, the ring master will assume that the ring is not operational and will initiate protection measures thereby unblocking its previously blocked (secondary) ring port. This action will immediately create a loop jeopardizing all communication in the ring network.

The **object** to be solved is to overcome the disadvantage as stated before and to provide an approach that is capable of handling ring networks and link aggregation thereby avoiding the creation of (temporary) loops within the ring network topology.

This problem is solved according to the features of the independent claims. Further embodiments result from the depending claims.

In order to overcome this problem a method is provided that can be run in a network or on a network component, in particular a ring master of a ring network. The network comprises several network elements (also referred to as nodes) that are connected via a ring, wherein one of the network element is a ring master comprising a primary port and a secondary port. The method comprises the steps:

- a failure is detected by the ring master;
- the ring master checks for a second message and based on the content of the second message unblocks the secondary port.

As the ring master checks and/or in particular waits for the second message it can be detected whether or not the secondary port of the ring master can be unblocked without creating a loop in the ring network.

In an embodiment, the ring master unblocks its secondary port if the second message indicates that there is no broken link within a link aggregation.

Hence, if there is no broken link in the link aggregation, the ring master can immediately unblock its secondary port in order to maintain traffic flow throughout the ring network.

The second message can be a PhyDown message (indicating that a physical layer is broken-down) of the link aggregation.

It is also an embodiment that the link aggregation covers at least one segment of the network.

According to a further embodiment, the link aggregation comprises at least two links in parallel, wherein upon failure of one link of the link aggregation the remaining at least one link conveys traffic that was destined to be transmitted via the failed link.

In yet an embodiment, the failure is detected by the ring master if at least one first message does not arrive at the primary port or at the secondary port of the ring master.

In another embodiment, the at least one first message is a test message (also referred to as a health check message).

In a further embodiment, the at least one first message is sent by the ring master.

Hence, the ring master can send a test message via at least one of its ports and receive it at the respective other port after a delay (time for the signal to be conveyed through the ring network). Hence, if the test message does not arrive within a predetermined period of time, the ring master may notify a failure.

However, in a next embodiment the failure corresponds to a loss of at least one first message, in particular to at least one test message.

Hence, if one or a predetermined number of test package(s) do(es) not arrive at the ring master within a certain time, the ring master assumes a failure that in this case corresponds to a loss of test packages (i.e., a loss of at least one first message).

It is an embodiment that the at least one first message is sent by the ring master via its primary port and via its secondary port.

In a further embodiment, the at least one first message is sent via a control virtual local area network (control VLAN).

In another embodiment the method comprises the steps:

- if the second message indicates that there is a broken link within the link aggregation the ring master waits a predetermined period of time;
- if the failure persists after the predetermined period of time, the ring master unblocks its secondary port.

Hence the second message indicates that the link aggregation has lost at least one link. However, there may be links remaining within this particular link aggregation that may still be able to convey the traffic. Usually, link aggregation needs some time until the still active links are able to convey the traffic of the failed link. After such period of time, the traffic is going to flow normally again. In such case it would have been fatal for the ring master to unblock its secondary port, because due to the recovery of the link aggregation this would have lead to a loop within the ring network.

However, if the period of time that is usually necessary for the link aggregation to redistribute the traffic among the still active links will be over and still there will be no at least one first message received at a port of the ring master, this is a very strong indication that the whole link aggregation (not only a single link within the link aggregation) may be down due to a severe link failure. In such case, the ring master can initiate protection thereby unblocking its secondary port.

If the time for the link aggregation to reconfigure passes and the at least one first test message is received at a port of the ring master (again), the ring master does not unblock its secondary port. In this case, waiting for that predetermined period of time avoided a precipitate unblocking of the ring master's secondary port that would have led to a loop in the ring network.

The problem stated supra is also solved by a device comprising a processor unit that is arranged and/or equipped such that the method as described herein is executable on said processor.

In an embodiment, said device is a communication device, in particular a network element or a ring master.

Also, the problem stated above can be solved by a communication system comprising the device as described herein.

Embodiments of the invention are shown and illustrated in the following figures:

Fig.3 shows a ring network comprising network elements and a link aggregation that connects two network elements of the ring network;

Fig.4 shows a flow chart illustrating the steps of a ring master to perform to check whether or not it can initiated ring protection in a ring network comprising a link aggregation.

Fig.3 shows a ring network comprising network elements (or nodes) 301 to 304, wherein node 301 is a Ring Master of this ring network. The Ring Mater 301 comprises a primary port P and a secondary port S. In normal operation, the secondary port S of the Ring Master 301 is blocked.

Network element 303 and network element 304 are connected via a link aggregation 308 comprising several links. In case of a failure 307 at the link aggregation 308, the Ring Master 301 has to decide whether the whole link aggregation 308 is broken down or only a single link within the link aggregation is defective and hence the link aggregation 308 will recover after a predetermined period of time. In the latter case in order to avoid a loop in the ring network, the Ring Master 301 should not initiate link protection, i.e. it should not unblock its secondary port S.

The Ring Master 301 sends test messages (also referred to as health check messages) via its primary port P and via its secondary port S. Such test messages are preferably sent via a control VLAN. By receiving the test messages at its respective port, the Ring Master is aware of the health of the ring. If test messages are missing, a link failure may have occurred.

In case test messages are not longer received at the port(s) of the Ring Master 301 and in case no Link-Down message is received, the physical interfaces of the network elements within the ring network may be operational. However, there may be a problem on a different OSI layer within the ring network, e.g. a physical link failure in a path using link aggregation may have occurred.

If a single link within a link aggregation is broken down, such failure should not be corrected by an Ethernet ring protection mechanism. Such protection should only be provided if all links within the link aggregation are defective.

If not all links of a link aggregation fail, the traffic can be redirected over the still operational physical links of this link aggregation. This redirection operation takes some time, e.g., one second, to be completed.

If the test messages were conveyed via a particular link of the link aggregation that became defective, the Ring Master 301 will not receive these test messages until link aggregation mechanisms finish redirections. In order not to unblock its secondary port S, the Ring Master 301 has to be informed that there is a chance for a link aggregation to recover the failure within a predetermined period of time.

In Fig.3, the network elements 303 and 304 each has ports that are configured with link aggregation 308. If one of the network elements 303 or 304 recognizes a failure at the link aggregation, it sends a PhyDown message to the Ring Master 301 via its opposite interface/port.

Due to missing test messages the Ring Master 301 can determine whether or not the ring network is healthy. The Ring Master 301 shall act as follows:

a) Ring network is not healthy AND PhyDown message received: There is a failure that may be protected by the link aggregation 308 itself. The Ring Master 301 temporarily suppresses protection actions.

b) Ring network is not healthy AND No PhyDown message received: There is no failure of or within the link aggregation 308, the Ring Master should initiate protection actions, i.e. unblock its secondary port.

c) Ring network is healthy AND PhyDown message received: There is a failure that is protected by the link aggregation 308, the Ring Master 308 does not have to protect the ring network.

Fig.4 shows a flow chart illustrating various states as well as necessary actions depending on certain messages that result from failures within a ring network.

In a step 401 the Ring Master may detect a failure in the ring due to a loss of test messages and switch to a step 402 thereby checking whether a PhyDown message has been received. If no PhyDown message has been received, the failure of a link cannot be restored by the link aggregation. Hence, the Ring Master has to protect the ring thereby unblocking its secondary port (step 405).

If the Ring Master receives a PhyDown message it branches to a step 403 waiting for a predetermined time x (required by the link aggregation to distribute the traffic among its still operative links). After this time x the Ring Master checks whether test messages are (still) missing. If this is the case, it is branched to the step 405 thereby initiating protection of the ring and unblocking the Ring Master's secondary port.

If no test messages are lost, the link aggregation has recovered and there is no need to protect the ring. Hence, it is branched to step 401.

Advantages of the approach provided herewith:

One advantage of this solution is that the ring master determines whether a link failure is related to a single link within a link aggregation and hence it can be avoided to immediately initiate a protection of the ring thereby creating a loop within the ring network.

The time period the ring master has to wait (see step 403 in Fig.4) should be long enough to ensure that the traffic in the link aggregation can be redistributed among the operational links and that test messages can be received at the port(s) of the ring manager.

Claims:

1. A method to be run in a network
   - wherein the network comprises several network elements that are connected via a ring;
   - wherein one of the network elements is a ring master comprising a primary port and a secondary port;

   comprising the steps:
   - a failure is detected by the ring master;
   - the ring master checks for a second message and based on the content of the second message unblocks the secondary port.

2. The method according to claim 1, wherein the ring master unblocks its secondary port if the second message indicates that there is no broken link within a link aggregation.

3. The method according to claim 2, wherein the link aggregation covers at least one segment of the network.

4. The method according to any of claims 2 or 3, wherein the link aggregation comprises at least two links in parallel, wherein upon failure of one link the remaining at least one link conveys traffic that was destined to be transmitted via the failed link.

5. The method according to any of the preceding claims, wherein the failure is detected by the ring master if at least one first message does not arrive at the primary port or at the secondary port of the ring master.

6. The method according claim 5, wherein the at least one first message is a test message.

7. The method according to any of claims 5 or 6, wherein the at least one first message is sent by the ring master.

8. The method according to any of claims 5 to 7, wherein the at least one first message is sent by the ring master via its primary port and via its secondary port.

9. The method according to any claims 5 to 8, wherein the at least one first message is sent via a control virtual local area network (VLAN).

10. The method according to any of the preceding claims comprising the steps:
- if the second message indicates that there is a broken link within the link aggregation the ring master waits a predetermined period of time;
- if the failure persists after the predetermined period of time, the ring master unblocks its secondary port.

11. The method according to claim 10 comprising the step:
- if there is no failure after the predetermined period of time, the ring master maintains the secondary port blocked.

12. A device comprising a processor unit that is arranged such that the method according of any of the preceding claims is executable on said processor.

13. The device according to claim 12, wherein said device is a communication device, in particular a network element or a ring master.

14. Communication system comprising the device according to any of claims 12 to 13.

FIG 1

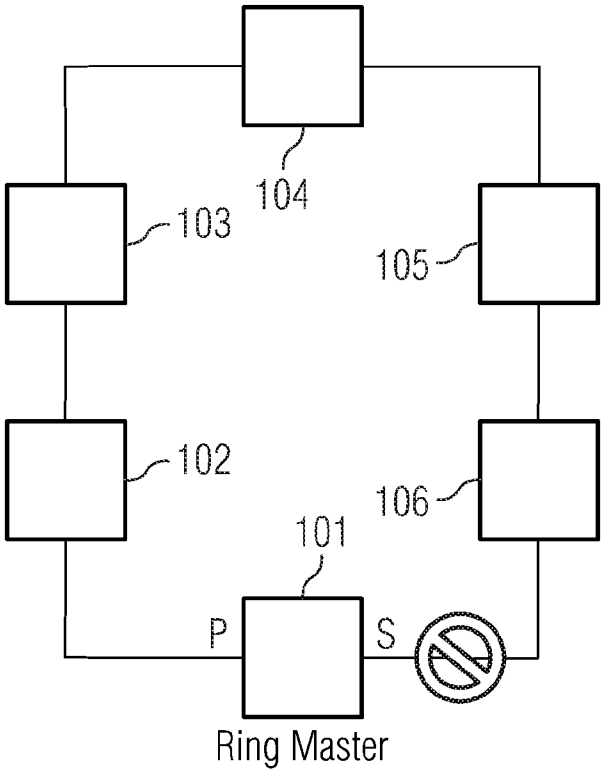

104
103
105
102
106
101
P
S
Ring Master

FIG 2

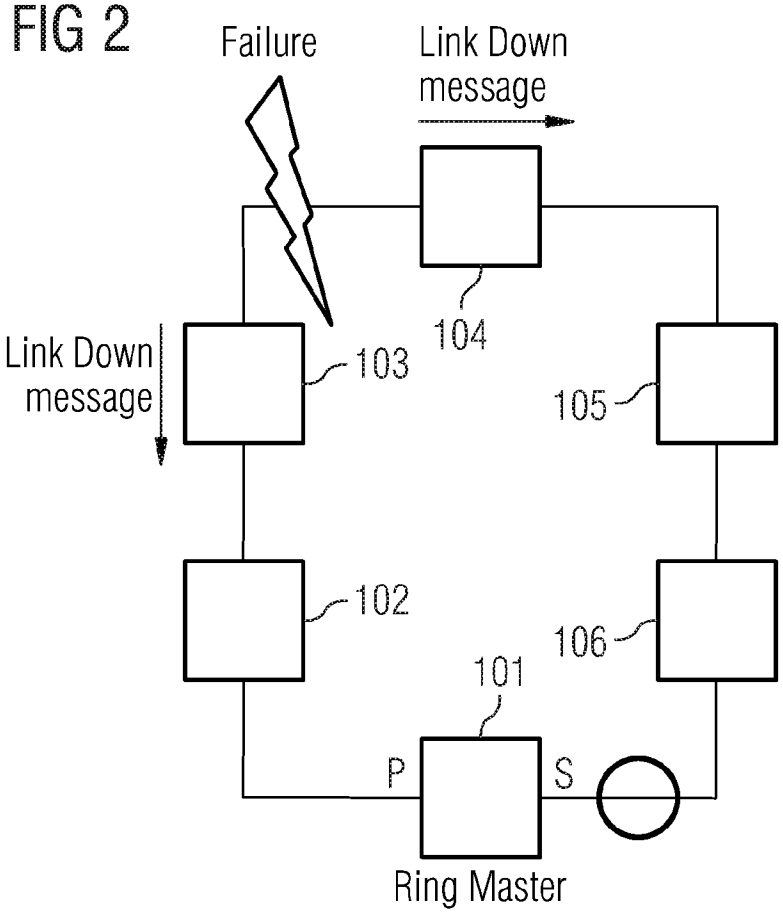

Failure
Link Down message
104
Link Down message
103
105
102
106
101
P
S
Ring Master

FIG 3

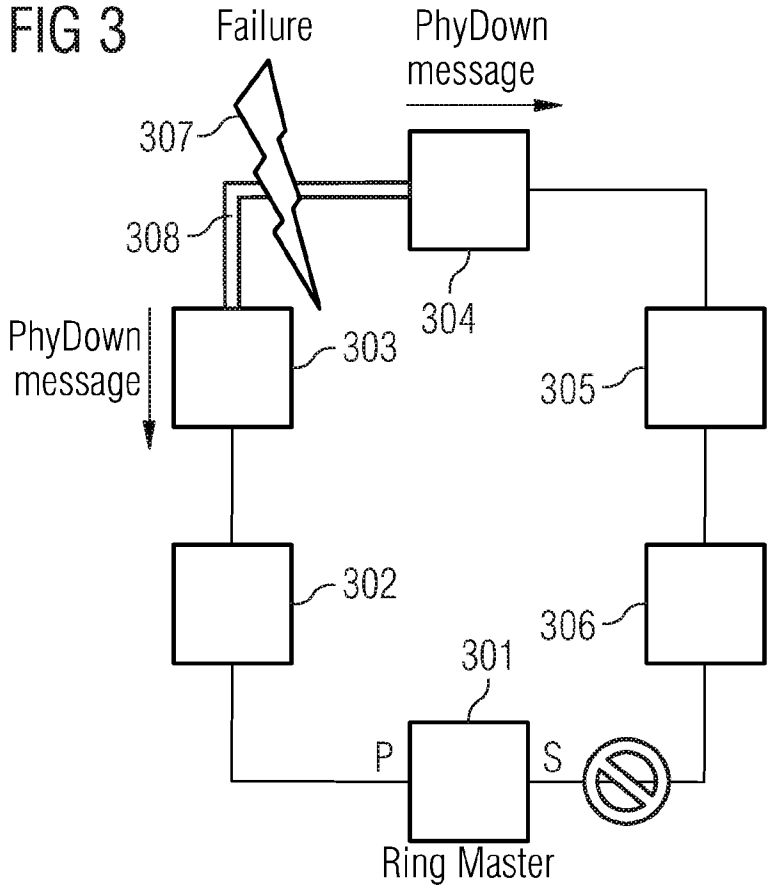
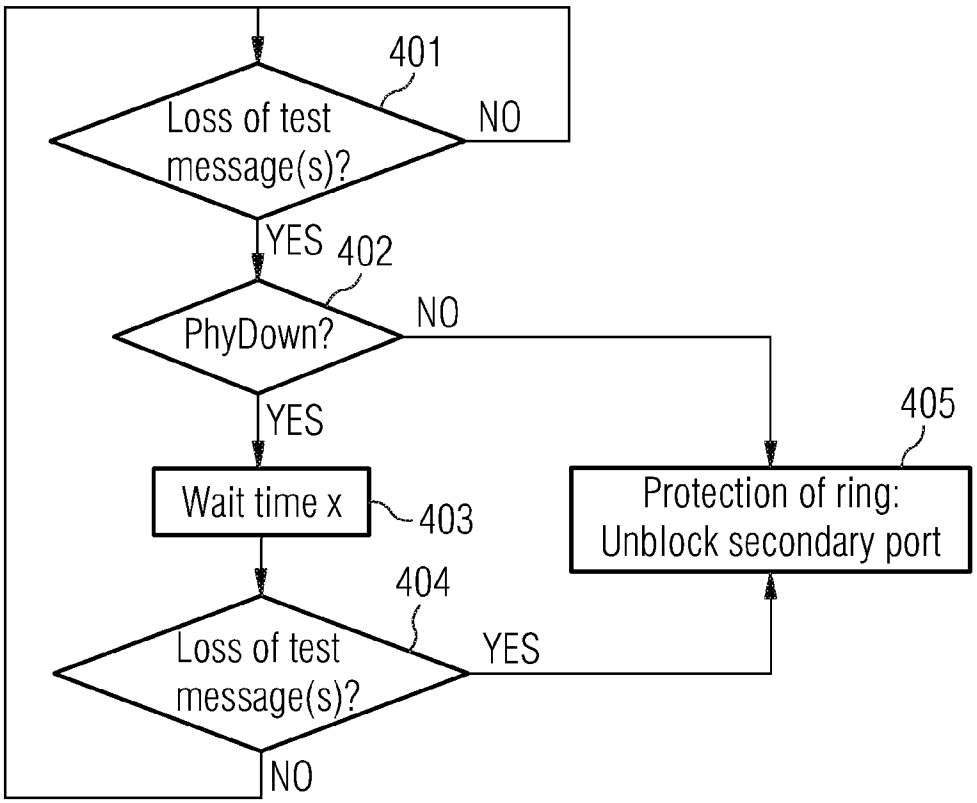

Failure
PhyDown message
307
308
304
PhyDown message
303
305
302
306
301
P
S
Ring Master

FIG 4


401
Loss of test message(s)?
NO
YES
402
PhyDown?
NO
YES
405
Wait time x
403
Protection of ring: Unblock secondary port
404
Loss of test message(s)?
YES
NO

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2008/060246

A. CLASSIFICATION OF SUBJECT MATTER
INV. H04L12/40 H04L12/56

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 H04J

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, PAJ, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	EP 1 062 787 A (SIEMENS AG [DE]; HIRSCHMANN RICHARD GMBH CO [DE] SIEMENS AG [DE]; HIRS) 27 December 2000 (2000-12-27) paragraphs [0004] - [0013]	1-14
A	IEEE802.3AD: "Amendment to Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications - Aggregation of Multiple Link Segments" 30 March 2000 (2000-03-30), IEEE , PAGES 1, 95-96, 108,114 , XP002465332 the whole document	1-14

☐ 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.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
22 September 2008	26/09/2008

Name and mailing address of the ISA/	Authorized officer
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	Siebel, Christian

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2008/060246

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 1062787 A	27-12-2000	AT 248476 T	15-09-2003
		CA 2323429 A1	16-09-1999
		CN 1292193 A	18-04-2001
		DE 19810587 A1	16-09-1999
		DK 1062787 T3	22-12-2003
		WO 9946908 A1	16-09-1999
		ES 2207180 T3	16-05-2004
		NO 20004478 A	10-11-2000
		US 6430151 B1	06-08-2002