DISTRIBUTED BACKBONES WITH GIGABIT ETHERNET

R.García, J. Pons, M.P. Malumbres and J. Prades
Department of Computer Engineering (DISCA)
Technical University of Valencia
46022 Valencia, Spain
E-mail: {roman, jpons, mperez}@disca.upv.es, jprades@dcom.upv.es

KEYWORDS
Spanning Tree, Backbones, Loops, Gigabit Ethernet.

ABSTRACT

The current standard switches are based on the Spanning Tree (ST) protocol. The most important restriction is that ST switches cannot work when the topology has active loops. A new protocol is proposed in this paper for Gigabit Ethernet switches, to use in the last stage of a fat tree in order to allow a final backbone with active loops. So, rings, mesh and other regular active loop topologies can be used to connect the Gigabit switches in order to obtain better performance results. The proposed protocol is named ALOR for Active Loops and Optimal Routing (as loops imply alternative paths, the ALOR protocol uses only optimal routing).

INTRODUCTION

Switches can be used to build a diameter-limited network, usually named "Switched LAN" or "extended LAN". The current standard switch is a transparent switch. Transparent means that the stations do not need to use special software to work with switches. The most attractive feature of transparent switches is their easy installation procedure and null maintenance. This paper is focused in an old restriction associated with switches, their inability to work with loops (Perlman 1999).

A switch is a smart hub. It learns where the stations are, so it can forward frames to their destination using the appropriate paths. The learning process is simple:

a) Station-n transmits a frame.
b) Switch-j receives the frame on port-i. The switch reads the “source address” field of the frame and learns that station-n can (and must) be reached using port-i.

The learning process is, obviously, a continuous (non stop) process. When a switch do not know where a destination station is, it simply sends a copy of the frame for all its ports but the one on which frame was received. Finally, all switches in the LAN learn the way (port) to reach any active station and route frames in consequence.

There is an important restriction in order to do the learning process feasible; loops are forbidden! Why? Because a loop implies alternative paths, so a station can be detected by multiples ports in the same switch and this confuse the learning process. And, more important, some frames (i.e. broadcast frames) would be caught in the loop infinitely with no solution.

Therefore, the current standard switches use ST algorithm to transform any topology into a tree. So, switches have no problems with the learning process and broadcast frames flooding the LAN in a normal way. But the tree has a very important problem; it is not a good interconnection topology. In fact, it is very bad.

The figure 1 (a) shows a tree with three hierarchy levels. Root-switch is in the first (top) level of the hierarchy. At the bottom there are the leaf-switches. They connect with stations. If only one technology is used (i.e. 10 Mbps Ethernet) it is clear that an excess of traffic will saturate switches near to root. In order to alleviate this problem, engineers have designed powerful switches with multiple ports (24 ports per switch, and even more, are usual) and/or use “fat tree” topologies.

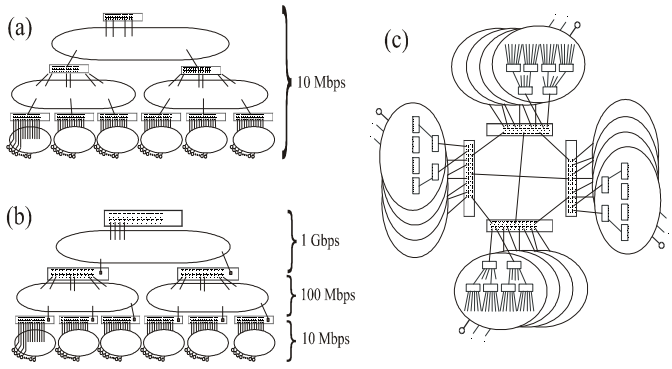

Figure 1: (a) A tree topology (b) A fat tree topology (collapsed backbone on top) (c) Fat tree with ALOR (distributed backbone on top)

Figure-1 (b) shows a typical fat tree based in Ethernet technologies. A problem in a fat tree is that the best performing technology has to be kept for the backbone LAN, and cannot be used in the rest of the LANs in order to balance traffic and bandwidth. A variant of this problem occurs when it is necessary to update the end-stations technology. For example, currently it is usual to install Fast Ethernet (100 Mbps) at the stations. So, it is only possible to make a fat tree with two hierarchical levels (1Gbps on top and 100 Mbps down). Considering the maximum number of ports per switch (24 ports is the common available) it is a restriction in order to connect a great number of stations (like, for example, in a campus LAN).

Note that the problem is that the backbone with gigabit Ethernet switches is a collapse backbone. The backbone is formed for only one gigabit switch. If we finally need to use three o more gigabit Ethernet switches as backbone

(distributed backbone), we have again the problem of the tree. (Seifert 1998)

This paper proposes a new protocol (ALOR) for gigabit switches that allow the use of active loop topologies. Therefore, strongly connected regular topologies, like meshes, as well as irregular topologies with active loops, can be used as distributed backbones. As loops imply alternative paths, the ALOR protocol uses optimal routing.

ALOR works on top of ST protocol. It uses information gathered by ST. It is a distributed protocol but, in this case, only for gigabit switches. Although ALOR is proposed for gigabit switches, it can be proposed for any kind of switches. And not only at the top of a fat tree but also at any point in the network. (García and Duato 1998) (García et al. 1998)

ACTIVE LOOPS AND OPTIMAL ROUTING PROTOCOL (ALOR)

The ALOR protocol is based on the idea that switches can learn "which stations are associated with each gigabit switch" or, in other words, “where each station is”. This is the main difference with respect to the ST based standard, in which the switches only learn "from which direction" the station has been listened to. Thus, the ST is not strictly required, and the most favorable route can be used considering all the existing lines (links). The optimal route criterion is based on number of hops.

This paper assume the following:

a) A gigabit switch is a switch in witch all the ports work at 1 Gbps. So, switches with have one port at 1 Gbps and N ports at 100 Mbps are not considered gigabit switches. These switches do not take part in the distributed backbone and are out of ALOR protocol.
b) The root switch of the LAN is a gigabit switch. Otherwise it has no sense.
c) Equally, all the gigabit switches must have identity numbers in coherence with the rest of the switches of the LAN. So, if root-switch crash, the new root-switch selected by the ST protocol will be another gigabit switch.

Determining which ones of the ports of a gigabit switch is in the giga-tree.

A gigabit switch has all its ports working at 1 Gbps., but only some of them are involved in the giga-tree (the distributed backbone). It is necessary to define an automatic method that allows the gigabit switch to know which ones are those ports. ALOR protocol proposes that gigabit switches exchange *ALOR configuration messages* (frames) with the following format:

destination address = ALOR group (multicast address),
source address = MAC address of the source switch,
type = ALOR id. ,
data = transmission port status OR acknowledge.

The gigabit switches must accept ALOR configuration messages, and any other ALOR messages, even if they are received from a port in blocking status.

The ALOR protocol uses configuration messages as follow:

a) Every gigabit switch needs to transmit the message for all its ports only one time.
b) When a message arrives at a port of a gigabit switch, it must to acknowledge it.
   These messages are very important in the correct operation of ALOR protocol, so this acknowledgement is to make the protocol more robust.
c) Root-switch transmits first
d) Rest of gigabit switches transmit configuration messages when receive the first ALOR configuration message from a neighbor.

This process is repeated only when the Spanning Tree detect a topology change.

Note: An ALOR configuration message cannot pass from a gigabit switch to another gigabit switch through a normal-switch (Figure 2). There are two reasons:

a) It would suppose that there is a loop in the topology because any gigabit switch must reach another gigabit switch through root. The ST protocol guarantee there is no active loops.
b) A normal-switch does not accept ALOR messages from a blocking port. It only will accept ST protocol messages.

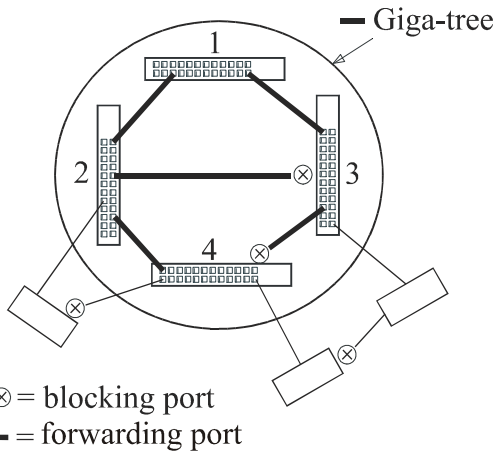

Figure 2: Giga-tree, a tree at the top level of a campus backbone

Preliminary Definitions

Port identification

ALOR takes decisions according with the status of every gigabit switch in the giga-tree. This status can be known from the state of the ports involved in the giga-tree in every gigabit switch. According to the ST protocol (IEEE 1993), four states are defined for a switch port: *blocking*, *listening*, *learning* and *forwarding*.

A *root port* is defined as the one used by the switch to reach the root of the ST and *designated port* is also defined as ports in forwarding state others than root port. Both, root and designated ports are in forwarding state.

These definitions are sufficient to know the location-connection of each port within the tree and in the fat-tree.

Corollary 1: Every switch, other than the root-switch, has just one root port.
Corollary 2: In the root-switch all the ports are designated ports

Definition 2: Leaf-switch

A switch is a "leaf switch" if it is at the end of the tree hierarchy

<< A switch know it is a leaf switch if it has a root port and the rest of ports (in the giga-tree) are blocking. >>

This definition is important because the leaf switches are the ones that will initiate the ALOR learning process, as discussed later.

Definition 3: Intermediate-switch.

A switch is "intermediate switch" if it is between the root switch and a leaf switch, so, in the middle of a tree hierarchy.

<< A switch knows it is an intermediate switch if it has a root port and one or more designated ports. >>

ALOR: Learning

ALOR Learning Fundamentals

The ALOR protocol learning process is based on the tree generated by the ST protocol, and evolves from the ST leaves towards the root. This is therefore a bottom-up process.

The ALOR fundamental is:

<< A gigabit switch is proprietary of all the stations it listen from all its ports others than the ports in the giga-tree. Thus, it can associate a "cost to reach" equal to zero to the MAC address of the source station (hop count is the simplest metric, but other metrics are also possible). >>

But the main goal of the learning process is to share the information among all the switches of the switched-LAN. Thus it is necessary to plan a spreading strategy to obtain an ordered full propagation. Switch knowledge is transmitted to the neighboring switches through ALOR location messages. Basically it contains a list of the new stations (MAC-addresses) and the cost (hops) to reach them.

The learning process consists of the following steps:

1) Bottom-Up process: This process is initiated by the leaf-switches. They transmit their knowledge (in *ALOR location messages* defined later) to the switches
   a) on the higher level of the hierarchy using the root port and
   b) on the others branches of the tree (lateral propagation) using the blocking ports.

   Any switch other that the leaf-switches waits to receive an ALOR location message for all the designated ports before to repeat the process.

   Finally the bottom-Up process stops at the root switch. At that point, the root switch has a full knowledge of all the active stations in the switched-LAN and their location (cost). Note that the routes known by root are optimal since the ST is an optimal tree.

It is necessary a second learning/propagation top-down phase, in order to allow the root switch to spread its knowledge to the rest of the switches. Then all the switches will know every new station and location (cost) across the tree.

2) Top-Down process: This process is initiated by the root-switch. A switch transmit their knowledge to the switches
   a) on the lower level of the hierarchy using the designated ports and
   b) on the others branches of the tree (lateral propagation) using the blocking ports.

   A switch repeats the top-down process when it receives an ALOR location message by its root port. Obviously, the process stops at the leaf-switches.

Learning Example

Figure 3 shows a campus LAN using switches. At the top level of the fat tree topology the campus LAN uses four gigabit switches.

Note: A restriction of gigabit Ethernet (1000BASE-SX) is a maximum distance supported of 525 m. (using 50 ?m. multimode fiber), so another of the reasons to use more than only one gigabit switch could be to cover a large campus.

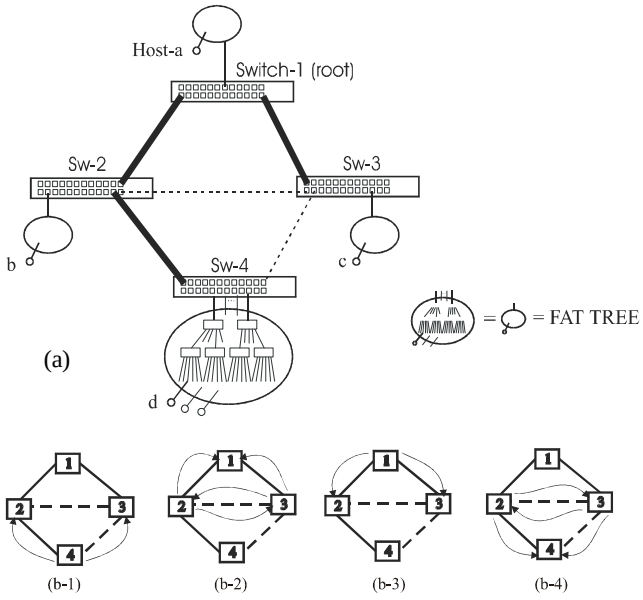

Figure 3: ALOR protocol evolution. (a) Giga-tree in the example. (b-1) First ALOR location messages; transmission and reception (tx-1 and rx-1). (b-2) tx-2 and rx-2. (b-3) tx-3 and rx-3. (b-4) tx-4 and rx-4

Fat lines between gigabit switches show the ports enabled by the ST and the dotted lines show the ports blocking. The rest of the ports of the switches are supposed connected to non-gigabit switches but giga-to-fast switches.

Table 1: ALOR evolution in the example

	Switch-1				Switch-2					Switch-3					Switch-4			
	2	3	i	M+	1	3	4	i	M+	1	2	4	i	M+	2	3	i	M+
tx-1			a	a0i				b	b0i				c	c0i			d	d0i
rx-1 & tx-2							d0		d14			d0		d14				
rx-2 & tx-3	b0 d1	c0 d1		b12 c13 d22 d23		c0 d1			c13		b0 d1			b12				
rx-3 & tx-4					a0 b1 c1 d2				a11	a0 b1 c1 d2				a11				
rx-4						a1 b1 c0 d1					a1 b0 c1 d1				a1 b0 c1 d1	a1 b1 c0 d1		a22 a23

The table 1 shows the evolution of the ALOR learning process. For each switch, the table shows a column with the number of the port from which it receives the ALOR message. To make the example simpler it is numbered "port-n" the port that connects with the switch-N (so, "2" means "port-2" and it is the port to switch-2). The column labeled "i" resume the rest of the ports that are not implicated with the input/output of ALOR messages, but are the ports that connect with the rest of the fat tree, so with the final stations. The column labeled "M+" is the cache memory where the results of the ALOR learning process are summarized.

The example supposes that the ST is already formed and the gigabit switches know which ports are involved in the giga-tree. Initially, the cache memory is empty. Four stations "a", "b", "c" and "d" transmit (i.e. broadcast a frame). The frames reach the corresponding gigabit switches and ALOR learns, in each switch, that a new station can be reached by port-i with cost=0 (i.e. in switch-1 column-i, "a" means that station "a" has been detected). ALOR store in cache <station><cost><by port>. In the example, Switch-1 store "a0i" (host-a can be reached at cost 0 by port-i).

Row labeled "tx-1" (tx=transmit) in table 1 shows the first ALOR location messages are transmitted by switch-4 (a leaf-switch). It transmits a message to switch-2 and another copy to switch-3 (Figure-3 (a)). In the data field, message sends all new data in its cache memory "M+". In the example, switch-3 sends "d0".

The next row of the table 1 labeled "rx-1, tx-2" (rx=reception) shows two steps:

a) In the first one, switches 2 and 3 receive the message by their port-4 containing data "d0". Them, both switches learn that its neighbor switch-4 can reach station-d with cost=0, so if they can reach switch-4 with cost=1 (one hop), then they can reach station-d with cost=0+1=1. Both switches store in cache "d14" ("I can reach station-d with cost=1 by my port-4).

b) The second step in this row is the transmission that switches 2 and 3 make (figure-3 (b-2)). In this point switch-2 sends an ALOR location message to its gigabit switch neighbors 1 (bottom-up propagation) and 3 (lateral propagation). Switch-2 sends data ("b0", "d1"). Approximately at the same time switch-3 does the corresponding. Switch-3 sends data ("c0", "d1").

The rest of the process (next rows of the table) is a repeat of those steps. As singular point, it is important to highlight that "M+" can (and must) store more than one route to a single station. For example, switch-1 in row labeled "rx-2, tx-3"store "d22" and "d23", so it knows it has two optimal routes to reach station-d in with cost=2, one by port-2 (switch-2) and another one by port-2 (switch-3). The same in the last row of the table in switch-4 with station-a

Learning fidelity criterion

ALOR switches acquire their knowledge through transmission of location messages. We should consider the implications of lost location messages due to transmission errors.

In case of transmission errors, the cross information that each switch has about the neighbors will not be coherent with the information recorded by these neighbors. But this is not, in fact, a big problem. When a gigabit switch does not know an optimal route (ALOR) for a station it always will use the normal spanning tree information.

A situation that can be done is the following (see figure-3(a)): Suppose that Switch-4 want to transmit a frame to station-c and does not know that an optimal route exists through its blocking pot to switch-3. Then switch-4 will send the frame through the spanning tree towards switch-2. Now, it is possible that switch-2 knows an optimal route to station-c through its blocking port towards switch-3. So, there is no a big problem that an ALOR location message can be lost.

Can a switch know an optimal route that cross through a switch that does not know this optimal route? The answer is: No!

Optimality principle:
If switch-Y is on the optimal path from switch-X to station-Z, then the optimal path from Y to Z is a sub-path of the optimal path from X to Z.

As corollary of this principle, it is no possible that X knows the optimal route to Z if Y does not know it and transmit its knowledge.

Information Expiry Time

Like in the ST protocol, the information learnt by switches has an expiry time. ALOR uses the same expire criterion as ST. There is a long cache time (5 to 15 min.) for the normal operation and a short cache time (3 to 15 sec.) when a topology change is produced.

The expiry time is only controlled by the owner switch. When the station-related information is no longer valid, the proprietary switch will set an infinite cost associated with that station in the next location message, (infinite = 255).

ALOR: Routing

Tunneling

Before proposing ALOR routing, it is important to remember that the ALOR protocol must be compatible with the ST protocol. Compatibility imposes a restriction: ALOR cannot transmit a frame using its original format. Otherwise, it will interfere with the normal learning process of the ST. To solve the problem, a tunneling technique is used.

The standard learning process on a transparent switch is based on the assumption that only one route exists toward each station. ALOR eliminates this restriction. So, it is necessary to encapsulate frames in order to distinguish normal frames from ALOR frames.

ALOR encapsulates the original frame in a new special frame format that is only recognized by ALOR switches. The tunneling frame adapted to Ethernet format is:

1. Dest. address = ALOR multicast address (6 bytes)
2. Source address = Switch address (6 bytes)
3. Type = ALOR protocol (2 bytes)
4. Data = Station original frame

The "destination address" field is a globally known multicast group address that identifies all ALOR switches. The "source address" field is the MAC address of the ALOR switch that transmits the frame. The "protocol type" field identifies the ALOR protocol and the "data" field contains the original frame. The maximum size of the new frame is 1528 bytes, slightly longer than the maximum standard size (1514 bytes). This does not cause any problem because only ALOR switches handle these frames.

Ones the tunneling frame reaches the destination gigabit switch the frame is send through the destination fat tree without tunneling.

ALOR Routing

ALOR routing is performed in a distributed way. When a station frame reaches a gigabit switch, it is checked in ALOR cache for the destination station. If there is a route, and it is different of the route through the ST, then the frame is tunneled and sent through the corresponding port, else ALOR leave to ST the job.

An interesting case arises when the switch knows two or more optimal routes. In that case traffic could be distributed in a proportional way. For example, if there are two optimal routes, traffic can be split fifty-fifty but this policy can produce a bad collateral effect if it is done carelessly. If frames can reach a destination through different routes, it is possible a second frame reach to destination before the first one. Indeed, this is a situation that can be done in any distributed protocol that continuously learns new routes. Note that ST protocol either guarantee this situation.

CONCLUSIONS

A new protocol ALOR is proposed. ALOR is very simple and easy to implement. It works on top of the Spanning Tree (ST) protocol and allows that gigabit Ethernet switches work with active loop topologies. It is an important improvement over the ST protocol that allows topologies with loops but blocks ports in switches in order to obtain a tree.

As a tree is not a good interconnection topology, fat trees are the usual solution to enhance performances using standard switches. Gigabit switches are designing to work as collapsed backbones, so if more than three gigabit switches are used the problem come out again. Then, ALOR is an alternative to consider.

REFERENCES

García R., J. Duato, JJ.Serrano. 1998. “A new Transparent Bridge Protocol for LAN Internetworking using Topologies with Active Loops”, Proceeding of the 1998 International Conference on Parallel Processing (ICPP98).

García, R., J. Duato. 1998. “Suboptimal-Optimal Routing for LAN Internetworking using Transparent Bridges”, International Journal of Foundations of Computer Science Vol. 9 No. 2 (1998) 139-156.

IEEE. 1993, *Mac Bridges*, ANSI/IEEE Std. 802.1D, ISO/IEC 10038

Perlman, R. 1999. *Interconnection Networks (2º Edition)*, Addison Wesley.

Seifert, R. 1998. *Gigabit Ethernet.* Addison Wesley.