Performance Evaluation of Recursive Network Architecture for Fault-tolerance

[1]Minho Shin, [1]Raheel Ahmed Memon, [1]Yeonseung Ryu*, [2]Jongmyung Rhee and [3]Dongho Lee

[1]Department of Computer Engineering, Myongji University, Korea
[2]Department of Information and Communication Engineering, Myongji University, Korea
[3]Agency of Defense Development, Korea
ysryu, mhshin {@mju.ac.kr}

Abstract. Network fault tolerance is one of the most important capabilities required by mission-critical systems such as the naval Combat System Data Network (CSDN). In this paper, we present performance evaluation results of a fault-tolerant network scheme called Recursive Scalable Autonomous Fault-tolerant Ethernet (RSAFE). The primary goal of RSAFE scheme is to provide network scalability, and autonomous fault detection and recovery within given a time. We show that proposed recursive architecture can support a large number of nodes while guaranteeing the fail-over time requirement.

Keywords: Fault-tolerant Ethernet, Large-scale network, Fail-over time, Mission-critical systems

1 Introduction:

In the network based mission critical systems such as unmanned vehicles, military weapon and aviation equipment control systems, Ethernet connectivity must provide fault tolerance to all the constituents for the smooth operations of such systems. There have been a lot of software-based fault tolerant approaches that adopts a heartbeat mechanism for failure detection [1-6].

In our previous work, we presented a fault tolerant Ethernet scheme called *Recursive Scalable Autonomous Fault tolerant Ethernet* (RSAFE) [7]. In RSAFE, a large network is divided into various subnets (a subnet contains limited number of nodes), further limited number of subnets are grouped together to form different groups, then groups are recursively combined to form levels, until only one group remains in the highest level. RSAFE uses a heartbeat mechanism for fault detection. The heartbeat mechanism may cause large bandwidth consumption as the size of network grows. But, in our proposed approach by dividing the large network into small subnets and limiting the size of subnet, heartbeat mechanism can be implemented efficiently.

*corresponding author.

This work was supported by Defense Acquisition Program Administration and Agency for Defense Development under the contract UD070019AD.

In this work, we perform theoretical evaluations to analyze the number of nodes that can be supported in proposed RSAFE while providing fail-over functionality within a given time. According to our evaluation, to maintain the failover latency below 1 second, RSAFE can support up to 2400 nodes for subnet size of 16, 4700 nodes for subnet size of 32, and 10000 nodes for subnet size of 64. Therefore, proposed recursive architecture can support a large number of nodes while guaranteeing the fail-over time requirement. In section 2, we describe related works. In section 3, we present our RSAFE scheme. We then show analysis study in Section 4. Finally, we summarize the work in Section 5.

2 Related Works

In [1], Scalable Autonomous Fault-tolerant Ethernet (SAFE) scheme was proposed. SAFE divides large network into several subnets. In order to bound fault recovery times, SAFE limits the number of nodes in a subnet and configures the subnets as a star network. Because SAFE scheme is based on star topology, if two switches in the center subnet are destroyed at the same time, the entire network's operations could be interrupted.

[6] proposed a recursive network scheme, called DCell, for interconnectivity in between exponentially increasing number of servers in data centers. High-level DCell is constructed from many low-level DCells while DCells at a same level are fully connected to each other. Since each server in DCell networks is connected to different levels of DCells via its multiple links, it becomes a rather costly solution.

3 RSAFE

We proposed RSAFE (Recursive SAFE) scheme in [7]. RSAFE is a low cost fault tolerant architecture for mission critical networks. The architecture of RSAFE is shown in Fig. 1. RSAFE comprises of *three basic building blocks*; Subnet, Group and Level. Together these components form a Recursive structure. The building algorithm for RSAFE network is described in detail in our previous work.

Fig. 1 (a) shows that a subnet is constructed from limited number of nodes and two switches. There are four possible paths for communication between two nodes. In the subnet one data path is defined as the primary path for communication between two nodes of a subnet. Rest of the paths will remain on standby and will activate in the case if the primary data-path encounter a fault.

Fig. 1 (b) shows that a limited number of subnets together form a group where each subnet has a dual connectivity with other subnets of the group. The connectivity is formed as second switch of Subnet_0 is connected to first switch of Subnet_1 and second switch of Subnet_1 is connected to first switch of Subnet_2, for last subnet the second switch of Subnet_n is connected to first switch of Subnet_0.

Fig. 1 (c) shows that a limited number of Groups together forming a Level. Once the Subnets and Groups build the next step is to build levels, Level0 is combination of limited number of groups, and Level1 to onward is the combination of different Levels. In Fig. 1, it is shown that Level0,0 only, but suppose that there Level0,m

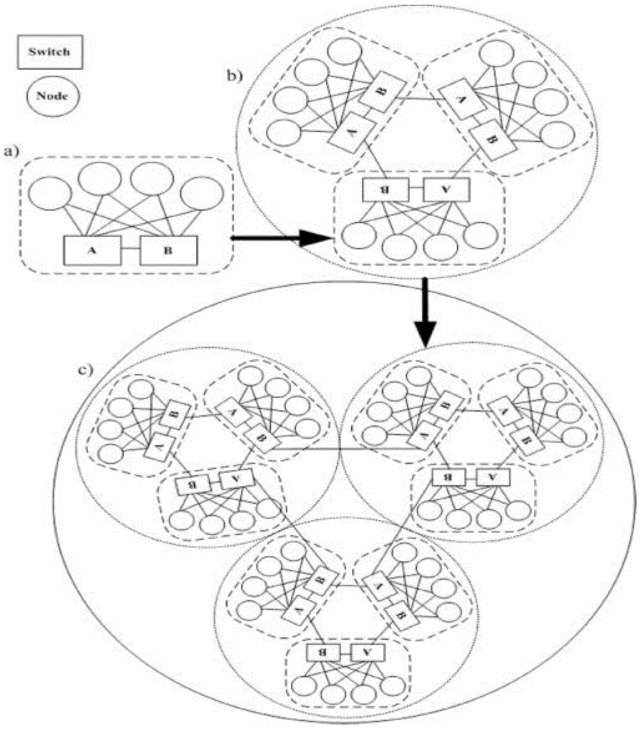

Fig. 1. Three basic building blocks of Recursive Scalable Autonomous Fault-tolerant Ethernet: a) Subnet, b) Group, and c) Level

levels are available then the Level1 will build by combining all the levels of Level0. Similarly all the required levels can be created recursively; we can say that, in this structure all the higher levels are constructed recursively from lower levels. And for connectivity between the Levels or groups the subnets are selected randomly from each level/group and connected to each other.

RSAFE basically detects network faults on the basis of subnet using heartbeat mechanism. A heartbeat message (HBM) is an Ethernet frame sent and received between nodes in each subnet. HBM can reach only the nodes in a subnet and cannot reach the outside nodes of the subnet because it is layer-2 data. Each node is responsible for detecting the faults of its belonging subnet and recovering from it.

We adopt our previously proposed solution in order to exchange the information outside the subnets [1]1, 3]. RSAFE manages master nodes in each subnet for inter-subnet fault recovery. Primary Master Node (PMN) is an active master node and Secondary Master Node (SMN) is a standby master node, in case if PMN fails then SMN can recover the fault. PMNs communicate with each other using IP packets to exchange the subnet status only when fault occurs. When a fault occurs in a subnet, the master node can detect it by heartbeat mechanism and notify to other master nodes to recover from the fault on the inter-subnet communication path. When the master node receives a notification from other master nodes, it in turn sends a notification to nodes in its subnet to recover from the fault.

4 Evaluation

4.1 Failover latency within a subnet

Eq. 1 shows the upper bound of the failure over time $_{\boldsymbol{T_{sub}}}$ within a subnet, where $\boldsymbol{T_{HBM}}$ is heartbeat interval:

$$T_{SUb} \sim 2 \text{ x } T_{HBM}. \tag{1}$$

If we increase the number of nodes in a subnet, the heartbeat interval should also be increased. This results in, however, increasing the fault recovery time because the node detects the faults when it has not received two consecutive heartbeats.

4.2 Failover latency across the network

In this section, we compute the upper bound of the failover time throughout the network. Let ***N*** denote the number of nodes in the network and ***d*** denotes the number of groups in each level. For brevity of analysis, we assume that $\boldsymbol{N} = \boldsymbol{d}^{\boldsymbol{k}}$ for some positive integer ***k***, which is called the depth of the network. Note that RSAFE is recursively structured; each level consists of ***d*** groups connected as a ring, and each group forms a one-step lower level. Each subnet is level-0 and the whole network is level-***k***. Without loss of generality, we assume that ***d*** is an even number. We assume that each link has the same latency of $\boldsymbol{T_L}$. Also note that two connected groups of the

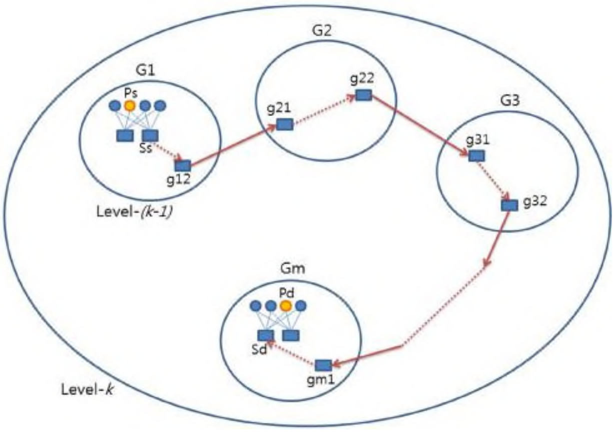

Fig. 2. Failover message path in Level ***k***. The PMN $\boldsymbol{P_s}$ in a faulty subnet sends a message to another PMN $\boldsymbol{P_d}$, which belongs to the group $\boldsymbol{G_m}$, ***m*** hops far from the $\boldsymbol{P_s}$ along the ring.

same level are connected via their ***gate-way switches*** chosen at random.

When a fault occurs in a subnet, the PMN recognizes the fault event within T_{sub} seconds. Then, the PMN sends a message to each PMN. Consider the highest level of the network (Fig. 2). As shown in the figure, the message travels from the source group (G_1) to the destination group (G_m) through the ***m-1*** intermediate groups (G_2, G_3, ..., $_{Gm-1}$) along the group ring. The path from the source PMN (P_s) to the destination PMN (P_d) consists of the first hop from the source PMN to the subnet switch ($P_s 4 S_s$), the path from the switch to the gateway switch $_{(ss4g12),}$ multiple hops between groups along the ring ($g_{12}4g_{21}$, $g_{22}4g_{31}$, ..., $_{gm-1,24gm1),}$ the paths from in-gateway to out-gateway in each intermediate groups ($g_{21}4g_{22}$, $g_{31}4g_{32}$, ...), the path from the in-gateway of the destination group to the subnet switch of the destination PMN ($g_{m1}4s_d$), the hop from the subnet switch to the destination PMN (s_d4P_d). Finally, the destination PMN sends a notification to each node of its subnet in one hop. Therefore, we can compute the upper bound of the fail-over-time latency by

$$T\ T\sim\sim\sim+{}^{T\sim}+T\sim T\sim\sim\sim$$

$$\sim\sim\sim+T\sim$$

where $_{Tgrp(k)}$ is the largest path length between switches within a group of level-***k***, i.e., the whole network.

Now we focus on computing $_{Tgrp(k).}$ Due to recursive structure, we get

$$TgrP = R + (R + {}^{1})Tg^{(}rP^{1)}$$

for ***i***>1, where ***R*** is the maximum hop count along the ring between two groups. When ***i***=0, $T_{grp}^{(0)}$ = 1 (one hop between two switches). By unfolding the recursive formula, we get

$$T$$

$$\sim\sim\sim=2\sim R+1\sim^{\sim}-1$$

Since ***R***=***d/2*** (***d*** is the ring size, so ***d/2*** is the path length between farthest groups) and ***k***=log$_d$***(N/m)*** where ***N*** is the number of nodes and ***m*** is the subnet size. By Eq. 1, we have an upper bound of the failover latency as

$$\%\sim\sim{}_{\&}'/)$$

$$T\sim\sim\sim\sim 2T\sim !+T\sim+2T\sim {}''{}^{d}{}_{2}+1\$$$

Fig. 3 shows $_{Tgrp(k),}$ the largest number of hop counts from the source switch to the destination switch as the number of nodes in the network increases. Considering 1 millisecond link delay, $T_{grp}^{(k)}$=100 is equivalent to the failover latency of 1 second. Three lines, from bottom to up, represent $_{Tgrp(k)}$ when the subnet size is 16, 32, and 64. To maintain the failover latency below 1 second, RSAFE can support up to 2400 nodes for subnet size of 16, 4700 nodes for subnet size of 32, and 10000 nodes for subnet size of 64.

5 Conclusion

Network fault tolerance is one of the most important capabilities required by mission-critical systems such as the naval Combat System Data Network (CSDN). In

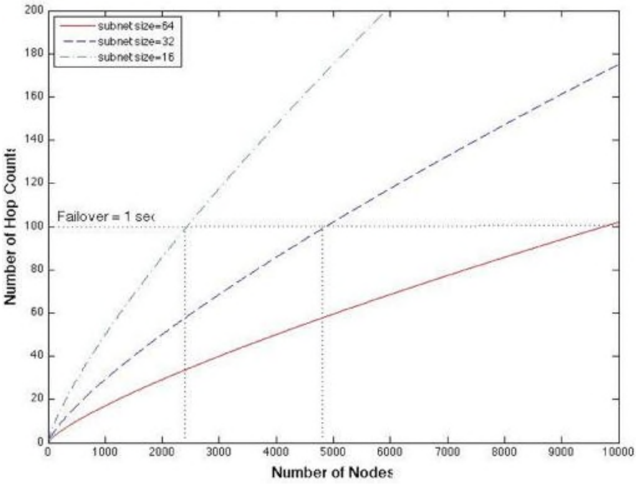

Fig. 3. Maximum hop count between two switches as the number of nodes increase. Multiplied by the link delay of 1 millisecond, it represents failover latency.

this paper, we presented theoretical evaluation results of our recursive fault-tolerant network scheme called RSAFE. In RSAFE, a large network is constructed recursively from subnets, groups, and levels. We showed that proposed recursive architecture can support a large number of nodes while guaranteeing the fail-over time requirement.

References

1. K.Y Kim, Y.S Ryu, J.M Rhee, and D.H Lee, "SAFE: Scalable Autonomous Fault-tolerant Ethernet," Proc. of the 11th International Conference on Advanced Communication Technology (ICACT), pp. 365-369, 2009.
2. H.A Pham, J.M Rhee, S.M Kim, and D.H Lee, "A Novel Approach for Fault Tolerant Ethernet Implementation," Proc. of the 4th International Conference on Networked Computing and Advanced Information Management (NCM 2008), Vol. 1, pp. 58-61, 2008.
3. H.A Pham, J.M Rhee, Y.S Ryu, and D.H Lee, "Performance Analysis for a Fault-Tolerant Ethernet Implementation based on Heartbeat Mechanism," Proc. of the 41th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, 2011.
4. S. Song, J. Huang, P. Kappler, R. Freimark, J. Gustin, and T. Kozlik, "Fault-Tolerant Ethernet for IP-Based Process Control Networks" Proc. of the 25th Annual IEEE International Conference on Local Computer Networks (LCN'00), pp. 116-125, 2000.
5. H.A Pham, J.M Rhee, Y.S Ryu, and D.H Lee, "An Adaptive and Reliable Data-Path Determination for Fault-Tolerant Ethernet Using Heartbeat Mechanism," Proc. of the 5th International Conference on Computer Sciences and Convergence Information Technology (ICCIT 2010), pp. 440-444, 2010.
6. C. Guo, H. Wu, K. Tan, L. Shei, Y. Zhang, and S. Lu. "Dcell: a Scalable and Fault-tolerant Network Structure for Data Centers" ACM SIGCOMM, 2008.
7. R.A. Memon, Y.S. Ryu, M.H. Shin, J.M. Rhee, and D.H. Lee, "Building Scalable Fault Tolerant Network," Proc. of the 14th International Conference on Advanced Communication Technology (ICACT), 2012.