2009 International Conference on Knowledge and Systems Engineering

A Three-tier Structured Model of Overlay Networks

Mei Wu[1,2]
[1]School of Electronic Engineering,
Dublin City University,
Dublin, Ireland
wumei@eeng.dcu.ie

Chanle Wu[2]
[2]School of Computing,
Wuhan University,
Wuhan, China
wuchl@whu.edu.cn

Martin Collier[1]
[1]School of Electronic Engineering,
Dublin City University,
Dublin, Ireland
collierm@eeng.dcu.ie

***Abstract*—One of the open research problems in the area of overlay networks is the creation of an adequate network topology and a proper network model. Current models mainly focus on traffic demands and loads between the nodes. However it is more challenging to find a model which supports scalability, adaptability and robustness in a heterogeneous environment. We introduce a new three-tier model based on regular-graph theory called STree and its two-stage joining mechanism. We compare STree with other models such as NICE, DTree and HMRB. Our results suggest that STree is a robust and scalable model for overlay networks in large, dynamic and heterogeneous environments.**

Keywords- structured; overlay networks

I. INTRODUCTION

As the Internet is growing, its inherent characteristics such as heterogeneity and local autonomy increase the difficulty of designing and developing new overlay networks. Besides these factors, designers also need to consider the dynamic growth of real-time applications (e.g., live video) and difficulties related to this type of network traffic (like randomness and burstiness). One of the open research problems in the area of overlay networks is the creation of an adequate network topology and a proper network model. To a large extent, the above factors limit the development of overlay network applications.

The main research objective of this paper is to develop a scalable model for overlay networks that will ensure reliability and robustness for core business applications. Such a model would be attractive for applications that provide large-scale real-time media streaming services. This article also reviews existing models of overlay networks.

In order to meet the above needs, we propose a robust and scalable structure for overlay networks.

The remainder of this work is organized as follows. Section II discusses related work. Section III formally introduces the three-tier model. Section IV shows how to construct the proposed model. Finally section V presents simulation results and summarizes the paper.

II. RELATED WORK

Network heterogeneity, robustness, scalability and high efficiency are topics of increasing interest in overlay network research. A great deal of research is focused on developing models and designing systems that provide large-scale live video services for a large number of users. Banerjee and Bhattacharjee [1] provide a comprehensive survey.

The current overlay network topologies can be classified into two categories. The first group includes tree-mesh topologies such as single-tree large-scale multicast schemes or multi-tree schemes. The second category is peer-to-peer networks that use special logic to setup the connections.

NICE [2] and Narada [3] are representatives of single-tree large-scale multicast schemes. In single-tree schemes leaf nodes do not share upload bandwidth with other nodes. Therefore, these approaches cannot achieve the maximum network throughput. Multi-tree schemes solve the robustness problems of the single-tree schemes, but they face a new problem. Because of the special logic structure, multi-tree schemes require high control overhead. This drawback limits the scalability of the whole system. SplitStream [4] and CoopNet [5] are representative multi-tree schemes. CAN [6] is an example of a scheme based on special logic. Schemes based on special logic have higher maintenance costs and have lower search overhead.

There are currently two models of overlay networks: structured and unstructured, namely, random topology model and aggregation topology model. From a layered architecture point of view, all these models are two-tier models based exclusively on a self-similar structure.

In the heterogeneous environment, a self-similar structure has obvious disadvantages. The random topology model belongs to the class of unstructured models; Bittorent, eMule [7] and PPlive [8] are representative examples. An unstructured model is not suitable for the heterogeneous environment where nodes have different traffic capabilities. It does not guarantee the quality of the service for the nodes and uses a lot of the

978-0-7695-3846-4/09 $26.00 © 2009 IEEE
DOI 10.1109/KSE.2009.26

network's bandwidth. On the other hand, an aggregation topology model has a larger network diameter in a single domain. The rigid structure constraint weakens the scalability and heterogeneity of the aggregation topology model [9]. Nowadays, all of the above problems are becoming more clear and evident.

III. MODEL

We propose a three-tier overlay network model. As shown in Figure 1, the three-tier model consists of three parts. These three layers (top to bottom) are:

Layer 1: Main Structured Network (MSN).

Layer 2: Domain Structured Network (DSN).

Layer 3: Unstructured Network (UN).

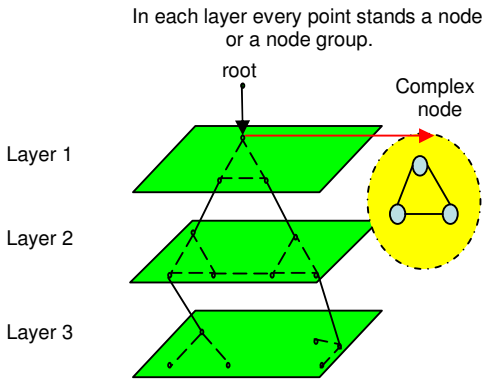

Figure 1. Three-tier model

MSN: the top layer provides the highest quality of service. It consists of nodes with the same degree which satisfy the given constraints of both bandwidth and delay. It is responsible for providing reliability and high quality of services for users.

DSN: the middle layer consists of nodes with the same degree. The role is to provide stable and reliable quality of services.

In addition to the above layers, we provide a supplementary layer — the unstructured network (UN).

UN: the bottom layer provides scalable services. It consists of nodes of various degrees. It is a random unstructured network to provide scalable quality of service, as shown in Figure 2.

IV. THREE-TIER MECHANISM

Under a dynamic environment, each node joins and leaves the network at random moments. This leads to dynamic changes in the network model. Therefore, we need to consider the spatial location information of the nodes.

Using the hierarchical concept and "mixed" searching methods, we design a two-stage node joining mechanism. In the first stage a single node joins the unstructured network. In the second stage it moves into the structured network which is in the middle or the top layer.

At the beginning we let all the nodes join the unstructured network. This ensures that all the nodes form a network and communicate with each other without considering the difference between their capabilities (e.g., available bandwidth, capacity, etc.). This allows nodes with different capabilities to provide their services more rapidly.

In order to provide better, reliable and high-quality multicast services to nodes, we move nodes in the networks from a lower layer to a higher layer. In our model nodes can offer higher quality of service assurances when they are in a structured layer.

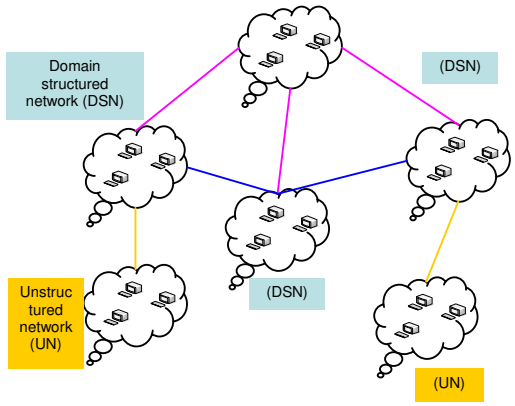

Figure 2. Three-tier network model

A. *Joining of Nodes*

We build an initial structured overlay network (DSN) and initialize the sets of nodes in both the unstructured network (DSN) and main structured network (MSN) to empty sets. When a new node requests admittance to the network, it first joins the unstructured network. Next, we have to decide when the node needs to be moved from the lower layer to the higher layer. We use three judging criteria, namely Definition 1, Inference 1 and Definition 2 below to decide this. We further have to decide when we release some nodes from the current networks and replace them with others.

Definition 1: Given a graph $G= (V, E)$, where G is a common degree of a regular graph (which describes our structured network), V is the set of vertices, E is the set of edges and n is the number of nodes in the network, we add nodes to the corresponding graphs of the structured network. If condition (1) occurs, we can move new nodes into the structured layers.

$$\forall i, \left(\sum_{i=1}^{n} deg\ ree(Vi)\right) mod\ 2 = 1 \qquad (1)$$

Definition 1 states that we move new nodes into the structured layers when the sum of the degrees of all the vertices in G is an odd number. Inference 1 decides upon the condition whether a graph meets the constraint of the regular degree when we add new nodes into the structured network. If the above condition is not true, then we need to add other nodes into the network until the graph integrity condition is met.

In order to judge whether a sequence of degrees of the vertices is graphical (the network is connected.), we use the following method. Given a non-increasing sequence S of n nonnegative integers, we construct a new sequence S' consisting of $(n-1)$ integers. S' is a judging condition to decide whether the current network meets the constraint of the graph.

Inference 1: Suppose S' was derived from S then S is graphical if and only if S' is graphical.

Proof: Reduction to absurdity. If the sequence S is not graphical, then S' does not exist. The whole proof is described in detail in [12].

We choose two parameters as the constraint conditions in Definition 2 namely the available bandwidth B of the nodes and the minimum delay time D in the overlay networks.

Definition 2: When conditions (2) and (3) are fulfilled a node (or a group of nodes) joins the graph that consists of nodes that form the Main Structured Network (MSN).

$$\varepsilon_{ij} \le B \tag{2}$$

$$D_{avg}^{max}(i) \le D \tag{3}$$

Where R_{ij} is the average data speed on the link between nodes i and j per unit time, B_{ij} is the link bandwidth between nodes i and j, $D_{avg}^{max}(i)$ is the maximum predicted average-delay when node i is the root and $\varepsilon_{ij} = R_{ij} / B_{ij}$ shows the degree of congestion on the link between nodes i and j.

The value of $D_{avg}^{max}(i)$ can be calculated according to the following formula:

$$D_{avg}^{max}(i) = D_i * Minimum\ \{((\log \frac{n-1}{C_i}) / \log C_{avg}), (n-1)/C_i\} \tag{4}$$

$$D_i = (\sum_{j=1}^{n} RTD(i, j)) / (n-1) \tag{5}$$

where D_i is the average round-trip delay time when node i received the messages from all other nodes, C_i is the degree of node i, n is the number of nodes, $RTD(i, j)$ is the round-trip delay between node i and j, $Minimum\{((\log((n-1)/C_i))/\log C_{avg}), (n-1)/C_i\}$ is the estimated value for the average maximum depth of the overlay tree with node i as its root and C_{avg} is the average degree of peer nodes in the communication zone of node i.

For simplicity in our model we do not consider the effect of connection time.

B. Departure of Nodes

Any given node is allowed to join and leave the system at any time. Single nodes and node groups can leave the network involuntarily in two ways: in the case of a system crash (or a terminated application), or a temporary and unexpected outage such as network overloads. In the latter case, since the conditions in Definition 2 for both the bandwidth B and the delay D are not satisfied we release these nodes (or node groups) from the MSN networks and replace them with others.

C. Node Substitution Operation

Modern real-time multimedia applications often require strict bandwidth and delay guarantees. Taking into consideration the strict delay constraints, the multicast routing performance is affected by the end hosts. It may lead to dynamic load imbalance among hosts if the optimal path is blocked and network congestion happens. When node failures occur, communication interruptions can happen. There are two ways to recover from such communication interruptions: by nodes switching or by links switching.

- Switching of Connected Nodes

Once the nodes leave the network, all the links related to these nodes have to be rebuilt. The switching process can happen at any given moment and this operation cannot be avoided.

- Switching of Links

When nodes are switched between networks (e.g., from unstructured network A to structured network B or from sub-network A to sub-network B), the time of links switching can be predicted in advance.

From a customer perspective, frequent interruptions during communication between users will become more inconvenient than rejected connection requests.

We use substitutions of nodes and node groups to accomplish the link switching. Substitution operation has two advantages. Firstly, it improves the backbone network performance by exchanging the nodes with poor stability and low bandwidth and replacing them with nodes and node groups that have better performance (e.g., bandwidth, relay time). Secondly, it reduces data recovery time. Whenever the node or link failure occurs, it recovers the network by substituting limited number of affected nodes and node groups.

V. RESULTS

In this section, we compare the experimental results of our proposed model with the NICE approach. In our comparison we take into consideration two different aspects: the initial tree construction quality and the responsiveness in the recovery planning process. A custom simulation model was developed using C++.

A. Experimental environment

Step1: Simulate the underlying network topology, and generate nodes.

Step2: Randomly choose one of the nodes as the root.

Step3: New nodes join and leave a multicast session at any given moment. Construct and maintain one STree.

B. Comparison of the initial tree construction quality

The path stretch is a parameter that reflects the overall performance and the quality of the overlay multicast. That is why in our experiments we compare our path stretch results with the results of the NICE model.

The path stretch is the ratio of the overlay routing path length along the multicast tree between two nodes to the shortest unicast path length between the same two nodes. It shows the relative delay cost of data transmission. We assume k = 3 for NICE in NICE experiment. Then we carry out 20 random tests. Experimental results are shown in Figure 3. Preliminary obtained results show that the initial overlay tree construction quality in the STree model is better than in the NICE model.

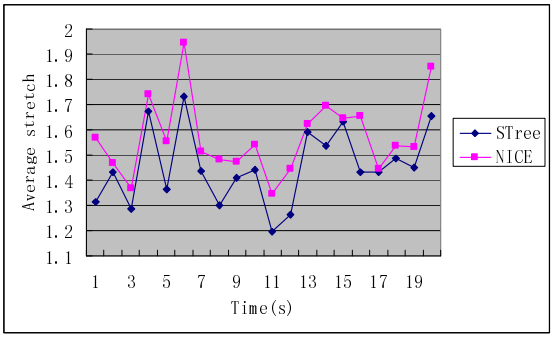

Figure 3. Average Stretch

C. *Comparison of the responsiveness in the recovery process*

When a node failure occurs, the system automatically adds a new node into the original tree and repairs the whole tree. We compare our model's results with the NICE results by using the Average Recovery Time (ART) parameter. We assume that A is an isolated node and node B is the father of node A.

ART = withdrawing time (T1) + request time (T2) + link building time between node B and node A (T3) + link building time between node A and other nodes among the system (T4) (7)

T1 - the node failure detection time. If node A did not get data from its adjacent nodes within one second, it means that adjacent nodes have failed.

T2 - the time needed to search for node B.

T3 - the link building time between node B and node A.

T4 - the link building time between node A and other nodes in the same subnet. Different protocols have different constraints on the node's outgoing (incoming) links. The number of links that node A needs to build to connect with its adjacent nodes differs depending on the protocol that is used.

We assume that the link building time is equal to t. There is one root node in our experiment.

As K=3, each node in NICE connects with the other ordinary nodes, See for example n=12 (root does not include in n.). So T3= t, T4 = (n-2)*t=10t, and ART (NICE) = T1 + T2 + t + 10t = T1+ T2 + 11t.

In our model, the degree of nodes is 3 in Layer one and Layer two. Therefore, we get ART (STree) = T1+ T2 + t +2t=T1+ T2 +3t. In general case ART (NICE) > ART (STree).

The average recovery time in our STree model is shorter than in the NICE model.

D. *Comparison of other models*

Next, we compare our approach with the following models: NICE [2], DTree [10] and HMRB [14].

- NICE

NICE [2] is a hierarchical model that uses clusters. Majority of the nodes belong to the bottom layer of the tiered structure. Only minor nodes belong to the highest level in the hierarchical structure. Once the center nodes fail, they have a huge impact on the performance of the whole system. The proposed scheme does not achieve the maximum network capacity, because the leaf nodes do not share upload bandwidths with other nodes that are higher in the hierarchy.

- DTree

DTree [10] is a degree pre-reserved hierarchical tree. It consists of one backbone tree and several domain trees. Whenever the center nodes of the backbone tree fail, they impact the whole system. Node's degree is pre-reserved in the backbone tree. This allows us to save time while searching for new parent nodes, but at the same time wastes bandwidth resources, as we need to pre-reserve links for temporary connections from other nodes.

In the actual environment, we need to efficiently use all of the backbone nodes resources to the full extent.

- HMRB

HMRB [14] is a two-tier model. It consists of nodes, super-nodes and the super-nodes groups. Nodes form a ring using the Chord protocol. One disadvantage of HMRB is its high maintenance cost, which results in poor scalability.

- STree

Our approach features robustness and connectivity. The concepts of regular degree MSN and its subnet are different from those presented in previous models.

It is based on the regular graph. The degree of all the nodes that belong to a single layer is the same and equal to k. With the increasing number of nodes in the system, the diameter of the graph also increases with O(log(n)) complexity. Each distribution sub-tree is divided into several isolated parts only in the case where we remove at least k nodes. When we randomly remove a certain number (of linear size) of edges or nodes in a sub-tree, the tree still maintains connectivity. Therefore, the model is robust.

Our model has a self-similar structure and supports substitution operation. The role of each node (or node groups) is similar and can be replaced at the top two layers. When nodes or node groups fail, the model heals itself by electing suitable alternative nodes (node groups). The nodes that failed do not affect the stability of the running model.

VI. FUTURE WORK

This paper proposed a three-tier model in order to achieve a robust and scalable overlay network. We gave a comparison with existing models. Our preliminary results suggest that our model is a scalable and robust solution for heterogeneous

networks. Simulation results in a large-scale network environment need to be obtained to confirm this. This will be addressed in our future work.

REFERENCES

[1] S. Banerjee, B. Bhattacharjee, "A comparative study of application layer multicast protocols", Univ. Maryland, College Park, Technical Report, USA, 2002.

[2] S. Banerjee, B. Bhattacharjee and C. Kommareddy, "Scalable application layer multicast", Proc. ACM Conf. on SIGCOMM'02, Pennsylvania, USA, pp. 205-217, 2002.

[3] Y. Chu, S. G. Rao, H. Zhang, "A case for end system multicast", Proc. ACM Conf. on SIGMETRICS, CA, 2000.

[4] M. Castro, P. Druschel, A.M. Kermarrec, A. Nandi, A. Rowstron, A. Singh, "SplitStream: high-bandwidth multicast in cooperative environments", Proc. 19th ACM symposium. on SOSP '03, New York, USA, pp. 292-303, 2003.

[5] CoopNet.http://research.microsoft.com/enus/um/people/padmanab/projects/CoopNet.

[6] S. Ratnasamy, P. Francis, M. Handley and R. Karp. "A scalable content-addressable network", Proc. ACM Conf. on SIGCOMM, CA, USA, 2001.

[7] EMule. http://emule.com.

[8] PPLive. http://www.pplive.com.cn.

[9] J. Zheng, S. Zhang, et al, "Research and simulation for peer selection strategy on P2P streaming", Journal of Computer Engineering and Design, pp. 5396-5399, 2007.

[10] N. Zhang, Y-C. Shi, B. Chang, "Degree pre-reserved hierarchical tree for multimedia multicast", Proc. IEEE Conf. on ICME2006, pp. 1209-1212, 2006.

[11] F. Buckley, and F. Harary, Distance in Graphs, Addison-Wesley, Redwood City, CA, 1990.

[12] N. Wormald, "Models of random regular graphs", In Surveys in Combinatorics, 1999.

[13] A.Mourad, M.Ahmed, "A scalable P2P model for optimizing application layer multicast", Proc. 08-6th IEEE Conf. on Computer Systems and Applications, Doha, Qatar, pp. 407-413, 2008.

[14] J-Z. Xu, Z-L. Yan, et al, "HMRB: Application layer multicast protocol based on hierarchical multi-rings", Journal of Microelectronics & Computer, 24(10), pp. 50-57, 2007.