A Framework for Balancing Load across Systems for Efficient Reliability

Mahendar Reddy Bheemreddy¹, V. Krishna²

¹pursuing M.Tech (IT), Holy Mary Institute of Technology and Science, Keesara, Affiliated to JNTU-Hyderabad, A.P, India
²working as Associate Professor & HOD (IT Department) in Holy Mary Institute of Technology and Science, Keesara, Affiliated to JNTU-Hyderabad, A.P, India

Abstract— Global content requests dynamically routed to the nearest content servers, increasing the uploading speed of page, maximizing the bandwidth and delivering same content regardless of load and network congestions. According to network traffic and number of nodes, the network algorithm chooses the optimal routing option to deliver the content with scalability and reliability. It provides redundancy during internet malfunctions. Content duplication also avoided against loss or degradation of data. The clients requests the content server to fetch the content by CDN. In this Content Delivery Network, to accomplish a scalable and reliable service it should maintain two different properties. First, load awareness to partition the requests over a group of servers that replicates the content, balanced load and network congestion.

Keywords— Put your keywords here, keywords are separated by comma.

I. INTRODUCTION

The usage of internet became massively over the years. In Today’s day-day life style the internet users downloads the different formats of data contents according to their wish. The file size for those content may ranges from few Megabytes to several Gigabytes. These contents are delivered to the users by Content Providers by using Content Delivery Networks (CDN) strategies, like Akamai, Limelight, CoralCDN and CodeeN to deliver the different types of contents to millions of users over Internet through content servers in distributed networks.

In this Content Delivery Network, to accomplish a scalable and reliable service it should maintain two different properties. First, load awareness to partition the requests over a group of servers that replicates the content, balanced load and network congestion. which results the increase in the deliver of same content to the requested users. to handle the flow of incoming requests (flash crowds) the replication is chosen dynamically when Slashdot effect occurs i.e., content becomes popular. Second, locality awareness which exploits the proximity relations in the network like geographical distance, topological distance or round-trip latency, while client requested to the content server. By maintaining network path short the content delivery network provides better QoS (Quality of Service) to the clients.

Designing of an CDN is an challenging task which make an accord between load awareness and locality awareness. In traditional CDN the request of an client is always redirected to geographically closest content servers which lacks the load balancing and results in the overload of content sever when popularity of local content increases. In Advanced CDNs which are mainly based on Distributed Hash Tables (DHTs) primarily targets on load balancing. network locality is only used as an tie-breaker in multiple servers. different strategies of CDNs like Akamai are used as recovery media with an knowledge on types of complex optimizations performed. In this paper, we proposed a new CDNs that captures the load and locality by using Load Aware Network Coordinates. LANCs is an coordinate that maps the network location and load on the content servers. In a decentralized manner CDN maps the client to most appropriate server dynamically by using LANCs. the content which is popular are replicated among the nearest content server with least load on it. for efficient design of CDN and to avoid ad-hoc solutions we combine locality and load of the server by using advanced mechanism of LANCs. The main contribution in our proposed work are:

1. Introduction to LANCs which shows CPU load...
and network congestion 2. design and implementation of CDN by routing client request to content server and replicate the content by using LANCs. Our approach results with a locality workload as to lower the request time rather then traditional content retrieval strategy.

II. CONTENT DELIVERY NETWORK

In Content delivery networks the content is delivered to the user through multiple servers which are located at different geographical locations for delivering static and streaming content. Global content requests dynamically routed to the nearest content servers, increasing the uploading speed of page, maximizing the bandwidth and delivering same content regardless of load and network congestions. according to network traffic and number of nodes, the network algorithm chooses the optimal routing option to deliver the content with scalability and reliability. it provides redundancy during internet malfunctions. content duplication also avoided against loss or degradation of data. the clients requests the content server to fetch the content by CDN, which are served by content servers according to there mapping strategy. the main goal of CDN is to minimize the requested time i.e., until the user retrieve the desired content.

Load Balancing, for a CDN to balance the load, there is an mechanism which balances the load between content servers. For example, we are redirecting the request to the least loaded servers and we are distributing all the requests uniformly around all content servers. if users content request goes to overloaded server then the user experiences the poor performance in retrieval of content.

In our paper, we distinguish between balancing of load for computational load and congestion of network. 1. Computational load at content server are achieved by processing the content requests. Which involves request parsing, cached content retrieval if there are any cache hit and content delivery to the users? 2. Congestion of Network is caused by limiting the network path capacity to carry traffic. Which results the throughput of the content which was delivered to clients as an bottleneck links. here the selection of different content server may increases the performance for retrieval of content. 3. at the access links of content server, alters all users requests to this server. In this case, CDN redirects the request away from congested server.

Locality balancing, if paths of network kept short then we can achieve the locality balancing. For example, CDN delivers the content from nearest servers with minimum load which was requested. In terms of geographical distance, latency, count of routing hops or overlapped address prefixes. By minimizing the length of network path, users experiences better quality-of-service (QoS). Its because: 1) low latencies offer shortest path, which results in TCP to produce high throughput by decreasing the transmission times of small content; 2) the improved throughput can be obtained by congestion hot-spots for short paths; 3) the reliability will be more for short paths as they involve lesser linking of network and routers; 4) network saturation and leaving capacity of network for other traffic can be decreased by shortest paths.

III. LOAD BALANCING NETWORK COORDINATES

Network coordinates are used to maintain locality aware. In network coordinates each node in the network maintains synthetic coordinates of n-dimensional according to measurement of round-trip latency for individual nodes. Load balancing, its very easy to identify the benefits of network coordinates by achieving the locality awareness, here we assume that both client and server are known to network coordinates and client request to the server directly which are located at the closest or nearest network coordinates and the content servers uses their neighbour servers to replicate the popular content which reduces the request times by replicating content. for latency prediction the direct measurements are less important in content delivery network. there will be marginal impact on total request time for huge content by choosing a nearest server which reduces the latency, the main goal of locality based balancing in content delivery network is to select content from the nearest content server for better performance. The nodes in the network coordinates are calculated by maintaining the latency into a space metric. the prediction between network coordinates is actual
communication latencies. To reflect changes in internet latencies are dynamically updated in network coordinates. The estimation of missing measurement is a main benefit of network coordinates. Without having the relative proximities it allows internet node for all measurements. The violation of triangular inequality in the internet latencies makes measurement without any errors which results in minimum accurate latency prediction. There is also an overload due to constant measurement while maintaining network coordinates in background. There is an algorithm called Vivaldi algorithm which maintains the network coordinates it is a decentralized algorithm with help of spring relaxation technique it computes network coordinates. The network coordinate nodes are designed as mass points and latencies as springs between individual nodes. The nodes of network coordinates are changed when node attract and repel with other node.

IV. EQUATIONS

Load balancing, its very easy to identify the benefits of network coordinates by achieving the locality awareness. Here we assume that both client and server are known to network coordinates and client request to the server directly which are located at the closest or nearest network coordinates and the content servers use their neighbour servers to replicate the popular content which reduces the request times by replicating content. Load balancing, its very easy to identify the benefits of network coordinates by achieving the locality awareness. Here we assume that both client and server are known to network coordinates and client request to the server directly which are located at the closest or nearest network coordinates and the content servers use their neighbour servers to replicate the popular content which reduces the request times by replicating content. For latency prediction the direct measurements are less important in content delivery network for latency prediction the direct measurements are less important in content delivery network. There will be marginal impact on total request time for huge content by choosing a nearest server which reduces the latency. The main goal of locality based balancing in content delivery network is to select content from the nearest content server for better performance.

V. DYNAMIC LOAD BALANCING

For a parallel delivery of content in content delivery networks is an most important problem for efferent delivery of content to the clients. To overcome this problem we can schedule the distribution process for replicating to the other content servers dynamically of efficient delivery of content. The main advantage of balancing the load on content servers are to complete the task as early as possible, minimizing accessing time of the client to get desired content which was requested by client. Form implementing dynamic load balancing we use different methods they are Round Robin, Least Connect, perceptive algorithm. In round robin method on succession of each node the balancer sends or replicates the content request to each node according to the traffic among the Content Network. In least connect, here the replication will be done according to the number of connections opened in the content network and replicates according to current no. of open connections in the content server. In perceptive algorithm the requests are replicated according to the response time of each individual nodes in the content delivery network.

Load balancer, there are different variety of features maintained by load balancer according to software and hardware. The main fundamental of load balancer is to replicate the incoming replicate the incoming content requests over number of backend content servers. The features of load balancers are asymmetric load, here the replication of content servers are done by calculating the ratio of work loaded served servers manually. Priority activation, the content server is made to online if the number of available servers are dropped from the content network or else the servers which are in standby.

VI. CDN DESIGN

Based on load aware network coordinates the designing and architecture of content delivery network is done and show how the requests among the content server passes and how they are replicated among themselves.
Global content requests dynamically routed to the nearest content servers, increasing the uploading speed of page, maximizing the bandwidth and delivering same content regardless of load and network congestions. According to network traffic and number of nodes, the network algorithm chooses the optimal routing option to deliver the content with scalability and reliability. It provides redundancy during internet malfunctions. Content duplication also avoided against loss or degradation of data. The clients requests the content server to fetch the content by CDN, which are served by content servers according to their mapping strategy. The main goal of CDN is to minimize the requested time i.e., until the user retrieve the desired content.

VII. CONCLUSION

In this Content Delivery Network, to accomplish a scalable and reliable service it should maintain two different properties. First, load awareness to partition the requests over a group of servers that replicates the content, balanced load and network congestion. Load balancing, its very easy to identify the benefits of network coordinates by achieving the locality awareness. Here we assume that both the client and server are known to network coordinates and client request to the server directly which are located at the closest or nearest network coordinates and the content servers uses their neighbour servers to replicate the popular content which reduces the request times by replicating content.

VIII. REFERENCES


Author Profile

Mahendar Reddy Bheemreddy, pursuing M.Tech (IT) from Holy Mary Institute of Technology and Science, Keesara, Ranga Reddy Dist., Affiliated to JNTU-HYDERABAD.

V. Krishna, Associate Professor & HOD (IT Department) at Holy Mary Institute of Technology and Science, Keesara, Ranga Reddy Dist., Affiliated to JNTU-HYDERABAD