www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issue 12 Dec. 2016, Page No. 19385-19391

Obtaining Good Heuristics for Big Data Broadcasting Problem Using Novel Pipeline Approach.

Prof. Priyadarshani Kalokhe, Abhishek Kumbhar, Jyoti Shelke, Rinkal Bari, Snehal Shewale
hodcomp.acem@gmail.com
abhi6034@gmail.com
jyotishelke107@gmail.com
Rinkalbari19@gmail.com
snehals2424@gmail.com

Abstract – *Today big data computing is a critical challenge faced by many industries .Everyday up to terabytes to petabytes of the data is generated in the cloud computing paradigm. Thus there is a drastic demand for building a service to distribute, manage and process massive data sets that has risen tremendously. Engineers and researchers are dealing with these large dataset. In this paper, we investigate the big data broadcasting problem for single source node to broadcast a big chunk of data to a set of nodes with the objective of minimizing the maximum completion time. These nodes may be present in the same data center or across geo distributed data centers. We model the big data broadcasting problem into a* ***LOCKSTEP BROADCAST TREE (LSBT).****The main idea of this model is to define a basic unit of upload bandwidth r, such that a node with capacity c broadcast a data to a set of nodes at the rate r. We divide the broadcast data into m chunks. These data chunks can then be broadcasted using a pipeline approach.*

KEYWORDS: *Big data computing, delivery algorithm, cloud computing.*

1. INTRODUCTION

Today the major efforts are being taken on big data computing. The advances in the ICT technologies such as computation, communication and storage have resulted in enormous data sets in business, science and the society being generated and analyzed to explore the value of those data. Today engineers and researchers are dealing with terabytes and petabytes of datasets in cloud computing paradigm. For example, the industries such as Google, Yahoo, and Amazon collect huge amount of data every data for proving information services freely to people in useful ways. Thus the demand for building a distributed service stack to efficiently distribute, manage and process massive datasets has risen drastically. Earlier there were several efficient techniques proposed to manipulate huge amount of data ranging from terabytes to petabytes of data on as many as thousands of machine. For big data computing data transmission overhead is a significant factor of a job completion time. It is shown that the total amount of data transmission time occupies approximately one third of the jobs running time in the hadoop tracing logs of Facebook.

There are lot of application domains which widely apply broadcasting operations such as scientific data distribution, database transaction logs backup, latest security patches, multimedia streaming applications among distributed data centre. Since, the size of data become so enormous, the impact of broadcasting operation also becomes increasingly significant. Therefore

In this paper we focus on big data broadcasting operation which is one of the most essential communication mechanism distributed system.

1.1 Potential application of LSBT

There are at least three broad applications where LSBT can be applied:1)topology control in BitTorrent like systems; 2) data broadcasting in cloud computing software stack; 3) energy conservation in peer-assisted content delivery services. We consider these in the context of network systems that are heterogeneous network environments. LSBT could be useful in topology control in BitTorrent-like systems. BitTorrent is a peer-to-peer application that aims to enable the fast and efficient distribution of large files among a large group of nodes. In BitTorrent, each peer maintains a constant number of concurrent upload connections (usually five).

LSBT can be integrated into the cloud computing software stack. For example, Apache Hadoop1 is a software framework that allows for the distributed processing of large data sets across clusters with thousands machines. Thus an efficient and scalable way to disseminate a large volume of data among machines is a significant challenge in Hadoop [8]. Our LSBT can be integrated into the delivery services of Open Stack software stack.

2. Related work

It focus on the big data broadcasting operation that is one of the most essential communication mechanisms in distributed systems. There are a lot of application domains that widely apply broadcasting operations, such as scientific data distributions [9], database transaction logs backups, the latest security patches, multimedia streaming applications, and data replica or virtual appliance deployment [3] among distributed data centers. Since the size of data becomes so enormous, the impact of broadcasting operation also becomes increasingly significant. We introduce the novel LockStep Broadcast Tree (LSBT) to model the Big Data broadcast problem [4], [5]. LSBT is a broadcast tree where data chunks can be sent in a pipelined fashion with a good throughput.

2.1 MapReduce: Simplified Data Processing on Large Clusters

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a map function that processes a key/value pair to generate as set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model. Programs written in this function a style are automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system.

Advantage:

This technique is simple and powerful. Interface that enables automatic parallelization and distribution of large-scale computations, combined with an implementation of this interface that achieves high performance on large clusters of commodity PCs.

2.2 Data Replication in Data Intensive Scientific Applications with Performance Guarantee

From this paper we demonstrate that the distributed caching technique significantly outperforms an existing popular file caching technique in Data Grids, and it is more scalable and adaptive to the dynamic change of file access patterns in Data Grids. Replication is an effective mechanism to reduce file transfer time and bandwidth consumption in Data Grids placing most accessed data at the right locations can greatly improve the performance of data access from a user's perspective

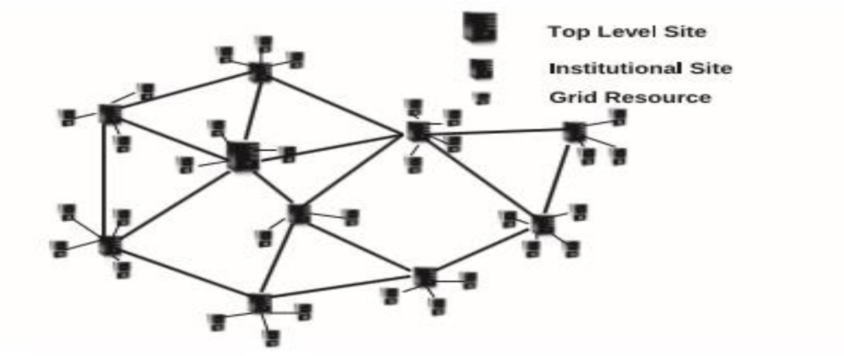

Fig 1 – Peer to Peer connectivity.

From this we selected centralized data replication algorithm which is a greedy algorithm. First, all Grid sites have all empty storage space (except for sites that originally produce and store some files). Then, at each step, it places one data file into the storage space of one site such that the reduction of total access cost in the Data Grid is maximized at that step. The algorithm terminates when all storage space of the sites has been replicated with data files, or the total access cost cannot be reduced further.

2.3 Efficient Multimedia Broadcast for Heterogeneous Users in Cellular Networks

The Efficient Multimedia Broadcast and Multicast Services (MBMS) to heterogeneous users in cellular networks imply adaptive video encoding, layered multimedia transmission, optimized transmission parameters, and dynamic broadcast area definition.A multi-dimensional approach for broadcast area definition, which provides an effective solution. By using multi-criteria K-means clustering, our scheme provides users with high levels of Quality-of-Experience (QoE) of multimedia services. Adaptive video encoding and allocation of radio resources (i.e., time-frequency resource blocks, and modulation and coding scheme) are performed based on user spatial distribution, channel conditions, service request, and user display capabilities.

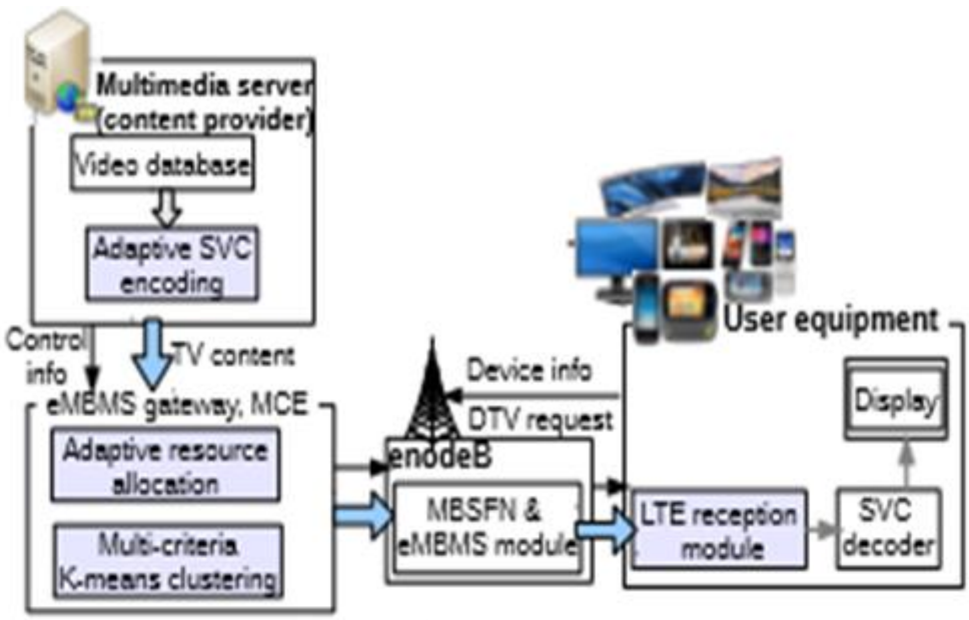

Fig 2 – Multimedia data transfer

This approach accounts for heterogeneous display capabilities and channel conditions of the users in the network, as well as for different multimedia service requests.

2.4 Product Comparisons with Improved Reliability

In this we model the Big-data Broadcasting to build a Scalable, Efficient and Precise system for service level comparison between products in Market. The product comparisons for users request will b given using well known TSV (Tab Separated Value) Format is used. The proposal of TSV format gives service comparison, recommendations in a tabular format based on the user's request. Here we model the service level comparisons and completes the banking process from a single gateway to all service providers. With this single gateway banking process the users need not to maintain multiple E-commerce accounts. Here we model Map Reduce Algorithm. The Broadcasted data's are received using different web services. Map Reduce collects data from different web servers and gives service level comparisons and recommendations for the selected products. Recommendations for purchasing a product are given using CBR (case based recommendations) and the transaction process can be done. Hence, our applications stand unique and do not rely on the single service provider.

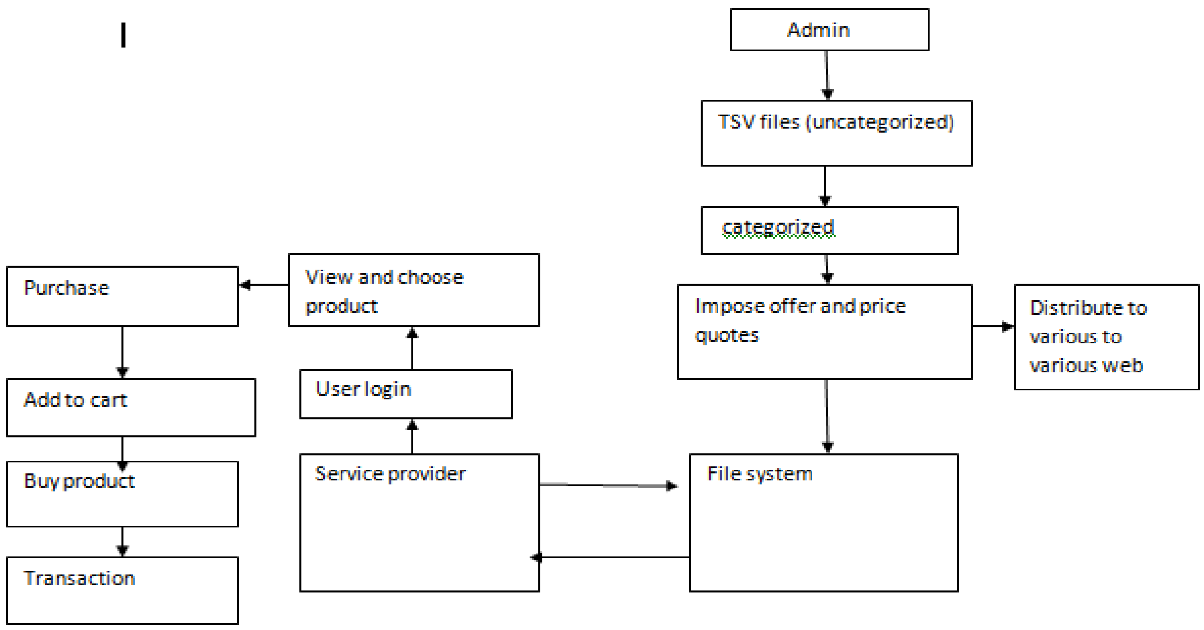

Fig 3 – Product transition diagram

2.5 Managing Data Transfers in Computer Clusters with Orchestra

Cluster computing applications like MapReduce and Dry transfer massive amounts of data between their computation stages. These transfers can have a significant impact on job performance, accounting for more than 50% of job completion times. Despite this impact, there has been relatively little work on optimizing the performance of these data transfers, with networking researchers traditionally focusing on per-flow traffic management. We address this limitation by proposing global management architecture and a set of algorithms that (1) improve the transfer times of common communication patterns, such as broadcast and shuffle, and (2) allow scheduling policies at the transfer level, such as prioritizing a transfer over other transfers. Using uproot type implementation, we show that our solution improves broadcast completion times by up to 4.5×compared to the status quo in Hadoop. We also show that transfer-level scheduling can reduce the completion time of high priority transfers.

2.6 Analyzing and Improving a BitTorrent Network's Performance Mechanisms

We find BitTorrent to be remarkably robust and scalable at ensuring high uplink bandwidth utilization. It scales well as the number of nodes increases, keeping the load on the origin server bounded, even when nodes depart immediately after completing their downloads. The LRF policy performs better than alternative block-choosing policies in a wide range of environments (e.g., flash crowd, post-flash crowd situations, small network sizes, etc.) By successfully getting rid of the last block problem, it provides a simpler alternative to previously proposed source coding strategies, e.g. Second, we find that BitTorrent shows sub-optimal behavio

2.7 Broadcasting on Large Scale Heterogeneous Platforms under the Bounded Multi- Port Model

We consider the classical problem of broadcasting a large message at an optimal rate in a large scale distributed network under the multi-port communication model. From an optimization point of view, we aim both at maximizing the throughput (i.e., the rate at which nodes receive the message) and minimizing the degree of the participating nodes, i.e., the number of TCP connections they must handle simultaneously. The goal is to organize data transfers so as to maximize the throughput (or minimize the make span for a given message size). On the other hand, in the context of large scale Internet level platforms, the goal is to find the topology (i.e., the overlay network) that maximizes the throughput. The one-to-all broadcast, or single-node broadcast, is the most primary collective communication pattern: initially, only the source processor holds the data that needs to be broadcast; at the end, there is a copy of the original data residing at each processor
Advantage - Minimizing the degree of the participating nodes, to organize data transfers so as to maximize the throughput.

2.8 An Agent-Based Approach for Modeling Peer to Peer Networks

This article has demonstrated how the use of ABM tools such as Net Logo can indeed provide many advantages over classical Peer-to-Peer simulation tools. P2P systems can mainly be divided into three different architectures, namely, centralized, decentralized structured and decentralized unstructured .Agent-based Modeling (ABM) is a type of computational modeling for simulating the actions and interactions of autonomous agents [7]. These agents can be either individual or collective entities such as organizations or groups. In this section, an introduction to Agent-Based Modeling will be provided. Also, we shall discover and discuss the differences between ABM approaches and classical discrete-event approaches for modeling P2P networks.ABM can be used to develop agents that are proactive, autonomous and intelligent. Proactiveness refers to agents' capability of taking the initiative, enabling them to communicate with other agents and make decisions on their own. Most of the conventional discrete-event simulations are not designed to deal with such a model as the number of events can be exponentially increased, which makes the model inefficient and hard to analyze [7], [8]. Agent-based modeling provides a description of complex systems in the P2P paradigm with a higher level of abstraction.This can provide a better understanding of the P2P paradigm, which, in turn, may prove helpful for improving some its required aspects

2.9 P2P Streaming Capacity under Node Degree Bound

We focus on the limit of P2P streaming rate under node degree bound, i.e., the number of connections a node can maintain is upper bounded**.** P2P streaming can be modeled as multiple multicast trees superimposed on top of the overlay graph. The streaming capacity problem depends on the constraints on the graph and tree properties. The P2P streaming capacity problem on P2P overlay graph with capacity constraints and degree bounds on *nodes*. Thus it is different from a large body of work focusing on underlay graph with node degree constraints and capacity constraints on links

3. Conclusion

We conclude that from the different algorithms and systems we learnt how efficiently we can broadcast the data ,share the data in centralized system, distributed system, peer 2 peer networks. We have introduced a framework that is prepared to do naturally extricating information parallelism from synchronising applications. Our work contrasts from former work by having the capacity to concentrate such parallelism with wellbeing sureties in the vicinity of administrators that can be stateful , particular, and client denied. We have exhibited that these procedures can scale with accessible assets and exploitable parallelism.

4. References

[1] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," Proc. Oper. Syst. Design Implementation, 2004, pp. 137–150.

[2] D. Nukarapu, B. Tang, L. Wang, and S. Lu, "Data replication in data intensive scientific applications with performance guarantee," IEEE Trans. Parallel Distrib. Syst., vol. 22, no. 8, pp. 1299– 1306, Aug. 2011.

[3] M. Banikazemi, V. Moorthy, and D. Panda, "Efficient multimedia broadcast for heterogeneous users in cellular networks " in Combinatorial Algorithms. New York, NY, USA: Algorithmics Press, 1972

[4] M. Chowdhury, M. Zaharia, J. Ma, M. I. Jordan, and I. Stoica, "Managing data transfers in computer clusters with orchestra," Proc. ACM Special Interest Group Data Commun., 2011, pp. 98–109.

[5] A. R. Bharambe, C. Herley, and V. N. Padmanabhan, "Analyzing and improving a bittorrent networks performance mechanisms," in Proc. IEEE Int. Conf. Comput. Commun., 2006, pp.

[6] O. Beaumont, L. Eyraud-Dubois, and S. K. Agrawal, "Broadcasting on large scale heterogeneous platforms under the bounded multi-port model," in Proc. IEEE Int. Symp. Parallel Distrib. Process. Symp., 2010, pp. 1–11.

[7] K.-S. Goetzmann, T. Harks, M. Klimm, and K. Miller, "Agent based approach for modeling peer-to-peer networks," Proc.

[8] L. Massoulie, A. Twigg, C. Gkantsidis, and P. Rodriguez, "P2P streaming capacity under node degree bound," IEEE 30th Int. Conf. Dist. Comput. Syst., pp. 587–598, 2010.

[9] R. Thommes and M. Coates, "Bittorrent fairness: Analysis and improvements," in Proc. Workshop Internet, Telecommun. Signal Process., Dec. 2005, pp. 9–14.

[10] O. Beaumont, L. Eyraud-Dubois, and S. K. Agrawal, "Broadcasting on large scale heterogeneous platforms under the bounded multi-port model," in Proc. IEEE Int. Symp. Parallel Distrib. Process. Symp., 2010, pp. 1–11.

[11]K. Wang, J. Li, and L. Pan, "Fast file dissemination in Peer-to-peer networks with upstream bandwidth constraint," Future Generation Comput. Syst., vol. 26, pp. 986–1002, Jul. 2010.