

LSBT-A NEW IMPEND FOR BIG DATA BROADCASTING

P.Naresh¹, V.Krishna², P.Rajyalakshmi³

¹Assistant Professor in CSE, HITS, Hyderabad

²Associate Professor in CSE, HITS, Hyderabad

³Assistant Professor in CSE, HITS, Hyderabad

Abstract

In contemporary period huge amount of data will be stored on cloud server, these types of data is referred as Big-Data. Big data computing is a complicated issue for broadcasters and engineers because they are dealing with datasets of petabytes in the cloud. To manage process and broadcast these massive data has increased drastically. Here we propose an approach to broadcast the data by large chunks to group of nodes that can reduce the maximum completion time. Here we model a Lock Step Broadcast Tree (LSBT) which will define an upload bandwidth(r) and each node capacity(C) at c/r per children. After these data sets divided into chunks, and are broadcasted in a pipeline manner. In homogeneous network environment the capabilities of each node is C , and uplink rate r which gives less maximum completion time. In heterogeneous networks capabilities $C_1, C_2 \dots C_n$, so optimal uplink rate r will be calculated by $O(n \log_2 n)$ algorithm there after we construct a LSBT. Finally our results show less computational complexity and less maximum completion time over other broadcasting techniques.

Keywords: chunks, computational complexity

1. INTRODUCTION

Now-a-days large amount of data is going to store in cloud server. These data further managed and processed in cloud [10]. Big data computing and broadcasting is a challenging issue for IT industry [1].in recent days engineers and researchers are dealing with heterogeneous database[2][3].companies like amazon,flipcart,google maintains huge amount of data for providing services to their users. In ancient days there are several techniques to broadcast and process huge amount of data. Google presents MapReduce framework [4] and Bigtable [5] for storing formatted data on hundreds of machines. There are many issues while developing such applications like data decomposition, exchanging huge data among number of nodes. Data transmission overhead is a key factor for Big data computing.

Big data broadcasting is an essential mechanism to communicate nodes in distributed systems. There are many applications that broadcast data but the data size varies pet bytes. Since data size becomes huge and enormous, there is an impact on broadcasting operations. Here we focus on big data broadcasting problem in heterogeneous networks'. In these types of networks, several nodes have different uploading capabilities. The problem is how nodes receive data in minimum total transmission time.

Assume that there are n nodes (n_1, n_2, \dots, n_n) in heterogeneous network and their upload capabilities are different i.e(c_1, c_2, \dots, c_n) measured in kilobytes per second. The source data is divided into 's' chunks of equal size. Here we introduce lockstep broadcast tree to model big data broadcasting problem. [6][7].LSBT is a tree where data chunks transferred in a pipelined manner with a less maximum transmission time. The theme of this tree model is

to define upload bandwidth r ,and uplink of each node is divided into different connections. After that broadcast data divided into s chunks and broadcast down through nodes in pipeline manner. In homogeneous networks all nodes have same uplink capacity c , and uplink rate r ,then the LSBT is $c/2$. In heterogeneous networks we use $O(n \log_2 n)$ algorithm for selecting uplink rate r , there after we construct an optimal LSBT.

2. BACKGROUND AND RELATED WORK

S data chunks are held by a single source node in a network. Here data broadcasting problem comes at sending these 's' data chunks towards 'n' nodes attaining less time. This time is totally depended on uplink capacities of nodes. In this paper we mainly focus on broadcasting problem in heterogeneous networks. Earlier so many researches done on non chunk based approach. Khuller and Kim [6] proved that problem by a single chunk in heterogeneous networks. They also showed that Fastest-Node-First method for solving this problem and showed less computation time. Among all Chunks based models, optimal solution showed in paper [7] by uplink sharing model and by implementing a mixed integer linear programming. Many decentralized systems are proposed to transfer chunks via overlay topology. Bittorrent[8],splitsteam[9] are examples of overlay approaches.

In order to reduce the complexity of broadcast problem [6][7], we introduce lockstep broadcast tree. By this we achieve minimum completion time by optimizing bandwidth r . Here data is divided into chunks and sent in a pipeline manner. This LSBT is very useful for host applications some of them are topology control, data broadcasting in cloud, energy conservation in peer to peer data delivery services. In previous studies the LSBT was implemented for

homogeneous network systems. In these systems all nodes are having same uploading capabilities(c). In paper [7] each node costs one unit of time, then maximum completion time of optimal solution is $s+\lceil\log_2 n\rceil$, where s is number of chunks and n the number of nodes. In the next we discuss about this problem in heterogeneous network systems by LSBT approach.

3. LSBT IMPLEMENTATION

Here we implement LSBT in heterogeneous networks, which having different upload capabilities.

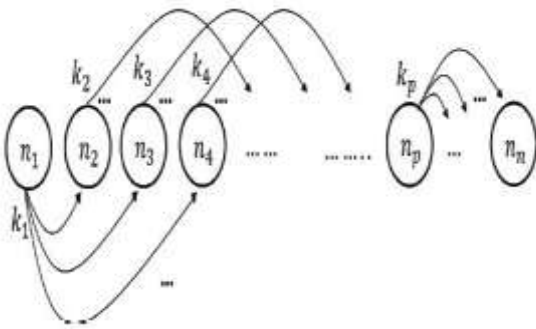


Fig 1.Building a LSBT

First we use algorithm 1 to find out the candidate set based on capabilities C , and upper bound and lower bound values of r . There after we implement algorithm 2 for optimal LSBT based on r values. Fig 1 shows the illustration of constructing a LSBT.

Algorithm 1 discretization algorithm for the candidateset
Input: a set of upload capacities c and the upper and lower bounds of r

```

BEGIN
UnionSet ← empty
CandidateSet ← empty
for i ← 1 to n do
UnionSet ← UnionSet ∪ ci
end for
UnionSet ← Sort(UnionSet)
for j ← 1 to n - 1 do
for all u in UnionSet do
r ← u=k
if r ≥ upper then
Continue
Else if r < lower then
Break
End if
CandidateSet ← CandidateSet ∪ r
End for
End for
Return CandidateSet
END
Output: CandidateSet

```

After implementing candidate set we find out bandwidth r n maximum completion time D by using below algorithm 2.

Algorithm 2 The r -search algorithm for the optimal LSBT

Input: a set of upload capacities c and *CandidateSet*

```

BEGIN
CandidateSet ← Sort(CandidateSet)
D ← ∞
r ← empty
for h ← 1 to 2(⌊log2 n⌋ + 1) do
rt ← 1
lt ← Sizeof(CandidateSet)
while rt ≤ lt do
mid ← ⌊(rt + lt)/2⌋
r ← CandidateSet[mid]
t ← GLSBT(c; r)
if lt - rt = 1 and t:Height = h then
d ← t:BroadcastingTime
if d < D then
r ← r
D ← d
end if
else
break
end if
if h ≤ t:Height then
rt ← mid
else
lt ← mid - 1
end if
end while
end for
return r and D
END
Output: r and D

```

4. RESULT ANALYSIS

In this paper we show the performance of our LSBT by using numerical analysis. Here we implemented two approaches .one is FNF approach and another is LSBT approach. We first analyse algorithm as the number of nodes in network increases the size of candidate set slowly vary. Compared with naive approach our approach gives good performance and this gives good heuristic for reducing candidate set.

We now show the maximum completion time for two algorithms. We select the networks with the number of nodes (n) 100,1000,10000 nodes. Size of file is 50 mb and chunks are 100. Figure 2(a) shows number of nodes versus maximum completion time. As going on increasing no of nodes the maximum completion will reduce. Figure 2(b) shows the computation time as number of nodes increases.

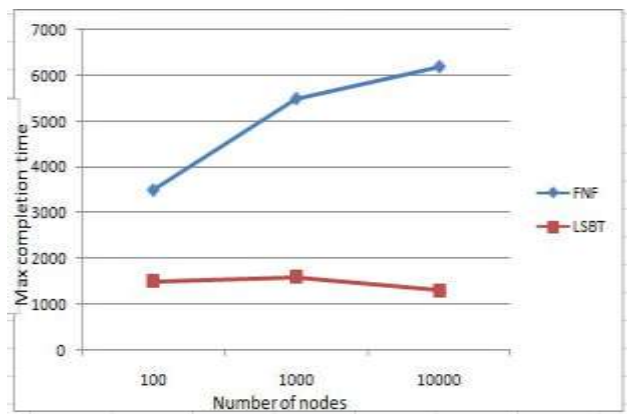


Fig 2(a) Max completion time

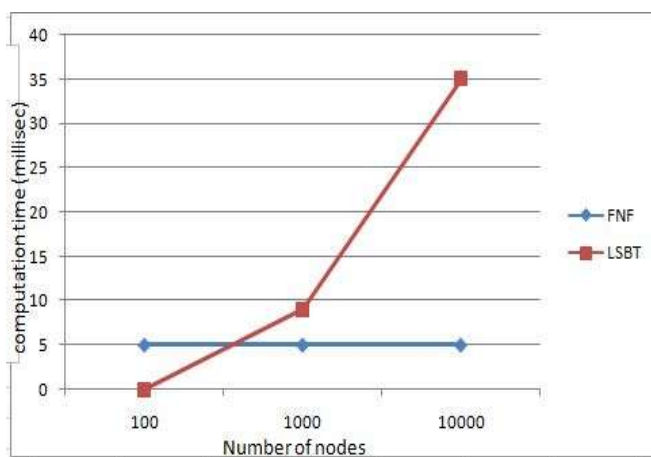


Fig 2(b) Computation time

5. CONCLUSION

In this paper we focused on big data broadcasting problem by LSBT algorithm. Here we implement both in homogeneous and heterogeneous network environments. In homogeneous network environment the capabilities of each node is c , and uplink rate r which gives less maximum completion time. In heterogeneous networks capabilities c_1, c_2, \dots, c_n , so optimal uplink rate r will be calculated by $O(n \log_2 n)$ algorithm there after we construct a LSBT. Our results show less computational complexity and less maximum completion time over other broadcasting techniques. Future work involves more heuristic methods to minimize computational complexity and reduce overheads.

REFERENCES

- [1] Chi-Jen Wu, Chin-Fu Ku, Jan-Ming Ho, *IEEE Member*, Ming-Syan Chen, *IEEE Fellow* "A novel approach for efficient big data broadcasting".
- [2] R. E. Bryant, R. H. Katz, and E. D. Lazowska, "Big-data computing: Creating revolutionary breakthroughs in commerce, science, and society," *In Computing Research Initiatives for the 21st Century.*, 2008.
- [3] A.Szalay and J. Gray, "2020 computing: Science in an exponential world," *Nature 440, 413-414, March*, 2006.

- [4] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," *Proc. of Operating Systems Design and Implementation (OSDI)*, 2004.
- [5] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, , and R. E. Gruber, "Bigtable: A distributed storage system for structured data," *Proc. of Operating Systems Design and Implementation (OSDI)*, 2006.
- [6] S. Khuller and Y.-A. Kim, "Broadcasting in heterogeneous networks," *Algorithmica*, vol. 48, no. 1, Mar. 2007.
- [7] J. Munding, R. Weber, and G. Weiss, "Optimal scheduling of peer-to-peer file dissemination," *Journal of Scheduling*, vol. 11, no. 2, 2008.
- [8] B. Cohen, "Incentives build robustness in bittorrent," *Proc. of ACM P2PECON*, 2003.
- [9] M. Castro, P. Druschel, A.-M. Kermarrec, A. Nandi, A. Rowstron, and A. Singh, "Splitstream: High-bandwidth multicast in a cooperative environment," *Proc. of ACM SOSP*, 2003.
- [10] S. Saroiu, K. P. Gummadi, and S. D. Gribble, "A measurement study of peer-to-peer file sharing systems," *Proc. of Multimedia Computing and Networking (MMCN)*, 2002.