# 《A Novel Pipeline Approach for Efficient Big Data Broadcasting》阅读报告

姓名：牛菲菲 学校：天津大学

电话：15122269673 邮箱：[niufeifei@tju.edu.cn](mailto:niufeifei@tju.edu.cn)

博客地址：

<http://www.cnblogs.com/niufeifei/p/7183944.html>

1.论文基本信息

|  |  |  |  |
| --- | --- | --- | --- |
| 作者: | [Chi-Jen Wu](http://dl.acm.org/author_page.cfm?id=81371593289&coll=DL&dl=ACM&trk=0&cfid=960596217&cftoken=37926279) | [Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan](http://dl.acm.org/inst_page.cfm?id=60005429&CFID=960596217&CFTOKEN=37926279) | |
|  | [Chin-Fu Ku](http://dl.acm.org/author_page.cfm?id=99658623018&coll=DL&dl=ACM&trk=0&cfid=960596217&cftoken=37926279) | [Institute of Information Science, Academia Sinica, Taipei, Taiwan](http://dl.acm.org/inst_page.cfm?id=60013651&CFID=960596217&CFTOKEN=37926279) | |
|  | [Jan-Ming Ho](http://dl.acm.org/author_page.cfm?id=81552514656&coll=DL&dl=ACM&trk=0&cfid=960596217&cftoken=37926279) | [Institute of Information Science and the Research Center of Information Technology Innovation, Academia Sinica, Taipei, Taiwan](http://dl.acm.org/inst_page.cfm?id=60013651&CFID=960596217&CFTOKEN=37926279) | |
|  | [Ming-Syan Chen](http://dl.acm.org/author_page.cfm?id=81508693932&coll=DL&dl=ACM&trk=0&cfid=960596217&cftoken=37926279) | [Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan](http://dl.acm.org/inst_page.cfm?id=60005429&CFID=960596217&CFTOKEN=37926279) | |
| 期刊: | | |
|  | | |
| IEEE Transactions on Knowledge and Data Engineering | | |
| Volume 28 Issue 1, January 2016  Pages 17-28 | | |

名词解释

同质网络环境：网络中每个节点具有相同的上传能力c

异质网络环境: 网络中每个节点的上传能力c不一定相同

FNF启发式：最快节点优先启发式技术[32]是用于在异构网络中查找广播树的常用集中式算法。该算法简单易用。我们简要地描述如下：在FNF启发式的每次迭代中，它从已经接收到消息的节点集合选择一个发送者和，从未接收到消息的集合中选择接收者。显然，在每次迭代中，FNF启发式需要做出两个决定。首先，它必须决定哪个发送者要将消息发送到新的接收者。第二个决定是在尚未添加到树中的节点之间选择新的接收者。 FNF启发式是始终选择最快的发送者，最快的接收者是快速传递消息的最佳方法。通过这种方式，FNF启发式可以生成可以实现广播操作的贪心树。请注意，FNF启发式算法仅限于一条消息。因此，我们使用它逐个广播多个块。

DIM-Rank：DIM-Rank技术是一种用于异构网络中多块传播的集中式启发式算法。在DIM-Rank中，使用集中式调度器基于所有相关节点的成对带宽和时延来管理节点之间的数据块传递。 另一方面，调度器用于在每次迭代中选择发送者，接收者和数据块。 在DIM-Rank中，有两个调度指标，以帮助调度器做出决策。 首先，Chunk Spread是具有相同数据块的节点的总数。 第二，Node Rank是一个节点包含的所有数据块的总和的倒置，例如，如果节点包含很少数据块或许多块，则其秩更高。 在每次迭代中，DIM-Rank技术首先将节点的按node rank从最低到最高排序，然后按照上传容量（从高到低）对节点进行排序。 排序后，调度器从排序列表中为每个发送者选择接收者，并选择lowest-spread块进行传播。

2.论文要解决的问题及其重要性

这篇文章主要研究从单源节点向一组节点传播大量数据的大数据广播问题，目标是使得最大完成时间最小化。

近几年来，随着计算机和信息技术的迅猛发展和普及应用，行业应用系统的规模迅速扩大，行业应用所产生的数据呈爆炸性增长。动辄达到数百TB甚至数十至数百PB规模的行业/企业大数据已远远超出了现有传统的计算技术和信息系统的处理能力，因此，寻求有效的大数据处理技术、方法和手段已经成为现实世界的迫切需求。

在过去的十年，出现了几种有效的技术来操作在数万台机器上的大量的数据，从T到P级别。例如，谷歌展示了一个分布式计算框架，叫MapReduce来有效处理大规模的数据并起初Bigtable用于早上千台机器上存储结构数据。这些技术允许用户实现数据并行。在发展MapReduce应用上，有许多重要问题，比如设计有效的策略用于数据分解，负载均衡，以及在大量节点之间的数据交换。尤其是，对于大数据计算，数据传输开销是完成时间的重要因素。例如，facebook的hadoop跟踪日志显示数据传输的时间占据将近三分之一的总时间。

这篇论文主要集中于分布式系统中最重要的通信机制之一—大数据广播操作。有许多的领域广泛应用广播操作，例如科学数据分发，数据库事务日志备份，最新的安全补丁，多媒体流应用和数据副本或分布式数据中心的虚拟设备部署。随着数据大小变大，广播操作的影响也因此而变得重要。异构网络环境中考虑大数据传播问题，在这个条件下，节点会有不同的上传能力。大数据传播问题就是节点如何以相对较短的总传输时间来获得给定的大数据。

假设在异构网络环境中有n个节点，，, ,…,. 是传播源，将所要传播的数据平均分成m块，将这些数据传输给其他所有的节点。上传能力记为,, ,…,（KBps）. 并且假设每个节点的下载能力不小于上传能力。

主要集中于以下问题:

1. 具有固定上传速率的单覆盖树与广播操作的关系是什么?
2. 在异构网络中，如何构造单覆盖树以实现最大完成时间的最小化？

3.论文提出的解决方法，效果

这篇文章介绍了LockStep Broadcast Tree (LSBT)来模拟大数据传播问题。同时还提出了多项式时间算法来选择构建最优LSBT的最优上传速率。LSBT是一个能以较好的吞吐量以流水线方式传播数据块的广播树。

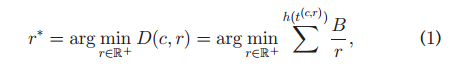
LSBT模型的主要思想是定义一个上传带宽的基本单位r，使得具有上传能力c的节点以速率r向个子节点传输。r就是作为LSBT问题的一部分需要优化的参数。我们进一步将传播数据分成m 块，然后这些块就可以沿着LSBT以流水线的方式传播。在同质网络环境中，每一个节点有相同的上传能力c，可以证明LSBT的最优上传速率为c/2或者c/3，使得最大完成时间最小。在异构网络环境中，我们提出一种复杂度为算法来选择最优上传速率，并构造最优LSBT。

（1）LSBT

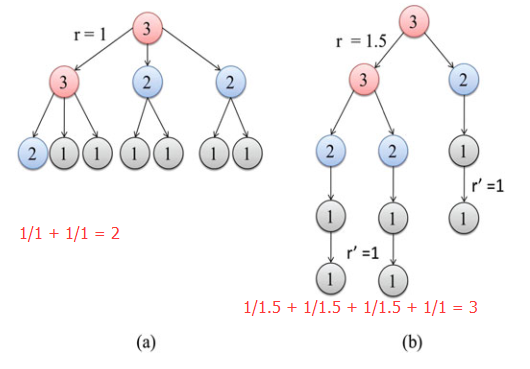
单个LSBT的性能目标：通过优化LSBT节点的基本带宽分配，r，来实现最小完成时间。给定n个节点N={，, ,…,},每个节点通过上传能力和块的大小B连接到网络中, LSBT问题就是决定每一个上传连接的上传带宽, 来构建LSBT t，每一个节点为其子节点分配上传带宽，以此来最小化传输一个数据块的最大完成时间D。

对于节点上传连接数目取决于上传能力，比如： =

最大完成时间D的定义如下：



代表上传能力c与上传带宽r组成的LSBT，h()返回LSBT 的高度。这个式子只计算了数据块从根节点至叶节点的传播时延。另外，LSBT模型通过建立一个单一的传播树来解决数据传播问题，树的节点以流水线的模式传播数据块。因此，最大完成时间D是每一层LSBT 的数据块（如：B/r）传播时间的加和。如下图所示：给定一个有11个节点的集合，上传容量分别为{3, 3, 2, 2, 2, 1, 1, 1, 1,1, 1}，通过将他们按照边的数目非递增顺序排序选择可以创建一个最优LSBT。



（2）最优LSBT

给定一组上传能力c，目的是找到一个最优LSBT，使得数据块可以以流水线的方式传播。这篇文章对同构以及异构网络环境进行了彻底的分析。

1）同质网络环境

同质网络系统中，所有节点的上传能力相同，假设为c，每个周期内每个节点只能向其他节点传送一个数据块，每个节点可以同时向k个其他节点发送数据块。根据Mundinger’s的定理1，最大完成时间的最优解决方案是m + , m代表数据块的个数，n代表节点的个数。

定理1（Mundinger’s Theorem）在同质网络环境中，向n个节点发送m个数据块的最小时间周期为m + 。

在同质网络环境中，节点的上传能力均相同，因此D可以表示为如下：



令G =

D = G

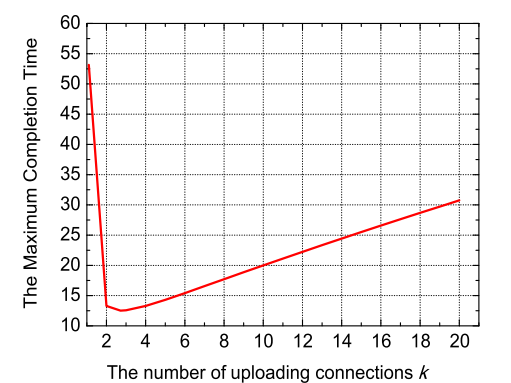
令 = 0,

= - = 0.

lnk = 1, k=e.

因此，在离散模型中有如下定理：

定理2. 在同质网络环境中，LSBT的的最优值为c/2或c/3使得LSBT最大完成时间最小.



由图可知，当k=e时取最小值。

2）异质网络环境

在异质网络环境中，节点的上传能力有可能是不同的。首先给出了对于给定速率r构建最优LSBT的算法，然后给出的上下边界。最后给出复杂度为算法来选择构建最优LSBT的最优上传带宽。

算法1.GLSBT：给定节点N={，, ,…,}，以及上传能力，r代表LSBT的上传速率，根据求取每一个节点的度，时间复杂度为O(n)。

引理1.GLSBT的时间复杂度为O(n).

定理3.给定上传速率，若LSBT t是最优的，则任何子节点的出度都不超过其父节点的出度。

引理2.（下限）在异质网络环境中，的下限大于等于, 对1<in.

引理3.（上限）在异质网络环境中， 的上限小于,对所有的1in

引理4.给定r的候选集，任何一个的取值都会影响LSBT t的高度。

引理5.所有节点的出度大于2的LSBT的高度一定小于n.

定理4.任何最优LSBT的高度不大于2×n.

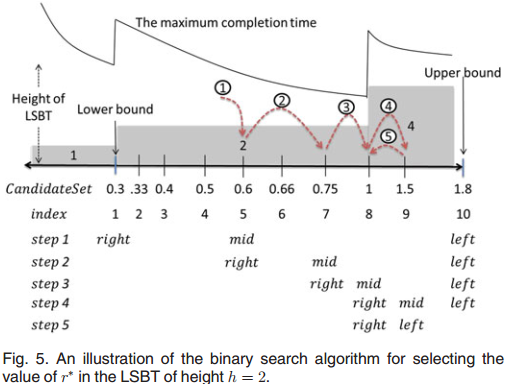
算法2.求取的候选集：通过u/k求取候选集CandidateSet，代表所有的的可能值。

算法3.搜索选择构建最优LSBT的最优：首先将CandidateSet进行排序，然后根据定理4对LSBT的高度进行遍历，求取以及，通过对比返回最小的值，由此求得最优LSBT以及最小时间D。

定理5.搜索算法的时间复杂度为。

定理6.搜索算法（算法3）可以准确的求取任何给定的上传能力c所对应的最优LSBT。

定理7.在最优LSBT中，完成数据块的传播需要的最少步数为O(m+logn)。

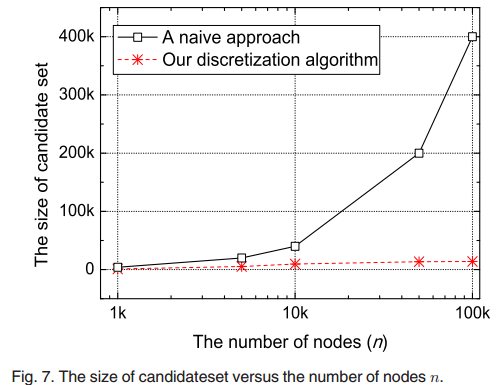


如上图演示了上传能力集合：{3, 3, 2, 2, 2, 1, 1, 1, 1,1, 1}利用算法3进行二分搜索的过程。

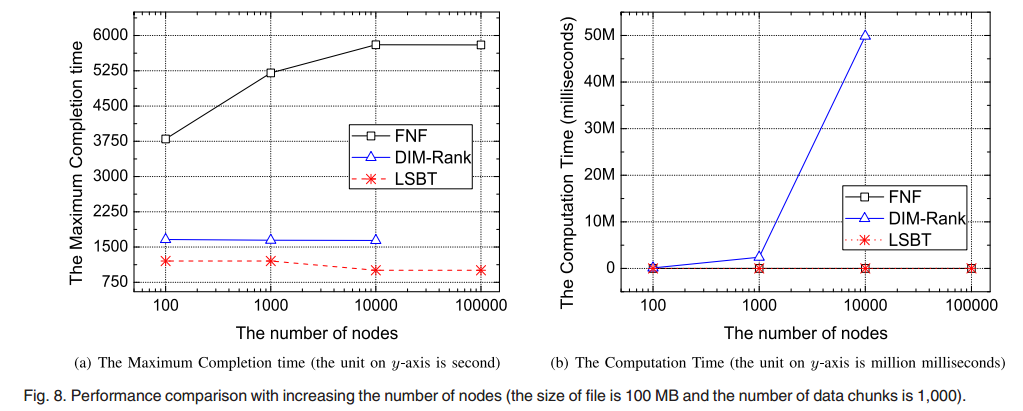
该算法总的时间复杂度为O(2××n)

实际实验结果证明这种方法相较于其他的有效方法，有较少的最大完成时间和更低的计算复杂度。

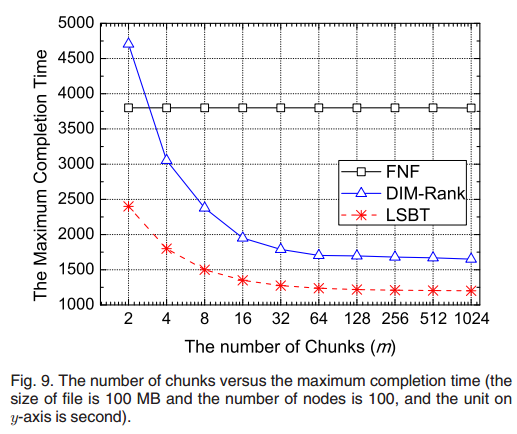
这篇论文主要实现了FNF启发式，DIM-Rank启发式以及LSBT三种方式，对这三种集中式方式进行数值评估对比。



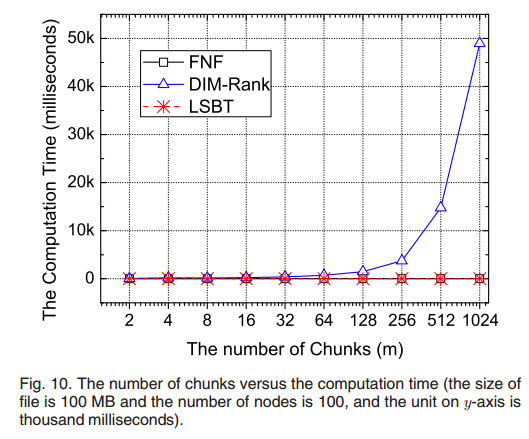
（1）上图对比了当要广播的节点数目增加时候选集的大小变化。通过对比初始方法与本文的离散算法，随着节点的数目增加，本文的算法可以大规模减小网络中候选集的大小。



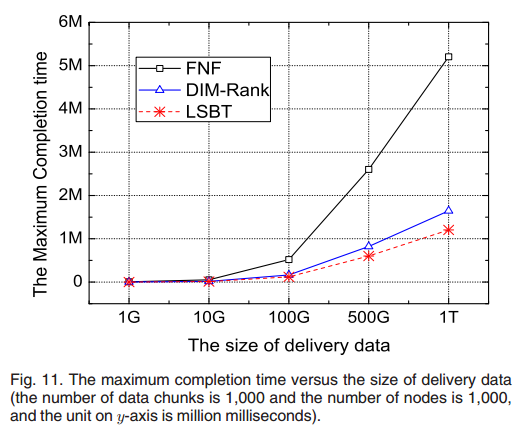
（2）上图对比了各种场景下三种算法的最大完成时间。对比n = 100，1,000，10,000和100,000个节点的网络,文件大小为100 MB，数据块数为1,000。 图8a显示了每个算法将文件广播到所有节点的总时间。图8b显示出了每个算法规划广播工作的计算时间。通过仿真结果，LSBT执行最好，而FNF启发式显示了较差的性能，因为FNF没有利用流水线的优势。DIM-Rank的计算时间是重要的，这是因为DIM-Rank的时间复杂度是。在n = 10,000网络中，DIM-Rank的计算时间需要近15个小时（在使用Intel Xeon 2.33 GHz和8 GB RAM的服务器上）。



（3）上图对比了数据块数（m）增加时的效果。文件的大小为100 MB，节点数为100.可以看到FNF的结果不取决于m的值。LSBT的最大完成时间比另外两种算法执行的最大完成时间显着降低（至少约60％）。



（4）上图对比了数据块数（m）增加对计算时间的影响。文件大小为100 MB，节点数为100.LSBT和FNF的结果不依赖于m的值，LSBT的计算时间显著低于由DIM-Rank算法执行的时间。



（5）上图显示了递送数据大小增加时的效果。数据块的数量为1,000，节点数为1,000。当传送数据的大小从1千兆字节增加到1兆字节时，LSBT的最大完成时间显著低于另外两个算法执行的最大完成时间。通过仿真结果，LSBT显示出在异构数据中心网络中提供高性能大数据传输的能力.

4.论文提出该解决方案的动机

LSBT是一个能以较好的吞吐量以流水线方式传播数据块的广播树,可以非常适合很多应用。比如：

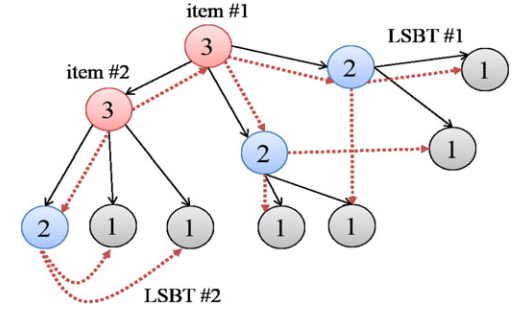
1. BitTorrentlike中的拓扑控制
2. 云计算软件堆栈中的数据广播
3. 同伴辅助内容传送服务节能

LSBT算法对BitTorrent-like的系统中拓扑控制会非常有用。BitTorrent是一个点对点的应用，旨在在一大组的节点之间进行大文件的快速有效的传送。在BitTorrent中，每个对等体都保持不断的并发上传连接数（通常为5个）。然而，如何在BitTorrent中确定适当数量的并发上传仍然是一个挑战。 所提出的LSBT算法深入探讨了如何在BitTorrent-like系统中选择并发上传数量。

大数据广播问题类似于上传共享问题，所以可以通过LSBT建模解决。

5.该解决方案有何不足之处

（1）这篇论文只针对一个LSBT进行讨论，但是在大数据计算中，一个数据通常被分配至多个数据源，目前的LSBT算法没有解决分区数据源问题。将当前LSBT算法应用于分区数据源环境的最直接的方法之一是为每个数据项构建多个LSBT广播树。例如，如下图所示，在多个数据源环境中有两个数据项，即item #1和item＃2，为这些数据项构建两个LSBT广播树，来广播多个数据源.

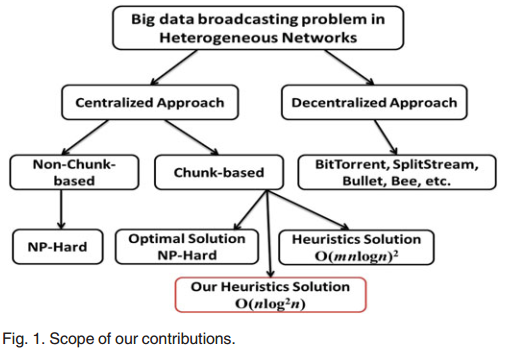


（2）此外，在inter-datacenter网络中，如何基于物理网络拓扑构建最佳LSBT也有待研究。

(3)根据这篇论文的介绍，的二分搜索时间复杂度为O(2××n)，通过对h的逐一遍历求取，计算时间浪费比较多，可以想办法求取r与h与D的关系式，通过线性规划进行求取最优解。

6.以前解决该问题的方法以及不足之处

假设m个大小相等的数据块最初属于网络中的同一个单源节点，数据广播问题就是在符合上传能力限制的条件下，以尽可能少的时间将这m个数据块传输至n个节点。这篇文章主要讨论异质网络环境中的大数据广播问题。



对于集中式方法，非基于块方法，Khuller 和 Kim提出了NP-hard来最小化异构网络环境中单块广播完成时间的问题。作者还表明，最快节点优先（FNF）启发式方法性能至少为1.5倍，FNF在许多情况下为单块广播提供了最佳解决方案。Liu表明，FNF启发式方法只在两类节点中是最优的。然而，当数据由多个块组成时数据广播问题更复杂，并且仍然存在一个待解决的问题：是否可以通过多项式时间算法求解多个块的数据广播问题？

在基于块的方法中，文献[12]提出了一种上行链路共享模型，并将数据广播问题作为混合整数线性规划（MILP）提出。然而，随着线性规划中的变量数量呈指数增长n和m，这种方法对于大的n和m是不实际的。 Goetzmann等人表明，如果对等能力是异构的和对称的，这个问题就变成强NP-hard。 最近的结果[19]提出了两种启发式算法来调度节点之间的数据块传输。 两个集中算法的时间复杂度是.

对于分散式方法，已经提出了许多通过覆盖拓扑传播块的分散系统。使用基于覆盖的方法，节点保持一组链接到其他节点的覆盖链路并在相邻节点之间交换块。BitTorrent, SplitSteam, Bullet 和 Bee就是一些基于覆盖的方法的例子。在文献[23]中，作者表明，蜜蜂可以通过模拟来接近异构网络中最大完成时间的下限。

上文提及对FNF启发式，DIM-Rank启发式以及LSBT三种方式的实现对比，LSBT综合来看是目前最优的。

7.总结

这篇文章从算法角度讨论了经典的数据广播问题。将此问题形成LSBT模型，同时考虑了设计一个有固定上传速率的单一覆盖树以及模型建立的最大完成时间。这项工作是首次对有固定上传速率的单覆盖树和最大完成时间在异质网路环境中关系的研究。同时还提出了多项式时间算法来选择构建最优LSBT的最优上传速率。算法的时间复杂度为。未来的工作涉及为数据广播问题取得良好的启发式。这个问题更有挑战性的是多个LSBT，我们将其作为未来研究的方向。此外，在inter-datacenter网络中，如何基于物理网络拓扑构建最佳LSBT也有待未来的研究。

初次接触大数据相关内容，不甚了解，相关术语解释如有误，还望指正，非常感谢！