一种内容分发网络中的快速复制方案*)

赵 进 张福炎
(南京大学计算机科学与技术系 南京 210093)

摘 要 由于在内容分发网络中,将大文件从一台服务器复制到其他服务器需要耗费大量的时间。本文首先对内容分发网络中的复制问题进行了形式化描述,然后提出了一种分布式的方案 NCOM,用于减小复制时间。方案的创新性在于,NCOM 在 CDN 中构建一个 Mesh 结构,利用多路径传输数据块,提高速率;同时也利用 Network Coding 技术来避免需要从不同路径调度不同数据块所带来的协调开销。实验结果表明,与现有方案相比,NCOM 可以显著减小复制时间。

关键词 内容分发网络,复制,Network coding

Achieving Fast Data Replication in Content Distribution Networks

ZHAO Jin ZHANG Fu-Yan
(Department of Computer Science and Technology, Nanjing University, Nanjing 210093)

Abstract It is time-consuming to replicate large-size files to a large number of edge servers in content distribution network. In this paper, a distributed scheme NCOM is proposed to provide fast data replication. Starting from problem formulation, NCOM is designed to organize the edge servers into an overlay mesh and propagate the packetized data blocks in the mesh. The novelty of this scheme lies in that network coding is introduced to alleviate the block coordination between edge servers during data propagation. NCOM is evaluated by simulations, which shows that the introduction of network coding to content distribution brings significant benefits in reducing replication time.

Keywords Content distribution network, Replication, Network coding

1 引言

Internet 上的很多应用,比如热点视频文件的下载、软件补丁的更新等,会同时有大量用户。采用单一服务器的方案由于服务器端的处理能力和带宽会成为瓶颈,引起用户访问的时延增大。作为一种有效减小网络时延的方案,内容分发网络(Content Distribution Network, CDN)吸引了广泛应用,例如 Akamai[1]。CDN 是由一组地理上分散的边界服务器(edge server)构成。用户对某一个内容的访问,被重新定位到离用户最近的服务器。这样,传输的时延和网络资源的消耗就减小了。

当 CDN 中某一台服务器节点上有新内容发布或者有内容更新的时候,该源节点需要发起一个复制(replication)请求,把新内容分发到其他节点上,以保证 CDN 服务器上内容的一致性。很显然,如果源节点同时与每个其他节点都建立一个单播连接,方案的扩展性不好。IP 多播具有很好的扩展性,但是,目前 Internet 上路由器的 IP 多播功能没有得到广泛的支持。如何在有大量服务器的 CDN 中有效地复制内容成为一个具有挑战的问题。现有的方案大都采用应用层多播技术(Application Layer Multicast, ALM)在 CDN 中建立一个或者多个树结构来复制内容。FastReplica[2] 首先将需要复制的内容分成 N 个小块,然后分别把这些块发到 N 个不同的接受者节点,每个接受者再与其他的 $N-1$ 个接收者相互交换数据块。实际上,FastReplica 是建立了 N 个高度为两层的树结构来复制内容。SPIDER[3] 假设 Internet 核心网络中有专门的高带宽转发节点,CDN 利用这些额外的节点来构建多个树参与复制。但是,实际的 Internet 中,使用这些专门转发节点的代价是很高的。FastReplica 和 SPIDER 都假定每个 CDN 节点只能转发所收到的数据块。

本文提出了一种分布式的方案 NCOM(Network Coding in Overlay Mesh)来提高 CDN 中的复制速度。方案的创新性在于:首先,方案将 CDN 节点构建成 Mesh 结构,而不是树结构来复制;其次,充分利用节点的计算能力,使用 Network Coding 技术[6]来传输数据,每个节点不是简单地转发所收到的数据块,而是对收到的数据块先编码后再转发给其他节点。每个节点只要收到足够多的数据块,就能恢复原始内容,从而不需要协调从不同节点获取不同的数据块。

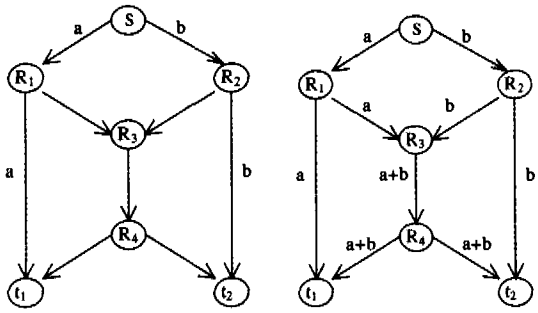
在对等网(P2P)中,也有内容复制的相关工作,比如 BitTorrent[5] 和 Byers 等人[4]提出的基于 Fountain 编码的方案。这些方案的每个节点都需要维护当前的收到的数据块的状态,以避免从不同的其它节点获得重复的数据块。本文提出的方案与这些 P2P 网络中的方案也存在显著的区别。首先是应用的场景不同,CDN 节点相对稳定而且具有较高的带宽,而 P2P 网络相对比较动态。其次我们的方案不需要协调从不同节点获得不同数据块,每个节点只需要收到足够多的块就可以恢复源内容。LION[8] 提出了将 Network Coding 应用在 P2P 网络中,但主要是研究利用分层技术来解决 P2P 网络中异构性问题。

*)自然科学基金项目资助(编号 60103013)。赵 进 博士研究生,研究方向包括应用层多播、Network Coding 技术;张福炎 教授,博士生导师,研究方向包括多媒体技术及其应用系统、多媒体网络、数字化图书馆。

本文第2节形式化描述CDN中的复制问题;第3节详述分布式的复制方案;第4节进行试验比较和分析;最后总结全文。

2 问题描述

将CDN网络用有向图$G(V,E)$表示,其中V是节点集,E是边集。边(i,j)表示节点i和j之间的链路,C_{ij}表示(i,j)的可用带宽。假设需要更新的内容位于节点S,复制过程需要以尽可能高的速率将内容发送到其他节点,以减小复制时间。需要指出的是,出于存储代价的考虑,CDN的策略可能只把内容复制到CDN网络中最合适的部分节点。CDN中复制的问题可以描述成:从节点S到节点通过网络G到集合T的最大速率问题,其中$T\subseteq V\backslash\{S\}$。此问题可以规约到NP-hard的Packing Steiner Tree问题[10],因此求最大速率也是NP-hard的。2000年Network Coding理论[6]的提出,使最大速率问题在多项式时间可解。与传统的路由相比,Network Coding允许网络的中间节点进行编码。图1演示了利用Network Coding的好处。网络中的链路都是单位带宽,节点S需要发送两个比特a和b到节点t_1和t_2,如果用传统的路由发送,需要两次才能发送完,而利用Network Coding,在节点R_3处加入编码功能,符号"+"表示异或,这样t_1和t_2就能一次同时恢复出a和b。



(a) 传统路由　　(b) network coding

图1 节点S发送a和b到t_1和t_2

利用Network Coding,S到整个集合T的最大速率$R(S,T)$等于S到T中任意节点t的独立最大速率$R(S,t)$中最小的值:

$$R(S,T)=\min_{t\in T}R(S,t) \tag{1}$$

也就是说如果每个节点t都能独立达到速率R,那么,利用Network Coding,整个组也能达到速率R。CDN中的复制问题也就转变成了在CDN图G中找数据传输子图G',使得在G'中每个t都有速率R,并且使得R最大。

下面分析每个节点t的独立最大速率。对t,假定链路(i,j)上承载的从S到t的数据流为

$$f_{ij}(t)\geqslant 0,\forall t\in T,\forall (i,j)\in E \tag{2}$$

由于每条链路有带宽限制,因此

$$f_{ij}(t)\leqslant C_{ij},\forall (i,j)\in E \tag{3}$$

传入节点i的数据流是所有以i为终点的链路上数据流的和,表示为

$$f_{\to i}(t)=\sum_{\{j|(j,i)\in E\}}f_{ji}(t) \tag{4}$$

同理,在节点i,传出的数据流为:

$$f_{i\to}(t)=\sum_{\{j|(i,j)\in E\}}f_{ij}(t) \tag{5}$$

因为t获得的速率为R,所以传入t的数据流应该为R,而t又是数据流的终点,所以传出的数据流应该为0。同理,因为S是数据流的源,所以传入S的数据流为0,而传出S的数据流为R。G中的其他节点,根据数据流守恒,传入的数据流和传出的数据流应该相等,所以

$$\begin{cases} f_{\to S}(t)=0 \\ f_{t\to}(t)=0 \\ f_{S\to}(t)=R,\forall t\in T,\forall i\in V-\{S\}-T \\ f_{\to t}(t)=R \\ f_{\to i}(t)-f_{i\to}(t)=0 \end{cases} \tag{6}$$

如果每个t都能独立获得速率R,那么利用Network Coding,整个T都能获得速率R。最大速率问题可以形式化为如下的优化问题:

$$\begin{aligned} &\max R \\ &\text{s. t.} \\ &(2)(3)(4)(5)(6) \end{aligned} \tag{7}$$

该问题是一个线性规划。

3 分布式的方案

尽管最大速率的优化问题是多项式时间可解的,但是求解该问题需要预先知道CDN的全局网络结构,用于集中式计算出最优的传输路径,因此在真实的网络中应用是不实际的。本文提出一种分布式的启发式方案NCOM(Network Coding in Overlay Mesh),用于快速复制。该方案将CDN中的节点组织成一个应用层的Overlay Mesh结构,并且利用Network Coding传输数据。现有的CDN内容复制方案,大都通过构建树结构来传输内容数据,由于树结构上的每一个节点只有一个上游节点,因此每个节点只能单路径获取数据,节点的可用带宽显然没有充分利用。与现有方案不同的是,NCOM通过建立Overlay Mesh来充分利用可用带宽,每个节点同时从多个上游节点获取数据。

NCOM建立Overlay Mesh的过程如下:每个节点t都有一个连接度d,最多有d个其他节点连接到t。当S有内容更新时,将通过CDN的管理策略发起一个复制请求,CDN管理策略会选择一部分节点从S获取更新的内容,以保证一致性。如果节点t需要参与复制过程,t首先向S发送消息,请求加入Mesh,S记录下t,并且返回N个已经在Mesh中的其他节点给t。通过测量与N个节点间的可用带宽,t从中选择K个节点作为在Mesh中的连接邻居,其中$K/2$个带宽最大的节点和$K/2$个随机的节点。如果选择的某个节点i的连接度d已经满了,t将请求i返回i在Mesh中邻居节点。这样t可以获得Mesh中的其他成员列表,从而递归地获得具有连接度的节点。选择随机的节点作为邻居可以保证Mesh具有较好的连接性,避免Mesh成为多个小的团(cluster)。而选择带宽好的节点可以保证节点在Mesh中获得较好的传输速率。通过这种方法,每个节点都有多个上游和下游节点。Mesh的建立是分布式的,每个节点只需要维护自己的邻居,具有控制开销小的优点。

当Mesh建立好后,NCOM利用随机线性Network Coding技术[7]进行内容数据的传输。由于CDN中的节点具有额外的处理能力和存储能力,因此Network Coding可以在应用层实现。每个节点将从上游节点收到的所有数据块进行编码后再传给下游节点。与Reed-Solomon编码类似,Network Coding对数据块的操作也是在Galois域上的。源节点S将需要更新的内容X分成h个相等的原始数据块$X=(x_1,x_2,$

…,x_h),每个数据块,都伴随一个 h 维编码向量(α_1,α_2,…,α_h)^T,表示该块与 h 个原始数据块进行线性组合编码的系数关系,例如 x_1 可以用编码向量$(1,0,\cdots,0)^T$ 表示。编码向量被包含在数据块的包头里面,随数据块一起传输,由于数据块的大小比编码向量大得多,因此编码向量的开销很小。当 S 需要发送一个数据块的时候,S 随机生成一个 h 维编码向量,对 h 个原始数据块在 Galois 域上做编码操作,生成编码过的数据块,连同编码向量发到下游节点。每个节点需要维护一个缓冲区,用来保存所收到的编码过的数据块,当节点 i 从上游节点收到一个数据块后,i 把收到的数据块保存到缓冲区。假定 i 的缓冲区有 K 个数据块 $Y=(y_1,y_2,\cdots,y_K)$,其中 y_j($j=1,\cdots,K$)的编码向量为 $A_j=(\alpha_{j1},\alpha_{j2},\cdots,\alpha_{jh})^T$,当 i 需要发送一个数据块到下游时,i 随机产生一个 K 维向量$(\beta_2,\beta_2,\cdots,\beta_K)$,将缓冲区中已有的 K 个数据块再进行编码,然后生成新的数据块,发送给下游节点,新数据块的编码向量为$(A_1,A_2,\cdots,A_K)\cdot(\beta_1,\beta_2,\cdots,\beta_K)$,仍然是 h 维的。节点 i 收到的 K 个数据块的编码向量所组成的矩阵 $A=(A_1,A_2,\cdots,A_K)$,显然有 $Y=X\cdot A$,当 A 的秩达到 h,就能通过消元法解出对应的方程组,从而恢复原始内容 X。需要说明的是,编码向量元素都是从一个足够大的 Galois 域上选取的,如 GF(2^8),编码操作都是在 Galois 域上进行。

用图 1(b)为例子,假设 S 把内容分成两个数据块 a 和 b,对应 $h=2$,然后选择向量$(m,n)^T$ 对 a 和 b 进行编码组合,生成新块发送到下游节点。图例中,S 发送到 R_1 的块对应的向量为$(1,0)^T$,而发送到 R_2 的为$(0,1)^T$。当 R_3 收到两个编码后的数据块后,再选择向量$(1,1)^T$ 进行编码,发送到下游。当 t_1 收到 2 个编码后的块后,对应的编码向量所组成的矩阵($(1,0)^T$,$(1,1)^T$)的秩为 2,因此 t_1 就可以恢复出原始的 a 和 b,同理 t_2。

4 实验仿真

为了验证 NCOM 的有效性,我们用 C++ 实现了模拟器,底层网络是用 GT-ITM[9] 的 Transit-Stub 模型生成。网络包含 2000 个节点,网络链路的时延随机设置为 20～100ms,链路的可用带宽随机设置为 10～100Mbps。我们比较了 NCOM 与 FastReplica 的复制时间。一个节点的复制时间定义为该节点从源节点开始复制,到完成复制所消耗的时间。最大复制时间也就是整个组从复制请求开始,到所有参与节点都完成复制所消耗的时间。为了比较的公平性,两种方案都采用同样的环境配置和参与节点,图中的曲线都是 50 次平均值。

首先的实验场景是比较 NCOM 和 FastReplica 的复制时间的累积分布函数(cumulative distribution function, CDF)。从网络中随机选取 200 个节点作为 CDN 的参与节点,复制一个 500MB 的文件到这些参与节点。图 2 显示了 NCOM 和 FastReplica 的复制时间 CDF。可以看出,大约在 300s,NCOM 的所有节点都完成了复制,比 FastReplica 要快。这是由于 NCOM 利用多路径,并且用 Network Coding 传输,提高了传输的速率,从而减小复制时间。

然后的实验场景是比较 CDN 的网络规模对最大复制时间的影响。我们随机选取 10 至 100 个节点作为 CDN 的参与节点,然后比较 NCOM 和 FastReplica 在各种规模下的最大复制时间。图 3 显示了实验结果,NCOM 在各种网络规模下都要比 FastReplica 所消耗的时间少。

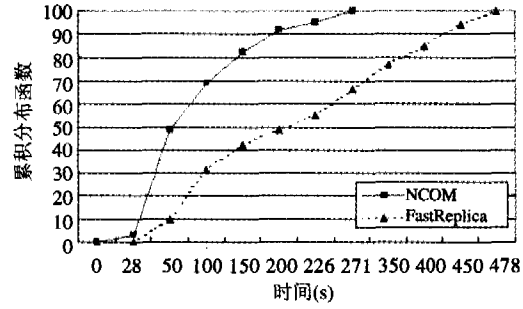


图 2 累积分布函数比较

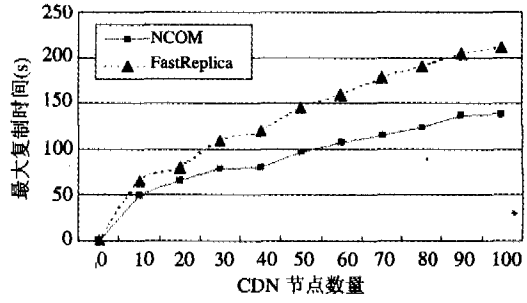


图 3 最大复制时间比较

实验结果表明,将 Network Coding 用于 CDN 的复制,可以带来好处,减小复制时间。

结论 实现快速复制是保证 CDN 网络中不同节点间内容一致性的关键技术之一。本文提出了一种利用 Network Coding 技术在 CDN 网络中提供快速复制的方案 NCOM。该方案是分布式的,具有低复杂性。实验结果表明,该方案能显著减小复制时间。

参 考 文 献

1 Akamai. http://www.akamai.com

2 Cherkasova L, Lee J. FastReplica: Efficient Large File Distribution Within Content Delivery Networks. In: Proc. of the 4th USENIX Symposium on Internet Technologies and Systems, Seattle WA, 2003. 1204～1216

3 Ganguly S, Saxena A, Bhatnagar S, et al. Fast Replication in content distribution overlays. In: Proc. of IEEE INFOCOM. Miami FL: IEEE press, 2005. 2246～2256

4 Byers J, Considine J, Mitzenmacher M, et al. Informed content delivery across adaptive overlay networks. IEEE/ACM Trans on Networking, 2004, 12(5): 767～780

5 Cohen B. BitTorrent. http://www.BitTorrent.org

6 Ahlswede R, Cai N, Li S, et al. Network information flow. IEEE Trans on Information Theory, 2000, 46(5):1204～1216

7 Chou P, Wu Y, Jain K. Practical network coding. In: Proc. of 41st Allerton Conf on Communication Control and Computing, Monticello, IL, 2003

8 Zhao J, Yang F, Zhang Q, et al. LION: layered overlay multicast with network coding. IEEE Trans on Multimedia (accepted to appear)

9 Zegura E, Calvert K, Bhattacharjee S. How to model an internetwork. In: Proc. of IEEE INFOCOM, San Francisco, CA: IEEE press, 1996. 594～602

10 Jain K, Mahdian M, Salavatipour M. Packing Steiner trees. In: Proc. ACM-SIAM SODA, Baltimore, MD, 2003. 266～274