cs 591B1: ALGORITHMIC ASPECTS OF COMPUTER NETWORKS **Spring 2002**

Lecture 12 — April 22

Lecturer: Jeffrey Considine and Stan Rost Boston University *Scribe: Thomas A Florio*

Following an improvised introduction by Jeffrey Considine, Stan Rost (MIT) and Jeffrey Considine presented material from their work with Prof. John Byers and Michael Mitzenmacher (Harvard University), to appear in SIGCOMM 2002. Ignoring issues of topology management, their work "Informed Content Delivery across Adaptive Overlay Networks" focuses on encoding content to take advantage of collaborative down loads between peers which do not have the whole file.

12.1 Evolution of Content Delivery

To detail the motivation for their work, the presenters gave four broad categories of past content delivery schemes.

Prehistoric Delivering content over point to point connections, the classic model of internet communications, has several serious flaws. First, the demands on the network in terms of bandwidth are unreasonable-many copies of the same information may traverse the same physical links in the network. Further, the "client-server" model inherently poses serious demands on the CPU of the server, which must maintain an active connection to each client. Lastly, transfer rates are limited by the end to end characteristics of the path from server to client.

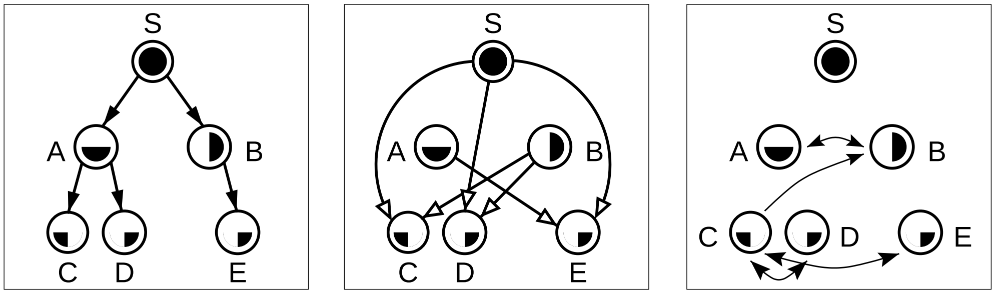
Mythological A near optimal solution to the content delivery problem is found in the form of IP-level multicast. In this model, network resources are utilized optimally: a tree is constructed over the IP network, and one copy of the data is dispatched from the server and replicated by routers in the network as it disseminates to the clients at the leaves of the tree. However, the presenters have reason to label this strategy "mythological". By requiring routers to replicate packets, it violates the internet paradigm of maintaining state and intelligence on the edges of the network. Further, this requires infrastructure changes that have thus far not been implemented.

Classic The classic content delivery scheme is end-system multicast. In this scheme, an overlay multicast tree is constructed of unicast links. As end-systems now represent nodes in the tree, specialized hardware is not required and infrastructure changes are not necessary. The presenters further argue that such a solution is adaptable to the topological changes inherent in the internet. As with any overlay network, there is no guarantee that the mapping of virtual, unicast tunnels to underlying physical links will be optimal.

Trendy The last category, and the one in which the subject of the presenters' work falls, is peer to peer networks. In a peer to peer solution to the content delivery problem, and indeed in the presenters' work, leverage is gained by virtue of the fact that the topology is more connected than that of a multicast tree. This higher degree of connectivity motivates the next discussion.

12.2 Motivation

Tree structured topologies provide good qualities to a content delivery scheme: the ancestor is responsible for delivering content to its children, handling retransmissions, etc. However, this structure also imposes limitations. For example, the bandwidth on the path from server to client appears monotonously decreasing, and the overall transmission rate is limited to that of the bottleneck bandwidth on the path from server to receiver. If the scheme is able to leverage a more richly connected topology than a tree, then "perpendicular bandwidth" between siblings or nodes of the same level in the tree structure exists. This observation is the central motivation for the presentation- perpendicular bandwidth implies the possibility of parallel down loads.



(a) Multicast Tree (b) Parallel Downloads (c) Collaborative Transfers

Figure 12.1. Content Delivery sub-schema. Shaded regions indicate portions of received content.

Parallel downloads usually refers to a situation in which client receives content from two servers. However, a peer with only partial content may still be a viable provider of usable data, as in figure 12.1. A group of peers who collectively contain the entire content can collaborate to individually acquire the whole, in a process of *collaborative transfers*, as in figure 12.1. Assuming peers have not received exactly the same content, such a process can be considered *reconciling* the sets of received content at each host. Thus, peers must intelligently collaborate to use bandwidth effectively.

Note: The presenters argue that such a solution applies to *content delivery networks* of sufficient bandwidth. For example, a CDN of modem users would be unable to leverage additional connections- most of their bandwidth would be utilized in a single connection.

12.3 Environmental Challenges

A solution to the problem of content delivery must contend with challenges stemming from the nature of the internet and those of overlay networks. In the internet, connections are asynchronous, and bandwidth, loss rates and disconnection rates are heterogeneous. Further, the network is transient- links become overloaded or are lost completely. Adaptive overlays in some cases exacerbate these problems: hosts may disconnect and changes to the overlay topology may pre-empt valid physical layer connections. The internet also provides an enormous number of hosts, any of which may become clients in the content delivery scheme.

Thus, scalable support is necessary for migration and fault tolerance. The Digital fountain approach [5] provides such support. For the purposes of the talk, under the digital fountain approach, files are divided into *input symbols* which are `XOR`'ed together to create output symbols. The purpose of such a procedure is that if the file constitutes n input symbols, then any n output symbols need be received to reconstruct the original file. Thus, the Digital fountain approach absorbs differences between clients, and provides reliable multicast data transfer. Further, such a procedure is stateless- symbols can be encoded on the fly. An additional property of the Digital fountain approach is additivity: senders generating different output symbols based on different sources of randomness are uncorrelated. This implies that parallel downloads from multiple servers requires no organization.

Additionally, this inherent environmental fluidity provides opportunity for hosts receiving identical content to reconcile sets. Two hosts receiving identical content may have different sets of received data if one host has been in the session longer. Two hosts subjected to independent loss rates will have "gaps" of missing data which do not overlap. Two hosts with differing transfer rates will have disparate working sets, and the one with the higher rate is a potential source of useful data for the other.

12.4 Reconciliation and Informed Delivery

Two main means of reconciling working set differences between two peers are described:

1. Speculative Transfers/Recoding

   In this method, a host makes an "educated guess" as to the degree of redundancy necessary for re-encoding and transferring useful information. The degree of redundancy is determined by an estimate of the overlap of working sets of two peers made by *Coarse-grained Reconciliation* which utilizes random sampling or *minwise permutations.*

2. Reconciled transfers

   In this method, a host filters symbols deemed redundant after a process of *Fine-grained Reconciliation*, which utilizes a searchable working set summary (like Bloom Filters [1] or *Approximate Reconciliation Trees*).

Fine-grained reconciliation techniques are more resource intensive than coarse-grained techniques, but provide more certainty about the common symbols in two peers' working sets.

12.4.1 Coarse-grained Reconciliation

Two methods of coarse-grained reconciliation are discussed:

Random Sampling k elements from the working set are selected randomly and are transported to the peer. Here there are several drawbacks. The receiver must search for each element it receives in its working set, requiring a search time of $O(\lg\lg n)$ using interpolation search. Further, it would not be possible to easily measure overlap between multiple peers.

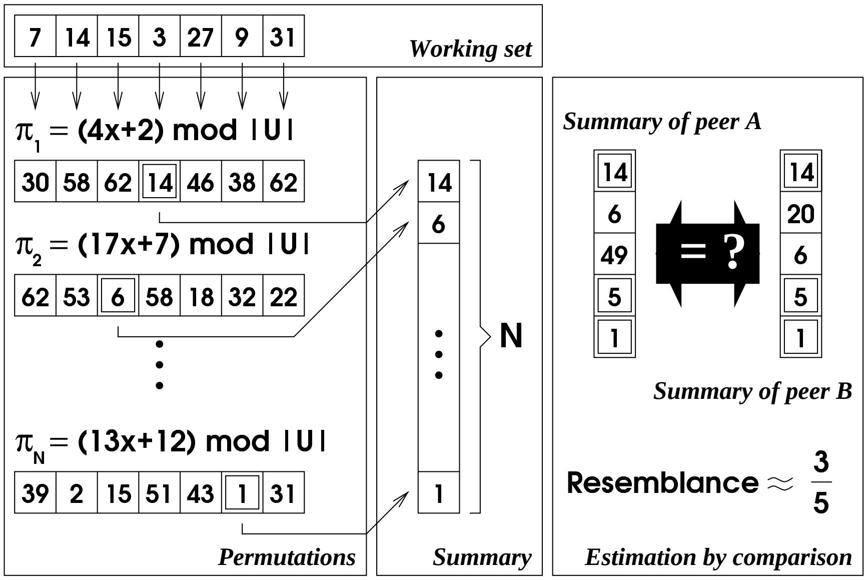


Figure 12.2. Example of Minwise summarization as used for coarse-grained reconciliation.

Minwise Summarization k simple permutations are applied to the working set of symbols on both hosts (treated as integers), and the minimal element of each resultant set is transfered. The estimated size of the set of common symbols is thus the number of identical minima from the two hosts. In theory, said permutations would ideally be random, but such a constraint is infeasible. Simple linear permutations of the form $a \cdot x + b mod |U|$ where U is the universe of symbols suffice. For an example, see figure 12.2.

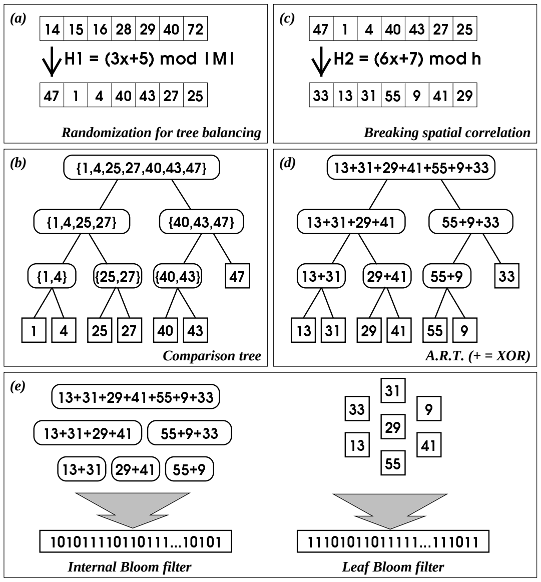


Figure 12.3. A visual example of the process of building and encoding an Approximate Reconciliation Tree.

12.4.2 Fine-grained Reconciliation

Two methods of fine-grained reconciliation are discussed:

A Bloom Filter Approach A peer encodes its working set into a Bloom filter which it then transmits to its peer. The receiving peer searches the Bloom filter for each elements in its working set, and sends those it does not find. The false positive probability of Bloom filters is mitigated by the fact that a false positive in this case leads the peer to not send a useful symbol- a useless symbol can never be sent using this approach. As the probability of a false positive is low, only a small number of useful symbols will not be sent. Further, since the symbols incorporate redundant information, there are other symbols which may be transferred in their place.

Approximate Reconciliation Trees A peer internally builds a tree structure with logarithmic depth in the following manner: First, it applies a hash function to each element of its working set to ensure an even distribution over the key space. Next, it builds a binary tree called a *Merkle Tree* as follows: The root represents the entire working set, hashed, in order. The left child contains the hashed elements of the working set in order that are in the first half of the key space of the parent node, and the right child has those of the second half. This recurses to the leaves, which are individual elements of the working set, hashed.

To make comparisons of nodes easy, each hashed element of a node has a second hash function applied to it, and the resultant values are combined using fast `XOR` operations.

Finally, the internal nodes of the resultant approximate reconciliation tree are encoded into a Bloom filter, and the leaves are encoded into a separate Bloom filter. These two filters are thus transfered to the receiving peer, where the receiver compares the nodes of its approximate reconciliation tree to the Bloom Filter. An failure to match an internal node with the Bloom filter indicates a set difference at a lower level in the tree.

For a visual example of this process, see figure 12.3

Separating the leaves and internal nodes into two separate Bloom filters allows the false positive rate of the internal nodes to be controlled. A false positive would terminate a search along a tree path, and a difference in the working sets may not be observed. The false positive rate can further be controlled by specifying a number of addition consecutive matches that must occur before the search down the tree path can be terminated.

The chief advantage of the tree approach as compared to a solution involving strictly Bloom filters is that it allows for faster searching of elements in the working set difference. Tree based searches run in time $O(\lg |W|)$ where W is the working set of the node running the search. The strict Bloom filter approach runs in time $O(|W|)$.

12.5 Conclusion

In conclusion, the presenters demonstrated how current distribution schemes which do not consider the peers of a node who have partial content as viable sources of new data have overlooked a potential performance enhancement. Further, changes to the overlay network are expected, and in the case of potential for set reconciliation, are actually of benefit. Thus, the method scales even over lossy traffic media, like wireless links.

12.6 Non-bibliographic Acknowledgments

Credit goes to Jeffrey Considine, for providing the figures used in this document, the original presentation slides, and an advance copy of [4].

Bibliography

[1] B. Bloom. "*Space/time trade-offs in hash coding with allowable errors*," Communications of the ACM, 13(7):422-426, 1970.

[2] M. Mitzenmacher. "*Compressed Bloom Filters*," in Proceedings of PODC 2001.

[3] J. Byers, et al. "*Informed Content Delivery Across Adaptive Overlay Networks*," To appear in SIGCOMM 2002.

[4] J.Byers, et al. "*Fast Approximate Reconciliation of Set Differences.*" Preliminary version, to be submitted to SODA 2003.

[5] J. Byers, et al. "*A Digital Fountain Approach to Asynchronous Reliable Multicast.*" To appear in IEEE Journal on Selected Areas in Communications, 2002. (Journal version of the ACM SIGCOMM '98 paper.)

[6] J. Considine, S. Rost "*Informed Content Delivery in Adaptive Overlay Networks.*" Lecture Slides, April 22, 2002.