RZ 3436 (# 93720) 07/15/02
Computer Science 5 pages

Research Report

Efficient Topology-Aware Overlay Network

Marcel Waldvogel* and Roberto Rinaldi

IBM Research
Zurich Research Laboratory
8803 Rüschlikon
Switzerland

*e-mail: mwl@zurich.ibm.com

Efficient Topology-Aware Overlay Network

Marcel Waldvogel and Roberto Rinaldi

Abstract

Peer-to-peer (P2P) networking has become a household word in the past few years, being marketed as a work-around for server scalability problems and "wonder drug" to achieve resilience. Current widely-used P2P networks rely on central directory servers or massive message flooding, clearly not scalable solutions. Distributed Hash Tables (DHT) are expected to eliminate flooding and central servers, but can require many long-haul message deliveries. We introduce Mithos, an overlay network that only uses minimal routing information and is directly suitable for normal and DHT addressing. Unlike other schemes, it also efficiently provides locality-aware connectivity, thereby ensuring that a message reaches its destination with minimal overhead and highly efficient forwarding. The service can in addition be used to support third-party triangulation to point to close replicas of data or services. Its addressing can be mapped directly into a subspace of the IPv6 addresses.

1 Introduction

The computing world is experiencing a transition from fixed servers and stationary desktop PCs to connected information appliances and ubiquitous connectivity, profoundly changing the way we use information. With cellular data communication, Bluetooth, and IEEE 802.11b (WiFi), the need for a global system that supports these new communication patterns becomes more pressing day by day. Two main patterns can be identified: First, Internet routing table size is surging, and second, direct serverless communication is gaining importance.

Routing Table Size. The ever increasing size of the Internet routing tables calls for new ways in network protocols. Although the introduction of Classless Inter-Domain Routing (CIDR) [FLYV93] enabled large-scale aggregation of routing information and thus provided a respite in the exponential growth of routing and forwarding tables for several years, the expansion has resumed in the first half of 2001 with full strength. Among the reasons given for the increased growth rates are the exhausting of preallocated address ranges, proliferation of always-on connected devices, and, probably most significantly, the tendency for businesses and even small Internet Service Providers (ISPs) to become multi-homed. This fact of being connected to multiple upstream providers breaks the hierarchy model behind CIDR, which is necessary for its aggregation to be efficient.

Serverless Communications. While services such as Napster brought publicity to the term peer-to-peer (P2P), serverless communication only started becoming popular when Napster's demise became a possibility. The events of September 11, 2001, have further shown that centralized servers and thus single points of failure should be avoided when system reliability and availability are business-critical. Serverless systems of the first generation heavily relied on flooding as the prime mechanism to query the distributed directory and to support connectivity when network components become unavailable. The second generation being designed now is based on distributed hash tables (DHTs) to allow direct addressing once the ID of the resource, such as document or service, is known.

Although many theoretical schemes for minimizing routing information have been proposed and many designs for DHTs have recently become prominent discussion topics, we are unaware of any practical and efficient system combining both. In this paper, we introduce Mithos, a novel mechanism that combines both, and provides additional benefits, such as its ability to use IPv6 as a native transport mechanism and its support for third-party triangulation.

Unlike other systems that map Internet topology to Cartesian coordinates [FJP+99, NZ02], Mithos, in full P2P spirit, uses *every node* in the entire network also as a topology landmark. This helps achieve accuracy and efficiency without the overhead of numerous dimensions or full-mesh probing of all landmarks. Instead, directed incremental probing is used to find a near-optimal placement, as will be explained below.

In Mithos, routing table size is minimized because every node only needs to know its direct neighbors; transitive routing enables messages to reach any destination nevertheless. To achieve this, Mithos employs a novel approach to routing in multi-dimensional irregular meshes, which is key to achieving minimum routing table size while guaranteeing connectivity.

The remainder of the paper is organized as follows. Section 2 introduces and describes the concepts behind Mithos. Section 3 presents early results from our simulation environment. Related work is discussed in Section 4, and conclusions are drawn in Section 5.

2 Mithos Design

The basic idea of Mithos is to embed the network into a multi-dimensional space, with every node being assigned a unique coordinate in this space. This is similar to interconnects used in many high-performance parallel computers, enabling optimal global routing with simple knowledge of the local coordinate gradients, i.e., which links lead to higher/lower coordinates in which dimensions. Unlike parallel computers, however, the mesh used for Mithos connectivity is not regular, in order to accommodate dynamic membership as well as to represent locality.

These goals are established for every new node in a three-phase process:

1. Finding close-by nodes and establishing a neighborhood
2. Assigning an ID to the newcomer based on this neighborhood
3. Establishing links with the neighborhood

The individual phases are discussed in more detail below.

2.1 Finding Neighbors

To ensure that neighbors in the overlay network are also close in the "underlay" network, a distance metric and a location process need to be defined. We chose network delay between two nodes as metric for measuring distances, but any metric establishing geometry-like foundations would be suitable, including any metrics typically used in routing protocols, independent of their Quality-of-Service (QoS) awareness. Examples include physical distance, monetary link cost, or the bandwidth a TCP-compliant stream would achieve.[1] Independent of the metric used, the value is referred to as *distance* below.

It is well known that connectivity and connection parameters are not necessarily symmetric or transitive in the Internet, especially when multiple autonomous systems (AS) are involved [S+99]. Nevertheless, these metrics provide a reasonable basis for an overlay network. When setting up a sufficiently dense overlay network whose goal is to minimize these connection parameters on a per-link basis, the overlay will adapt itself, trying to get optimal service from the underlay.

When searching for neighbors, the natural choice would be to perform an expanding ring search using a multicast mechanism [FJM+95]. Although the protocols were defined more than a decade [DC90], multicast is still only available as an experimental platform in the Internet, if at all. Therefore, the neighborhood location process has to revert to using unicast.

For bootstrapping, Mithos requires a candidate member to know how to contact (at least) one of the existing members. A nonempty subset of these members is used as the first set of candidate neighbors. Then, knowledge from within the overlay network is used to locate the actual neighborhood as follows. Each candidate neighbor is first asked for its direct neighbors, then these neighbors are probed for their distance according to the metric chosen for the overlay system. The best node is then used as the new candidate neighbor. This process is iterated until no further improvement can be achieved, effectively following the distance gradient (Figure 1).

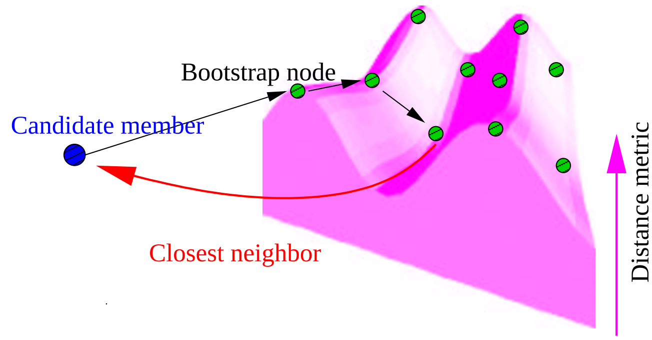

Figure 1: Finding neighbors

As this process is prone to terminate at a local instead of the global minimum, local minima must be recognized and avoided. For Mithos, this is currently done by probing all nodes that are two steps away from the current minimum before giving up. If a better candidate neighbor is found, the iterative process continues.

2.2 ID Assignment

Now that one of its neighbors has been selected, it is necessary to actually assign an ID to the candidate member. This ID selection process is critical, as an inappropriate assignment will eventually create many local minima, preventing an efficient neighborhood location in the future.

Mithos uses the distances measured during the last step of neighborhood establishment as a basis for ID assignment. The two closest nodes found in the process, their neighbors, and the corresponding distances are used in this computation, which requires no further communications.

For ID calculation, virtual springs are established between the candidate member and its fixed neighbors. The tension of each spring is set to be inversely proportional to the distance measured. Then this virtual equivalent of a physical system is allowed to settle, achieving the minimum energy state. This minimum energy location of the candidate node in the multidimensional space is directly used for its ID.

Now that an ID has been established, distances are *computed* in ID space, no longer requiring *measurements* (and thus message exchanges) according to the distance metric.

2.3 Linking Options

The final step is the establishment of peering relationships between neighbors. To evaluate the possible options for interconnecting neighbors, we established the following criteria:

1. Minimum routing table size;
2. efficient connectivity, full reachability; and
3. fast and simple forwarding algorithm.

[1] When setting up a system, care should be taken that the metric chosen is relatively stable for the duration of the P2P network.

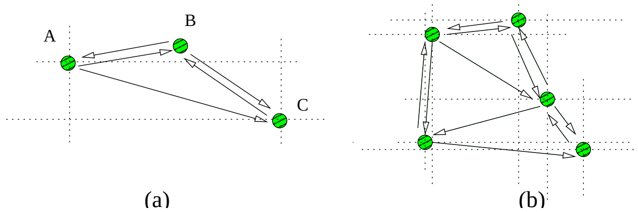

Figure 2: Example quadrant links in 2-space

These goals would be readily achieved by the strongly regular hypercube or hypertorus interconnect used in many parallel computers. In the presence of network dynamics, the regularity requirement would need to be significantly weakened. Our criterion of maintaining locality between neighbors completely breaks the dynamic supercomputer analogy. Furthermore, locality can lead to some local clustering effects, which need to be dealt with. Alternatives to rectangular connectivity in dynamic, locality-preserving environments are described and evaluated below.

Closest to axis. Along each axis in each direction, find a node that is closest to this axis and establish a link. Then, use the traditional hypertorus forwarding mechanism when delivering messages.

Quadrant-based. Each node establishes a link to the closest neighbor in each quadrant.[2] When forwarding, the next hop is chosen as the neighbor in the same quadrant as the final destination. This can done by computing the difference vector between the current node and the destination, and using the bit vector of the resulting d sign bits (one per dimension) as an index into the next-hop table.

Rectangular subdivision. Each node is assigned an enclosing axis-parallel multi-dimensional rectangle [RSS02]. Forwarding is done to the rectangle abutting at the point where the vector to the destination intersects with the current node's rectangle boundary.

Delaunay triangulation. Establish links according to a Delaunay triangulation of the nodes. Forward analogous to the previous whose vector is angularly closest to the destination vector.

All of these approaches *typically* achieve small routing tables, although in the worst case (for all but the *axis* mechanism) a single node could have all other nodes in the system as neighbors.

The connectivity is efficient, except when using *closest to axis*, which fails to locate off-axis nodes closer than the next on-axis node.

Forwarding lookups are optimal for the *quadrants* solution, as the final decision can be made by a simple indexed array access. Forwarding is still very good for the *axis* method, but as this method is unable to find all nodes without the aid of another algorithm, we consider it impractical, *Rectangles* and *Delaunay* base their decisions on angular calculations and comparisons, requiring expensive multiplications and multidimensional range searches.

[2] We use the term "quadrant" as a generic term, even when the number of dimensions, d, does not equal 2. All quadrants are determined relative to the current node.

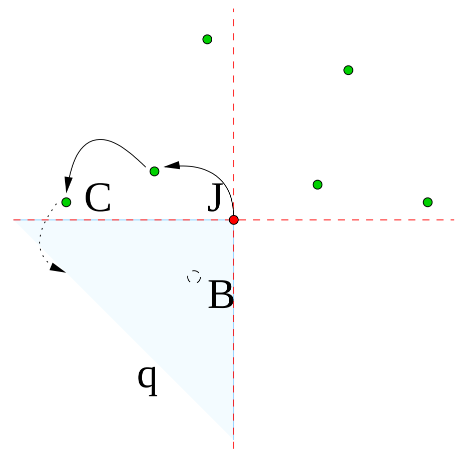

Figure 3: Finding neighbors in all quadrants

We therefore decided to use a *quadrant-based* mechanism, as it easily fulfilled all the criteria.

2.4 Establishing Quadrant Links

Before describing how to achieve quadrant-based links, we first evaluate some of their properties. Figure 2 shows two excerpts of networks situated in 2-space. Looking at Figure 2 (a), even though A has C as its closest southeast neighbor, C does not consider A as its closest northwest neighbor, resulting in asymmetric links. Fortunately, this asymmetry has no functional drawbacks during forwarding, as all nodes can still be reached efficiently. However, it needs to be taken into account when establishing the links. To simplify the description, the routing and link establishment process establishes bidirectional links, even though some of them will be used only unidirectionally when forwarding. Thus, the forwarding database remains minimum.

When the joining node J has established its ID, the sum of neighbors that helped it establish its ID may have no information about the best neighbor in all of J's quadrants. This can be because J's final position is out of range of the nodes' knowledge, or due to the asymmetry of the routing. Also, even though J might know of a node in each quadrant, this does not necessarily imply that this node is also the node closest to J. Therefore, J needs to identify the best neighbors in the region. The mechanism to achieve this is based on ideas similar to the perimeter walk used in Greedy Perimeter Stateless Routing (GPSR) [KK00], but has been extended to higher dimensions.

Now that a complete neighborhood has been established, it must be ensured that links are established to the closest neighbors, in order to guarantee correct forwarding operation. Thus the second phase tries to locate a closer neighbor by starting at the known neighbor and scanning towards all quadrant borders (Figure 4).

This second phase is an even further generalization of GPSR [KK00]. It currently uses parallel path processing,

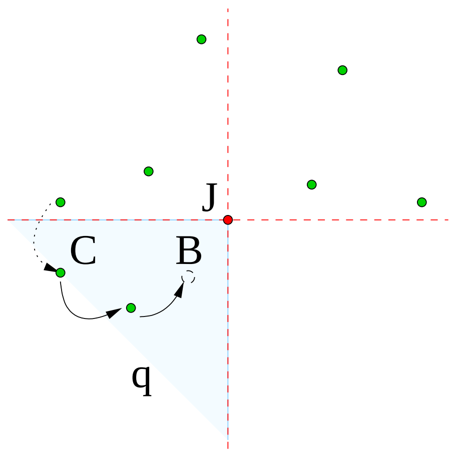

Figure 4: Finding the best neighbor in a quadrant

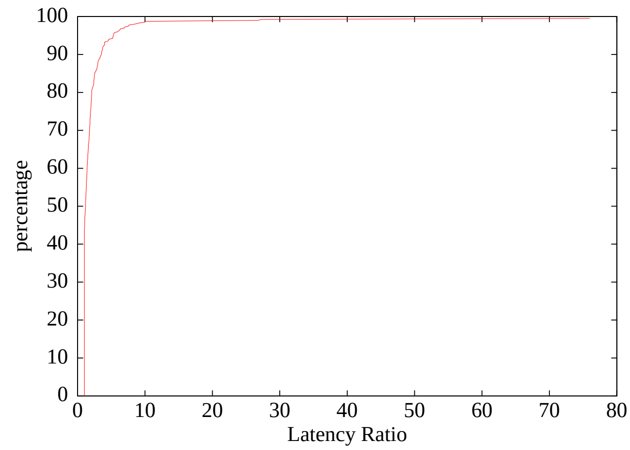

Figure 5: Latency ratio from the local/global minimum for each joining node (CDF)

which we expect can be optimized further by taking into account further geometric properties of the node relationships. Our early simulations have revealed that in the vast majority of cases, the best neighbors are already known from the merge step.

Serialization of multiple join events is only necessary if they involve the same neighborhood. As the steps requiring serialization all operate only on highly local areas with short distances, serializing them is not expected to become a bottleneck, although we are looking at ways to improve that.

2.5 Priming the Overlay Network

Starting the network from a single node using the mechanisms described above can lead to a very uneven utilization of the available space. To initialize the constants and provide enough initial points required for the spring forces algorithm, the network is primed with a small number of nodes appropriately distributed throughout the space the overlay network should span. These initial nodes are preferentially selected from early nodes interested in joining the system, but we envision that appropriate public landmarks could also be used to bootstrap the system.

3 Results

Preliminary results indicate that the above algorithms work very well. Figure 5 shows the quality of the minimum-finding algorithm. Despite its simple heuristics, the results are very encouraging. The test network consisted of 10,000 nodes in the underlay network (generated using the INET topology generator[3]) and 1000 nodes in the four-dimensional overlay network. About half of the nodes are optimally placed and more than 90% of the nodes are less than a factor of 5 in delay from their minimum. Further analysis reveals that this is often due to the small absolute delay.

Figure 6 compares the overhead of end-to-end path lengths under different numbers of dimensions (the same underlay network was used, but this time, only 200 nodes are placed in the overlay network for simulation efficiency). As can be seen, already at four dimensions, more than 97% of the paths are less than a factor of 3 from optimal. This is in contrast to non-P2P localization algorithms which require more dimensions and do not provide an efficient addressing scheme at the same time.

We expect better placement heuristics to further improve these results at potentially even further savings during node placement. More of our early results can be found in [Rin02].

4 Related Work

In the areas of Cartesian mapping of the Internet, the papers by Francis et al. [FJP+99] and, more recently, by Ng and Zhang [NZ02] use landmarks and measurements for triangulation. They use only a limited set of landmarks for location and do not provide for building a locality-aware overlay network.

On the other hand, numerous overlay networks have been proposed that employ a variety of mechanisms. Typical overlay networks for use in P2P systems have been designed to perform searches in a minimum number of steps, without taking the network latency of these steps into account. The most comprehensive work comparing different latency metrics was recently published by Ratnasamy et al. [RHKS02]. Owing to the limited number of landmarks used, their results are much less promising than ours.

5 Conclusions and Future Work

By having all nodes in the P2P network provide neighborhood location service through a directed, efficient search, we are able to create an overlay network whose connectivity is close to the optimum achievable with full topology knowledge. In contrast to other approaches, Mithos does not require full topology knowledge, even the forwarding and routing information is minimum and can be used in a highly efficient manner. At the same time, Mithos provides a powerful addressing scheme directly suitable for use in DHTs.

[3] Available from `http://topology.eecs.umich.edu/inet/`.

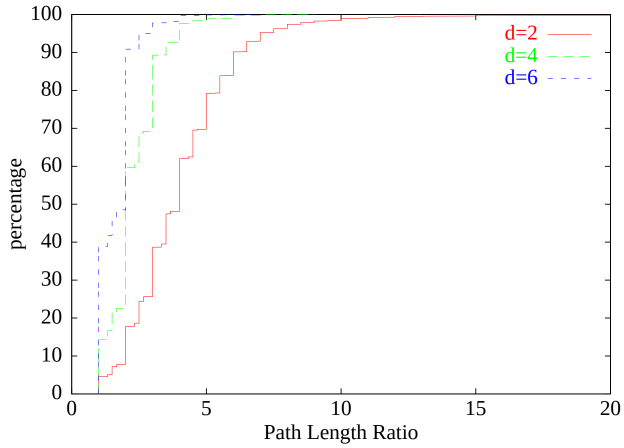

Figure 6: Path length ratios with 2, 4, and 6 dimensions (CDF)

We also believe that such addresses could be directly used in a dedicated P2P subspace of the IPv6 address space, e.g., by using six dimensions of 16 bits each.

In the future, we will investigate the dynamic behavior of the network and how to handle asymmetric underlay failures. We also plan to employ metrics obtained from real networks, including metrics other than pure delay. Further topics include optimizations of the "local minimum" and "spring forces" heuristics, as well as evaluating "asymmetric" dimensions, such as local and non-wrapping dimensions, which we expect to provide significant further gains.

References

[DC90] Stephen Deering and David R. Cheriton. Multicast routing in datagram internetworks and extended LANs. *ACM Transactions on Computer Systems*, 8(2):85–110, May 1990.

[FJM+95] Sally Floyd, Van Jacobson, Steve McCanne, Lixia Zhang, and Ching-Gung Liu. A reliable multicast framework for light-weight sessions and application level framing. In *Proceedings of ACM SIGCOMM*, pages 342–356, September 1995.

[FJP+99] Paul Francis, Sugih Jamin, Vern Paxson, Lixia Zhang, Daniel F. Gryniewicz, and Yixin Jin. An architecture for a global Internet host distance estimation service. In *Proceedings of IEEE INFOCOM*, pages 210–217, New York, NY, USA, March 1999.

[FLYV93] Vince Fuller, Tony Li, Jessica Yu, and Kannan Varadhan. Classless Inter-Domain Routing (CIDR): An address assignment and aggregation strategy. Internet RFC 1519, September 1993.

[KK00] Brad Karp and H. T. Kung. GPSR: Greedy perimeter stateless routing for wireless networks. In *Proceedings of MobiCom*, pages 243–254, August 2000.

[NZ02] T. S. Eugene Ng and Hui Zhang. Predicting Internet network distance with coordinates-based approaches. In *Proceedings of IEEE INFOCOM*, pages 170–179, New York, NY, USA, June 2002.

[RHKS02] Sylvia Ratnasamy, Mark Handley, Richard Karp, and Scott Shenker. Topologically-aware overlay construction and server selection. In *Proceedings of INFOCOM*, June 2002.

[Rin02] Roberto Rinaldi. Overlay network with efficient routing and topology. Diploma thesis, Institut Eurécom and Università degli Studi di Milano, June 2002.

[RSS02] Sylvia Ratnasamy, Scott Shenker, and Ion Stoica. Routing algorithms for DHTs: Some open questions. In *Proceedings of First International Workshop on Peer-to-Peer Systems (IPTPS)*, 2002.

[S+99] Stefan Savage et al. Detour: A case for informed Internet routing and transport. *IEEE Micro*, 19(1):50–59, January 1999.