High Performance, Robust, Secure and Transparent Overlay Network Service

Yair Amir, Claudiu Danilov, and Cristina Nita-Rotaru

Department of Computer Science
Johns Hopkins University
3400 North Charles St.
Baltimore, MD 21218 USA
{yairamir, claudiu, crisn}@cs.jhu.edu

1 Introduction

The technology underlying the Internet is a considerable engineering achievement. Over the last thirty years, it has proven its scalability, robustness and efficiency in providing global connectivity to a vast number of end nodes. A large part of the Internet technology success can be attributed to a few guiding rules. Among these are the reliance on one relatively simple network protocol (IP) and the end-to-end approach to reliability that is based on stateless intermediate nodes (TCP/IP).

The success of the Internet brings with it some problems: it is almost impossible to deploy new protocols or improve existing protocols in ways that break backward compatibility. As a result, much of the work in network protocol design is confined to improving some aspects of TCP/IP while maintaining backward compatibility and the ability to inter-operate with different versions of the protocol. Completely new methods, even if proven superior, have little chance to be adopted.

In addition, it is very difficult to add new services at the network level, as that will affect many different equipment and software providers. For example, adding the IP Multicast capability, first suggested around 1988, took almost a decade to be broadly supported (although rarely used).

The end-to-end approach to reliability and congestion control as implemented by TCP/IP has proven extremely scalable, as intermediate routers need not maintain any state for the vast amount of sessions routed through them. This comes with a price in efficiency. For example, when a message is lost on one of the intermediate links, it has to be recovered from the source rather than the last router that originally received it.

The overlay networks concept presents a promising way to address the above deficiencies. An overlay network instantiates a virtual network on top of a physical network using application-level routers (or daemons), where each overlay link is potentially composed of several physical links. While the application-level routers can execute any protocol between them, they are viewed by the physical network as an application. As a consequence, the overlay level is very flexible in the use of peer protocols. When a better routing or flow control protocol is invented, it can be immediately deployed in that level without affecting the lower level Internet. Deploying a new service (e.g. multicast) in the overlay level is much easier as it affects only the overlay routers and translates to regular unicast on the Internet.

Overlay networks suffer from two main drawbacks. First, they incur some overhead every time the message is processed by one of the overlay routers, which actually requires delivering the message to the application level, processing it, and sending a new message toward the next overlay router. Second, the placement of overlay routers in the topology of the physical network is often far from optimal, because the creator of the overlay network rarely has control over the physical network (usually the Internet) or even the knowledge about its actual topology.

In spite of its drawbacks, we view the overlay networks approach as key to facilitating new network services required by future applications as well as improving the performance of existing network services in confined, less general environments.

We propose a secure messaging infrastructure for unicast and multicast with near-optimal performance and stronger semantics, beyond what is naturally achieved over the Internet, and completely transparent to the application. This infrastructure is constructed by long-lived daemons, running as user-space programs. The daemons dynamically create and maintain a logical overlay network. The algorithms on this overlay network are not bounded by the standard Internet protocols. Rather, they exploit the more limited scalability of the overlay network (as opposed to the global Internet) to optimize performance (e.g. latency, throughput). Our infrastructure can still scale to much higher numbers than, for example, group communication systems, as we maintain much weaker global guarantees.

The target application domain is essentially any application that uses the Internet.

Relevant to our work are virtual private networks (VPN) [13] and the overlay network approaches in the USC/Xbone [12, 11] and the MIT/RON [9, 8]. VPN, Xbone and RON provide transparent service to the application. This property is critical for such a system to be practical. Our work is similar in this respect.

Virtual private networks usually focus on the security aspects, instantiating secure tunnels between different sites in order to extend the firewall domain over wide area networks. Our messaging infrastructure extends that by also focusing on the message performance and semantics such as low latency reliability, and near optimal routing and flow control. The USC/Xbone system instantiates a logical IP network on top of the current IP network (there can be multiple layers of overlays). This work is orthogonal to ours in that our infrastructure can be deployed in any of the Xbone logical layers. MIT/RON's goal is to provide resilience by utilizing redundant paths. Our goal includes resiliency, but also focuses on optimizing the performance of the reliable messaging latency and overall network throughput, while addressing security concerns.

Next, we discuss the main properties of our proposed infrastructure and some initial thoughts on how each of them can be achieved: high performance, scalability and robustness, security, and transparency.

2 Technical Approach

Our messaging infrastructure instantiates an overlay network that is dynamically constructed such that each overlay link connects two overlay nodes running our daemons over the underlying physical network (e.g. the Internet). We present a concise description of the main desired properties and some techniques to achieve them.

2.1 High Performance

The messaging infrastructure currently optimizes both the latency of a reliable service and the global flow control. Work on near-optimal dynamic multi-path routing is in progress.

We use buffers at intermediate overlay daemons to ensure hop-by-hop reliability [6]. This way, the end-to-end latency for lost packets is improved because of two reasons: we can detect the loss much faster on the hop compared with an end-to-end detection (which has to be at least the diameter of the network), and we can recover the loss much faster, as we only need to recover it from the last hop that received it, rather than from the source.

The above technique can achieve substantial latency improvement. For example, assume that a diameter of a US-wide network is 50 milliseconds and there are five hops connecting end nodes, each with 10 milliseconds latency. An end-to-end recovery of a lost message will take at least 50 milliseconds to detect, 50 more milliseconds to request the lost message from the source, and 50 milliseconds to recover it, for a total of 150 milliseconds. Moreover, messages that follow the lost packet will also be delayed, as they will not be delivered at the destination until the lost packet is recovered. In contrast, in our overlay approach, it will take only 20 additional milliseconds to recover one loss on one hop, in addition to the 50 milliseconds of dissemination, for a total of 70 milliseconds.

We use a cost-benefit framework to implement global flow control over the overlay network [5]. We make use of an exponential cost function for each overlay link so that the cost of the link increases exponentially as the capacity of the link is depleted. When the resource is not utilized at all its cost is zero and when it is fully utilized, its cost is prohibitively expensive. We assign benefit based on the parameter we want to optimize, which is the sending throughput or receiving throughput (these are not the same in the case of multicast). We have proved that this scheme is competitive with a logarithmic ratio compared with the optimal offline algorithm and have obtained good results both in simulations (using ns2 [2]) and in practice (over CAIRN [1] and Emulab [3]).

2.2 Scalability and Robustness

Strong semantics services (e.g. membership) come at the expense of overhead and scalability. Our approach is to base our messaging infrastructure on looser semantics, which is a better fit for the unicast model and for the weaker-semantics multicast. Our Spread Toolkit [7, 4] provides an efficient solution for stronger, group communication semantics.

Scalability with the number of nodes in the regular Internet is an important issue. However, the distributed infrastructure embedded in the Internet (IP multicast) is not scalable with the number of groups both in terms of its limited global IP address space and in terms of the state per group that has to be maintained in every intermediate router. Instead of maintaining state and sending control traffic per application connection, we keep track of the links in the overlay network. This approach scales well with the number of application sessions, groups and connections, with the expense of an overhead proportional to the size of the overlay network or even the number of immediate neighbors (depending on the desired semantics).

When an Internet route or an overlay daemon fails, our system reconfigures itself quickly, reestablishing routes if possible, allowing applications to use the networking services with little interruption since no global membership has to be agreed upon.

2.3 Security

We identify two types of security concerns. One is the authentication, confidentiality and integrity of the information passed between daemons (user data, routing information. etc). The other is the authentication and access control between client applications and the overlay network daemons.

We create on-demand secure tunnels over the dynamic overlay network to provide confidentiality. Since the channels are established between daemons, the cost of security is amortized for the various applications. The entity authentication between daemons can be addressed by using certificates. A certificate-based approach allows us to avoid the need for global knowledge regarding the potential participants of the overlay network. However, it requires the presence of a certificate authority in order to issue the certificates. One of the problems that we face, is designing a scalable solution to revoke certificates in a completely distributed environment.

Another important problem is protecting routing information. Current solutions use either expensive methods such as digital signatures or more efficient techniques such as hash chains or HMAC [10] that lack source non-repudiation. We intend to use a mixed scheme that exploits the advantages of both, as appropriate.

2.4 Transparency

We consider two approaches to provide seamless integration of our infrastructure with existing applications. The first approach intercepts the application's messaging in the socket evel, directs messages to our messaging infrastructure, and finally delivers them to the application on the receiver(s) side. The second approach instantiates a logical Internet router that gets all of the packets originating from the machine, routes them regularly if they are not addressed to overlay network participants, and handles them if they are.

We will enable new applications that want to use more powerful semantics to specify their needs using the signaling support provided by the operating system, such as setsockopt or out-of-band messages.

3 Conclusions

Overlay networks provide a promising practical method to address the lack of flexibility in the current Internet protocols both in terms of introducing new peer protocols and new services. We believe that careful engineering can make the associated overhead of overlay networks insignificant compared with the benefit gained by its use.

We are developing a prototype system that encapsulates the overlay network paradigm and the above discussed techniques. This prototype is designed to allow other researchers to experiment with their own protocols and services in an overlay network environment.

References

1. CAIRN network. http://www.cairn.net.
2. The Network Simulator - ns-2. http://www.isi.edu/nsnam/ns/.
3. The Utah network emulation facility. http://www.emulab.net.
4. Amir, Y., Ateniese, G., Hasse, D., Kim, Y., Nita-Rotaru, C., Schlossnagle, T., Schultz, J., Stanton, J., and Tsudik, G. Secure group communication in asynchronous networks with failures: integration and experiments. In *Proceedings of the 20th IEEE International Conference on Distributed Computing Systems* (Taipei, Taiwan, April 2000), pp. 330–343.
5. Amir, Y., Awerbuch, B., Danilov, C., and Stanton, J. Flow control for many-to-many multicast: A cost-benefit approach. In *IEEE Open Architectures and Network Programming (OpenArch)* (New York, New York, USA, June 2002).
6. Amir, Y., Danilov, C., and Stanton, J. A low latency, loss tolerant architecture and protocol for wide area group communication. In *Proceedings of the International Conference on Dependable Systems and Networks* (June 2000), pp. 327–336.
7. Amir, Y., and Stanton, J. The Spread wide area group communication system. Tech. Rep. 98-4, Johns Hopkins University, Center of Networking and Distributed Systems, 1998.
8. Andersen, D. G. Resilient overlay networks. Master's thesis, Massachusetts Institute of Technology, May 2001.
9. Andersen, D. G., Balakrishnan, H., Kaashoek, M. F., and Morris, R. Resilient overlay networks. In *18th ACM SOSP* (October 2001).
10. Menezes, A., van Oorschot, P., and Vanstone, S. *Handbook of Applied Cryptography*. CRC Press, 1996.
11. Touch, J. Dynamic internet overlay deployment and management using the X-bone. *Computer Networks* (July 2001), 117–135.
12. Touch, J., and Hotz, S. The X-bone. In *the 3rd Global Internet Mini-Conference at Globecom* (November 1998), pp. 59–68.
13. Yuan, R., and Strayer, W. T. *Virtual Private Networks: Technologies and Solutions*. Addison-Wesley, 2001.