IAB Workshop on Stack Evolution in a Middlebox Internet (SEMI) Report
draft-iab-semi-report-00

Abstract

   The Internet Architecture Board (IAB) through its IP Stack Evolution
   program, the Internet Society, and the Swiss Federal Institute of
   Technology (ETH) Zurich hosted the Stack Evolution in a Middlebox
   Internet (SEMI) workshop in Zurich on 26-27 January 2015 to explore
   the ability to evolve the transport layer in the presence of
   middlebox- and interface-related ossification of the stack. The goal
   of the workshop was to produce architectural and engineering guidance
   on future work to break the logjam, focusing on incrementally
   deployable approaches with clear incentives to deployment both on the
   endpoints (in new transport layers and applications) as well as on
   middleboxes (run by network operators). This document summarizes the
   contributions to the workshop, provides an overview of the discussion
   at the workshop, as well as the outcomes and next steps identified by
   the workshop. The views and positions documented in this report are
   those of the workshop participants and do not necessarily reflect IAB
   views and positions.

Status of This Memo

1. Introduction

   The transport layer of the Internet has becomed ossified, squeezed
   between narrow interfaces (from BSD sockets to pseudo-transport over
   HTTPS) and increasing in-network modification of traffic by
   middleboxes that make assumptions about the protocols running through
   them. This ossification makes it difficult to innovate in the
   transport layer, through the deployment of new protocols or the
   extension of existing ones. At the same time, emerging applications
   require functionality that existing protocols can provide only
   inefficiently, if at all.

   To begin to address this problem, the IAB, within the scope of its IP
   Stack Evolution Program, organized a workshop to discuss approaches
   to de-ossifying transport, especially with respect to interactions
   with middleboxes and new methods for implementing transport
   protocols. Recognizing that the end-to-end principle has long been
   compromised, we start with the fundamental question of matching paths
   through the Internet with certain characteristics to application and
   transport requirements.

   We posed the following questions in the call for papers: Which paths
   through the Internet are actually available to applications? Which
   transports can be used over these paths? How can applications
   cooperate with network elements to improve path establishment and
   discovery? Can common transport functionality and standardization
   help application developers to implement and deploy such approaches
   in today's Internet? Could cooperative approaches give us a way to
   rebalance the Internet back toward its end-to-end roots?

   The call for papers encouraged a focus on approaches that are
   incrementally deployable within the present Internet. Identified
   topics included the following:

   o Development and deployment of transport-like features in
      application-layer protocols

   o Methods for discovery of path characteristics and protocol
      availability along a path

   o Methods for middlebox detection and characterization of middlebox
      behavior and functionality

   o Methods for NAT and middlebox traversal in the establishment of
      end-to-end paths

   o Mechanisms for cooperative path-endpoint signaling, and lessons
      learned from existing approaches

   o Economic considerations and incentives for cooperation in
      middlebox deployment

   The SEMI workshop followed in part from the IAB's longer term
   interest in the evolution of the Internet and the adoption of
   Internet protocols, including the Internet Technology Adoption and
   Transition workshop [RFC7305], "What Makes for a Successful Protocol"
   [RFC5218], back to Deering's plenary talk [deering-plenary] at IETF
   51 in 2001 and before.

1.1. Organization of this report

   This workshop report summarizes the contributions to, and discussions
   at the workshop, organized by topic. We started with a summary of
   the current situation with respect to stack ossification, and
   explored the incentives which have made it that way and the role of
   incentives in evolution. Many contributions were broadly split into
   two areas: middlebox measurement, classification, and approaches to
   defense against middlebox modification of packets; and approaches to
   support transport evolution. All accepted position papers and
   detailed transcripts of discussion are available at
   https://www.iab.org/activities/workshops/semi/.

   The outcomes of the workshop are discussed in Section 6, including
   progress after the workshop toward each of the identified work items
   as of the time of publication of this report.

2. The Situation in Review

   At the time of Deering's talk in 2001, network address translation
   (NAT) was identified as the key challenge to the Internet
   architecture. Since then, the NAT traversal problem has been largely

   solved, but the boxes in the middle are getting smarter and more
   varied.

   SEMI, as the IP Stack Evolution program in general, is far from the
   first attempt to solve the problems caused by middlebox interference
   in the end to end model. Just within the IETF the MIDCOM, NSIS, and
   BEHAVE efforts have addressed this problem, and the TRAM working
   group is updating the NAT traversal outcomes of MIDCOM to reflect
   current reality.

   We believe we have an opportunity to improve the situation in the
   present, however, due to a convergence of forces. While the tussle
   between security and middleboxes is not new, the accelerating
   deployment of cryptography for integrity and confidentiality makes
   many packet inspection and packet modification operations obsolete,
   creating pressure to improve the situation. There is also new energy
   in the IETF around work which requires transport layer flexibility
   we're not sure we have (e.g. WebRTC) as well as around flexibility
   at the transport interface (TAPS).

3. Incentives for Stack Ossification and Evolution

   The current situation is, of course, the result of a variety of
   processes, and the convergence of incentives for network operators,
   content providers, network equipment vendors, application developers,
   operating system developers, and end users. Moore's Law makes it
   easier to deploy more processing on-path, network operators need to
   find ways to add value, enterprises find it more scaleable to deploy
   functionality in-network than on endpoints, and middleboxes are
   something vendors can vend. These trends increases ossification of
   the network stack.

   Any effort to reduce the resulting ossification in order to make it
   easier to evolve the transport stack, then, must consider the
   incentives to deployment of new approaches by each of these actors.

   As Christian Huitema [huitema-semi] pointed out, encryption provides
   a powerful incentive here: putting a transport protocol atop a
   cryptographic protocol atop UDP resets the transport versus middlebox
   tussle by making inspection and modification above the encryption and
   demux layer impossible. Any transport evolution strategy using this
   approach must also deliver better performance or functionality (e.g.
   setup latency) than existing approaches while being as deployable as
   these approaches, or moreso.

   Indeed, significant positive net value at each organization where
   change is required - operators, application developers, equipment
   vendors, enterprise and private users - is best to drive deployment

of a new protocol, said Dave Thaler, pointing to [RFC5218]. All
tussles in networking stem from conflicting incentives unavoidable in
a free market. For upper layer protocols, incentives tend to favor
protocols that work anywhere, use the most efficient mechanism that
works, and are as simple as possible from an implementation,
maintenance, and management standpoint. For lower layer protocols,
incentives tend toward ignoring and or disabling optional features,
as there is a positive feedback cycle between being rarely used and
rarely implemented.

4. The Role and Rule of Middleboxes

Middleboxes are commonplace in the Internet and constrain the ability
to deploy new protocols and protocol extensions. Engineering around
this problem requires a "bestiary" of middleboxes, a classification
of which kinds of impairments middleboxes cause and how often,
according to Benoit Donnet [edeline-semi].

Even though the trend towards Network Function Visualization (NFV)
allows for faster update-cycle of middleboxes and thereby more
flexibility, the function provided by middleboxes will stay. In
fact, service chaining may lead to more and more add-ons to address
and manage problems in the network, in turn further increasing the
complexity of network management. Ted Hardie [hardie-semi] warned
that each instance may add a new queue and may increase the
bufferbloat problem which is contra-productive for new emerging
latency-sensitive applications. However, this new flexibility also
provides a chance to move functionality back to the end host.
Alternately, more appropriate in-network functionality could benefit
from additional information in application and path characteristics,
though this in turn implies a variety of complicated trust
relationships among nodes in the network. In any case, an increasing
trend of in-network functionality can be observed, especially in
mobile networks.

Costin Raiciu [raiciu-semi] stated that middleboxes make the Internet
unpredictable, leading to a trade-off between efficiency and
reachability. While constructive cooperation with middleboxes to
establish a clear contract between the network and the end might be
one approach to address this challenge, enforcement of contract in
less cooperative environments might require extensive tunneling.
Raiciu's contribution on "ninja tunneling" illustrates one such
approach.

5. Evolving the Transport Layer

For evolution in the transport layer itself various proposals have been discussed, reaching from the development of new protocols (potentially as user-level stacks) encapsulated in UDP as a transport identification sub-header to the use of TCP as a substrate where the semantics of TCP are relaxed (e.g. regarding reliability, ordering, flow control etc.) and a more flexible API is provided to the application.

Discussion on evolution during the workshop divided amicably along two lines: working to fix the deployability of TCP extensions ("the TCP Liberation Front") versus working to build new encapulation-based mechanisms to allow wholly new protocols to be deployed ("the People's Front of UDP"). David Black [black-semi] pointed out that UDP encapsulation has to be adapted and separately discussed for every use case, which can be a long and painful process. UDP encapsulation can be an approach to develop more specialized protocols that helps to address special needs of certain applications. However, Stuart Cheshire [cheshire-semi] (as presented by Brian Trammell) pointed out that designing a new protocol instead of fixing/extending TCP might not always solve the problem.

To address the extensibility problem of TCP, Bob Briscoe proposed Inner Space [briscoe-semi]. Here, the general principle is to extend layer X's header within layer X+1; in the case of TCP, additional TCP header and option space is provided within the TCP payload, such that it cannot presently be inspected and modified by middleboxes.

Further instead of only focusing on those cases there new extensions and protocols are not deployable, Micheal Welzl [welzl-semi] points out that there are also a lot of paths in the network that are not ossified. To enable deployment on these paths an end host would need to probe or use a happy-eyeball-like approach and potentially fallback. The TAPS working group implements the first step to decouples applications from transport protocols allowing for the needed flexibility in the transport layer.

6. Outcomes

The SEMI workshop identified several areas for further work, outlined below:

6.1. Minimal signaling for encapsulated transports

Assuming that a way forward for transport evolution in user space would involve encapsulation in UDP datagrams, the workshop identified that it may be useful to have a facility built atop UDP to provide

minimal signaling of the semantics of a flow that would otherwise be
available in TCP: at the very least, indications of first and last
packets in a flow to assist firewalls and NATs in policy decision and
state maintenance. This facility could also provide minimal
application-to-path and path-to-application signaling, though there
was less agreement exactly what should or could be signaled here.

The workshop did note that, given the increasing deployment of
encryption in the Internet, this facility should cooperate with DTLS
[RFC6347] in order to selectively expose information about traffic
flows where the transport headers and payload themselves are
encrypted.

To develop this concept further, it was decided to propose a non
working group forming BoF session, SPUD (Substrate Protocol for User
Datagrams), at the IETF 92 meeting in March in Dallas. A draft on
use cases [I-D.hardie-spud-use-cases], a prototype specification for
a shim protocol over UDP {{I-D.hildebrand-spud-prototype}, and a
separate specification of the use of DTLS as a subtransport layer
[I-D.huitema-tls-dtls-as-subtransport] were prepared following
discussions at SEMI, and presented at the BoF.

Clear from discussion before and during the SPUD BoF, and drawing on
experience with previous endpoint-to-middle and middle-to-endpoint
signaling approaches, is that any selective exposure of traffic
metadata outside a relatively restricted trust domain must be
declarative as opposed to imperative, non-negotiated, and advisory.
Each exposed parameter should also be independently verifiable, so
that each entity can assign its own trust to other entities. Basic
transport over the substrate must continue working even if signaling
is ignored or stripped, to support incremental deployment. These
restrictions on vocabulary are discussed further in
[I-D.trammell-stackevo-newtea].

There was much interest in the room in continuing work on an approach
like the one under discussion. It was relatively clear that the
state of the discussion and prototyping activity now is not yet
mature enough for standardization within an IETF working group. An
appropriate venue for continuing the work remains unclear.

Discussion contiues on the spud mailing list (spud@ietf.org). The
UDP shim layer prototype described by
[I-D.hildebrand-spud-prototype].

6.2. Middlebox measurement

   Discussion about the impairments caused by middleboxes quickly
   identified the need to get more and better data about how prevalent
   certain types of impairments are in the network. It doesn't make
   much sense, for instance, to engineer complex workarounds for certain
   types of impairments into transport protocols if those impairments
   are relatively rare. There are dedicated measurement studies for
   certain types of impairment, but the workshop noted that prevalence
   data might be available from error logs from TCP stacks and
   applications on both clients and servers: these entities are in a
   position to know when attempts to use particular transport features
   failed, providing an opportunity to measure the network as a side
   effect of using it. Many clients already have a feature for sending
   these bug reports back to their developers. These present
   opportunities to bring data to bear on discussion and decisions about
   protocol engineering in an Internet full of middleboxes.

   The HOPS (How Ossified is the Protocol Stack) informal birds of a
   feather session ("BarBoF") was held at the IETF 92 meeting in Dallas,
   to discuss approaches to get aggregated data from these logs about
   potential middlebox impairment, focusing on common data formats and
   issues of preserving end-user privacy. While some discussion focused
   on aggregating impairment observations at the network level, initial
   work will focus on making relative prevalence information available
   on an Internet-wide scope. The first activity identified has been to
   match the types of data required to answer questions relevant to
   protocol engineering to the data that currently is or can easily be
   collected.

   A mailing list (hops@ietf.org) has been established to continue
   discussion.

6.3. Guidelines for middlebox design and deployment

   The workshop identified the potential to update [RFC3234] to provide
   guidelines on middlebox design, implementation, and deployment in
   order to reduce inadvertent or accidental impact on stack
   ossification in existing and new middlebox designs. This document
   will be produced by the IAB IP Stack Evolution program, drawing in
   part on the work of the BEHAVE working group, and on experience with
   STUN, TURN, and ICE, all of which focus more specifically on network
   address translation.

6.4. Architectural guidelines for transport stack evolution

The workshop identified the need for architectural guidance in
general for transport stack evolution: tradeoffs between user- and
kernel-space implementations, tradeoffs in and considerations for
encapsulations (especially UDP), tradeoffs in implicit versus
explicit interaction with devices along the path, and so on. This
document will be produced by the IAB IP Stack Evolution Program; the
new transport encapsulations draft [I-D.trammell-stackevo-newtea] may
evolve into the basis for this work.

Further due to the underlying discuss on trust and a needed "balance
of power" between the end hosts and the network, the workshop
participants concluded that it is neccessary to define cryptographic
protocol based approaches to enable transport protocol extensibility.

6.5. Additional Activities in the IETF and IAB

The workshop identified the need to socialize ideas connected to
transport stack evolution within the IETF community, including
presentations in the transport and applications open area meetings on
protocol extensibility, UDP encapsulation considerations, and the
application of TLS/DTLS in order to prevent middlebox meddling. Much
of the energy coming out of the workshop went into the SPUD BoF (see
Section 6.1), so these presentations will be given at future
meetings.

There are also clear interactions between the future work following
the SEMI workshop and the IAB's Privacy and Security Program; Privacy
and Security program members will be encouraged to follow
developments in transport stack evolution to help especially with
privacy implications of the outcomes of the workshop.

6.6. Additional Activities in Other Venues

Bob Briscoe did an informal liaison of the SEMI workshop discussions
to the ETSI Network Function Virtualization (NFV) Industry
Specification Group (ISG) following the workshop, focusing as well on
the implications of end to end encryption on the present and future
of in-network functionality. In the ISG's Security Working Group, he
proposed text for best practices on middlebox access to data in the
presence of end to end encryption.

7. Security Considerations

This document presents no security considerations.

8. Acknowledgments

   The IAB thanks the SEMI Program Committee: Brian Trammell, Mirja
   Kuehlewind, Joe Hildebrand, Eliot Lear, Mat Ford, Gorry Fairhurst,
   and Martin Stiemerling. We additionally thank Prof. Dr. Bernhard
   Plattner of the Communication Systems Group at ETH for hosting the
   workshop, and the Internet Society for its support. Thanks to
   Suzanne Woolf for the feedback.

9. Attendees

   The following people attended the SEMI workshop:

   Mary Barnes, Richard Barnes, David Black, Marc Blanchet, Bob Briscoe,
   Ken Calvert, Spencer Dawkins, Benoit Donnet, Lars Eggert, Gorry
   Fairhurst, Aaron Falk, Mat Ford, Ted Hardie, Joe Hildebrand, Russ
   Housley, Felipe Huici, Christian Huitema, Jana Iyengar, Mirja
   Kuehlewind, Eliot Lear, Barry Leiba, Xing Li, Szilveszter Nadas, Erik
   Nordmark, Colin Perkins, Bernhard Plattner, Miroslav Ponec, Costin
   Raiciu, Philipp Schmidt, Martin Stiemerling, Dave Thaler, Brian
   Trammell, Michael Welzl, Brandon Williams, Dan Wing, and Aaron Yi
   Ding.

   Additionally, Stuart Cheshire and Eric Rescorla contributed to the
   workshop but were unable to attend.

10. Informative References

   [RFC3234] Carpenter, B. and S. Brim, "Middleboxes: Taxonomy and
              Issues", RFC 3234, February 2002.

   [RFC5218] Thaler, D. and B. Aboba, "What Makes For a Successful
              Protocol?", RFC 5218, July 2008.

   [RFC6347] Rescorla, E. and N. Modadugu, "Datagram Transport Layer
              Security Version 1.2", RFC 6347, January 2012.

   [RFC7305] Lear, E., "Report from the IAB Workshop on Internet
              Technology Adoption and Transition (ITAT)", RFC 7305, July
              2014.

   [I-D.hardie-spud-use-cases]
              Hardie, T., "Use Cases for SPUD", draft-hardie-spud-use-
              cases-01 (work in progress), February 2015.

   [I-D.hildebrand-spud-prototype]
              Hildebrand, J. and B. Trammell, "Substrate Protocol for
              User Datagrams (SPUD) Prototype", draft-hildebrand-spud-
              prototype-03 (work in progress), March 2015.

   [I-D.huitema-tls-dtls-as-subtransport]
              Huitema, C., Rescorla, E., and J. Jana, "DTLS as
              Subtransport protocol", draft-huitema-tls-dtls-as-
              subtransport-00 (work in progress), March 2015.

   [I-D.trammell-stackevo-newtea]
              Trammell, B., "Thoughts a New Transport Encapsulation
              Architecture", draft-trammell-stackevo-newtea-01 (work in
              progress), May 2015.

   [black-semi]
              Black, D., "UDP Encapsulation: Framework Considerations
              (https://www.iab.org/wp-content/IAB-uploads/2014/12/
              semi2015_black.pdf)", January 2015.

   [briscoe-semi]
              Briscoe, B., "Tunneling Through Inner Space
              (https://www.iab.org/wp-content/IAB-uploads/2014/12/
              semi2015_briscoe.pdf)", January 2015.

   [cheshire-semi]
              Cheshire, S., "Restoring the Reputation of the Much-
              Maligned TCP (https://www.iab.org/wp-content/IAB-
              uploads/2015/01/semi2015-cheshire.pdf)", January 2015.

   [deering-plenary]
              Deering, S., "Watching the Waist of the Protocol Hourglass
              (https://www.ietf.org/proceedings/51/slides/plenary-1)",
              August 2001.

   [edeline-semi]
              Edeline, K. and B. Donnet, "On a Middlebox Classification
              (https://www.iab.org/wp-content/IAB-uploads/2014/12/
              semi2015_edeline.pdf)", January 2015.

   [hardie-semi]
              Hardie, T., "Network Function Virtualization and Path
              Character (https://www.iab.org/wp-content/IAB-
              uploads/2014/12/semi2015_hardie.pdf)", January 2015.

   [huitema-semi]
              Huitema, C., "The Secure Transport Tussle
              ([https://www.iab.org/wp-content/IAB-uploads/2014/12/](https://www.iab.org/wp-content/IAB-uploads/2014/12/semi2015_huitema.pdf)
              [semi2015_huitema.pdf](https://www.iab.org/wp-content/IAB-uploads/2014/12/semi2015_huitema.pdf))", January 2015.

   [raiciu-semi]
              Raiciu, C., Olteanu, V., and , "Good Cop, Bad Cop: Forcing
              Middleboxes to Cooperate ([https://www.iab.org/wp-content/](https://www.iab.org/wp-content/IAB-uploads/2015/01/ninja.pdf)
              [IAB-uploads/2015/01/ninja.pdf](https://www.iab.org/wp-content/IAB-uploads/2015/01/ninja.pdf))", January 2015.

   [welzl-semi]
              Welzl, M., Fairhurst, G., and D. Ros, "Ossification: a
              result of not even trying? ([https://www.iab.org/wp-](https://www.iab.org/wp-content/IAB-uploads/2014/12/semi2015_welzl.pdf)
              [content/IAB-uploads/2014/12/semi2015_welzl.pdf](https://www.iab.org/wp-content/IAB-uploads/2014/12/semi2015_welzl.pdf))", January
              2015.

Authors' Addresses

   Brian Trammell (editor)
   ETH Zurich
   Gloriastrasse 35
   8092 Zurich
   Switzerland

   Email: ietf@trammell.ch

   Mirja Kuehlewind (editor)
   ETH Zurich
   Gloriastrasse 35
   8092 Zurich
   Switzerland

   Email: mirja.kuehlewind@tik.ee.ethz.ch