            The Rise of the Middle and the Future of End-to-End:
          Reflections on the Evolution of the Internet Architecture

Status of this Memo

   This memo provides information for the Internet community. It does
   not specify an Internet standard of any kind. Distribution of this
   memo is unlimited.

Copyright Notice

Abstract

   The end-to-end principle is the core architectural guideline of the
   Internet. In this document, we briefly examine the development of
   the end-to-end principle as it has been applied to the Internet
   architecture over the years. We discuss current trends in the
   evolution of the Internet architecture in relation to the end-to-end
   principle, and try to draw some conclusion about the evolution of the
   end-to-end principle, and thus for the Internet architecture which it
   supports, in light of these current trends.

Table of Contents

1. Introduction

   One of the key architectural guidelines of the Internet is the end-
   to-end principle in the papers by Saltzer, Reed, and Clark [1][2].
   The end-to-end principle was originally articulated as a question of
   where best not to put functions in a communication system. Yet, in
   the ensuing years, it has evolved to address concerns of maintaining
   openness, increasing reliability and robustness, and preserving the
   properties of user choice and ease of new service development as
   discussed by Blumenthal and Clark in [3]; concerns that were not part
   of the original articulation of the end-to-end principle.

   In this document, we examine how the interpretation of the end-to-end
   principle has evolved over the years, and where it stands currently.
   We examine trends in the development of the Internet that have led to
   pressure to define services in the network, a topic that has already
   received some amount of attention from the IAB in RFC 3238 [5]. We
   describe some considerations about how the end-to-end principle might
   evolve in light of these trends.

2. A Brief History of the End-to-End Principle

2.1. In the Beginning...

   The end-to-end principle was originally articulated as a question of
   where best to put functions in a communication system:

      The function in question can completely and correctly be
      implemented only with the knowledge and help of the application
      standing at the end points of the communication system.
      Therefore, providing that questioned function as a feature of the
      communication system itself is not possible. (Sometimes an
      incomplete version of the function provided by the communication
      system may be useful as a performance enhancement.) [1].

   A specific example of such a function is delivery guarantees [1].
   The original ARPANET returned a message "Request for Next Message"
   whenever it delivered a packet. Although this message was found to
   be useful within the network as a form of congestion control, since
   the ARPANET refused to accept another message to the same destination
   until the previous acknowledgment was returned, it was never
   particularly useful as an indication of guaranteed delivery. The
   problem was that the host stack on the sending host typically doesn't
   want to know just that the network delivered a packet, but rather the
   stack layer on the sending host wants to know that the stack layer on
   the receiving host properly processed the packet. In terms of modern
   IP stack structure, a reliable transport layer requires an indication
   that transport processing has successfully completed, such as given

by TCP's ACK message [4], and not simply an indication from the IP
layer that the packet arrived. Similarly, an application layer
protocol may require an application-specific acknowledgement that
contains, among other things, a status code indicating the
disposition of the request.

The specific examples given in [1] and other references at the time
[2] primarily involve transmission of data packets: data integrity,
delivery guarantees, duplicate message suppression, per packet
encryption, and transaction management. From the viewpoint of
today's Internet architecture, we would view most of these as
transport layer functions (data integrity, delivery guarantees,
duplicate message suppression, and perhaps transaction management),
others as network layer functions with support at other layers where
necessary (for example, packet encryption), and not application layer
functions.

2.2. ...In the Middle...

As the Internet developed, the end-to-end principle gradually widened
to concerns about where best to put the state associated with
applications in the Internet: in the network or at end nodes. The
best example is the description in RFC 1958 [6]:

> This principle has important consequences if we require
> applications to survive partial network failures. An end-to-end
> protocol design should not rely on the maintenance of state (i.e.,
> information about the state of the end-to-end communication)
> inside the network. Such state should be maintained only in the
> endpoints, in such a way that the state can only be destroyed when
> the endpoint itself breaks (known as fate-sharing). An immediate
> consequence of this is that datagrams are better than classical
> virtual circuits. The network's job is to transmit datagrams as
> efficiently and flexibly as possible. Everything else should be
> done at the fringes.

The original articulation of the end-to-end principle - that
knowledge and assistance of the end point is essential and that
omitting such knowledge and implementing a function in the network
without such knowledge and assistance is not possible - took a while
to percolate through the engineering community, and had evolved by
this point to a broad architectural statement about what belongs in
the network and what doesn't. RFC 1958 uses the term "application"
to mean the entire network stack on the end node, including network,
transport, and application layers, in contrast to the earlier
articulation of the end-to-end principle as being about the
communication system itself. "Fate-sharing" describes this quite
clearly: the fate of a conversation between two applications is only

shared between the two applications; the fate does not depend on anything in the network, except for the network's ability to get packets from one application to the other.

The end-to-end principle in this formulation is specifically about what kind of state is maintained where:

> To perform its services, the network maintains some state information: routes, QoS guarantees that it makes, session information where that is used in header compression, compression histories for data compression, and the like. This state must be self-healing; adaptive procedures or protocols must exist to derive and maintain that state, and change it when the topology or activity of the network changes. The volume of this state must be minimized, and the loss of the state must not result in more than a temporary denial of service given that connectivity exists. Manually configured state must be kept to an absolute minimum.[6]

In this formulation of the end-to-end principle, state involved in getting packets from one end of the network to the other is maintained in the network. The state is "soft state," in the sense that it can be quickly dropped and reconstructed (or even required to be periodically renewed) as the network topology changes due to routers and switches going on and off line. "Hard state", state upon which the proper functioning of the application depends, is only maintained in the end nodes. This formulation of the principle is a definite change from the original formulation of the principle, about end node participation being required for proper implementation of most functions.

In summary, the general awareness both of the principle itself and of its implications for how unavoidable state should be handled grew over time to become a (if not the) foundation principle of the Internet architecture.

2.3. ...And Now.

An interesting example of how the end-to-end principle continues to influence the technical debate in the Internet community is IP mobility. The existing Internet routing architecture severely constrains how closely IP mobility can match the end-to-end principle without making fundamental changes. Mobile IPv6, described in the Mobile IPv6 specification by Johnson, Perkins, and Arkko [7], requires a routing proxy in the mobile node's home network (the Home Agent) for maintaining the mapping between the mobile node's routing locator, the care of address, and the mobile node's node identifier, the home address. But the local subnet routing proxy (the Foreign Agent), which was a feature of the older Mobile IPv4 design [8] that

compromised end-to-end routing, has been eliminated. The end node
now handles its own care of address. In addition, Mobile IPv6
includes secure mechanisms for optimizing routing to allow end-to-end
routing between the mobile end node and the correspondent node,
removing the need to route through the global routing proxy at the
home agent. These features are all based on end to end
considerations. However, the need for the global routing proxy in
the Home Agent in Mobile IPv6 is determined by the aliasing of the
global node identifier with the routing identifier in the Internet
routing architecture, a topic that was discussed in an IAB workshop
and reported in RFC 2956 [9], and that hasn't changed in IPv6.

Despite this constraint, the vision emerging out of the IETF working
groups developing standards for mobile networking is of a largely
autonomous mobile node with multiple wireless link options, among
which the mobile node picks and chooses. The end node is therefore
responsible for maintaining the integrity of the communication, as
the end-to-end principle implies. This kind of innovative
application of the end-to-end principle derives from the same basic
considerations of reliability and robustness (wireless link
integrity, changes in connectivity and service availability with
movement, etc.) that motivated the original development of the end-
to-end principle. While the basic reliability of wired links,
routing, and switching equipment has improved considerably since the
end-to-end principle was formalized 15 years ago, the reliability or
unreliability of wireless links is governed more strongly by the
basic physics of the medium and the instantaneous radio propagation
conditions.

3. Trends Opposed to the End-to-End Principle

While the end-to-end principle continues to provide a solid
foundation for much IETF design work, the specific application of the
end-to-end principle described in RFC 1958 has increasingly come into
question from various directions. The IAB has been concerned about
trends opposing the end-to-end principle for some time, for example
RFC 2956 [9] and RFC 2775 [12]. The primary focus of concern in
these documents is the reduction in transparency due to the
introduction of NATs and other address translation mechanisms in the
Internet, and the consequences to the end-to-end principle of various
scenarios involving full, partial, or no deployment of IPv6. More
recently, the topic of concern has shifted to the consequences of
service deployment in the network. The IAB opinion on Open Pluggable
Edge Services (OPES) in RFC 3238 [5] is intended to assess the
architectural desirability of defining services in the network and to
raise questions about how such services might result in compromises

of privacy, security, and end-to-end data integrity. Clark, et al.
in [10] and Carpenter in RFC 3234 [11] also take up the topic of
service definition in the network.

Perhaps the best review of the forces militating against the end-to-
end principle is by Blumenthal and Clark in [3]. The authors make
the point that the Internet originally developed among a community of
like-minded technical professionals who trusted each other, and was
administered by academic and government institutions who enforced a
policy of no commercial use. The major stakeholders in the Internet
are quite different today. As a consequence, new requirements have
evolved over the last decade. Examples of these requirements are
discussed in the following subsections. Other discussions about
pressures on the end-to-end principle in today's Internet can be
found in the discussion by Reed [13] and Moors' paper in the 2002
IEEE International Communications Conference [14].

3.1. Need for Authentication

Perhaps the single most important change from the Internet of 15
years ago is the lack of trust between users. Because the end users
in the Internet of 15 years ago were few, and were largely dedicated
to using the Internet as a tool for academic research and
communicating research results (explicit commercial use of the
Internet was forbidden when it was run by the US government), trust
between end users (and thus authentication of end nodes that they
use) and between network operators and their users was simply not an
issue in general. Today, the motivations of some individuals using
the Internet are not always entirely ethical, and, even if they are,
the assumption that end nodes will always co-operate to achieve some
mutually beneficial action, as implied by the end-to-end principle,
is not always accurate. In addition, the growth in users who are
either not technologically sophisticated enough or simply
uninterested in maintaining their own security has required network
operators to become more proactive in deploying measures to prevent
naive or uninterested users from inadvertently or intentionally
generating security problems.

While the end-to-end principle does not require that users implicitly
trust each other, the lack of trust in the Internet today requires
that application and system designers make a choice about how to
handle authentication, whereas that choice was rarely apparent 15
years ago. One of the most common examples of network elements
interposing between end hosts are those dedicated to security:
firewalls, VPN tunnel endpoints, certificate servers, etc. These
intermediaries are designed to protect the network from unimpeded
attack or to allow two end nodes whose users may have no inherent
reason to trust each other to achieve some level of authentication.

At the same time, these measures act as impediments for end-to-end
communications. Third party trust intermediaries are not a
requirement for security, as end-to-end security mechanisms, such as
S/MIME [15], can be used instead, and where third party measures such
as PKI infrastructure or keys in DNS are utilized to exchange keying
material, they don't necessarily impinge on end-to-end traffic after
authentication has been achieved. Even if third parties are
involved, ultimately it is up to the endpoints and their users in
particular, to determine which third parties they trust.

3.2. New Service Models

New service models inspired by new applications require achieving the
proper performance level as a fundamental part of the delivered
service. These service models are a significant change from the
original best effort service model. Email, file transfer, and even
Web access aren't perceived as failing if performance degrades,
though the user may become frustrated at the time required to
complete the transaction. However, for streaming audio and video, to
say nothing of real time bidirectional voice and video, achieving the
proper performance level, whatever that might mean for an acceptable
user experience of the service, is part of delivering the service,
and a customer contracting for the service has a right to expect the
level of performance for which they have contracted. For example,
content distributors sometimes release content via content
distribution servers that are spread around the Internet at various
locations to avoid delays in delivery if the server is topologically
far away from the client. Retail broadband and multimedia services
are a new service model for many service providers.

3.3. Rise of the Third Party

Academic and government institutions ran the Internet of 15 years
ago. These institutions did not expect to make a profit from their
investment in networking technology. In contrast, the network
operator with which most Internet users deal today is the commercial
ISP. Commercial ISPs run their networks as a business, and their
investors rightly expect the business to turn a profit. This change
in business model has led to a certain amount of pressure on ISPs to
increase business prospects by deploying new services.

In particular, the standard retail dialup bit pipe account with email
and shell access has become a commodity service, resulting in low
profit margins. While many ISPs are happy with this business model
and are able to survive on it, others would like to deploy different
service models that have a higher profit potential and provide the
customer with more or different services. An example is retail
broadband bit pipe access via cable or DSL coupled with streaming

multimedia. Some ISPs that offer broadband access also deploy
content distribution networks to increase the performance of
streaming media. These services are typically deployed so that they
are only accessible within the ISP's network, and as a result, they
do not contribute to open, end-to-end service. From an ISP's
standpoint, however, offering such service is an incentive for
customers to buy the ISP's service.

ISPs are not the only third party intermediary that has appeared
within the last 10 years. Unlike the previous involvement of
corporations and governments in running the Internet, corporate
network administrators and governmental officials have become
increasingly demanding of opportunities to interpose between two
parties in an end-to-end conversation. A benign motivation for this
involvement is to mitigate the lack of trust, so the third party acts
as a trust anchor or enforcer of good behavior between the two ends.
A less benign motivation is for the third parties to insert policy
for their own reasons, perhaps taxation or even censorship. The
requirements of third parties often have little or nothing to do with
technical concerns, but rather derive from particular social and
legal considerations.

4. Whither the End-to-End Principle?

Given the pressures on the end-to-end principle discussed in the
previous section, a question arises about the future of the end-to-
end principle. Does the end-to-end principle have a future in the
Internet architecture or not? If it does have a future, how should
it be applied? Clearly, an unproductive approach to answering this
question is to insist upon the end-to-end principle as a
fundamentalist principle that allows no compromise. The pressures
described above are real and powerful, and if the current Internet
technical community chooses to ignore these pressures, the likely
result is that a market opportunity will be created for a new
technical community that does not ignore these pressures but which
may not understand the implications of their design choices. A more
productive approach is to return to first principles and re-examine
what the end-to-end principle is trying to accomplish, and then
update our definition and exposition of the end-to-end principle
given the complexities of the Internet today.

4.1. Consequences of the End-to-End Principle

In this section, we consider the two primary desirable consequences
of the end-to-end principle: protection of innovation and provision
of reliability and robustness.

4.1.1. Protection of Innovation

One desirable consequence of the end-to-end principle is protection
of innovation. Requiring modification in the network in order to
deploy new services is still typically more difficult than modifying
end nodes. The counterargument - that many end nodes are now
essentially closed boxes which are not updatable and that most users
don't want to update them anyway - does not apply to all nodes and
all users. Many end nodes are still user configurable and a sizable
percentage of users are "early adopters," who are willing to put up
with a certain amount of technological grief in order to try out a
new idea. And, even for the closed boxes and uninvolved users,
downloadable code that abides by the end-to-end principle can provide
fast service innovation. Requiring someone with a new idea for a
service to convince a bunch of ISPs or corporate network
administrators to modify their networks is much more difficult than
simply putting up a Web page with some downloadable software
implementing the service.

4.1.2. Reliability and Trust

Of increasing concern today, however, is the decrease in reliability
and robustness that results from deliberate, active attacks on the
network infrastructure and end nodes. While the original developers
of the Internet were concerned by large-scale system failures,
attacks of the subtlety and variety that the Internet experiences
today were not a problem during the original development of the
Internet. By and large, the end-to-end principle was not addressed
to the decrease in reliability resulting from attacks deliberately
engineered to take advantage of subtle flaws in software. These
attacks are part of the larger issue of the trust breakdown discussed
in Section 3.1. Thus, the issue of the trust breakdown can be
considered another forcing function on the Internet architecture.

The immediate reaction to this trust breakdown has been to try to
back fit security into existing protocols. While this effort is
necessary, it is not sufficient. The issue of trust must become as
firm an architectural principle in protocol design for the future as
the end-to-end principle is today. Trust isn't simply a matter of
adding some cryptographic protection to a protocol after it is
designed. Rather, prior to designing the protocol, the trust
relationships between the network elements involved in the protocol
must be defined, and boundaries must be drawn between those network
elements that share a trust relationship. The trust boundaries
should be used to determine what type of communication occurs between
the network elements involved in the protocol and which network
elements signal each other. When communication occurs across a trust
boundary, cryptographic or other security protection of some sort may

be necessary. Additional measures may be necessary to secure the
protocol when communicating network elements do not share a trust
relationship. For example, a protocol might need to minimize state
in the recipient prior to establishing the validity of the
credentials from the sender in order to avoid a memory depletion DoS
attack.

4.2. The End-to-End Principle in Applications Design

The concern expressed by the end-to-end principle is applicable to
applications design too. Two key points in designing application
protocols are to ensure they don't have any dependencies that would
break the end-to-end principle and to ensure that they can identify
end points in a consistent fashion. An example of the former is
layer violations - creating dependencies that would make it
impossible for the transport layer, for example, to do its work
appropriately. Another issue is the desire to insert more
applications infrastructure into the network. Architectural
considerations around this issue are discussed in RFC 3238 [5]. This
desire need not result in a violation of the end-to-end principle if
the partitioning of functioning is done so that services provided in
the network operate with the explicit knowledge and involvement of
endpoints, when such knowledge and involvement is necessary for the
proper functioning of the service. The result becomes a distributed
application, in which the end-to-end principle applies to each
connection involved in implementing the application.

5. Internet Standards as an Arena for Conflict

Internet standards have increasingly become an arena for conflict
[10]. ISPs have certain concerns, businesses and government have
others, and vendors of networking hardware and software still others.
Often, these concerns conflict, and sometimes they conflict with the
concerns of the end users. For example, ISPs are reluctant to deploy
interdomain QoS services because, among other reasons, every known
instance creates a significant and easily exploited DoS/DDoS
vulnerability. However, some end users would like to have end-to-
end, Diffserv or Intserv-style QoS available to improve support for
voice and video multimedia applications between end nodes in
different domains, as discussed by Huston in RFC 2990 [16]. In this
case, the security, robustness, and reliability concerns of the ISP
conflict with the desire of users for a different type of service.

These conflicts will inevitably be reflected in the Internet
architecture going forward. Some of these conflicts are impossible
to resolve on a technical level, and would not even be desirable,
because they involve social and legal choices that the IETF is not
empowered to make (for a counter argument in the area of privacy, see

Goldberg, et al. [17]). But for those conflicts that do involve
technical choices, the important properties of user choice and
empowerment, reliability and integrity of end-to-end service,
supporting trust and "good network citizen behavior," and fostering
innovation in services should be the basis upon which resolution is
made. The conflict will then play out on the field of the resulting
architecture.

6. Conclusions

The end-to-end principle continues to guide technical development of
Internet standards, and remains as important today for the Internet
architecture as in the past. In many cases, unbundling of the end-
to-end principle into its consequences leads to a distributed
approach in which the end-to-end principle applies to interactions
between the individual pieces of the application, while the unbundled
consequences, protection of innovation, reliability, and robustness,
apply to the entire application. While the end-to-end principle
originated as a focused argument about the need for the knowledge and
assistance of end nodes to properly implement functions in a
communication system, particular second order properties developed by
the Internet as a result of the end-to-end principle have come to be
recognized as being as important, if not more so, than the principle
itself. End user choice and empowerment, integrity of service,
support for trust, and "good network citizen behavior" are all
properties that have developed as a consequence of the end-to-end
principle. Recognizing these properties in a particular proposal for
modifications to the Internet has become more important than before
as the pressures to incorporate services into the network have
increased. Any proposal to incorporate services in the network
should be weighed against these properties before proceeding.

7. Acknowledgements

Many of the ideas presented here originally appeared in the works of
Dave Clark, John Wroclawski, Bob Braden, Karen Sollins, Marjory
Blumenthal, and Dave Reed on forces currently influencing the
evolution of the Internet. The authors would particularly like to
single out the work of Dave Clark, who was the original articulator
of the end-to-end principle and who continues to inspire and guide
the evolution of the Internet architecture, and John Wroclawski, with
whom conversations during the development of this paper helped to
clarify issues involving tussle and the Internet.

8. Security Considerations

This document does not propose any new protocols, and therefore does not involve any security considerations in that sense. However, throughout this document, there are discussions of the privacy and integrity issues and the architectural requirements created by those issues.

9. Informative References

[1]   Saltzer, J.H., Reed, D.P., and Clark, D.D., "End-to-End Arguments in System Design," ACM TOCS, Vol 2, Number 4, November 1984, pp 277-288.

[2]   Clark, D., "The Design Philosophy of the DARPA Internet Protocols," Proc SIGCOMM 88, ACM CCR Vol 18, Number 4, August 1988, pp. 106-114.

[3]   Blumenthal, M., Clark, D.D., "Rethinking the design of the Internet: The end-to-end arguments vs. the brave new world", ACM Transactions on Internet Technology, Vol. 1, No. 1, August 2001, pp 70-109.

[4]   Postel, J., "Transmission Control Protocol", STD 7, RFC 793, September 1981.

[5]   Floyd, S. and L. Daigle, "IAB Architectural and Policy Considerations for Open Pluggable Edge Services", RFC 3238, January 2002.

[6]   Carpenter, B., Ed., "Architectural Principles of the Internet", RFC 1958, June 1996.

[7]   Johnson, D., Perkins, C. and J. Arkko, "Mobility Support in IPv6", Work in Progress.

[8]   Perkins, C., Ed., "IP Mobility Support for IPv4", RFC 3344, August 2002.

[9]   Kaat, M., "Overview of 1999 IAB Network Layer Workshop," RFC 2956, October 2000.

[10] Clark, D.D., Wroclawski, J., Sollins, K., and Braden, B., "Tussle in Cyberspace: Defining Tomorrow's Internet", Proceedings of Sigcomm 2002.

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

[12] Carpenter, B., "Internet Transparency", RFC 2775, February 2000.

[13] Reed, D., "The End of the End-to-End Argument?",
     http://www.reed.com/dprframeweb/
     dprframe.asp?section=paper&fn=endofendtoend.html, April 2000.

[14] Moors, T., "A Critical Review of End-to-end Arguments in System
     Design," Proc. 2000 IEEE International Conference on
     Communications, pp. 1214-1219, April, 2002.

[15] Ramsdell, B., Ed., "S/MIME Version 3 Message Specification", RFC
     2633, June 1999.

[16] Huston, G., "Next Steps for the IP QoS Architecture", RFC 2990,
     November 2000.

[17] Goldberg, I., Wagner, D., and Brewer, E., "Privacy-enhancing
     technologies for the Internet," Proceedings of IEEE COMPCON 97,
     pp. 103-109, 1997.

10. Author Information

Internet Architecture Board
EMail: iab@iab.org

IAB Membership at time this document was completed:

   Bernard Aboba
   Harald Alvestrand
   Rob Austein
   Leslie Daigle
   Patrik Faltstrom
   Sally Floyd
   Jun-ichiro Itojun Hagino
   Mark Handley
   Geoff Huston
   Charlie Kaufman
   James Kempf
   Eric Rescorla
   Mike St. Johns

11. Full Copyright Statement

Intellectual Property

Acknowledgement