Network Working Group
Internet-Draft
Intended status: Informational
Expires: December 29, 2021

P. Bottorff
Hewlett Packard Enterprise Co.
July 5, 2021

Multi-tenant Data Center Use Case for IPsec Load Balancing
draft-bottorff-ipsecme-mtdcuc-ipsec-lb-00

Abstract

IPsec is of increasing importance within data centers to secure tunnels used to carry multi-tenant traffic encapsulated using the Network Virtualization over L3 (NVO3) protocols. Encrypting NVO3 tunnels provides defense against bad actors within the physical underlay network from monitoring or injecting overlay traffic from outside the NVO3 infrastructure. When securing data center tunnels using IPsec it becomes crucial to retain entropy within the outer IPsec packet headers to facilitate load balancing over the highly meshed networks used in these environments. While entropy is necessary to support load distribution algorithms it is also important that the entropy codes used retain integrity of flows to prevent performance deterioration resulting from packet re-ordering. Here, we describe a use case for load balancing IPsec traffic within multi-tenant data centers.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

Load balancing is essential within data centers to achieve high utilization of the meshed networks common in these environments. Typically, the load on the network is scattered over the mesh network using a hash of the outer packet headers (i.e. a 5-tuple hash). When a tunneling protocol is used over a data center mesh network packets are addressed from tunnel end point to tunnel end point (e.g. servers) which does not provide the entropy required to spread the traffic over the data center mesh network. To provide load balancing support, tunnel protocols used in the data center need to provide entropy codes within their outer packet headers to support load balancing.

While spreading traffic over a data center mesh network mis-ordering of packet flows needs to be avoided to prevent slowing operations caused by packet order recovery. To retain flow alignment within tunneling protocols entropy codes need to be based on the flow of the encapsulated packets.

Multi-tenant data center using Network Virtualization over L3 (NVO3) [RFC7365, RFC8014, RFC8394] create virtual networks interconnecting virtual servers within an overlay on top of the physical underlay network. In these NVO3 networks virtual network packets are multiplexed into tunnels which extend over the physical underlay network. The encapsulations used to in the NVO3 tunnels (i.e. VxLAN, GENEVE, GUE) have an outer UDP header followed by an outer NVO3 header. The NVO3 header carries a key called the Virtual Network Identifier (VNI) which identifies the virtual network within the virtual overlay network. Each of the virtual overlay networks is a separate subnetwork which can have its own IP and L2 virtual address space.

Since a single NVO3 tunnel is used between communicating servers, any server to server connection has the same IP source, destination, and UDP destination port. To retain entropy for load balancing the NVO3 protocols use the UDP source port to hold a hash of the encapsulated inner packet. This outer UDP source port provides the entropy necessary for spreading traffic over the mesh based on the encapsulated flow. Other fields such as the IPv6 flow label could also be used, however these are not universally supported in data center switching infrastructures, while the use of UDP source port is broadly available in data center switches, routers, and middle boxes.

The NVO3 protocols isolate tenant virtual networks based on the VNI identifiers carried in the tunnel headers. Since any bad actor connected to the data center underlay network could spoof an encapsulation transporting a virtual network and any device in the middle of the communication can monitor the tenant networks, NVO3 networks must operate in a secure perimeter. With the rise of more aggressive bad actors it is desirable to provide secure connections for NOV3 tunnels to eliminate the threat of a server or switch within the data center underlay monitoring or interfering with the operation of virtual networks throughout the data center.

Encryption of [RFC4301, RFC4303, RFC7321] the NVO3 tunnels can provide protection against devices outside the virtual overlay from monitoring, spoofing or interfering with the virtual networks. This can be done using IPsec to encrypt the tunnels carrying virtual networks between servers. Since the tunnels can be encrypted using smart network interfaces this method can be very efficient, retaining the high performance required within data centers.

If we apply IPsec directly to the NVO3 tunnels the IP source and destination as well as the protocol type and SPI will be the same for each server to server communication, therefore we will lose the entropy needed to support the data center mesh network. Internet Draft “Encapsulating IPsec ESP in UDP for Load-balancing” [IPSEC-LB], which proposes using the source port of IPsec transport mode ESP in UDP encapsulated packets for entropy, provides the solution needed to support the use of IPsec to secure the tunnels used for NVO3 traffic in data centers. By using transport mode ESP in UDP encapsulation of NVO3 tunnels entropy can be provided by using the UDP source port just as was done originally for the NVO3 UDP encapsulations.

2. Generic Data Center Network Architecture

Figure 1 depicts a typical Clos mesh network [RFC7938] used in a data center. Each server is typically redundantly connected to a set of switches located at the Top of the Rack (ToR) switch (also called a leaf switch). Each of these switches is, in turn, connected to a set of switches for the row of racks commonly called the End of Row (EoR) switch (also called a spine switch). Typically these redundant links are managed either by L2 link aggregation protocols [IEEE-AX, IEEE-Q] or by L3 equal cost multi-path protocols [RFC7938]. The number of links interconnecting these layers varies depending on the bandwidth and resiliency requirements. The figure shows a two level hierarchy, however it is common for data centers to have a three level hierarchy where a similar Clos mesh is used to interconnect rows of servers.

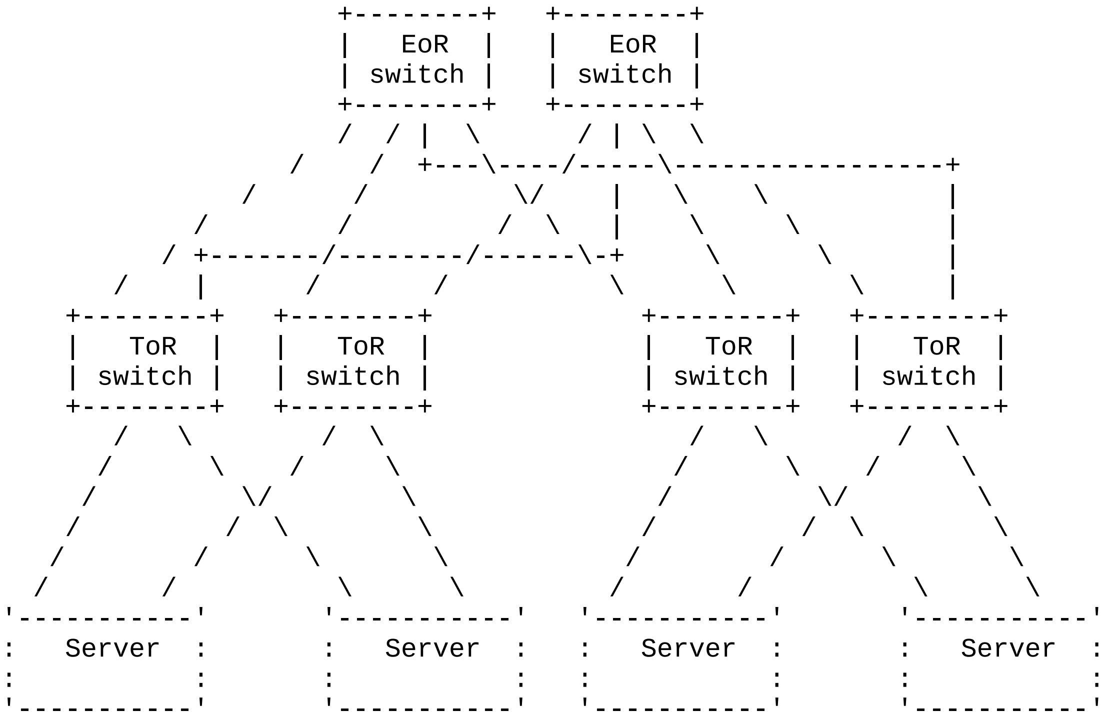

Figure 1: Typical Data Center Mesh Network

To distribute packets over the network paths typically a hash function is used to reduce the fields within the outer packet headers into a group of flows. The group is then allocated to a network link or path. In the simplest and most common implementation the distributions are done based on a hash. In more sophisticated implementation additional load data and timing information may be used to move flow groups based on load estimates.

3. Network Virtualization Over L3 (NVO3) Architecture

In an NVO3 multi-tenant data center the physical interconnect depicted in figure 1 is used as the underlay physical IP network where IP addresses are assigned to servers.

```
      +--------+                                    +--------+
      | Tenant +--+                            +----| Tenant |
      | System |  |                           (')   | System |
      +--------+  |    .................     (   )  +--------+
                  |  +---+           +---+    (_)
                  +--|NVE|---+   +---|NVE|-----+
                     +---+   |   |   +---+
                     / .    +-----+      .
                    /  . +--| NVA |--+   .
                   /   . |  +-----+   \  .
                  |    . |             \ .
                  |    . |   Overlay   +--+--++--------+
      +--------+  |    . |   Network   | NVE || Tenant |
      | Tenant +--+    . |             |     || System |
      | System |       .  \ +---+      +--+--++--------+
      +--------+       .....|NVE|.........
                            +---+
                              |
                              |
                    =====================
                      |               |
                  +--------+      +--------+
                  | Tenant |      | Tenant |
                  | System |      | System |
                  +--------+      +--------+
```

Figure 2: NVO3 Architecture Reference Diagram

The NVO3 protocols are used on top of the physical underlay network to create virtual networks which overlay the physical underlay network. The virtual networks are carried over the underlay encapsulated in an NVO3 encapsulation protocol such as GENEVE [RFC8926], VxLAN [RFC7348], or GUE. These encapsulations indicate the virtual network by encoding a Virtual Network Identifier (VNI) within the encapsulation header.

Figure 2 is a copy of the NVO3 architecture [RFC8014] reference diagram. In figure 2 the Network Virtual Edge (NVE) entities provide the tunnel terminations for the encapsulation protocols. The NVEs can be located within the server’s hypervisor, within smart NICs on the servers, or within switches of the physical network [RFC8394].

The tenant systems (TS) are virtualized servers. These may be virtual machines, containers, or physical servers that are connected over the virtual networks and multiplexed into NOV3 tunnel by the NVEs.

The network virtualization authority (NVA) manages the creation and configuration of virtual networks by configuring the NVEs.

4. Load Balancing Secure NVO3 tunnels

The NVO3 protocols isolate tenant virtual networks based on the Virtual Network Identifiers carried in the tunnel header. Since any bad actor inside the data center could spoof an encapsulation transporting a virtual network and any device in the middle of the communication can monitor the tenant networks, these networks are only secure when operating within a secure perimeter. With the rise of more aggressive bad actors it is desirable to provide secure connections for NVO3 tunnels to eliminate the threat of a server or switch within the data center underlay monitoring or interfering with the operation of overlay virtual networks.

Encryption of [RFC4301, RFC4303, RFC7321] the NVO3 tunnels provides protection against devices outside the virtual overlay from monitoring, spoofing or interfering with the virtual networks. This can be done using IPsec to encrypt the tunnels carrying virtual networks between servers. Since the tunnels can be encrypted using smart network interfaces this method can be very efficient, retaining the high performance required within data centers. However since the resulting tunnel headers don’t provide enough entropy to support load balancing over the data center mesh networks the resulting network bandwidth could be greatly reduced.

To solve the load balancing problem produced by securing the NVO3 tunnels using IPsec the method proposed in Internet Draft “Encapsulating IPsec ESP in UDP for Load-balancing” [IPSEC-LB] can be used. Applying I-D.draft-xu-ipsecme-esp-in-udp-lb the NVO3 tunnels are encapsulated as IPsec transparent mode ESP in UDP packets. Given the UDP header on the outside of the IPsec tunnel the source port can be used for entropy. By copying the source port from the original

NVO3 encapsulation into the IPsec ESP in UDP header it is possible to retain the flow identity of the encrypted virtual network. Since the NVO3 encapsulation source port contains an entropy code based on the encapsulated overlay packet the resulting packet will provide the entropy necessary to support load balancing in the data center mesh network.

5. Security Considerations

Here we use IPsec to secure NVO3 tunnels between data center NVEs to prevent attacks from servers or switches located within the data center physical underlay, however outside of the overlay networks or the NVE tunnel terminations and NVA managers. To fully secure a multi-tenant network additional security methods need to be used to prevent attackers from infiltrating the overlay infrastructure including the Tenant Systems, NVEs and NVA.

6. IANA Considerations

The Internet Draft “Encapsulating IPsec ESP in UDP for Load-balancing” [IPSEC-LB] proposes getting a new UDP destination port assignment for use with load balanced IPsec. The use of a new port would prevent existing implementations of IKE from operating with a load balanced transparent mode ESP in UDP stream. It does not appear this is necessary. Instead, the existing ESP in UDP port 4500 could be used, provided both ends of the connection are configured to exchanging ESP in UDP with an entropy code in the UDP source port. If the existing ESP in UDP port 4500 is used, then there are no IANA considerations since no new code points are necessary.

7. Conclusions

IPsec may be used to secure the underlay of a NVO3 multi-tenant data center by encrypting the NVO3 tunnels. To make IPsec a viable solution the IPsec tunnels need to provide load balancing.

By applying the proposal in Internet Draft “Encapsulating IPsec ESP in UDP for Load-balancing” [IPSEC-LB], entropy can be added to the IPsec packet header using the UDP source port of the ESP in UDP IPsec packets. In particular, the source port of the original NVO3 tunnel header can be copied to the new IPsec ESP in UDP source port providing the necessary entropy while retaining the flow identity of the encapsulated overlay packet.

8. Normative References

[IPSEC-LB] Xu, X., Hegde, S., Pismenny, B., Zhang, D., and Xia, L., “Encapsulating IPsec ESP in UDP for Load-balancing”, December 2020, <https://datatracker.ietf.org/doc/draft-xu-ipsecme-esp-in-udp-lb/>.

[IEEE-AX] "IEEE Standard for Local and Metropolitan Area Networks--Link Aggregation," in IEEE Std 802.1AX-2020 (Revision of IEEE Std 802.1AX-2014), vol., no., pp.1-333, 29 May 2020, doi: 10.1109/IEEESTD.2020.9105034.

[IEEE-Q] "IEEE Standard for Local and Metropolitan Area Network--Bridges and Bridged Networks," in IEEE Std 802.1Q-2018 (Revision of IEEE Std 802.1Q-2014), vol., no., pp.1-1993, 6 July 2018, doi: 10.1109/IEEESTD.2018.8403927.

[RFC768] Postel, J., "User Datagram Protocol", STD 6, RFC 768, DOI 10.17487/RFC0768, August 1980, <https://www.rfc-editor.org/info/rfc768/>.

[RFC1122] Braden, R., Ed., “Requirements for Internet Hosts – Communication Layers”, STD 3, RFC 112, DOI 10.17487/RFC1122, October 1989, <https://www.rfc-editor.org/info/rfc112/>.

[RFC1191] Mogul, J. and S. Deering, "Path MTU discovery", RFC 1191, DOI 10.17487/RFC1191, November 1990, <https://www.rfc-editor.org/info/rfc1191/>.

[RFC2003] Perkins, C., "IP Encapsulation within IP", RFC 2003, DOI 10.17487/RFC2003, October 1996, <https://www.rfc-editor.org/info/rfc2003/>.

[RFC3948] Huttunen, A., Swander, B., Volpe, V., DiBurro, L., and M. Stenberg, "UDP Encapsulation of IPsec ESP Packets", RFC 3948, DOI 10.17487/RFC3948, January 2005, <https://www.rfc-editor.org/info/rfc3948/>.

[RFC4301] Kent,S., Seo, K., “Security Architecture for the Internet Protocol”, RFC 4301, December 2005, <https://datatracker.ietf.org/doc/rfc4301/>

[RFC4303] Kent, S., “IP Encapsulating Security Payload”, RFC 4303, December 2005, <https://datatracker.ietf.org/doc/rfc4303/>

[RFC4821] Mathis, M. and J. Heffner, "Packetization Layer Path MTU Discovery", RFC 4821, DOI 10.17487/RFC4821, March 2007, <https://www.rfc-editor.org/info/rfc4821/>.

[RFC7296] Kaufman, C., Hoffman, P., Nir, Y., Eronen, P., and T. Kivinen, "Internet Key Exchange Protocol Version 2 (IKEv2)", STD 79, RFC 7296, DOI 10.17487/RFC7296, October 2014, <https://www.rfc-editor.org/info/rfc7296/>.

[RFC7321] McGrew, D., Hoffman, P., “Cryptographic Algorithm Implementation Requirements and Usage Guidance for Encapsulating Security Payload (ESP) and Authentication Header (AH)”, RFC 7321, August 2014, https://datatracker.ietf.org/doc/rfc7321/

[RFC7348] Mahalingam, M., et. al., “Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtual Layer 2 Networks over Layer 3 Networks”, RFC 7348, August 2014, <https://datatracker.ietf.org/doc/rfc7348/>

[RFC7365] Lasserre, M., et al, “Framework for Data Center (DC) Network Virtualization”, October 2014, <https://datatracker.ietf.org/doc/rfc7365/>

[RFC7938] Lapukhov, P., Premji, A., Mitchell, J., Ed., “Use of BGP for Routing in Large-Scale Data Centers”, August 2016, <https://datatracker.ietf.org/doc/rfc7938/>.

[RFC8014] Black, D., et al, “An Architecture for Data-Center Network Virtualization over Layer 3 (NVO3)”, December 2016, <https://datatracker.ietf.org/doc/rfc8014>

[RFC8394] Li, Y., Eastlake, D., Kreeger, L. Narten, T., Black, D., “Split Network Virtualization Edge (Split-NVE) Control-Plane Requirements”, RFC 8394, May 2018, <https://datatracker.ietf.org/doc/rfc8394/>

[RFC8926] Gross, J., Ganga, I., Sridhar, T., “Geneve: Generic Network Virtualization Encapsulation”, November 2020, <https://datatracker.ietf.org/doc/rfc8926/>

9. Informative References

[RFC6438] Carpenter, B. and S. Amante, "Using the IPv6 Flow Label for Equal Cost Multipath Routing and Link Aggregation in Tunnels", RFC 6438, DOI 10.17487/RFC6438, November 2011, <https://www.rfc-editor.org/info/rfc6438>.

[RFC8200] Deering, S., Hiden, R., “Internet Protocol, Version 6 (IPv6) Specification”, STD 76, RFC 8200, July 2017, <https://datatracker.ietf.org/doc/rfc8200/>

[RFC8201] McCann, J., Deering, S., Mogul, J., Hiden, R., Ed., “Path MTU Discovery for IP version 6”, STD 87, RFC 8201, July 2017, <https://datatracker.ietf.org/doc/rfc8201/>

10. Acknowledgments

This document was prepared using 2-Word-v2.0.template.dot.

Authors' Addresses

Paul Bottorff
Aruba a Hewlett Packard Enterprise Co
8000 Foothill Blvd.
Roseville, CA 95747

Email: paul.bottorff@hpe.com