IPv6 NDP Table Exhaustion Attack

“The sky is falling,” but you can prevent it with simple configuration

Jeff S Wheeler – jsw@inconcepts.biz

Problem: Big subnets, small NDP table

- IPv6 /64 subnet is 2**64 addresses
  - 2**64 = 18,446,744,073,709,551,616
- Common layer-3 “Top-of-Rack” switch holds far fewer NDP entries
  - Juniper EX4200: ≦ 16,000
  - Cisco Nexus 5500: ≦ 6,500
- Even larger chassis switches hold relatively few
  - Many vendors are afraid to specify a value
  - Real figures typically range from 32k – 100k

Why does this matter?

- NDP entry is necessary to forward traffic at access layer, similar to IPv4 ARP entry
- Malicious DoS attack can **trivially** flood router’s NDP function, which can only resolve a finite number of host addresses per second
  - Policed to protect control-plane CPU
  - If it isn’t, you have a bigger problem
- Any host on a connected LAN can consume all space in NDP table
  - No way to store all possible entries in FIB (or DRAM)

Failure modes

- New NDP entries cannot be learned
  - Some routers break all interfaces, even if only one interface is targeted by an attack
  - Some routers break only the targeted interface
- Legitimate NDP entries are evicted from table
  - On a few routers (Juniper in particular)
- Normal operation, no affect
  - Zero routers – they are all vulnerable by design

…unless

- Don’t configure /64 subnets
  - Much like IPv4 ARP, most routers maintain a state table for IPv6 NDP resolutions in-progress
    - Often represented as *Incomplete* in CLI output
  - Resolution state effectively throttles NDP queries from the router to the targeted LAN without breaking new host learning, but only if this table is not full
  - Subnet with similar number of addresses to IPv4 subnet works just fine
    - IPv6 /120 ~ IPv4 /24

Isn’t /64 “the Standard?”

- VLSM and CIDR became “the Standard” as IPv4’s success exceeded its design basis.
- IPv6 was designed in the mid-1990s, and the Internet has evolved considerably since that time.
  - Catalyst 5500 with RSM was state-of-the-art
  - Essentially zero routers had IPv4 forwarding in ASIC
  - The current scale of DDoS attacks had not been conceived. There were no “botnets;” smurf was roughly the worst DoS attack method of that era.
  - Internet was not yet “mission critical,” “carrier grade,” etc. There was no VOIP, Netflix, or Google. Most people were afraid to use a credit card online.
  - Even junk e-mail was a relatively new concept!

What will break if I configure /120?

- SLAAC
  - Good tool, but not needed on every subnet/LAN
  - Especially not needed in datacenter network
  - Flat-out stupid on backbone links
- Anything else?
  - Detractors of this proposal have failed to demonstrate anything else breaking
    - Except devices which are broken in many other ways (such as end-user CPE)

What if I find things that break?

- Allocate /64, configure /120
  - NDP exhaustion attack should be fixed by vendors in the future
  - You can take advantage of /64 subnets when vendors make it safe and practical

Revisiting the “standard” argument

- Certain community members advocate the use of /64 subnets on every interface
  - regardless of its function
  - or the current or planned number of hosts
- There is no advantage to /64 on backbone links
- There is are several disadvantages (this is only one) to /64 on backbone links
- Yet these community members advocate /64 for these links anyway, stating that all subnets must be the same size
- These community members want to return to pre-VLSM era for no reason, and their input must be excluded (ignore them)
  - Sorry, Owen and Randy; this means you guys
  - Others do agree that /64 is not appropriate everywhere, but is useful somewhere. These people are right. I agree.

Illustration of outbound attack

- 1 PPS of traffic, with random IPv6 source address in your /64 subnet, is enough to fill your NDP table in layer-3 ToR switch (router)
  - These packets are **not** NDP packets, they can be any packets which cause router to learn a new NDP entry (virtually all packets from a previously unseen source address)
  - Hard to detect before something breaks
    - Routers lack logging (SNMP trap, syslog) upon learning new NDP entry
    - Also lack logging when NDP table is nearing or at 100% fill
- 1 PPS, really?
  - 3600 new NDP entries per hour, 16k table size
  - NDP expiration time is long, like ARP
    - Often there is no knob to adjust this aging timer
  - Threshold PPS = (table_size – normal_entries) / expiration_time
  - Nexus 5500 0.45 PPS; Juniper EX8200 6.94 PPS
  - Malicious host can send more than 1 PPS and break network in seconds (or milliseconds)
- What breaks?
  - All interfaces on the router lose the ability to learn new NDP entries (or refresh expired entries that haven't been active recently)
  - Some vendors evict NDP entries regularly, even if they have active traffic on them constantly (even upstream routers, busy servers, NAS)
    - This can break even constantly-busy, high-traffic hosts and services, IGP, BGP, etc.
    - "Some vendors" includes Juniper
  - Some vendors share IPv4 ARP and IPv6 NDP resource pool
    - **Both IPv4 and IPv6 will break on dual-stack routers, with no malicious IPv4 traffic**
- What else can happen?
  - Some foolish people have suggested using a new, random IPv6 source address for every outgoing TCP connection (or web browser page load, etc.) This supposed "privacy mechanism" (distinct from RFC3041 etc.) would unintentionally break the network

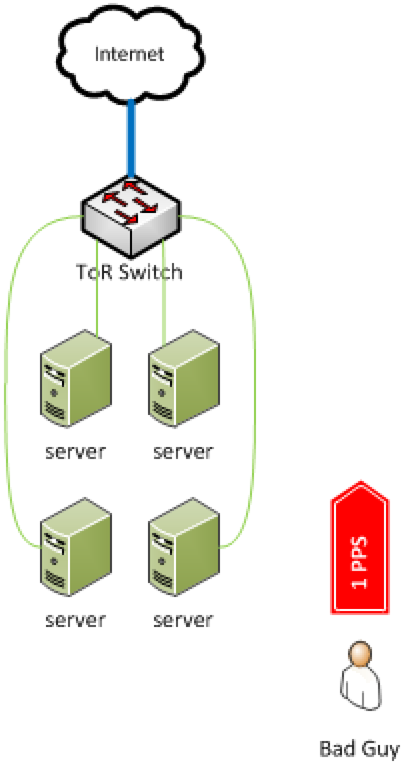
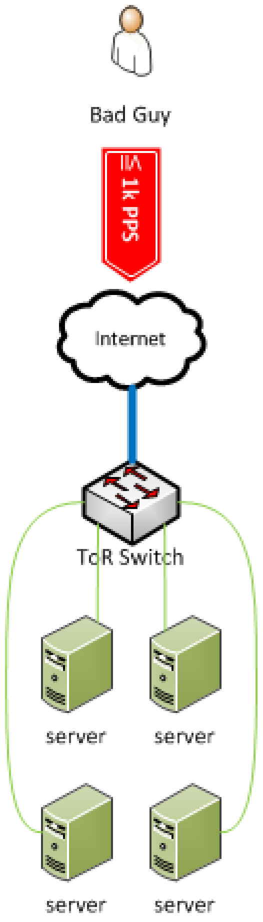


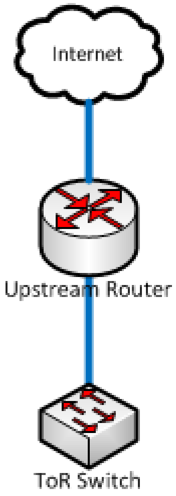
Illustration of inbound attack

- ≦ 1k PPS of ordinary packets toward random destinations within your /64
  - Looks very much like a “scan”
    - Does anyone think we won’t have network scanning attacks in IPv6?
      - Bad guys will at least look for things in the first addresses (::1 ::2 etc.) in many subnets
      - Even though they could never traverse an entire /64, they might scan the first hundred addresses of many /64s
  - Congests your NDP resolution mechanism
    - Policer will protect control-plane CPU and avoid filling LAN with large amount of NDP multicast transmissions
    - But policer will also hinder legitimate NDP resolver requests towards not-random hosts
    - ARP can (and does) have policer per destination IPv4 address
    - NDP can’t have policer per destination IPv6 address, there are 2**64 addresses in the subnet, and a table of *Incomplete* resolutions will simply fill up and churn as long as attack continues
  - Router cannot learn new NDP entries on the LAN
    - Any hosts which do not exchange traffic regularly enough to maintain NDP entry will “go missing” and will not be re-learned until attack stops
    - Routers which evict NDP entries and require an active refresh will eventually evict most or all legitimate hosts (probability-based on attack PPS, legitimate incoming PPS, resolver state table size, number of legitimate hosts, resolver time-out)
  - If a misconfigured host on the LAN responds to all NDP inquiries (promiscuous host, like “proxy arp”) it will look like an outbound attack



…it gets worse

- Traffic exchange between layer-3 ToR switch ("ToR") and Upstream Router ("Upstream") is very rare, compared to normal traffic exchange (downstream servers, end-users, malicious attack from LAN or Internet)
  - Traffic only comes from Upstream source IPv6 address for routing protocols (IGP, BGP if not sourced from loopback)
  - Traffic to next-hops may or may not hint resolution mechanism adequately to keep entry alive
  - Increases probability that NDP entry for Upstream may be aged out (for ToRs which require "refresh")
  - Decreases probability that NDP entry will be successfully re-learned (when "refresh" is required, interface flaps due to troubleshooting procedure, etc.)
  - Some platforms prioritize ARP/NDP resolution for addresses used in next-hops, improving situation
    - Except if routing protocols drop, Upstream will no longer be used as a next-hop (default route is gone, etc.) and special priority may be lost
- This risk condition does not affect all platforms
  - If we're going to tell people "the sky is falling," let's be truthful about all aspects of it



…and damage spills over to IPv4

- Some routers have a shared resource pool for IPv4 ARP and IPv6 NDP entries
  - This may be true in FIB, control-plane resolver ("*Incomplete*"), or both
- On these routers, IPv4 will also break
  - So it's not just a problem for end-users who happen to have IPv6

Static NDP is not a fix

- Increased operator maintenance for a questionable, platform-dependant benefit
- NDP table might be full when interface flaps
- If so, router may not evict an existing entry to make room (no prioritization/reservation)
  - If it did, worry about its eviction behavior for new dynamic entries!
- Not good for customer LANs
  - Customers will have to open ticket to update NDP entry!
- Not good for datacenter LANs
  - SysAdmins will have to open tickets just like customers
  - VPS cluster managers will have to configure network
    - or network must have integration with VPS migrations (some do)

How do you respond to attack?

- Attack originating from within your LAN
  - Clear ARP/NDP tables to restore functionality
  - Identify malicious host and disconnect it from LAN
    - This may be tricky if host is also churning its MAC addresses, which may be done slowly enough not to trigger port-security mechanism
  - All you can do is wait for smoke, fight the fire
    - Unless you redesign your access layer
    - or are comfortable fighting fires until vendors deliver new knobs for your routers and switches
- Attack originating from the Internet
  - Filter source address of malicious traffic (haha, right)
  - Configure static NDP entries as a stop-gap measure
    - Only works in VPS environment unless MACs move with VMs
    - Does not work if machines require ability to move IP addresses among them
      - High-availability mechanisms?
  - Hope that a table full of *Incomplete* resolutions will not impede installation of static NDP entries when interfaces flap up, otherwise they will never be installed
  - Keep up with static NDP entries until vendors deliver new knobs
    - or you get tired of this overhead, and redesign your access layer

Alternatives to /120

- SeND
  - Looks great on paper, few (zero?) implementations exist
- More knobs on routers and switches
  - Layer-3 routers need:
    - Per-interface policer for NDP requests
    - Configurable behavior when exceeding threshold
      - Check destination address against list of all IPv6 addresses “ever” seen on the LAN
      - Do not send NDP request unless address has been seen before (likely to resolve successfully)
    - Configurable limit of NDP entries per MAC address
    - Configurable limit of NDP entries per interface
    - Configurable NDP **reservation** per interface
      - Ensure that some NDP table space will always be available when interfaces flap to UP state; for example, core-facing interface
      - Router could implement as on-demand eviction to satisfy reserved entry
    - Configurable logging upon learning new NDP entry
      - To identify problems and streamline troubleshooting
  - Layer-2 switches need:
    - Configurable, per-port, long-term policer for new layer-3 source address introduction
      - Configurable aging, threshold and violate action
      - This is basically as complex to implement as per-source-IP counters on each port
    - Or longer-term, per-port policer for new layer-2 source address (MAC) introduction
      - Configurable aging, threshold, and violate action
      - Intended to match up with ARP/NDP timers in router, as opposed to CAM timers in switches

Why is no one talking about this?

- IPv6 DoS is not being observed on “Internet scale” yet
  - Single IPv4 DDoS events exceed all IPv6 inter-domain traffic (made up statistic that is doubtlessly correct; peak IPv6 traffic at AMS-IX remains substantially below 10 Gbps)
- Problem requires new features in both switches and routers to be solved for /64
  - Features which have no IPv4 analogue
- Networks can simply choose not to deploy /64
  - Many have already made this choice
- New problem introduced by IPv6 design choice: the days of IPv4 subnets larger than router ARP tables are largely a distant memory
  - NDP tables will never be able to hold 2**64 entries!

Why should you care now if there is no IPv6 DoS?

- Unless vendors deliver needed fixes before NDP attack DoS appears, you will have to **re-design your entire access layer** to defend your network
  - and re-configure all your core interfaces
  - and coordinate with all your customers
  - and rush something “IPv6 Fundamentalists” claims is non-standard, smaller subnets, into production without any testing period or time to properly adjust provisioning tools
- Some developers are already considering using a unique IPv6 source address for each outgoing TCP connection (foolish expansion on “privacy extensions”)
  - If implemented, these hosts will inadvertently DoS their own gateway by creating garbage NDP entries

Symptom of a much bigger problem

- "Standard" IPv6 deployment practices include a serious, well-known, widely-acknowledged design flaw; yet "standards" community has willfully ignored this issue for 10+ years
  - IPv6 is largely being driven by a "Fundamentalist" mindset; so-called "experts" believe the original protocol and implementation recommendations, as written in the mid-1990s, must never change
  - "Fundamentalists" treat something that does not work as if there is no room for changing the /64 "standard" to an "option"
    - End-users still can't get IPv6
    - Most SOHO CPE still has no support
    - Most call centers are still completely untrained to support IPv6
    - Some transit-free ISPs continue to have no IPv6 transit product
    - Some ISPs are still telling their users they "have no plans to support IPv6" because they "have plenty of IPv4 addresses" (they don't get it)
  - Amazingly, /120 works correctly in substantially all routers and OSes
      - Vendors understood this would be necessary for years

The same “Fundamentalists” say …

- There will never be 6to6 NAT
- BGP will be obsolete for non-ISPs in favor of IPv6’s built-in multi-homing
- Non-ISPs won’t need IPv6 RIR allocations because IPv6 renumbering is “easy” due to classful addressing
- We won’t need DHCP, because SLAAC takes its place
- etc., etc., etc.

“Standards” community is broken

- Most operators have understood this for years
- Didn’t matter in the 1990s, because real problems happened before IPv4 Internet was truly mission-critical
- Does matter now; all indications are that IPv6 is the only practical solution to IPv4 depletion
- Vendors must do what they have always done
  - Ignore standard when standard is broken
  - Give customers practical options
  - Let standards bodies catch up to real-world
- Operators must do what we have always done
  - Use available vendor knobs to ensure network function
  - Request more knobs, work-around current limitations

Comments?

- jsw@inconcepts.biz