Residential IPv6 CPE

What Not to Do and Other Observations

Mark Smith
Nextgen Networks
mark.smith@nn.com.au

September 2011

Agenda

- Context.
- Overview of generic IPv6 residential broadband service.
- Residential IPv6 CPE testing and evaluation methodology.
- IPv6 CPE issues encountered, why they may have occurred, and how they were resolved.
- Recent and near future IETF residential IPv6 CPE developments.

Context

Context

- Left Adam Internet in November 2009 to pursue long held intention of moving to Melbourne.
- Contract opportunity became available at Internode in February 2010. Worked at Internode through to May 2011.
- Joined Nextgen Networks in Melbourne in July 2011.

@Internode

- Primary focus was working on the evolution of Internode's IPv6 trial environment onto Internode's production BRAS/LNS/BNG platform.
  - Service provisioning for residential and SOHO broadband customers for IPv4 and/or IPv6, via RADIUS.
  - IPv4 and IPv6 traffic accounting reported via RADIUS.

@Internode

- Secondary focus was to work with a few residential CPE vendors, to help them with their IPv6 implementations.
  - Internode's goal was to sell residential CPE that supported IPv6, where “supported” means actually worked out of the box with their IPv6 trial/production environment.
    - Not the token, “It supports IPv6 (when we get around to implementing all the required features and fixing the bugs).”

Overview of generic IPv6 residential broadband service.

Use Minimum IPv6 Features

- Use the most minimal and likely to be commonly available IPv6 features to provision residential IPv6 services.
  - .au is a customer chooses and owns CPE market, feature sets aren't known or controlled by ISP.
  - It's early days in IPv6 residential deployment, some of the more “exotic” methods ISPs may like to use to provision IPv6 may not be widely available.
  - Use IPv6 baseline features to provide the service.

ISP Residential IPv6 Service

Assuming PPPoE/PPP broadband service, ISP provides -

- Global and individual /64 prefix for each PPP link/session.
- CPE end PPP link/session IPv6 address via Stateless Address Autoconfiguration (SLAAC), with prefix announced via IPv6 Router Advertisements (R.A. M bit off).
- Number of IPv6 /64 subnets for CPE downstream LAN interfaces, provided via DHCPv6 Prefix Delegation - /56 (256 x /64s) suggested by RFC6177, “IPv6 Address Assignment to End Sites”. (Remember, no need for address expansion via NAPT)
- Other service parameters (DNS addr(s), NTP, SIP, etc.) via DHCPv6 (R.A. O bit on).

Stable Internal Customer Network Addressing

- So what about stable addressing for customers' internal network, independent of ISP's delegated prefix?
  - Delegated prefix's lifetime could expire if ISP's outage was long enough.
- RFC1918 addresses in IPv4.
- RFC4193, “Unique Local IPv6 Unicast Addresses” - ULAs.
  - Similar to IPv4 RFC1918, but better.
  - IPv6 site-locals deprecated in 2004 (see RFC3879), as they had the same issues two overlapping and interconnected RFC1918 addressing domains had i.e. 192.168.0.1 is where?

Stable Internal Customer Network Addressing - ULAs

- ULAs come from outside global unicast address space i.e. 2000::/3
- Form is -
  - 7 bits ULA prefix – fc00::
  - 1 “L” bit indicating Global ID locally or globally assigned
  - 40 bits Global Identifier
  - 16 bits Subnet IDs
  - 64 bits Interface IDs (a.k.a. host addresses)

Stable Internal Customer Network Addressing – ULAs (cont.)

- Central Global IDs not specified yet, so **fd00::/8** is current ULA prefix, indicating local Global ID assignment.
- Global ID needs to be as globally unique as possible to avoid chance of address space collisions if and when two or more ULA addressing domains (networks) interconnect.
  - Big (pseudo-)random 40 bit number after fd00::/8, making a /48. Formula suggested in RFC4193.
- Not to be leaked globally (hard to aggregate on purpose!).
- Use in **parallel** with global IPv6 prefix(es).

Residential IPv6 CPE testing and evaluation methodology.

Worlds largest pile of IPv6 CPE?
(My desk @ Internode, Feb 2011)

Testing and Evaluation Methodology

- Initially, testing and evaluation wasn't expected to be as involved as it ended up being. The initial expectation was mainly to “have a bit of a look” at the residential CPE IPv6 implementation from one vendor. Consequently it was somewhat informal.
- Once they realised they would get free help with their implementations, most vendors we dealt with were very responsive.
  - One vendor in particular was sending me new firmware to look at 24 to 48 hours after I reported issues.
- In hindsight, it would have been better to come up with a more formal testing and evaluation plan, including documented test plans.

3 Test and Evaluation Phases

- Testing and evaluation took place in 3 broad phases -
  1. Test, evaluate and provide feedback to the vendor until it successfully connects to the IPv6 service, and can access the IPv6 Internet.
  2. Evaluate and provide feedback to the vendor regarding RFC compliance for the various functions.
  3. Test, evaluate and provide feedback on recovery from common failure modes e.g. ADSL link failure.
- Not strict boundaries though, e.g. evaluation of RFC compliance also took place during steps 1 and 3.

Importance of Phase 2 – RFC Compliance

- Non-RFC-compliant devices can sometimes work due to the Robustness Principle -
  - “"Be liberal in what you accept, and conservative in what you send" (RFC1122).
- One device misbehaving on the network might be appear to be tolerable.
- The problem comes about when there might be 100s or 1000s of devices misbehaving at the same time, possibly triggered by an external event e.g. power outage at an exchange.
- There seems to be a common enough mentality that “if it's working (I can access the Internet), it must be ok.” On the testing and evaluation list, that is stopping at phase 1.

Importance of Phase 2 – RFC Compliance (cont.)

- RFC compliance should ensure the device not just works, but works well.
  - Interoperates with other implementations.
  - Tolerant of other devices' misbehaviour
  - Will self correct itself if impacted by a fault, when the cause of the fault is removed.

Importance of Phase 3 – Recovery from common failures

- This phase may also be missed if the approach is “it works so it must be ok”. People performing this sort of “testing” are testing in the best case scenario, and ignoring the common worst case scenarios.
- An example. An ADSL router “tested” by non-operational people didn't bring PPP back up after an ADSL drop out. Not too much of an issue for a normal ADSL customer, although they'll get annoyed at having to manually reboot their router occasionally. A bit more serious for a Naked DSL customer wanting to receive VoIP calls …

IPv6 CPE issues encountered, why they may have occurred, and how they were resolved.

Big Picture

- IPv6 is not just IPv4 with bigger addresses.
  - Don't need to expand address space via NAT methods.
  - Addresses/prefixes have lifetimes (preferred/valid), and they may age out.
    - To assist with phasing in and phasing out address space a.k.a. renumbering.
  - Designed to support multiple addresses on an interface at once.
    - IPv4 supported this “by accident”, not by design.

Big Picture (cont.)

- As a result of IPv6 being more than just IPv4 with bigger addresses, CPE vendors need to change their mental models of how CPE operates.
  - TCP and UDP are not the only transport protocols (e.g. SCTP, DCCP). IPv4 NAT prevented them from being widely used.
  - NAT traversal techniques shouldn't need to be used to support peer-to-peer architecture applications.
  - ISPs may phase in new addressing for customers and phase out old addressing over time (e.g. months), rather than using service disconnect/reconnect to facilitate renumbering.

Unsolicited RAs sent too often.

- IPv6 RAs are used to convey a number of network layer parameters, including link MTU, link assigned prefixes, whether to use stateful or stateless addressing, how to acquire other non-network layer host parameters, and if the RA announcer is a default router.
- Unsolicited RAs are periodically multicast to all nodes to reliably refresh information that ages out.
- RFC4861 Unsolicited RA announcement default periodic interval should vary between 3 and 10 minutes.
- One implementation was sending Unsolicited RAs every 5 seconds.
  - Reduces battery life of wireless mobile devices (laptops, smart phones) as host CPU is interrupted to process them.

DHCPv6-PD prefix lifetimes not decremented on LAN interface.

- The RA announced prefixes need to have their preferred and valid lifetimes decremented at the same rate and in parallel with the DHCPv6-PD supplied prefix.
  - i.e., the preferred and valid lifetime values in the RA Prefix Information Option are reduced upon each announcement
- This ensures that the LAN /64 prefixes expire at approximately the same time as the DHCPv6-PD supplied aggregate prefix.

DHCPv6-PD prefix lifetimes not decremented on LAN interface. (cont.)

- Multiple implementations were not decrementing the RA prefix lifetimes. DHCPv6-PD RFC, RFC3633, is a bit brief and vague on this. RFC6204 makes it clear.
- One implementation was completely ignoring the DHCPv6-PD lifetimes, and setting the values in all RAs to a valid lifetime of 7200 seconds, and a preferred lifetime of 3600 seconds, the RFC4862 minimums.
- ISPs need to be able to rely on customer LAN RA prefixes expiring at approximately same time as DHCPv6-PD supplied aggregate prefix so they can manage a phased renumbering process.

Set the Cur Hop Limit value in RAs to 255

- RAs can convey the initial IPv6 unicast packet hop count used by hosts.
  - The current default is 64 hops, which is enough for the current Internet's diameter.
  - If that needed to be changed in the future, it can be updated via RAs, rather than by patches or manual configuration changes to individual hosts.
- One CPE implementation was setting this to 255, the maximum. This would cause packets to loop more than necessary if they encountered a forwarding loop.
- There may be 1000s of these CPE, and therefore 1000s of hosts with bad initial hop counts values because of this bug.

Set the Cur Hop Limit value in RAs to 255 (cont.)

- Likely to be a misinterpretation of the “ICMPv6 255 hop count” trick.
- Some ICMPv6 messages, such as RAs, must not be forwarded, or if they accidently are, must not be accepted.
- Each time an IPv6 packet is forwarded, it's Hop Count must be decremented.
- A received Hop Count value of 255 ensures that the packet has not been forwarded.
- An RA's outer IPv6 Hop Count is set to 255, but the Cur Hop Limit value inside the RA should be set to either 0 (use the host's stack default) or 64.

ULAs with random part of all zeros

- One implementation would announce ULAs with an all zero random part, when IPv6 on the WAN interface went down i.e. would attempt to “swap” ULAs for globals, rather than make ULAs constant and independent of WAN IPv6 delegated prefix.
- Effectively makes ULAs the deprecated IPv6 site-locals.
- Will suffer from all the issues identified in RFC3879, “Deprecating Site Local Addresses.”
- In particular, two “all zero” ULA domains that interconnect will need to have one renumber to avoid addressing collisions.
- Don't understand how this could happen, RFC4193 is very clear about the purpose and creation of the random ID part.

CPE Firewall rules only allow specifying inbound TCP/UDP

- Other possibly useful transport layer protocols exist -
  - SCTP – Stream Control Transport Protocol
    - Connection oriented like TCP, multi-homed, channel and message based. First RFC published in 2000.
  - DCCP – Datagram Congestion Control Protocol
    - Basically a congestion controlled version of UDP. First RFC published in 2006.
- IPv4 NATs have constrained their deployment as they can't translate the addresses in their headers/payloads.

CPE Firewall rules only allow specifying inbound TCP/UDP (cont.)

- One IPv6 CPE implementation only allowed specifying inbound TCP/UDP ports, others only added ICPMv6 to this list.
- SCTP and DCCP should be added to this list if possible i.e. SCTP or DCCP port specifiers.
- Being able to specify a permitted transport protocol number would allow for all future transport layer protocols.
- Being able to permit all transport protocol numbers for people who are happy to rely on their hosts' firewall.
  - IOW, firewall only protects the CPE (host based firewall for CPE itself), but forwards all IPv6 traffic unfiltered, acting as a pure IPv6 router.

No “Stateless” DHCPv6 Server on CPE LAN Interface

- “Stateless” DHCPv6 is used to provide “other”, non-network parameters to hosts, such as IPv6 DNS server address(es), NTP server addresses and SIP server addresses.
  - “Stateful” DHCPv6 combines this function with database-driven host address assignment. Hosts are informed of the choice to use Stateful DHCPv6 by setting the M bit in RAs.
  - Stateless DHCPv6 is a subset of “Stateful” DHCPv6, using a simple two packet Information Request/Reply transaction. Hosts are informed of the choice to use Stateless DHCPv6 to acquire “other” parameters by setting the O bit in RAs, leaving the M bit off.
  - Both Stateful and Stateless DHCPv6 described in RFC3315, RFC3736 provides more specific Stateless DHCPv6 implementation advice.

No “Stateless” DHCPv6 Server on CPE LAN Interface (cont.)

- At a minimum, a CPE should facilitate IPv6 address assignment via RA based Stateless Address Autoconfiguration (SLAAC) and “other” parameter assignment via stateless DHCPv6 on it's LAN interfaces.
  - LAN interface Stateless DHCPv6 parameters may have been acquired from ISP during Stateful DHCPv6-PD transaction on WAN interface and/or configured via CPE web interface.
- Lack of a Stateless (or Stateful) DHCPv6 server on the LAN interface means “other” parameters have to be configured on hosts manually, a far less user friendly experience than under IPv4 (and that wasn't really all that user friendly anyway).

No ICMPv6 Destination Unreachables when WAN Interface down.

- Dual stack hosts with native IPv4 and native IPv6 addresses will prefer IPv6 if a DNS lookup returns both AAAA (IPv6) and A (IPv4) records.
- Current hosts don't treat ULA addresses as special, and therefore consider they have IPv6 Internet connectivity if they have ULAs (or any other IPv6 address).
- Consequently they will prefer IPv6 over IPv4, even if the WAN link doesn't support IPv6 (i.e. an IPv4 only PPP link).

No ICMPv6 Destination Unreachables when WAN Interface down (cont.)

- If the host attempts to connect to an remote AAAA (IPv6) destination that is unreachable, the CPE should issue an ICMPv6 Destination Unreachable back to the host.
- The host will then immediately resort to attempting to access the A (IPv4) destination.
- If the CPE does not issue an ICMPv6 Destination Unreachable, the host will have to time out it's IPv6 attempts before resorting to IPv4.
- IPv6 timing out depends on the host's IPv6 implementation, however it can be in the order of 30 to 180 seconds, which is a very, very bad IPv6 user experience compared to the IPv4 one.

No ICMPv6 Destination Unreachables when WAN Interface down (cont.)

- Alternative and additional approach to this problem is the so called “happy eyeballs” method.
- Details still being developed, however broadly, application or host stack attempts to connect to to both AAAA (IPv6) and A (IPv4) roughly in parallel.
- Which ever of IPv6 or IPv4 connects first is continued, the other is abandoned.
- Over time, host or application builds up a cache of connection attempt success, and then subsequently only uses the protocol that succeeded.
- http://tools.ietf.org/html/draft-ietf-v6ops-happy-eyeballs-03

Ignored M bit state in received RAs on the WAN interface.

- Stateful (via DHCPv6) or Stateless Addressing (SLAAC via RAs) are addressing options for the WAN link.
- Internode chose SLAAC as implementations of Stateful DHCPv6 clients in CPE were and generally still are rare.
  - RADIUS used to provide or acquire details of the prefix assigned to customers. A stateful DHCPv6 address database was not necessary.
- Internode's choice was expressed by leaving the M or Managed Addressing bit switched off in the RAs sent to customers' CPE.
- One CPE ignored the status of the RA M bit, and attempted to acquire an address via Stateful DHCPv6, causing IPv6 address assignment to fail.

Sent RA MTU option with 1460 value.

- RA MTU option can be used to lower the MTU used by hosts if the default link MTU is too large.
- An example is CPE with a 1492 PPP/PPPoE WAN MTU could announce LAN RAs with MTU option set to 1492, avoiding “dumbbell (<big><small><big>) MTU” PMTU issues when accessing Internet.
- Drawback is that the hosts use the MTU for all traffic, including between themselves on the LAN, potentially reducing LAN throughput.
- One CPE was announcing an RA MTU option with a value of 1460, regardless of the PPP/PPPoE link MTU of 1492.

Firmware changes not sent upstream.

- One vendor did a good job at addressing a number of issues with their implementation.
- They had passed phases 1 and 2, and had a few changes left to pass phase 3.
- At this point they decided to stop working on their firmware, as their upstream OEM was going to release a major new release with “IPv6 support”.
- After a delay of approximately 4 months, the new upstream firmware was released, and had none of the IPv6 changes that the vendor had spent time and effort making.

IETF Developments

- RFC6204 - “Basic Requirements for IPv6 Customer Edge Routers”, April 2011.
  - IPv6 requirements for basic scenario of single CPE with a WAN link and directly attached LAN links (e.g. Ethernet and wifi).
- “CPE bis” - “Advanced Requirements for IPv6 Customer Edge Routers”
  - Additional requirements to suit QoS, multicast, DNS, routed network in the home, transition technologies

IETF Developments (cont.)

- RFC6092 - “Recommended Simple Security Capabilities in Customer Premises Equipment (CPE) for Providing Residential IPv6 Internet Service.”
  - Advice on IPv6 security functions that can be implemented in CPE.

Questions?

Thanks!

Thanks for listening!

Thanks to Internode for letting me do this presentation.

Thanks to Nextgen for sending me up here to present it.