Disaggregating CGNAT

In-line CGNAT and multi-service edge routing on the same open switch

An Introduction to NAT

Motivation

IPv4 addresses, by virtue of being 32 bits long each, are not available in large enough numbers to meet all of today's requirements for Internet connected devices. This problem was recognized early on and a private address space (RFC 1918) consisting of prefixes 10/8, 172.16/12 and 192.168/16 was carved out of the IPv4 address space to allow for reuse within private domains, such as Autonomous Systems (ASes), and can be freely used within those domains. The corollary is that these addresses have a restricted scope and cannot be used to access services across the Internet. There are a few other spaces carved out additionally, but most of the remaining addresses are used for public addressing. Addresses from this public address space have been traditionally allocated by the Regional Internet Registries (RIRs) in blocks of 24-bit or lesser prefixes to organizations on a first-come first-serve basis and the last of these prefixes had been allocated by the end of 2019. The prefixes allocated early were in chunks of /8, /16 and /24 to universities, research and governmental organizations with much lower eventual requirements. As a consequence, there is an uneven and skewed distribution of prefixes, leading to a scarcity of IPv4 addresses.

However, since the demand for these addresses is strong marketplaces have sprung up where a single public IPv4 address have traded in the $40 - $60 range since 2021. The insatiable customer need for public address coupled with high pricing and the fact that Communication Service Providers (CSPs) have subscribers, each of who may need access to a public address to access Internet services, running into thousands depending upon the size of the provider, has made the economics of providing such large number of public IP addresses unviable.

To address this need, a technique called Network Address Translation (NAT) has been used to minimize the need for public IPv4 addresses by reusing them across multiple subscribers.

It should be noted that IPv6 addresses, which are widely available, avoid the pitfalls associated with IPv4 addresses and specifically do away with the NAT requirement completely. However, the majority of Internet traffic is still on IPv4 and unlikely to migrate to IPv6 in the foreseeable future, necessitating NAT implementation in topologies. The subsequent discussion is focused on IPv4 which is subject to NAT translation and unless explicitly stated otherwise, address refers to IPv4 address.

NAT Mechanism

Let us delve deeper into how NAT enables conserving public addresses. To understand this, let us first understand how a subscriber accessing services, such as visiting a website on the Internet, works.

- Subscriber communication takes place over UDP or TCP, two transport protocols that help establish end-to-end sessions between the subscriber's address (source address) and the web server's address (destination address).
- When a website which runs http(s) is accessed, a TCP session is established to the well-known application TCP port on the destination address using a TCP port on the source address allocated by the subscriber's network stack.
- A session is thus identified by the source address, source port and a destination address and destination port. Once a session is established, the content is downloaded from the server over the session and rendered in the browser.
- It is possible for a source to establish multiple sessions with a destination as long as it uses a different source port for each session. Certain other applications use UDP in much the same way using UDP ports. The ports, both TCP and UDP, are 16-bit identifiers of which the first 1024 are well known and reserved for specific applications and the other ports are used to establish connections to the well-known ports.
- Once the relevant content is downloaded, the TCP session has served its purpose and it is torn down. The source port is reserved on the subscriber's network stack for the duration of the session and cannot be reused during this time.
- Once a session is torn down, the source port is reused to establish other sessions.

It needs to be noted that even though a single address has 216, i.e. 65536 source ports at its disposal, a given subscriber typically uses no more than a few 100 ports (sessions) at a time with the other ports remaining unused during this period. NAT uses this property to enable the sharing of a single public IPv4 address among multiple subscribers. Please note that processes such as DNS, TLS, etc. are also involved in the communication, however they are not relevant and do not interfere with NAT operation and hence have been glossed over in this discussion.

In a NAT enabled access network, a subscriber is assigned an address from the private address space. As before, the subscriber uses this address to establish a session with the web server. The Broadband Network Gateway (BNG)/NAT device which lies between the subscriber and the web server detects session establishment and assigns an available port from a public address it owns and records the mapping between the private address/port and public address/port. The device modifies the packet to replace the private source address and port with the public source address and port. Although very often referred to as NAT, this technique is actually Network Address Port Translation (NAPT).

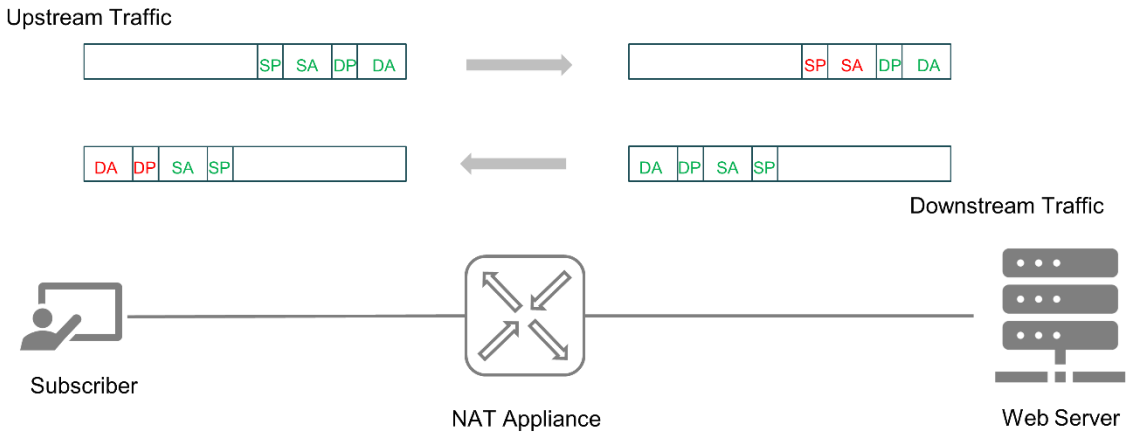

Figure 1: NAPT in action

The session can be now established with the web server. When the web server responds to the subscriber using this public address/port combination the packet reaches the BNG which hosts and has advertised the public address route. The BNG which has recorded the corresponding private address and port from the upstream interaction modifies the destination address and port to the private address/port before forwarding the traffic towards the subscriber. The BNG thus acts as a transparent intermediary allowing a subscriber to connect to the web server using a private address. Apparently it looks like we have added an extra step of address translation without much benefit as the subscriber ultimately utilizes a public address/port until we realize that the BNG services multiple subscribers. The BNG can efficiently allocate different ports belonging to the same public address across multiple subscribers thereby significantly improving the utilization of the public address and reducing the need for public addresses many times over. This technique is called NAT44 since it involves translating from a private IPv4 address to a public IPv4 address.

We omitted certain details from the description to keep things simple. Let us now address those details to outline the anatomy of the service in practice.

- Typically, a subscriber connects to a CSPs network through a Customer Premises Equipment (CPE) which is a small router, and the CPE gets an IP address assigned from the CSP.
- The CPE then assigns addresses from the private address space to hosts that exist behind it and does NAT44 translation using the Provider assigned address.
- Since Private address space is already claimed by the CPE, it follows that the address assigned to the CPE has to be from a different address space. The IETF has reserved the 100.64/10 address space for this purpose - i.e. to be used by providers to assign addresses to the CPEs.
- Thus, the CPE performs the first NAT44 at its end on the transit traffic to 100.64/10 space and the BNG does the subsequent NAT44 translation to public address space before forwarding the traffic towards the Internet and vice versa.

This solution is referred to as Carrier Grade NAT (CGNAT) or NAT444 since there are two IPv4 address translations taking place between the host and the Internet service. Considering the large number of subscribers that it supports, a typical BNG handles NAT translations running into millions of TCP/UDP sessions at a time.

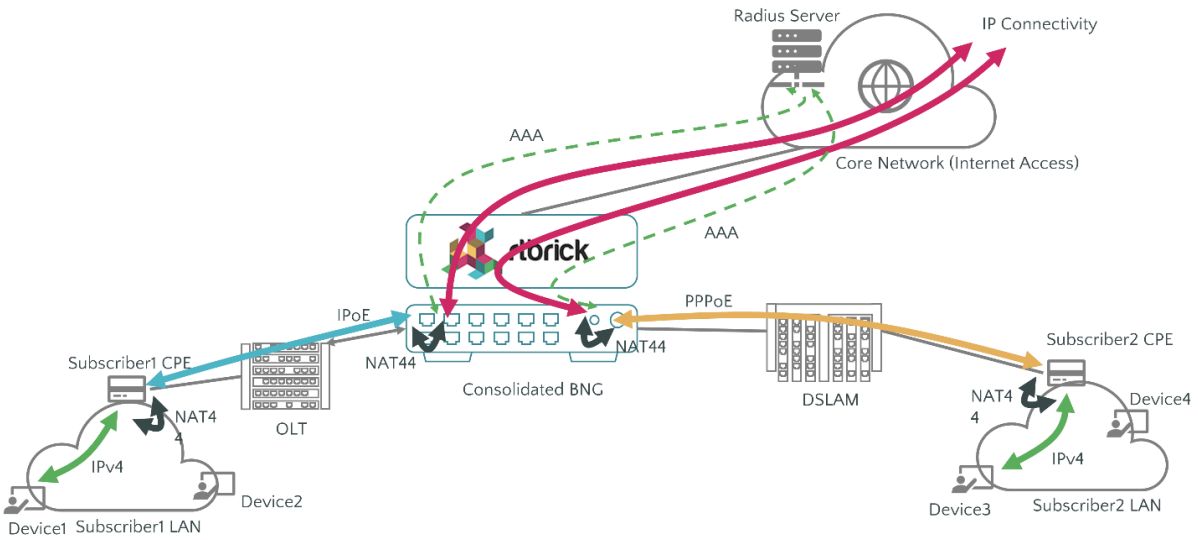

Figure 2: CGNAT on RBFS

RtBrick's CGNAT Solution

RtBrick's Network Operating System, RtBrick Full Stack (RBFS) is a composable, cloud-native Layer 3 Network Operating System (NOS) that is packaged as a container with a runtime environment and is enabled on Broadcom DNX Qumran2 bare metal switches to address various large scale Layer 3 Routing use-cases.

This software is available as a Consolidated BNG (C-BNG) package. CBNG is a MultiService Edge Router that provides a combined BNG as well as a Carrier Grade IP/MPLS Router on a single device. A switch with this image can provide subscriber termination and rich set of triple play services such as IPoE, PPPoE, AAA, L2TP, L2BSA, Lawful Interception, Hierarchical QoS, Subscriber ACLs, Walled Garden, High Availability with Hot Standby, IPTV services, as well as routing using BGP, IGPs and MPLS. The solution is also optimized for AI-Ops with highly scalable streaming telemetry support.

Depending upon the hardware selected, and the number of queues allocated per subscriber, the subscribers can be scaled to tens of thousands while hosting a full Internet Routing Table. A unique characteristic of RtBrick's CGNAT solution is that NAT is a Software-Defined network function and implemented in the chipset with Broadcom's SDK to deliver CGNAT functionality in-line **(fully in the same packet processing pipeline as the other functions)** in the data plane without any additional chipset resources.

This solution is currently available on Ufispace S9600-72XC and the Edgecore AGR420 (AS7946-74XKSB) open switches.

Benefits derived from Implementation in Hardware

Chipset Capability enabling Full Throughput

RBFS solution is implemented using the Qumran2 NPUs thereby allowing it to scale to a high subscriber count and throughput while retaining the ability to handle multiple network events and significant route churn.

A unique feature of RBFS CGNAT is that it follows the above template and has been seamlessly implemented in the NPU thereby allowing a CSP to maintain all the speeds and feeds of the switch along with benefits of NAT flow-based service. The solution can NAT the complete throughput, which is 2.4Tbps on today's chipsets. Additionally, the switch houses an external TCAM called the OP2(NLA 16K) co-processor which allows the solution to support a massive 4.5+ Million NAT entries. In the market there are multiple appliances or service line cards

that provide CGNAT functionality that can often scale to multi-million NAT entries, but they are often based on the x86 based forwarding plane which restricts them to supporting tens of Gbps to no more than low hundreds of Gbps of throughput.

The RBFS solution differentiates itself by being able to support NAT at the entire throughput of the switch (an order of magnitude greater than any of the current solutions), enabled via a software defined container software completely in-line without impact to any of the switch functions.

Investment Protection

RBFS enables complete plug-n-play capability of all the software features which in turn allows us to package the software in a 'lego block' function to work either as an IP/MPLS router, BNG access switch and a CGNAT device or combined form-factor as a Multiservice Edge Router. Since the data plane chipset resources are finite, each additional capability brings in different scaling considerations, however the investment of a CSP in the bare-metal switch is completely protected as one or more functions are enabled using RBFS.

Miniaturization and Lower Power Consumption

The hardware to deploy CGNAT are 2-RU switches. These already use highly power-efficient chips and consume as little as 0.15W/Gbps. Competing CGNAT solutions are available either as standalone 1-2 RU appliances that have a low throughput and high power consumption or are available as service line cards on large chassis-based routers. All of this results in additional cost to operate CGNAT functions. There is no other competing solution, to the best of our knowledge, that houses multiple in-line network functions on 1 to2 RU switches at such a high scale and throughput.

Summarizing the numbers shared in a table:

Parameter	Throughput	1-Q Subscribers	4-Q Subscribers	NAT entries	Routes (IPv4)	Routes (IPv6)
Ufispace S9600-72XC	2.4 Tbps	48K*	28K*	4.5M	1M	300K
Edgecore AGR420	2.4 Tbps	48K*	28K*	4.5M	1M	300K

**Please note that either 1-Q subscribers or 4-Q subscribers can be deployed but not both together*

Rich Policy support

Rich NAT Policy Options

CSPs have separate service plans that cater to different price/value segments. RtBrick recognizes this CSP requirement to treat subscribers differently depending upon the plan they subscribe to. To enable this, the C-BNG config data model provides the ability to create different subscriber profiles each of which cater to service in accordance to its service plan. Once a subscriber is identified as subscribed to a particular plan, either through AAA authentication or local configuration, the services defined in the specific profile are instantiated for the subscriber within the BNG. This functionality has been extended to CGNAT as well, and RBFS provides a rich tapestry of options to support such service plans. Let us explore the broader Policy features provided by RBFS.

Public Address Allocation and Port Pool Assignment

Public addresses are allocated in pools to the BNG to be used for CGNAT. A pool is a set of contiguous addresses defined by a starting and an ending address. UDP and TCP port blocks from each address are then allocated to subscribers. Each pool can be configured with a port block size starting from 64 in powers of 2 up to 2048 entries. Therefore it is possible to configure separate pools for different profiles to enable subscribers to receive NAT entries as per their plan.

Furthermore, it is possible to chain pools to one another so that once a pool gets exhausted, the next pool in the chain can be used for allocating port blocks. This enables assignment of public addresses to the BNG device on the go. In addition to this, subscribers can request for multiple port blocks to be assigned to them, subject to a maximum configured limit.

These allocation policies help ensure that public addresses can be very efficiently utilized while enabling specific treatment to subscribers on that plan.

Deterministic NAT

It is possible that subscribers can get port blocks from different addresses, although the allocation algorithm minimizes the chance of this happening. However, CSPs may not want to allocate port blocks from more than one public address to a subscriber. This is called deterministic NAT and is available in RBFS. The flip-side of enabling this is that a subscriber may not receive an additional block, despite not reaching its max limit, if the blocks from the public IP address get exhausted. The implementation places this trade-off decision in the hands of the CSP.

Port Recirculation

Ports from an allocated port block used for TCP/UDP sessions are reused once the sessions are torn down. A configurable idle timeout for both TCP and UDP sessions is used so that timed-out entries can be reused for other sessions. While such timeout mechanisms provide reuse of port entries, this can be made more efficient under certain circumstances. It is often seen that most TCP/UDP sessions are short-lived and while a timeout will eventually free the corresponding port entries for reuse, it is not necessarily the most efficient way of doing this. For TCP sessions, the software tracks FIN requests sent from either end and when it detects that such a request has been issued, it frees up the port sooner than taking the timeout approach. This results in more efficient use of the port entries and hence fewer port blocks would need to be allocated to subscribers than otherwise.

High Availability with Hot Standby

To provide high customer experience (CX) to subscribers and reduce opex through reducing truck-rolls, RBFS can be deployed in a High Availability configuration where the OLT is connected to two different BNG devices. If a network event occurs, such as a primary BNG failure or the link failure to primary BNG, the subscriber sessions are automatically switched over to the standby BNG with the subscribers hardly noticing the impact. More importantly, urgent and expensive truck-rolls can be avoided thereby reducing expenditure. CGNAT works seamlessly with this feature.

Conclusions

NAT is a table stakes feature for CSPs deploying broadband Internet services. CSPs typically implement BNG and CGNAT as separate functions on different network elements. This requires higher CAPEX in terms of upfront cost and recurring OPEX in terms of power, cooling and space requirements, not to mention the added operational complexity. Commercially available solutions impose restrictions on bandwidth, lack of uniformity in user experience as they are dependent on software, and not as power efficient. Many established vendors provide CGNAT functionality on service line-cards that can only be slotted into large chassis-based routers. RtBrick has provided a solution on merchant silicon that addresses these shortcomings to provide a highly scalable, open and flexible CGNAT solution.