Designing a Mikrotik CGNAT solution that scales

He is crazy right?

Presented by: Matthew Enger

Agenda

1. Who are we?
2. What are we trying to solve?
3. Let's make it happen
4. CGNAT static port allocation setup
5. Testing
6. Conclusion

Who are we?

Matthew Enger

Network Engineer – Self Taught

Owner – Leaptel

Chair – Internet Association of Australia

Decided it was a good idea to start an ISP and Leaptel was originally using Mikrotik for everything. Nowadays we use Mikrotik for specific applications

Email: m.enger@team.leaptel.com.au

Who is Leaptel?

- Small(ish) (~70k subscribers) retail service provider
- Privately owned
- Network throughput over 300G nationally aggregated daily
- NBN, Opticomm, Redtrain, Lyman Networks, ASN and more
- Focusing on Australian Based Support, good quality experience
- International routing optimised for low latency (and we are happy to look at issues if you find any)

Leaptel Wholesale!

Recently launched the Leaptel Wholesale brand

Offering all the good things:

- Transit
- Inter-capital
- Layer 2 and 3 NBN TC4
- Enterprise Fibre (EE, Telstra, Fibreconex, Vocus)
- Voice

Built on top of the Leaptel reliable network

Come talk to us if you are interested!

What are we trying to solve?

What are we trying to solve?

- Provide CGNAT solutions for broadband subscribers
- Redundancy
- Scale
- Cost effective
- Law enforcement logging
- Take our learnings from commercial offerings and improve our current design

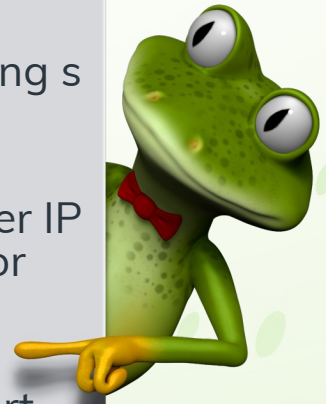
Law enforcement logging

NAT rule with logging turned on:

- Every new connection generates a log event
- Parse logs
- Store lots of data
- Need to parse data quickly for law enforcement request (hard)

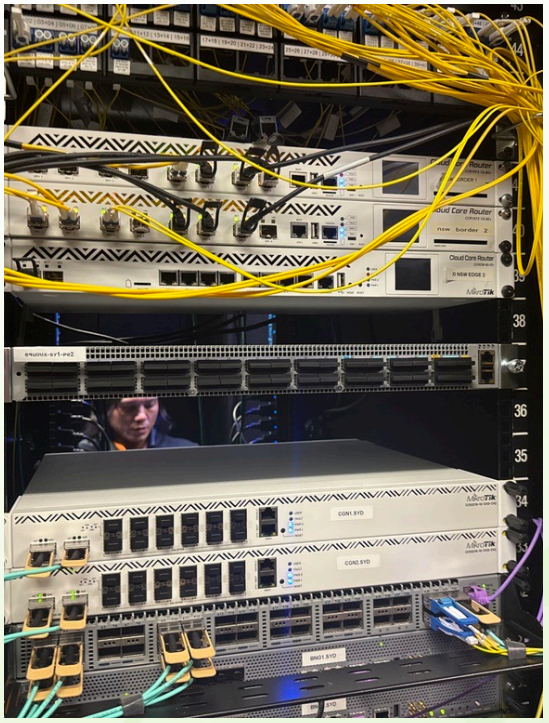
NAT with fixed port allocations:

- We setup static port mapping s of 250 ports mapping to a single CGNAT customer IP
- We log the CGNAT customer IP on radius (just like you do for non-CGNAT)
- When a law enforcement request comes in, we convert the public IP and port to the CGNAT IP
- Very low record keeping volume

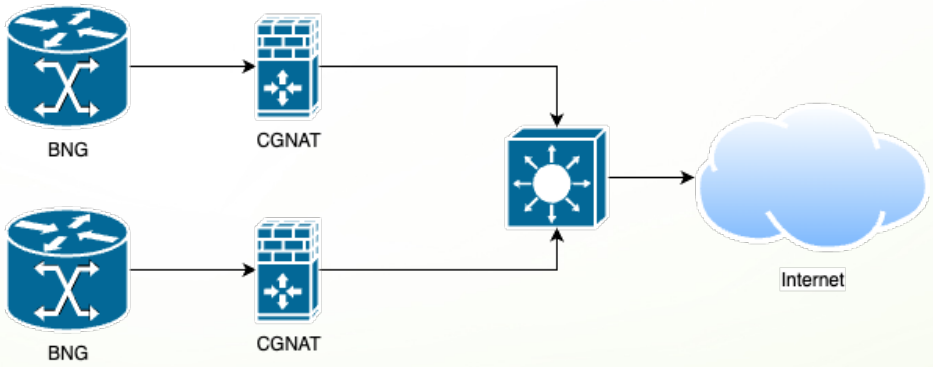


What about Leaptel?

- We are using this solution currently (August 2025)
- We use fixed port allocation (250 ports)
- For our larger sites on the east coast, we are moving to commercial solution
- We are using the hardware removed from these sites to bolster and expand our other sites.



Leaptel – Current Design



- Each CGNAT box is tied to a BNG (direct connect)
- CGNAT fails you have to really work to move the traffic (hard)

Introducing Mikrotik CCR2216-1G-12XS-2XQ

2 x 100G QSFP28 ports

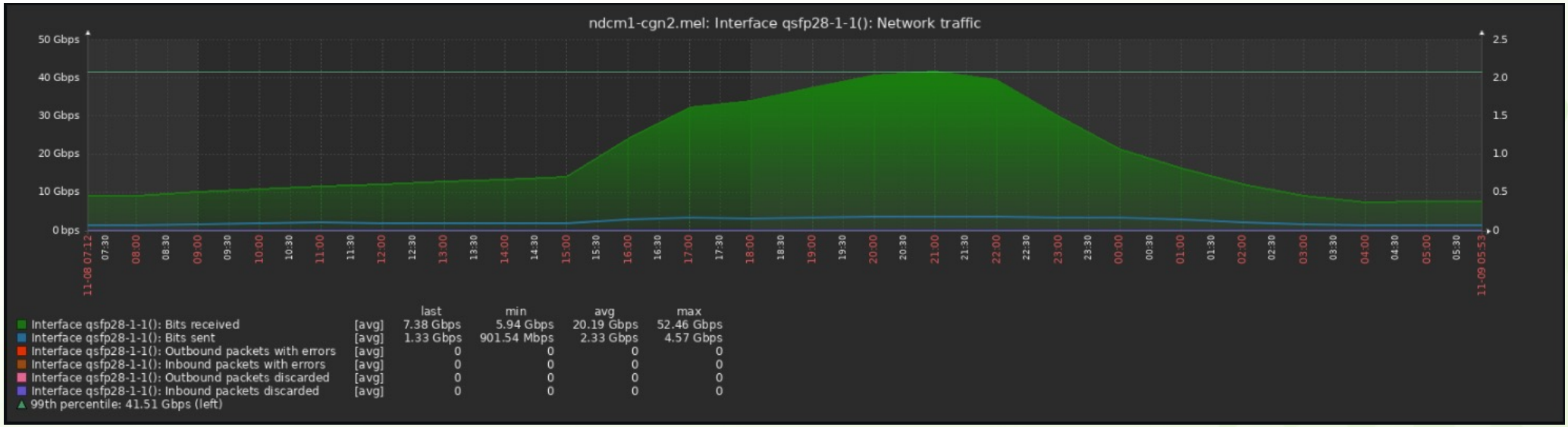
12 x 25G SFP28 ports

16 x ARM AL72400 CPU @ 2000 Mhz

16G RAM

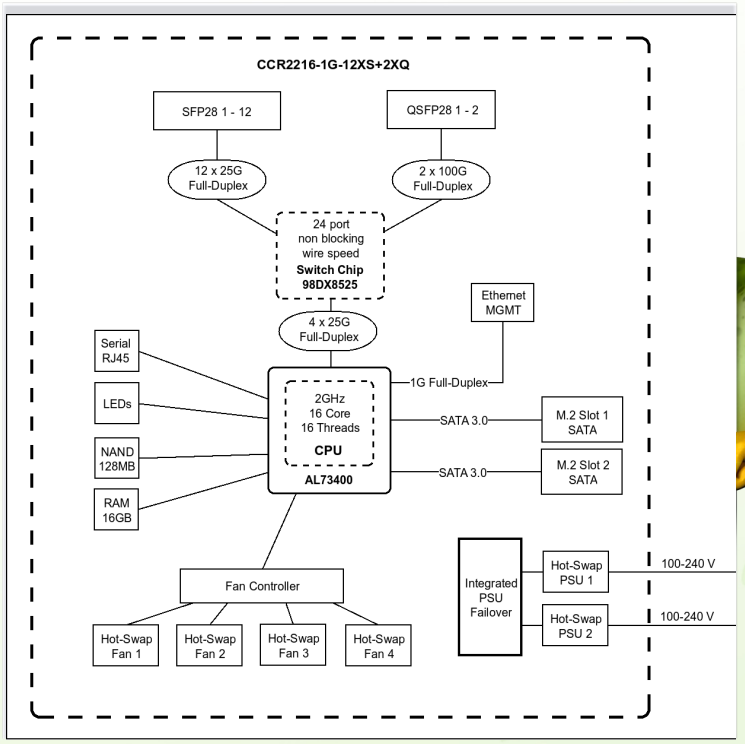
What can it do?

~42G ******* conditions apply!!!!

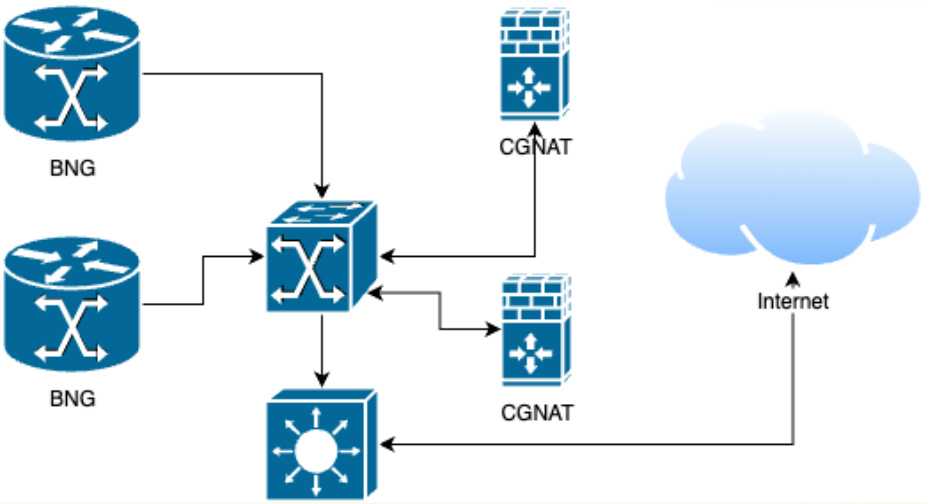


Limiting factors

- CPU (~75% on some cores)
- Connection tracking table size – max entries is 1,048,576
- Does not support dynamic port block allocation
- 4 x 25G paths to CPU



Leaptel – Future Design



- Removing 1 to 1 relationship between BNG and CGNAT -
- Introducing failover for if (when) a CGNAT appliance fails

Leaptel

Let's make this happen!

Technical Specs

- VRRP towards BNG with one CGNAT appliance ready and operational and the other ready to cut over
- BGP towards layer 3 switch with AS-Prepends to handle primary and secondary
- Static allocations, 200 ports per IP using script found at https://wiki.mikrotik.com/Manual:IP/Firewall/NAT#Carrier-Grade_NAT_(CGNAT)_or_NAT444
- Limitation: design assumes that either CGNAT box is online or offline. If it loses an interface this design won't fail over

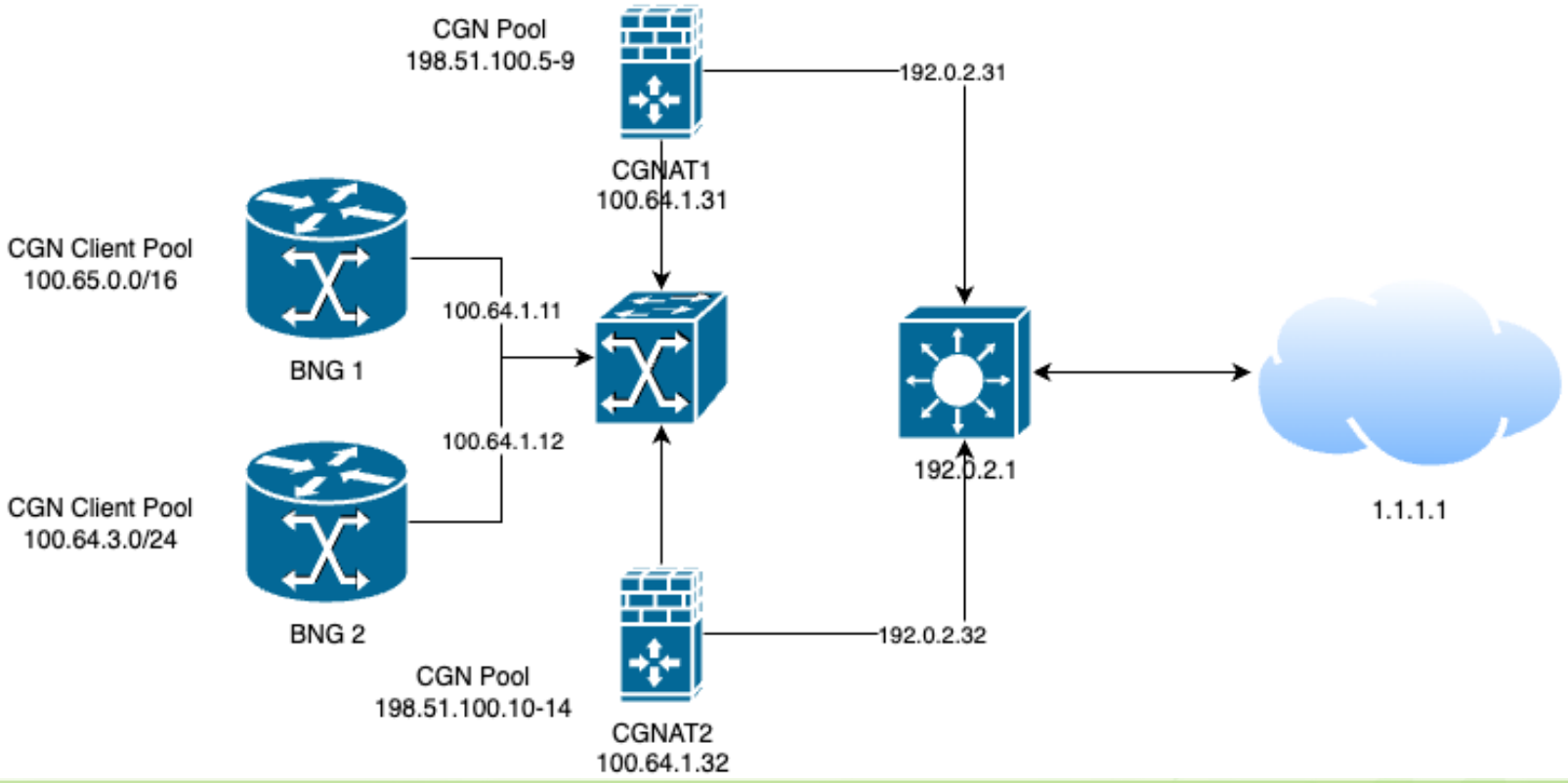
My Test Rig

Using spare parts I made this:

- Top Router is a pretend BNG
- Middle two Mikrotiks are for CGNAT
- Bottom Arista used as layer 3 switch to test public side
- Connected to each other by 10G links

Upgraded all Mikrotiks to 7.19.3

Test Rig Deployment



Configure the CGNAT link network

CGN1:

```
/interface/vlan
add interface=sfp-sfpplus1 vlan-id=20
name=internet

/interface/vrrp
add interface=cgnat-link name=cgn1-vrrp
priority=200
add interface=cgnat-link name=cgn2-vrrp
priority=100 vrid=2

/ip/address
add interface=cgnat-link address=100.64.1.31/24
add interface=cgn1-vrrp address=100.64.1.21/32
add interface=cgn2-vrrp address=100.64.1.22/32
```

CGN2:

```
/interface/vlan
add interface=sfp-sfpplus1 vlan-id=10
name=cgnat-link

/interface/vrrp
add interface=cgnat-link name=cgn1-vrrp
priority=200
add interface=cgnat-link name=cgn2-vrrp
priority=100 vrid=2

/ip/address
add interface=cgnat-link address=100.64.1.32/24
add interface=cgn1-vrrp address=100.64.1.21/32
add interface=cgn2-vrrp address=100.64.1.22/32
```

Configure the Internet link network

CGN1:

/interface/vlan

add interface=sfp-sfpplus1 vlan-id=20 name=internet-link

/ip/address

add interface= internet-link address=192.0.2.31/24

/ip/route

add dst-address=0.0.0.0/0 gateway=192.0.2.1

CGN2:

/interface/vlan

add interface=sfp-sfpplus1 vlan-id=20 name= internet-link

/ip/address

add interface= internet-link address=192.0.2.32/24

/ip/route

add dst-address=0.0.0.0/0 gateway=192.0.2.1

Configure the CGNAT Pool

CGN1:

```
/interface/bridge
add name=cgnat-pool

/ip address
add interface=cgnat-pool
address=198.51.100.5/32
...
add interface=cgnat-pool
address=198.51.100.14/32
```

CGN2:

```
/interface/bridge
add name=cgnat-pool

/ip address
add interface=cgnat-pool
address=198.51.100.5/32
...
add interface=cgnat-pool
address=198.51.100.14/32
```

Internet side BGP

Mikrotik on ROS 7 has four components:

1. IP Firewall Address List
2. Filter Rule
3. Template
4. Connection

IP Firewall Address List

Same on both devices:

```
/ip firewall address-list
add address=198.51.100.10 list=cgn2
add address=198.51.100.11 list=cgn2
add address=198.51.100.12 list=cgn2
add address=198.51.100.13 list=cgn2
add address=198.51.100.14 list=cgn2
add address=198.51.100.5 list=cgn1
add address=198.51.100.6 list=cgn1
add address=198.51.100.7 list=cgn1
add address=198.51.100.8 list=cgn1
add address=198.51.100.9 list=cgn1
```

Adjust connection tracking

Same on both devices:

/ip firewall connection tracking

set tcp-established-timeout=31m

This will cause stale connections to clear much faster than the system default of 1 day. We have found that this allows the scaling of the connections table to handle 7000+ customer services fairly well

Filter Rule

CGN1:

```
/routing filter rule
add chain=eBGP-out disabled=no rule="if (dst in cgn1) {accept}"
add chain=eBGP-out disabled=no rule=\
    "if (dst in cgn2) {\
    \nset bgp-path-prepend 2; accept\
    \n}"
```

CGN2:

```
/routing filter rule
add chain=eBGP-out disabled=no rule="if (dst in cgn2) {accept}"
add chain=eBGP-out disabled=no rule="if (dst in cgn1) {\
    \nset bgp-path-prepend 2; accept\
    \n}"
```

BGP Template

CGN1:

/routing bgp template

add as=64500 disabled=no name=default output.filter-chain=eBGP-out router-id=192.0.2.31 routing-table=main

CGN2:

/routing bgp template

add as=64500 disabled=no name=default output.filter-chain=eBGP-out router-id=192.0.2.32 routing-table=main

BGP Connection

CGN1:

/routing bgp connection

add as=64500 connect=yes disabled=no listen=yes local.address=192.0.2.31 .role=ebgp name=layer3-switch output.filter-chain=eBGP-out .redistribute=connected \

remote.address=192.0.2.1/32 .as=134090 router-id=192.0.2.31 routing-table=main templates=default

CGN2:

/routing bgp connection

add as=64500 connect=yes disabled=no listen=yes local.address=192.0.2.32 .role=ebgp name=layer3-switch \

output.filter-chain=eBGP-out .redistribute=connected remote.address=192.0.2.1/32 .as=134090 router-id=192.0.2.32 \

routing-table=main templates=default

BGP End Result

```
test-sw#show ip bgp vrf internet
BGP routing table information for VRF internet
Router identifier 1.1.1.1, local AS number 134090
Route status codes: s - suppressed, * - valid, > - active, E - ECMP head, e - ECMP
                    S - Stale, c - Contributing to ECMP, b - backup, L - labeled-unicast
                    % - Pending BGP convergence
Origin codes: i - IGP, e - EGP, ? - incomplete
RPKI Origin Validation codes: V - valid, I - invalid, U - unknown
AS Path Attributes: Or-ID - Originator ID, C-LST - Cluster List, LL Nexthop - Link Local Nexthop

          Network                Next Hop              Metric  AIGP       LocPref Weight  Path
 * >      198.51.100.5/32        192.0.2.31            0       -          100     0       64500 i
 *        198.51.100.5/32        192.0.2.32            0       -          100     0       64500 64500 i
 * >      198.51.100.6/32        192.0.2.31            0       -          100     0       64500 i
 *        198.51.100.6/32        192.0.2.32            0       -          100     0       64500 64500 i
 * >      198.51.100.7/32        192.0.2.31            0       -          100     0       64500 i
 *        198.51.100.7/32        192.0.2.32            0       -          100     0       64500 64500 i
 * >      198.51.100.8/32        192.0.2.31            0       -          100     0       64500 i
 *        198.51.100.8/32        192.0.2.32            0       -          100     0       64500 64500 i
 * >      198.51.100.9/32        192.0.2.31            0       -          100     0       64500 i
 *        198.51.100.9/32        192.0.2.32            0       -          100     0       64500 64500 i
 * >      198.51.100.10/32       192.0.2.32            0       -          100     0       64500 i
 *        198.51.100.10/32       192.0.2.31            0       -          100     0       64500 64500 i
 * >      198.51.100.11/32       192.0.2.32            0       -          100     0       64500 i
 *        198.51.100.11/32       192.0.2.31            0       -          100     0       64500 64500 i
 * >      198.51.100.12/32       192.0.2.32            0       -          100     0       64500 i
 *        198.51.100.12/32       192.0.2.31            0       -          100     0       64500 64500 i
 * >      198.51.100.13/32       192.0.2.32            0       -          100     0       64500 i
 *        198.51.100.13/32       192.0.2.31            0       -          100     0       64500 64500 i
 * >      198.51.100.14/32       192.0.2.32            0       -          100     0       64500 i
 *        198.51.100.14/32       192.0.2.31            0       -          100     0       64500 64500 i
```

CGNAT Static Port Allocations Setup

Setup

- We need to setup rules all the rules on both CGNAT1 and CGNAT2 in case of failover
- BNGs will be statically set to the VRRP ip address so that things continue in case of a device failing
  - BNG1 to CGNAT1 (100.64.1.21)
  - BNG2 to CGNAT2 (100.64.1.22)
- We need to build a table for doing lookups for law enforcement

Build mapping table

Excel is your friend...

	A	B	C	D	E	F	G	H
1	BNG	Public IP	Port Start	Port Finis	CGNAT IP			
2	BNG1	198.51.100.5	2000	2249	100.65	0	10	100.65.0.10
3	BNG1	198.51.100.5	2250	2499	100.65	0	11	100.65.0.11
4	BNG1	198.51.100.5	2500	2749	100.65	0	12	100.65.0.12
5	BNG1	198.51.100.5	2750	2999	100.65	0	13	100.65.0.13
6	BNG1	198.51.100.5	3000	3249	100.65	0	14	100.65.0.14
7	BNG1	198.51.100.5	3250	3499	100.65	0	15	100.65.0.15
8	BNG1	198.51.100.5	3500	3749	100.65	0	16	100.65.0.16
9	BNG1	198.51.100.5	3750	3999	100.65	0	17	100.65.0.17

Mikrotik Script

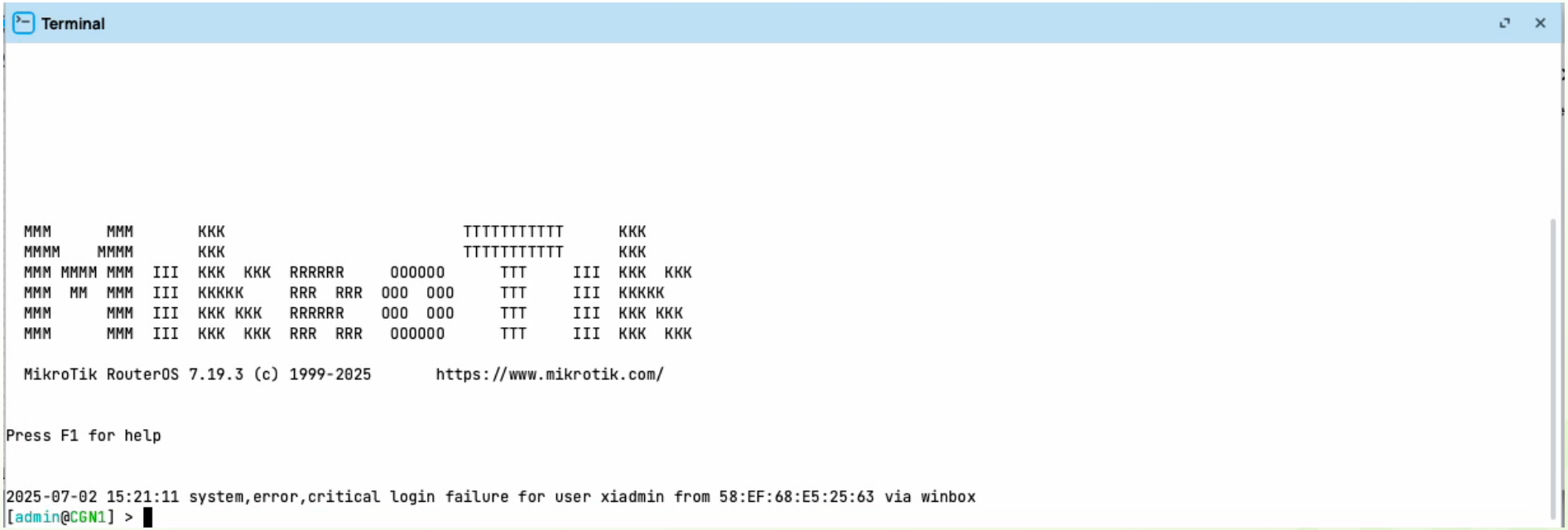
```
:global sqrt do={
  :for i from=0 to=$1 do={
    :if (i * i > $1) do={ :return ($i - 1) }
  }
}

:global addNatRules do={
  /ip firewall nat add chain=srcnat action=jump jump-target=xxx \
    src-address="$($srcStart)-$($srcStart + $count - 1)"
  :local x [$sqrt $count]

  :local y $x
  :if ($x * $x < $count) do={ :set y ($x + 1) }
  :for i from=0 to=$x do={
    /ip firewall nat add chain=xxx action=jump jump-target="xxx-$($i)" \
     src-address="$($srcStart + ($x * $i))-$($srcStart + ($x * ($i + 1) - 1))"
  }

  :for i from=0 to=($count - 1) do={
    :local prange "$($portStart + ($i * $portsPerAddr))-$($portStart + (($i + 1) * $portsPerAddr) - 1)"
    /ip firewall nat add chain="xxx-$($i / $x)" action=src-nat protocol=tcp src-address=($srcStart + $i) \
     to-address=$toAddr to-ports=$prange
    /ip firewall nat add chain="xxx-$($i / $x)" action=src-nat protocol=udp src-address=($srcStart + $i) \
     to-address=$toAddr to-ports=$prange
  }
}
```

https://wiki.mikrotik.com/Manual:IP/Firewall/NAT#Carrier-Grade_NAT_(CGNAT)_or_NAT444

Generate function calls

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	BNG	Public IP	Port Start	Port Finis	CGNAT IP												
2	BNG1	198.51.100.5	2000	2249	100.65	0	10	100.65.0.10		$addNatRules count=240 srcStart=100.65.0.10 toAddr=198.51.100.5 portStart=2000 portsPerAddr=250							
242	BNG1	198.51.100.6	2000	2249	100.65	0	250	100.65.0.250		$addNatRules count=240 srcStart=100.65.0.250 toAddr=198.51.100.6 portStart=2000 portsPerAddr=250							
482	BNG1	198.51.100.7	2000	2249	100.65	1	234	100.65.1.234		$addNatRules count=240 srcStart=100.65.1.234 toAddr=198.51.100.7 portStart=2000 portsPerAddr=250							
722	BNG1	198.51.100.8	2000	2249	100.65	2	218	100.65.2.218		$addNatRules count=240 srcStart=100.65.2.218 toAddr=198.51.100.8 portStart=2000 portsPerAddr=250							
962	BNG1	198.51.100.9	2000	2249	100.65	3	202	100.65.3.202		$addNatRules count=240 srcStart=100.65.3.202 toAddr=198.51.100.9 portStart=2000 portsPerAddr=250							
1202	BNG2	198.51.100.10	2000	2249	100.66	0	10	100.66.0.10		$addNatRules count=240 srcStart=100.66.0.10 toAddr=198.51.100.10 portStart=2000 portsPerAddr=250							
1442	BNG2	198.51.100.11	2000	2249	100.66	0	250	100.66.0.250		$addNatRules count=240 srcStart=100.66.0.250 toAddr=198.51.100.11 portStart=2000 portsPerAddr=250							
1682	BNG2	198.51.100.12	2000	2249	100.66	1	234	100.66.1.234		$addNatRules count=240 srcStart=100.66.1.234 toAddr=198.51.100.12 portStart=2000 portsPerAddr=250							
1922	BNG2	198.51.100.13	2000	2249	100.66	2	218	100.66.2.218		$addNatRules count=240 srcStart=100.66.2.218 toAddr=198.51.100.13 portStart=2000 portsPerAddr=250							
2162	BNG2	198.51.100.14	2000	2249	100.66	3	202	100.66.3.202		$addNatRules count=240 srcStart=100.66.3.202 toAddr=198.51.100.14 portStart=2000 portsPerAddr=250							
2402	BNG2	198.51.100.15	2000	2249	100.66	4	186	6.4.186		$addNatRules count=240 srcStart=100.66.4.186 toAddr=198.51.100.15 portStart=2000 portsPerAddr=250							

Configure the Mikrotik



ICMP

CGN1:

/ip/firewall/nat

add action=src-nat chain=xxx protocol=icmp to-addresses=198.51.100.5 place-before=0

CGN2:

/ip/firewall/nat

add action=src-nat chain=xxx protocol=icmp to-addresses=198.51.100.10 place-before=0

Fasttrack

CGN1:

/ip firewall filter

add action=fasttrack-connection chain=forward connection-state=established,related hw-offload=yes

add action=accept chain=forward connection-state=established,related

FastTrack	
Without	With
360Mbps	**890Mbps**
CPU 100%	**CPU 86%**
44% CPU on firewall	6% CPU on firewall

* tested on RB2011 with single TCP stream

Testing

Reboot Test

Reboot CGNAT1, how long till CGNAT 2 takes over?

The default VRRP timeout is 1 second x 3

So the unit should detect a failure and take a maximum of 3 seconds to switch over

```
[admin@BNG] > /ping src-address=100.65.0.10 1.1.1.1
  SEQ HOST                                     SIZE TTL TIME       STATUS
    0 1.1.1.1                                    56  63 734us
    1 1.1.1.1                                    56  63 431us
    2 1.1.1.1                                    56  63 395us
    3 1.1.1.1                                    56  63 647us
    4 1.1.1.1                                    56  63 432us
    5 1.1.1.1                                    56  63 428us
    6 1.1.1.1                                    56  63 436us
    7 1.1.1.1                                    56  63 448us
    8 1.1.1.1                                                timeout
    9 1.1.1.1                                                timeout
   10 1.1.1.1                                                timeout
   11 1.1.1.1                                                timeout
   12 1.1.1.1                                    56  63 737us
   13 1.1.1.1                                    56  63 647us
   14 1.1.1.1                                    56  63 622us
   15 1.1.1.1                                    56  63 424us
    sent=16 received=12 packet-loss=25% min-rtt=395us avg-rtt=531us
   max-rtt=737us
```

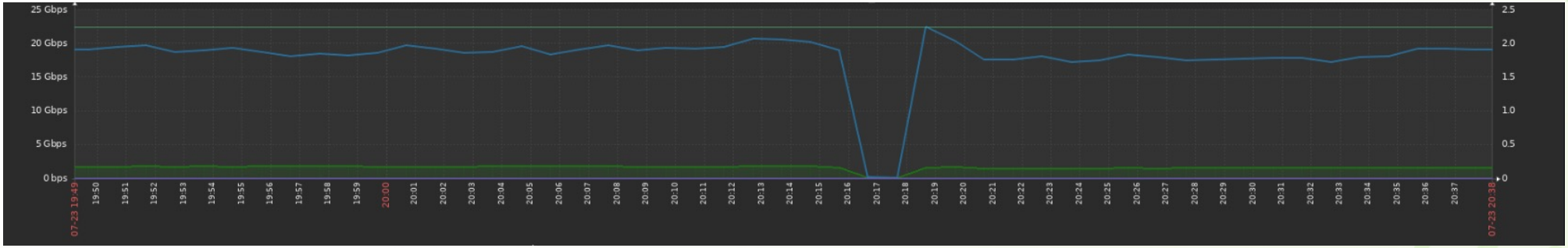
BGP

```
test-sw#show ip bgp vrf internet
BGP routing table information for VRF internet
Router identifier 1.1.1.1, local AS number 134090
Route status codes: s - suppressed, * - valid, > - active, E - ECMP head, e - ECMP
                    S - Stale, c - Contributing to ECMP, b - backup, L - labeled-unicast
                    % - Pending BGP convergence
Origin codes: i - IGP, e - EGP, ? - incomplete
RPKI Origin Validation codes: V - valid, I - invalid, U - unknown
AS Path Attributes: Or-ID - Originator ID, C-LST - Cluster List, LL Nexthop - Link Local Nexthop

          Network                Next Hop              Metric  AIGP       LocPref Weight  Path
 * >      198.51.100.5/32        192.0.2.32            0       -          100     0       64500 64500 i
 * >      198.51.100.6/32        192.0.2.32            0       -          100     0       64500 64500 i
 * >      198.51.100.7/32        192.0.2.32            0       -          100     0       64500 64500 i
 * >      198.51.100.8/32        192.0.2.32            0       -          100     0       64500 64500 i
 * >      198.51.100.9/32        192.0.2.32            0       -          100     0       64500 64500 i
 * >      198.51.100.10/32       192.0.2.32            0       -          100     0       64500 i
 * >      198.51.100.11/32       192.0.2.32            0       -          100     0       64500 i
 * >      198.51.100.12/32       192.0.2.32            0       -          100     0       64500 i
 * >      198.51.100.13/32       192.0.2.32            0       -          100     0       64500 i
 * >      198.51.100.14/32       192.0.2.32            0       -          100     0       64500 i
```

Connection Tracking Reset

Does it matter?

When a CGNAT appliance cuts over connection tracking resets. Does it matter?



Above was a random reboot of an appliance.

Customer impact was less than 2 minutes without new design. New design this should be down to seconds. (Maybe, QUIC will actually delay this to 30 seconds)

Conclusion

Conclusion

Good

- Cheap solution ($5k AUD per box)
- Quick cutover
- Works very well ☺

Challenging

- Scale limits
- Static mapping (compared to other vendors using dynamic port allocation)
- Will randomly reboot but recovers quickly
- Need to figure out how to distribute on BNG side across appliances

Thank you / Q&A

Any questions?

Contact info: Matthew Enger <m.enger@team.leaptel.com.au>

Thank you

Any questions?

Contact info: Matthew Enger <m.enger@team.leaptel.com.au>