RIPE 76
Marseille, France
14 - 18 May 2018

Towards IPv6 only: A large scale lw4o6 deployment (rfc7596) for broadband users @AS6799

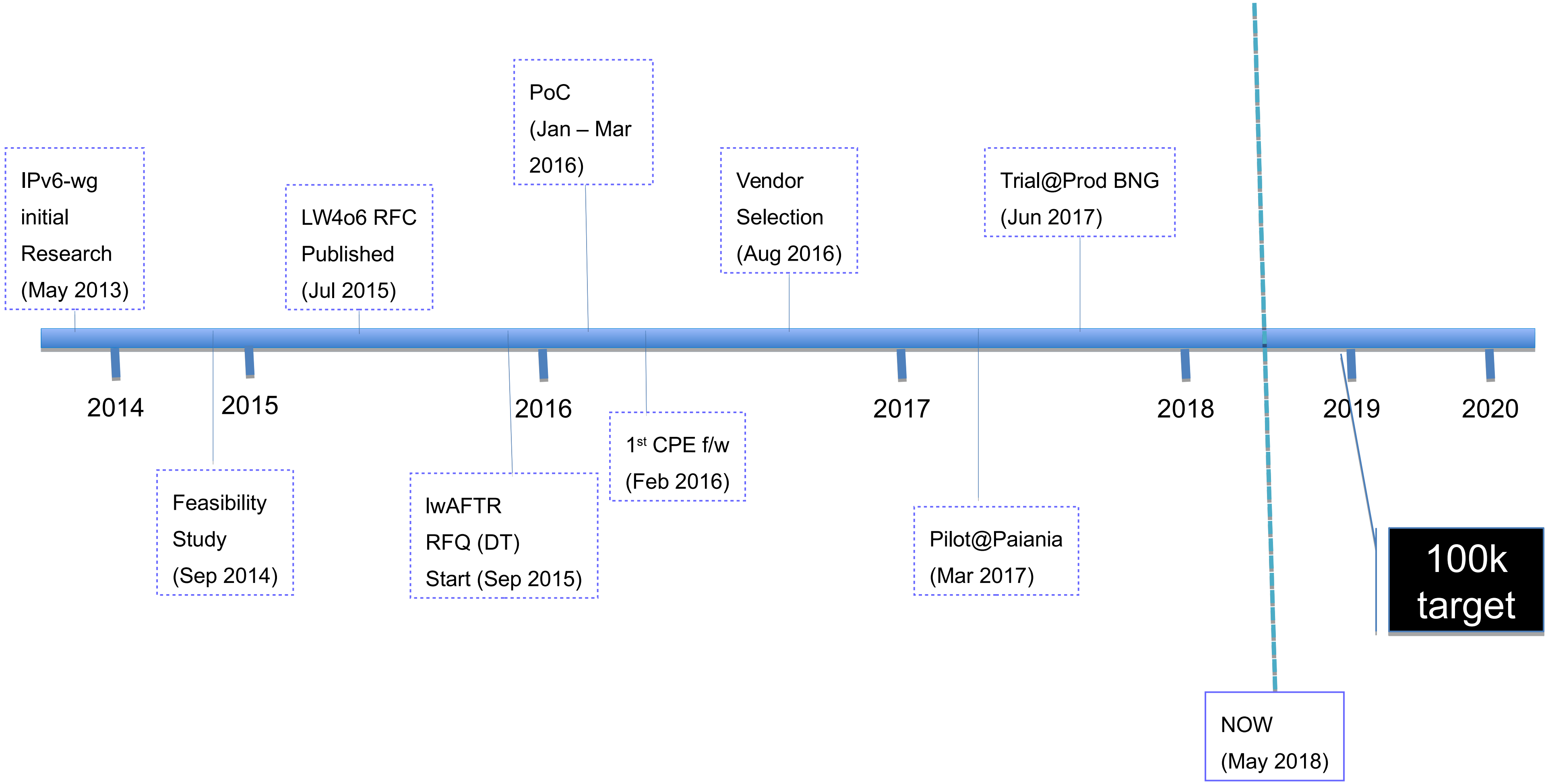
Kostas Zorbadelos
Yannis Nikolopoulos
Senior Systems & Network Engineers

OTE IP Network Engineering

Contents

- Problem Definition
- IPv6-only Service
- lw4o6 overview
- lwAFTR VNF
- Deployment @ Production Network
- Challenges / Experiences
- Future Work

Timeline of events @AS6799



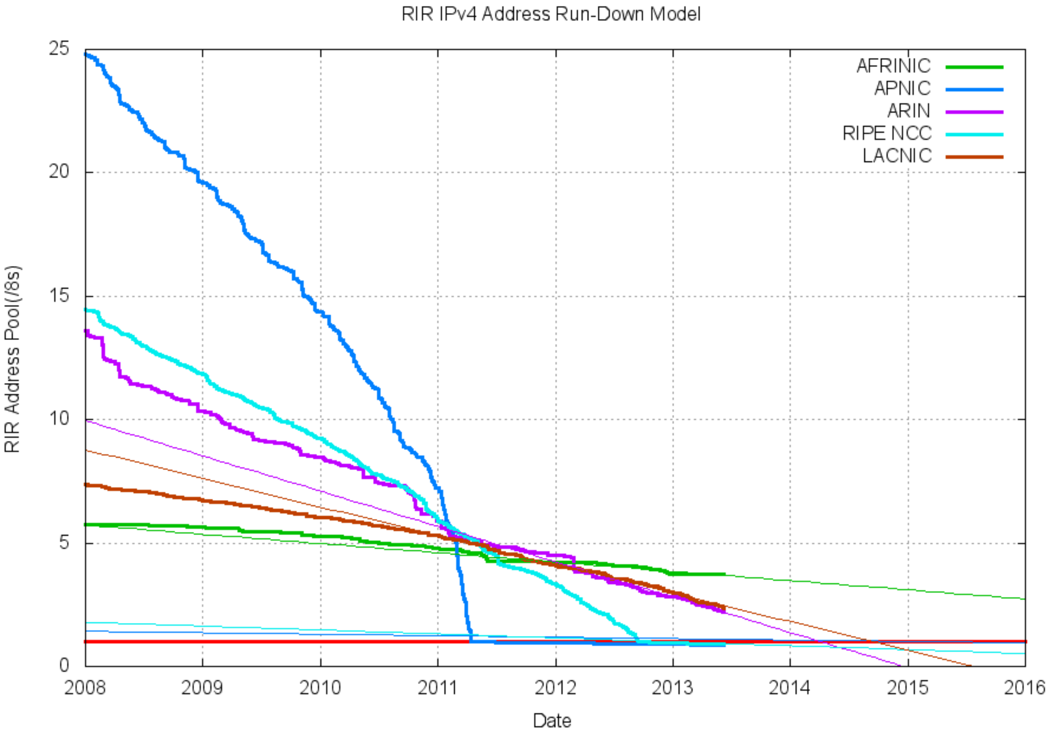
Public IPv4 Exhaustion

Public IPv4 Exhaustion was and still is the main driver

- Around 2012 RIRs really started worrying about the imminent exhaustion

- RIPE began allocation from its last /8 around September 2012

- At the time, ~500k addresses were available from OTE's public IPv4 pool

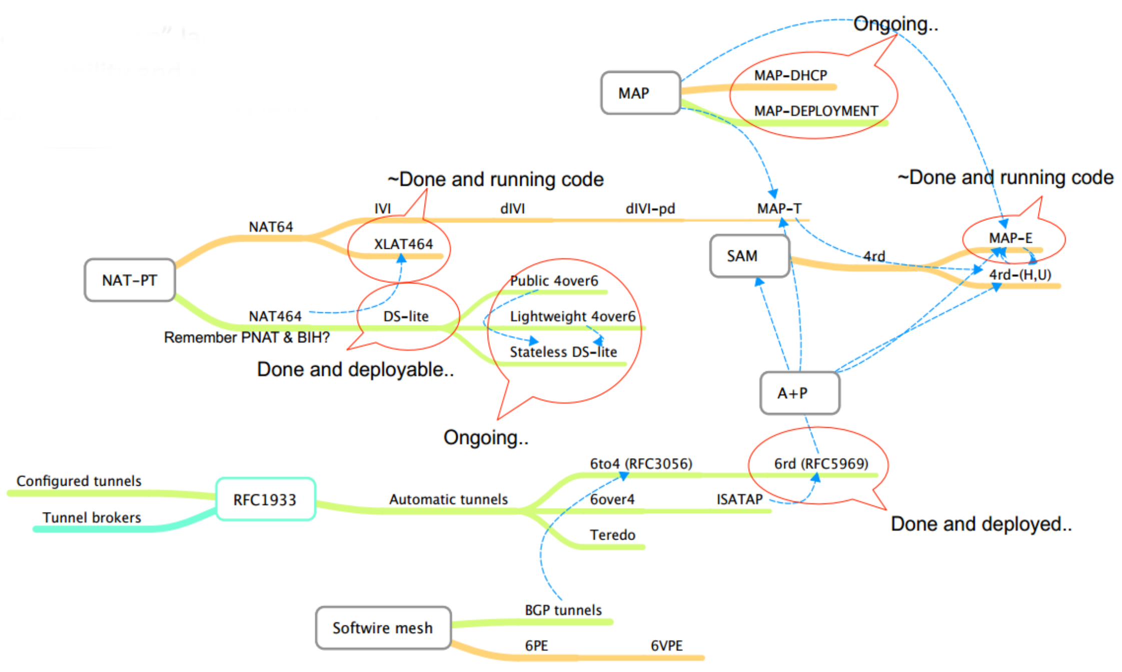


Public IPv4 Exhaustion – cont.

Public IPv4 Exhaustion was and still is the main driver

- OTE started researching possible transition techniques on May 2013, as part of the (now defunct) IPv6 Working Group
- These included:
  - **DS-Lite**
  - NAT64
  - 6rd
  - XLAT464
  - **MAP-(E/T)**
  - **Lightweight 4over6**

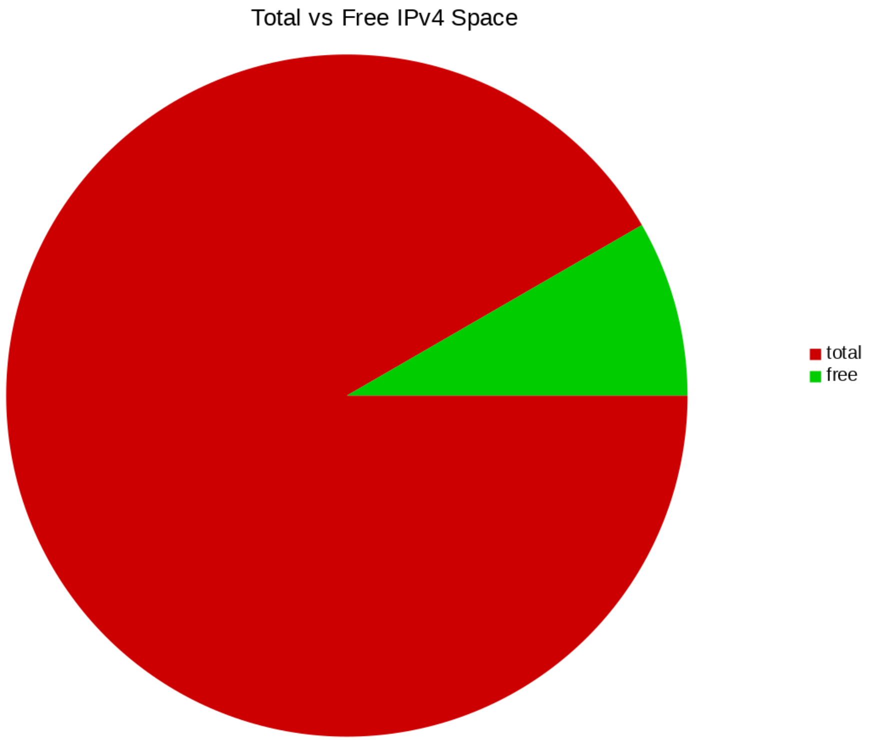
Evolution of IPv6 transition technologies in IETF



Public IPv4 Exhaustion – IPv6-only feasibility

Feasibility Study

- Around September 2014, OTE's public pool was running low (~150k addresses left)
- A feasibility study was produced
- Main proposal was to move forward with an IPv6-only residential service, as a mitigation
- IPv4 was to be treated as a service over the IPv6 network
- If time became of essence, a temporary CGN (NAT444) solution was to be deployed



IPv6-only Service (desired) Main Characteristics

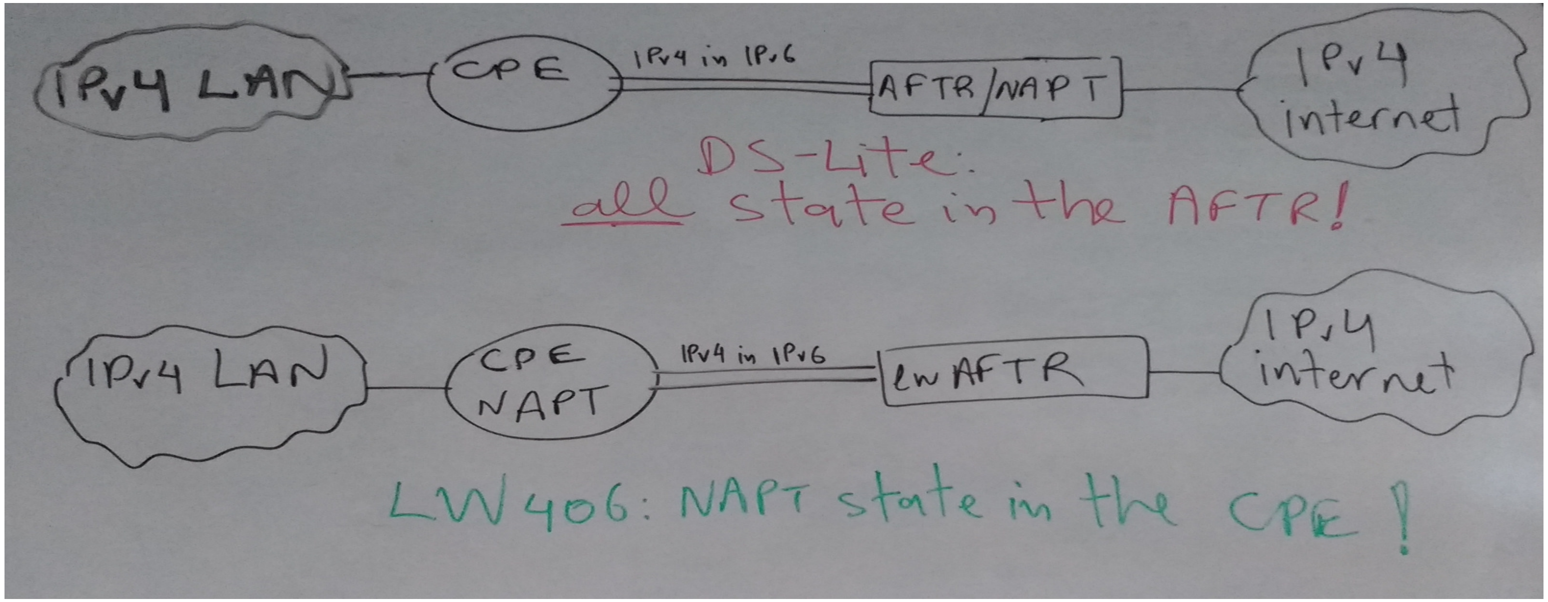
- IPv4 as a service (IPv4aaS) carried over the IPv6 network
- Stateless
- Distributed
- Flexible
- Possibility to completely remove IPv4 in the future
- Minimize Logging (data retention)
- Virtualized (scalability of functions)
- SIMPLE design

MAP and LW4o6 were the main contenders

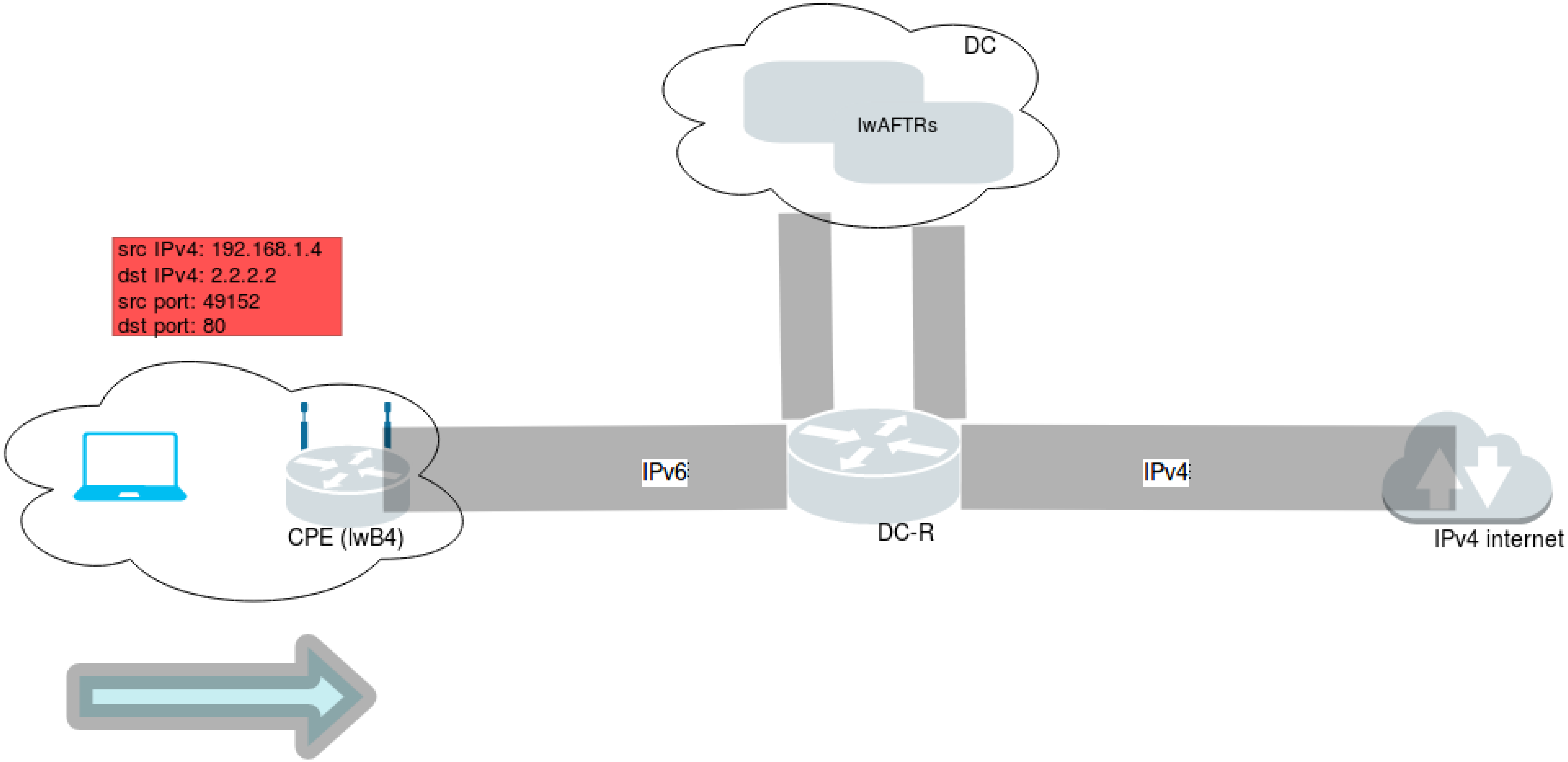
LW4o6 chosen as part of DT focus (Terrastream architecture)

LW4o6 Overview

- rfc7596 - Extension to DS-Lite
- Moves the Network Address and Port Translation (NAPT) function from the (centralized) AFTR to the tunnel client lwB4 (function on the CPE)
- Provisioning of necessary parameters to CPEs via DHCPv6
- stateless nature of lwAFTR
- Encapsulation/decapsulation via a mapping table (shared IPv4 address, port range, lwB4 tunnel endpoint)
- All IPv6 traffic follows the existing native path

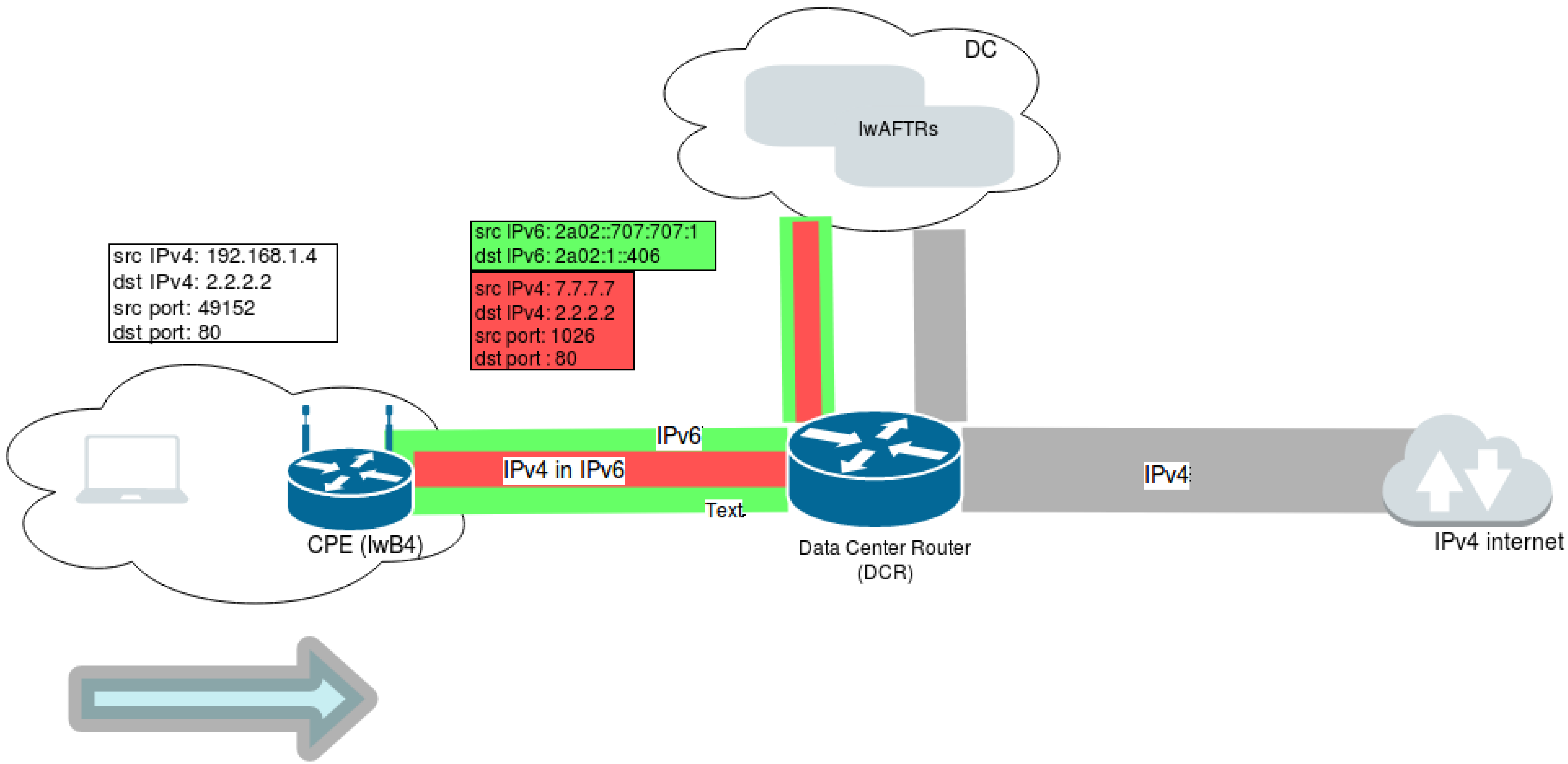


Lw4o6 Overview (cont.)



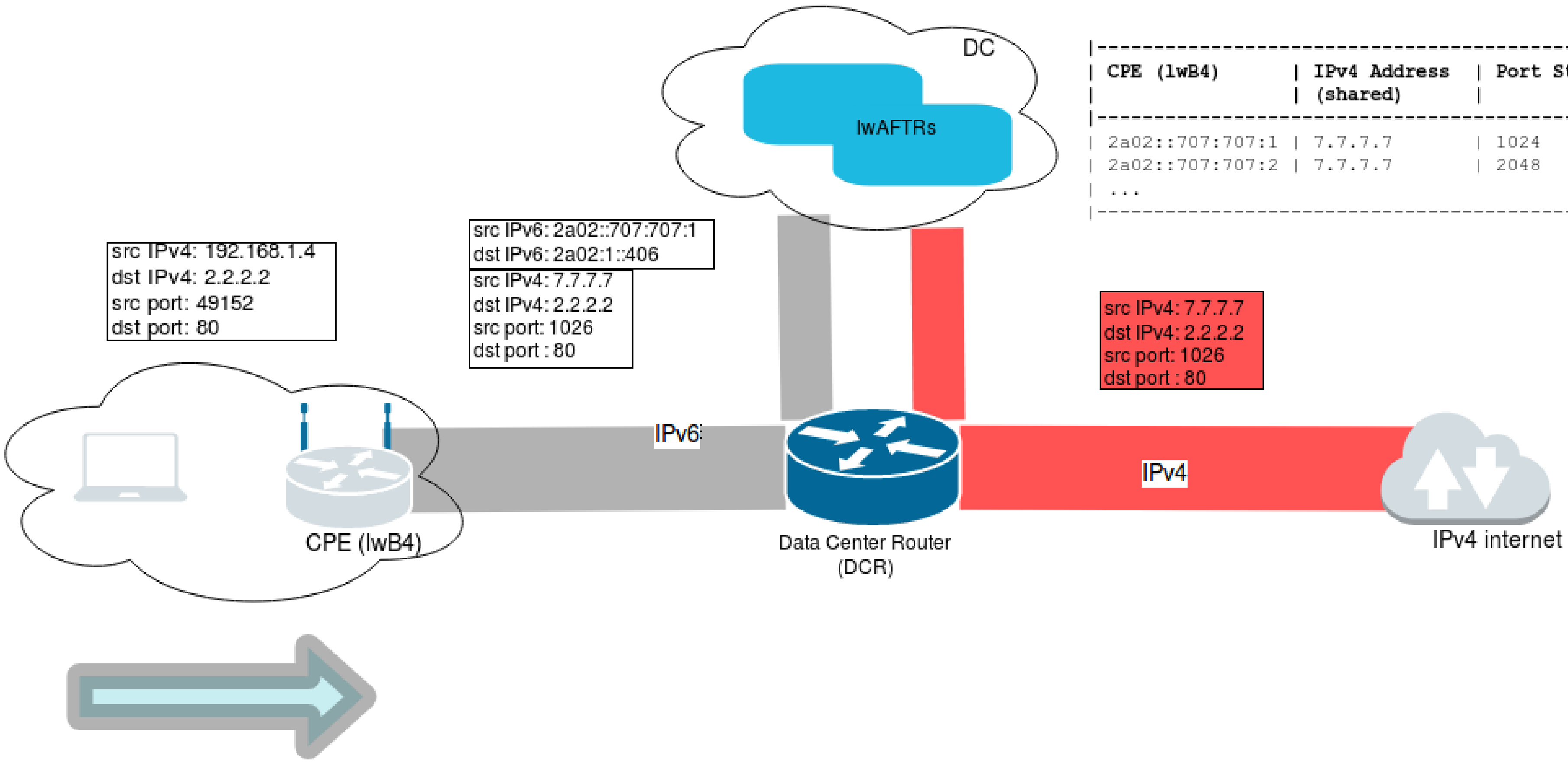
- Main components: lwAFTR, lwB4
- Only Encap/Decap performed at lwAFTR, based on binding table rules

Lw4o6 Overview (cont.)



- Main components: lwAFTR, lwB4
- Only Encap/Decap performed at lwAFTR, based on binding table rules

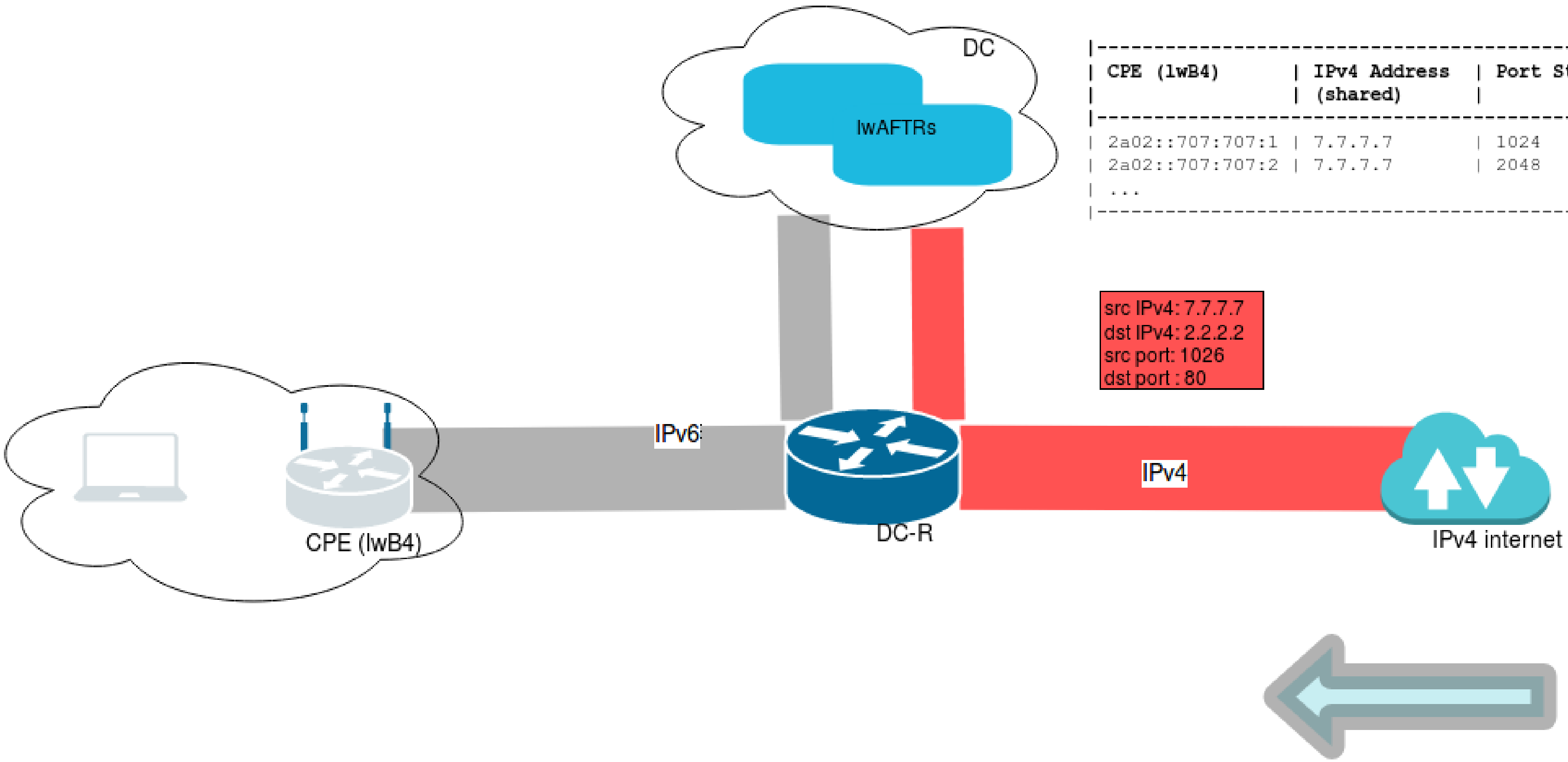
Lw4o6 Overview (cont.)



CPE (lwB4)	IPv4 Address (shared)	Port Start	Port End
2a02::707:707:1	7.7.7.7	1024	2047
2a02::707:707:2	7.7.7.7	2048	3071
...			

- Main components: lwAFTR, lwB4
- Only Encap/Decap performed at lwAFTR, based on binding table rules

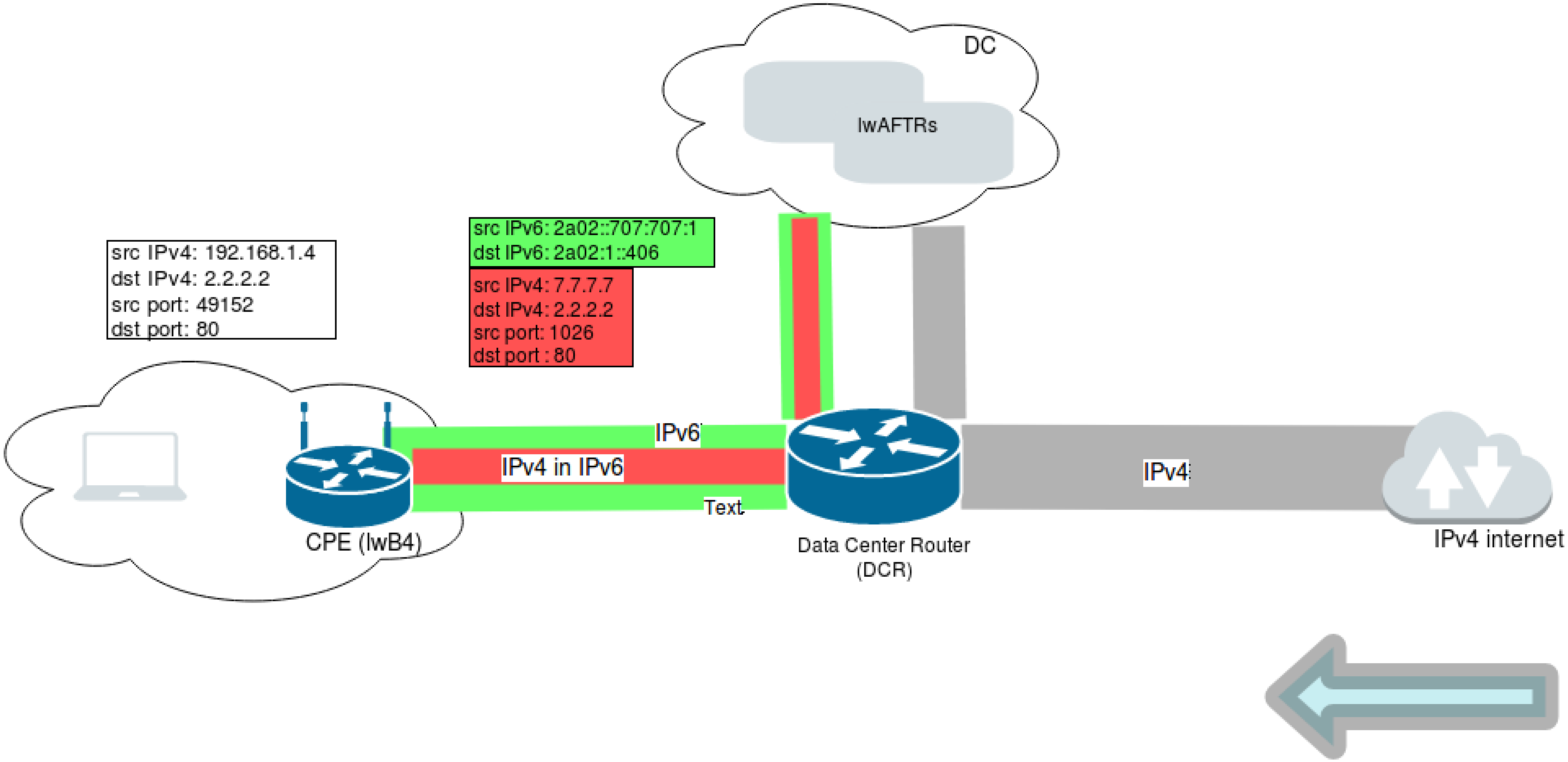
Lw4o6 Overview (cont.)



CPE (lwB4)	IPv4 Address (shared)	Port Start	Port End
2a02::707:707:1	7.7.7.7	1024	2047
2a02::707:707:2	7.7.7.7	2048	3071
...			

- Main components: lwAFTR, lwB4
- Only Encap/Decap performed at lwAFTR, based on binding table rules

Lw4o6 Overview (cont.)

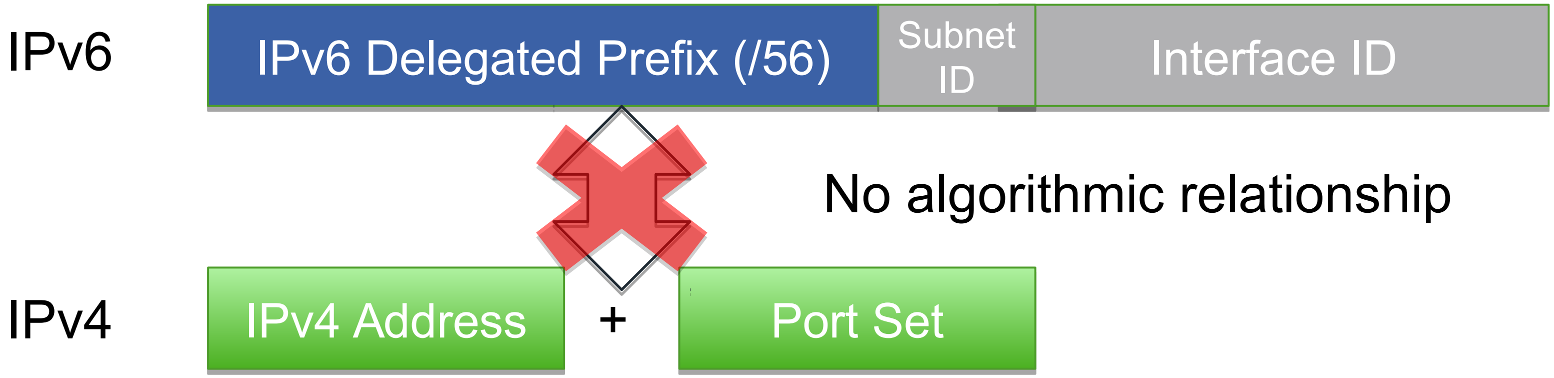


- Main components: lwAFTR, lwB4
- Only Encap/Decap performed at lwAFTR, based on binding table rules

Lw4o6 Overview - Provisioning

Binding Table @AFTR

IPv6 address for a single lwB4	Public IPv4 address	Restricted port set
2a02::707:707:1	7.7.7.7	1024-2047
...	...	...



- No algorithmic relationship between IPv4 and IPv6 – greater flexibility
- Much simpler from a planning/implementation perspective
- Harder provisioning ;-)

CPE (lwB4 function)

- CPE part (as well as provisioning) was excluded from the RFQ process
- LW4o6 support (lwB4 function) had to be implemented
  - Tunnel endpoint
  - Support for (new) DHCPv6 options
- Development was done for a single target CPE
- lwB4 function and LW4o6 support in general, is now mandatory for all (future) company/group CPEs

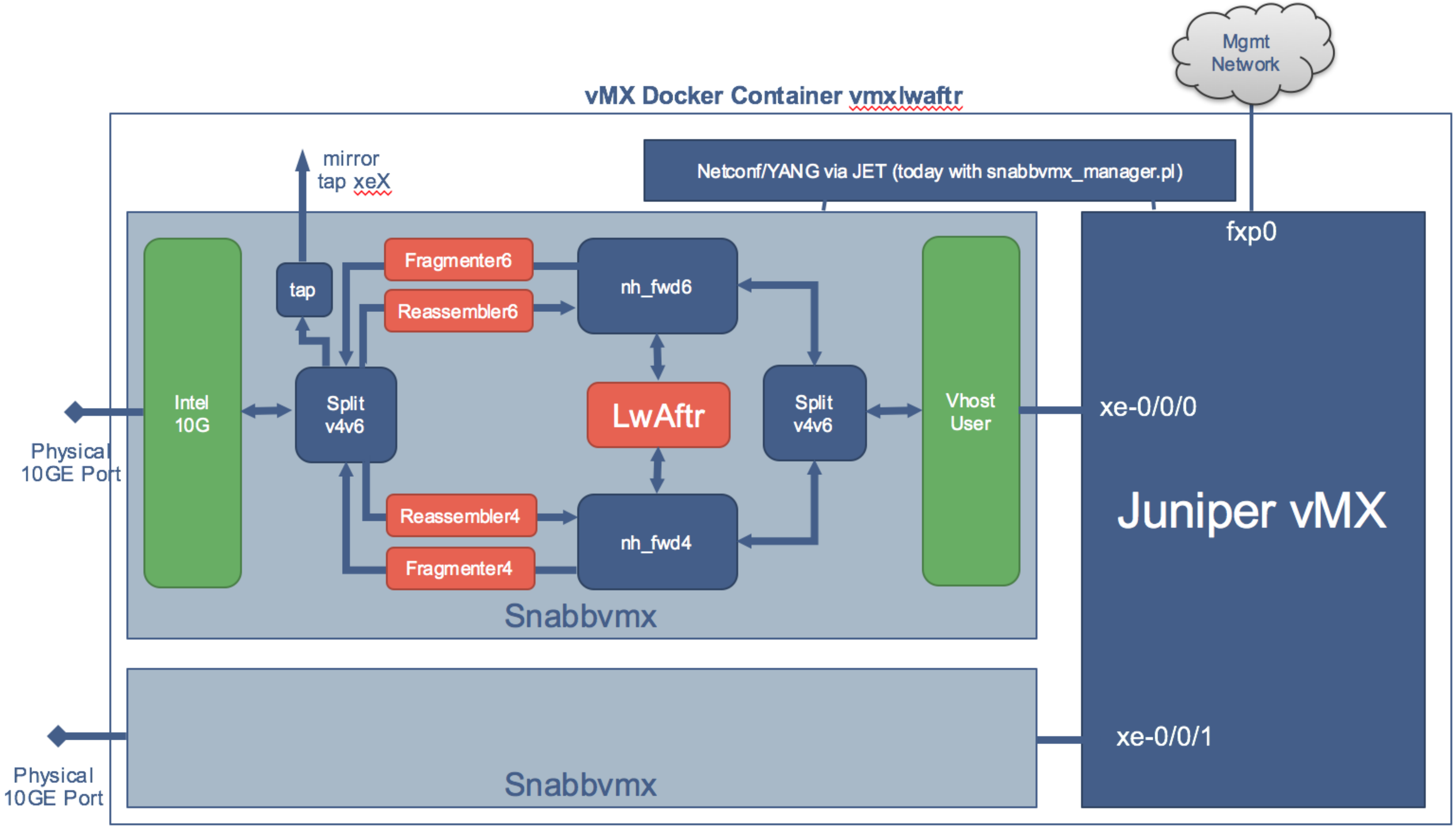
RFQ Process

- Deliverable: *a lwAFTR VNF running on COTS H/W that could accommodate multi-10Gbps traffic with predictable performance*
- 4 Vendors participated
- No solutions were mature, development was done in parallel with the RFQ mostly
- All solutions were tested in OTE labs
- After more than a year, a single vendor was selected

Snabb

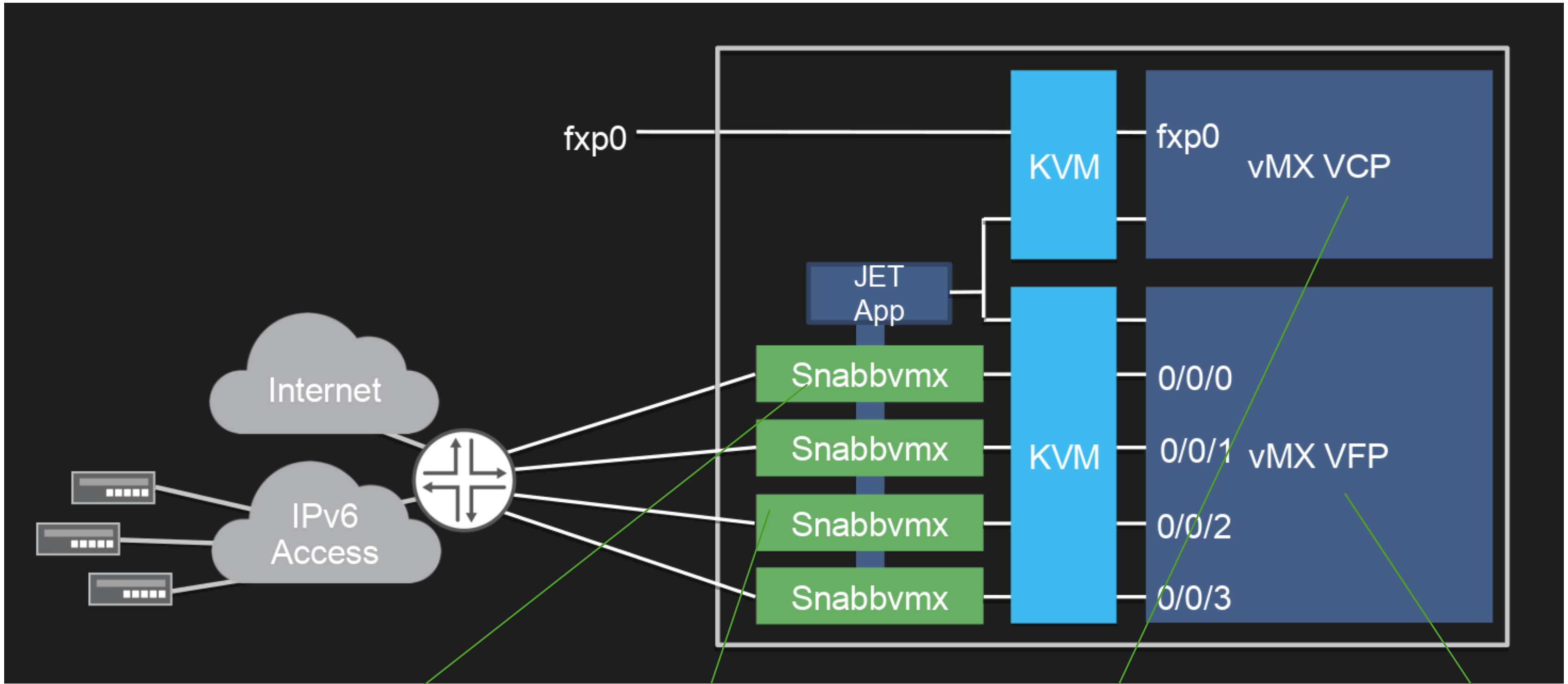
- Snabb (formerly "Snabb Switch") is a simple and fast packet networking toolkit.
- With Snabb you can implement networking applications using the Lua(JIT) language.
- Open Source project, creator Luke Gorrie (has an affiliation with DT (consultant at Terrastream project)
- Ethernet I/O with no kernel overhead ("kernel bypass" mode). Also called user-mode networking
- Can create data-plane applications achieving line rate performance on 10G and beyond

Snabb - VMX



Mgmt Network
vMX Docker Container vmxlwaftr
mirror tap xeX
Netconf/YANG via JET (today with snabbvmx_manager.pl)
fxp0
tap
Fragmenter6
Reassembler6
nh_fwd6
Intel 10G
Split v4v6
LwAftr
Split v4v6
Vhost User
xe-0/0/0
Physical 10GE Port
Reassembler4
nh_fwd4
Fragmenter4
Juniper vMX
Snabbvmx
Physical 10GE Port
Snabbvmx
xe-0/0/1

lwAFTR VNF (Snabb+Juniper) – High Level



encap/decap performed

Separate Snabb process per physical NIC

Separate VM for control plane functions (mgmt, routing etc)

Separate VM for forwarding plane

Packaged as a docker container

lw4o6 deployment - components

- CPE support (focused on a single model, the one with the highest user penetration)
- BNG configuration (for DHCPv6 proxy)
- RADIUS configuration per user
- Central DHCPv6 in DC locations (2 locations in Athens)
- lwAFTR in DC locations (2 locations in Athens)
- Monitoring / measurements
- Provisioning / automation scripts

BNG

- Cisco equipment
- Target 2 specific IOS-XR versions (current and future deployment)
- Communicated and requested missing / desired DHCPv6 features to vendor
- DHCPv6 server or proxy functionality per PPP class of users, end-user identification in DHCPv6 messages

  [Interface-Id (option 18 , rfc 3315), Relay Agent Remote-ID (option 37, rfc 4649), Relay Agent Subscriber-ID (option 38, rfc 4580) and Client Link-Layer Address (Option 79, rfc 6939)]

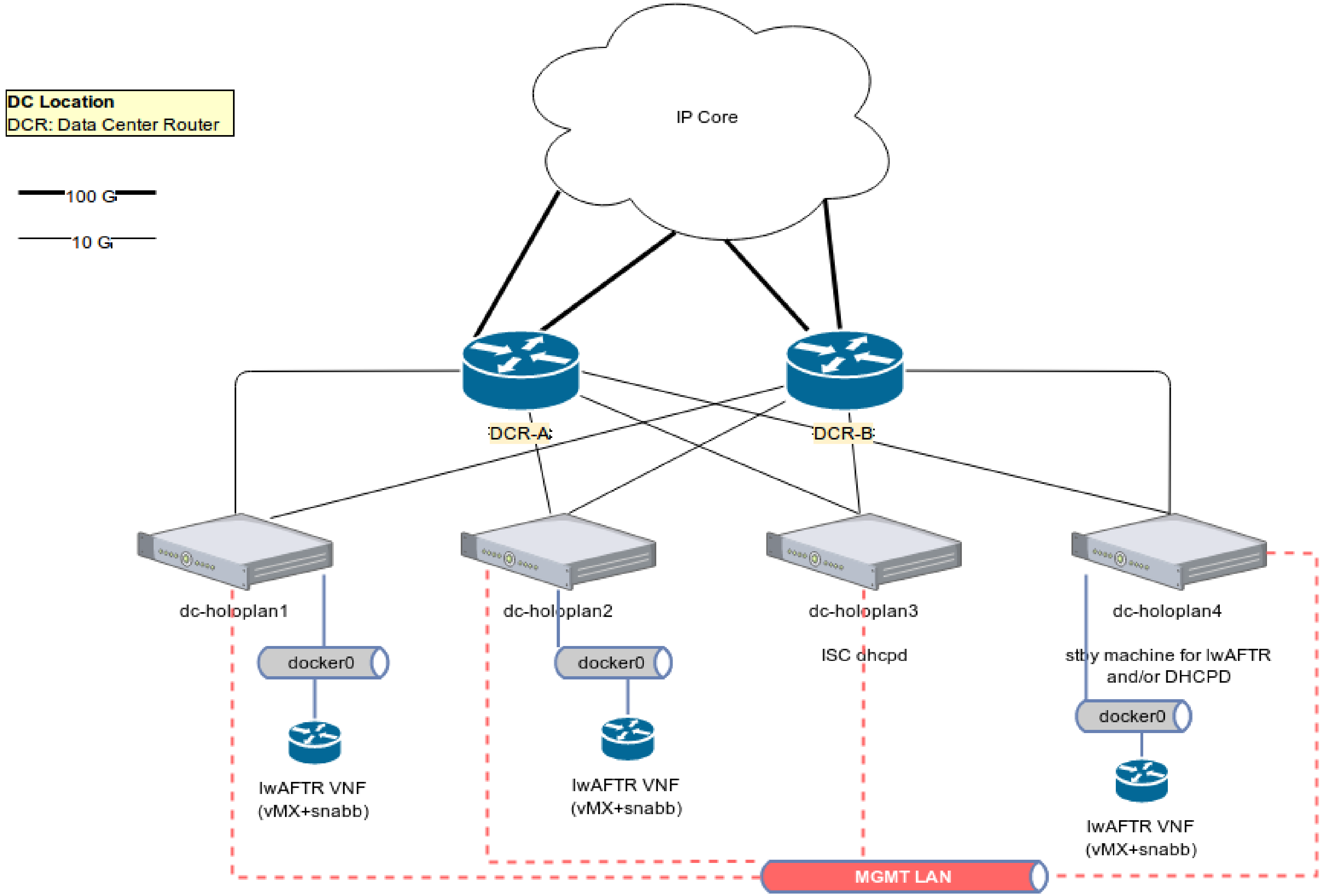
RADIUS

- FreeRADIUS setup with custom config
- Introduced a new RADIUS profile 'lw4o6'
- The profile disables IPv4 during PPP and instructs DHCPv6 on BNG
- Selected users during the provisioning procedure get the profile in their LDAP entry
- A disconnection is required for the new profile to be activated

DHCPv6

- lw4o6 provisioning on the CPE happens over DHCPv6
- DHCPv6 options provide the tunnel endpoint (S46_BR), the public IPv4, tunnel source prefix (S46_V4V6BIND) and port range (S46_PORTPARAMS) all encapsulated in S46_CONT_LW (96)
- BNG acts as DHCPv6 server in the dual stack case, does not support lw4o6 options
- A central DHCPv6 server was introduced for lw4o6 support
- ISC dhcpd initially with 'binary' config generated via script
- High availability using 2 servers in each location

lw4o6 DC location - schema



Service provisioning

- Initial service deployment in a few BNGs
- Custom developed scripts
- Target users activated in batches
- TR069 platform provides an initial report of users with correct CPE
- A script selects a batch (various criteria applied)
- For each user in batch CPE TR069 URL is changed, RADIUS profile activated in LDAP and a PPP disconnection is performed
- Provisioning occurs outside "official" IT flows

Automation

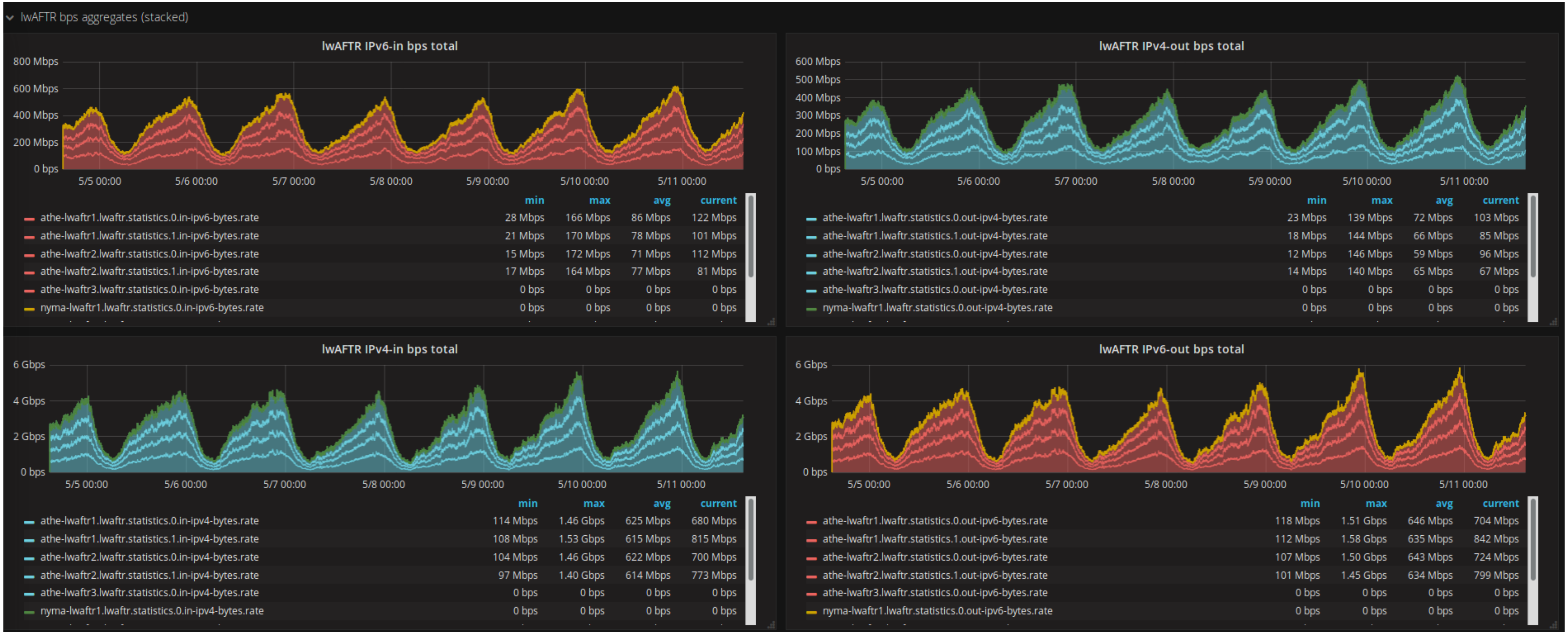
- lwAFTR binding table and DHCPv6 configuration need to match
- Developed scripts to automate the process of config generation
- Binding table upload and commit in lwAFTRs is automated as well
- Web interface provided to help desk to revert user in the original dual stack setup in case of problems

Monitoring / measurements

- Utilized open source open-nti from Juniper for lwAFTRs, contributed patches upstream
- Contains multiple open source components in a docker container (InfluxDB, Grafana, …)
- Added few extra stuff in our nagios infrastructure
- A migration of the extra stuff (DHCPv6 mostly) is underway to a next-gen setup (grafana, prometheus)

Measurements (schema)

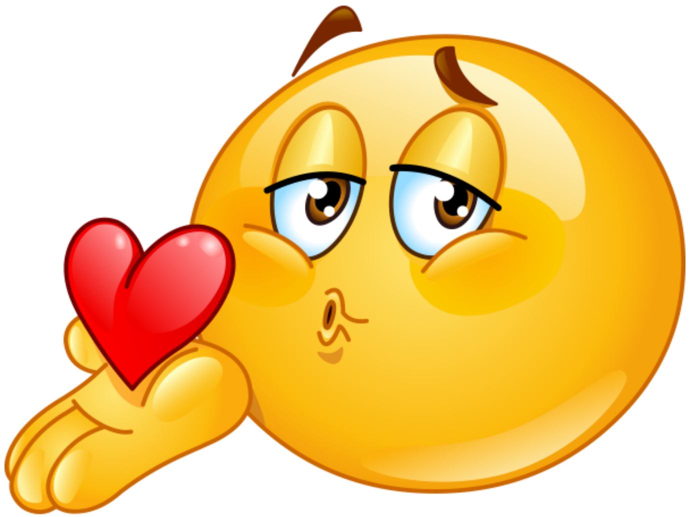
- 12 BNGs in production, ~18500 lw4o6 users, 2 DC locations



<HUMOR>

Problems?

</HUMOR>



Challenges

- Users with services over IPv4 (e.g. IP cameras) [expected]
- CPE dragons
- Selected and not global BNG service deployment causes issues (e.g. user dslam port migration)
- Unknown scalability of current ISC dhcpd deployment
- lwAFTR issues @live traffic [mitigated]
- No possibility for a 'static' IPv6 offering in current setup
- Higher risk due to a forced change to a central DHCPv6 server for all users of the target BNG [old IOS-XR]

Operational Experiences

- CPE Issues (e.g VoIP, DHCPv6,...)
- (IPv4) address sharing issues
- Fragmentation / MTU [no noticable issues]
- No dependency between IPv4-IPv6 addressing (but planning ahead recommended)
- **Flexibility** in routing
  - **anycast** IPv6 AFTR endpoint
  - IPv4 (public ranges)
- Can easily add extra capacity (VNF on-a-stick lwAFTR)

Future work

- Lighten / offload current CGN setup
- Move all BNGs to new IOS-XR version
- Evaluation of ISC Kea for a scalable central DHCPv6 server to overcome current issues with the DHCPv6 infrastructure
- Expand deployment with more users / BNGs
- Improve service provisioning [most work done]
- Promote deployment on international fora / collaborate with other operators

Future work (contd)

- Different classes of service (groups with different port ranges)
- Expose Port range to end users (facilitate port-fwding)
- Provisioning via RADIUS (draft-ietf-softwire-map-radius)
- Configuration/Operation via YANG (draft-ietf-softwire-yang)
- Support Unified IPv4-in-IPv6 CPE (rfc8026)

Questions?

References

- IETF Softwires WG https://www.ietf.org/mailman/listinfo/softwires
- Dual-Stack Lite Broadband Deployments Following IPv4 Exhaustion https://tools.ietf.org/html/rfc6333
- Lightweight 4over6: An Extension to the Dual-Stack Lite Architecture https://tools.ietf.org/html/rfc7596
- https://github.com/snabbco/snabb
- Snabb switch podcasts in Ivan Pepelnjak's Software Gone Wild http://feed.ipspace.net/podcast/software.gone.wild
- Juniper lwAFTR https://github.com/Juniper/vmx-docker-lwaftr
- Juniper open-nti https://github.com/Juniper/open-nti